The University of Sheffield
Department of Automatic Control and Systems Engineering

Data-driven National Strategic Traffic Assignment Models for Road Network Congestion Management

Alexander Roocroft

September 2022

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Declaration

All work presented within this thesis is the author's own work except where specific reference has been made to the work of others.

Research material produced during the course of this PhD has been peer-reviewed or is currently undergoing peer-review for publication. Chapters which contain this material, with the permission of the co-authors, are indicated in Section 1.4.1.

Alexander Roocroft
September 23, 2022

Acknowledgements

Firstly, I would like to thank my supervisors Giuliano Punzo and Azfar Ramli for their continuous guidance and effort during this PhD. It is thanks to their experience and knowledge of research that this thesis was completed. The work of this PhD was partially undertaken at A*STAR research agency in Singapore. I would like to particularly thank them for their assistance in extending my attachment at A*STAR and support through the disruption caused by the COVID-19 pandemic.

Also, I would like to thank Martin Mayfield and Daniel Coca for their additional supervisor input at the beginning of the PhD.

I am grateful to A*STAR and The University of Sheffield for providing the funding for this research. Also, many thanks to MWay Comms and National Highways for supplying access to the traffic data used in this thesis.

Lastly, I would like to thank my friends and family for their advice and encouragement. Especially my parents, for everything they have done.

Abstract

The national management of road congestion is a complex and multi-disciplinary challenge. In previous decades the solution was to build more capacity; however, in densely populated countries this is no longer an easy option due to the cost and environmental issues entailed. Pro-active traffic management is one key to improving the performance of the road infrastructure going into the future, when not only the economic productivity but also the environmental impacts of transport will be under increasing scrutiny. Models to analyse congestion need to be developed that can be effectively applied to large national networks within the constraints of accuracy, efficiency and data-privacy.

This work seeks to investigate how to use the existing cross-sectional traffic data that highway authorities readily have access to for the creation of data-driven traffic assignment models. These models can assess the performance of key national road infrastructure and strategic interventions to reduce congestion. Such data is currently used to reactively manage traffic with action taken after congestion has started.

This work first looks at extracting the building blocks of a data-driven model for the English motorway network. This includes a degenerate topographic representation via map generalisation. Techniques for the estimation of the key components of traffic assignment models are then developed to work with the data restrictions. The use of density-based road-specific congestion functions is proposed and compared to the state of the art to enable the efficient and accurate calculation of traffic patterns. A new technique utilising network modularity community detection is developed that divides the network and estimates the demand profile of its drivers from the measured road flows, reducing the network size restrictions of current approaches. Finally, the developed techniques are applied to a national strategic road network to evaluate network inefficiency from selfish driving and potential targeted intervention strategies.

Contents

List of Figures

List of Tables

Terms and Abbreviations

ANPR Automatic Number Plate Recognition. 37, 42, 85

BiLev Bi-Level optimisation problem. 42, 43, 80, 81, 83

BPR Bureau of Public Roads. 47, 48, 51, 74, 75, 113, 117, 132, 133, 137, 141–143, 147, 148, 150, 153, 157, 159, 183, 189, 190

CAVs Connected Autonomous Vehicles. 58

CBA Cost-Benefit Analysis. 52, 53, 148

DFS Depth First Search. 89, 91, 100

GLS Generalised Least Squares. 42–45, 59, 60, 78, 101, 102, 105, 106, 118, 126, 187

HCM US Highways Capacity Manual. 75, 76

I-VI Inverse Variational Inequality. 138, 139

Inv-Opt Inverse Optimisation. 52, 59, 137, 140, 145, 147, 148, 150, 151, 183, 184

MAPE Mean Absolute Percentage Error. 112

MIDAS Motorway Incident Detection and Automatic Signalling. 16, 18, 20, 22, 23, 38, 39, 53, 57, 60, 85, 86, 92–94, 117, 131, 145, 152, 153, 176, 181, 184, 186–188

NH National Highways. 16, 53, 85

NTIS National Traffic Information Service. 23, 40, 61, 85, 86, 88, 89, 92, 145, 159, 181, 188

O-D Origin-Destination. 17, 18, 20–23, 29, 31–33, 36–45, 51, 55–57, 59, 60, 62, 64, 65, 68–72, 77, 78, 80, 81, 83, 95, 96, 101–104, 115, 117, 118, 126, 129–131, 138, 147, 152, 154–156, 159, 163, 172, 173, 182, 184, 187, 188, 190

POA Price of Anarchy. 32, 54–58, 148, 153–155, 163, 167, 169, 172, 173, 176, 178, 179, 185

RCC Regional Control Centre. 92

SO System-Optimal. 26, 32, 37, 52, 54, 56–58, 65–67, 69, 70, 131, 153–156, 163, 164, 167, 169, 172, 185

SRN Strategic Road Network. 16, 17, 19, 20, 22, 23, 38–40, 48, 57, 59–61, 85, 86, 88, 124, 129, 132, 150, 152, 153, 176, 179, 181–184, 186, 189, 190

TA Traffic Assignment. 17–23, 25, 27–29, 31–40, 42, 46, 47, 50–53, 56, 59, 60, 64, 72, 74, 76, 88, 95, 96, 102, 104, 113, 117, 129–132, 135, 137, 143, 145, 148, 150, 152, 153, 156, 157, 159, 178–186, 188, 190

TAP Traffic Assignment Problem. 21, 31–33, 41, 43, 45, 48, 50, 53, 65, 66, 70–72, 81–84, 109, 117, 130, 137, 138, 140, 145, 147, 150, 157, 188

TMU Traffic Monitoring Units. 85, 187

TSTT Total System Travel Time. 54, 65, 147, 148, 153–156, 158, 159, 163, 167, 169, 172, 173, 176, 178

UE User-Equilibrium. 31, 32, 37, 43, 54, 57, 58, 65–70, 81, 95, 96, 105, 109, 112–114, 117, 123, 132, 137, 138, 140, 145, 147, 150, 153–159, 163, 164, 167, 169, 172, 173, 182, 184, 185

VDT Vehicle Distance Travelled. 153, 154

VI Variational Inequality. 137, 138

VMS Variable Message Signals. 53, 85, 186, 188

Chapter 1

Introduction

1.1 Background

The problem of congestion on road networks

The management of vehicle traffic flows is a complex and multi-disciplinary challenge. The economic losses in the UK due to congestion on roads has been estimated by an INRIX report to be £8 billion in 2021 and this is predicted to rise [1]. Conventionally, the solution to traffic congestion taken by transport planners was to build more roads and increase their capacity. However, due to induced demand effects, where increased capacity attracts new traffic, this is not always effective [2]. Furthermore, in densely populated countries, new construction is often no longer viable due to the cost and environmental issues attached [3, 4].

Many road authorities around the world rely on reactive approaches to managing congestion. For instance, much of the English Strategic Road Network (SRN) is constantly monitored by the Motorway Incident Detection and Automatic Signalling (MIDAS) system [5]. The system uses sensors spaced less then 500m apart to provide information to National Highways (NH) about traffic on the roads. Currently, the system is used in a reactive way with actions taken in response to congestion, using techniques such as varying speed limits and modifying the number of

lanes on affected road sections [6, 7]. According to statistics reported by the UK Office of Rail and Road [8], apart from disruption due to the COVID-19 pandemic, since 2015, average delays on the English SRN have been steadily increasing, a trend that coincides with steadily rising vehicle numbers. This results in a reduction in road users' satisfaction with the network. As such, there is a need to take a broader pro-active approach to road management, anticipating congestion patterns across the entire network rather than just implementing reactive changes on localized sections of road. This requires significant improvements in congestion analysis methods aimed at strategic planning on large-scale national road systems. Pro-active intelligent traffic management is key to improving the performance of the road transport infrastructure in the future, when not only the economic productivity but also the environmental impacts of transport will be under increasing scrutiny.

Approaches to analysing congestion

In the literature of previous research focusing on road traffic there are two main branches of modelling, each with different purposes [9]. One is traffic flow propagation, which looks into the dynamic evolution of traffic flow patterns. This approach is more suitable for traffic engineering applications such as designing traffic light intersections. The other is Traffic Assignment (TA), which is suitable for strategic transport planning. It focuses on the routes drivers take on roads and is used to analyse equilibrium flow patterns for making decisions on network investments such as capacity improvements. It models the road system as a large-scale complex flow network and is capable of the holistic appraisal of congestion at the entire network level. This thesis focuses on the analysis of national road systems using the latter network-based TA models.

Research gap in data-driven traffic assignment models

As will be described in more detail in Chapter 2, current approaches to building data-driven TA models and computing their two key components, which are appropriate congestion functions and estimated Origin-Destination (O-D) demand matrices, are lacking in accuracy and applicability within the context of modelling large-scale national networks. New methods are therefore needed to unlock the potential of existing

direct traffic measurement systems such as MIDAS, which are largely free of privacy restrictions and readily accessible through routine collection by highway authorities. Section 2.3.1 details how alternative traffic data such as travel surveys used for estimating O-D demand matrices have two major drawbacks. Firstly, they entail high costs and require significant time to collate, which means that they cannot easily reflect the latest demand profiles at higher frequency (e.g. annually). Secondly, travel surveys only ever sample a small percentage of actual driver trips, whereas automated data collection systems constantly monitor all traffic passing their sensors.

The generation of full TA models for large scale networks using purely raw cross-sectional sensor data (e.g. loop detector) has mainly only been explored by Zhang *et al.* [10]. However, the authors of this model have identified limitations in accuracy and network size relating to the calculation of its key inputs.

Currently, O-D matrix generation using solely cross-sectional flow data is computationally intractable for large networks. This is due to problems with the number of decision variables used in the process of prior O-D matrix estimation (details are provided in Section 2.4). As such, this thesis proposes a network simplification procedure in order to reduce the problem size by clustering road network nodes together through a modularity-based community detection algorithm. The simplification needs to reduce the computational difficulty of prior O-D estimation whilst maintaining the necessary information needed to build a reasonable model. However, the detailed implementation of O-D matrix estimation on a partitioned network has never been explored before. In Chapter 5, the optimal method of network simplification to produce the O-D input is determined and compared to other developed implementations based on computational requirements and TA model accuracy. This method is capable of reducing a large highway network to a tractable size whilst maintaining reasonable result accuracy.

Section 2.5 details how existing methods of choosing congestion functions are typically too generalized. With the availability of detailed data of actual congested traffic such as that provided by MIDAS, this thesis explains the need to develop road-specific density-based fitting meth-

ods, and tests the approach on an actual large highway network. Previous research has indicated that there are accuracy gains from fitting different congestion functions to individual road segments and using traffic density to fit these functions on the congested regimes. However, it is not clear from the existing literature which mathematical formulations of congestion functions are most suitable for this kind of fitting, particularly in the specific context of national road system analysis. In Chapter 6, the optimal approach to calculating these congestion functions is determined through a full comparative analysis of different congestion function forms and their impact on TA results, balancing accuracy against computational cost.

Currently in the literature (as described in Section 2.6), empirical studies into the effect on network performance due to selfish behaviour from drivers and targeted road interventions have been limited to simulated data, small regional networks or simplified models. Having developed methodological improvements for extracting TA model inputs from cross-sectional data, a full national network TA model for accurate congestion analysis can now be produced. To demonstrate the importance and relevance of the developed models, the loss in performance from selfish routing and the impact of potential congestion reduction strategies are quantified for the case study of the English SRN in Chapter 7. Such analysis illustrates the benefits of the model, in that it allows the analysis of the network at the whole system level as well as at smaller scales such as individual road segments or clusters of multiple neighbouring road segments.

The work in this thesis will use the MIDAS data available on the English SRN in a novel way to develop the appropriate data-driven techniques needed for traffic assignment modelling on large-scale national networks. As detailed in Section 2.3, it is routinely collected by NH without the major privacy issues of other data systems. This results in it having a useful level of accessibility that could provide real utility for transport planners. It is used at the core of this thesis as it is representative of a direct national traffic measurement system that provides cross-sectional data. It does have some limitations relating to the extent of the road system it can provide data for. For instance, its spatial coverage is restricted to the main roads of the SRN, so it does not cover the

large number of non-SRN routing options available to drivers. Also, the loop detectors it relies on can be prone to errors, which can reduce temporal data coverage when sensors are malfunctioning. Despite this, it has sufficient coverage and scale to represent a large contiguous portion of the English SRN, which makes it a good candidate for developing the data-driven techniques for extracting TA model inputs for large national road systems. Such techniques developed with the MIDAS data could then be transferable to national road systems in other countries.

1.2 Research Aim and Objectives

The aim of this thesis is to develop accurate and computationally efficient methods to analyse overall traffic congestion on national road networks solely using cross-sectional sensor data, such as that provided by MIDAS and other loop detector systems.

This aim will be supported by the following objectives:

1. Develop the methodology to create the topographic representation of a strategic road network for TA models and process traffic sensor data compatible with the data available on the English SRN (**Chapter 4**).

2. Develop the methodology to calculate the O-D demand matrix from cross-sectional sensor data alone for national scale networks, which can estimate demand patterns on the topographic representation at different scales (**Chapter 5**).

3. To evaluate, through accuracy and computational cost, different methods of selecting congestion functions on a national network scale derived solely from cross-sectional sensor data (**Chapter 6**).

4. Create an accurate data-driven TA model of the English SRN monitored by MIDAS to analyse how efficient its routing is and the impact of road specific interventions (**Chapter 7**).

The first objective can be classed as data processing and preparation, the second and third objectives are model development, and the fourth objective is model application.

1.3 Thesis Outline and Structure

A brief summary of the main content and results of each chapter is provided below.

Chapter 1 presents the background context to this thesis. It outlines the need for further developments in the tools available for the strategic planning of road systems to reduce congestion. The data-driven approach to traffic assignment is introduced along with the aim and objectives of the thesis relating to its improvement.

In Chapter 2, a literature review is presented to assess the current state of the knowledge in strategic road traffic analysis. The review covers the different types of traffic models and data types currently available. It is found that the current tools for analysing strategic road systems could be improved through a data-driven approach, utilising the data collected routinely by highway authorities. The chapter describes the use of data-driven methods for the generation of TA models for large scale networks using purely cross-sectional data. It explains the limits on result accuracy and size of networks analysed from the techniques used to obtain the key inputs of congestion functions and O-D demand matrices. It discusses current approaches to improving road network performance and how improved TA models can enhance this. The literature review is summarised to frame the needed research to develop data-driven TA for better strategic national planning.

The general methodology for the formulation of a data-driven TA model is outlined in Chapter 3. This includes the mathematical definition of the network used for analysis, the formulation of the Traffic Assignment Problem (TAP) and a description of the algorithm used to provide solutions to it. It presents alternative forms of congestion function and describes the current techniques for obtaining an O-D demand matrix from cross-sectional data. This formulation is used as the base for the developments of the thesis.

The network and traffic data sets of the English SRN are described in Chapter 4. These data sets are used for the real-world empirical experimentation that provides the evidence for the results of the thesis. The chapter describes the techniques used to process the data for use in TA models.

Chapter 5 contains the development of a novel method of obtaining the prior O-D matrix for a large-scale road system solely from cross-sectional data. This is done through utilising network modularity to partition the network into communities, which reduces the computational complexity of the problem. Alternative ways of combining the partitioned prior matrix estimates are investigated, including changing the scale of network analysis. These are tested on artificially-generated networks and the real-world English SRN.

Chapter 6 contains a series of experiments to develop road-specific congestion function fitting. It evaluates, through accuracy, the choices of function form for use with flow-based and density-based fittings on the sample of real-world roads from the English SRN. It then tests the most suitable function for its impact on TA results and compares it to the current state of the art on accuracy and computational cost.

Chapter 7 applies the developed data-driven TA modelling techniques to a large representation of the English SRN covered by the MIDAS system. This is made possible by the results of the preceding chapters. It showcases the utility that the created model has in understanding the strategic analysis of congestion at the national level.

The thesis concludes in Chapter 8 with an overall discussion of the research findings and recommendations for future work.

1.4 Summary of Contributions to Knowledge

This thesis makes a clear contribution to research in transport planning by filling a research gap in practical methods of data-driven model creation for the strategic national analysis of congestion. The novel contributions to knowledge of this thesis correspond directly to the fulfilment of the objectives and can be summarised as follows:

1. A set of techniques are developed which effectively clean and process the National Traffic Information Service (NTIS) and MIDAS data sets of the English SRN to allow such map and sensor data to be used for the purposes of data-driven TA modelling of a national road system. This includes an algorithm for the simplification of the NTIS data set into a degenerate arterial road topographic representation. Also, it includes a procedure for processing MIDAS data to remove problematic measurements and enable its utilisation for the estimation of TA model inputs.

2. A novel method is developed for calculating O-D matrices from flow counts which utilises modularity-based community detection to optimally partition a road network. It enables existing prior O-D estimation techniques to be applied with reasonable accuracy to sizes of networks previously unattainable due to computational requirements, and to be applied at different scales.

3. A density-based, road-specific fitting approach for constructing congestion functions is developed which, when applied to a real-world large SRN, provides improved accuracy for TA model results with much lower computational requirements compared to the state of the art. This makes the proposed approach the best currently available for the analysis of large-scale strategic networks.

4. Utilising the methodological advances of the thesis, an accurate data-driven TA model of the English SRN monitored by MIDAS is produced to illustrate the utility of the developed modelling approach. By applying it, the effects of rerouting drivers and making network changes are quantified for a real-world large-scale road system, including at the level of different network areas and in-

dividual roads. This provides new insight into the merits of the different options available to transport planners for improving network performance. Based on the literature reviewed in this thesis, such an analysis has not been done on a similar road system at this scale before.

1.4.1 Research Output

The research featured in this thesis has also contributed to the following conference and journal publications:

- **Alexander Roocroft**, Muhamad Azfar Ramli, Giuliano Punzo. Efficient Computation of Optimal Traffic Assignment in Nationwide Highway Networks from Raw Loop Detector Data. In: Transportation Research Board 100th Annual Meeting, Washington D.C., 2021.

  Material prepared for this article is featured in Chapters 2, 3, 4, 6.

- **Alexander Roocroft**, Giuliano Punzo, Muhamad Azfar Ramli. Link Count Data-driven Static Traffic Assignment Models Through Network Modularity Partitioning. Transportation, Springer. 2023 (accepted for publication).

  Material prepared for this article is featured in Chapters 2, 3, 4, 5.

- **Alexander Roocroft**, Muhamad Azfar Ramli, Giuliano Punzo. Improved Data-Driven Optimal Traffic Assignment Through Density-Based Road-Specific Congestion Function Estimation. IEEE Access, IEEE. 2023 (under review).

  Material prepared for this article is featured in Chapters 2, 3, 4, 6.

- **Alexander Roocroft**, Giuliano Punzo, Muhamad Azfar Ramli. System Optimal Routing and Distribution of Benefits on National Road Networks. In: 16th World Conference on Transport Research, Montreal, Canada, 2023.

  Material prepared for this article is featured in Chapters 2, 3, 4, 7.

Chapter 2

Review of Literature on Strategic Road Traffic Analysis

2.1 Modelling Road Traffic

Transportation engineering has a rich literature in many different types of models operating at different scales to try to accurately recreate the behaviour of road traffic. For the purposes of strategic planning on national road networks the model types can be split into traffic flow propagation, traffic state estimation and TA models [11].

2.1.1 Flow Propagation Models and Traffic State Estimation

Traffic flow propagation models are typically divided into three main levels depending on the scale of their underlying processes: *micro*-, *meso*- and *macro*-scopic. All models have parameters which need to be calibrated to reduce model errors. The choice of model depends on the type of data available for calibration and the purpose of the model [12].

Microscopic models simulate the behaviour of each vehicle on the road using equations to represent behaviours like driver gap acceptance and lane changing at the vehicle level. However, they require highly per-

sonalised data about the vehicle dynamics and driver behaviour which often limits the accuracy of the model parameters, preventing their application to large scale networks where errors can compound [12].

Mesoscopic models are the intermediate level of description, examples include headway distribution and gas-kinetic [13]. They are able to produce more realistic reproductions of vehicle interactions than macromodels but they need detailed car-tracking data which is not feasible for large-scale national road systems [12].

Macroscopic models are the least detailed and only model the aggregate traffic flow characteristics such as traffic density and average speed. They have the advantage over microscopic models in terms of lower computational requirements while preserving most of the essential traffic behaviour [12]. Since macroscopic models operate with average, aggregate values they can be calibrated with cross-sectional data (e.g. loop detectors, Bluetooth sensors). They are used often at the whole road level to understand how the average flow of the traffic changes over distance and time, showing phenomena like shockwaves for traffic jams [9]. Examples include Daganzo's Cell Transmission Model which uses the kinematic wave theory of traffic flow, treating traffic as a compressible fluid. Macroscopic flow propagation models can be used for looking at traffic behaviour on large networks; however, they are intractable for calculating the System-Optimal (SO) flows in a network analysis [14].

There has been research into expanded macroscopic models at the network level. This includes network fundamental diagrams which describe the relationship between average density in a network and its outflow [15]. These however, being aggregate models of the network as a whole, do not have the detail at the individual road level so cannot analyse each road's influence on the performance of the network.

Traffic state estimation is another established research area of macroscopic modelling. It is the process of estimating the macroscopic traffic variables (i.e. flow, speed or density) on roads from partially observed data based on how they are expected to evolve over time [16]. Such models are used mostly in short-term traffic management, typically extrapolating available data from different sources and combining it [17, 18].

Recent work has looked into applying machine learning and neural networks to forecast predictions [19]. However, the scope of such models does not include analysing the TA patterns at the network-level.

2.1.2 The Network Analysis Approach of Traffic Assignment Models

Traffic Assignment (TA) is a network-level modelling approach aimed at predicting and analysing the distribution of drivers on a road system across their available routes [9]. It does not look into how traffic flows within the roads but in how vehicles flow around the system from a whole network perspective. This approach is largely born from research within the area of transport planning models, with recent additions from computer science, economics and algorithmic game theory [20]. Essentially, it can be thought of as deconstructing complex road systems and re-examining them as a network, like any other large network such as the internet. It has the advantage of being able to analyse large numbers of roads without trying to replicate the minutiae of the complex network, instead merely linking the travel time (or other cost) on a road to the number of vehicles present at that time, and estimating how drivers choose their routes based on the present conditions.

Traffic flow propagation models use simulation over a period of time to see how the traffic evolves at different scales. Instead, the network analysis approach works with equilibria relating to the pattern of assignment, looking into how changes to the network affect the traffic patterns and associated travel costs [9].

The approach has the advantage over aggregate network fundamental diagrams of providing insight into how the individual roads are involved in the overall performance of the network. This allows analysis of how traffic on each road is affected by different changes within the road and elsewhere on the network [21]. Also, this permits investigation of road-specific improvement works (e.g. maintenance, capacity addition, road reconstruction), which will facilitate its use for strategic planning [10].

2.1.3 Graph Theory

The network analysis of roads through TA relies on graph theory. Graph theory is a field of mathematics which underpins system representations of networks. TA models utilise directed graphs (a.k.a digraphs). A directed graph G is defined as a non-empty finite set of nodes (a.k.a vertices) $V(G)$ and a finite family $A(G)$ of ordered pairs of elements of $V(G)$ named edges (a.k.a links, arcs) [22]. $V(G)$ is the node set and $A(G)$ the edge family of G. An edge (v, w) from node v to w is usually written as vw. An example of a directed graph is shown in Figure 2.1 with the node set $V(G)$ as $\{a, b, c, d\}$ and edge set $E(G)$ as $\{ab, ba, ac, ca, bc, cb, bd, db\}$.

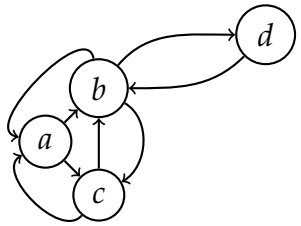


Figure 2.1: *Example of four node directed graph with eight edges.*

An undirected graph is known as a simple graph. A simple graph G is defined as a non-empty finite set of nodes (a.k.a vertices) $V(G)$ and a finite set $E(G)$ of distinct unordered pairs of distinct elements of $V(G)$ named undirected edges [22].

In general, nodes represent entities and edges are the interactions between them. In road systems, graph theory can be used to generalise the topological information of the network. Nodes are often locations or junctions and edges are sections of road. Many techniques at the core of TA, such as shortest path algorithms (e.g. Dijkstra's) [23], utilise graph theory for its ability to analyse complex systems.

2.2 Traffic Assignment Models

Typically, traffic patterns on road networks are modelled using a framework consisting of four-stages, often referred to as the four-step travel

model. The stages are i) trip generation, ii) trip distribution, iii) modal split and iv) route assignment (see Figure 2.2).

The first three steps are concerned with the estimation of travel demand. The first step, trip generation, determines the total number of trips produced from, or attracted to, zones of the network based on population and land-use information. The number of trips is the network's travel demand. It forms the input of the trip distribution. This aims to allocate the trips from individual trip origins to destinations with respect to the overall zonal trip production and attraction from the first step. The third step, modal split, uses the costs of alternative modes of transport to divide the trip demand between the options available. It produces an estimated trip demand for each transport mode between each origin and destination on the network. This is represented by an O-D demand matrix [11]. In the final step, route assignment, the estimated demand is distributed across the available routes between the origins and destinations, using the generalised costs of each route which are often not fixed. Generalised cost can include many different 'expenses' for the driver, from fuel usage to vehicle emissions, most commonly it includes travel time [11]. If the model assumes elastic demand, such that the demand for travel depends on the cost, then the costs for travellers in the last step can alter the choices made in the mode-split and trip distribution steps [11].

The process of the four-step model is shown in Figure 2.2 where the steps are grouped according to their role. TA can sometimes be referred to as the latter three steps working in a loop, which depend on the transportation system. In this thesis, TA only refers to the route assignment stage. The first three steps are not used in the models of this thesis and lie outside its scope. The O-D matrix is instead estimated directly from measured traffic data on the roads only for an assumed single mode of private vehicles. To simplify the analysis, it is assumed that the demand is not elastic and the costs of route assignment do not feedback to trip distribution or mode split. Although this may reduce overall model accuracy, it improves computational requirements and allows analysis to meet the objectives of the thesis.

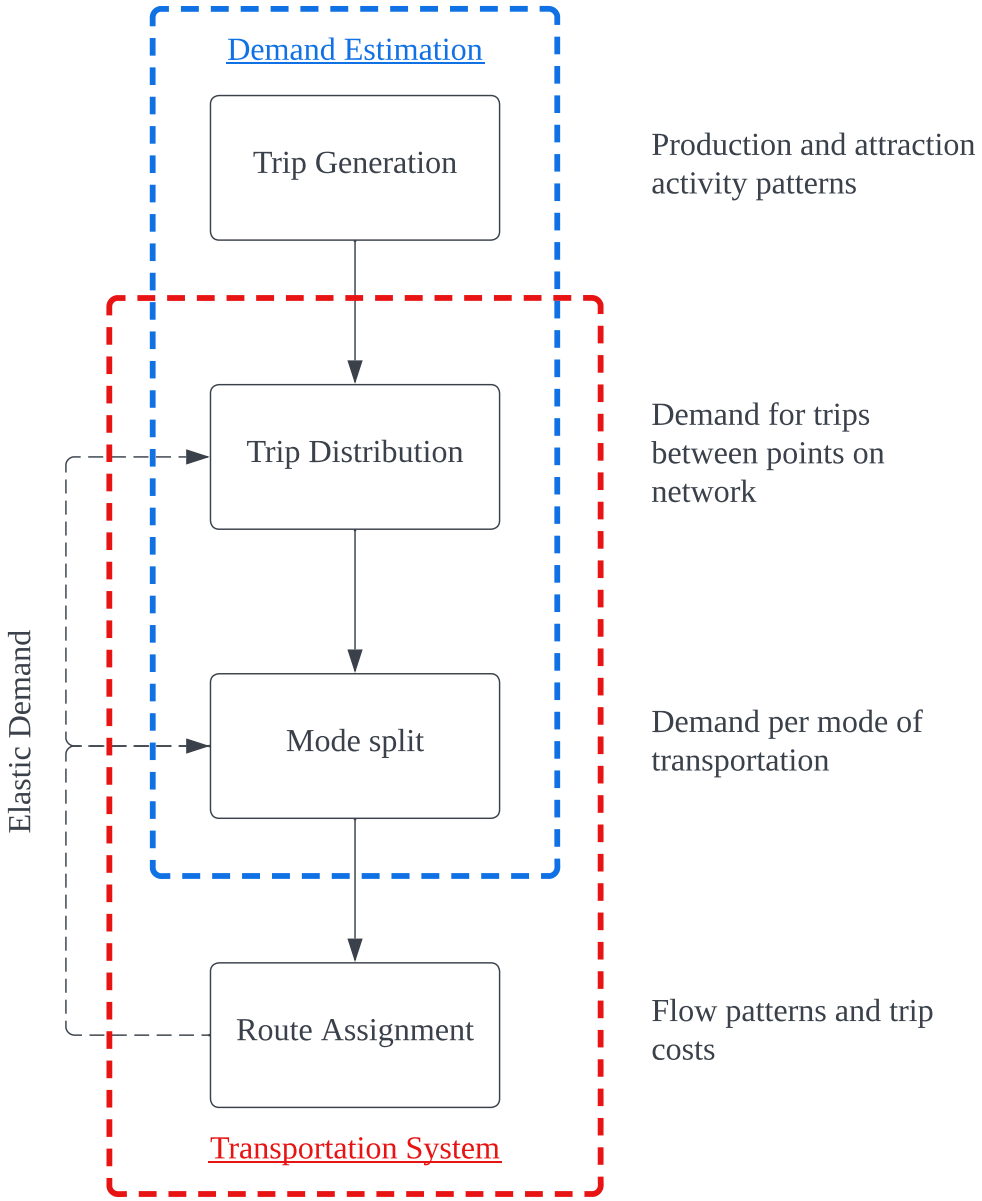


Figure 2.2: *The interaction of the different stages of the conventional four-step travel model.*

2.2.1 The Traffic Assignment Problem

The TAP is concerned with allocating the estimated demand for travel between the O-D pairs onto the routes connecting them so that route flows and costs are in an equilibrium. It is assumed that the drivers choose routes based on their generalised cost and that the cost depends on the number of drivers due to congestion. There are different behavioural rules for the equilibria of the flows and costs.

The first of which, User-Equilibrium (UE), is based on Wardrop's first principle developed in the 1950s [24], which asserts:

"The journey times on all routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route."

This implies, the UE flow pattern is one in which the drivers pursue their individual best route, seeking to minimise their own travel cost without considering the impact on the overall performance of the system. In TA literature it is frequently named as the Wardrop equilibrium. It is similar to a game-theoretic Nash equilibrium of the flows, a flow assignment pattern were no individual driver can improve their cost unilaterally (see [9] for an explanation of the slight difference). This flow pattern is frequently observed in driver behaviour on real-world networks and throughout this work it is assumed that it matches the observed flows as commonly done in other models [25, 10].

The UE pattern assumes that travellers know the costs of the routes available to them. While this may be true if they have assistance from navigation technology, the assumption is not always correct. An alternative assumption is that drivers aim to unilaterally minimise their perceived cost which includes a random error into the cost of routes, leading to a stochastic user-equilibrium [11]. This is outside the scope of the analysis of this thesis.

Another widely used equilibrium pattern is System-Optimal (SO), which is based on Wardrop's second principle [24]:

"The average journey time is a minimum"

This is understood as the global cost of all drivers is minimised through assigning them to the routes which allow them to reach their destination with the lowest overall travel cost. To quantify the difference in cost between UE and SO routing policies and highlight the loss of performance on the network, the Price of Anarchy (POA) is used as a metric to compare the two (Section 2.6.2) [20].

Beckmann *et al.* provided the first mathematical formulation of the TAP as an optimisation problem which was proven to find a unique solution [26]. This mathematical formulation of the TAP is covered in Chapter 3. A TA model is formulated to determine the optimal traffic distribution and associated costs to reflect the behaviour of the drivers, whether UE or SO, through different objective functions in an optimisation formulation. Solving the TAP involves iterating between route choice (when the lowest cost routes are chosen) and network loading (when the route costs are adjusted based on the assigned flows) [11].

2.2.2 Assignment Algorithms

Different methods to solve the optimisation problem for both the UE and SO flow patterns with non-linear cost functions on a network structure have been proposed [27]. Several approaches for solutions have been applied, from heuristic techniques such as all-or-nothing and incremental assignment to more advanced algorithms.

There are three main types of algorithm for solving the static TAP: edge-based, route-based and bush-based [28].

Route-based algorithms (e.g. [29]) use the O-D pair separability of the TAP to output route flows. For each iteration, only the flows of one O-D pair are moved and the others remain fixed. It requires all the routes and route flows to be stored, which can lead to larger memory requirement compared to the edge-based algorithms; however, the convergence to a solution is faster.

Bush-based algorithms (e.g. [30] and [31]) also use the O-D pair separability of the TAP and produce route flows, but further reduce the problem to individual origins instead of the O-D pairs. For each iteration, flows are moved within a directed connected acyclic subnetwork named a bush. Each origin has its own bush representing the paths from that origin to all possible destinations. The approach's fast convergence and improved memory efficiency compared to route-based makes it a superior choice for larger networks. However, bush-based algorithms can be more difficult to implement [29].

Edge-based algorithms were the first of the three types to be developed. For an iteration, flows are moved between edges. They can only produce results for edge flows (not route flows) and their convergence to higher precision results is slower than route-based and bush-based. However, their memory requirements are very low, much lower than route-based and bush-based. The best example of an edge-based algorithm is Frank-Wolfe [32].

First applied to the fixed demand TAP in [33], the Frank-Wolfe algorithm is used frequently in practitioner software [34] and academic studies [35, 36, 37, 38]. Its variations usually differ in how the step-size between iterations are calculated (e.g. Method of Successive Averages) [24]. It is used in this thesis for its ease of implementation and low memory requirements for large networks (Section 3.2.1).

2.2.3 Alternative Traffic Assignment Model Assumptions

There have been many different types of TA models created. TA models are often roughly categorized into static and dynamic models, however, there is more to understanding the classification than just temporal considerations. Models share many features, however, they vary on how they treat spatial assumptions and driver behaviour as well as time assumptions. This is comprehensively explained in the review of TA models in [39].

The review in [39] states the spatial assumptions relate to capacity and storage constraints on the network. In this regard, four model classes exist: unrestrained, capacity-restrained, capacity-constrained, capacity

and storage-constrained. Capacity-constrained models have a hard limit on edge capacity which cannot be surpassed, instead such models reroute traffic to roads with spare capacity or incorporate queues. Storage-constrained models have a maximum storage per edge, leading to excess flow moving to upstream roads in spill-backs if the size of the queue exceeds the storage limit. In capacity-restrained models, flow can exceed the capacity of a road and queues are not explicitly included. Instead the delays of queues are represented through larger travel times at flows surpassing capacity. Unrestrained models are rarely used and have fixed travel times. Within these model classes the assumptions vary further with the choice of fundamental diagram, which describes the relationship between edge flow, density and speed in different steady-states of traffic. Also, models vary on whether delays are added at intersections through turn flow restrictions.

Behavioural assumptions in models can be mostly grouped as one-shot, all-or-nothing or equilibrium [39]. One-shot models do not contain feedback from previous travel time experience, often a single advanced network loading is applied based on route choice proportions from a simpler model. All-or-nothing models have all drivers taking the fastest available route for the current travel times, it is mostly used as a sub-model within equilibrium models. Equilibrium models are the most capable and are as described in Section 2.2.1. The behavioural assumptions can vary further on decision-making with deterministic or stochastic models. These relate to whether the drivers are assumed to have perfect route information and rationality. Also, models can vary with how route cost is updated and considered as drivers complete their trips.

TA models with different time assumptions can be classified into static, semi-dynamic and dynamic. The assumptions consider interactions within time periods, including the speed of propagation of traffic states (i.e. congestion) and the speed of vehicles propagating through the network. Also they consider interactions across time periods, relating to residual traffic transferred for incomplete trips [39].

Static models simplify the travel demand to being constant over a single time period for route choice and network loading, often a multi-hour peak period (e.g. 4-8pm evening commute). They assume traffic outside

the period has no influence, route choices are stationary and all trips are completed within a period. The flows they produce are the average over the period [39].

Semi-dynamic models consider multiple time periods in the analysis, often in one hour time slices (e.g. 6-7am, 7-8am, etc.). They can be thought of as a series of static models which pass residual traffic, such as vehicles in queues, into following periods. They are more able to represent time changes in demand than static models [39].

Dynamic models often consider smaller multiple time periods (e.g. 15 mins) to represent varying demand and route choice with greater time resolution. Route choice is usually stationary for a time period but the network loading often uses simulation models with even smaller time steps to more accurately represent temporal traffic variations. Dynamics can consider day-to-day and within-day effects. The day-to-day adjustment accounts for road users interpreting information about the network and changing their departure times and route choices. The within-day version models the traffic flows as they dynamically develop over a day, based on the drivers' route choices and departure times. There is feedback from the within-day realisation for the day-to-day process, such as when the user decides to depart the next day [34].

The most advanced TA models are ones which spatially constrain capacity and storage, are temporally fully dynamic, and include equilibrium behaviour [39]. Current research is more focused on improving the accuracy of TA models through more realistic approaches with a particular focus on dynamic TA. In cases when the input data is highly accurate and computational times are not limiting (e.g. small networks), it would be better to use more complex dynamic TA if detailed congestion results are required. Many dynamic TA models incorporate the traffic flow propagation models in an iterative procedure. The main issue is that the computing requirements of dynamic assignment models are orders of magnitude larger than static assignment, making them intractable for the national level scale [40]. Whilst complex dynamic models have the advantage of modelling phenomena such as spillbacks across multiple road segments due to extended congestion, in their current state they lack the convergence properties required to be useful in the con-

text of strategic planning [41, 42]. They are more suited for operational management and short-term planning. Their application in large-scale strategic planning remains relatively rare.

The most common type of TA model used for strategic planning is static TA. These are more specifically capacity-restrained equilibrium static TA models. Some key assumptions of these models include: no flow capacity or storage constraints; no turn flow restrictions; no residual traffic transfer between periods; infinite vehicle propagation speeds; perfectly rational drivers with complete information; instantaneous travel time consideration; infinite forward wave speeds and no backward waves [39]. Such limited assumptions restrict the capability of these models and the accuracy of their results in many situations. However, capacity-restrained equilibrium static TA still has many advantages for specific applications compared to more complex models. These include computational efficiency, analytical accountability, mathematical tractability, a greater robustness to errors in input data and easier calibration of model parameters (e.g. static demand profile) [41, 42].

Capacity-restrained equilibrium static TA models are suitable for high-level preliminary screening of strategic interventions to create a shortlist which can be explored with further more exact analysis. They are preferred when the model inputs are uncertain and rapid analysis of a large number of different TAs is needed, such as within strategic intervention planning and O-D demand adjustment. The models are more widely used in planning and are able to estimate different macroscopic flow patterns on large-scale real-world transportation networks with greater efficiency. The analysis of this thesis is aimed at the strategic analysis of large networks so this approach is chosen.

2.2.4 Model Inputs

With the increase in availability of real-world traffic data there has been a trend in research towards utilising it for the TA model inputs [10, 43]. The data-driven approach of this thesis seeks to create a complete network model of a road transportation system.

The three key inputs of a static TA models include: 1) the road network structure being modelled, consisting of road segments (edges) and junctions (nodes); 2) the O-D matrix of traffic demand within the network; 3) the congestion functions of travel delays being assumed for the model. These are used to produce the primary outputs of expected vehicle traffic flows and travel times across each road segment for UE and SO traffic patterns. The estimation of these inputs from data is considered in the following sections.

2.3 Road Traffic Data Sources

A key challenge for data-driven TA models is determining what data to use as input, this can be grouped into traffic data and map data.

2.3.1 Traffic Data

The measurement of real-world traffic is essential to understanding its behaviour. The various types of field data that can be collected have led to improvements in planning and management of the complex systems of transportation networks. Two main categories of field measurements for traffic moving on the roads are floating car data and cross-sectional data [44].

Floating car data provides information on the full trajectories of vehicles on the network and monitors how a vehicle's behavior changes over the length of a journey. New types of collection for this data are emerging, most commonly it is collected from sources such as mobile phone GSM, GPS and Automatic Number Plate Recognition (ANPR) [45, 46, 47]. GPS data collects the coordinates of vehicles at specified time intervals which are then matched to the road network. With less precision, the triangulation of mobile phone GSM signals of vehicle drivers and passengers also estimates the coordinates of vehicles. ANPR tracks vehicles at different points through licence plate identification as they travel across the road network. This requires the installation of overhead gantries which could entail higher installation and maintenance costs.

One key advantage of floating car data over cross-sectional data is that it provides routing information which could improve the estimation of the O-D demand matrix. However, common types of floating car data currently have issues relating to commercial restrictions, privacy and road network integration that limit their accessibility for data-driven modelling [48]. Such issues often limit the number of recorded vehicles to a small sample of the total road users, so the estimation of macroscopic quantities (e.g. flow, density) requires extra modelling.

The methods of obtaining cross-sectional traffic data are more established than floating car data, as the technology used has been around since the early 1960s [4]. The data are obtained from stationary sensors installed at fixed locations along selected roads and measure the vehicles which pass by. They aggregate traffic over set time periods and can provide macroscopic measurements of flow, speed, occupancy and headway at that fixed location.

A common way of collecting cross-sectional traffic data is through the installation of inductive loop detectors. These sensors are frequently installed under the road surface across the highway systems of various countries around the world [49] and they work by detecting the presence of vehicles using electromagnetic induction as they pass overhead. Unfortunately, data obtained from loop detectors can be noisy and error-prone for a range of reasons, from being confused by multiple axles to being damaged by roadworks [50]. Whilst emerging alternative methods to detect vehicle counts such as radar sensors or Bluetooth-based sensing technology are already available [51], loop detectors remain widely used on major road systems worldwide. This includes the MIDAS system of the English SRN which mostly utilises loop detectors [5] (see Chapter 4).

The computation of average traffic measurement values aggregated from loop detector data such as the average speed for the given hour or the average traffic density for a given segment of road may not be truly accurate for use in TA models. This is because measurement systems using loop detectors provide temporal averages of speed which is inexact to use in the calculation of density [52]. Truly accurate space-mean measurements of density are only practically available with aerial pho-

tography [53]. Currently research is looking into the use of aerial drones for recording traffic, which could be an effective future option for these types of measurements [54].

Other types of data used for traffic analysis relate to demand estimation and can be used to complement the lack of routing information within cross-sectional data. An established way of obtaining O-D matrices is through manual surveys of road users, such as household or roadside interviews. However, these can be expensive and time-consuming, which leads to them having low sample rates and frequency. As a result they are prone to high sampling bias risk and missed movements [55]. On closed-system highways, usually where a toll has to be paid to enter the system, the smart card or toll gate information can be used to estimate trip information [43, 56]; however, this does not apply to open-systems such as the English motorway network. Further approaches include utilising zone-based activity and socio-economic data to simulate approximate theoretical demands [57, 58].

Cross-sectional data from loop detectors has its aforementioned drawbacks, however, it is often publicly available [49] and its anonymous nature avoids the privacy concerns that are restrictive for other forms of data collection. Loop detectors are installed and operated by many highway authorities around the world. The accessible nature of the routinely collected data is still, in many ways, an untapped resource for the strategic planning and improvement of road performance. The MIDAS system is a good example of this. With its wide coverage of a large area of the contiguous English SRN, it is amenable to the creation of a data-driven TA model and that is why it is used to deliver the objectives of this thesis.

2.3.2 Network Structure Map Data

To operate at the scale of overall route selection, TA models need a simplified version of the true road network. This simplified network has some of the detail of junctions and roundabouts removed to produce a degenerated arterial road topographic representation. The models are not concerned with the detailed movements of drivers as they navigate

junctions, as for example microscopic simulations would be used to investigate changes to traffic lights.

The map data used to produce the topographic representation can come from sources such as Google Maps and OpenStreetMap. For the case of the English SRN, the NTIS Network and Asset model represents the road network and links it to the different sensors and infrastructure used by the managing authorities. The NTIS model is at a level of detail unnecessarily high for TA models, which limits its ability to analyse the system at large scale with agility. Large national road networks including the English SRN are intractable without a network reduction approach [59].

Automated map generalisation of complex road junctions has been the subject of previous research to reduce and simplify road map data by identifying the nodes and edges associated with interchanges on the road network. The work in [60] first developed a method of using clustering on nodes of a specific order to identify junctions. Other work has expanded this by using extra information such as the angle between the edges to better identify the structures of the characteristic intersections used in clustering [61]. Similar methods are used in this work to process the map data to produce the topographic representation of the TA model.

2.4 Origin-Destination Demand Matrix

Origin-Destination (O-D) demand matrix estimation is a key challenge for static TA models and road transportation planning. O-D demand matrices represent the number of trips taken by drivers between distinct origins and destinations on the road network within a specific analysis time period [62]. As outlined in Section 2.3.1, there are a range of data sources from surveys to mobile phone signals which can be directly used for O-D estimation. When that data is unavailable due to established issues (e.g. privacy, sample size) then the cross-sectional data measured on roads from sensors such as loop detectors can be utilised. However, this does not provide any information on the routes drivers take. Techniques in the literature exist which can use flow counts from cross-

sectional data to estimate O-D demand without the additional need for surveys or historic trip data [63].

Attempting to estimate the O-D matrix solely from mean traffic flows entails problems relating to identifiability [64]. Usually, for a given road network the flows $\mathbf{x}$ on the edges can be expressed as a function of the O-D demand $\mathbf{g}$.

$$\mathbf{x} = assign(\mathbf{g}) \ , \tag{2.1}$$

where *assign* is the process of drivers making route choices through the solution of the TAP. If the demand is not known then $\mathbf{g}$ could be estimated through the inverse of the *assign* process in a kind of reverse of the TAP.

$$\mathbf{g} = assign^{-1}(\mathbf{x}) \ , \tag{2.2}$$

As the number of edge flow counts is almost always less than the number of O-D demand pairs to be estimated, it is difficult to know which vehicles on a road are travelling between which O-D pairs [64]. The equation is undetermined and a unique solution unattainable. In practical terms this means an infinite number of O-D demand patterns could reproduce the observed flows.

Additional information is required, which is provided via a prior matrix, borrowing a concept from Bayesian statistics relating to prior belief in evidence. Sometimes referred to as the target or historic matrix, the prior matrix can be estimated in many ways (e.g. surveys); its purpose is to provide structural information on the demand to restrict the number of solutions to those close to the real O-D matrix.

For cross-sectional traffic data, network tomography-based approaches such as [55, 65, 66, 67] attempt to use the stochastic nature of traffic counts to estimate prior O-D demands using multiple samples of edge flows on the network for the estimation time period. Assuming the Pois-

son distribution of demands and a non-congested network, the Generalised Least Squares (GLS) method as formulated in [55, 64] is a practical version of this approach, which has been applied to real world highway networks in static TA models [10]. Other related flow count techniques are reported to have superior accuracy, however, they require additional data sources such as privacy sensitive ANPR [68, 69, 70]. Although its assumptions may be strong [71], due to its relatively lower computational requirements compared to the other network tomography-based approaches, the GLS method is useful for gaining a prior matrix to be subsequently refined to include the effects of congestion through O-D adjustment algorithms [72, 73].

O-D adjustment methods aim to combine the information of the prior matrix with flow counts on individual roads. This is often used to update demand matrices derived from surveys with more recent measurements of the vehicle flows on the roads, avoiding the expense of a new survey. They are formulated as optimisation problems with an objective function that aims to balance the twin goals of finding a new O-D matrix that matches the observed flows through TA, and not moving the O-D matrix too far from the prior matrix. In current research, there are many alternative approaches to formulating this problem, some with different weightings and distance metrics that affect the balance between objectives, along with alternative algorithms to find the optimal solution [74].

A key classification can be made by distinguishing between methods, with demand-independent and demand-dependent assignment of demand. Earlier attempts at O-D adjustment [75, 76] made the assumption that the assignment matrix mapping demand to flows does not depend on demand. This allows an explicit linear relation between demand and edge flows, typically based on the shortest paths by distance. This makes the problem a convex optimisation with a guaranteed convergence to a global optimum. However, this assumption only holds for non-congested networks.

To account for congestion, the assignment matrix has to depend on the demand which leads to an implicit relationship between flows and demand. The problem becomes a Bi-Level optimisation problem (BiLev), where the upper-level problem seeks the demand matrix that optimises

the objective function, and the lower-level problem solves the TAP for UE flows to obtain the assignment matrix for each iteration of demand estimate. The most widespread algorithms for solving the BiLev are heuristic gradient descent–based [72, 77, 73], where at each iteration an estimate of the gradient of the objective function updates the optimal demand matrix, assuming the assignment obtained through the lower-level is locally constant. Alternative algorithms do not include the locally constant assumption, for instance incorporating Taylor-series expansions, however, they are less feasible for application to large networks sizes [78, 79]. Alternatives to gradient descent such as genetic algorithm approaches are also in use [36]. The genetic algorithms have more flexibility with the choice of objective function, however, they have higher computational requirements than gradient descent [74].

The GLS method of prior matrix estimation and the BiLev gradient descent algorithm are used throughout this thesis, their mathematical formulations are detailed in Section 3.5.

The estimate of the prior matrix is instrumental to the solution of the BiLev; however, GLS and network tomography-based techniques in general have difficulties working with large network sizes due to high space and time complexity in the involved processes. Previous real-world applications of GLS have been limited to 34 node networks with single routes between O-D pairs [10]. Other network tomography-based approaches have been applied to smaller sized road networks [67, 80].

In [10], previous attempts to apply the GLS method to a larger number of nodes have involved zoning and 'landmark' subnetworks. The zoning approach groups nodes within the same geographic area into separate zones. These zones are assigned a single dummy node as their origin or destination to reduce the computational difficulty of the O-D matrix estimation. This approach is limited as, by grouping based on geographic proximity, the zones include nodes which are disconnected and remote from each other in terms of network distance. The approach using the 'landmark' subnetworks estimates the prior matrix from a simplified version of the network with less nodes and edges and uses it as an estimate of the prior matrix for a more detailed network, with the O-D pairs unique to the detailed network set to zero. It was suggested

that prior matrix estimates of a number of different simplified networks could be combined to produce a better prior matrix for the detailed network; however, so far only a single landmark subnetwork of 8 nodes has been used to estimate a 22 node network. There is no exploration into the effect of combining multiple subnetwork estimates on the produced accuracy and computational requirements of the resulting O-D matrix.

In Chapter 5 a novel method is proposed to apply edge flow count O-D estimation to large-scale real-world road networks. This is accomplished by partitioning the network into communities of smaller subnetworks to apply the GLS method to obtain a prior matrix. An analysis is carried out to determine how partitioning a road network into a range of sizes affects accuracy and computational requirements.

2.4.1 Partitioning Network Methods

The partitioning approach employed in this thesis uses community structure detection, which aims to use the information contained within the topology of networks to identify communities and potentially their hierarchical organization. Many networks representing complex systems contain a modular structure where the nodes cluster into communities (i.e. modules or clusters) of relatively high density connections with fewer connections between them [81] (e.g. Figure 2.3). A well-known performance measure to detect such community structure is network modularity. A higher network modularity indicates the nodes have dense connections within their communities and sparse connections to nodes in other communities, indicating stronger community structure and partitions [82].

One of the most widely used algorithms to evaluate community detection with modularity, which is an NP-complete problem [83], is the Louvain algorithm which allows the evaluation of a hierarchy of community partitions to be made [84]. A resolution parameter can determine the size of clusters that are identified. Applied to a road network, this can group areas of the network into clusters which are internally well-connected and externally less strongly. Basing community detection and the resulting partitioning on modularity utilises the network distance and not geographic distance between pairs, which can be dif-

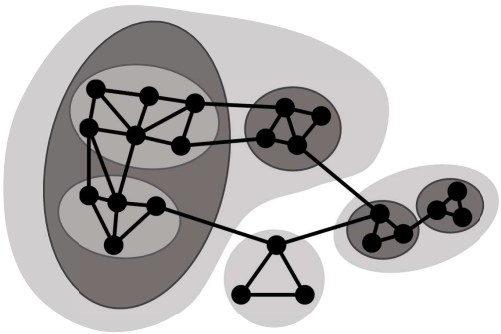


Figure 2.3: *An example of a simple graph, where edge length represents edge weight, with hierarchical communities (shaded areas). Based on example from [81].*

ferent. The grouping of nodes closer together on the network benefits the GLS estimation as the method does not account for geography constraints explicitly [67].

The Louvain algorithm can find an estimate of the optimal community partition with good computational efficiency compared to alternative algorithms such as Fast Newman [85]. Further refinements of the algorithm have been developed to avoid disconnected communities (e.g. Leiden algorithm) and issues with modularity's resolution limit (e.g. constant Potts model) [86]. However, Louvain is sufficient for use on the road networks analysed in this thesis. It has the benefit of providing a hierarchical structure of communities by varying a resolution parameter, which is important for the application to investigating partitioning effects on O-D estimation.

Previously, modularity and Louvain have been used to investigate high-level spatial and temporal patterns in travel demand when the demand is known, finding a strong relation between demand and geographic closeness of O-D pairs [87]. This provides evidence that the structure of travel demand could work with partitioned estimation.

Other works in transport literature have partitioned road networks with different approaches, utilising it for microscopic simulation [88], macroscopic fundamental diagrams [89, 90], parallel static TAP solution [59]

and traffic management through travel speed correlation [91]. It appears previous research has not used partitioning the road network via network modularity for flow-count demand estimation within static TA.

In Chapter 5, the work develops several ways of applying partitioning to the estimation problem. The partitions are used as the basis for reducing the road network down to a smaller, degenerated network with single nodes representing each community. Such a model could be integrated into infrastructure models, including NISMOD in the UK [92], which work at the scale of large urban areas but lack accurate treatment of traffic modelling. The partitions are also used within non-degenerate approaches, which preserve the road network in full but utilise the different scales of analysis, internal and external to the partitions, to estimate a full network demand matrix with increased agility.

2.5 Congestion Functions

Congestion functions are positive, smooth, monotonically increasing functions [93] used to produce flow assignment patterns through the network loading stage of TA. Also referred to as volume-delay functions, accurate congestion functions are key to static TA models as they connect the cost of traversing an edge to the vehicle flows on an edge. Most commonly the cost is travel time which increases slowly at lower flows but, as congestion becomes more significant, increases more rapidly at higher flows [24]. An example of such a function linking the flow to the travel time of a road can be seen in Figure 2.4.

There have been many different candidate forms for the functional shape of congestion functions. They mostly have the same components linking travel time t to the vehicle flows x through a free-flow travel time t^0 and a travel time multiplier. They take the form:

$$t(x) = t^0 f\left(\frac{x}{m}\right) \ , \tag{2.3}$$

where $f(\cdot)$ is the travel time multiplier, a strictly increasing and continuously differentiable function dependent on the saturation rate which is the flow x divided by the flow capacity m of that edge. The definitions

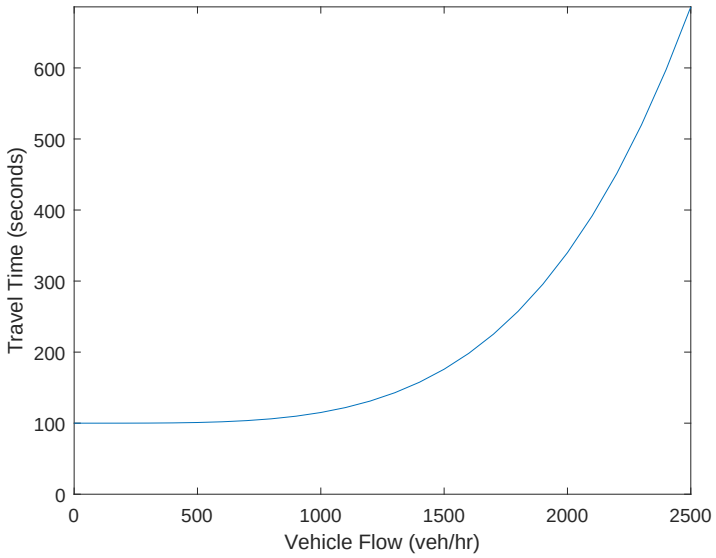


Figure 2.4: *An example of a positive, smooth, monotonically increasing function used as a typical congestion function linking vehicle flow (veh/hr) to travel time (seconds) for a road.*

of flow capacity and free-flow travel time are discussed in Section 3.4.1. The various types of congestion function usually differ in the form of travel time multiplier, which have coefficients that can be calibrated to improve the model.

A key aspect of congestion functions is that the saturation rate can go above 1, such that flow is allowed to exceed the capacity of a road, which by the definition of capacity (i.e. the maximum flow on a road) is not possible. The behaviour of congestion functions is accepted to be not fully realistic, however, for application in large-scale strategic models they have clear advantages (Section 2.2.3). The aim of congestion functions is to reproduce the TA observed on the network and it is understood that they do not directly include phenomena such as bottlenecks [41].

The detailed formulations of the congestion functions utilised in analysis of this thesis are described in Section 3.4. The most common form of congestion function is the Bureau of Public Roads (BPR) formulation, which was developed in the 1960s based on empirical measurements

taken from highways in the USA [94, 95]. Different candidate forms for congestion functions have been developed in the attempt to improve BPR and to better incorporate road characteristics, these include Davidson [96], Conical [93] and Akçelik [97]. However, for long uninterrupted stretches of road such as motorways, the standard simple BPR formulation offers well understood advantages [98]. Its simple, algebraically tractable form, amenable for use in the TAP [93] and its widespread adoption make it a good choice for use on national SRNs such as the English motorway network.

In academic studies, the BPR formulation commonly uses a set of standard coefficients ($\alpha = 0.15$, $\beta = 4$) which were suggested at its inception. However, transport agencies also frequently use their own set of coefficients for congestion functions, either area-wide but fitted to their specific network [99, 100], or based on road type [101]. In countries such as England and the USA, highway authorities recommend using an empirically modified version of the Akçelik function for motorways, which depends on empirical factors relating to road features [101, 102]. However, these generic formulations covering a wide range of situations may not be updated regularly as a significant amount of traffic data is required for proper fitting to be done.

Previous work has attempted to improve the fitting of congestion functions to local conditions network-wide through different approaches such as by using simulated data [103] as well as empirical data [98, 104, 105, 106]. However, these approaches again utilize a single set of network-wide coefficients and therefore do not account for the real-world variability in the flow-delay relationships on roads. Evidence shows that the relationship between flow and travel time in congested conditions is highly dependent on road design [97].

2.5.1 Road-specific and Hyper-critical Traffic Fitting

As a solution to this problem, road-specific fitting has seen the use of uncongested traffic flow and travel time data from automatic detectors [107]. However, this is only viable for the non-congested traffic, where flow and travel time readings increase together. Such traffic flows are called hypo-critical where the critical point is the traffic density thresh-

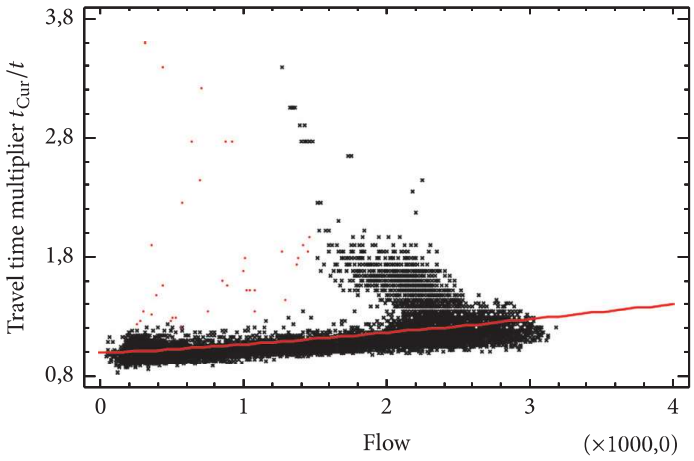


Figure 2.5: *Example of an attempt to fit a congestion function using flow and travel time data for a road. The plot shows hourly flow (veh/hr) vs. travel time multiplier (-) (i.e. non-dimensional time). The red points are identified outlier measurements. The red line is the fitted Bureau of Public Roads (BPR) congestion function, which can be seen to fit the higher travel times poorly. Reproduced from [108].*

old after which travel time increases despite a decrease in flow due to congestion phenomena (e.g. queuing), as such, this approach struggles to represent congested traffic.

An example of an attempt to fit a congestion function using flow and travel time data for a road from [108] can be seen in Figure 2.5. It can be seen that the congestion function fitted has a very shallow gradient and does not fit the data well.

Being defined both below and above the critical flow, congestion functions may offer misleading delay predictions in hyper-critical traffic. To avoid such misunderstanding, the functions can be thought to use *flow demand* to calculate the travel time of a road. This can be interpreted as the number of vehicles wishing to use the road, accepting that whenever the flow demand is greater than capacity, then significant delays are experienced [109].

Fittings biased towards the hypo-critical flow often lead to overestimating travel time and flows for congested conditions and underestimation for non-congested conditions [108]. Moreover, numerical convergence of the TAP is affected by the function's gradient. It has been suggested that the congestion function should be understood as merely a tool for the solution of the TAP as opposed to a reliable tool for estimating travel time [110], which is of paramount importance in testing policies concerned with vehicle speeds (e.g. air quality and emissions, cost-benefit analysis) [111].

More accurate representations of the hyper-critical traffic regime have been obtained estimating the non-observable flow demand from observable measurements, such as flow count [100]. These include incorporating queue-based theory into the estimation [111], using the queues measured by the loop detectors at bottlenecks to provide the number of vehicles surplus to capacity, which can then be used as an approximation of the flow demand to match to the observed travel time on an edge. While promising, this approach proved hard to implement, being difficult to measure traffic flow at every bottleneck point in non-stationary real-world traffic.

Alternatively, traffic density can be taken as a proxy for flow demand. By using a density-based approach to congestion function fitting, the work in [108] introduced a more realistic estimation of congestion functions, which can account for both hypo-critical and hyper-critical traffic regimes. Such a density-based approach leads to a lower mean absolute percentage error in speed prediction than the queue-based approach across different types of road [109].

An example of an attempt to fit a congestion function using density and travel time data for a road from [108] can be seen in Figure 2.6. It can be seen that the congestion function fitted has a better fit than in Figure 2.5 and a shape more similar to that required by a congestion function illustrated in Figure 2.4.

Prominent works on hyper-critical fitting and road-specific functions do not investigate quantifying the impact of their formulations on the TA problem for full real-world networks [107, 108]. Experiments in such

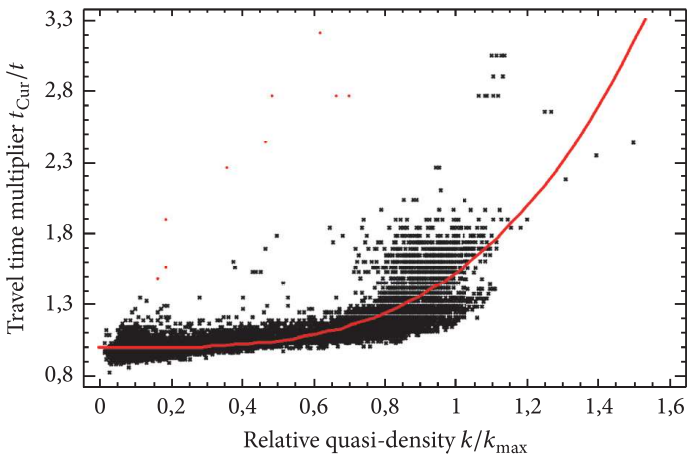


Figure 2.6: *Example of an attempt to fit a congestion function using density and travel time data for a road. The plot shows hourly relative quasi-density (-) vs. travel time multiplier (-) (i.e. both are non-dimensional). The red points are identified outlier measurements. The red line is the fitted Bureau of Public Roads (BPR) congestion function, which can be seen to fit the higher travel times better than Figure 2.5. Reproduced from [108].*

research are limited to a small number of roads or traffic sensors. Notably, the method in [108] is only applied to a single road sensor on a single road and the method in [107] is only applied to 24 road sites. Furthermore, both [107] and [108] only investigate the suitability of the commonly used BPR formulation in their experiments, with which [108] only tested with an incorrect goodness-of-fit metric for the non-linear least squares fitting applied.

An open research question remains in identifying the most accurate choice of function form for density-based fitting and how that compares to flow-based fittings on a range of highways across a large-scale network, with a particular emphasis on their performance within TA. The analysis in Chapter 6 investigates this using different congestion function forms in a data-driven static TA model.

The data-driven TA model used in the analysis (Chapter 3) utilises the investigated congestion functions in the calculation of the traffic flow patterns and within an O-D matrix congestion adjustment. A prominent

example of this family of data-driven approaches is the static TA model used in [10] to calculate flow patterns on the Eastern Massachusetts highway network.

The model in [10] utilises a technique of fitting a congestion function using Inverse Optimisation (Inv-Opt) that fits only a single congestion function to all edges on a network specific to a particular time period of the day. Utilising a computationally intensive optimisation problem formulation, the method is only applied to a small network with 8 nodes and 24 edges.

In progressing the results in [10], the density-based road-specific congestion functions are compared against this state-of-the-art method in a large network TA. The analysis in Chapter 6 compares the effect of using density-based fitting with Inv-Opt on the computation time and accuracy of recreating the observed traffic within TA. The performance informs the suitability for modelling SO flows and use in a fully data-driven model for strategic traffic analysis.

2.6 Improving Road Network Performance

The analysis of real national road network performance is key to understanding the best ways to tackle congestion. In the strategic planning of infrastructure, before transport projects are commenced, economic feasibility analysis is undertaken in the form of transport appraisal. This appraisal can be qualitative in nature, however, most authorities focus on quantitative assessment using tools such as the NISMOD model in the UK [92] and established Cost-Benefit Analysis (CBA). Planners assessing potential projects using CBA often assent to those with highest benefits compared to costs [112]. There are many steps in a CBA, which include collecting together the alternative projects, identifying the various associated costs and benefits, and assigning them a monetary value.

By simulating the changes in traffic patterns and associated delays across the whole network, TA contributes highly to the balance of each project's costs and benefits. For example, UK's Department for Transport include travel time savings, vehicle operating cost and pollution

(i.e. noise, air) in their appraisals. These are based on predicted flows, average speed or travel time taken from TA models [113].

Currently, highway management authorities such as NH in England do not utilise traffic monitoring systems such as MIDAS to inform data-driven TA models for a strategic pro-active approach to analysing congestion. The implementation of intervention methods by NH through MIDAS is currently reactive in its nature, detecting warning signs of traffic delays and intervening to lessen further congestion [6].

The MIDAS system uses a series of algorithms which operate the Variable Message Signals (VMS) when thresholds on flow, speed or occupancy are breached. For example, when the traffic is detected as stationary or slow-moving, reduced speed limits are displayed on upstream VMS to try to try to prevent queues from growing. NH can reduce the speed from 70mph to 60mph with the message '*Queue ahead*' and further down to 40mph with '*Queue caution*' [7]. The use of reduce speed limits also intends to harmonise traffic across lanes and reduce lane changing to help prevent additional accidents following initial incidents.

There is a need for tools which can utilise the existing measurement infrastructure to inform strategies for pro-active, more effective congestion reduction. A data-driven TA model would be able to provide insight into network performance and the costs associated with delays.

The work of this thesis focuses on applying such a model to the national analysis of targeted network changes and routing efficiency. In doing so, it demonstrates how the proposed techniques are scalable and amenable for use in large scale problems.

2.6.1 Targeted Network Improvements

Within CBA, rival investment options can be compared through sensitivity analysis. This can take the form of adapting the TAP to find the network edges which would benefit the most from investment. For example, static TA models have been used in such a sensitivity analysis with road capacity and free-flow travel time [10]. Such analysis aims to

reveal the edges that would most reduce the overall travel time on the network under UE conditions to prioritise edge-specific improvements.

Braess' Paradox is the name given to situations where adding a new road to a network can counter-intuitively increase the average travel time [114, 115, 116]. Research has been carried out to investigate the closure or partial closure of edges in the road network to improve travel times. There is evidence showing that, by using the Braess' Paradox concept in reverse, edges referred to as Inverse Braess can be identified, which cause the average travel time for road users to decrease when they are removed [21].

2.6.2 Routing Efficiency

As an alternative to making physical changes to the network, potential improvements to congestion can be made from directing drivers to use the network more efficiently. To measure the network performance of the equilibria reached by drivers on the roads, researchers have introduced a metric known as the POA.

The concept of the POA was introduced to measure the inefficiency of equilibria in routing on networks [117]. In the context of transport systems where the cost is travel time, the POA is the ratio of the Total System Travel Time (TSTT) from the selfish UE routing to the co-ordinated system optimal routing, assessing the performance of the transport network as it currently operates compared to the best possible overall routing pattern. It is a dimensionless metric equal to or greater than one that quantifies the relative inefficiency that a road system has due to the non-cooperative behaviour of its users. The higher the value, the greater the potential cost savings that could be made from re-routing traffic in a SO way.

POA research was originally focused on application to selfish routing on internet networks, with data transmission replacing vehicle flows [118, 119]. Since its introduction, much research was carried out to develop its theoretical bounds and properties within the computer science and game theory fields [120, 121, 122]. Much theoretical research has

been targeted at understanding the possible POA values depending on the type of congestion functions used. For example, it is proven analytically that the POA cannot exceed 4/3 for linear congestion functions and 2.151 for polynomials of degree 4 (e.g. BPR with standard coefficients: α=0.15, β=4) [123]. In the case of road networks where the free-flow travel time contributes a significant proportion of total journey time, the maximum is proven to be lowered to 1.365 for BPR with standard coefficients [124]. Understanding the factors which influence the magnitude of inefficiency is an ongoing area of research [125]. These include the network topology, traffic demand, marginal costs and complex interactions between flows on routes [117, 126].

A smaller area of POA research has focused on testing the theoretical advances on real transportation networks [126]. Technological developments in traffic data collection have created opportunities to measure the cost of uncoordinated driving on large-scale transportation systems. Different methodological approaches have been used to quantify the empirical static analysis POA using real-world measurements and additional synthetic data.

In [20], the authors investigated the POA on road networks taken from three cities: Boston, New York and London. Using varied synthetic demand between a single O-D pair, they found POA values between 1.24 - 1.30. As these results are for single O-D pairs, they do not consider a realistic demand profile.

Another study applied varying demand levels to reference O-D demand matrices for real-world road networks in the cities of Santiago de Chile, Anaheim and Chicago [25]. It found the POA never exceeded 1.06-1.09 as the β coefficient for BPR was changed (β= 1,4,8). Also, the analysis in [127] found a POA of 1 for inner Beijing and investigated how it varies with changing numbers of O-D pairs loading the network, in an alternative way of varying demand. These values are considerably lower than the maximum theoretical values predicted. However, the investigations into varying demand concur with the numerical study done on the Sioux Falls test network in [126] that POA is low for small and large demands, peaking at mid-range values.

For studies based on real-world measurements, the POA has been estimated to be relatively higher. Using crowd-sensing individual commuter data without a TA model, a POA between 1.11 and 1.22 for Singapore has been obtained [128]. The study attempted to acquire the SO routing via Google Directions API which does not have the analytical accountability of a TA model. Also, the lack of a TA model in this study limits the experimentation that could be done, such as varying demand.

The study in [10] used a data-driven static TA model for the prediction of POA via speed data of cars. It found a POA for the Eastern Massachusetts highway network of up to 2.4, with an average across a month of 1.5. This study had several limitations. The estimated vehicle flows from time-average speed data using Greenshield's model has limited accuracy. Also, the calculated POA compared the modelled SO costs to the costs of the measured flows, which introduced the modelling inaccuracy into the POA values. The study was limited by the size of the network, due the high computational demands in the method used to obtain the O-D matrices and congestion functions. The congestion functions were data-driven but homogeneous for all edges.

The results of static analysis POA studies have a range of results, however, they all have in common considerably lower values than the theoretical bounds. A recent study attempted to apply a dynamic TA micro-simulation model to calculate POA for a small test network and found values on average of 1.6-2.6 [117], closer to the theoretical maxima. The study varied the micro-simulation parameters modelling driver behaviour (e.g. reaction time) and found values up to 3.4. By using micro-simulation the model incorporates phenomena such as spillbacks and blockages which are reported to affect the POA, however, the limitations of dynamic TA restrict its application to smaller networks.

Current research mainly focuses on the travel time cost when assessing the road network performance. However, transport planners have a range of other policy goals to consider, from minimising harmful vehicle emission (e.g. CO_2, NO_x) to reducing fatal accidents [3]. It would be useful to see how routing efficiency varies when the analysis includes these other costs, however, to develop the methods of this thesis the scope is restricted to time.

Furthermore, using a single metric to measure a complex system such as a road network is limited. Using a range of metrics can help to build a better picture of network performance and routing efficiency. There have been developments in alternative metrics to the POA which assess the routing in different ways. One example is Travel Time Index (a.k.a Free-flow Index), used in practice in the USA, which compares measured total travel time to the total travel time under free-flow conditions (i.e. no congestion) [124, 129]. In [129], Regret and Consistency are used to evaluate the inefficiencies of the network at an individual commuter level. Regret measures the time-saving a commuter could have made with perfect traffic routing information. Consistency measures the proportion of users choosing the same daily routes, indicating whether the traffic has reached UE. They both need routing data for individual network users.

Additionally, the POA-delay metric has been developed in [130]. In road networks often the minimum travel cost a driver experiences for their journey is a large proportion of the total cost compared to the additional cost from delays. This has the effect of dwarfing POA values and reducing their insight. The POA-delay metric focuses solely on the delay component of travel cost and how that changes with re-routing. By removing the impact of large minimum travel costs, it provides insight into the delay cost savings that can actually be controlled through re-routing.

To meet Objective 4 of this thesis, it is necessary to expand network routing inefficiency analysis to the national scale of the English SRN. This requires efficient methods for O-D matrix and congestion function estimation as these are limiting factors to network size. Using MIDAS data also brings accuracy benefits as it requires no data conversions (i.e. speed-to-flow). The use of the POA and POA-delay metrics will provide a broader understanding of routing efficiency.

Implementing re-routing

Whilst the implementation of SO routing is outside the scope of this thesis, this section outlines some potential options. Technology relating to vehicle automation and connectivity is currently developing and in

coming years will open up new opportunities to implement system-optimised traffic management.

The progress of fully Connected Autonomous Vehicles (CAVs) is currently not at the level required for optimised flow implementation. However, options now available to implement a SO routing pattern on the network include driver navigation advice (e.g. through mobile phones) and road pricing. Road pricing uses economic levers to influence traffic to a more optimised state. They have the ability to reduce congestion, improve air quality and raise revenue for transport systems [131, 132, 133].

There has been research into the development of using road pricing specifically to reduce the POA. The development of CAVs is leading to more advanced pricing strategies. The effect of different types of pricing schemes (e.g. distance-based, edge-based) are dependent on how CAVs are developed and implemented [134]. Prior to CAVs availability, taking advantage of improved communication technology in vehicles, micro-road pricing has been developed [14, 135, 136]. This has the ability to set individualised and adaptable marginal-cost road price values for every edge in a network. Such prices can be based on the observable traffic conditions to lead UE to align with SO if there is a mechanism to redistribute the collected charges to the users (e.g. via tax reductions).

There are several issues relating to using road pricing to manage congestion. There are technical issues involving estimating the optimal prices, such as various types of road-users having different values of time [137]. Also, the effect the road prices have on the flows in the network takes time to reach an equilibrium state as the drivers adjust to the new costs. Over this time the network demand and infrastructure can change, which requires the optimal road prices to be calculated again. For example, in Singapore the road prices are only updated every three months, which reduces their effectiveness [138]. Evidence from Singapore's cordon pricing scheme shows that the implementation can cause the congestion to be moved elsewhere, such as just outside where and when the charges are applied. Furthermore, there are problems relating to public perceptions of road pricing, including a perceived invasion of privacy and equity for road-users [139, 133].

2.7 Summary

From the comprehensive review of the relevant literature, it can seen that TA modelling is essential to strategic planning on national road systems such as the English SRN. There has been much previous research into TA models, with many alternative variations proposed with different capabilities. Section 2.2 outlined the alternative types of TA models and why, for the strategic analysis of congestion on a large scale, capacity-restrained equilibrium static models are the chosen option. The key inputs of such models are congestion functions and O-D demand matrices. There has been research into developing methods to extract data-driven versions of these inputs. However, sections 2.4 and 2.5 showed current approaches lack in accuracy, computational efficiency or applicability in the context of large networks. Currently there is a growing amount of traffic data available which can be used to improve TA model accuracy. However, it is clear there is a need for improved methods that can utilise the existing measurement infrastructure which produce accessible cross-sectional data. As described in Section 2.3.1, such cross-sectional data lacks routing information, however, it is free of the privacy concerns which limit the use of alternative data types.

After reviewing the relevant literature, the following is needed to achieve the aim of the thesis:

- A new technique is needed to apply the current state-of-the-art network tomography techniques for estimating prior O-D matrices solely from cross-sectional data to large networks. Such a method needs to provide a prior O-D estimate which can be used to obtain, via an adjustment procedure, an O-D matrix suitable for large-scale TA flow pattern analysis. All approaches surveyed in this chapter, including GLS described in Section 2.4, present issues relating to computational cost at larger network sizes.

- An investigation is needed into accurate congestion function estimation methods which can produce road-specific functions to account for individual roads' responses to congestion. As described in Section 2.5.1, data-driven congestion function methods such as Inv-Opt have only been applied to small networks and current

methods of road-specific function estimation lack accuracy in congested conditions.

To meet these needs the following is carried out. Firstly, in Chapter 5, a technique is proposed, which uses community detection via network modularity to experiment with alternative size network partitions to which GLS prior O-D estimation is applied. It investigates the effects different ways of using the partitions within the O-D estimation have on the computational efficiency and accuracy of flow pattern results. It considers these effects when applied to artificially-generated networks of different sizes and real-world data from the English SRN. Secondly, in Chapter 6, investigations are carried out, assessing road-specific density-based congestion function fitting on a large-scale national road network. One part focuses on finding the most accurate form of congestion function to fit density and travel time data across the range of roads comprising the English SRN, comparing it to flow-based fittings. The other part compares the effect such a fitting has on TA accuracy and computational requirements compared to the leading alternatives for similar data-driven models.

With these methods enabling more accurate and computationally efficient data-driven TA model inputs, the collected MIDAS data on the English SRN can be utilised for the strategic analysis discussed in Section 2.6. In Chapter 7, to showcase the analysis these improved models can produce on a large representation of the English SRN, a series of rapid results are presented which explore the effects of targeted interventions and re-routing traffic more efficiently.

Chapter 3

Model Formulation

3.1 Preliminaries and Notation

3.1.1 Notation

In this work all the vectors are column vectors. For example, the column vector $\mathbf{x}$ is written as $\mathbf{x} = \{x_i, ..., x_{dim(\mathbf{x})}\}$, where dim($\mathbf{x}$) is the dimension of $\mathbf{x}$. To denote the transpose of a matrix or vector a 'prime' is used (e.g. $\mathbf{x}'$). $\mathbb{R}_+$ denotes the set of all non-negative real numbers. Matrix $\mathbf{Q} \geq \mathbf{0}$ or vector $\mathbf{x} \geq \mathbf{0}$ indicates that all entries of a matrix $\mathbf{Q}$ or vector $\mathbf{x}$ are non-negative. Also, $|\mathcal{X}|$ represents the cardinality of a set $\mathcal{X}$, and $[\![\mathcal{X}]\!]$ is used for the set $\{1, ..., |\mathcal{X}|\}$.

3.1.2 Network Definition

For the English SRN, the NTIS model edges and nodes are grouped into superedges and supernodes which are used to create the simplified topographic representation. Each supernode is a group of NTIS model nodes which comprise motorway junctions. Each superedge is a collection of the NTIS model edges which comprise each carriageway between the junctions. After the process of node and edge combination, the supernodes and superedges which constitute the simplified topographic representation are referred to its nodes and edges.

Based on the simplified topographic representation, the road network is modelled as a directed graph with a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{A}$. The model assumes the graph is strongly connected and is defined by the node-edge incidence matrix with $\mathbf{N} \in \{0, 1, -1\}^{(|\mathcal{V}| \times |\mathcal{A}|)}$. On road networks in general, and the English SRN in particular, there is a path between all pairs of nodes so the assumption is valid.

The set of all O-D pairs on the network is denoted by $\mathcal{W} = \{\mathbf{w}_i : \mathbf{w}_i = (w_{si}, w_{ti}), i = 1, ..., |\mathcal{W}|\}$, where w_{si} is the origin node and w_{ti} is the destination node of O-D pair i. The amount of travel demand between any single O-D pair $\mathbf{w} = (w_s, w_t)$ is represented by $\hat{d}^{\mathbf{w}} \geq 0$. Using this, $\mathbf{d}^{\mathbf{w}} \in \mathbb{R}^{|\mathcal{V}|}$ is defined as a vector with all zeros except for two entries of $-\hat{d}^{\mathbf{w}}$ for node w_s and a $\hat{d}^{\mathbf{w}}$ for node w_t. Then, $\mathbf{d}^{\mathbf{w}_i}$ is a demand vector for O-D pair i, which can be combined for all O-D pairs to create the O-D demand matrix represented using $\mathbf{D} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{W}|}$.

For the demand estimation in Section 3.5, the O-D demand matrix $\mathbf{D}$ is denoted in a simplified vector form as $\mathbf{g} = (g_i; i \in [\![\mathcal{W}]\!])$ with each g_i equivalent to $\hat{d}^{\mathbf{w}_i}$. $\mathcal{R}_i$ is the index set of simple routes (without cycles) connecting O-D pair $i \in [\![\mathcal{W}]\!]$, each $r \in \mathcal{R}_i$ is a different sequence of edges which connect the O-D pair. The flow on a specific r^{th} route between O-D pair $i \in [\![\mathcal{W}]\!]$ is denoted by y_{ir}.

Let $\mathbf{x} \in \mathbb{R}_+^{|\mathcal{A}|}$ be the vector of the total edge flow x_a on edge $a \in \mathcal{A}$. Then the set of feasible flow vectors $\mathcal{F}$ is defined by:

$$\mathcal{F} \stackrel{\text{def}}{=} \{\mathbf{x} : \exists\, \mathbf{x}^{\mathbf{w}} \in \mathbb{R}_+^{|\mathcal{A}|}\ s.t. \mathbf{x} = \sum_{\mathbf{w} \in \mathcal{W}} \mathbf{x}^{\mathbf{w}}, \mathbf{N}\mathbf{x}^{\mathbf{w}} = \mathbf{d}^{\mathbf{w}}, \forall \mathbf{w} \in \mathcal{W}\}, \qquad (3.1)$$

where $\mathbf{x}^{\mathbf{w}}$ indicates the flow vector attributed to O-D pair $\mathbf{w}$. This implies that the total flow vector $\mathbf{x}$ is consistent with the demands $\mathbf{d}^{\mathbf{w}}$ between all O-D pairs.

Let $\mathbf{y}_i \in \mathbb{R}_+^{|\mathcal{R}_i|}$ be the vector of the total route flow y_{ir} on route $r \in \mathcal{R}_i$ between O-D pair $i \in [\![\mathcal{W}]\!]$. The sum of route flows, y_{ir}, equals the total demand of OD pair, i:

$$\sum_{r \in \mathcal{R}_i} y_{ir} = d^{\mathbf{w}_i} \qquad \forall i \in [\![\mathcal{W}]\!]. \tag{3.2}$$

The relation between the route and edge flows are expressed as:

$$x_a = \sum_{i \in [\![\mathcal{W}]\!]} \sum_{r \in \mathcal{R}_i} \delta_{ir}^a y_{ir} \qquad \forall a \in \mathcal{A}, \tag{3.3}$$

where δ_{ir}^a is the indicator taken from the edge-route incidence matrix $\mathbf{B} \in \{0,1\}^{(|\mathcal{A}| \times (|\mathcal{W}| \cdot |\mathcal{R}|))}$. For each $a \in \mathcal{A}$, $i \in [\![\mathcal{W}]\!]$, $r \in \mathcal{R}_i$, this indicator is defined as:

$$\delta_{ir}^a = \begin{cases} 1, & \text{if route } r \in \mathcal{R}_i \text{ uses edge } a, \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

Each route $r \in \mathcal{R}_i$ has an associated cost for its traversal denoted by $c_{ir}(y_{ir})$, this is assumed to be equal to the sum of the costs of the edges $t_a(x_a)$ defining the route (i.e. additive costs),

$$c_{ir}(y_{ir}) = \sum_{a \in \mathcal{A}} \delta_{ir}^a t_a(x_a) \qquad \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!]. \tag{3.5}$$

Also, the cost $t_a(x_a)$ on edge $a \in \mathcal{A}$ is assumed to be independent of the flow on any other edge (i.e. separable costs). In this work, it is assumed that the only cost is travel time and $t_a(x_a)$ for edge $a \in \mathcal{A}$ is provided by its congestion function (Section 3.4).

Assuming the traffic is under steady-state conditions (i.e. all vehicles have constant homogeneous speed and spacing), the fundamental relation is assumed to hold [95, 140]:

$$k_a = \frac{x_a}{v_a}, \tag{3.6}$$

where k_a is the traffic density and v_a is the spatial average speed on an edge $a \in \mathcal{A}$.

Symbol	Definition
$\mathcal{V}$	Set of Nodes
$\mathcal{A}$	Set of Edges
$\mathcal{W}$	Set of O-D Pairs
$\mathcal{F}$	Set of Feasible Flow Vectors
$\mathcal{R}_i$	Set of Simple Routes for O-D pair $i \in [\![\mathcal{W}]\!]$
$\mathcal{J}$	Set of Time-bin Average Flow Vector Samples
N	Node-Edge Incidence Matrix
$\mathbf{g}$	O-D Demand Vector
x_a	Flow on Edge $a \in \mathcal{A}$
t_a	Travel time (Cost) on Edge $a \in \mathcal{A}$
v_a	Average speed on Edge $a \in \mathcal{A}$
k_a	Density on Edge $a \in \mathcal{A}$
y_{ir}	Flow on Route $r \in \mathcal{R}_i$
c_{ir}	Cost of Route $r \in \mathcal{R}_i$
$\mathbf{B}$	Edge-Route Incidence Matrix

Table 3.1: *Notation for Network Definition*

The methods described in the following sections use different days of flow data on the network. They are seen as 'snapshots' of the network at different points in time, with $|\mathcal{J}|$ samples of the edge flow vector $\mathbf{x}$. $j \in [\![\mathcal{J}]\!]$ where j is the index of different snapshots of the network with corresponding average time-bin hourly flows.

The time-bins are discrete periods of the day used for approximating demand in the static TA model. In the model, three time-bins relating to workdays (Monday-Friday) are used: Morning (AM) 6am-10am; Midday (MD) 10am-4pm; Evening (PM) 4pm-8pm.

A collection of the network variables is provided in Table 3.1.

3.2 Traffic Assignment Optimisation Problem

To analyse the routing of traffic, the aggregate total network cost through TSTT, $\mathcal{L}$, is used. For the network, the TSTT is defined by:

$$\mathcal{L}(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} x_a t_a(x_a), \tag{3.7}$$

where the flows x_a can be obtained through the flow vectors of the UE, $\mathbf{x}_a^{UE}$, or SO, $\mathbf{x}_a^{SO}$, calculated routing patterns. The congestion functions t_a, assumed to be positive and strictly increasing, are defined in Section 3.4. Also, the measured TSTT can be calculated from the measured time-bin edge flows and travel times.

The aggregate network cost is used in the following formulation of the TAP optimisation which finds the SO flow pattern with the assumptions as defined in Section 3.1. For the fixed demand case [24, 11]:

$$\begin{aligned}
\min_{\mathbf{x} \in \mathcal{F}} \quad & T^{SO}(\mathbf{x}) = \sum_{a \in \mathcal{A}} x_a t_a(x_a), && && \text{(3.8a)}\\
s.t. \quad & \sum_{r \in \mathcal{R}_i} y_{ir} = d^{\mathbf{w}_i} && \forall i \in [\![\mathcal{W}]\!], && \text{(3.8b)}\\
& y_{ir} \geq 0 && \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!], && \text{(3.8c)}\\
& x_a = \sum_{i \in [\![\mathcal{W}]\!]} \sum_{r \in \mathcal{R}_i} y_{ir} \delta_{ir}^a && \forall a \in \mathcal{A}, && \text{(3.8d)}
\end{aligned}$$

where the objective function used (Equation 3.8a) is one which minimises the total system cost of all vehicles on the network. This is equivalent to Wardrop's second principle for SO flows. The linear constraints of the formulation are shown in Equations 3.8b to 3.8d [11]. Equation 3.8b is a set of flow conservation constraints making the total route flows between an O-D equal to its demand, such that all trips have to be assigned. Equation 3.8c ensures that route flows are non-negative as so physically realistic. Equation 3.8d links the route flows to the edge flow through the edge-route incidence matrix. This is needed as the objective

function is defined in terms of edge flows but the constraints are defined in terms of route flows. As the solution to this optimisation finds the lowest total network cost, it is equivalent to Wardrop's second principle defining SO as minimising average journey time overall.

The objective function of the SO optimisation can be expressed as the sum of the integrals of the marginal total costs of each edge $\bar{t}_a(x_a)$. Such that [24]:

$$\bar{t}_a(x_a) \stackrel{\text{def}}{=} \frac{d}{dx_a} x_a t_a(x_a) = t_a(x_a) + x_a t'_a(x_a), \tag{3.9a}$$

$$\sum_{a \in \mathcal{A}} x_a t_a(x_a) = \sum_{a \in \mathcal{A}} \int_0^{x_a} \bar{t}_a(s) ds. \tag{3.9b}$$

The marginal total cost, $\bar{t}_a(x_a)$, of an edge a at flow x_a can be interpreted as the increase in total cost on that edge from an additional driver joining that edge (mathematically it is an infinitesimal flow unit). On the right hand side of Equation 3.9a the total marginal cost is broken down into two components: $t_a(x_a)$ is the the marginal private cost; $x_a t'_a(x_a)$ is the marginal external cost. The difference between the UE and SO patterns results from the individual driver's avoidance of paying all the cost that they contribute to the system's total travel cost. From an economics perspective, a driver's private cost does not equal their total cost to the system (a.k.a social cost) due to the external costs (a.k.a externalities) they do not pay.

Based on this understanding of the difference between SO and UE, the UE TAP was developed by Beckmann *et al.* as a mathematical optimisation problem which satisfied the conditions of Wardrop's first principle [26]. With the same constraints as Equation 3.8, the UE optimisation problem with fixed demand is [24]:

$$\min_{\mathbf{x}\in\mathcal{F}} \quad T^{UE}(\mathbf{x}) = \sum_{a\in\mathcal{A}} \int_0^{x_a} t_a(s)ds, \tag{3.10a}$$

$$s.t. \quad \sum_{r\in\mathcal{R}_i} y_{ir} = d^{\mathbf{w}_i} \qquad \forall i \in [\![\mathcal{W}]\!], \tag{3.10b}$$

$$y_{ir} \geq 0 \qquad \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!], \tag{3.10c}$$

$$x_a = \sum_{i\in[\![\mathcal{W}]\!]} \sum_{r\in\mathcal{R}_i} y_{ir}\delta_{ir}^a \qquad \forall a \in \mathcal{A}. \tag{3.10d}$$

The objective function T^{UE} (Equation 3.10a) is the sum of the integrals of the congestion functions (the marginal private cost) for each edge. This is often referred to as Beckmann's transformation or function and it is often stated that it does not have a direct intuitive behavioural interpretation as the SO objective function does [11]. Instead it is an artificial form of objective function used because its first-order optimality conditions match the definition of UE conditions [141].

This can be seen through formulating the Lagrangean function and associating a set of multipliers $\boldsymbol{\mu} = (\mu_i)$ with the constraints in Equation 3.10b [24]:

$$L(\mathbf{y}, \boldsymbol{\mu}) \stackrel{\text{def}}{=} T^{UE}(\mathbf{x}(\mathbf{y})) + \sum_{i\in[\![\mathcal{W}]\!]} \mu_i (d^{\mathbf{w}_i} - \sum_{r\in\mathcal{R}_i} y_{ir}). \tag{3.11}$$

As the constraints in Equation 3.10d are definitional, used to formulate T^{UE} as a function of route flow, the only remaining constraints are the non-negativity of the route flows (Equation 3.10c). Using the Karush-Kuhn-Tucker conditions relating to inequality-constrained optimisations, the stationary points of the Lagrangean (Equation 3.11) state [141]:

$$y_{ir}\frac{\partial L(\mathbf{y},\boldsymbol{\mu})}{\partial y_{ir}} = 0, \qquad \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!], \tag{3.12a}$$

$$\frac{\partial L(\mathbf{y},\boldsymbol{\mu})}{\partial y_{ir}} \geq 0, \qquad \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!], \tag{3.12b}$$

$$\frac{\partial L(\mathbf{y},\boldsymbol{\mu})}{\partial \mu_i} = 0, \qquad \forall i \in [\![\mathcal{W}]\!], \tag{3.12c}$$

$$y_{ir} \geq 0, \qquad \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!]. \tag{3.12d}$$

By using the relation between edge and route flows in Equation 3.10d,

$$\frac{\partial T^{UE}(\mathbf{x}(\mathbf{y}))}{\partial y_{ir}} = \sum_{a\in\mathcal{A}} \frac{\partial T^{UE}}{\partial x_a}\frac{\partial x_a}{\partial y_{ir}}(\mathbf{x}(\mathbf{y})) = \sum_{a\in\mathcal{A}} \delta_{ir}^{a} t_a(x_a) = c_{ir}(\mathbf{y}). \tag{3.13}$$

This shows that the partial derivative of T^{UE} with respect to route flow, y_{ir}, at a given flow is equal to the route cost for route r between O-D pair i. Using Equation 3.13 and assuming the congestion functions are positive, the optimality conditions in Equation 3.12 become:

$$y_{ir}(c_{ir}(\mathbf{y}) - \mu_i) = 0, \qquad \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!], \tag{3.14a}$$

$$c_{ir}(\mathbf{y}) - \mu_i \geq 0, \qquad \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!], \tag{3.14b}$$

$$\sum_{r\in\mathcal{R}_i} y_{ir} = d^{\mathbf{w}_i} \qquad \forall i \in [\![\mathcal{W}]\!], \tag{3.14c}$$

$$y_{ir} \geq 0, \qquad \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!], \tag{3.14d}$$

$$\mu_i \geq 0, \qquad \forall i \in [\![\mathcal{W}]\!], \tag{3.14e}$$

where μ_i can be interpreted as the lowest cost route between O-D pair i. These first-order conditions are necessary for the optimality of $\mathbf{y}$ in the UE optimisation problem. Looking at the physical meaning of these conditions reveals the equivalence of the optimisation problem to the definition of Wardrop's first principle. The first group of the conditions

(Equation 3.14a) state that if a route is used it has to be equal to the lowest cost route, as either $y_{ir} = 0$ or $c_{ir}(\mathbf{y}) = \mu_i$. The second group of conditions (Equation 3.14b) constrain the cost of the unused routes to being more than the shortest cost route. The other conditions are the flow conservation and non-negativity constraints. These conditions match with the definition of UE that all routes between an O-D pair that are utilised have the same cost, and the cost of the unused routes is at least as large. This means that no driver can unilaterally save cost by changing routes as all alternatives are equal or higher [11].

The equivalence between the optimality conditions and the conditions for UE means that the UE conditions are satisfied at any stationary point of the optimisation formulation. It can be shown that the formulation only has one stationary point that is a minimum. The uniqueness of the solution can be established by the convexity of the optimisation formulation which has been proven [24]. The objective function is the sum of integrals of assumed increasing functions. Integrals of increasing functions are strictly convex and sums of strictly convex functions are also strictly convex. Therefore, T^{UE} is strictly convex and only has a single minimum [11].

The uniqueness of the SO optimisation problem has been proven in a similar way to UE [11]. Although not explicitly formulated, if the first-order conditions for the SO optimisation formulation (Equation 3.8) were derived they would be analogous to those derived for UE. They resemble the conditions in Equation 3.14 but the route cost $c_{ir}(\mathbf{y})$ is replaced with the marginal total cost on a route $\tilde{c}_{ir}(\mathbf{y})$ and lowest route cost μ_i is replaced with the lowest marginal total cost route $\tilde{\mu}_i$. The relationship between SO and UE can be interpreted as being, if $t_a(x_a)$ is substituted for $\bar{t}_a(x_a)$ in the objective function T^{UE} it is then equivalent to T^{SO} [24].

3.2.1 Flow Pattern Calculation

The UE (Equation 3.10) and SO (Equation 3.8) optimisation problems can be evaluated for predicted flow patterns using the Frank-Wolfe algorithm.

The Frank-Wolfe algorithm was originally developed for solving convex quadratic problems. Applied to a problem on a bounded polyhedral feasible set, it alternates between solving a linear program defined by the tangential approximation of the objective function and a line search which minimises the objective over the line segment from the solution of the linear program. The linear sub-problem defines a lower bound on the optimal value which moves closer over iterations [24].

For application to the TAP, it has the steps outlined in Algorithm 1 [24]. The Frank-Wolfe algorithm implemented in this thesis uses a convergence criterion based on the size of relative gap between consecutive iterations [24]. A non-dimensional relative gap of $\epsilon = 10^{-5}$ is used for the convergence of the edge flows, as that has been shown to be sufficient in previous analysis [42]. T is defined as T^{SO} (Equation 3.8) for SO flow patterns and T^{UE} (Equation 3.10) for UE flow patterns.

The initial solution $\mathbf{x}^0$ is obtained by setting the flows on the edges to zero, essentially an all-or-nothing assignment based on free-flow travel times (see Section 3.4.1). Key to the implementation of the algorithm on the TAP is the separability of the TAP constraints for the subproblem in step 1. This allows it to separate into independent shortest path calculations between each O-D pair based on the route costs at that iteration. Although the TAP is based on route flows, all the routes are not needed to be known. Instead only the shortest paths at each iteration are needed. To solve the search direction generation, find the shortest path between each O-D pair and add the associated demand to the edges constituting the route. The flows from each O-D pair combine to give the solution of the linear programming subproblem which is used to obtain the search direction. For calculating the shortest paths, the weights of the edges are found through the congestion functions, $t_a(x_a)$, for UE. For SO, the edge weights are calculated through the edge total marginal cost, $\bar{t}_a(x_a)$ (Equation 3.9b).

Algorithm 1: Frank- Wolfe Algorithm [24]

Step 0: (*Initialization*) Let $\mathbf{x}^0$ be a feasible solution to the TAP, LBD=0, ϵ>0, and k=0.

Step 1: (*Search direction generation*) Let:

$$\underline{T}(\mathbf{x}) \stackrel{\text{def}}{=} T(\mathbf{x}^k) + \Delta T(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k). \tag{3.15}$$

Solve the linear programming subproblem via all-or-nothing assignment of demand to the shortest path of each O-D pair based on $\mathbf{x}^k$,

$$\begin{aligned} \min \quad & \underline{T}(\mathbf{x}), && && (3.16)\\ s.t. \quad & \sum_{r\in\mathcal{R}_i} y_{ir} = d^{\mathbf{w}_i} && \forall i \in [\![\mathcal{W}]\!],\\ & x_a = \sum_{i\in[\![\mathcal{W}]\!]} \sum_{r\in\mathcal{R}_i} y_{ir}\delta_{ir}^a && \forall a \in \mathcal{A},\\ & y_{ir} \geq 0 && \forall r \in \mathcal{R}_i, i \in [\![\mathcal{W}]\!]. \end{aligned}$$

Let $\mathbf{f}^k$ be its solution. Then $\mathbf{p}^k = \mathbf{f}^k - \mathbf{x}^k$ is the search direction.

Step 2: (*Convergence Check*) Let $LBD := max\{LBD, \underline{T}(\mathbf{f}^k)\}$. If,

$$\frac{T(\mathbf{x}^k) - LBD}{LBD} < \epsilon$$

then terminate, $\mathbf{x}^k$ is the approximate solution. Otherwise, continue.

Step 3: (*Line Search*) Find step length, l^k, which is the solution of the one-dimensional problem:

$$min\{T(\mathbf{x}^k + l^k\mathbf{p}^k) | 0 \leq l^k \leq 1\}.$$

Step 4: (*Update*) Let $\mathbf{x}^{k+1} = \mathbf{x}^k + l^k\mathbf{p}^k$.

Step 5: (*Convergence check*) If

$$\frac{T(\mathbf{x}^{k+1}) - LBD}{LBD} < \epsilon.$$

then terminate, $\mathbf{x}^{k+1}$ is the approximate solution. Otherwise, let $k := k + 1$, go to Step 1.

The line search in step 3 finds the step length to adjust the flows on each edge to move the solution towards the optimum. After the flows are adjusted, if the convergence criterion is not met, the shortest paths are recalculated to find a new search direction. Many of the variations of the Frank-Wolfe algorithm focus on alternative choices for this line search. Simplified methods such as the method of successive averages, use a predetermined step length based on the iteration number [24].

3.3 Process of Model Creation

The process to calculate a traffic flow assignment from a TA model derived from cross-sectional data consists of four steps. After collecting the data and associating them to the network model, a simplified topographic representation of the motorway network is extracted. This is the network model on which the O-D matrix can be calculated as well as the congestion functions for the solution of the TAP, which is then addressed in the fourth and final step. This process is presented in Figure 3.1.

3.4 Congestion Function Formulations

Accurate congestion functions are key to TA models as they connect the travel time t_a to the vehicle flow demand $\check{x}_a$ on edge $a \in \mathcal{A}$. Note the difference between flow demand $\check{x}_a$ and flow x_a. Flow demand indicates the number of vehicles wishing to use the edge in a period of time, which due to congestion may be greater than the flow ($\check{x}_a \geq x_a$). As discussed in Chapter 2, flow cannot exceed capacity however, flow demand can.

In the network model congestion functions take the form:

$$t_a(\check{x}_a) = t_a^0 f\left(\frac{\check{x}_a}{m_a}\right) , \tag{3.17}$$

where t_a^0 is the free-flow travel time of an edge $a \in \mathcal{A}$ and $f(\cdot)$, also known as the travel time multiplier, is a strictly increasing and continuously differentiable function dependent on the flow demand $\check{x}_a$ divided by the flow capacity m_a of that edge $a \in \mathcal{A}$ (i.e. the saturation rate).

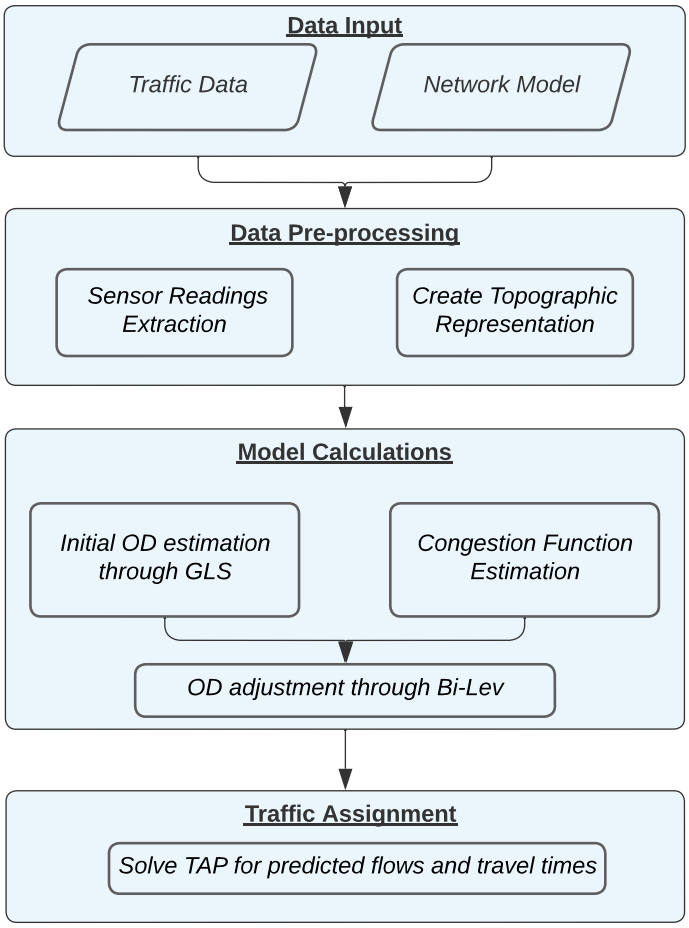


Figure 3.1: *Procedure to obtain data-driven Traffic Assignment (TA) model on real-world road networks from traffic data.*

There are several main candidates for the form of the travel time multiplier $f(\cdot)$ which have been developed with the required traits for TA.

BPR

The BPR equation is consistent with Equation 3.17 and is widely used in TA models [25, 20]. In its more general form it is:

$$t_a = t_a^0 \left(1 + \alpha (\frac{\check{x}_a}{m_a})^{\beta} \right) , \tag{3.18}$$

where the values of α and β are coefficients commonly taken as 0.15 and 4, respectively [11]. This form of BPR with these standard coefficients will be referred to as BPR-Standard.

Conical

Conical is a function that was developed after BPR to address some of its perceived issues relating to equilibrium assignment convergence [93]. Unlike BPR, which was based on empirical observations, Conical was derived mathematically using basic algebra and geometry to meet postulated requirements of well-behaved congestion functions. In addition to the general requirements of congestion functions to be positive, smooth and monotonically increasing, Conical was developed with additional requirements. Notably, this included limitations on the steepness of the curve and its coefficients having compatibility with BPR.

The form of the function is:

$$t_a \quad = \quad t_a^0 \left(2 + \sqrt{\alpha^2 (1 - \check{x}_a / m_a)^2 + \beta^2} - \alpha (1 - \check{x}_a / m_a) - \beta \right), \tag{3.19}$$

where,

$$\beta = \frac{2\alpha - 1}{2\alpha - 2} , \quad \alpha > 1.$$

In Conical, α is described as the steepness of the function and can be interpreted as similar to β in BPR.

Akçelik

Akçelik is a function form which has been used in practice to develop the BPR further, to represent better the hyper-critical and hypo-critical flow regimes [97]. In previous works it has been shown to be more accurate representing road facilities with signalised intersections [98]. The form of the function used in this work is a simplified version based on [111] and [142]:

$$t_a \quad = \quad t_a^0 + \left[0.25 \left((\frac{\check{x}_a}{m_a} \; - \; 1) \; + \; \sqrt{\left((\frac{\check{x}_a}{m_a} - 1)^2 + J\frac{\check{x}_a}{m_a} \right)} \right) \right], \quad (3.20)$$

where J is the delay parameter to be fitted, in this work it subsumes constants present in other formulations. In this equation t_a is average travel time per unit distance (hr/km) and t_a^0 is free-flow travel time per unit distance (hr/km).

Exponential

In other works, exponential functions are used as an alternative to BPR as they possess a similar shape [111]. In this work, the form of exponential function used is:

$$t_a = t_a^0 e^{\alpha(\check{x}_a/m_a)^\beta} \; , \quad (3.21)$$

where the values of α and β are coefficients to be fitted.

3.4.1 Calculating Fundamental Traffic Parameters from Data

In addition to the form of the travel time multiplier which is investigated in Chapter 6, the performance of congestion functions also depends on the values chosen for other key parameters specific to each road.

Capacity

The capacity of an edge on the network is defined in the US Highways Capacity Manual (HCM) [101] as:

"The capacity of a system element is the maximum sustainable hourly flow rate at which persons or vehicles reasonably can be expected to traverse a point or a uniform section of a lane or roadway during a given time period under prevailing roadway, environmental, traffic, and control conditions."

The vagueness of the definition and the lack of an agreed methodology for the calculation of the capacity on a road has led to many studies interpreting capacity in different ways. The most widely used is the average of high traffic flows. Its simplicity makes it preferred for the purposed of national road systems with a potentially large number of edges. Studies employing the method use different criteria as the cut-off for which average high traffic flow constitutes the capacity. For example, [111] uses 99^{th} percentile and in [108] it is the 95^{th} percentile. However, the reasons for the choice are often unclear.

In this work, as in [143], the proposed congestion function estimation uses the maximum of the observed hourly-average flows on an edge as its capacity. This is based on the assumption that additional stochastic variables, such as weather and road conditions, always negatively affect the capacity, and the true value for a road is hardly ever reached. Capacity is treated as being a deterministic maximum value reduced by random factors. For the purpose of the TA model only the deterministic maximum part is used so the model can reflect the ideal road performance.

Free-flow speed and time

The free-flow speed of an edge on the network is defined in the HCM [101] as:

"The theoretical speed when the density and flow rate on the study segment are both zero"

HCM expands that this can be seen as a constant speed in traffic flows between 0 and 1,000 vehicles per hour per lane. This is limited as there is uncertainty regarding constant speeds in low flow traffic.

As with road capacity, free-flow speed is calculated in alternative ways in different studies and research continues into improving its estimation [144]. Some studies use posted speed limits without detailed field measurements of speeds. For example, [113] sets the free-flow speed of light vehicles as 111kph (approx. the 70mph speed limit) for two-lane motorways. In [108], free-flow speed is taken as the 85^{th} percentile of measured speeds. In [143], the free-flow speed is estimated through a least squares fit on the flow versus density data for measurements with a speed greater than 55mph.

In this thesis, the free-flow speed of an edge is obtained by taking the 95^{th} percentile of the observed speeds as in [107]. This is a quick way to obtain an estimate which accounts for localised factors specific to the road (e.g. road curvature) which have influence but are not captured by the posted speed limit. The free-flow travel time used in the congestion function is then obtained by converting it through the edge length.

3.5 Data-driven Origin-Destination Demand Matrix Estimation

For the calculation from traffic flow measurements, the O-D matrix estimation problem can be understood as an optimisation problem and formulated as:

$$\min_{\mathbf{g}\geq\mathbf{0}} \quad F(\mathbf{g}) \stackrel{\text{def}}{=} \gamma_1 F_1(\mathbf{g}, \mathbf{g}^0) + \gamma_2 F_2(\mathbf{x}(\mathbf{g}), \tilde{\mathbf{x}}) \tag{3.22a}$$

$$s.t. \qquad \mathbf{x} = assign(\mathbf{g}), \tag{3.22b}$$

where $F_1(\mathbf{g}, \mathbf{g}^0)$ is a distance measure between estimated demand vector $\mathbf{g}$ and the initial prior demand vector $\mathbf{g}^0$, $F_2(\mathbf{x}(\mathbf{g}), \tilde{\mathbf{x}})$ is a distance measure between estimated edge flows $\mathbf{x}(\mathbf{g})$ and observed edge flows $\tilde{\mathbf{x}}$. γ_1 and γ_2 are weighting factors which can be dependent on the relative confidence in the prior matrix and measured flows, or the scale each distance measure operates on. Lastly, *'assign'* is the process of drivers making route choices that is dependent on the demand.

Algorithms are available which adjust the demand vector so that the resulting assigned flows are as close as possible to the observed flows whilst the demand also remains as close as possible to the prior matrix. Essentially the measured flows can be thought of as correcting or updating a prior O-D matrix.

3.5.1 Estimating the Prior Demand Matrix

Applied to each of the time-bins, the GLS method is used to estimate the prior O-D demand matrices. It assumes the roads are uncongested so that for each O-D pair the route choices of the road users are independent of the traffic flows. It assumes an assignment matrix **A** linking demand to flow with the route choices predetermined.

$$\mathbf{x} = \mathbf{A}\mathbf{g}. \tag{3.23}$$

The GLS method assumes that the O-D trips are Poisson distributed and there is a maximum of one edge in each direction between pairs of nodes [55, 64]. To find the prior O-D demand matrix, first the feasible routes for each O-D node pair are found and used to create the edge-route incidence matrix **B**. The feasible routes are limited to the two shortest routes by distance, if available, as it is commonly the case for the majority of the route flows to use the best couple of choices [145]. The feasible routes for each O-D node pair are found using Yen's multiple shortest paths algorithm [146]. Next, define $\{\mathbf{x}^{(j)}; j \in [\![\mathcal{J}]\!]\}$ to be $|\mathcal{J}|$ observations of the flow vector on the edges and $\bar{\mathbf{x}}$ as their arithmetic average. Define $\boldsymbol{P} = [p_{ir}]$ to be the route choice probability matrix, p_{ir} as the probability that a traveller between O-D pair i uses route r. The assignment matrix is then $\mathbf{A} = \mathbf{B}\mathbf{P}'$. Next, vectorise the O-D demand matrix as **g** and define the sample covariance matrix by [10] :

$$S = \frac{1}{(|\mathcal{J}| - 1)} \sum_{j=1}^{|\mathcal{J}|} (\mathbf{x}^{(j)} - \bar{\mathbf{x}})(\mathbf{x}^{(j)} - \bar{\mathbf{x}})'. \tag{3.24}$$

Using this, the method requires solving the following optimisation problem:

$$
\begin{aligned}
(E0) \quad \min_{\mathbf{P}\geq\mathbf{0},\mathbf{g}\geq\mathbf{0}} \quad & \sum_{j=1}^{|\mathcal{J}|}(\mathbf{x}^{(j)} - \mathbf{B}\mathbf{P}'\mathbf{g})'S^{-1}(\mathbf{x}^{(j)} - \mathbf{B}\mathbf{P}'\mathbf{g}) && \text{(3.25a)}\\
s.t. \quad & p_{ir} = 0 \qquad \forall\ (i,r) \in \{(i,r) : r \notin \mathcal{R}_i\}, && \text{(3.25b)}\\
& \mathbf{P}\mathbf{1} = \mathbf{1}. && \text{(3.25c)}
\end{aligned}
$$

This optimisation minimises the weighted sum of the squared errors in the flow observations, however, directly solving it is difficult because of its complicated form of objective function. For a solution it is rewritten as two optimisation sub-problems sequentially. To do this, the substitution $\boldsymbol{\xi} = \mathbf{P}'\mathbf{g}$ is made and an arbitrary smooth scalar-valued function $h(\mathbf{P},\mathbf{g})$ is used to formulate the following [10] :

$$
(E1) \qquad \min_{\boldsymbol{\xi}\geq\mathbf{0}} \quad \frac{|\mathcal{J}|}{2}\boldsymbol{\xi}'\mathbf{Q}\boldsymbol{\xi} - \mathbf{b}'\boldsymbol{\xi}, \qquad \text{(3.26a)}
$$

$$
\begin{aligned}
(E2) \quad \min_{\mathbf{P}\geq\mathbf{0},\mathbf{g}\geq\mathbf{0}} \quad & h(\mathbf{P},\mathbf{g}) && \text{(3.26b)}\\
s.t. \quad & p_{ir} = 0 \qquad \forall\ (i,r) \in \{(i,r) : r \notin \mathcal{R}_i\}, && \text{(3.26c)}\\
& \mathbf{P}'\mathbf{g} = \boldsymbol{\xi}^0, && \text{(3.26d)}\\
& \mathbf{P}\mathbf{1} = \mathbf{1}, && \text{(3.26e)}
\end{aligned}
$$

where $\mathbf{Q} = \mathbf{B}'\mathbf{S}^{-1}\mathbf{B}$ and $\mathbf{b} = \sum_{j=1}^{|\mathcal{J}|}\mathbf{B}'\mathbf{S}^{-1}\mathbf{x}^{(j)}$.

$\boldsymbol{\xi}^0$ is the optimal solution to (E1), which is obtainable via quadratic programming, using a numeric solver (e.g. Gurobi [147]). The variable substitution removes the probability matrix from the problem, eliminating the constraints on $\mathbf{P}$ in (E0). The solution of (E2) returns a $\mathbf{P}$ which is

consistent with the solution of (E1) and $\boldsymbol{\xi} = \mathbf{P}'\mathbf{g}$, providing the optimal solution $(\mathbf{P}^0, \mathbf{g}^0)$. Although (E2) is written as an optimisation problem, $h(\mathbf{P}, \mathbf{g})$ is a dummy objective function which can be set equivalent to a zero constant, without affecting the results [10].

For the purposes of providing a prior matrix of the demand between each O-D pair from the vector $\mathbf{g}^0$, the solution of $\mathbf{g}^0$ is treated as the sum of the values of $\boldsymbol{\xi}^0$ which are related to the routes of the same O-D pair. Therefore, $\mathbf{P}$ is not needed.

The number of decision variables in the optimisation depends on the number of O-D pairs and routes. The size of the network is a restriction of the method that faces numerical difficulties at large network sizes, however, it is better than other formulations considered such as the maximum likelihood estimation method [55].

3.5.2 Demand Matrix Congestion Adjustment

To account for the effects of congestion and improve the accuracy of the initial estimate of the demand vector $\mathbf{g}$, the estimated congestion functions can be used to find an improved solution through a heuristic gradient-based algorithm [72, 73]. The assignment of the demand is then assumed to be dependent on the edge flows, such that:

$$\mathbf{x} = \mathbf{A}(\mathbf{g})\mathbf{g}. \tag{3.27}$$

This is not analytically solved, instead the calculation of the assignment matrix $\mathbf{A}$ is a separate optimisation problem within the main BiLev process. Based on Equation 3.22, the BiLev problem is expressed through the following formulation:

$$\min_{\mathbf{g}\geq\mathbf{0}} \quad F(\mathbf{g}) \stackrel{\text{def}}{=} \gamma_1 \sum_{i\in[\![\mathcal{W}]\!]} (g_i - g_i^0)^2 + \gamma_2 \sum_{a\in\mathcal{A}} (x_a(\mathbf{g}) - \tilde{x}_a)^2, \tag{3.28a}$$

$$s.t. \qquad \mathbf{x} = \mathbf{A}(\mathbf{g})\mathbf{g}, \tag{3.28b}$$

In the upper-level of the BiLev the optimal demand vector is sought through the objective function, subject to the lower-level optimisation of the assigned edge flows through solution of the UE TAP. To iterate a numerical solution, the gradient descent algorithm updates the demand vector at each iteration according the descent direction evaluated from the objective function gradient. The gradient of F(**g**) is:

$$\begin{aligned}\nabla F(\mathbf{g}) &= \left(\frac{\partial F(\mathbf{g})}{\partial g_i}; i \in [\![\mathcal{W}]\!]\right) \\ &= \left(2\gamma_1(g_i - g_i^0) + 2\gamma_2 \sum_{a\in\mathcal{A}} (x_a(\mathbf{g}) - \tilde{x}_a)\frac{\partial x_a(\mathbf{g})}{\partial g_i}; i \in [\![\mathcal{W}]\!]\right).\end{aligned} \tag{3.29}$$

The Jacobian matrix in the expression is simplified by assuming that the route choice probabilities are locally constant. Also, for an O-D pair $i \in [\![\mathcal{W}]\!]$, only the fastest route $\hat{r}(\mathbf{g})$ is considered. In each iteration of the algorithm, the edge travel times are updated based on the current demand vector after the flow vector has been found via solution of the UE TAP [10].

Assuming that the partial derivatives do exist [148], then the Jacobian is approximately [72]:

$$\frac{\partial x_a(\mathbf{g})}{\partial g_i} \approx \delta^a_{i\hat{r}(\mathbf{g})} = \begin{cases} 1 & \text{if a} \in \hat{r}(\mathbf{g}), \\ 0 & \text{otherwise.} \end{cases} \tag{3.30}$$

The assumption of locally constant route choice probabilities means edge flows are separable, which could lead to biased gradient approximations.

The simplification of the Jacobian via the assumption of only including the fastest route between each O-D pair is based on the widespread use of GPS navigation suggesting that route to most drivers. If more than one route was included then the route flows could not be uniquely determined by solving the TAP and this would reduce the accuracy of the Jacobian approximation from unstable route-choice probabilities [10].

Algorithm 2: BiLev Gradient Descent Algorithm

Input:

The road network $(\mathcal{V}, \mathcal{A}, \mathcal{W})$; the fitted functions $t(\cdot)$ for each edge; the initial estimated demand vector $\mathbf{g^0}$=$(g_i^0; i \in [\![\mathcal{W})]\!]$; five parameters $T \geq 1$, $\Theta \geq 1$, $\epsilon_1 \geq 0, \epsilon_2 > 0, l_{max} > 0$.

Initialization:

Take the demand vector $\mathbf{g}^0$ as the input, solve the TAP (using the Frank-Wolfe algorithm (Section 3.2.1)) to obtain $\mathbf{x}^0$. Set $l = 0$. If $F(\mathbf{g}) = 0$, stop; otherwise, go on to Step 1.

Step 1:

Computation of a descent direction. Calculate $\mathbf{h}^l = -\nabla F(\mathbf{g}^l)$ by Equation (3.29).

Step 2:

Calculation of a search direction. For $i \in [\![\mathcal{W}]\!]$ set:

$$\bar{h}_i^l = \begin{cases} h_i^l, & \text{if } (g_i^l > \epsilon_1) \text{ or } (g_i^l \leq \epsilon_1 \text{ and } h_i^l > 0) \\ 0, & \text{otherwise.} \end{cases}$$

Step 3:

Line search.

Step 3.1: Calculate the maximum possible step-size $\theta_{max}^l = min\{-g_i^l / \bar{h}_i^l; \bar{h}_i^l < 0, i \in [\![\mathcal{W}]\!]\}$. If $\bar{h}_i^l \geq 0$ for all $i \in [\![\mathcal{W}]\!]$, then $\theta_{max}^l = 0.001$.

Step 3.2: Determine $\theta^l = \arg\min_{\theta \in \mathcal{S}} F(\mathbf{g}^l + \theta \bar{\mathbf{h}}^l)$, where $\mathcal{S} \stackrel{\text{def}}{=} \{\theta_{max}^l, \theta_{max}^l/\Theta, \theta_{max}^l/\Theta^2, ..., \theta_{max}^l/\Theta^T)$.

Step 4:

Update and termination

Step 4.1: Set $\mathbf{g}^{l+1} = \mathbf{g}^l + \theta^l \bar{\mathbf{h}}^l$. Using $\mathbf{g}^{l+1}$ as the demand, solve the TAP to obtain $\mathbf{x}^{l+1}$.

Step 4.2: If $\frac{F(\mathbf{g}^l) - F(\mathbf{g}^{l+1})}{F(\mathbf{g}^0)} < \epsilon_2$, stop the iteration; otherwise, go on to Step 4.3.

Step 4.3: Set $l = l + 1$ and return to Step 1 if $l \leq l_{max}$.

The gradient-based algorithm is outlined in Algorithm 2. It is a variant of the algorithms proposed in [10], [72] and [73]. The BiLev optimisation problem is not convex due to the potential non-linearity in $\mathbf{x}(\mathbf{g})$ and so the solution is not expected to be a global minimum. A key concern is the accuracy of the calculated gradient to ensure the algorithm leads to a local minimum. If the gradient is calculated inaccurately due to computational inaccuracy, the descent of the algorithm may not occur in those iterations. For this reason it is important to have the evaluation of the TAP as accurate as possible.

The line search uses an Armijo-type procedure as it is computationally efficient [73, 149]. The search is initiated with a maximum possible step-size $\theta = \theta^l_{max}$ that leads to non-negative demands for all O-D pairs. For a value of θ, the objective function $F(\mathbf{g}^l + \theta\bar{\mathbf{h}}^l)$ is calculated. If the objective function is improved such that $F(\mathbf{g}^l) - F(\mathbf{g}^l + \theta\bar{\mathbf{h}}^l) > 0$, then the search is interrupted and that value of θ is the chosen step-size. If not, then a smaller step-size $\theta = \theta^l_{max}/\Theta^{\hat{T}}$ is tried in the evaluation of the objective function, such that $1 \leq \hat{T} \leq T$ where $\Theta \geq 1$ and $T \geq 1$ are specified parameters.

The choice of the parameters used is important for the accuracy of the result. The choice of $\gamma_1 = \gamma_2 = 1$ is made with both parameters equal to ensure the algorithm produces a reduction in the difference between the calculated and measured flows while equally considering the initial estimation of $\mathbf{g^0}$. As in [10], the values of $\epsilon_1 = 0$ and $\epsilon_2 = 10^{-20}$ are chosen to ensure the accuracy of the algorithm although they reduce convergence speed. Effectively, in practice, such a small value for ϵ_2 means that the algorithm runs until a maximum number of iterations $l_{max} = 30$, chosen to ensure convergence. In the solution of the TAP the accuracy of the relative gap is set to 10^{-5} to make sure the flow vector and therefore the gradient are sufficiently accurate.

The use of $\Theta = 10$ and $T = 8$ is to quickly reduce the step-size from θ^l_{max}. In [73], the authors remark that if the gradient is a descent direction then θ^l_{max} will usually be the best step-size to reduce the objective function the most. They state that if it is not the descent direction then the step-size should be quickly reduced and the smallest increase in objective function accepted, to move on to the next iteration's search direction

calculation. In the implementation of the algorithm used in this work, it starts from the largest step-size θ^l_{max} and decreases it until a reduction in objective function is obtained, the first step-size with a reduction is selected and the algorithm continues. This improves speed as it reduces the number of times the Frank-Wolfe algorithm is used to solve the TAP at each step-size. If a reduction in objective function is not obtained, then the smallest step-size is chosen.

In [10], the authors prove their version of the algorithm converges through including $0 \in \mathcal{S}$. This indicates that the non-negative objective function is non-increasing between iterations and convergence is guaranteed by the monotone convergence theorem [150]. However, by allowing small increases in the objective function, it allows all executions of the algorithm to run equally many iterations, which is useful for comparisons and in practice usually leads to lower final objective function values.

Chapter 4

Application on the English Strategic Road Network

4.1 English SRN Data Sets

4.1.1 NTIS

The NTIS Network and Asset Model contains information on the location and details of the different systems used by NH to monitor and control traffic on the SRN. The systems it covers include MIDAS sites, ANPR Camera sites, Traffic Monitoring Units (TMU) sites, VMS and Matrix Signals. Also, it includes the Network Model containing geospatial information on the edges, edge shapes and nodes which can be used in a graph representation of the roads. Various attributes are also available to determine the direction of travel, capacity and length of the associated weighted graph's edges [5].

The data is updated approximately fortnightly to account for changes to the network and to correct errors (e.g. duplicate sensor IDs). As the data is provided in DATEX II, a European standard format, the techniques applied to the English SRN could be applied with relative ease to other systems such as in Scotland [5].

Category	Length
1	<=5.2m
2	>5.2m and <=6.6m
3	>6.6m and <=11.6m
4	>11.6m
5	All lengths

Table 4.1: *Table of the MIDAS vehicle length categories [5].*

4.1.2 MIDAS Traffic Monitoring System

The traffic data used in this thesis is taken from the MIDAS system. The MIDAS system measures speed, flow, occupancy and headway at approximately 7000 sites on the main motorways of the English SRN. The data is given on a per-lane basis and aggregated over 1-minute intervals. It provides flow data for separate vehicle categories defined by length (see Table 4.1). However, the vehicle category data is not on a lane basis and not supplied for speed, occupancy and headway [5]. In most cases the MIDAS measurement site uses two inductive loops installed under the road surface, however, some use radar or alternative technology. They are spaced approximately every 500m. If one is faulty, the site cannot classify vehicle length or speed but can measure total flow, occupancy and headway. Occupancy could be used to estimate the speed as is done for single-loop detector systems [5].

There are in total approximately 9000 MIDAS sites installed on the motorways of the SRN, however only approximately 7000 are configured for measuring data [5].

The NTIS network comprising the main carriageways with relevant MIDAS sensor sites on the SRN is selected for analysis (Figure 4.1). This connects a selection of major cities in England; however, only the main part of the network is included in the analysis. There are small disconnected stretches of MIDAS monitoring other areas of the network (e.g. near Southampton) which are not included.

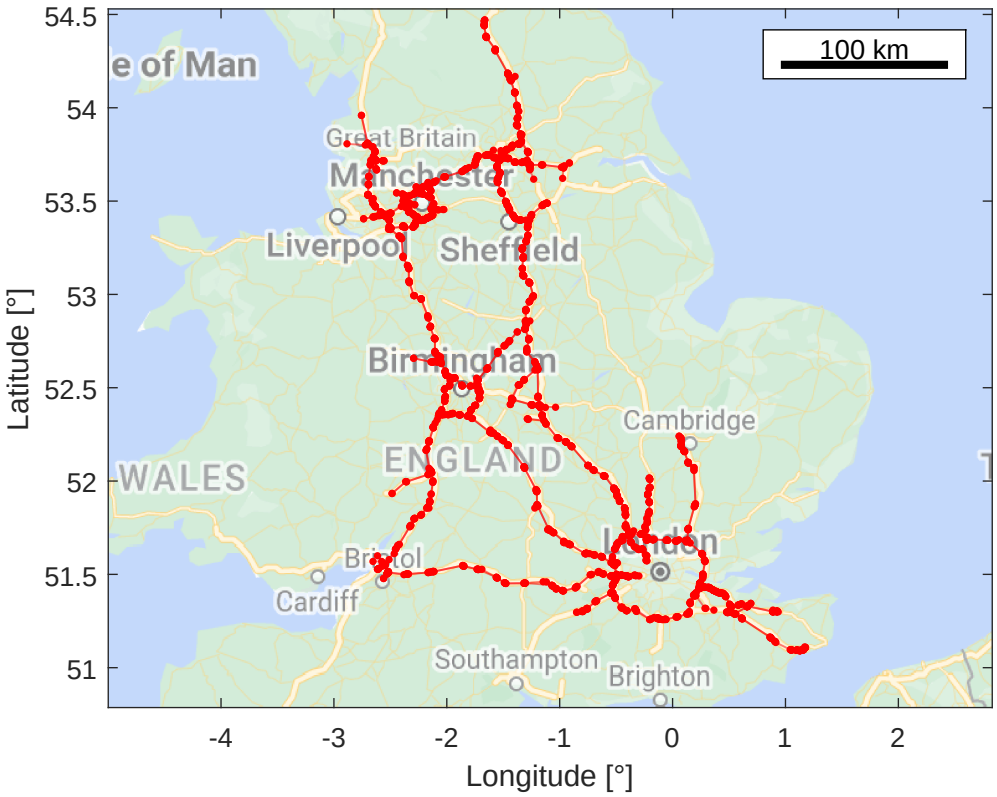


Figure 4.1: *Graph representation of the National Traffic Information Service (NTIS) model of the contiguous Strategic Road Network (SRN) covered by the Motorway Incident Detection and Automatic Signalling (MIDAS) system. Map underlay from Google Maps [151].*

Network	Nodes	Edges	O-D Pairs
E_1	30	70	870
E_2	73	156	5256
E_3	132	278	17292

Table 4.2: *Features of the topographic representations of the English Strategic Road Network (SRN).*

4.2 Network Graph Topographic Representation

The scale of the TA models is not concerned with navigation through the junctions between roads but instead with modelling the overall flows around the network. Therefore, a degenerate arterial road topographic representation is created for the SRN. This creates a processed version of the NTIS model with the junctions and interchanges simplified to single supernodes and the carriageways in-between grouped as single superedges. The use of the superedges involves averaging the flows recorded by the sensors on the edges which compose them. The roads are assumed independent in each direction of travel and represented by separate superedges. Furthermore, it is assumed that the network is an open-system without entry gates or intersection control devices such as traffic lights. The use of topographic simplification has the benefit of reducing computational complexity, however, the averaging of sensors along superedges may reduce model accuracy. As the superedges are intended to capture the edges between the junctions represented by supernodes, they should be accurate for the purposes of routing for a traffic assignment model, as the routes can only change at the junctions.

Three topographic representations of the SRN are presented in Figure 4.2. Table 4.2 describes the features of the subnetworks. The first, E_1, is a more simplified version of the second, E_2, which is a smaller network covering the central subnetwork of the SRN. E_1 and E_2 are used in the development of the methods in Chapters 5 and 6. E_3 is a larger representation of the national network used in the analysis of Chapter 7. The nodes and edges of the NTIS network that do not have suitable data over the analysis period are not included in the topographic representations. Due to such data limitations, this notably results in a single

pair of edges representing a good proportion of the M6 that connects Birmingham to near Manchester (approx. 53° N, 2.3° W).

4.2.1 Map Data Simplification

To create the topographic representation for analysis, an automated simplification approach was developed to work with the NTIS data. The process is outlined in Algorithm 3.

The interchanges of the road network which require simplification consist of intersections of different edges. The approach detects the characteristic intersections of edges that comprise the interchanges. The method uses the order of the nodes (i.e. how many edges connect to the node) and the NTIS-provided labels of the edges to identify the characteristic intersections. In the NTIS database there are two types of interchange edges to identify labelled slip roads and roundabouts (Figure 4.3). There is also a minor 'other' category. First, a search of all the edges of the network is applied to find those with the correct features to group their attached nodes into the interchange types. Hierarchical clustering is applied to each the nodes grouped under each interchange type (slip road, roundabout, other) separately with the geographic distance parameter 500-700m based on the previous research for intersection detection [61].

The new simplified network graph is comprised of NTIS edges grouped into superedges and NTIS nodes grouped into supernodes. A modified Depth First Search (DFS) algorithm is used to find neighbouring clusters and establish the superedges connecting the topographic graph.

The DFS algorithm [152] works by exploring the NTIS network from all the NTIS nodes in a cluster. It is modified to stop when it reaches a NTIS node identified as from another cluster. The NTIS edges of the shortest path taken to reach this other cluster's NTIS node comprise the superedge connecting the two supernodes.

Following the production of an estimate of the topographic representation, it is checked for errors and corrected if necessary.

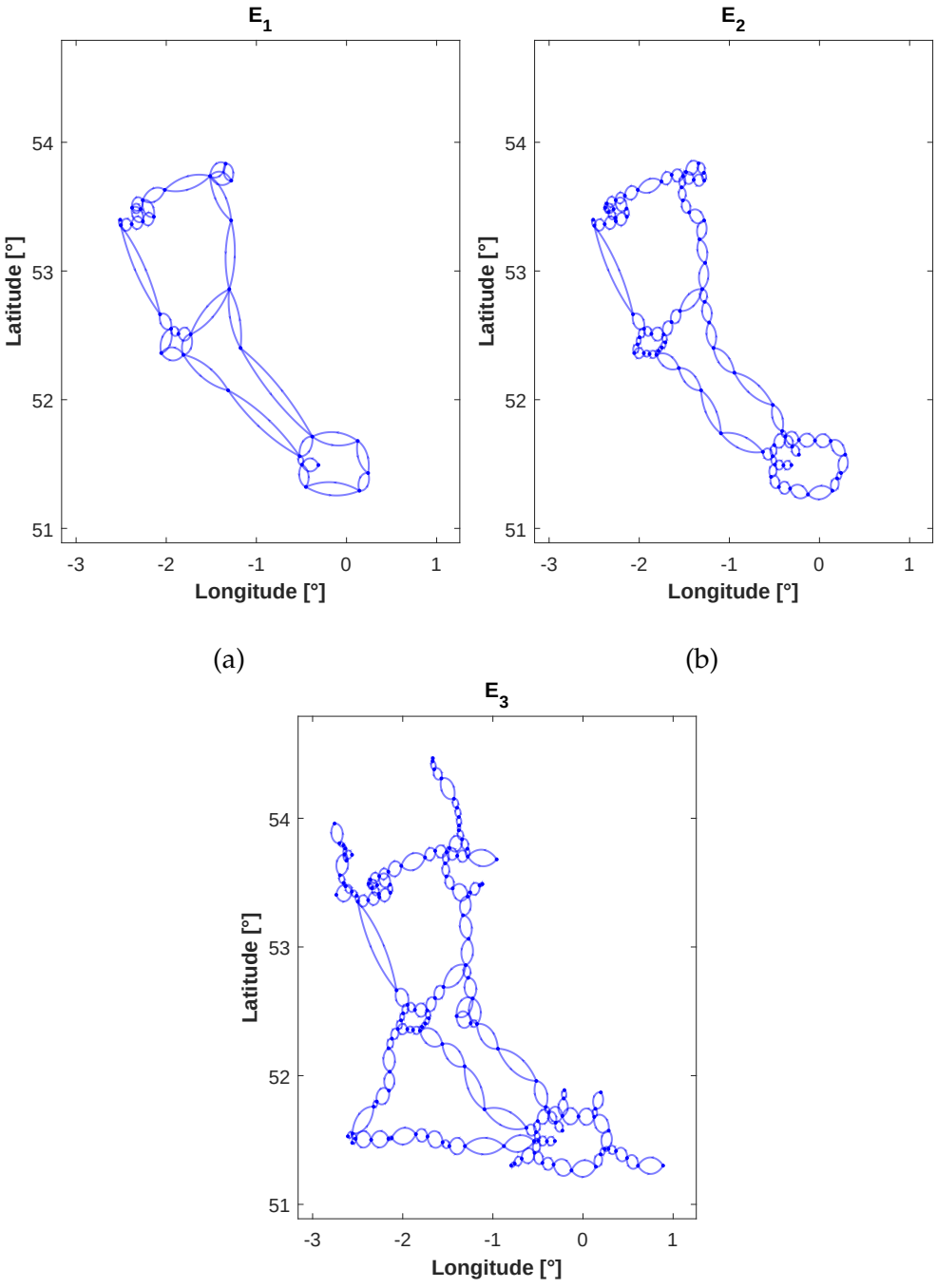


Figure 4.2: *Topographic subnetwork representations of the main roads connecting the Strategic Road Network (SRN). (a) E_1; (b) E_2; (c) E_3.*

Algorithm 3: Algorithm for Interchange Simplification

Input: The road network $(\mathcal{V}, \mathcal{A})$.

Step 1 - Initial Classification: For all edges, select the nodes attached to the edges which fit in the following categories:

1. *Roundabout*
2. *Slip Roads (with node order 1)*

Step 1.1 - Initial Clustering: Apply hierarchical clustering on node geographical position with distance parameters 700m and 500m to roundabout and slip road nodes, respectively.

Step 1.2 - Cluster Expansion: After clustering include in the roundabout category:

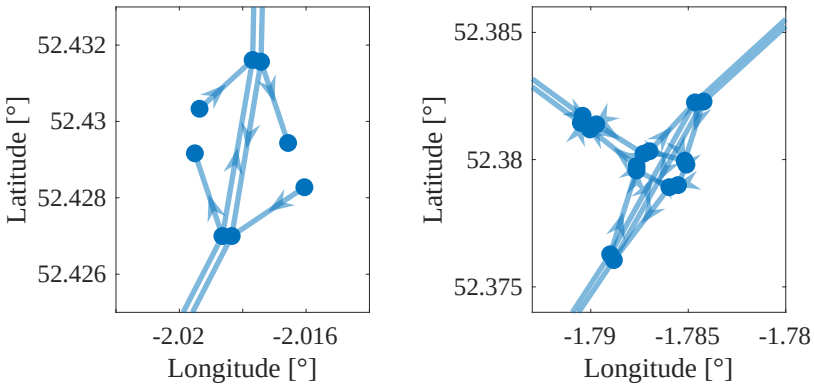
1. Nodes of entry or exit slip roads that connect to roundabout nodes.
2. The nodes of entry or exit slip roads that connect to those entry or exit slip roads
3. Nodes of connected edges with type "*mainCarriageway*" where the other node is order 1.

After clustering slip roads, the other node of the entry/exit slip road edge with an order greater than 1 is added to that cluster.

Step 2 - Other Classification and Clustering: Select all nodes with an order not equal to 2 in the 'Other' category. Apply hierarchical clustering with distance parameter 500m.

Step 3 - Supernode creation: Calculate the mean latitude and longitude of the nodes of each cluster that are grouped into supernodes.

Step 4 - Connect Clusters: Use the modified DFS to establish superedges connecting each supernode.



(a) (b)

Figure 4.3: *Examples of (a) slip roads and (b) roundabouts in the National Traffic Information Service (NTIS) model.*

4.3 MIDAS data extraction

MIDAS data from the available sensors is extracted and matched to the associated topographic edge through the NTIS data set which contains the information of the position of the sensors on the network. The matching process accounts for errors in the NTIS data set by using both the Regional Control Centre (RCC) and sensor ID to account for the small number of sensors across RCCs with duplicate non-unique IDs.

The flow data recorded are grouped into the time-bins defined in Section 3.1. For each time-bin, the mean hourly flow is calculated over the respective period. Within each time-bin the per-minute measurements are extracted and then the hourly mean values are calculated for the congestion function fitting in Chapter 6. The multiple lanes of flow are summed to get the total flow, the speed and occupancy are averaged across the lanes which are active. The hourly mean values are used to approximate the steady-state conditions of the traffic (i.e. all vehicles have constant homogeneous speed and spacing) [95]. The steady-state assumption is important for the fundamental relation of traffic to hold between flow, speed and density (see Section 3.1).

The data set was restricted to weekdays only. Data were omitted from dates which coincided with public holidays, days with major weather disruption (i.e. snow on 01/02/2019), and around the first and last week of the year due to potential Christmas and New Year holiday demand disruption (20/12/2018 - 07/01/2019). Nine months of data, September 2018 to May 2019, were used for the smaller networks featured in the analysis of Chapters 5 and 6. This was expanded to twelve months, September 2018 to August 2019, for the larger analysis in Chapter 7.

Loop detector data can be noisy [50, 153] and needs to be processed correctly to remove faults [154, 155]. When multiple sensors are available along the length of the same edge, the median flow readings are used. This both minimises the effect of outliers and filters out erroneous readings, which are those differing from the median by more than twice the median absolute deviation. This allows the central tendency of measured flows to be resistant to sensors with faults or which do not measure the main carriageway flow, even after the slipway sensors are excluded through their database names.

The data available from the MIDAS sensors can change over the time window of the data set. Once the data available for each edge have been assessed, days of data that are missing measurements (zero flow) for any of the edges in any of the time-bins are excluded from the final data set. If the sensors on a particular edge are consistently returning faulty readings then the topographic graph is amended to absorb that edge into its neighbour.

4.3.1 Traffic Density Calculation

Density is a key variable in traffic analysis, however, it is not measured directly by MIDAS. While flow, occupancy and arithmetic mean speed are available directly from MIDAS for each minute, the density can only be estimated indirectly [6, 52]. Traffic density is defined as a spatial average at a fixed time (i.e. the number of vehicles per kilometre of road), however, cross-sectional measurement systems only measure temporal averages at a fixed location (i.e. the loop detector site) [44]. Unlike flow and density, the speed can be defined in two different ways, spatial average (a.k.a. space-mean) and temporal average (a.k.a. time-mean).

Because density is defined as a spatial quantity, the fundamental relation (Equation 3.6) implicitly assumes that speed is a spatial average [140]. The use of temporal averages from cross-sectional loop detectors introduces systematic errors, such that faster vehicles are detected more frequently than slower vehicles leading to a bias towards higher speeds and lower densities [44].

Direct spatial average measurements of density are only practically available with aerial photography [44, 53]. With an understanding of the limitations of using MIDAS measurements, the fundamental relation could be used to obtain an approximation of the density via average flow and temporal average speed. However, in this work, the measured vehicle occupancy is used to approximate the traffic density as is commonly done in practice through [140, 156]:

$$k = \frac{\rho}{L_v + L_s} * C_{lanes} \ , \tag{4.1}$$

where ρ is the measured mean lane occupancy (the fraction of time the detector has a vehicle above). L_v is the mean length of vehicle and L_s is the length of sensor. C_{lanes} is the number of lanes for the road. For the MIDAS system, the sensor length is 2m. The mean vehicle length is calculated by using the vehicle class-specific flow data to calculate a weighted average vehicle length for each minute recorded by the system.

Chapter 5

Prior Origin-Destination Demand Matrix Estimation through Network Partitioning

In this chapter, a novel integrated and scalable method is presented to obtain O-D matrix estimations for large real-world highway networks and evaluate their performance producing accurate UE flow patterns with static TA models. This is accomplished by using network modularity as a basis for dividing up the road network into partitioned subnetworks to reduce the computational difficulty of the prior O-D matrix estimation problem. The technique is applied to the E_2 subnetwork (Section 4.2) and several theoretical networks. It is demonstrated that the incorporation of partitioned O-D estimation within UE flow pattern calculation has the effect of enabling reasonable estimates of the predicted flows and travel times compared to the unpartitioned case, while greatly reducing the computational requirements.

The primary outcomes of the chapter are summarized as:

- A novel method of producing O-D matrices from flow counts is proposed, which utilises network modularity to determine the optimal way to partition the network effectively and automatically.

- The novel method is applied in the calculation of UE flow patterns solely from cross-sectional data on real-world networks without the current size limitations of similar existing O-D estimation techniques.

- Different approaches to utilising the partitioning are investigated, one degenerates the network based on the partitioning, others use the partitioning to focus on estimating the prior matrix from the internal and/or external movements of the partitioned nodes. It is found that using within-the-partition internal estimates for the prior O-D calculation provides the best accuracy. Including the external between-the-partition estimates can help computation time.

Standard validation techniques are often inadequate to assess the effects of the partitioning on the O-D estimates [67]. Comparing the estimated O-D matrix to another validation data source, such as historic trip surveys, is problematic as that is still only a sample of the movements. It is impractical to account for all the movements on a large-scale road network for a ground-truth matrix. For this reason, the validation of the results is done via the relative accuracy, estimating the flow and travel times through the UE flow pattern of a derived static TA model.

5.1 Network Simplification

5.1.1 Network Partitioning

Partitioning is performed on the topographic representation based on community detection using network modularity with the Louvain algorithm [84].

Network modularity measures the relative density of edges inside communities compared to outside communities. It is defined as the difference between the fraction of edges which are between the nodes of a community and the fraction which would be expected if the edges were randomly distributed [157]. It is measured with a scale value ranging from -0.5 to 1 (non-modular to fully modular clustering). By achieving the optimal value for modularity (closest to 1) the results should be the best possible grouping of the network nodes.

The Louvain algorithm uses the following definition for modularity [84]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{z_i z_j}{2m} \right] \delta(c_i, c_j), \tag{5.1}$$

where, for all combinations of node pairs (including $i = j$), A_{ij} is the weight of an edge between nodes i and j taken as the inverse of the edge length; the sum of the weights of the edges attached to node i is represented by $z_i = \sum_j A_{ij}$. The δ-function $\delta(c_i, c_j)$ is 1 if $c_i = c_j$ and 0 otherwise, where c_i is the community to which node i is assigned. Also, $m = \frac{1}{2} \sum_{ij} A_{ij}$ is based on the total weight of all network edges.

As only pairs of nodes belonging to the same community contribute to the modularity calculation, the total modularity of the network can be expressed as the sum of the modularity of each community within a given partitioning. The modularity of a single community C is expressed as [84]:

$$Q_C = \frac{\Omega_{in}}{2m} - \left(\frac{\Omega_{tot}}{2m} \right)^2, \tag{5.2}$$

where Ω_{in} is the sum of edge weights inside community C for each node in the community (meaning each edge is considered twice), Ω_{tot} is the sum of edge weights connected to nodes in community C. Essentially, the first term of the equation is the probability that any edge in the network is fully within community C. The second term is the probability that any edge would have at least one node in community C. Hence, a high modularity indicates a higher number of edges within a community than would be expected if the network was random.

Part of the Louvain algorithm's high efficiency is due to the easy computation of the change in modularity ΔQ from adding or removing a node v into community C [84]. For adding a node to a community:

$$\Delta Q_{add} = \left[\frac{\Omega_{in} + 2z_{v,in}}{2m} - \left(\frac{\Omega_{tot} + z_v}{2m} \right)^2 \right] - \left[\frac{\Omega_{in}}{2m} - \left(\frac{\Omega_{tot}}{2m} \right)^2 - \left(\frac{z_v}{2m} \right)^2 \right], \tag{5.3}$$

and for removing a node from a community,

$$\Delta Q_{remove} = \left[\frac{\Omega_{in} - 2z_{v,in}}{2m} - \left(\frac{\Omega_{tot} - z_v}{2m} \right)^2 + \left(\frac{z_v}{2m} \right)^2 \right] - \left[\frac{\Omega_{in}}{2m} - \left(\frac{\Omega_{tot}}{2m} \right)^2 \right]. \tag{5.4}$$

$z_{v,in}$ is the sum of edge weights from v to other nodes in C. The first part of each expression is the modularity of the newly created community and the second part is the modularity of the original community and node separated. The change in modularity is evaluated via the net change from removing node v from its community and grouping it into its neighbouring community.

$$\Delta Q = \Delta Q_{add} + \Delta Q_{remove} \tag{5.5}$$

Each pass of the Louvain algorithm consists of two main phases, modularity optimisation and community aggregation [84, 86]. The process works by first finding small communities based on optimising modularity for single node communities through calculating the changes in modularity from their combination. Then those small communities are regrouped as aggregate nodes in a new network. The weights of the new edges between the aggregate nodes are the sum of the edge weights between nodes in the corresponding communities. Edges between nodes of the same community create self-loops. The same modularity optimisation process is applied to this new network to see if there are further increases in modularity from combining the new communities. The process iterates over multiple passes of these two phases, calculating modularity changes and regrouping nodes, until there are no more community changes and increases in modularity. The partition is then optimal

for the set resolution. The resolution is a parameter used in the algorithm implementation which controls the size of communities outputted from the hierarchical structure. The method is outlined in Algorithm 4.

Algorithm 4: Description of Louvain algorithm [84, 85]

Input: The network $(\mathcal{V}, \mathcal{A})$, resolution parameter r.

Initialisation:

For the initial partition, set each node v as a separate community C, namely, C_1: v_1, C_2: v_2,...,C_i: v_i.

Phase 1 - Modularity Optimisation:

Step 1.1: Identify all communities connected to node v_i, and calculate ΔQ from moving node v_i to each connected community. Node v_i is moved to the community with maximum ΔQ. If $\Delta Q \leq 0$ for all moves, v_i is not moved.

Step 1.2: Apply to all nodes until no nodes are left to move. This produces a layer of community partitioning.

Phase 2 - Community Aggregation:

Step 2.1: The communities produced in Phase 1 are merged into aggregate nodes $\hat{v}_i$ containing the nodes of the communities. The edges connecting $\hat{v}_i$ are derived from the sum of edges linking the original communities.

Step 2.2: Return to Step 1.1 and apply the steps to the aggregated network until there is no increase in modularity.

Output:

A multilevel community partition hierarchy of which the partition with the highest modularity is chosen as the final result for the provided inputs.

The topographic representation is partitioned using the inverse of the edge distances as the network edge weights, as opposed to the true road distance. This is so that nodes closer on the topographic network are treated as having a stronger connection. In the process, pairs of parallel edges that have opposite flow directions are replaced with undirected edges due to the Louvain implementation used [158]. This does not affect the final result due to carriageways being in identical pairs.

With efficiency for large networks, the Louvain algorithm finds different high modularity partitions for chosen resolutions of community detec-

tion. The resolution size is a parameter of the algorithm that affects the size of the communities, making it larger leads to a smaller number of partitions being produced with a greater number of nodes inside each one [84]. The size of the resolution is varied over a range to produce partition sizes from unpartitioned (resolution equals zero) to the largest partitions when there are only two separate communities (a resolution value which depends on network size). Not every resolution produces a unique number of communities; the lowest resolution that finds each unique number of communities is the one selected.

Each time the Louvain algorithm is run with the same inputs it can produce a variation on the exact partitioning produced due to randomized cluster initialization [87]. As partitioning is primarily used to find communities of different sizes, control of the exact nodes in each partition is not a great concern.

Once we have produced a partitioning result for the given resolution, the new community topographic representation is created from the groupings. The nodes of each partition are grouped into community supernodes. A modified DFS [152] is used to find the neighbours of each partition and establish the community superedges of a new community topographic representation. An example of the process can be seen in Figure 5.1. If multiple edges connect the partitions, then the mean distance of the edges weighted by mean measured flow in all time-bins is used as the community superedge distance, which is used to obtain the community superedge free-flow travel time. The sum of the flows and capacities on the constituent edges are used as the community superedge flow and capacity, respectively. As the partitions are adjacent to each other, it is often the case that only one edge forms the community superedge.

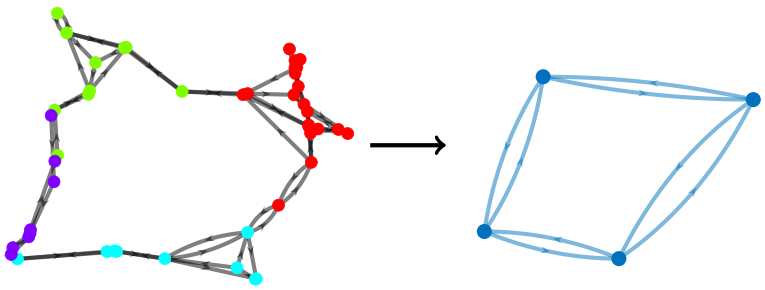


Figure 5.1: *Example of community topographic representation after partitioning using Louvain algorithm. Coloured nodes indicate different communities.*

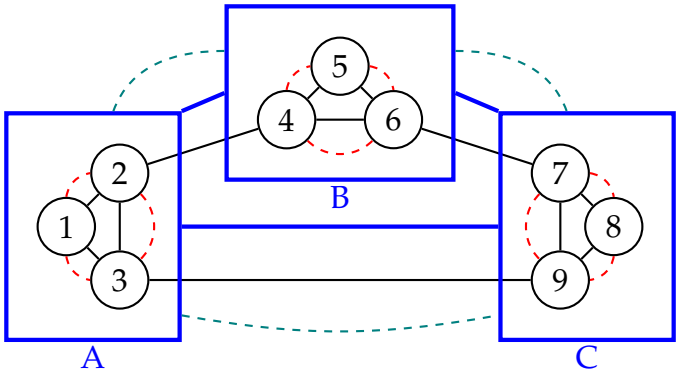


Figure 5.2: *Example nine node topographic network (black) partitioned into three communities. Community topographic is in blue. The green dashed lines are the partitions' external O-D movements, the red dashed are the partitions' internal O-D movements.*

5.1.2 Partitioned Network Demand Matrix Estimation

The community topographic representation can be used to obtain estimates of the uncongested prior O-D demand matrix using the GLS method. Four different ways of utilising the representation for this purpose are investigated: (i) degenerate; (ii) non-degenerate internal-only; (iii) non-degenerate external-only; (iv) non-degenerate internal-external combined. Figure 5.2 shows an illustrative example of a nine node undirected simple graph network to demonstrate the partition grouping with internal and external O-D movements.

(i) Degenerate

In the degenerate O-D estimation, the community topographic representation (Figure 5.2 - blue graph) is used as a substitute network for original topographic representation (Figure 5.2 - black graph). The O-D estimation and adjustment are applied to the flows and structure of the community topographic representation and not the original topographic representation.

In the nine node example, the partitioned community topographic representation is used to produce an O-D estimate, $\mathbf{H_{com}}$, for the partitions A, B and C.

$$\mathbf{H_{com}} = \begin{bmatrix} 0 & H_{com}^{AB} & H_{com}^{AC} \\ H_{com}^{BA} & 0 & H_{com}^{BC} \\ H_{com}^{CA} & H_{com}^{CB} & 0 \end{bmatrix}, \tag{5.6}$$

where each non-zero entry (e.g. H_{com}^{XY}) is an estimate of the demand travelling between the pair of partitions (e.g. X and Y) based on the edge flows of the community superedges (Figure 5.2 - green dashed lines).

This approach reduces the network size as shown in Figure 5.1. It loses the detail of individual road junctions but seeks to preserve some of the main network structure. $\mathbf{H_{com}}$ is used within the TA model to produce estimates of flows and travel times between the partitions on the community topographic representation.

(ii) Non-degenerate internal-only

The non-degenerate approaches aim to find an estimate of the demand for each O-D pair of the original topographic representation through breaking down the problem with the simpler community topographic representation.

The internal approach applies O-D estimation to separately estimate demands for the internal O-D pairs of each partition by applying GLS to

the flows and structure of that partition's subnetwork (Figure 5.2 - red dashed lines). For example, for Partition A in the nine node example, a matrix of demands $\mathbf{H}^{\mathbf{A}}_{\mathbf{int}}$ can be expressed:

$$\mathbf{H}^{\mathbf{A}}_{\mathbf{int}} = \begin{bmatrix} 0 & H^{12} & H^{13} \\ H^{21} & 0 & H^{23} \\ H^{31} & H^{32} & 0 \end{bmatrix}, \tag{5.7}$$

where each non-zero entry is an estimate of the demand travelling between the pair of nodes based on the edge flows of the topographic representation (Figure 5.2 - black graph). It follows the same form for other partitions.

For each partition, the O-D values between the internal nodes will be larger than what would be estimated if the whole unpartitioned network was being used, as all the flows are assumed to be going only between the internal nodes. This is corrected with the help of the O-D adjustment algorithm.

In the non-degenerate internal-only approach, the matrices of demands for each of the partitions are combined into a prior matrix $\mathbf{H}$ by assuming zero demand for the inter-partition O-D pairs. Such that for the nine node example the prior estimate is,

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}^{\mathbf{A}}_{\mathbf{int}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{\mathbf{B}}_{\mathbf{int}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H}^{\mathbf{C}}_{\mathbf{int}} \end{bmatrix}, \tag{5.8}$$

where $\mathbf{0}$ is a matrix of zeros the size of the inter-partition O-D pairs. For the example in Figure 5.2, there are three nodes in each partition and nine inter-partition O-D pairs between a permutation of two partitions. This corresponds to a 3x3 matrix for $\mathbf{0}$.

(iii) Non-degenerate external-only

The non-degenerate external-only approach uses the external partition O-D estimate, $\mathbf{H}_{\mathbf{com}}$, obtained from the community topographic representation. The external partition O-D demands are divided equally between the nodes that comprise the relevant partitions to spread the

demand amongst the O-D pairs of the topographic representation (Figure 5.2 - black graph).

To obtain estimates for the inter-partition demands, the community O-D matrix demands H_{com} are divided by the number of topographic O-D pairs that comprise each partition pair. For example, for partition pair AB, the number of nodes in A, u^A, is three and the number of nodes in B, u^B, is three so the number of O-D pairs is $u^{AB} = u^A * u^B = 9$. The value for each pair is then $H^{AB}_{com}/9$. Then, in matrix form, for partition pair AB with $\mathbf{l}_A$ as a column vector of ones the length of u^A, and $\mathbf{l}_B$ as a column vector of ones the length of u^B,

$$\hat{\mathbf{H}}^{\mathbf{AB}}_{\mathbf{ext}} = \frac{H^{AB}_{com}}{u^{AB}} \mathbf{l}_A \mathbf{l}'_B. \tag{5.9}$$

External-only assumes zero values for the demands between the O-D pairs internal to the partitions, resulting in the following prior matrix,

$$\mathbf{H} = \begin{bmatrix} \mathbf{0} & \hat{\mathbf{H}}^{\mathbf{AB}}_{\mathbf{ext}} & \hat{\mathbf{H}}^{\mathbf{AC}}_{\mathbf{ext}} \\ \hat{\mathbf{H}}^{\mathbf{BA}}_{\mathbf{ext}} & \mathbf{0} & \hat{\mathbf{H}}^{\mathbf{BC}}_{\mathbf{ext}} \\ \hat{\mathbf{H}}^{\mathbf{CA}}_{\mathbf{ext}} & \hat{\mathbf{H}}^{\mathbf{CB}}_{\mathbf{ext}} & \mathbf{0} \end{bmatrix}, \tag{5.10}$$

where $\mathbf{0}$ is a matrix of zeros the size of the intra-partition O-D pairs. For the example in Figure 5.2, there are three nodes in each partition and six O-D pairs between them. As in Equation 5.7, the demand from a node to itself is included but set to zero. This then corresponds to a 3x3 matrix for $\mathbf{0}$.

(iv) Non-degenerate internal-external combined

In the non-degenerate internal-external combined approach, a prior matrix is formed using both internal and external estimations without any O-D demands assumed zero:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}^{\mathbf{A}}_{\mathbf{int}} & \hat{\mathbf{H}}^{\mathbf{AB}}_{\mathbf{ext}} & \hat{\mathbf{H}}^{\mathbf{AC}}_{\mathbf{ext}} \\ \hat{\mathbf{H}}^{\mathbf{BA}}_{\mathbf{ext}} & \mathbf{H}^{\mathbf{B}}_{\mathbf{int}} & \hat{\mathbf{H}}^{\mathbf{BC}}_{\mathbf{ext}} \\ \hat{\mathbf{H}}^{\mathbf{CA}}_{\mathbf{ext}} & \hat{\mathbf{H}}^{\mathbf{CB}}_{\mathbf{ext}} & \mathbf{H}^{\mathbf{C}}_{\mathbf{int}} \end{bmatrix}. \tag{5.11}$$

In all the non-degenerate approaches, the prior matrix $\mathbf{H}$ is used in the O-D adjustment algorithm to produce a final O-D demand matrix which is used in a static TA model for the whole topographic network.

5.1.3 Example of the application of the partitioned network demand estimation

To illustrate the process of applying the partitioned network demand estimation, a small six node example network is used (Figure 5.3). The matrices are calculated using the degenerate approach and the three non-degenerate approaches.

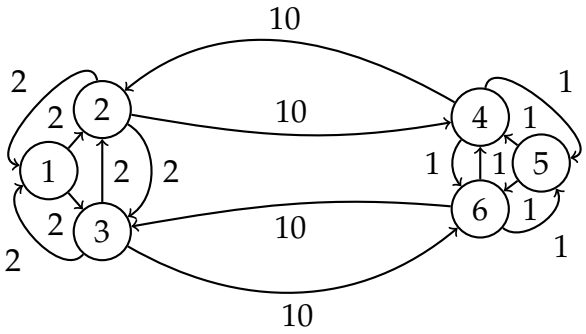


Figure 5.3: *Diagram of a small test network. The network contains 6 nodes, 16 edges. Edge labels are the edge weights. Node labels are the node numbers.*

A ground-truth O-D matrix is created for the specified network that randomly assigns a number between 0 and 10 to each O-D pair. With the assigned O-D matrix, the average flows on the network are created by using the Frank-Wolfe algorithm to solve for UE (Equation 3.10). This provides an average flow on each edge that is used to generate a sample of days of flows by using a random Poisson generator. The number of simulated days is set to be the number of edges in the network multiplied by 2.5. This flow sample is then used in the same processes described in the methodology.

The network is taken to be uncongested so the congestion functions used are just the edge distances (independent of flow). The edge weights are chosen to ensure every O-D pair has a unique single shortest path. In this example, a single route is assumed between each O-D pair for the application of GLS for prior matrix estimation. The UE effectively assigns each O-D demand to the shortest path for that O-D. These simpli-

fications and the assumption of Poisson-distributed daily flows make an idealised example that suits the assumptions of the GLS method.

For the example, the created ground-truth O-D matrix, $\mathbf{D}^{true}$, is:

$$\mathbf{D}^{true} = \begin{bmatrix} 0 & 1 & 2 & 3 & 10 & 3 \\ 1 & 0 & 4 & 9 & 1 & 2 \\ 2 & 10 & 0 & 10 & 10 & 3 \\ 3 & 4 & 9 & 0 & 7 & 3 \\ 1 & 3 & 8 & 7 & 0 & 9 \\ 5 & 4 & 6 & 7 & 1 & 0 \end{bmatrix} \tag{5.12}$$

This produces the ground-truth mean flow vector for the 16 edges, $\bar{\mathbf{x}}^{true}$, where $\bar{x}_{ij}^{true}$ indicates the flow from node i to node j:

$$\bar{\mathbf{x}}^{true} = \begin{bmatrix} \bar{x}_{12}^{true} \\ \bar{x}_{13}^{true} \\ \bar{x}_{21}^{true} \\ \bar{x}_{23}^{true} \\ \bar{x}_{24}^{true} \\ \bar{x}_{31}^{true} \\ \bar{x}_{32}^{true} \\ \bar{x}_{36}^{true} \\ \bar{x}_{42}^{true} \\ \bar{x}_{45}^{true} \\ \bar{x}_{46}^{true} \\ \bar{x}_{54}^{true} \\ \bar{x}_{56}^{true} \\ \bar{x}_{63}^{true} \\ \bar{x}_{64}^{true} \\ \bar{x}_{65}^{true} \end{bmatrix} = \begin{bmatrix} 14 \\ 5 \\ 5 \\ 4 \\ 25 \\ 7 \\ 10 \\ 26 \\ 15 \\ 18 \\ 14 \\ 11 \\ 17 \\ 28 \\ 21 \\ 11 \end{bmatrix} \tag{5.13}$$

Using the values of $\bar{\mathbf{x}}^{true}$ as the mean flows in a Poisson generator, a 40-day sample of artificial flows is generated, $\{\mathbf{x}_{(j)}^{true}; j \in [\![\mathcal{J}]\!]\}$ where $|\mathcal{J}| = 40$.

The example network is partitioned into two communities (Figure 5.4) using the Louvain algorithm with a resolution parameter of 0.5 (Algorithm 4). The weights of the community topographic edges are the average of the constituent topographic edges. The ground-truth demand matrix for the community topographic is the sum of the inter-partition demands of the topographic representation, such that, $\mathbf{D}^{true}_{com}$, is:

$$\mathbf{D}^{true}_{com} = \begin{bmatrix} \sum_{i=1}^{3}\sum_{j=1}^{3} D^{true}_{i,j} & \sum_{i=1}^{3}\sum_{j=4}^{6} D^{true}_{i,j} \\ \sum_{i=4}^{6}\sum_{j=1}^{3} D^{true}_{i,j} & \sum_{i=4}^{6}\sum_{j=4}^{6} D^{true}_{i,j} \end{bmatrix} = \begin{bmatrix} 0 & 51 \\ 43 & 0 \end{bmatrix} \tag{5.14}$$

The ground-truth flows on the community topographic, $\bar{\mathbf{x}}^{true}_{com}$, are the sum of the ground-truth flows on the constituent topographic edges, such that:

$$\bar{\mathbf{x}}^{true}_{com} = \begin{bmatrix} \bar{x}^{true}_{24} + \bar{x}^{true}_{36} \\ \bar{x}^{true}_{42} + \bar{x}^{true}_{63} \end{bmatrix} = \begin{bmatrix} 51 \\ 43 \end{bmatrix} \tag{5.15}$$

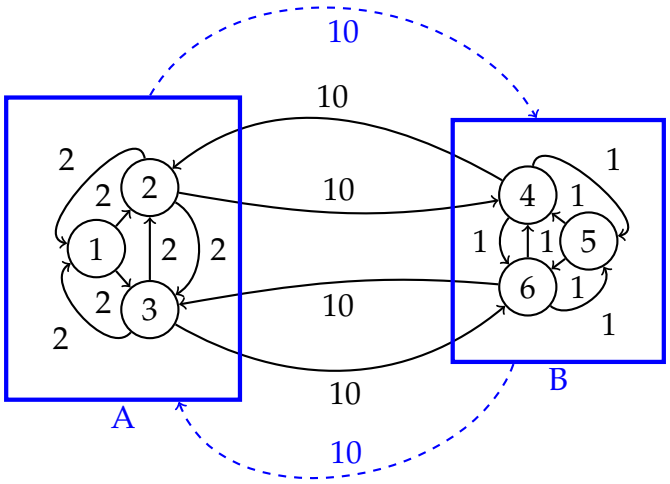


Figure 5.4: *Diagram of small test network partitioned into two communities. Community topographic is in blue. Edge labels are the edge weights. Node labels are the node numbers.*

For the partitioned network, the estimate of the prior matrix for the community topographic, $\mathbf{H_{com}}$, is:

$$\mathbf{H_{com}} = \begin{bmatrix} 0 & H_{com}^{AB} \\ H_{com}^{BA} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 50.32 \\ 43.60 & 0 \end{bmatrix} \tag{5.16}$$

This is the prior matrix to be used in the degenerate method. Using $\mathbf{H_{com}}$, the estimates for the external approach, $\hat{\mathbf{H}}_{\mathbf{ext}}$, can be obtained with $u^{AB} = u^{BA} = 9$:

$$\hat{\mathbf{H}}_{\mathbf{ext}}^{\mathbf{AB}} = \frac{H_{com}^{AB}}{u^{AB}} \mathbf{1}_A \mathbf{1}_B' = \begin{bmatrix} 5.59 & 5.59 & 5.59 \\ 5.59 & 5.59 & 5.59 \\ 5.59 & 5.59 & 5.59 \end{bmatrix} \tag{5.17}$$

$$\hat{\mathbf{H}}_{\mathbf{ext}}^{\mathbf{BA}} = \frac{H_{com}^{BA}}{u^{BA}} \mathbf{1}_B \mathbf{1}_A' = \begin{bmatrix} 4.84 & 4.84 & 4.84 \\ 4.84 & 4.84 & 4.84 \\ 4.84 & 4.84 & 4.84 \end{bmatrix} \tag{5.18}$$

Also, the estimates for the internal approach, $\mathbf{H_{int}}$, are obtained for A and B:

$$\mathbf{H}_{\mathbf{int}}^{\mathbf{A}} = \begin{bmatrix} 0 & H^{12} & H^{13} \\ H^{21} & 0 & H^{23} \\ H^{31} & H^{32} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 12.77 & 4.64 \\ 4.80 & 0 & 2.97 \\ 5.66 & 9.64 & 0 \end{bmatrix} \tag{5.19}$$

$$\mathbf{H}_{\mathbf{int}}^{\mathbf{B}} = \begin{bmatrix} 0 & H^{45} & H^{46} \\ H^{54} & 0 & H^{56} \\ H^{64} & H^{65} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 18.27 & 9.98 \\ 10.03 & 0 & 15.72 \\ 20.40 & 7.95 & 0 \end{bmatrix} \tag{5.20}$$

With these estimates, the prior matrices, $\mathbf{H}$, for the four different applications can be obtained:

(i) Degenerate:

$$\mathbf{H}^i = \mathbf{H_{com}} = \begin{bmatrix} 0 & 50.32 \\ 43.60 & 0 \end{bmatrix} \tag{5.21}$$

(ii) Internal-only:

$$\mathbf{H}^{ii} = \begin{bmatrix} \mathbf{H}^{\mathbf{A}}_{\mathbf{int}} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{\mathbf{B}}_{\mathbf{int}} \end{bmatrix} = \begin{bmatrix} 0 & 12.77 & 4.64 & 0 & 0 & 0 \\ 4.80 & 0 & 2.97 & 0 & 0 & 0 \\ 5.66 & 9.64 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 18.27 & 9.98 \\ 0 & 0 & 0 & 10.03 & 0 & 15.72 \\ 0 & 0 & 0 & 20.40 & 7.95 & 0 \end{bmatrix} \tag{5.22}$$

(iii) External-only:

$$\mathbf{H}^{iii} = \begin{bmatrix} \mathbf{0} & \hat{\mathbf{H}}^{\mathbf{AB}}_{\mathbf{ext}} \\ \hat{\mathbf{H}}^{\mathbf{BA}}_{\mathbf{ext}} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 5.59 & 5.59 & 5.59 \\ 0 & 0 & 0 & 5.59 & 5.59 & 5.59 \\ 0 & 0 & 0 & 5.59 & 5.59 & 5.59 \\ 4.84 & 4.84 & 4.84 & 0 & 0 & 0 \\ 4.84 & 4.84 & 4.84 & 0 & 0 & 0 \\ 4.84 & 4.84 & 4.84 & 0 & 0 & 0 \end{bmatrix} \tag{5.23}$$

(iv) Internal-external combined:

$$\mathbf{H}^{iv} = \begin{bmatrix} \mathbf{H}^{\mathbf{A}}_{\mathbf{int}} & \hat{\mathbf{H}}^{\mathbf{AB}}_{\mathbf{ext}} \\ \hat{\mathbf{H}}^{\mathbf{BA}}_{\mathbf{ext}} & \mathbf{H}^{\mathbf{B}}_{\mathbf{int}} \end{bmatrix} = \begin{bmatrix} 0 & 12.77 & 4.64 & 5.59 & 5.59 & 5.59 \\ 4.80 & 0 & 2.97 & 5.59 & 5.59 & 5.59 \\ 5.66 & 9.64 & 0 & 5.59 & 5.59 & 5.59 \\ 4.84 & 4.84 & 4.84 & 0 & 18.27 & 9.98 \\ 4.84 & 4.84 & 4.84 & 10.03 & 0 & 15.72 \\ 4.84 & 4.84 & 4.84 & 20.40 & 7.95 & 0 \end{bmatrix} \tag{5.24}$$

Using these respective prior matrices in the O-D adjustment algorithm (Section 3.5.2) together with $\bar{\mathbf{x}}_{true}$ as the observed flow vector, the estimates of the demand matrix, $\hat{\mathbf{D}}$, can be obtained. This can be used to solve the UE TAP to obtain estimates of, $\hat{\mathbf{x}}$, the flow column vector (written as row vectors to save space). The results for the four approaches are:

(i) Degenerate:

$$\hat{\mathbf{D}}^{i} = \begin{bmatrix} 0 & 50.32 \\ 43.60 & 0 \end{bmatrix} \tag{5.25}$$

$$\hat{\mathbf{x}}^{i\prime} = [50.32, 43.60] \tag{5.26}$$

(ii) Internal-only:

$$\hat{\mathbf{D}}^{ii} = \begin{bmatrix} 0 & 9.84 & 1.89 & 4.10 & 0.50 & 4.35 \\ 2.79 & 0 & 3.62 & 7.03 & 3.43 & 3.28 \\ 2.98 & 9.85 & 0 & 4.00 & 5.15 & 7.10 \\ 2.44 & 4.43 & 4.09 & 0 & 14.67 & 6.23 \\ 0.55 & 2.56 & 5.14 & 8.15 & 0 & 13.03 \\ 5.15 & 1.34 & 7.83 & 17.30 & 6.00 & 0 \end{bmatrix} \tag{5.27}$$

$$\begin{aligned} \hat{\mathbf{x}}^{ii\prime} = [&14.44, 6.23, 5.77, 3.62, 18.33, 8.13, 9.85, 20.60, 11.31, \\ &18.60, 13.60, 11.27, 18.18, 22.21, 22.64, 11.15] \end{aligned} \tag{5.28}$$

(iii) External-only:

$$\hat{\mathbf{D}}^{iii} = \begin{bmatrix} 0 & 1.01 & 0 & 4.71 & 7.06 & 5.01 \\ 0.20 & 0 & 1.94 & 3.70 & 6.06 & 4.50 \\ 0.26 & 4.81 & 0 & 8.94 & 6.83 & 5.34 \\ 2.07 & 2.54 & 6.45 & 0 & 2.35 & 0.80 \\ 3.77 & 4.23 & 9.27 & 1.70 & 0 & 3.62 \\ 5.91 & 6.13 & 5.65 & 3.60 & 1.49 & 0 \end{bmatrix} \tag{5.29}$$

$$\hat{\mathbf{x}}^{iii\prime} = [12.78, 5.01,\ 6.04, 1.94, 26.03, 6.17, 4.81, 26.12, 18.74,\ 15.47, 11.75, 9.70, 12.89, 27.29, 18.67, 8.33] \tag{5.30}$$

(iv) Internal-external combined:

$$\hat{\mathbf{D}}^{iv} = \begin{bmatrix} 0 & 10.35 & 2.40 & 4.74 & 1.84 & 5.23 \\ 2.99 & 0 & 3.43 & 7.16 & 4.26 & 4.28 \\ 3.56 & 9.79 & 0 & 4.93 & 5.80 & 7.47 \\ 3.10 & 4.91 & 5.00 & 0 & 15.37 & 7.09 \\ 1.38 & 3.19 & 5.78 & 8.31 & 0 & 13.62 \\ 5.78 & 2.38 & 7.88 & 17.87 & 6.29 & 0 \end{bmatrix} \tag{5.31}$$

$$\hat{\mathbf{x}}^{iv\prime} = [16.93, 7.63,\ 7.46, 3.43, 22.28, 9.35, 9.79, 23.42, 14.96,\ 21.47, 16.38, 12.88, 19.40, 24.45, 25.18, 12.09] \tag{5.32}$$

Outputs for demand estimation without using partitioning

To compare the results, the matrix and flows are also calculated without the use of partitioning, that is using the O-D matrix estimation and adjustment on the whole topographic graph. The unpartitioned results are:

$$\mathbf{H}^{whole} = \begin{bmatrix} 0 & 9.47 & 2.55 & 1.48 & 3.07 & 1.56 \\ 3.16 & 0 & 1.99 & 16.17 & 1.48 & 2.10 \\ 3.61 & 6.75 & 0 & 3.37 & 2.32 & 18.78 \\ 0.95 & 10.98 & 3.34 & 0 & 11.53 & 3.68 \\ 2.67 & 1.03 & 2.48 & 8.22 & 0 & 9.69 \\ 0.79 & 2.24 & 19.11 & 12.49 & 6.17 & 0 \end{bmatrix} \tag{5.33}$$

$$\hat{\mathbf{D}}^{whole} = \begin{bmatrix} 0 & 9.46 & 2.85 & 1.18 & 3.38 & 1.46 \\ 3.05 & 0 & 2.82 & 15.88 & 1.80 & 2.75 \\ 4.25 & 8.07 & 0 & 4.17 & 2.44 & 18.38 \\ 0.29 & 10.43 & 4.24 & 0 & 12.14 & 4.62 \\ 2.02 & 0.49 & 3.81 & 8.23 & 0 & 11.06 \\ 1.40 & 2.89 & 19.07 & 13.69 & 6.70 & 0 \end{bmatrix} \tag{5.34}$$

$$\hat{\mathbf{x}}^{whole\prime} = [14.02, 4.31, 5.36, 2.82, 24.99, 5.65, 8.07, 26.45, 16.13, 17.33, 11.61, 10.75, 14.87, 28.53, 20.76, 9.14] \tag{5.35}$$

Comparing the results

To compare the results for the flows produced, relative errors for the UE assignment prediction are used. The Mean Absolute Percentage Error (MAPE) for flows is calculated by:

$$MAPE^x = \frac{1}{n} \sum_{a=1}^{n} \frac{|\hat{x}_a - \bar{x}_a^{true}|}{\bar{x}_a^{true}}, \tag{5.36}$$

where a indicates one of the 16 edges of the network ($n = 16$). In the degenerate case, $\bar{x}_{com}^{true}$ is used for comparison ($n = 2$).

The results for the partitioned demand estimation approaches are presented together with unpartitioned estimation in Table 5.1. It can be seen that degenerate estimation provides a very low MAPE, which can be expected as the network is reduced to a simple two node network, making the result trivial. It can be seen that the unpartitioned approach provides a lower MAPE than the non-degenerate approaches. This is due to the whole network being considered in all the calculation steps, for both O-D prior estimation and adjustment.

	MAPE (%)
Unpartitioned	9.63
Internal-only	11.74
External-only	17.41
Internal-External combined	18.97
Degenerate	1.36

Table 5.1: *Comparison of the different approaches to demand estimation when applied to a small test network.*

5.2 Results

In the following section, the effect of applying the different demand estimation approaches is investigated on different sizes of partitions to understand the impact on result accuracy and computational requirements.

5.2.1 Testing the different applications of the partitioning on the Sioux Falls benchmark network

To test the different applications of network partitioning over a range of different partition sizes, they are first applied to a benchmark network using artificially-generated data. The testing is done on the Sioux Falls (USA) network commonly used in TA model testing with its network data taken from [159].

The Sioux Falls network provides a small 24 network with a sample ground-truth O-D matrix (Figure 5.5). Capacity is provided for each edge in the network data. It uses BPR with standard coefficients for the congestion functions on all edges. The length is provided for each edge of the network, from which the free-flow travel time is obtained using 70 mph as the free-flow speed.

In a similar process to Section 5.1.3, the average flows on the network were created by using the Frank-Wolfe algorithm to solve for UE for the ground-truth O-D matrix (Equation 3.10). This provided an average flow on each edge that could be used to generate a sample of days of flows by using a random Poisson generator. The number of simulated

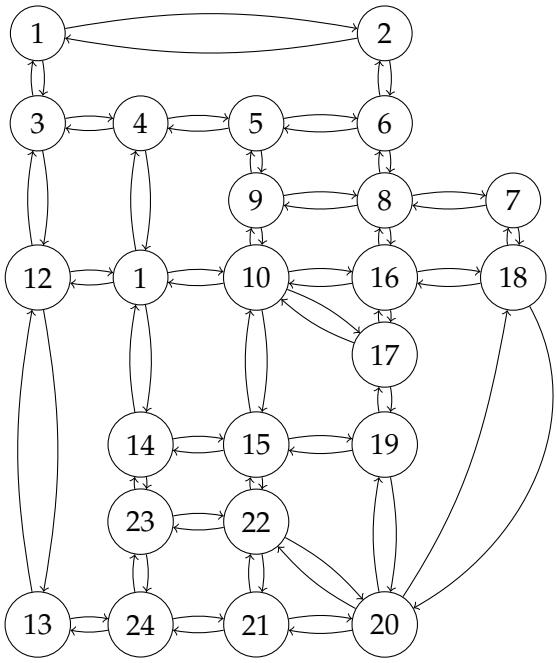


Figure 5.5: *Diagram of the Sioux Falls benchmark network. The network contains 24 nodes, 76 edges. Network data (e.g. edge capacities, lengths etc.) obtained from [159]. Node labels are the node numbers.*

days was set to be the number of edges in the network multiplied by 2.5. This flow sample was then treated as the observed flow pattern for testing.

Relative errors in the UE flow prediction are used to evaluate the performance. The Absolute Percentage Errors (APE) are calculated as:

$$APE_a^x = \frac{|\hat{x}_a - \bar{x}_a^{true}|}{\bar{x}_a^{true}}, \tag{5.37}$$

for each edge a in the network. The notation is the same as in Eq. 5.36.

To investigate the effect of partition resolution on each of the types of partitioned network demand estimation techniques, the Louvain resolution parameter was varied to produce different resulting partitions.

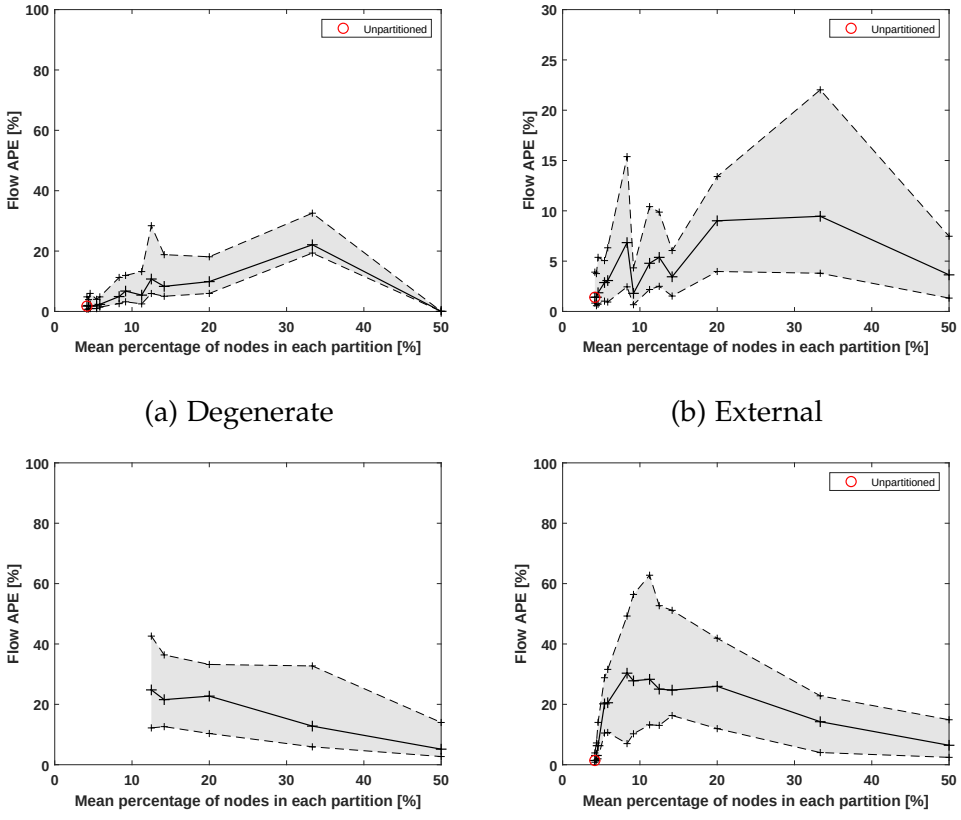
The results for flow prediction error for the four estimation approaches can be seen to exhibit different patterns as the size of the partitions varies (Figure 5.6). The error in prediction can be compared to the result for the unpartitioned case, which is a benchmark for the methods. The unpartitioned case gives the same value for all methods except internal-only, for which it was unattainable as each community only contains one node in that case.

Comparing the different partitioning approaches, it can be seen that the flow prediction accuracy for degenerate varies less for the partitions with a smaller percentage of the total nodes inside (a larger number of partitions); however, as the size of the partitions increases, the flow prediction has a larger variance between partition sizes. The relative error for flow is low for the largest partition size. As in Section 5.1.3, this can be attributed to the network being degenerated to a two node, two edge network so the demand prediction becomes trivial.

Several trends can be seen for the non-degenerate approaches (internal-only; external-only; internal-external combined). As the percentage of average nodes in a partition increases, the results for the external-only method show a broadly increasing trend in flow prediction error. This is due to the prior matrix increasingly basing the individual O-D movements on a smaller subset of topographic edges. Less information is available so the prior matrix moves further from its best estimate, which is the unpartitioned case.

For internal-only, as the size of the partitions increases to include more nodes, flow prediction improves. However, the results were not available for the smallest sizes of partitions. This is because the estimate of the prior matrix was too inaccurate for the O-D adjustment process to converge.

For internal-external combined, it can be seen that there is a degradation in accuracy for flow prediction from the unpartitioned case to approximately 10% of total nodes. After this, the results improve with increases in the partition size before they start to level off, matching the accuracy of internal-only at the largest partition size.



(a) Degenerate

(b) External

(c) Internal

(d) Internal and External

Figure 5.6: *Plot of Absolute Percentage Error in user-equilibrium flow prediction for each partition size investigated on the Sioux Falls benchmark network. Solid line is median error and dashed lines indicate the IQR. Lines are used as visual aid for the individual point results.*

5.2.2 Accuracy of different applications of the partitioning on the English SRN

To understand the performance of partitioned network demand estimation on real-world road traffic, analysis was carried out on the E_2 subnetwork using MIDAS data taken from September 2018 to May 2019. The BPR formulation with standard coefficients ($\alpha = 0.15$, $\beta = 4$) was used for all edges to isolate the effect of congestion function choice. To investigate the effect of partition resolution on each of the types of partitioning matrix estimation techniques, the Louvain resolution parameter was varied to evaluate the effect on the TA model accuracy and computation requirements of the resulting partitions. Without a ground-truth O-D demand matrix, making an assessment based on flow and travel time estimation is a practical way to validate the accuracy of the calculated O-D matrices. The computation time results in this section and Section 5.2.5 refer to the time taken to calculate the O-D demand matrix and solve the TAP.

Relative errors in the flow and travel times of the UE assignment prediction are used to evaluate the performance. The Absolute Percentage Errors (APE) are calculated as:

$$APE_a^t = \frac{|t_{p,a}^{user} - t_{p,a}^{obs}|}{t_{p,a}^{obs}}, \tag{5.38}$$

for travel time, while

$$APE_a^x = \frac{|x_{p,a}^{user} - x_{p,a}^{obs}|}{x_{p,a}^{obs}}, \tag{5.39}$$

is used for flows. For each time-bin p and edge a, $x_{p,a}^{obs}$ is the observed flow and $t_{p,a}^{obs}$ is the travel time derived from observed speed. The values are the mean within each time-bin over the fitting period. $t_{p,a}^{user}$ is the predicted travel time derived from the congestion function using $x_{p,a}^{user}$, which is the edge flow value predicted by the model through solving the UE TAP with the calculated O-D matrix.

The results for the four estimation approaches can be seen to exhibit different patterns as the size of the partitions varies (Figure 5.7 and 5.8),

similar to those previously found in Section 5.2.1. The error in flow and travel time prediction can be compared to the result for the unpartitioned case, which is a benchmark for the methods. The unpartitioned case gives the same value for all methods except internal-only, for which it was unattainable as each community only contains one node in that case.

Comparing the different approaches for using the partitioning, it can be seen that there is considerably different behaviour between degenerate and non-degenerate approaches (Figure 5.7). The flow prediction accuracy for degenerate varies less for the partitions with a smaller percentage of the total nodes inside (a larger number of partitions); however, as the size of the partitions increases, the flow prediction has a larger variance between resolutions. The relative error for flow is low for the largest partition size. This can be attributed to the network being degenerated to a two node, two edge system so the demand prediction through GLS becomes trivial. It can be seen that the time prediction accuracy for the degenerate method stays broadly similar before decreasing slightly as the partitions become larger and less numerous.

Between the other non-degenerate methods (internal-only; external-only; internal-external combined), in Figures 5.7 and 5.8 several trends can be seen. With internal-only, as the size of the partitions increases to include more nodes the results for both flow and time improve up to the 11% point. Between 11-50% the median is approximately constant. In Figure 5.9, the computation time for internal-only also begins to level off past the 11% point. This implies the results for using the internal-only approach are similar for the 11-50% partition size range in both accuracy and computation time. The results for internal-only were not available for the smallest five resolutions of partitioning. This is because the estimate of the prior matrix was too inaccurate for the O-D adjustment process to converge.

As the percentage of average nodes in a partition increases, the results for the external-only method show a broadly linear increase in error for flow and time prediction as well as computation time. This is due to the prior matrix increasingly basing the individual O-D movements on a smaller subset of topographic edges. Less information is available

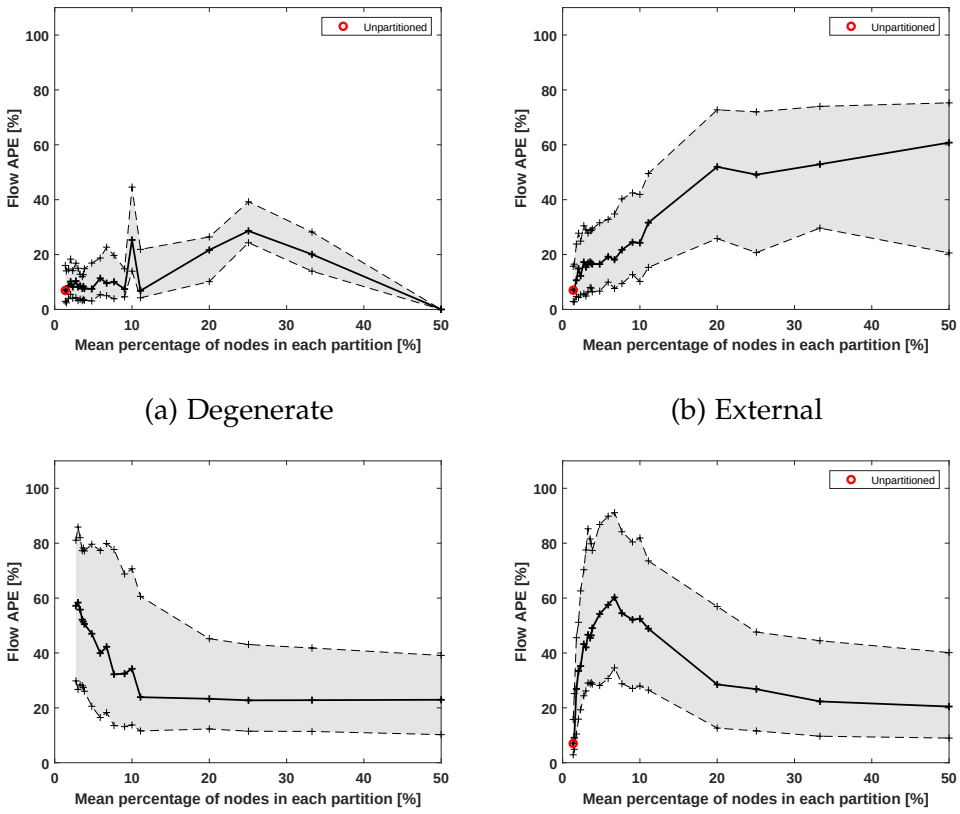
so the prior matrix moves further from its best estimate which is the unpartitioned case.

When internal and external estimates are combined to create the prior matrix, it can be seen that there is a degradation in accuracy for flow and time prediction from the unpartitioned case to the point of approximately 7% of total nodes. After this, the results for both flow and time improve with increases in the partition size before they start to level off. At the largest partition size it can be seen that the accuracy matches the internal-only result but with less computation time (Figure 5.9).

For all the approaches, the travel time APEs are generally lower than the flow APEs. The differences in patterns between flow and travel time can be explained by the non-linearity of the congestion functions that are used to obtain the UE modelled travel times from the UE modelled flows.

To provide some context to the errors obtained, the results can be compared with other recent work such as [67]. The analysis in that work used a more capable but computationally demanding method of network tomography than GLS to obtain an O-D matrix for a smaller network in Melbourne, Australia city centre (23 nodes, 54 edges). They found a mean APE of 24.18% for flow using simulated data. Using travel time data from Uber and Syic, they found mean APEs in travel time predictions of 18-33%. Although these results are not directly comparable, as the networks and methods are different, it highlights the approximate size of error for current static TA models using network tomography O-D estimation on real-world networks.

For the road subnetwork, the memory requirement of the four techniques for all partition sizes did not vary much, staying between 20.8-20.9 GB in all cases. The road network is not particularly large (73 nodes, 156 edges) so memory is not the concern. The calculations for the results were all performed on a Dell PowerEdge C6320 with 2.4GHz Intel Xeon E5-2630 v3 CPU. The implication of the results is that the best option would be to use the largest partition possible with the internal-external combined or the internal-only methods.



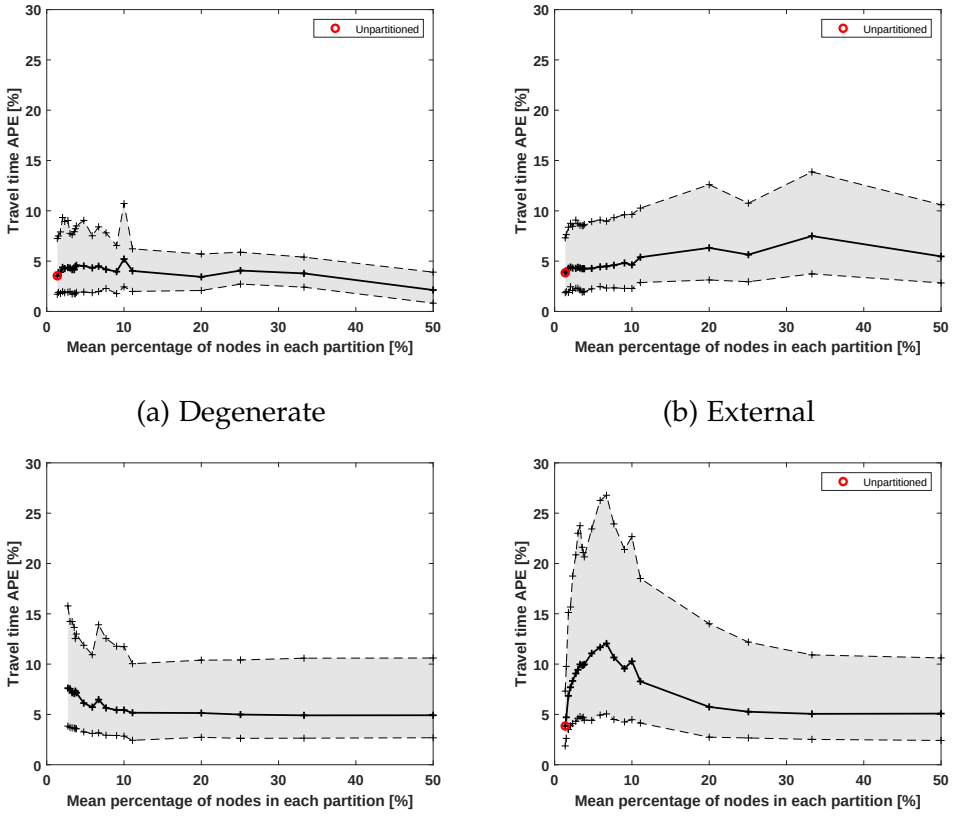
(a) Degenerate

(b) External

(c) Internal

(d) Internal and External

Figure 5.7: *Plot of Absolute Percentage Error in user-equilibrium flow prediction for each partition size investigated on the E_2 subnetwork for September 2018 to May 2019. Solid line is median error and dashed lines indicate the IQR. Lines are used as visual aid for the individual point results.*



(a) Degenerate

(b) External

(c) Internal

(d) Internal and External

Figure 5.8: *Plot of Absolute Percentage Error in user-equilibrium travel time prediction for each partition size investigated on the E_2 subnetwork for September 2018 to May 2019. Solid line is median error and dashed lines indicate the IQR. Lines are used as visual aid for the individual point results.*

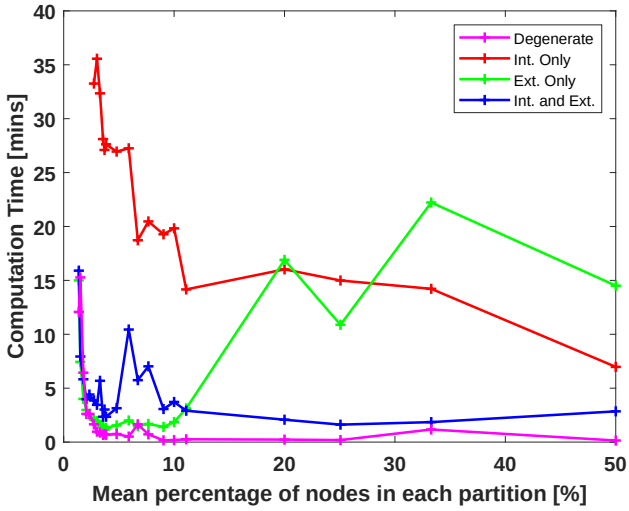


Figure 5.9: *Computation time of results for each partition size investigated on the E_2 subnetwork for September 2018 to May 2019. The time includes the calculation of the Origin-Destination (O-D) demand matrix and solution of the Traffic Assignment Problem (TAP). Lines are used as visual aid for the individual point results.*

5.2.3 Creating the Artificially-generated Networks for Further Investigation

To simulate the artificially-generated networks to further analyse the methods, the nine-node example in Figure 5.2 was used as a building block. The single undirected edges of the simple graph were replaced with edges in both directions which are assigned equal distances. The process added another of the nine-blocks to the network connecting a random node on the existing network to a random node on the new nine-node block. The random chosen nodes were limited to the nodes with order less than 6 (in and out combined). In the example of the process shown in Figure 5.10, this restricted the connections to nodes 1, 5 and 8. A larger distance for the dual edges connecting the blocks than those within the nine-node unit is used.

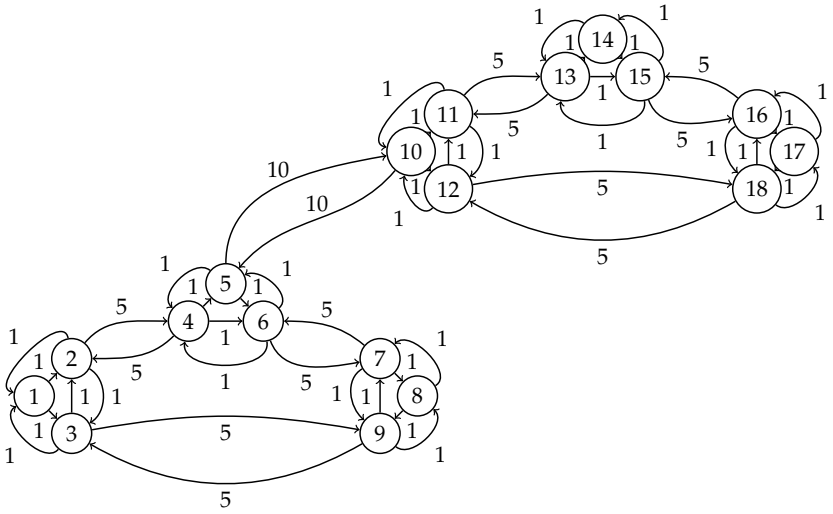


Figure 5.10: *Example of nine node weighted directed graph used to build a more complex artificially-generated network. Edge labels are the edge weights. Node labels are the node numbers.*

This approach was chosen to create the networks as it represents a suitable approximation of how conurbations connect together and it contains a visible modular structure amenable to the methods applied.

A number of network blocks were connected to make the required size of test network (in multiples of nine). Once the network was specified, as in Sections 5.1.3 and 5.2.1, an O-D matrix was created for the network which randomly assigned a number between 0 and 10 to each O-D pair. The network was taken to be uncongested so the congestion function used was just the edge distance (independent of flow).

With the assigned O-D matrix the average flows on the network were created by using the Frank-Wolfe algorithm to solve for UE (Equation 3.10). This provided an average flow on each edge which could be used to generate a sample of days of flows by using a random Poisson generator. The number of simulated days was set to be the number of edges in the network multiplied by 2.5. This flow sample was then used in the same processes described in the methodology to generate results.

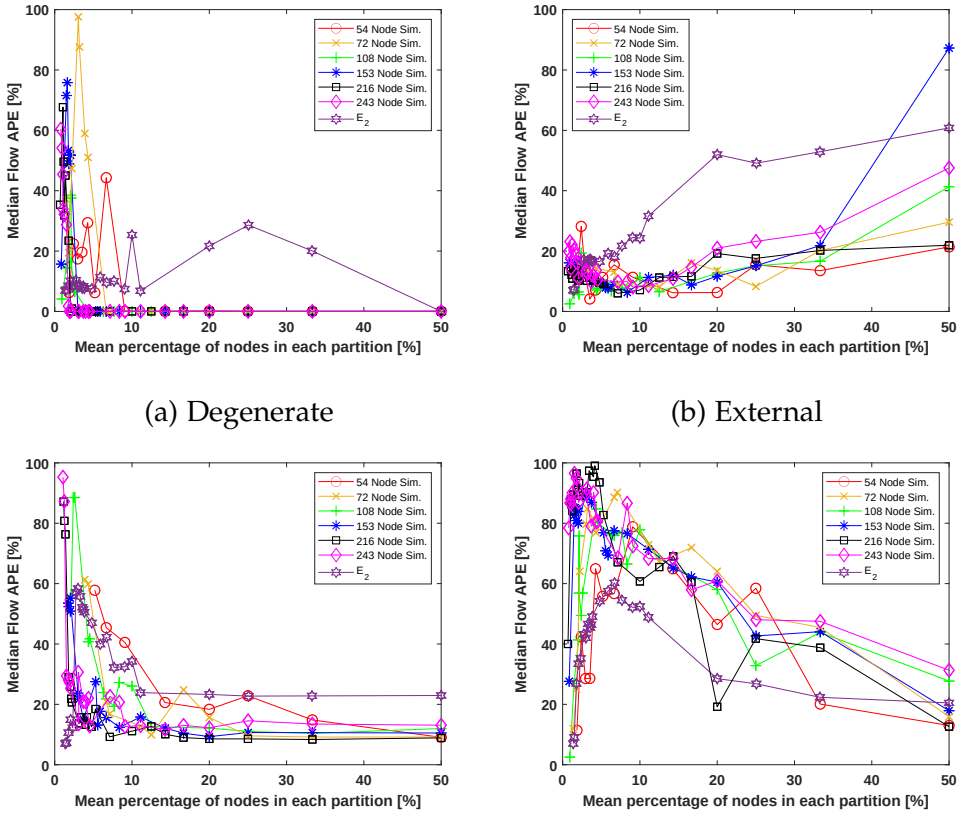
For each size of network, three iterations were trialled. The random aspect to the network creation did not have a considerable effect on the results.

5.2.4 Comparison of the Results with Different Sized Artificially-generated Networks

To investigate how the size of the network influences the results of the different methods, the same tests were carried out on additional artificially-generated networks of a range of sizes (Section 5.2.3). The analysis was carried out on simulated flow data without travel time.

Similar trends to the English SRN network can be seen when the techniques are applied to the artificially-generated networks (Figure 5.11). For internal-only, there is a peak in error for small partition sizes with no results produced for the smallest partitions. The internal-only results level out after around 11-13% of nodes (i.e. eight or nine partitions). For external-only there is a steady increase in flow error as the partition size increases. The results for the internal-external combined method show the same characteristic triangle shape with an initial increase followed by a decrease in error.

For the degenerate method, the trend is different for the artificially-generated networks. With the artificially-generated networks, there is a peak in error between 0-10% and then the error reduces to almost zero for the larger partition sizes. This can be attributed to the artificially-generated networks having no congestion and the simulated flows being created with a Poisson distribution, so that for the smaller network sizes (larger partitions) very accurate estimates of the demand are obtained.



(a) Degenerate

(b) External

(c) Internal

(d) Internal and External

Figure 5.11: *Plot of Median Absolute Percentage Error in user-equilibrium flow prediction for each partition size investigated on different artificially-generated networks and the 73 node E_2 network. Lines are used as visual aid for the individual point results.*

5.2.5 Computational Requirements

The computational requirements of the partitioning approaches were investigated using the artificially-generated networks.

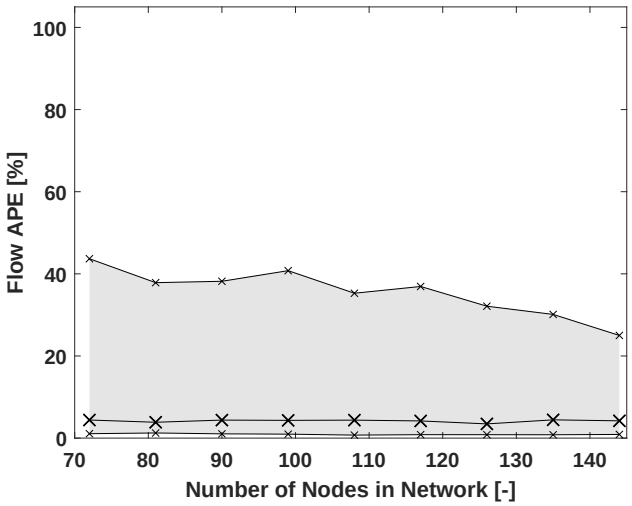
Computational requirements without partitioning

When the GLS method of O-D estimation is applied to a network without partitioning being used, it can be seen in Figure 5.12 that the median error in flow prediction remains constant as the network size grows, but the required computational time and memory increase steeply. For the results in Figure 5.12, the O-D estimation and adjustment algorithms are being applied to the entire network. Due to the steeply increasing computational requirements, there is a limit on the number of nodes that O-D estimation can be applied to at one time.

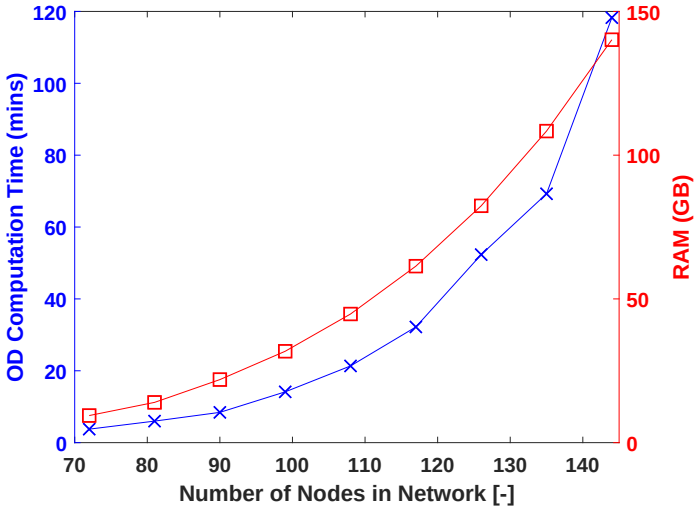
Computational requirements on larger networks with partitioning

The analysis of the artificially-generated networks was expanded to larger sizes for the internal-only and internal-external combined methods which are the best performers of the non-degenerate partitioning approaches. As the networks grow in size it can be seen in Figure 5.13 (a) that the memory requirements for both methods increase at the extreme ranges of partitioning. Comparing between Figures 5.12 and 5.13, the effectiveness of using partitioning to reduce the computational requirements for larger networks can be seen. For example, by using two partitions (internal-only and internal-external) the 243 node network has a similar RAM requirement and computation time to the unpartitioned 135 node network.

At very small partitions the memory requirements increase very steeply. The 216 and 243 node networks were unable to be calculated unpartitioned, this is due to the size of memory required and limitations with the Gurobi solver used. Of most interest is the increase in memory at the largest partition sizes. It can be seen that as the total network size grows, the memory for the larger partitions starts to become very high as each subnetwork within a partition is larger. This has the implication that for larger networks it would be best to choose smaller and more numerous partitions, using the internal-only method as the errors are

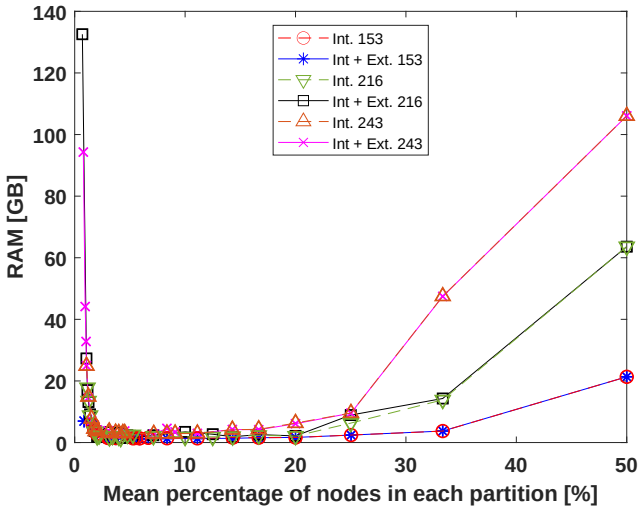


(a) User Equilibrium Flow Prediction

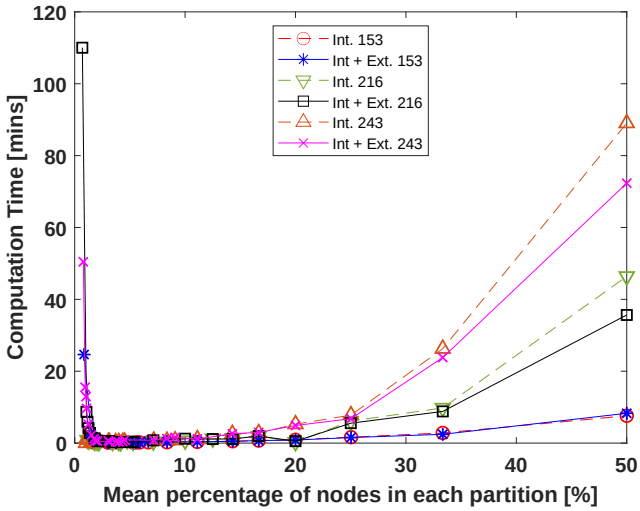


(b) Computational Requirements

Figure 5.12: *(a) Flow prediction error and (b) computational requirements for a range of network sizes when the Origin-Destination (O-D) estimation and adjustment are applied to a range of networks without the use of partitioning. In (a) the solid line is median error and dashed lines indicate the IQR. Lines are used as visual aid for the individual point results.*



(a) Memory



(b) Computation Time

Figure 5.13: *Computational requirements for each partition size investigated for a 153, 216 and 243 node artificially-generated network: (a) Memory; (b) Computation Time. Lines are used as a visual aid for the individual point results.*

smaller than internal-external for intermediate partition sizes. The optimal size and number of partitions depends on the size of the overall network and the number of routes between O-D pairs. Computation time shows a similar trend to memory for the two methods (Figure 5.13 (b)).

5.3 Summary

In this chapter a method of network partitioning through modularity was developed to estimate O-D demand matrices for large road networks to be used in static TA models. It was applied to a benchmark network, the central subnetwork of the English SRN and several artificially-generated networks to allow different levels of partition resolution to be tested for their effects on prior O-D estimation using cross-sectional traffic data.

The TA model error results showed similar patterns across networks for different approaches to combining partitions in the O-D estimation. The degenerate approach showed that partitioning the network into smaller networks representing communities of nodes was tolerable to reduce the size of the network being analysed. Within the non-degenerate approaches, using the external-only approach broadly led to larger errors as the size of partitions increased, due to the prior O-D matrix being based on increasingly simplified versions of the network. In contrast to this, the internal-only approach exhibited a decrease in errors for larger partition sizes. This is due to larger partitions providing more accurate demand estimates of the network within each partition to combine into the prior matrix. Likewise, for very small partition sizes, the internal-only approach was not able to calculate TA results as the demand estimates were too inaccurate. Combining internal and external estimates led to a triangular-shaped profile in errors. Unlike internal-only it was able to calculate results for very small partition sizes and at larger partition sizes it had smaller errors than external-only, more similar to internal-only.

For the case of the English SRN it was found that the lowest errors were achieved with internal-only and internal-external combined using the largest partition size (50%, i.e. two communities). Internal-external

combined had an advantage over internal-only in terms of computation time. The results for the artificially-generated networks revealed that for very large networks, where using large size partitions is still computationally infeasible, it would be better to use the internal-only approach with a number of communities three and greater unless the communities contain such a small proportion of the nodes that the errors are considerably affected.

In the context of the key findings from the literature review (Chapter 2), the novel network partitioning approach to prior O-D estimation developed in this chapter unlocks the ability to use network tomography-based approaches using cross-sectional data alone (e.g. [55, 65, 66, 67]) to create data-driven static TA models. The results show that modularity-based partitioning has the potential to obtain reasonable levels of TA accuracy with lower computation requirements compared to what can be obtained using network tomography without partitioning. This could allow transport planners to obtain the key TA model input of the O-D matrix without the need for expensive travel surveys [63] or privacy-sensitive floating car data [48]. This is especially useful in the case of the non-degenerate approaches. By investigating the alternative ways of applying the partitions, alternative use cases were developed. In the degenerate approach, the partitions were used as the basis for reducing the size of the road network and resulting TAP problem. Such an approach could be integrated into national infrastructure models (e.g. NISMOD in the UK [92]), to improve their traffic modelling accuracy at the scale of large areas.

Chapter 6

Road-specific Density-based Congestion Function Fitting

In TA models, the choice of congestion functions is fundamentally important to the accuracy of the results. In addressing the impact of congestion functions on the TA modelling, this chapter considers solely using cross-sectional data to develop congestion functions for use in a fully data-driven model for strategic planning of interventions on national road networks.

This chapter presents a novel TA formulation to utilize density-based fitted functions that are road-specific, in order to improve the accuracy of model prediction and calculation time over large highway networks. The technique is applied to the E_2 sample subnetwork (Section 4.2) connecting the main metropolitan areas in England, using traffic speed, occupancy and flow count data obtained from the MIDAS system over the period September 2018 to May 2019.

The congestion functions are used in the calculation of the traffic flow pattern and within an O-D matrix congestion adjustment. The effect of using density-based fitting of congestion functions on the computation time and accuracy of recreating the observed flows is compared to the state of the art. The performance informs the suitability for modelling SO flows.

It is shown how the BPR is the best candidate for function form when the individual road congestion functions are validated. Also, it is demonstrated that the incorporation of the proposed density-based BPR fitted congestion function compares favourably to other state-of-the-art methods for calculating UE flows and associated travel times. At the same time, it remains computationally tractable and applicable to model large-scale real-world major road networks.

Specifically, the primary outcomes from this chapter are:

- Identifying the functional form of congestion function that best captures the delay-flow demand relation on a large sample of road segments on the English SRN, comparing between the density-based and flow-based fitting approaches.

- The calculation of TA traffic patterns by including density-based road-specific congestion function fitting which is suited to large scale real-world networks.

- The benchmarking of the method against the current state of the art for use on SRNs, resulting in a favourable comparison, especially in the trade-off between accuracy and computational time.

6.1 Density-based Fitting of Congestion Functions

To estimate the form of travel time multiplier $f(\cdot)$ using observed traffic, a density-based fitting can be used to calculate specific functions for individual edges of the network. The symbols used in this section are as defined in Sections 3.1 and 3.4. The method presented below is an advancement of the approach proposed in [108], suitable to the purpose of this thesis.

While congestion functions model flow as an increasing function of travel time, the nature of congestion on roads means that, past the onset of congestion, flows decrease with increasing travel time. The onset of congestion corresponds to the flow reaching capacity, for an edge $a \in \mathcal{A}$

the density at this point $k_a(m_a)$ is called critical. It separates the hyper-critical ($k > k_a(m_a)$) from the hypo-critical ($k < k_a(m_a)$) regime. In the hyper-critical regime the travel time increases with decreasing flow and the flow-to-capacity ratio does not exceed unit value, as the congestion function would instead indicate (Figure 6.1 (a)). This makes fitting a monotonic congestion function to flow-time data for use in the TAP not possible. When the non-dimensionalised travel time (t_a/t_a^0) is plotted against density, the hyper-critical regime has travel times which increase as density increases so the congestion function can be fitted (Figure 6.1 (b)).

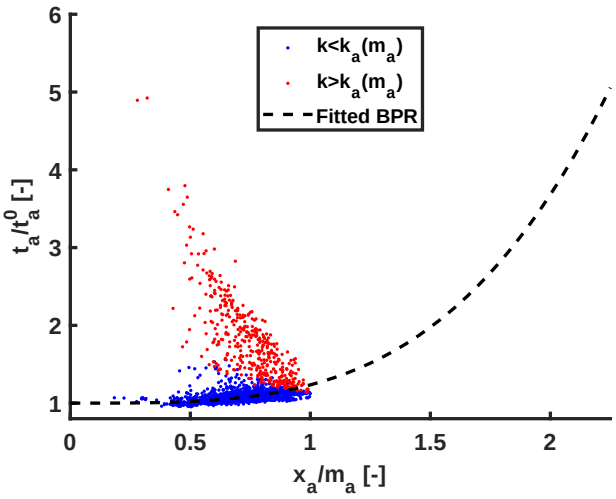
Density-based fitting uses data to estimate the traffic density which can transform the congestion function to a form which a curve can be fitted and specific estimates of parameters obtained. The method assumes that the flow demand $\check{x}_a$ of the congestion function is proportional to density k_a such that the following mapping is assumed:

$$\check{x}_a = m_a \frac{k_a}{k_a(m_a)}. \tag{6.1}$$

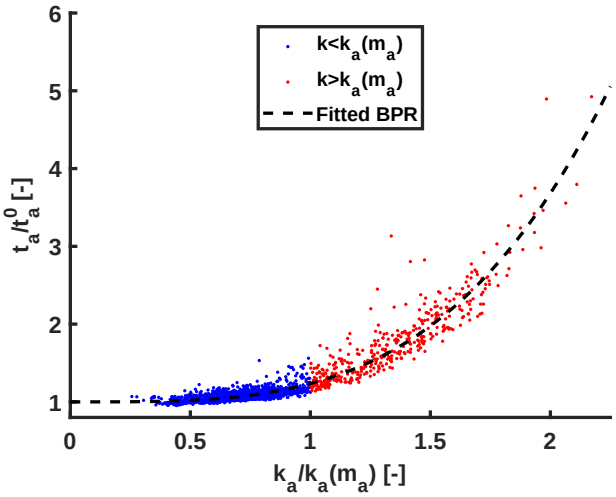
Flow is proportional to density at low flows in the hypo-critical regime of the fundamental diagram (Figure 6.2), however, in the hyper-critical regime it is not. The aim is to use density as a proxy for the number of vehicles wishing to use an edge, $\check{x}_a$, assuming Equation 6.1 holds for the hyper-critical regime. This mapping is applied to different congestion function forms to fit them using density. In the example of BPR, the mapping transforms Equation 3.18 into Equation 6.2. The same α and β values can be used for both equations. The BPR expressed in terms of density is:

$$t_a = t_a^0 \left(1 + \alpha \left(\frac{k_a}{k_a(m_a)}\right)^{\beta}\right), \tag{6.2}$$

where k_a is density of an edge $a \in \mathcal{A}$ and critical density is $k_a(m_a)$. By using the relation of speed to edge length l and travel time ($v = l/t$), the BPR expression reformulated in terms of average speed and density is:



(a) Flow-based



(b) Density-based

Figure 6.1: *Example of hypo-critical (blue) and hyper-critical (red) observations for: (a) Hourly non-dimensional travel time (t_a/t_a^0) against hourly flow/capacity (x_a/m_a); (b) Hourly non-dimensional travel time (t_a/t_a^0) against hourly density/critical density ($k_a/k_a(m_a)$). The Bureau of Public Roads (BPR) function is fitted to the data in (b) and also plotted in (a) for comparison. The measurements are the hourly average traffic of an edge in the E_2 subnetwork on the weekdays selected for analysis between September 2018 and May 2019.*

$$\hat{v}_a = \frac{v_a^0}{\left(1 + \alpha\left(\frac{k_a}{k_a(m_a)}\right)^{\beta}\right)}, \tag{6.3}$$

where, for an edge $a \in \mathcal{A}$, the free-flow speed is v_a^0, and $\hat{v}_a$ is the computed theoretical speed. The values of α and β can be fitted using a non-linear least-squares approach which computes the sum of squared difference between modelled and observed speeds in the set of observations $\mathcal{D}$:

$$\arg\min_{\alpha,\beta} \left(\sum_{d \in \mathcal{D}} |\hat{v}_{a,d} - v_{a,d}|^2 \right). \tag{6.4}$$

The data used in the estimation include all daytime measurements together. Night-time data are excluded as they rarely present congested flow conditions and so would bias the result. The per-minute observations are averaged with 60-min mean values to remove outliers to steady-state conditions.

During fitting, the authors of [108] suggested applying a weighting, as the number of data points recorded for hyper-critical congested flow ($k > k_a(m_a)$) is dwarfed by that of hypo-critical flow ($k < k_a(m_a)$). In this work the values are not weighted, as experiments with the data showed the effect to be limited on the TA results. The same authors also suggest including critical density and free-flow speed as variables to optimise in Equation 6.4; however, the effect on the results of this approach was also limited so is not used. Time-bin specific congestion functions were considered, however, this reduced the number of data points and led to a worse performance than combining all daytime measurements.

6.1.1 Critical Density

The capacity and the critical density estimates in [108] do not coincide with each other. The capacity is taken as the 95^{th} percentile of the measured flows but critical density is taken as the density corresponding to

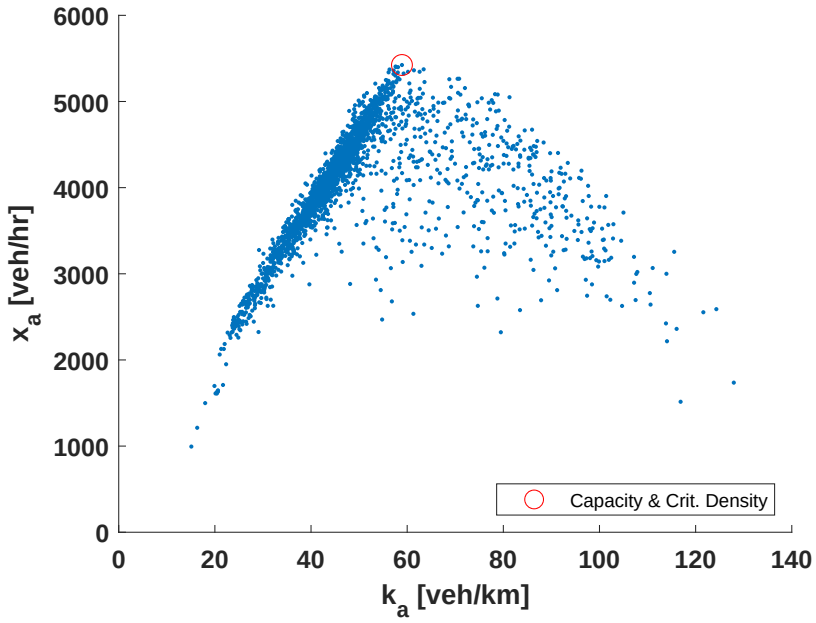


Figure 6.2: *Example Fundamental Diagram for obtaining capacity m_a (which is taken as maximum observed flow rate x_a) and associated critical density $k_a(m_a)$ (denoted by the red circle in the plot) from the hourly flow (x_a) vs. estimated hourly density (k_a). The measurements are the observed hourly average traffic of an edge in the E_2 subnetwork on the weekdays selected for analysis between September 2018 and May 2019.*

the peak of the fundamental diagram (i.e. maximum flow). The ratio of capacity to critical density is integral to the linear mapping between flow and density in Equation 6.1. Due to the noisy nature of loop detector data, automatically fitting the exact critical density is challenging, as seen in attempts at automated fitting of fundamental diagrams [153]. The work in [108] was not aimed at automated fitting.

As described in Section 3.4.1, the capacity is taken as the maximum hourly flow. Then, the critical density $k_a(m_a)$ is obtained as the minimum density, across all the observations, corresponding to that capacity flow. This is obtained from the peak of the rising free-flow branch of the fundamental diagram (Figure 6.2).

6.1.2 Selection of Edges to Apply Road-specific Fitting

The edges that road-specific fitting is applied to are best limited to those with suitable data measurements to improve accuracy. Those considered not suitable are: 1) edges without data in the hyper-critical regime (17 edges in E_2); 2) Edges with the peak of the flows missed by the sensor system (1 edge in E_2); 3) edges where it appears multiple speed limits have been in operation (1 edge in E_2). The edges can have a combination of the problems.

Using the fundamental diagram plots of the mean hourly traffic for each edge, the edges in these categories are selected by inspection. In total, 18 (12%) of the edges of E_2 are unsuitable.

6.2 Inverse-Optimisation Congestion Function Estimation

Acting here as a performance benchmark for the data-driven TA model, the Inverse-Optimization (Inv-Opt) of the TAP estimates a general function $f(\cdot)$ for all edges $a \in \mathcal{A}$ for each analysis time-bin (e.g. AM 6am-10am) [10]. Incorporating support vector regression with a polynomial kernel [160], it aims to find this congestion function such that the resulting calculated UE flows are as close as possible to the observed measurements.

The implemented code is taken from [161], which includes an additional normalisation constraint to match the total cost of the fitted congestion function on all edges to that of a BPR with standard coefficients. The Inv-Opt method involves reformulating the TAP in what is described as the forward problem. This is used to create the Inv-Opt problem.

The Forward Problem

The forward problem uses a formulation of the TAP different to that in Equation 3.10. There is an alternative way of solving the same problem by formulating it as a Variational Inequality (VI), a technique derived from mechanics [162, 163].

The technique aims to find a solution $\mathbf{x}^* \in \mathcal{F}$ to the VI problem, $VI(\mathbf{t}, \mathcal{F})$:

$$\mathbf{t}(\mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*) \geq 0, \quad \forall \mathbf{x} \in \mathcal{F}, \tag{6.5}$$

where $\mathbf{t}(\cdot)$ is assumed strongly monotone on $\mathcal{F}$ and continuously differentiable on $\mathbb{R}_+^{|\mathcal{A}|}$. $\mathcal{F}$ is assumed non-empty and contains an interior point (i.e. Slater's condition for strong duality to hold for convex optimization) [164]. These assumptions mean that there is a UE solution of the network, which is the unique solution to $VI(\mathbf{t}, \mathcal{F})$ [24].

The Inverse Problem

The method aims to use the VI formulation in an inverse problem procedure, named the Inverse Variational Inequality (I-VI). This procedure assumes the measurements of the time-bin flows on the roads are the user-optimal flows and solutions to the TAP for a specific congestion function and O-D demand. They are seen as 'snapshots' of the network at different points in time, with $|\mathcal{J}|$ samples of the edge flow vector $\mathbf{x}$. Solving the I-VI aims to find the congestion function such that each flow observation is as close to an equilibrium as possible. To account for noise in the measured flow data the approach uses the ϵ-approximate solution to $VI(\mathbf{t}, \mathcal{F})$, assuming $\epsilon > 0$ [10]:

$$\mathbf{t}(\hat{\mathbf{x}})'(\mathbf{x} - \hat{\mathbf{x}}) \geq -\epsilon, \quad \forall \mathbf{x} \in \mathcal{F}. \tag{6.6}$$

The I-VI problem aims to find a function $f(\cdot)$ to use as travel time multiplier, so that observed flow vector $\mathbf{x}^{(j)}$ is an ε_j-approximate solution to the VI for each $j \in [\![\mathcal{J}]\!]$; where j is the index of different snapshots of the network with corresponding observed flows, node-edge incidence matrix and O-D pairs. Denoting $\boldsymbol{\epsilon} = (\epsilon_j; j \in [\![\mathcal{J}]\!])$, define the I-VI problem as minimizing $\mathcal{L}_2$ norm of $\boldsymbol{\epsilon}$ over the selection of $\mathbf{t}$ and $\boldsymbol{\epsilon}$:

$$\min_{\mathbf{t}, \boldsymbol{\epsilon}} \ ||\boldsymbol{\epsilon}|| \tag{6.7a}$$

$$s.t. \quad \mathbf{t}(\mathbf{x}^{(j)})'(\mathbf{x} - \mathbf{x}^{(j)}) \geq -\epsilon_j, \quad \forall \mathbf{x} \in \mathcal{F}^{(j)}, j \in [\![\mathcal{J}]\!], \tag{6.7b}$$

$$\epsilon_j > 0, \quad \forall j \in [\![\mathcal{J}]\!]. \tag{6.7c}$$

In order to solve this, the function $f(\cdot)$ must be expressed in a Reproducing Kernel Hilbert Space, $\mathcal{H}$, which allows the following formulation of the I-VI problem [160]:

$$\min_{f,\mathbf{z},\boldsymbol{\epsilon}} \ ||\boldsymbol{\epsilon}|| + \gamma \, ||f||^2_{\mathcal{H}} \tag{6.8a}$$

$$s.t. \quad \mathbf{e}_a^{'}\mathbf{N}_j^{'}\mathbf{z}^{\mathbf{w}} \leq t_a^0 f(\frac{x_a}{m_a}), \quad \forall \mathbf{w} \in \mathcal{W}^{(j)}, a \in \mathcal{A}^{(j)}, j \in [\![\mathcal{J}]\!], \tag{6.8b}$$

$$\sum_{a \in \mathcal{A}^{(\mathrm{j})}} t_a^0 x_a f(\frac{x_a}{m_a}) - \sum_{w \in \mathcal{W}^{(\mathrm{j})}} (\mathbf{d}^{\mathbf{w}})' \mathbf{z}^{\mathbf{w}} \leq \epsilon_{\mathrm{j}}, \forall j \in [\![\mathcal{J}]\!], \tag{6.8c}$$

$$f(\frac{x_a}{m_a}) \leq f(\frac{x_{\hat{a}}}{m_{\hat{a}}}), \ \forall \ a, \ \hat{a} \in \cup_{j=1}^{|\mathcal{J}|} \mathcal{A}^{(j)}, s.t. \ \frac{x_a}{m_a} \leq \frac{x_{\hat{a}}}{m_{\hat{a}}}, \tag{6.8d}$$

$$\boldsymbol{\epsilon} \geq \mathbf{0}, \ \ f \in \mathcal{H}, \ \ f(0) = 1. \tag{6.8e}$$

In this formulation $\boldsymbol{\epsilon} = (\epsilon_j; j \in [\![\mathcal{J}]\!])$ and $\mathbf{z} = (\mathbf{z}^{\mathbf{w}}; \mathbf{w} \in \mathcal{W}^{(j)}, j \in [\![\mathcal{J}]\!])$ are decision vectors. $\mathbf{z}^{\mathbf{w}}$ is a dual variable interpreted as the 'price' of $\mathbf{d}^{\mathbf{w}}$. $\mathbf{e}_a$ is a vector with an entry of 1 for edge a and zeros for all other edges. γ is a regularisation parameter which controls the generalisation properties of $f(\cdot)$ (i.e. the tightness of fit). Note, x_a here is the observed flow on edge $a \in \mathcal{A}^{(j)}$ for $j \in [\![\mathcal{J}]\!]$.

This problem involves an optimisation over functions, solving it requires specifying $\mathcal{H}$ and through that the class of $f(\cdot)$ by selecting a polynomial reproducing kernel [165]. The application of the method in [166] makes this choice as it matches their intuition of how congestion behaves on edges. The polynomial kernel is written for some choice of $c \geq 0$, polynomial order $n \in \mathbb{N}$ and two variables q_1, q_2 in the bounded domain of $\mathcal{H}$ as:

$$\phi(q_1, q_2) = (c + q_1 q_2)^n = \sum_{i=0}^{n} \binom{n}{i} c^{n-i} q_1^i q_2^i. \tag{6.9}$$

This is used to reformulate the previous formulation through [165, Equations (3.2), (3.3), and (3.6)], leading to:

$$\min_{\boldsymbol{\beta},\mathbf{z},\boldsymbol{\epsilon}} \; ||\boldsymbol{\epsilon}|| + \gamma \sum_{i=0}^{n} \frac{\beta_i^2}{\binom{n}{i} c^{n-i}} \tag{6.10a}$$

$$s.t. \quad \mathbf{e}_a' \mathbf{N}_j' \mathbf{z}^{\mathbf{w}} \leq t_a^0 \sum_{i=0}^{n} \beta_i (\frac{x_a}{m_a})^i, \quad \forall \mathbf{w} \in \mathcal{W}^{(j)}, a \in \mathcal{A}^{(j)}, \forall j \in [\![\mathcal{J}]\!], \tag{6.10b}$$

$$\sum_{a \in \mathcal{A}^{(\mathrm{j})}} t_a^0 x_a \sum_{i=0}^{n} \beta_i (\frac{x_a}{m_a})^i \; - \sum_{w \in \mathcal{W}^{(\mathrm{j})}} (\mathbf{d}^{\mathbf{w}})' \mathbf{z}^{\mathbf{w}} \leq \epsilon_{\mathrm{j}}, \quad \forall j \in [\![\mathcal{J}]\!], \tag{6.10c}$$

$$\sum_{i=0}^{n} \beta_i (\frac{x_a}{m_a})^i \leq \sum_{i=0}^{n} \beta_i (\frac{x_{\hat{a}}}{m_{\hat{a}}})^i, \; \forall a, \hat{a} \in \cup_{j=1}^{|\mathcal{J}|} \mathcal{A}^{(j)}, \; s.t. \; \frac{x_a}{m_a} \leq \frac{x_{\hat{a}}}{m_{\hat{a}}}, \tag{6.10d}$$

$$\boldsymbol{\epsilon} \geq \mathbf{0}, \beta_0 = 1. \tag{6.10e}$$

An optimal β^* is obtained by solving this quadratic programming problem that parameterizes the estimator of the travel time multiplier function $\hat{f}(\cdot)$ for an edge $a \in \mathcal{A}$ with:

$$\hat{f}(\frac{x_a}{m_a}) = \sum_{i=0}^{n} \beta_i^* (\frac{x_a}{m_a})^i = 1 + \sum_{i=1}^{n} \beta_i^* (\frac{x_a}{m_a})^i. \tag{6.11}$$

The method involves a three-fold cross-validation of the choice of hyperparameters c, n and γ, where the congestion functions are calculated for a range of hyperparameters on three separate subsets of the data, and then their accuracy in recreating the observed flows is evaluated through calculating the UE TAP. The set of hyperparameters with the smallest error on average across the the three data subsets is selected. The process of k-folds cross-validation aims to reduce overfitting of the function so that it generalises well to data other than that it was fitted to [167]. A larger number of folds would reduce prediction errors, however, three folds are used to reduce total computation times.

The number of edges of E_2 (156) is too many to compute Inv-Opt. The largest previously considered network for Inv-Opt composed of 24 edges. In line with previous work, the function fitted to the simplified E_1 is used for E_2 [10].

6.3 Results

6.3.1 Choice of Function for Density-based Fitting

In this section, the alternative forms for fitting the congestion functions are tested on subnetwork E_2 for the period September 2018 to May 2019. The fitting was applied to BPR, Conical, Akçelik and Exponential (Section 3.4) with the dependent variable trialled with density, all the flow and only the hypo-critical flow regime (hypo-flow). As a non-linear least squares regression is used to fit the parameters, the goodness-of-fit is assessed using the Root Mean Square Error (RMSE) of the predicted speed values on each edge.

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}\left(v_{i,a}^{obs} - \hat{v}_{i,a}\right)^2}{n}}, \tag{6.12}$$

where the $v_{i,a}^{obs}$ is the observed hourly average speed during time-bins AM, MD and PM. $\hat{v}_{i,a}$ is the predicted average speed based on the observed hourly flow. i is the hourly observation and n is the total number of observations.

The range of RMSE values for the fittings on the selected 138 edges of the E_2 sample subnetwork show that across the different approaches to fitting, density-based fitting has the lowest RMSE values (Figure 6.3). Within density-based, the lowest errors are found for the BPR and Exponential function forms, with BPR slightly lower. For flow and hypo-flow fittings, it can be seen that all forms perform similarly.

From these results it can be concluded that BPR is the best choice of congestion function form within the tested functions as it consistently has a low error when applied systematically using density-based fitting. This finding is in line with previous research which has suggested that BPR is the most accurate when fitted with flow for uninterrupted flow facilities such as motorways [98].

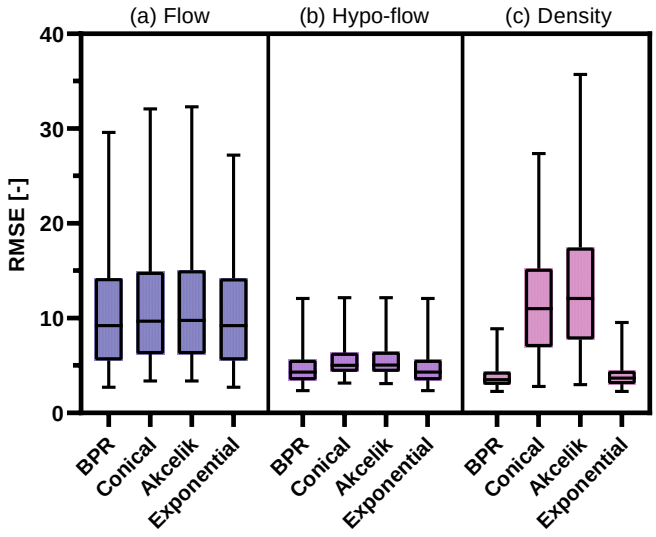


Figure 6.3: *Distribution of Root Mean Square Error (RMSE) values with fitting different function forms to the edges of the E_2 subnetwork with different traffic variables: (a) Flow; (b) Hypocritical flow only; (c) Density. The whiskers of the boxplots represent the range, the box is the IQR and median. The predicted average speeds, based on the observed hourly flows, are compared with the observed hourly speeds using measurements from the weekdays selected for analysis between September 2018 and May 2019.*

BPR and Exponential have a similar shape which explains why they perform similarly. Due to the slightly better result of BPR in Figure 6.3 (c) and the wide adoption of BPR in transport planning software, it is BPR that is used for further analysis of density-based fitting in this chapter.

6.3.2 BPR Parameter Range and Correlation

The range of values that are fitted to the BPR function is of interest as they impact on the convergence of the TAP.

It can be seen that the distribution of α and β are different when fitting with density, flow and hypo-flow (Figure 6.4). The flow-based fitting can be seen to have the largest spread of values. A large amount of the β

values are close to 1 for the flow-based and hypo-flow-based fittings, this could have a negative impact on the convergence of TAP as the functions may not sufficiently encourage the redistribution of vehicles exceeding capacity on those edges. The fitting of β is limited to not going below 1, as this would cause convergence issues in the TAP. The β values for the density-based fitting are all greater than 2, which means they should sufficiently encourage the redistribution of vehicles exceeding capacity. Furthermore, they are in accordance with the values of β mostly used in practice, typically between 2 and 12 [106].

A key difference between the BPR parameters fitted with density and those fitted with flow is that the Pearson rank correlation coefficient shows that the values of α and β are essentially independent for density-based fitting (r=0.009), whereas for flow (r=0.693) and hypo-flow (r=0.727) fitting there is a large amount of positive correlation.

As can be seen in Figure 6.1, for fitting to flow and hypo-flow there are no data points with a saturation rate above 1 (x_a/m_a>1) because flow cannot exceed capacity. This means there is an absence of information for fitting the region of the BPR where $\check{x}_a/m_a > 1$. This region has the most influence on β, the parameter which represents how quickly delays increase in hyper-critical conditions. The value of α has more influence representing delays in the hypo-critical regime $\check{x}_a/m_a \leq 1$. When fitting to flow and hypo-flow, both α and β are fitted to the same data points with saturation rates of less than 1 (x_a/m_a<1), which leads to them exhibiting correlation. However, with its approximation of flow demand in the hyper-critical regime, density-based fitting does not have this problem and its parameters can be fitted with almost no correlation. The density-based fitting's β values can more accurately represent how flow demand increases in congested hyper-critical conditions.

6.3.3 Fitted Capacity and Free-flow Speed

In addition to the choice of congestion function form and parameters, the road capacity and free-flow speed (and travel time) have a strong impact on TA results.

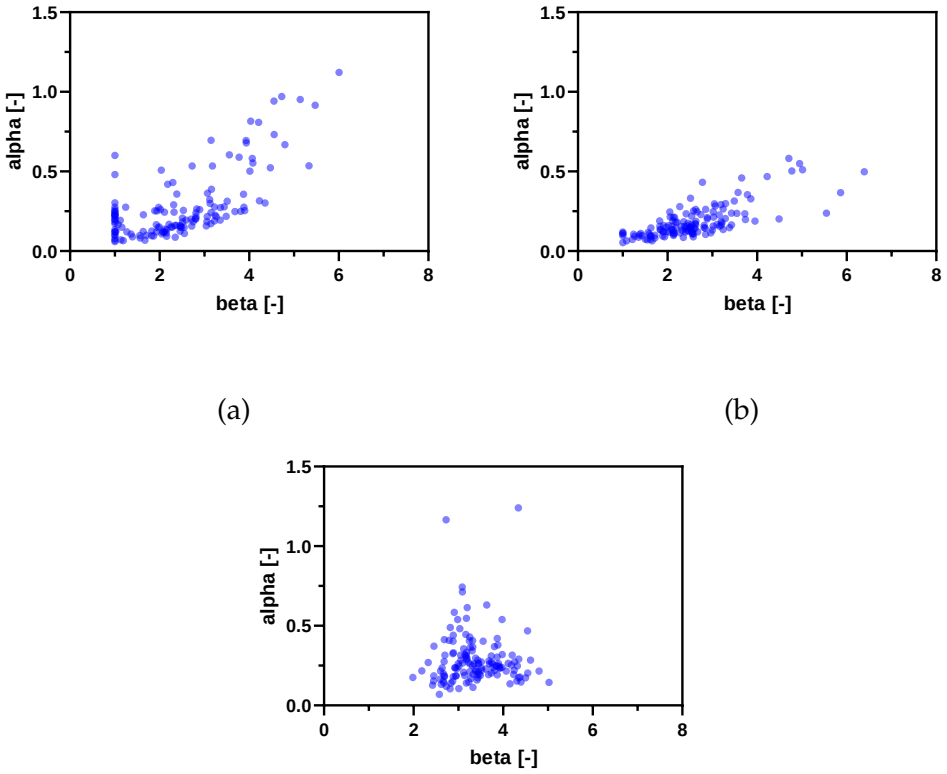


Figure 6.4: *Distribution of Bureau of Public Roads (BPR) formulation coefficient values for fitting with different traffic variables: (a) Flow; (b) Hypo-flow; (c) Density. The data used to obtain the values are the observed hourly average traffic on the E_2 edges for the weekdays selected for analysis between September 2018 and May 2019.*

On the E_2 subnetwork, the distribution of fitted capacities shows a wide range of values, from around 2700 to 8300, with the highest frequency of capacities between 4000 and 8000 (Figure 6.5 (a)). Compared to the capacity values provided by the NTIS model, which tend to group around either 6500 or 8500, there is a wider spread of values. On average the fitted values are 17% lower than the values provided in the NTIS model, which is not as large a difference as the 66% average decrease of fitted values from reference table values in [107].

The free-flow speed results show a spread between 90-120 km/h around the 113 km/h (70mph) speed limit (Figure 6.5 (b)). This result shows that using a free-flow travel time based on posted speed limit is inaccurate, as there is variation between the roads which could be dependent on road-specific features such as road condition, curvature and position in the network. Accuracy in estimating the resulting free-flow travel time is instrumental in the solution of the TAP.

6.3.4 User-equilibrium Assignment Prediction

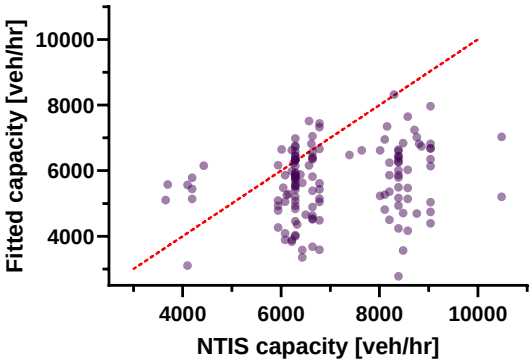
Relative errors in the flow and travel times of the UE assignment prediction are used to evaluate the performance of the density-fitted BPR method of congestion function estimation (BPR-Density). A comparison is made using BPR with standard coefficients ($\alpha = 0.15$, $\beta = 4$) for all edges (BPR-Standard) and the benchmark of Inv-Opt previously used in this type of data-driven static TA model. The comparison is made on subnetwork E_2 using MIDAS measurements from the period September 2018 to May 2019. For BPR-Density, the edges with the problematic data identified in Section 6.1.2 assume the standard coefficients of $\alpha = 0.15$ and $\beta = 4$. All methods take the NTIS values of capacity for these edges.

As in Chapter 5, the Absolute Percentage Errors (APE) are calculated as:

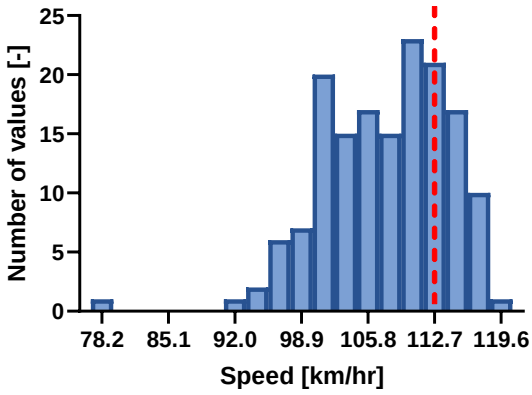
$$APE_a^t = \frac{|t_{p,a}^{user} - t_{p,a}^{obs}|}{t_{p,a}^{obs}}, \tag{6.13}$$

for travel time, while

$$APE_a^x = \frac{|x_{p,a}^{user} - x_{p,a}^{obs}|}{x_{p,a}^{obs}}, \tag{6.14}$$



(a)



(b)

Figure 6.5: *Fitted parameters for: (a) Comparison between fitted capacity against National Traffic Information Service (NTIS) capacity (points are semi-transparent, dashed line represent values where fitted capacity matches the capacity specified by the NTIS); (b) Histogram of free-flow speed (dashed line: posted speed limit). The data used to obtain the values are the observed hourly average traffic on the edges of E_2 for the weekdays selected for analysis between September 2018 and May 2019.*

is used for flows. For each time-bin p and edge a, $x^{obs}_{p,a}$ is the observed flow and $t^{obs}_{p,a}$ is the travel time derived from observed speed. The values are the mean within each time-bin over the fitting period. $t^{user}_{p,a}$ is the predicted travel time derived from the congestion function using $x^{user}_{p,a}$, which is the edge flow value predicted by the model through solving the UE TAP with the calculated O-D matrix. The partitioning methods of Chapter 5 were not used on E_2 as its size allowed an O-D matrix to be obtained unpartitioned.

The inaccuracy of the UE assignment prediction is also assessed through the error in TSTT, $\mathcal{L}_{error}$, such that:

$$\mathcal{L}^{p}_{error}(\mathbf{x}^{ue}) = \frac{\sum_{a\in\mathcal{A}} x^{user}_{p,a} t^{user}_{p,a} - \sum_{a\in\mathcal{A}} x^{obs}_{p,a} t^{obs}_{p,a}}{\sum_{a\in\mathcal{A}} x^{obs}_{p,a} t^{obs}_{p,a}}, \tag{6.15}$$

for time-bin p. This calculation combines the errors in flow and time prediction and has particular relevance for analysis based on aggregate total system cost as features in Chapter 7.

It was found that the UE flows produced after the O-D adjustment was applied to the prior matrix were similar for BPR-Density compared to Inv-Opt and BPR-Standard (Figure 6.6 (a)). Overall the errors for the methods in all time-bins are not statistically different when tested with a one-way ANOVA. The similarity of flow prediction error between methods can be expected, as the aim of the O-D adjustment is to adjust the O-D matrix to make the resulting UE flows match as closely to the observed flows, which they are all effectively able to do. Partly, it could be due to the limited routing choices and the large distances involved on the network; the routing is dominated by free-flow travel times which are the same for each of the methods.

There is a difference in the results when comparing average edge travel time for the UE flow pattern (Figure 6.6 (b)). The results of one-way ANOVA tests between the three methods show a statistically significant difference in performance. There is an improvement from using the edge-specific density-based BPR fittings compared to the network-wide values of Inv-Opt and the standard BPR values.

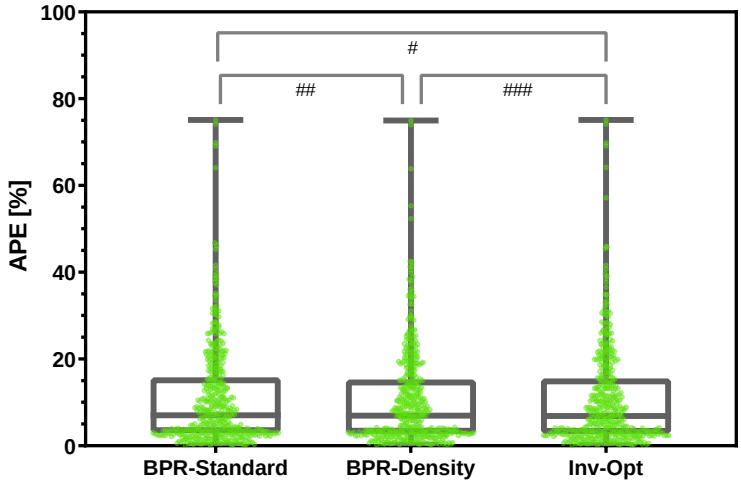
		Mean Flow Error (veh/hr)	Mean Time Error (s)	Mean Flow APE (%)	Mean Time APE (%)	$\mathcal{L}_{error}(\mathbf{x}^{ue})$ (%)
AM	BPR-Standard	-197.8	-24.0	10.9	7.4	-7.8
	BPR-Density	-228.7	-10.2	10.7	5.2	-5.3
	Inv-Opt	-202.3	-27.4	11.0	8.0	-9.0
MD	BPR-Standard	-194.5	-12.5	11.0	4.3	-4.9
	BPR-Density	-191.9	+1.2	10.7	3.9	-1.4
	Inv-Opt	-199.2	-18.0	11.0	5.3	-6.5
PM	BPR-Standard	-186.5	-23.1	10.7	8.0	-7.7
	BPR-Density	-151.8	-8.9	10.4	6.9	-3.3
	Inv-Opt	-192.5	-28.0	10.5	8.8	-9.4

Table 6.1: *Time-bin specific user-equilibrium prediction error statistics for all edges on the E_2 network during the analysis period September 2018 to May 2019. The mean errors refers to the mean across all the edges of E_2.*

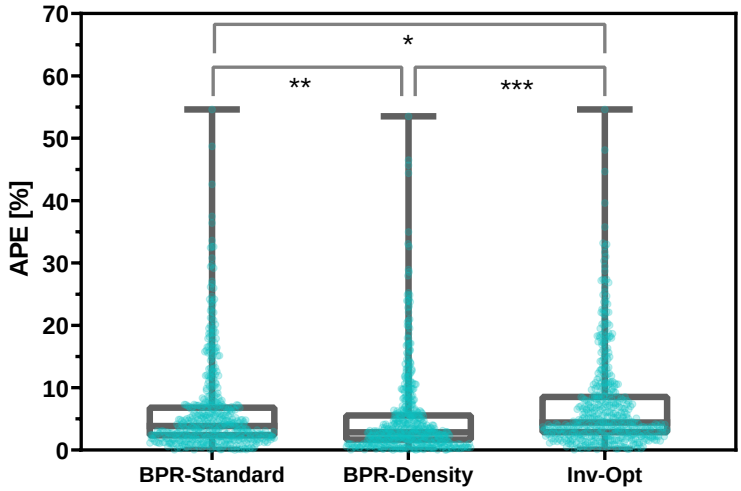
The same trends for flow and travel time accuracy are consistent within the time-bin comparisons (Table 6.1), where it is evident that BPR-Density has the best performance in estimating the travel time. In Table 6.1, the mean error is the mean of the difference between measured and calculated values over all edges. It can be seen that the mean flow errors are all negative, as well as most of the mean time errors, indicating there is a systematic underestimation of TA results regardless of congestion function. BPR-Density has the lowest values of $\mathcal{L}_{error}(\mathbf{x}^{ue})$ in all time-bins, indicating its superior total network cost prediction. This implies BPR-Density would be more accurate for use in CBA and POA calculations. The values of $\mathcal{L}_{error}(\mathbf{x}^{ue})$ are negative in all cases indicating the models systematically underestimate TSTT. The underestimation of the models is likely due to underestimates for the demand leading to less vehicles on the roads.

6.3.5 Function Fitting Computation Time Comparison

BPR density-fitting compares particularly favourably to Inv-Opt when considering the computation time taken to estimate the congestion functions. Inv-Opt took 247,420s for all three time-bins, whereas BPR fitting with density took 22s in total for all the edges. The Inv-Opt result is for computing equations for all three time-bins but does not include the



(a)



(b)

Figure 6.6: *Absolute Percentage Error (APE) in prediction of (a) flow and (b) travel time for user-equilibrium assignment with alternative congestion function estimation methods. Results are for all time-bins and edges on E_2, using data from the weekdays selected for analysis between September 2018 and May 2019. The whiskers of the boxplots represent the range, the boxes are the IQR and median. The points are the individual errors for each edge and time-bin. The p-values of one-way ANOVA tests are, #: p=0.998, ##: p=0.949, ###: p=0.966, *: p=0.237, **: p=0.035, ***: p=1.2e-4.*

time taken for the cross-validation to set the hyperparameters. Hence, the difference in overall time is even greater.

For Inv-Opt, the large increase in computation time as the network size increases, makes the method impractical for large scale national-level direct calculation, although the results suggest a simplified network may be a reasonable substitute regarding flow pattern accuracy.

The fitting of the congestion functions was performed on a Dell PowerEdge C6320 with 2.4GHz Intel Xeon E5-2630 v3 CPU and 24GB RAM.

6.4 Summary

This chapter developed the efficient computation of accurate congestion functions specific to individual roads on a national highway network solely using cross-sectional data. The functions were fitted on a central subnetwork of the English SRN and applied in a data-driven static TA model. The results indicate that this can be an effective approach to the solution of the static TAP for use in traffic analysis.

In the context of the key findings from the literature review of Chapter 2, the results of this chapter fill a gap in the knowledge regarding the best choice for selecting congestion functions that are computationally efficient and accurate. Despite the availability of more advanced models of congestion function form (e.g. Akçelik), for the case of the tested uninterrupted highways of the England SRN, it was shown that fitting the BPR function with a density-based approach provides the best fit, and parameters most suitable for use in a TA model. The results demonstrated the benefits of using a density-based approach rather than the flow-based fitting previously tested in a road-specific context in [107], providing improvements in fitting accuracy and parameter correlation. Furthermore, the results of the chapter showed an improvement in UE travel time prediction for the road-specific density-based BPR fitting compared to the conventional BPR-Standard and the Inv-Opt method previously applied in such a type of TA model in [10]. BPR-Density's improved accuracy compared to Inv-Opt suggests it may prove advantageous to fit a function based on road characteristics rather than the time of day. Lastly, the computational time for fitting the functions of BPR-

Density was considerably lower than Inv-Opt, currently regarded as a state-of-the-art technique, leading to the conclusion it is more suitable for application to large national road networks.

Chapter 7

Strategic National Traffic Analysis

The previous chapters developed methods to extract the key inputs of large-scale, national-level data-driven TA models, congestion functions and O-D matrices. This chapter takes those advances and applies them to a larger area of the English SRN to analyse traffic routing on a national scale. The test network, E_3 (Figure 4.2), represents the motorway network covered by the MIDAS system. The results obtained are not for the whole road system in England, but this strategically important subnetwork. The analysis of E_3 is applied to the year period September 2018 to August 2019. This is a larger window of time than the nine months used in Chapters 5 and 6.

As a static analysis on a subnetwork of the complete transport system, there are limits to what the model's results represent. However, this chapter showcases a number of rapid results for the purpose of high-level strategic analysis of national road infrastructure. The goal of this chapter is to demonstrate how the developed TA model can be used to understand the impacts of rerouting and network changes on congestion, informing strategic choices for traffic authorities.

The analysis in this chapter seeks to:

- Fit a suitable data-driven TA model to a large representation of the English SRN monitored by the MIDAS system.
- Use this model to quantify a national-scale POA for rerouting on the English SRN. Also, to investigate the impact of minimum route costs on rerouting savings through the alternative metric of POA-Delay.
- Use the network representation of the TA model to investigate the distribution of rerouting cost savings around the English SRN.
- Investigate the effect of changing demand levels on routing efficiency.
- Examine, through a sensitivity analysis on the English SRN, the potential improvements to congestion from targeted changes to edge capacity and free-flow travel time. Also, to examine the effect on congestion of specific road closures.

7.1 Routing Efficiency

To analyse the effect of rerouting traffic from UE to SO, the aggregate cost on the network is used. This is calculated as the TSTT, $\mathcal{L}$, as in Section 3.2. For the network, the TSTT is defined by:

$$\mathcal{L}(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} x_a t_a(x_a), \tag{7.1}$$

where the flows can be obtained from UE, x_a^{UE}, or SO, x_a^{SO}, routing patterns (Equation 3.10 and 3.8). The congestion functions t_a are obtained from the density-based BPR fitting (Chapter 6). The TSTT only considers the total time of traversing all edges. Other costs not considered in the TSTT include the vehicle operating and pollution costs. An alternative aggregate measure to TSTT would be the Vehicle Distance Travelled

(VDT) (a.k.a. Vehicle Miles or Kilometres Travelled), which is used in other transport analysis [168]. VDT is obtained through multiplying the flow by the distance of each edge and would be relevant if vehicle operating costs were considered.

With the calculated flow patterns, UE and SO, the POA can be calculated through:

$$POA \stackrel{\text{def}}{=} \frac{\mathcal{L}(\mathbf{x}^{UE})}{\mathcal{L}(\mathbf{x}^{SO})}, \tag{7.2}$$

where $\mathcal{L}(\mathbf{x}^{UE})$ is the UE TSTT and $\mathcal{L}(\mathbf{x}^{SO})$ is the SO TSTT.

An alternative metric to POA is the POA-delay. This removes the effect of long free-flow travel times on the POA. It is useful in the case of highway networks, as generally the minimum cost of travelling long routes can dominate the TSTT and push the POA ratio to unity. It focuses solely on the extra delay component of travel cost, which is the part that can be controlled on the network. The POA-delay is derived and defined as follows based on the work in [130].

Firstly, define the POA in terms of route flows y and route costs c, as defined in Section 3.1:

$$POA \stackrel{\text{def}}{=} \frac{\mathcal{L}(\mathbf{x}^{UE})}{\mathcal{L}(\mathbf{x}^{SO})} = \frac{\sum_i \sum_r y_{ir}^{UE} c_{ir}(y_{ir}^{UE})}{\sum_i \sum_r y_{ir}^{SO} c_{ir}(y_{ir}^{SO})}, \tag{7.3}$$

then simplify the equation to a single O-D pair i on the network with R routes for a demand g for either the UE or SO patterns. Next, routes $r = 1, 2, ..., R$ are ordered by their costs at zero route flow (i.e. free-flow), such that $c_1(0) \leq c_2(0) \leq ... \leq c_R(0)$. The numerator (or equivalently the denominator) of Equation 7.3 is then:

$$\mathcal{L}_i(\mathbf{x}) = \sum_{r=1}^{R} y_r c_r(y_r) = \sum_{r=1}^{R} y_r[c_r(0) + c_r(y_r) - c_r(0)] \tag{7.4a}$$

$$= \sum_{r=1}^{R} y_r c_r(0) + \sum_{r=1}^{R} y_r[c_r(y_r) - c_r(0)]. \tag{7.4b}$$

Define $\gamma_r = c_r(0) - c_1(0)$ for $r = 1, 2, 3, \ldots, R$ to represent the additional free-flow costs of the routes longer than the shortest route $r = 2, 3, \ldots, R$ for this O-D pair. Then $c_1(0) \leq c_1(0) + \gamma_2 \leq ... \leq c_1(0) + \gamma_R$. Substituting γ_r into the first term of $\mathcal{L}_i(\mathbf{x})$ in Equation 7.4b gives:

$$\mathcal{L}_i(\mathbf{x}) = \sum_{r=1}^{R} y_r[c_1(0) + \gamma_r] + \sum_{r=1}^{R} y_r[c_r(y_r) - c_r(0)] \tag{7.5a}$$

$$= \sum_{r=1}^{R} y_r c_1(0) + \sum_{r=2}^{R} y_r \gamma_r + \sum_{r=1}^{R} y_r[c_r(y_r) - c_r(0)] \tag{7.5b}$$

$$= c_1(0)g + \sum_{r=2}^{R} y_r \gamma_r + \sum_{r=1}^{R} y_r[c_r(y_r) - c_r(0)]. \tag{7.5c}$$

Equation 7.5c shows the cost between an O-D pair can be broken down into the sum of three components: the free-flow travel cost of all demand using the shortest route; the additional free-flow travel cost from flow using longer routes than the shortest route; and the delay costs associated with congestion on all routes. All three components feature in the numerator and denominator of the POA calculation; however, the first term is independent of whether routing is UE or SO. When this free-flow shortest route term is large in comparison to the other terms, it pushes the POA towards 1. As an alternative metric to POA, the POA-delay avoids this effect through subtracting the the shortest route free-flow travel costs between all O-D pairs from both UE and SO TSTT. It is is expressed as:

$$POA_{delay} \stackrel{\text{def}}{=} \frac{\mathcal{L}_{delay}(\mathbf{x}^{UE})}{\mathcal{L}_{delay}(\mathbf{x}^{SO})} = \frac{\mathcal{L}(\mathbf{x}^{UE}) - \sum_{i \in \mathcal{W}} c_{i1}(0) g_i}{\mathcal{L}(\mathbf{x}^{SO}) - \sum_{i \in \mathcal{W}} c_{i1}(0) g_i}. \tag{7.6}$$

This expression can be calculated by, first, finding the shortest route free-flow travel time between each O-D pair and multiplying it by its associated O-D demand. Then, the sum of this for all O-D pairs is subtracted from the TSTT (Equation 7.1) for both UE and SO flow patterns to obtain the numerator and denominator, respectively. Although the metric is derived through route flows and costs, it can be calculated using the TSTT derived from edge-based TA, as only the shortest route is needed.

Network Cost Distribution

The travel cost on a single edge $a \in \mathcal{A}$, $\mathcal{L}_{edge}(x_a)$, is defined as:

$$\mathcal{L}_{edge}(x_a) \stackrel{\text{def}}{=} x_a t_a(x_a). \tag{7.7}$$

The difference between the edge cost for UE and SO flow patterns is then defined as:

$$\Delta\mathcal{L}^a_{edge} \stackrel{\text{def}}{=} \mathcal{L}_{edge}(x^{UE}_a) - \mathcal{L}_{edge}(x^{SO}_a). \tag{7.8}$$

The zonal cost difference is used to analyse the cost difference between zones of the network to see how the rerouting benefits are distributed across the network. This is defined as:

$$\Delta\mathcal{L}_z \stackrel{\text{def}}{=} \frac{\mathcal{L}(\mathbf{x}^{UE}_z) - \mathcal{L}(\mathbf{x}^{SO}_z)}{|\mathcal{A}^z|}, \tag{7.9}$$

where for UE and SO flow patterns, $\mathcal{A}^z$ is the set of edges in zone z and $\mathcal{L}(\mathbf{x}_z) = \sum_{a\in\mathcal{A}^z} x_a t_a(x_a)$. The difference in the total costs in the zone from rerouting is divided by the number of edges in the zone $|\mathcal{A}^z|$ to give a mean zonal cost difference per edge.

Marginal External Costs

Understanding the distribution of the marginal costs on edges is relevant for potential road pricing schemes. As covered in Chapter 3, marginal external costs are the costs of the drivers which are not experienced by the driver themselves but contribute to the TSTT. For static

TA models with BPR congestion functions, the values of the marginal external cost are easy to calculate. For an edge $a \in \mathcal{A}$ with a road-specific BPR congestion function, the marginal external cost, $\check{x}_a t'_a(\check{x}_a)$, is calculated by:

$$\check{x}_a t'_a(\check{x}_a) = \frac{t_a^0 \beta_a \alpha_a}{{m_a}^{\beta_a}} \check{x}_a^{\beta_a}, \tag{7.10}$$

where the parameters α_a and β_a are the BPR coefficients fitted to edge a. The other symbols are as described in Section 3.4.

7.1.1 Sensitivity Analysis

Road Parameters

To investigate the sensitivity of the network to individual edge improvements, which could enable the prioritisation of intervention, analysis is conducted on the effect of slight changes to edge free-flow travel time t_a^0 and capacity m_a.

Defining $\mathbf{t}^0 \stackrel{\text{def}}{=} (t_a^0; a \in \mathcal{A})$ and $\mathbf{m} \stackrel{\text{def}}{=} (m_a; a \in \mathcal{A})$, the objective function of the TAP for the UE flow pattern is [10]:

$$V(\mathbf{t}^0, \mathbf{m}) \stackrel{\text{def}}{=} \min_{\mathbf{x} \in \mathcal{F}} \sum_{a \in \mathcal{A}} \int_0^{x_a} t_a^0 f(\frac{s}{m_a}) ds. \tag{7.11}$$

For each edge $a \in \mathcal{A}$, Equation 7.11 can be partially differentiated to obtain:

$$\frac{\partial V(\mathbf{t}^0, \mathbf{m})}{\partial t_a^0} = \min_{\mathbf{x} \in \mathcal{F}} \sum_{a \in \mathcal{A}} \int_0^{x_a} f(\frac{s}{m_a}) ds, \tag{7.12a}$$

$$\frac{\partial V(\mathbf{t}^0, \mathbf{m})}{\partial m_a} = \min_{\mathbf{x} \in \mathcal{F}} \sum_{a \in \mathcal{A}} \int_0^{x_a} t_a^0 f'(\frac{s}{m_a})(-\frac{s}{m_a^2}) ds, \tag{7.12b}$$

where $f'(\cdot)$ is the derivative of $f(\cdot)$. From Equations 7.12a and 7.12b is can be seen that usually $\frac{\partial V(\mathbf{t}^0, \mathbf{m})}{\partial t_a^0} > 0$ and $\frac{\partial V(\mathbf{t}^0, \mathbf{m})}{\partial m_a} < 0$, implying that typically a slight reduction of t_a^0 and increase of m_a reduces the objective

function value V (lowering congestion). From this, the slight change in free-flow travel time on an edge $a' \in \mathcal{A}$ is taken as $\Delta t^0_{a'} \stackrel{\text{def}}{=} -0.2 * min(t^0_a; a \in \mathcal{A})$ and the slight change in capacity on an edge $a' \in \mathcal{A}$ is taken as $\Delta m_{a'} \stackrel{\text{def}}{=} 0.2 * min(m_a; a \in \mathcal{A})$. These have been shown to be suitable changes in previous sensitivity analysis [10].

To analyse the effect that the changes have on the total system, the flow vectors which are obtained in the UE flow pattern calculations on the changed networks are used as input in the calculation of TSTT differences. The following flow vectors are required:

$$\mathbf{x}_{base} = \arg\min_{\mathbf{x}\in\mathcal{F}} \sum_{a\in\mathcal{A}} \int_0^{x_a} t^0_a f(\frac{s}{m_a}) ds, \tag{7.13}$$

$$\mathbf{x}_{tt,a'} = \arg\min_{\mathbf{x}\in\mathcal{F}} \Big[\sum_{a\in\mathcal{A}, a\neq a'} \int_0^{x_a} t^0_a f(\frac{s}{m_a}) ds + \int_0^{x_{a'}} (t^0_{a'} + \Delta t^0_{a'}) f(\frac{s}{m_{a'}}) ds \Big], \tag{7.14}$$

$$\mathbf{x}_{m,a'} = \arg\min_{\mathbf{x}\in\mathcal{F}} \Big[\sum_{a\in\mathcal{A}, a\neq a'} \int_0^{x_a} t^0_a f(\frac{s}{m_a}) ds + \int_0^{x_{a'}} t^0_{a'} f(\frac{s}{m_{a'} + \Delta m_{a'}}) ds \Big], \tag{7.15}$$

where $\mathbf{x}_{base}$ is the flow on the unchanged network, $\mathbf{x}_{tt,a'}$ is the flow on the network with a change to the travel time on edge $a' \in \mathcal{A}$, and $\mathbf{x}_{m,a'}$ is the flow on the network with a change to the capacity on edge $a' \in \mathcal{A}$. Then, for each edge $a' \in \mathcal{A}$ the change in TSTT is:

$$\Delta\mathcal{L}_{tt,a'} = \mathcal{L}(\mathbf{x}_{base}) - \mathcal{L}(\mathbf{x}_{tt,a'}) \tag{7.16}$$

$$\Delta\mathcal{L}_{m,a'} = \mathcal{L}(\mathbf{x}_{base}) - \mathcal{L}(\mathbf{x}_{m,a'}) \tag{7.17}$$

For $a \in \mathcal{A}$ approximately $\Delta\mathcal{L}_{tt,a'} \propto \partial V(\mathbf{t}^0, \mathbf{m})/\partial t^0_a > 0$ and $\Delta\mathcal{L}_{m,a'} \propto |\partial V(\mathbf{t}^0, \mathbf{m})/\partial m_a| > 0$.

Road Closures

In addition to investigating the sensitivity of the network to edge-specific improvements, analysis is done on the effect of removing individual edges from the network to simulate road closures. The flow on the network with a removed edge $e \in \mathcal{A}$ is calculated by:

$$\mathbf{x}_{edge,e} = \arg\min_{\mathbf{x}\in\mathcal{F}^e} \sum_{a\in\mathcal{A}^e} \int_0^{x_a} t_a^0 f(\frac{s}{m_a})ds, \tag{7.18}$$

where for each edge $e \in \mathcal{A}$, the network is changed by its removal to create a subset of edges $\mathcal{A}^e \subset \mathcal{A}$ and feasible flow vectors $\mathcal{F}^e \subset \mathcal{F}$. Using this and $\mathbf{x}_{base}$ defined in Equation 7.13, the change in TSTT is then calculated by:

$$\Delta\mathcal{L}_{edge,e} = \mathcal{L}(\mathbf{x}_{base}) - \mathcal{L}(\mathbf{x}_{edge,e}) \tag{7.19}$$

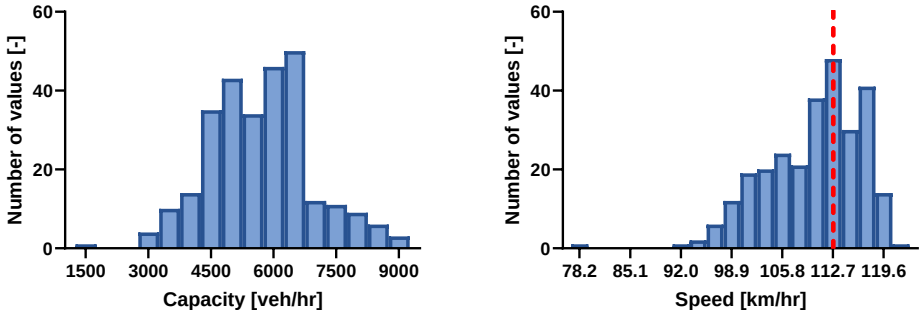
7.2 National Model Fitting

Using the methods described in previous chapters, a data-driven TA model is fitted to the topographic E_3 network (Figure 4.2 (c)). The MIDAS data is taken from the 12 month period September 2018 - August 2019 for the time-bins AM, MD and PM.

The density-based BPR congestion function fitting method outlined in Chapter 6 is applied to the E_3 network. It is applied to 231 suitable edges of the network (83% of total), with these edges selected as in Section 6.1.2. The results have a similar profile to the fitting results of E_2 (Figure 7.1). As in Section 6.3.4, the edges without suitable data use standard coefficients of $\alpha = 0.15$ and $\beta = 4$, and take the NTIS values of capacity.

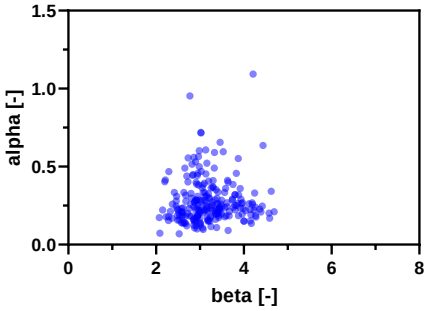
To obtain the O-D matrix, the internal-only partitioning method of Chapter 5 is used with two partitions (Figure 7.2). Without partitioning, suitable O-D matrices cannot be calculated due to numerical difficulties relating to the size of E_3.

As in Chapter 6, the fitted congestion functions and O-D matrices are used to obtain the errors in the resulting UE traffic pattern compared to the observed edge flows and travel times (Figure 7.3 and Table 7.1). The validation shows a larger error in flow prediction than E_2, however, the travel time error is relatively low and the overall error in TSTT, $\mathcal{L}_{error}(\mathbf{x}^{ue})$, is only 7.9% on average across the time-bins.



(a)

(b)



(c)

Figure 7.1: *Fitted parameters on the edges of the E_3 topographic network: (a) Capacity; (b) Free-flow speed (red-dash line is speed limit); (c) Bureau of Public Roads (BPR) coefficients. The data used to obtain the values are the observed hourly average traffic on the edges of E_3 for the weekdays selected for analysis between September 2018 and August 2019.*

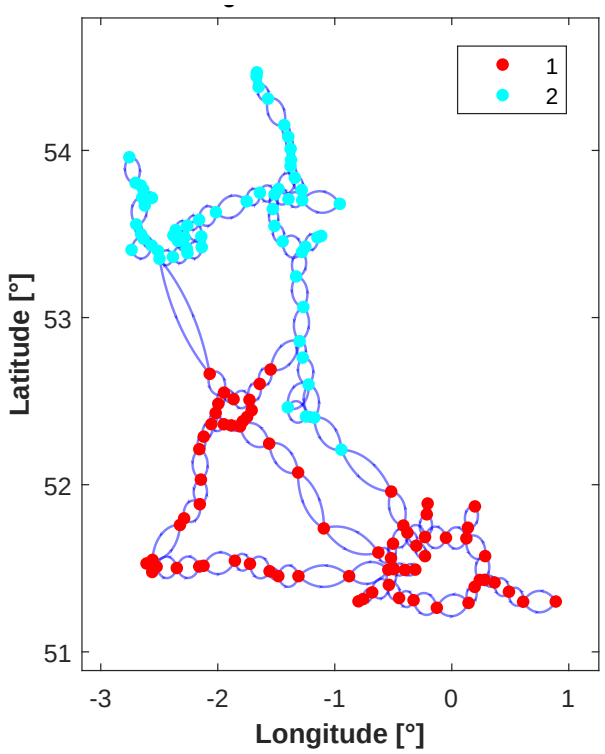


Figure 7.2: *Topographic network E_3 partitioned into two subnetworks.*

	Mean Flow Error (veh/hr)	Mean Time Error (s)	Mean Flow APE (%)	Mean Time APE (%)	$\mathcal{L}_{error}(\mathbf{x}^{ue})$ (%)
AM	-163	-0.5	32.8	8.8	-7.8
MD	-48	+4.5	30.9	8.1	-4.6
PM	-219	-1.2	32.6	9.5	-11.2

Table 7.1: *Time-bin specific User-Equilibrium (UE) prediction error statistics for all edges on the E_3 network during the analysis period September 2018 to August 2019. The mean errors refers to the mean across all the edges of E_3.*

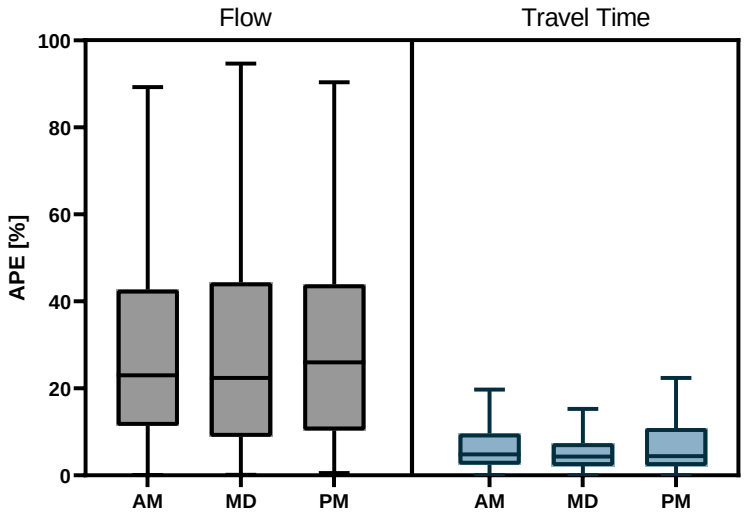


Figure 7.3: *Absolute Percentage Error (APE) in prediction of (a) flow and (b) travel time for User-Equilibrium (UE) assignment in each time-bin on the E_3 topographic network. Results are for the edges on E_3, using data from the weekdays selected for analysis between September 2018 and August 2019. The whiskers of the boxplots are calculated using the Tukey method [169], the boxes are the IQR and median.*

7.3 Results

7.3.1 Routing Efficiency

The routing efficiency of the E_3 network is assessed for the daytime time-bins AM (6 am - 10 am), MD (10 am - 4 am) and PM (4 pm - 10 pm). Additionally, it is assessed for the nighttime time-bins NT-A (12 am - 6 am), and NT-B (8 pm - 12 am) to investigate the efficiency when demand is lower. After calculating the total costs for the UE and SO flow patterns on E_3, the POA is found to be very low (Table 7.2). On average, the POA for the three day time-bins (AM, MD, PM) is 1.0013. The POA values for the night time-bins are even lower. This indicates that the E_3 network does not have a large potential for savings in overall congestion from rerouting selfish drivers in a more system optimised way. This may be due to the demand profile, congestion functions or network structure. It can be seen that the TSTT values, $\mathcal{L}(\mathbf{x})$, are much lower in the night time-bins than the day time-bins. This is due to the greatly reduced vehicle numbers in those periods (in particular for the early hours of NT-A).

Part of the reason for the low POA values is the impact of large free-flow travel times, which can distort the calculated POA and lead to low values. The POA-delay is calculated to remove the effect of the free-flow travel time of the shortest route between O-D pairs. The values for POA-delay show a mean value of 1.0136 and a maximum of 1.0151 for the day time-bins, indicating that there is up to a 1.5% decrease in delay costs from rerouting during the day (Table 7.3). This is an order of magnitude larger than the POA, however, it is still only a small change. It can be seen that the POA-delay is also similarly higher for the night time-bins, however, this could be misleading. It can also be seen that the $\mathcal{L}_{delay}(\mathbf{x})$ values are very low. This indicates that most of the cost in the night time-bins is associated with the minimum route costs and there is little absolute benefit from rerouting at these times. As the potential for rerouting is much less for the night time-bins, the rest of this chapter focuses on the analysis of the three day time-bins.

Time Bin	$\mathcal{L}(\mathbf{x}^{ue})$ [hr]	$\mathcal{L}(\mathbf{x}^{SO})$ [hr]	POA [-]
NT-A	13810	13809	1.0001
AM	87797	87662	1.0015
MD	87835	87709	1.0014
PM	82921	82831	1.0011
NT-B	28963	28959	1.0002

Table 7.2: *Price of Anarchy (POA) for E_3 during the twelve months of September 2018 to August 2019. The weekday periods covered by the time-bins are NT-A (12 am - 6 am), AM (6 am - 10 am), MD (10 am - 4 am), PM (4 pm - 10 pm), NT-B (8 pm - 12 am).*

Time Bin	$\mathcal{L}_{delay}(\mathbf{x}^{ue})$ [hr]	$\mathcal{L}_{delay}(\mathbf{x}^{so})$ [hr]	POA-delay [-]
NT-A	15.4	15.1	1.0199
AM	9075.2	8940.1	1.0151
MD	8784.9	8658.4	1.0146
PM	8210.8	8120.7	1.0111
NT-B	191.7	187.2	1.0240

Table 7.3: *Price of Anarchy-delay (POA-delay) for E_3 during the twelve months of September 2018 to August 2019. The weekday periods covered by the time-bins are NT-A (12 am - 6 am), AM (6 am - 10 am), MD (10 am - 4 am), PM (4 pm - 10 pm), NT-B (8 pm - 12 am).*

Distribution of routing costs

To investigate how the traffic cost is redistributed between UE and SO flow patterns, the difference in the cost on each edge, $\Delta\mathcal{L}^a_{edge}$ (Equation 7.8), is plotted in Figure 7.4. Other quantities such as the saturation rate, x_a/m_a, and delay factor, $t_a(x_a)/t^0_a$, could be used to show the effect of the different routing; however, the difference in edge cost is the most relevant as it captures both flow and travel time together.

The plots in Figure 7.4 show that there are many edges which do not have routing differences, for example between edges 220 to 250. These

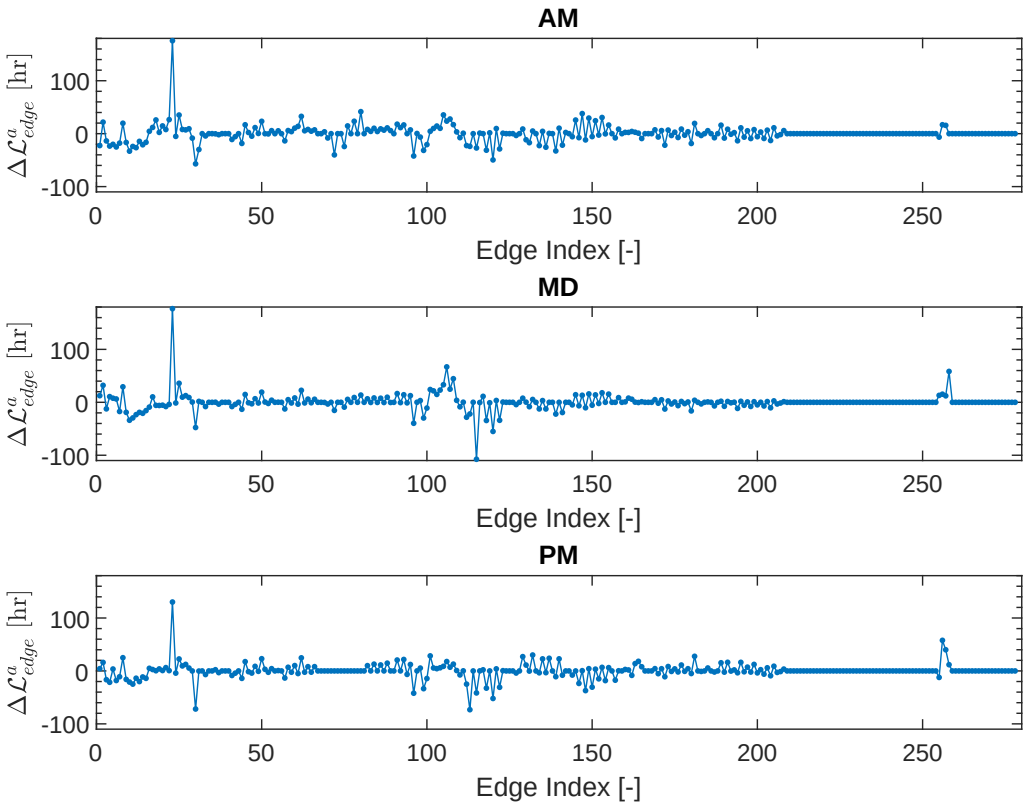
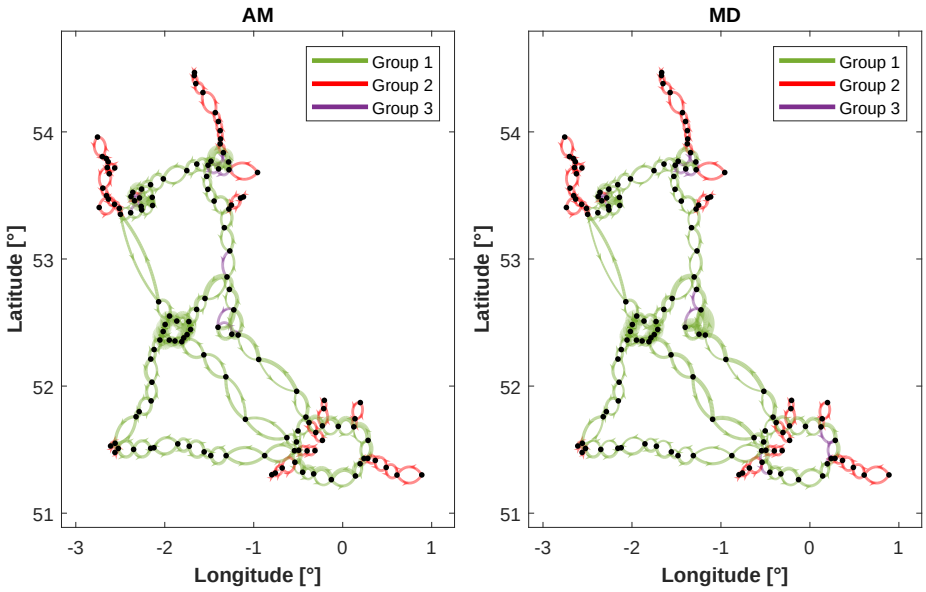
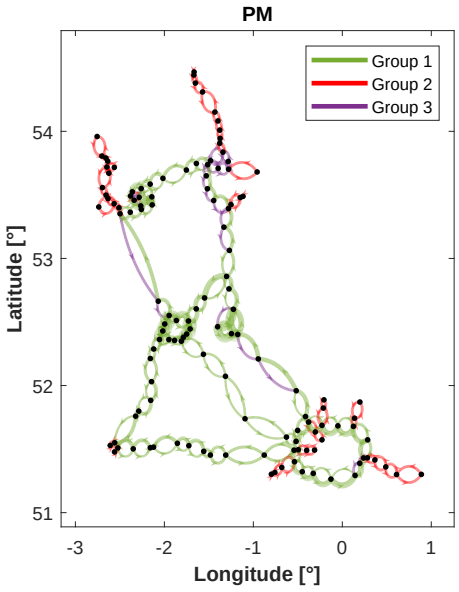


Figure 7.4: *Difference in the total cost on each edge of the network between User-Equilibrium (UE) and System-Optimal (SO) flow patterns in each time-bin. Results are for edges on E_3, using data from the weekdays selected for analysis between September 2018 and August 2019.*



(a)

(b)



(c)

Figure 7.5: *Edges on E_3 that have cost differences between User-Equilibrium (UE) and System-Optimal (SO) flow patterns during the analysis period September 2018 to August 2019: (a) AM; (b) MD; (c) PM. Edge thickness is proportional to change in edge cost.*

edges often correspond to the areas of E_3 that only have single routes. These are called bridge edges and removing them would prevent certain nodes from being accessible from the rest of the network.

The edges without cost differences are highlighted in red and purple in Figure 7.5. The red Group 2 edges are the bridge edges which only have one route available to their attached nodes, they do not have a difference in any time-bin. The edges in the purple Group 3 change across time-bins. This shows that the edges in the main subnetwork that are not 'bridges' yet present no cost difference, change depending on the demand profile of the time-bin considered.

The marginal external cost for each edge and time-bin is shown in Figure 7.6. As explained in Section 3.2, this is the amount of additional cost in the form of time that needs to be considered on each edge to achieve the SO flow pattern using the Frank-Wolfe algorithm. By adding this cost to the edge travel time, it makes the edges which have higher marginal external costs and contributions to TSTT less attractive to drivers, pushing them to choose routing patterns with optimal system total cost. In practice, this could be achieved through monetary road pricing using a value of time conversion. Notably, some of the highest marginal external costs are on bridge edges. High road prices on them would entail implementation difficulties as they only have one route to their associated nodes. Drivers wishing to access these nodes would have no alternative choice of route, meaning the charges would not directly affect the routing. Instead, these road prices would likely have unconsidered secondary effects on mode choice and patterns of demand, as drivers may not want to use the available road system for affected journeys. This highlights a problem for any road network with bridge edges which reduce routing opportunities. However, for the E_3 network, it is a simplification of the full road network and alternative routes would be available in reality.

7.3.2 Effect of Rerouting in Zones

Although the POA on the E_3 network is small, the change in costs from rerouting UE flow patterns to SO is distributed differently across the

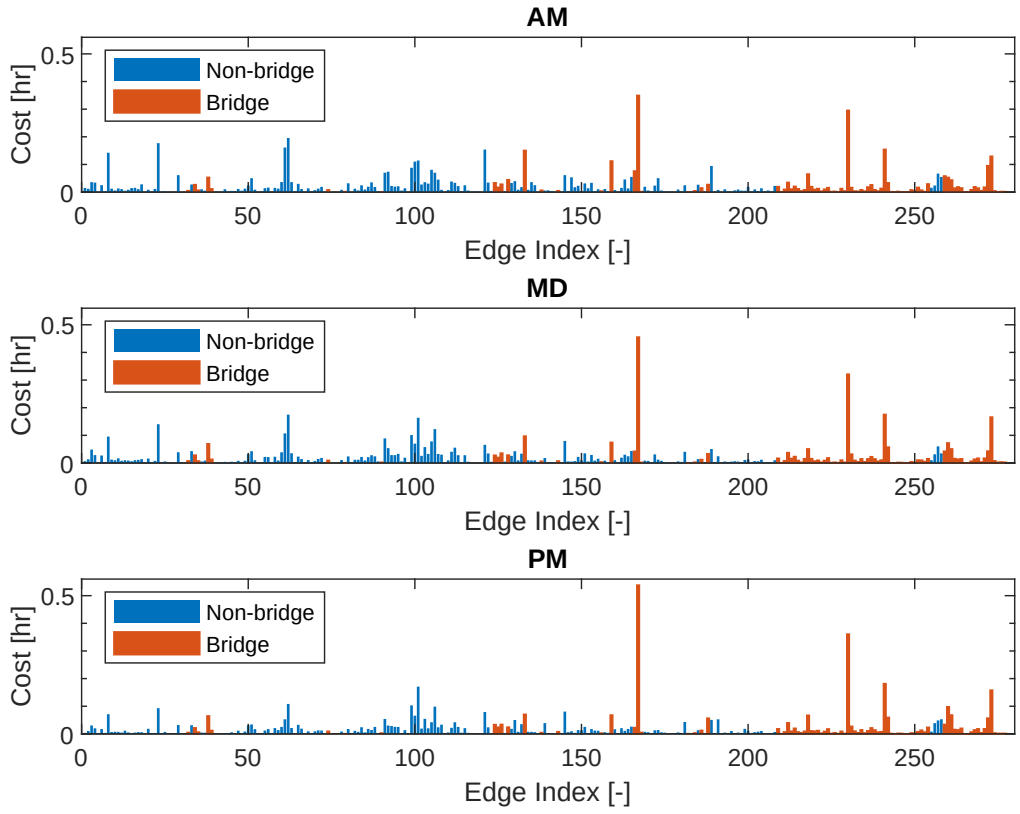


Figure 7.6: *Marginal external cost for each edge and each time-bin on the E_3 network for the analysis period September 2018 to August 2019. Bridge edges on the network (those with single routes) are indicated with red bars, non-bridge edges are indicated with blue bars.*

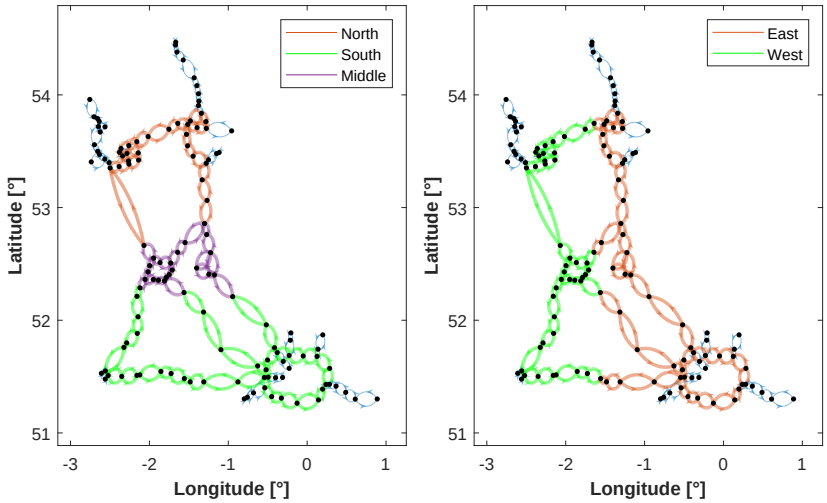
network. For the network overall, the total costs decrease when changing the routes taken by drivers, however, the benefits of this are not evenly distributed. For example, in Figure 7.4 it can be seen that the edges 1-28 have some of the highest redistribution of costs. These edges correspond to the roads around the Birmingham area.

To investigate the pattern of redistribution of costs from rerouting, in Figure 7.7 a sample of zones are identified on E_3. These split the network into broad geographic regions (e.g. North, South) and also city regions (e.g. London). The zones discount the bridge edges identified (Figure 7.5) to only leave the 'core' subnetwork of E_3. The parts of the network that contain the bridge edges decrease the opportunity to lower the TSTT through rerouting, as they contribute to the total demand but cannot vary routing between UE and SO.

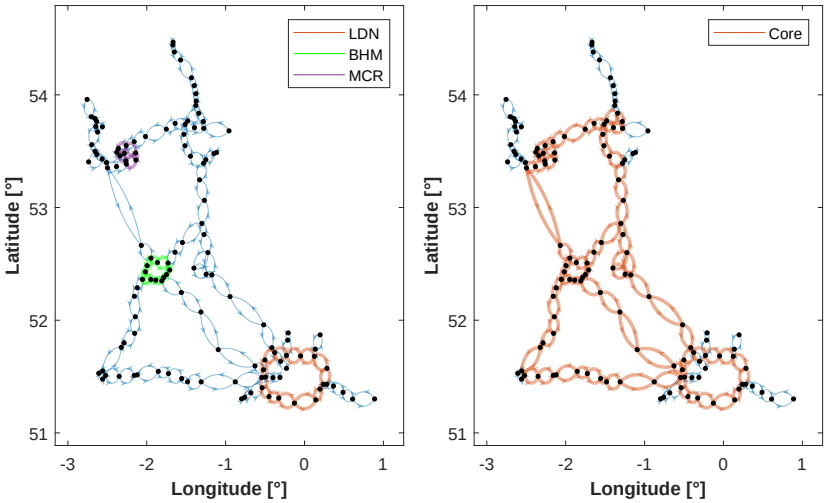
In Figure 7.8 (a), it can be seen that there are variations in the changes to the zonal costs from rerouting between different zones and time-bins. For example, between North, South and Middle it is apparent that North benefits the most from rerouting, in particular in the AM time-bin. This comes at the expense of the edges in the Middle zone, however, in the MD time-bin that zone does have a slight positive reduction in cost. In Figure 7.8 (b), it can be seen that, of the three zones, the highest mean edge marginal external cost is within the Middle region, highlighting that if a marginal cost road pricing scheme was implemented then higher charges would be paid in the region with the least reduction in time cost.

When the subnetwork is divided between the East and West zones, it can be seen that in AM and PM the West zone benefits more than the East. However, in the MD time-bin East benefits slightly more than West. In the city zones it is clear that Birmingham (BHM) has the largest reduction in cost of all the three, however, London (LDN) has higher mean edge marginal external costs.

In the Core zone, with the bridge edges removed from E_3, there is a positive reduction in cost across all time-bins. Using the Core zonal cost totals in the POA calculation leads to higher values than those obtained for E_3 (AM 1.0021 ; MD 1.0020 ; PM 1.0015).

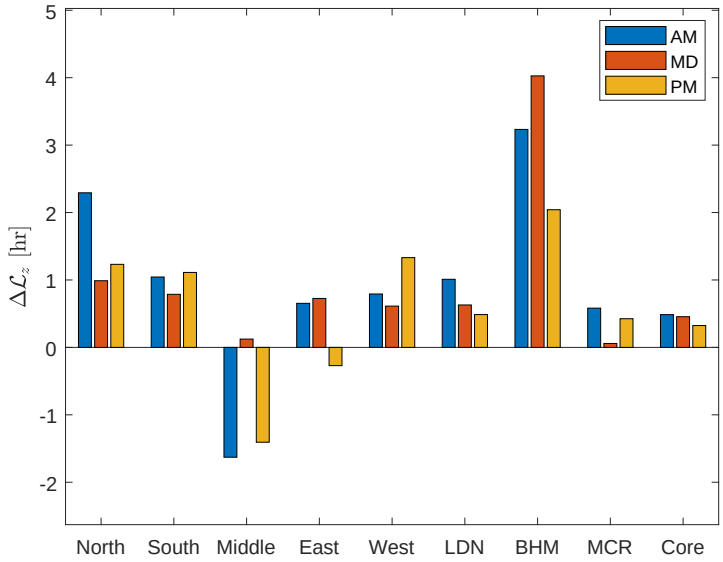


(a) North, South, Middle (b) East, West

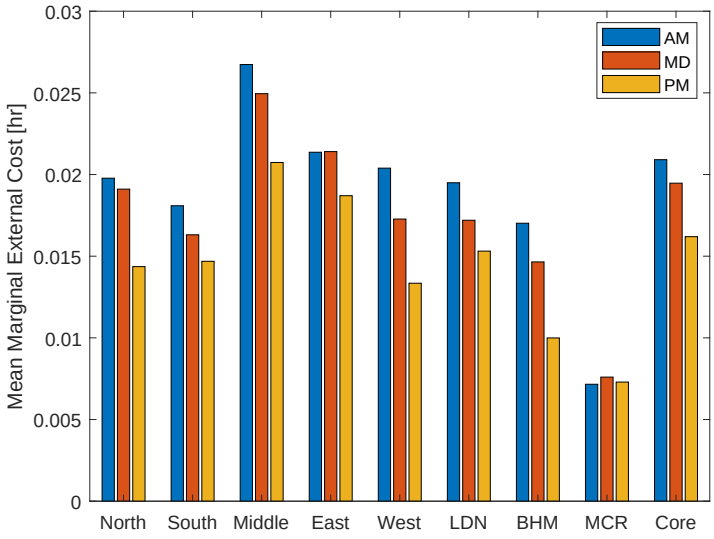


(c) Cities (d) Core (without bridge edges)

Figure 7.7: *Zonal divisions of the E_3 topographic representation. (a) North, South, Middle; (b) East and West; (c) Cities- London (LDN), Birmingham (BHM), Manchester (MCR); (d) Core subnetwork without bridge edges. The edges with thin blue lines are excluded from the zones.*



(a) Mean zonal cost difference



(b) Mean marginal external cost

Figure 7.8: *Zonal distribution on E_3 of cost differences between User-Equilibrium (UE) and System-Optimal (SO) flow patterns in each time-bin during the analysis period September 2018 to August 2019: (a) Mean zonal cost difference per edge; (b) Mean marginal external cost. The mean is across all the edges of the zone.*

The zone with the clearest benefit from SO rerouting consists of the 24 edges which represent roads surrounding Birmingham (BHM). When the total costs of this zone are included in the POA analysis, the values are considerably higher than for E_3 (AM 1.0203 ; MD 1.0262 ; PM 1.0154). BHM is clearly a beneficiary of rerouting on the network, however, the Middle zone of which it is a part of, is a zone which loses overall from rerouting. This is due to the position of the zones, the structure of the network and the profiles of demand. Depending on how the network is divided into zones, different aggregate results are obtainable. Overall, the network benefits from rerouting, however, users in distinct areas of the country with different needs from the transport system would perceive the changes differently. Further understanding of the causes that lead to different zones benefiting or losing out due to rerouting is beyond this analysis; however, future work could build on this insight.

7.3.3 Routing Efficiency with Changing Demand

Changes to the demand profile on the E_3 network causes changes to the routing of the traffic, which is clear from the previous analysis.

To investigate how the volume of demand affects the routing efficiency, the O-D matrix for each time-bin was scaled by a factor ranging from 0 to 5 in steps of 0.1 [126]. This demand multiplier was applied to all the O-D pairs equally. It is a linear change in total demand on the network, however, it causes an exponential increase in the TSTT (Figure 7.9 (a)).

In Figure 7.9 (a), it appears that costs for UE and SO flow patterns are very similar as the demand is varied. However, when the cost differences are plotted in isolation (Figure 7.9 (b)), it can be seen that the difference is increasing with the multiplier, implying the efficiency of the network degrades.

The calculated POA for each level of demand increases up to around 2 to 2.5 times the original demand before decreasing and levelling off at higher demand levels (Figure 7.10 (a)). This decrease at higher demands is due to the increase in the total network cost in Figure 7.9 (a) outstripping the increasing difference between the UE and SO total network costs (Figure 7.9 (b)). The peak POA of all the time-bins is 1.0055 for

AM, approximately five times the value for the original demand profile, although still low.

The POA-delay for the network, with the effect of minimum travel costs removed, has a different profile than POA (Figure 7.10 (b)) and the scale of results is an order of magnitude higher. The largest values are for demands less than the original O-D matrix, suggesting that there are more opportunities for lowering delays through rerouting when the number of drivers is lower. The difference in profile compared to POA highlights the impact that discounting minimum route costs has. At lower demands there are higher delay cost savings to be made by rerouting however, these are dwarfed by the high minimum route costs keeping the POA low. At higher demands, the POA-delay levels off at a lower value which is similar to the POA values for that demand level. This is again due to the difference in total network cost between routing patterns being outstripped by the increase in total network cost for both routing patterns (the same as in Figure 7.9).

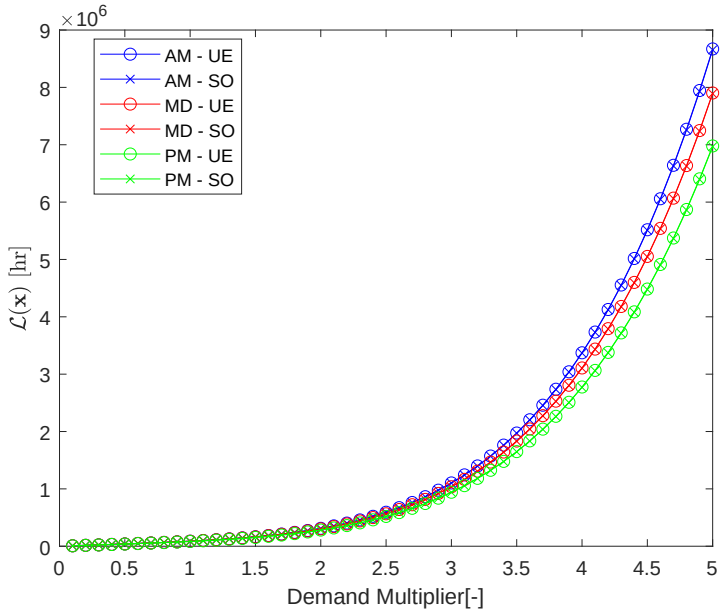
7.4 Sensitivity Analysis

7.4.1 Road Parameters

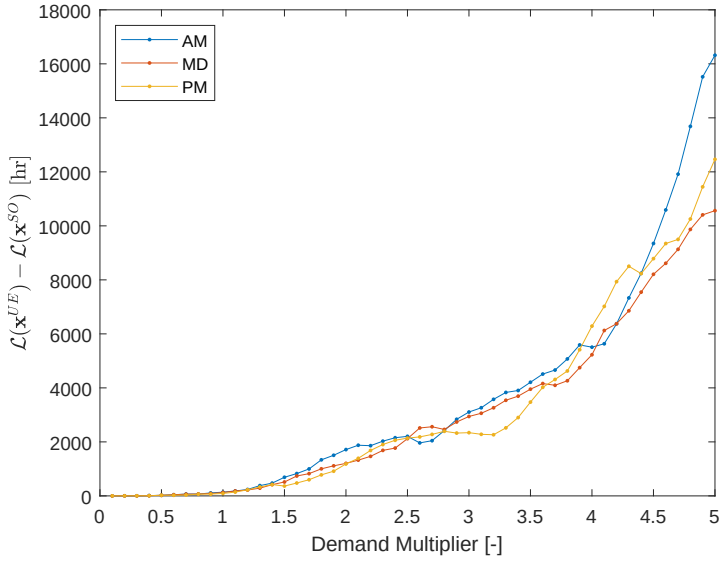
As rerouting the traffic appears not to have a large potential to reduce congestion overall, an alternative is to investigate the sensitivity of the TSTT under UE conditions to slight changes in the edge capacity and free-flow travel time.

The results for edge capacity changes, averaged over the time-bins, show that such changes result in a positive improvement in costs for the majority of edges (Figure 7.11 (a)). For a few edges the changes are slightly negative. There are several edges where there are significant improvements in TSTT, which indicates these edges would be best targeted for capacity-based interventions to improve performance.

The results for free-flow travel time changes, averaged over the time-bins, show that such changes almost always results in a positive improvement in congestion (Figure 7.11 (b)). There are less edges where there are significant improvements in TSTT compared to capacity, how-

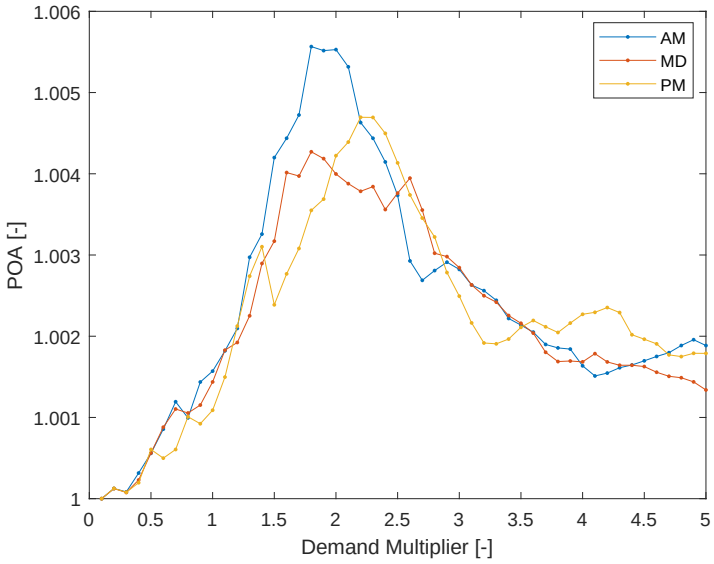


(a) Total System Travel Time

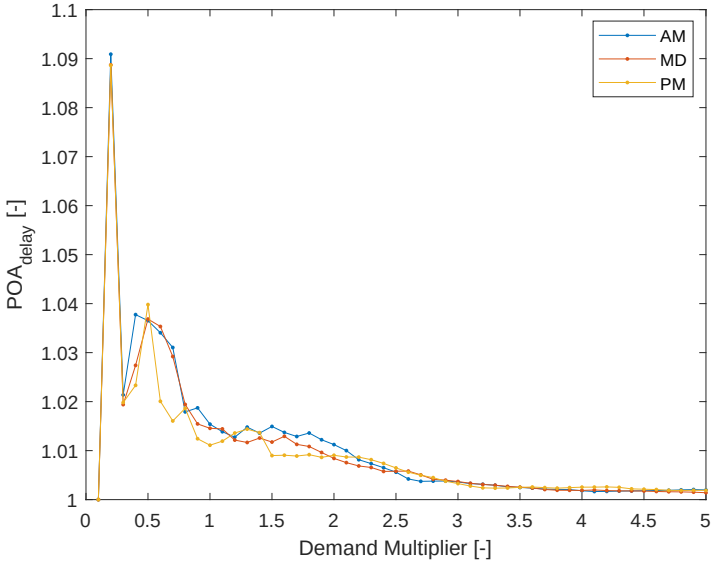


(b) Total System Travel Time difference

Figure 7.9: *Effect of varying demand on Total System Travel Time (TSTT) for E_3: (a) TSTT on E_3 as the demand is varied from 0 to 5 times the original demand matrix; (b) TSTT difference between User-Equilibrium (UE) and System-Optimal (SO) on E_3 as the demand is varied from 0 to 5 times the original demand matrix.*



(a) POA



(b) POA-delay

Figure 7.10: *Routing efficiency on E_3 as the demand is varied from 0 to 5 times the original demand matrix: (a) Price of anarchy (POA) and (b) Price of anarchy - Delay (POA-delay). Note that the scale of POA-Delay is an order of magnitude higher than POA. At higher values of Demand Multiplier, the metrics approach similar values.*

ever, it is also clear that certain edges would be better targets for intervention.

In both the free-flow travel time and capacity sensitivity analysis, the peak edge is 167. This corresponds to a bridge edge (on M4 motorway) going into London from the ring road (M25 motorway).

7.4.2 Road Closures

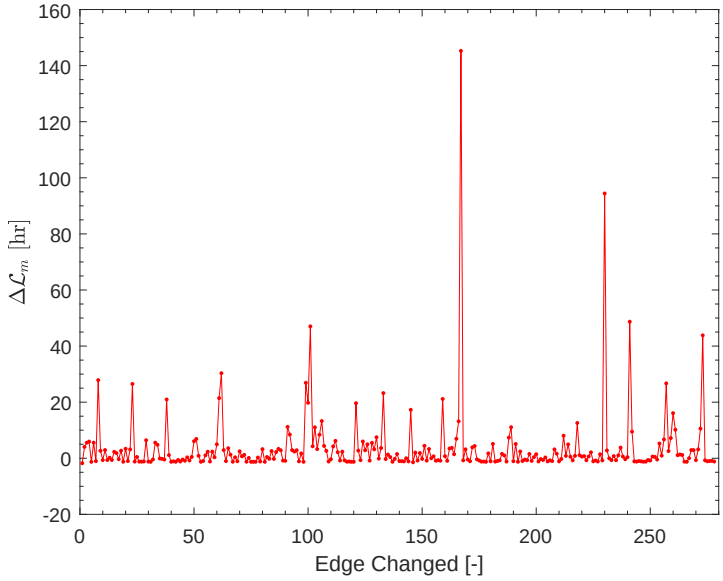
The sensitivity of congestion on E_3 to the removal of each edge can inform the impact of potential incidents and construction works.

For all the edges which are not bridge edges, the mean change in TSTT, averaged across time-bins, is negative (Figure 7.12). As such, there are no identified Inverse Braess edges which would improve congestion if removed. It can be seen that the removal of certain edges has a greater negative effect on congestion than others. The removal of any bridge edge leads to a lower TSTT as the number of trips completed is reduced from the disconnection of the network, so these results should be discounted. The edge removal with the largest increase in congestion is 51, which goes out of the west of Manchester (M56 motorway).

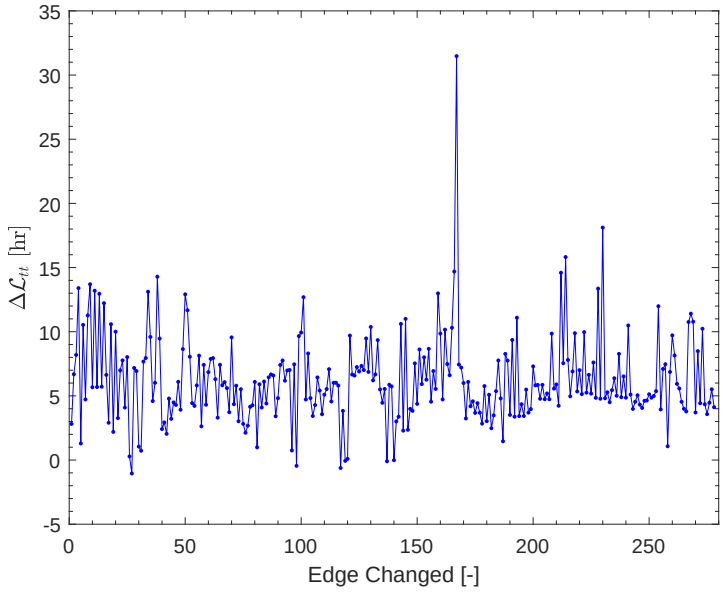
7.5 Summary

This chapter provided a series of insightful results which were obtained rapidly from cross-sectional data via the advances of Chapters 5 and 6. It proffered a high-level structural analysis of the motorway system of the English SRN covered by the MIDAS system over a year.

From the analysis, it was shown through the POA metric that the opportunity to improve overall congestion on the full-size network E_3 through rerouting selfish drivers was limited in the analysis period. When compared to the previous empirical POA studies discussed in Chapter 2, the values obtained were small. For example, the values were much smaller than the 1.5 average POA found in [10] for the Eastern Massachusetts network. By using the alternative metric of POA-delay, it was calculated that there were larger savings from rerouting with the minimum cost of trips discounted. With the investigation into the effect of



(a) Capacity



(b) Free-flow travel time

Figure 7.11: *Sensitivity analysis on E_3 for (a) capacity and (b) free-flow travel time. The data used to obtain the results are from the weekdays selected for analysis between September 2018 and August 2019. The values are the mean of all time-bins. Lines are for visual guidance only.*

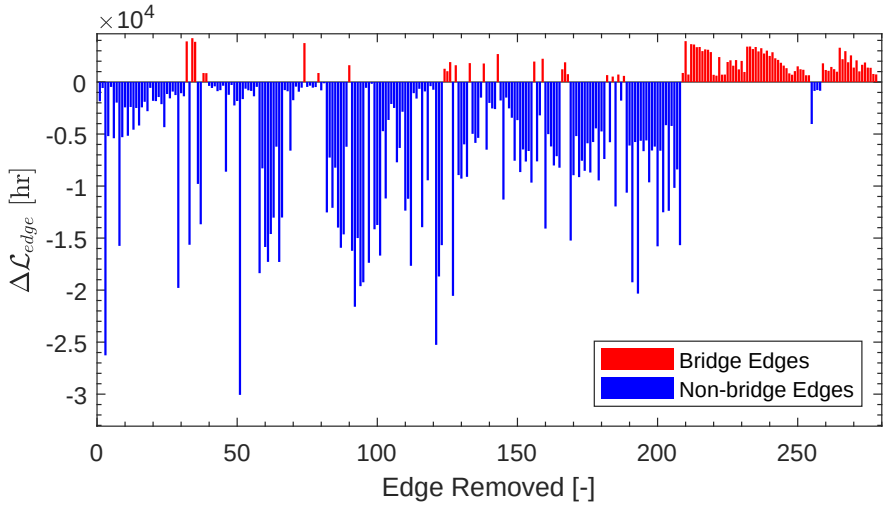


Figure 7.12: *Sensitivity analysis on E_3 for road closures. The data used to obtain the results are from the weekdays selected for analysis between September 2018 and August 2019. The values are the mean of all time-bins. Lines are for visual guidance only.*

varying demand on the network, the results showed that POA increased with a general increase in demand up until the demand is around twice the current levels. However, POA-delay had a different response to the demand changes, it saw an improvement if the demand was decreased. These results are in accordance with previous applications of POA-delay on synthetic data as found in [130]. The created TA model was used to investigate the distribution of the costs and benefits of rerouting across the road system, which was found to vary strongly with different areas and time-bin demand profiles. This distribution did not match the distribution of marginal external costs on the network.

Furthermore, the TA model was used to investigate the impact of changes to the network. Improving the capacity and free-flow travel time of several edges led to reductions in total network costs. However, the removal of any edge (i.e. road closure) on the core subnetwork was shown to increase network costs. Unlike the network tested in [21], there were no Inverse Braess' edges that caused the TSTT to decrease when removed.

The results further the understanding of rerouting on real-world road networks, however, as the analysis was only applied to the roads of the SRN it simplifies the road system of England by not including the non-SRN routing options. The analysis highlights key results such as the zonal distribution of rerouting benefits and the impact of bridge edges. Results such as these are based on the SRN, however, they could be transferred to any network with similar attributes, such as those containing bridge edges. They would likely only be more pronounced with greater network coverage. For instance, the inclusion of more routing options by incorporating the non-SRN road network in the model would be expected to increase the benefits of re-routing.

The work in this chapter showcased some of the potential analysis that the developed data-driven TA model can be used for. It shows it has the possibility to be used in further empirical research into congestion and strategic planning on real-world road networks. As a tool, it could augment existing theoretical research relating to POA (Section 2.6).

Chapter 8

Conclusions and Future Work

8.1 Conclusions

Within the broad context of transportation planning there is currently an important opportunity to take advantage of the increasing amounts of data and computing power available to improve the tools used for strategic decision making and management of road infrastructure. The coming technological advances in intelligent transportation need tools to understand traffic patterns on real-world networks. The aim of this thesis was to develop accurate and computationally efficient methods to analyse overall traffic congestion on national road networks solely using cross-sectional sensor data. In this section it is demonstrated that this aim was achieved.

The contributions of the thesis can be divided into three parts. The first focused on processing the data that would be used in a data-driven TA model. The second part developed the methods for extracting model inputs from cross-sectional traffic data. Finally, the third part quantified the extent to which application of the developed methods to a national network could be used for strategic network analysis. Four objectives were set for the thesis in Section 1.2, these are now reconsidered to show how the thesis aids the research community in achieving important contributions.

8.1.1 Data Processing

Data not carefully processed can contain errors which affect the performance of any model using it as input. As such, the first steps in the process of deriving the data-driven TA model (Figure 3.1) were data input and processing. To meet the target of **Objective 1**, the processes for extracting the topographic network representation from the NTIS data set and suitable traffic data from the MIDAS data set were developed, as shown in Chapter 4.

In Section 4.2, the need for a degenerate arterial representation of the English SRN to form the network at the base of the data-derived TA model was outlined. Section 4.2.1 described an algorithm developed from leading methods in map generalisation, specifically to be applied to the NTIS data set, which represents the physical infrastructure of the SRN. This process assisted in the production of three suitable topographic representations of different sizes and detail for the investigations using the MIDAS data set conducted in the later chapters.

Section 4.3 described the process of extracting MIDAS data from the relevant sensors and matching it to the topographic representations. It covered the process of averaging the traffic variables to approximate steady-state conditions on multi-lane highways, and assignment to weekday time-bins for static demand analysis. The data was filtered to remove days with unrepresentative traffic due to public holidays and extreme weather events. Where multiple sensors were available on a topographic edge, they were used to filter out erroneous measurements from loop detectors with problems. For the topographic representations, the processed MIDAS data was used to make adjustments to the graphs based on the availability of suitable traffic data.

The techniques developed for the processing of the MIDAS and NTIS data revealed key information to retain and that which could potentially be disregarded, necessary to produce an operative data-driven TA model of a national road system for strategic analysis.

8.1.2 Prior Origin-Destination Demand Matrix Estimation through Network Partitioning

In proposing a method for partitioning a road system through network modularity in Chapter 5, its potential was demonstrated for the calculation of the key O-D demand input for static TA models from cross-sectional data. This opened up the opportunity to estimate flow patterns for large national highways systems without the need for other data sources and met the target of **Objective 2**.

The results in Section 5.2 showed that partitioning the network into small communities of nodes was tolerable for a degenerate approach to reduce the size of the network being analysed. This degenerate approach would be well suited for use in infrastructure assessment models such as NISMOD [92] where the scale of analysis is more coarse, for instance at the inter-city level.

The non-degenerate approach is useful for application in more detailed traffic planning and producing performance comparisons of different national road systems. Applying partitioning in a non-degenerate way showed that a similar level of error in UE flow and travel time predictions was obtained by dividing the network into a small number of larger partitions. The results showed that the best accuracy results came from using only the internal O-D estimates of the partitions for the larger partition sizes. However, the results showed that by also including the external partition estimates, there was a reduction in computation time for a similar error in some cases. For the English SRN case study, it appeared that the best option was to partition the network into two large communities. In very large artificially-generated networks where the size was such that two community partitions were still infeasibly large, the results in Sections 5.2.4 and 5.2.5 showed that for community partitions numbering three and greater it was better to use the internal-only approach, unless the communities contained such a small proportion of the nodes that the flow error started to rise (approx. 11-13% of nodes, i.e eight or nine community partitions).

The performance of the methods was assessed by the prediction accuracy of the TA models using the estimated O-D matrices. The O-D ma-

trices produced, while not required to provide a close representations of the true demand profile, created a suitable demand input for the TA model which could recreate the observed traffic to varying levels of accuracy.

8.1.3 Road-specific Density-based Congestion Function Fitting

The optimal fitting of density-based road-specific congestion functions for a large national road system was identified, with the numerical investigation highlighting the benefits of using BPR functions towards this scope. This form of fitting was shown to have superior TA model accuracy compared to the state of the art for national networks solely using cross-sectional sensor data. The investigations in Chapter 6 which produced these results met the target of **Objective 3**.

In Section 6.3.1, from the testing of density-based fitting of congestion functions on a subnetwork of the English SRN, it was found that the most accurate choice of function for use in that type of fitting was BPR. Despite BPR's perceived issues, for the case of unsignalised motorways it fitted the shape of the data better than other candidates. Alternative functions such as the tested exponential equation with a similar shape to BPR also performed well. The results showed the effect that density-based fitting had compared to flow-based and hypo-critical flow-based fitting. The density-based approach was able to fit the two parameters of BPR largely independently, whereas they were correlated when they were fitted with flow-based approaches. This was due to a lack of information for fitting the hyper-critical flow regime. From the results, it was concluded that, at the network level, the density-based approach to fitting BPR was systematically superior for reliably obtaining the parameters.

In applying a density-based method, BPR-Density, for evaluating the congestion functions in a TA model, the analysis highlighted its potential advantages over more conventional (i.e. BPR-Standard) and more computationally intensive methods (i.e. Inv-Opt). The results in Section 6.3.4 showed that BPR-Density had a clear advantage over both BPR-Standard and Inv-Opt in the estimation of the travel time accuracy

when applied in a TA model to predict the UE assignment. This improved accuracy would have a positive impact on model application in cost-benefit analyses and TA-based emissions models.

Compared to BPR-Density, Inv-Opt did have the advantage that it only required flow data to fit the function, which could have been useful if that was the only data type available. Also, while it was not possible to apply Inv-Opt to the full size network, the functions fitted to the reduced subnetwork were of reasonable accuracy when applied to the more complex subnetwork. This could work well with the degenerate approach to O-D estimation presented in Chapter 5. However, these advantages may have been outweighed by the algorithmic complexity, which translated into large differences in computation time. The results in Section 6.3.5 revealed a very large difference in time to compute the functions between Inv-Opt and BPR-Density, strongly suggesting the latter is more suited to the purpose of strategic transport planning on large SRNs, where the functions may need to be updated regularly. It appeared that the travel time for an individual edge was dependent on the physical features of the road, such as how it was connected to other roads. As these features often do not change throughout the day, BPR-Density's superior performance suggests it may prove advantageous to fit a function based on road characteristics, rather than the time of day.

8.1.4 Strategic National Traffic Analysis

By creating an accurate data-driven TA model of the English SRN monitored by MIDAS and using it for system level analysis of rerouting traffic and network changes, Chapter 7 delivered **Objective 4**.

From the national analysis of England's SRN, it was shown in Section 7.3.1 that overall the opportunity to improve congestion on the E_3 subnetwork through only rerouting drivers was limited. Overall, the saving in total time across the network was around 0.1%. However, by utilising the alternative metric of POA-delay, it was calculated that with the minimum cost of trips discounted, there was a 1-2% improvement in delays to journeys. While this was still a small number, over the course of a year at the national level, such an amount may not be insignificant.

The reasons for the small improvements available from rerouting selfish drivers into a SO pattern can be attributed to the following. Firstly, the structure of the network contained a number of spurs that consisted of bridge edges; these were only accessible for one route, which reduced the potential rerouting cost differences. In general, the routing opportunities on the network were low. Secondly, the demand on the network was the average over a year for each of the time-bins. As such, the congestion modelled was not the worst case peak congestion when demand on individual days was at its highest. Thirdly, the only cost considered in the analysis was time. The inclusion of all the other costs involved in road transportation such as operating costs (e.g. fuel, maintenance), road accidents and pollution (e.g. CO_2, NO_x) would have likely increased the differences in total costs from rerouting.

The created TA model allowed the investigation in Section 7.3.2 into how the costs and benefits of rerouting were distributed across the road system. It was seen that different areas of the network benefited more than (or sometimes at the expense of) other areas. This distribution of the benefits varied for the different times of day and demand patterns. The distribution did not align with the distribution of marginal external costs on the network. This highlighted how the model could be used to further understand the impact of rerouting on individual regions and the fairness of any potential rerouting scheme, including those based on marginal external cost road pricing. The analysis of zones also showed the improvements from rerouting for the core subnetwork with the bridge edges removed.

An argument in favour of rerouting UE patterns to SO could be further supported if the demand was to be increased even in small measure. In Section 7.3.3, the results showed that POA increased with a general increase in demand, up until the demand was around twice current levels. Although the POA-delay did not improve with the same increase in demand, it was shown that there would be an improvement in delay savings if the demand was to decrease.

Furthermore, in Section 7.4.1 the model was used to investigate the potential impact of changes to the network. The analysis showed how possible benefits were obtainable from improving the capacity of sev-

eral edges in particular. This could be done through extra lane provision, however, further investigation would be needed, particularly into any induced demand effects [2]. The analysis also highlighted how the model could be used to prioritise interventions on the network to the free-flow travel time, such as varying the speed limit. The road closure analysis in Section 7.4.2 demonstrated that congestion would increase if any edge on the non-bridge core subnetwork was removed. The relative amount of congestion varied between the edges, which provided insight into the impact of incidents or construction on the network regarding its resilience. From this analysis, it can be concluded that the English SRN did not have any sections which could have been closed to improve performance (i.e Inverse Braess edges), which is reasonable given its purpose as a strategic trunk system of the broader English road transport network.

8.2 Recommendations for Future Research

Whilst working on this thesis, various areas of research concerning the development of data-driven TA models using cross-sectional data were identified, which in the future could be undertaken. There is also space for further research into empirical analysis of traffic patterns on real-world road systems. This section outlines the research recommendations for further work that could be pursued to extend this thesis.

Future Work on Data Sets and Processing

Future research could continue to improve the simplification of map data for topographic representations, perhaps using further techniques in machine learning it could improve the automation of junction detection.

There are many opportunities to improve cross-section and loop detector data cleaning methods used on the MIDAS data set. For example, with access to VMS data, times when actions such as temporary speed limits were in place could be identified, which could help with handling data of unrepresentative traffic. Furthermore, more robust algorithms could be employed to improve the identification of faults in the loop detector data, which has its own unique set of challenges.

The scope of the analysis in this thesis could be expanded by the inclusion of other data sets in addition to MIDAS. Other parts of the English road network are covered by systems different to MIDAS, such as TMU. Also, using regional and city traffic data would open the possibility of expanding the research into alternative types of road systems. Access to this extra data would create the opportunity to assess the developed methods over broader road networks than in this thesis.

Future Work on Cross-sectional Data-driven OD Estimation

Future research could investigate further ways of utilising the partitions other than the internal, external and degenerate covered in this thesis. Such alternatives could look into combining nodes from separate partitions in different combinations.

Also, future work could look to apply the type of multi-scale demand estimation developed in this thesis with alternative techniques of prior O-D estimation used on the partitions. Improvements to the GLS method could be a way of increasing accuracy and reducing computational requirements. Further experiments on the maximum number of feasible routes used in the GLS method could improve understanding of its application.

Furthermore, the methods of obtaining the prior matrix could be trialled with alternative adjustment procedures, such as genetic-based algorithms, which would allow greater experimentation with distance metrics. Also, research could look into how to incorporate separate terms and weightings in the O-D adjustment for the internal and external partition estimations of the prior matrix.

A worthwhile future investigation would be to look into the impact of partition size on the number of days of flow data required to obtain a result of suitable accuracy with the developed methods.

Additional cross-sectional data could be used to augment the O-D estimates, such as ramp flow measurements to constrain the total vehicles entering or leaving the network at a node. Further research could also incorporate other non-traffic data sources to inform the division between

the O-D pairs of the externally estimated O-D movements. For example, in the AM period a greater share of demand could be distributed to the destinations where more employment is located.

Future Work on Congestion Function Estimation

The congestion functions analysed in this thesis were for single class traffic. This could be expanded to look into the effect of multiple classes (e.g. cars, trucks, etc.) on the fittings obtained and TA results. As larger vehicles, such as trucks, are known to have a larger impact on congestion [170], the expansion of the function estimation to include multiple classes would an option for increasing model accuracy.

In future work, improvements to the estimation of congestion functions using density-based fitting may be possible with more accurate calculations of traffic density through additional data to MIDAS capable of accurate spatial measurements.

The use of density to represent congested conditions in TA models has been shown to be effective. The accuracy of TA could potentially be improved by using a TAP fully based on density, such as formulated in [171]. However, at present the methodology is not sufficiently developed for accurate and efficient calculation of traffic patterns using these formulations. Future work could fill this gap and expand the data-driven methods in this thesis into a full density-based TA model.

In this thesis, the density-based BPR fitting uses the standard coefficients and NTIS capacities for the edges which do not have suitable data for fitting. With additional data to MIDAS, potentially from VMS data, the edges with multiple speed limits could be included after data cleaning. Further work could look into techniques to automate the identification of edges with unsuitable data. Also, it could look into estimating the coefficients and capacities of such edges from those fitted to other edges on the network using additional attributes (e.g. road grade).

Future Work for Strategic National Traffic Analysis Applications

Future work could expand scope of the analysis in Chapter 7 by including greater road coverage to increase routing options, alternative time

periods other than year-average static demand, and include more transportation costs (e.g. emissions).

Additional sensitivity analysis could look into the impact of changes to the BPR parameters on individual edges to further understand how network changes at the edge level impact transport costs.

The analysis conducted in Chapter 7 looked at historical traffic patterns. The O-D matrices derived from measured traffic flows could be used as the input in online applications of traffic monitoring and future predictions of traffic scenarios. Future work could look into adapting the methods of this thesis for such purposes.

Furthermore, there are many opportunities for further research utilizing the highly transferable nature of this thesis' contributions. Future work could apply analysis similar to that in this thesis to data from numerous other countries (e.g. [49]), to compare the network performance of the English SRN. This could enable the evaluation of different government policies towards national road infrastructure.

Finally, the techniques developed for extracting computationally efficient and accurate inputs for traffic assignment models could have commercial implications. Evaluating better congestion functions and demand profiles using widely available cross-sectional data could enable transport planners and consultancies to improve their models' size and accuracy without the need for expensive trip surveys, privacy-sensitive proprietary routing data, and excessively large computing resources. Engineering consultancies could use the techniques of this thesis to conduct better strategic transport analysis at a lower cost.

8.3 Closing Remarks

This thesis developed accurate and computationally efficient methods to analyse traffic congestion on national road systems solely using cross-sectional sensor data. The methods were tested on a real-world strategic road system and it was demonstrated that they were capable of producing insightful results for planning at the national level.

The thesis first developed techniques to extract the building blocks of a data-driven model for the English SRN. This included data processing to produce suitable traffic data input and a topographic representation of the network. Then, techniques for the estimation of the key components of TA models were developed to work with the data restrictions. The new technique of utilising partitioning to estimate the O-D demand matrix removed the network size restrictions of previous approaches whilst maintaining similar accuracy. The use of density-based road-specific fitting of BPR functions was found to be the optimal choice to enable the efficient and accurate calculation of national traffic patterns. These developed techniques unlocked the ability to use the available data sets for the strategic analysis of the English SRN. This analysis quantified the inefficiency from selfish driving and the impact of potential targeted interventions, illustrating the utility of the developed model to the transport planning community.

Future work has been identified which could advance the knowledge contributed in this thesis. With further development of the methods and additional types of data, following models could build on this thesis and help provide the strategic planning necessary to tackle congestion in a changing world.

Chapter 9

Bibliography

[1] INRIX. INRIX Research: Global Traffic Scorecard. Technical report, December 2021.

[2] K. M. Hymel, K. A. Small, and K. V. Dender. Induced demand and rebound effects in road transport. *Transportation Research Part B: Methodological*, 44(10):1220–1241, 2010.

[3] A. D. May and C. A. Nash. Urban congestion: A European perspective on theory and practice. *Annual Review of Energy and the Environment*, 21(1):239–260, November 1996.

[4] L. A. Klein, D. R. P. Gibson, and M. K. Mills. *Traffic Detector Handbook*. Federal Highway Administration, Turner-Fairbank Highway Research Center, 3 edition, 2006.

[5] National Highways. National Traffic Information Service DATEX II Service v12. Technical report, London, 2022.

[6] Highways Agency. NMCS2 MIDAS Outstation Algorithm Specification. Technical Report TR 2177, Issue H, March 2009.

[7] C. Lomax. Highways England policy for the use of Variable Signs and Signals (VSS). Technical report, Highways England, London, 2018.

[8] Office of Rail and Road. Benchmarking Highways England: 2020 progress report. Technical report, Office of Road and Rail, London, February 2021.

[9] S. Maerivoet and B. D. Moor. Transportation planning and traffic flow models. Technical Report 05-155, Katholieke Universiteit Leuven, July 2005.

[10] J. Zhang, S. Pourazarm, C. G. Cassandras, and I. C. Paschalidis. The Price of Anarchy in Transportation Networks: Data-Driven Evaluation and Reduction Strategies. *Proceedings of the IEEE*, 106(4):538–553, 2018.

[11] Y. Sheffi. *Urban transportation networks: Equilibrium analysis with mathematical programming methods*. Prentice-Hall Inc, Englewood Cliffs, N.J., 1 edition, 1985.

[12] V. R. Melnikov, V. V. Krzhizhanovskaya, A. V. Boukhanovsky, and P. M. Sloot. Data-driven Modeling of Transportation Systems and Traffic Data Analysis during a Major Power Outage in the Netherlands. In *Procedia Computer Science*, 2015.

[13] S. P. Hoogendoorn and P. H. Bovy. Generic gas-kinetic traffic systems modeling with applications to vehicular traffic flow. *Transportation Research Part B: Methodological*, 2001.

[14] G. Sharon, J. P. Hanna, T. Rambha, M. W. Levin, M. Albert, S. D. Boyles, and P. Stone. Real-time adaptive tolling scheme for optimized social welfare in traffic networks. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, volume 2, pages 828–836, 2017.

[15] V. L. Knoop and S. P. Hoogendoorn. Network Transmission Model : a dynamic traffic model at network level. *93rd Annual Meeting of the Transportation Research Board*, 2014.

[16] T. Seo, A. M. Bayen, T. Kusakabe, and Y. Asakura. Traffic state estimation on highway: A comprehensive survey. *Annual Reviews in Control*, 43:128–151, 2017.

[17] Y. Wang and M. Papageorgiou. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transportation Research Part B: Methodological*, 39(2):141–167, 2005.

[18] J. Van Lint and S. P. Hoogendoorn. A robust and efficient method for fusing heterogeneous data from traffic sensors on freeways. *Computer-Aided Civil and Infrastructure Engineering*, 25(8):596–612, 2010.

[19] Y. Kim, P. Wang, Y. Zhu, and L. Mihaylova. A Capsule Network for Traffic Speed Prediction in Complex Road Networks. *2018 Symposium on Sensor Data Fusion: Trends, Solutions, Applications, SDF 2018*, 2018.

[20] H. Youn, M. T. Gastner, and H. Jeong. Price of anarchy in transportation networks: Efficiency and optimality control. *Physical Review Letters*, 101(12):1–4, 2008.

[21] J. Ivanchev, S. C. Litescu, D. Zehe, M. Lees, H. Aydt, and A. Knoll. Hard and Soft Closing of Roads Towards Socially Optimal Routing. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2018-November:3499–3504, 2018.

[22] R. J. Wilson. *Introduction to Graph Theory*. Prentice Hall/Pearson, New York, 2010.

[23] S. Lee Loh, C. Kim Gan, T. Han Cheong, S. Salleh, and N. H. Sarmin. An overview on network diagrams: Graph-based representation. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 8(2):83–86, May 2016.

[24] M. Patriksson. *The Traffic Assignment Problem: Models and Methods*. Dover Publications, Mineola, N.Y., 2 edition, 2015.

[25] L. De Grange, C. Melo-Riquelme, C. Burgos, F. González, and S. Raveau. Numerical Bounds on the Price of Anarchy. *Journal of Advanced Transportation*, 2017:1–9, 2017.

[26] M. J. Beckmann, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*. RAND Corporation, Santa Monica, CA, 1956.

[27] S. C. Dafermos and F. T. Sparrow. Traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards, Section B: Mathematical Sciences*, 73B(2):91, 1969.

[28] T. Potuzak and F. Kolovsky. Parallelization of the B static traffic assignment algorithm. *Ain Shams Engineering Journal*, 13(2):101576, 2022.

[29] J. Xie, Y. M. Nie, and X. Liu. A greedy path-based algorithm for traffic assignment. *Transportation Research Record*, 2672(48):36–44, 2018.

[30] R. B. Dial. A path-based user-equilibrium traffic assignment algorithm that obviates path storage and enumeration. *Transportation Research Part B: Methodological*, 40(10):917–936, 2006.

[31] H. Bar-Gera. Origin-based algorithm for the traffic assignment problem. *Transportation Science*, 36(4):398–417, 2002.

[32] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[33] L. LeBlanc, E. Morlok, and W. Pierskalla. An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research*, 9:309–318, 10 1975.

[34] D. E. Boyce, H. S. Mahmassani, and A. Nagurney. A retrospective on Beckmann, McGuire and Winsten's Studies in the Economics of Transportation. *Papers in Regional Science*, 84(1):85–103, 2005.

[35] T. Rambha, S. D. Boyles, A. Unnikrishnan, and P. Stone. Marginal cost pricing for system optimal traffic assignment with recourse under supply-side uncertainty. *Transportation Research Part B: Methodological*, 110:104–121, 2018.

[36] Q. Wang and J. Wang. A bi-level multi-objective optimization model for micro-circulation road networks in the open block area considering traffic pollution and intersection delays. *IEEE Access*, 9:129278–129292, 2021.

[37] A. Schneck and K. Nökel. Accelerating traffic assignment with customizable contraction hierarchies. *Transportation Research Record*, 2674(1):188–196, 2020.

[38] X. Chen, Z. Liu, and I. Kim. A parallel computing framework for solving user equilibrium problem on computer clusters. *Transportmetrica A: Transport Science*, 16(3):550–573, 2020.

[39] M. C. Bliemer, M. P. Raadsen, L. J. Brederode, M. G. Bell, L. J. Wismans, and M. J. Smith. Genetics of traffic assignment models for strategic transport planning. *Transport Reviews*, 37(1):56–78, January 2017.

[40] Y.-C. Chiu, J. Bottom, M. Mahut, A. Paz, R. Balakrishna, T. Waller, and J. Hicks. *Dynamic Traffic Assignment: A Primer*. Transport Research Board, Washington, D.C., June 2011.

[41] L. Brederode, A. Pel, L. Wismans, E. de Romph, and S. Hoogendoorn. Static traffic assignment with queuing: model properties and applications. *Transportmetrica A: Transport Science*, 15:179–214, January 2019.

[42] P. N. Patil, K. C. Ross, and S. D. Boyles. Convergence behavior for traffic assignment characterization metrics. *Transportmetrica A: Transport Science*, 17(4):1244–1271, 2021.

[43] X. Zeng, X. Guan, H. Wu, and H. Xiao. A data-driven quasi-dynamic traffic assignment model integrating multi-source traffic sensor data on the expressway network. *ISPRS International Journal of Geo-Information*, 2021.

[44] M. Treiber and A. Kesting. *Traffic Flow Dynamics: Data, Models and Simulation*. Springer, Berlin, Heidelberg, 2013.

[45] D. Cvetek, M. Mustra, N. Jelušić, and L. Tisljaric. A survey of methods and technologies for congestion estimation based on multisource data fusion. *Applied Sciences*, 11:2306, March 2021.

[46] A. Landmark, P. Arnesen, C.-J. Södersten, and O. Hjelkrem. Mobile phone data in transportation research: methods for benchmarking against other data sources. *Transportation*, 48:1–23, October 2021.

[47] Y. Liao, S. Yeh, and J. Gil. Feasibility of estimating travel demand using geolocations of social media data. *Transportation*, 49(1):137–161, 2022.

[48] V. Mahajan, N. Kuehnel, A. Intzevidou, G. Cantelmo, R. Moeckel, and C. Antoniou. Data to the people: a review of public and proprietary data for transport models. *Transport Reviews*, 0(0):1–26, 2021.

[49] GraphHopper. GraphHopper Open Traffic Collection, 2022. Available online at: `https://github.com/graphhopper/open-traffic-collection`, last accessed on 01/09/2022.

[50] D. M. Bramich, M. Menéndez, and L. Ambühl. Fitting empirical fundamental diagrams of road traffic: A comprehensive review and comparison of models using an extensive data set. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–24, 2022.

[51] A. Bhaskar and E. Chung. Fundamental understanding on the use of bluetooth scanner as a complementary transport data. *Transportation Research Part C: Emerging Technologies*, 37:42–72, 2013.

[52] V. Knoop, S. P. Hoogendoorn, and H. van Zuylen. Empirical differences between time mean speed and space mean speed. In *Traffic and Granular Flow '07*, pages 351–356, Berlin, 2009. Springer.

[53] B. J. Park, T. Kim, I. Yang, J. Heo, and B. Son. A method for measuring accurate traffic density by aerial photography. *Journal of Advanced Transportation*, 49(4):568–580, June 2015.

[54] E. Barmpounakis and N. Geroliminis. On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment. *Transportation Research Part C: Emerging Technologies*, 111:50–71, 2020.

[55] M. L. Hazelton. Estimation of origin-destination matrices from link flows on uncongested networks. *Transportation Research Part B: Methodological*, 34(7):549–566, 2000.

[56] E. Cascetta, A. Papola, V. Marzano, F. Simonelli, and I. Vitiello. Quasi-dynamic estimation of o–d flows from traffic counts: Formulation, statistical validation and performance analysis on real data. *Transportation Research Part B: Methodological*, 55:171–187, 2013.

[57] A. Horni, K. Nagel, and K. Axhausen, editors. *Multi-Agent Transport Simulation MATSim*. Ubiquity Press, London, Aug 2016.

[58] Y. Ren, M. Ercsey-Ravasz, P. Wang, M. C. González, and Z. Toroczkai. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nature Communications*, 5, 2014.

[59] E. Jafari, V. Pandey, and S. D. Boyles. A decomposition approach to the static traffic assignment problem. *Transportation Research Part B: Methodological*, 105:270–296, 2017.

[60] W. A. Mackaness and G. A. Mackechnie. Automating the detection and simplification of junctions in road networks. *GeoInformatica*, 3(2):185–200, 1999.

[61] Q. Zhou and Z. Li. Experimental analysis of various types of road intersections for interchange detection. *Transactions in GIS*, 19(1):19–41, 2015.

[62] T. Abrahamsson. *Estimation of Origin-Destination Matrices Using Traffic Counts - A Literature Survey*. IIASA, Laxenburg, Austria, 1998.

[63] S. Bera and K. V. Rao. Estimation of origin-destination matrix from traffic counts: The state of the art. *European Transport - Trasporti Europei*, (49):2–23, December 2011.

[64] M. L. Hazelton. Some comments on origin-destination matrix estimation. *Transportation Research Part A: Policy and Practice*, 37(10):811–822, 2003.

[65] Y. Vardi. Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91(433):365–377, 1996.

[66] H. P. Lo, N. Zhang, and W. H. K. Lam. Estimation of an origin-destination matrix with random link choice proportions: A statistical approach. *Transportation Research Part B: Methodological*, 30(4):309–324, August 1996.

[67] S. Dey, S. Winter, and M. Tomko. Origin–destination flow estimation from link count data only. *Sensors*, 20(18), 2020.

[68] M. Rostami Nasab and Y. Shafahi. Estimation of origin–destination matrices using link counts and partial path data. *Transportation*, 47(6):2923–2950, 2020.

[69] X. Yang, Y. Lu, and W. Hao. Origin-Destination Estimation Using Probe Vehicle Trajectory and Link Counts. *Journal of Advanced Transportation*, 2017:4341532, 2017.

[70] K. Parry and M. L. Hazelton. Estimation of origin–destination matrices from link counts and sporadic routing data. *Transportation Research Part B: Methodological*, 46(1):175–188, 2012.

[71] C. Tebaldi and M. West. Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93(442):557–573, 1998.

[72] H. Spiess. A gradient approach for the O-D matrix adjustment problem. *Centre for research on transportation, University of Montreal, Canada*, 693:1–11, 1990.

[73] J. T. Lundgren and A. Peterson. A heuristic for the bilevel origin-destination-matrix estimation problem. *Transportation Research Part B: Methodological*, 42(4):339–354, 2008.

[74] S. N. Patil, K. N. Behara, A. Khadhir, and A. Bhaskar. Methods to enhance the quality of bi-level origin–destination matrix adjustment process. *Transportation Letters*, 0(0):1–10, 2022.

[75] E. Cascetta. Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. *Transportation Research Part B: Methodological*, 18(4):289–299, 1984.

[76] M. G. Bell. The estimation of origin-destination matrices by constrained generalised least squares. *Transportation Research Part B: Methodological*, 25(1):13–22, 1991.

[77] E. Codina and J. Barceló. Adjustment of O–D trip matrices from observed volumes: An algorithmic approach based on conjugate directions. *European Journal of Operational Research*, 155(3):535–557, 2004.

[78] R. Frederix, F. Viti, R. Corthout, and C. M. J. Tampère. New gradient approximation method for dynamic origin–destination matrix estimation on congested networks. *Transportation Research Record*, 2263(1):19–25, 2011.

[79] R. Frederix, F. Viti, and C. M. Tampère. Dynamic origin–destination estimation in congested networks: theoretical findings and implications in practice. *Transportmetrica A: Transport Science*, 9(6):494–513, 2013.

[80] M. L. Hazelton. Estimation of Origin-Destination Trip Rates in Leicester. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50(4):423–433, 2001.

[81] T. Aynaud and J.-L. Guillaume. Static community detection algorithms for evolving networks. In *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, pages 513–519, 2010.

[82] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[83] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikolski, and D. Wagner. On modularity - np-completeness and beyond. Technical report, Faculty of Informatics, Universitat Karlsruhe, 2006.

[84] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[85] J. Zhang, J. Fei, X. Song, and J. Feng. An improved louvain algorithm for community detection. *Mathematical Problems in Engineering*, 2021:1485592, Nov 2021.

[86] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019.

[87] D. Leeuwen, J. Bosman, and E. Dugundji. Network partitioning on time-dependent origin-destination electronic trace data. *Personal and Ubiquitous Computing*, 23, 11 2019.

[88] M. S. Ahmed and M. Hoque. Partitioning of urban transportation networks utilizing real-world traffic parameters for distributed simulation in SUMO. pages 1–4, December 2016.

[89] T. Dantsuji, S. Hirabayashi, Q. Ge, and D. Fukuda. Cross comparison of spatial partitioning methods for an urban transportation network. *International Journal of Intelligent Transportation Systems Research*, 18, December 2019.

[90] X. Lin and J. Xu. Road network partitioning method based on canopy-k means clustering algorithm. *Archives of Transport*, 54:95–106, June 2020.

[91] Q. Yu, W. Li, D. Yang, and H. Zhang. Partitioning urban road network based on travel speed correlation. *International Journal of Transportation Science and Technology*, 10(2):97–109, 2021.

[92] S. Blainey and J. Preston. Predict or prophesy? Issues and trade-offs in modelling long-term transport infrastructure demand and capacity. *Transport Policy*, 74:165–173, February 2019.

[93] H. Spiess. Conical volume-delay functions. *Transportation Science*, 24(2):153–158, May 1990.

[94] US Bureau of Public Roads. *Traffic assignment manual for application with a large, high speed computer.* U.S. Dept. of Commerce, Urban Planning Division, Washington, D.C., 1964.

[95] D. Branston. Link capacity functions: A review. *Transportation Research*, 10(4):223–236, 1976.

[96] K. B. Davidson. A flow–travel time relationship for use in transportation planning. *Proceedings, Australian Road Research Board*, 3:183–194, 1966.

[97] R. Akçelik. Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function. *Australian Road Research*, 21, January 1991.

[98] E. T. Mtoi and R. Moses. Calibration and Evaluation of Link Congestion Functions: Applying Intrinsic Sensitivity of Link Speed as a Practical Consideration to Heterogeneous Facility Types within

Urban Network. *Journal of Transportation Technologies*, 04(02):141–149, 2014.

[99] Transport for London. *Traffic modelling guidelines. version 3.0.* Transport for London, London, 2010.

[100] R. Moses, E. Mtoi, H. McBean, and S. Ruegg. *"Development of Speed Models for Improving Travel Forecasting and Highway Performance Evaluation", Final Report*. Florida Department of Transportation (FDOT), Florida, 2013.

[101] National Research Council (U.S.). Transportation Research Board. *HCM 2010: Highway Capacity Manual.* Transportation Research Board, Washington, D.C., 5th edition, 2010.

[102] UK Department for Transport. *TAG Unit M3.1 Highway Assignment Modelling*. Department for Transport, London, 2014.

[103] L. V. Leong. Effects of volume-delay function on time, speed and assigned volume in transportation planning process. *International Journal of Applied Engineering Research*, 11(13):8010–8018, 2016.

[104] Z. Irawan, T. Sumi, and A. Munawar. Implementation of the 1997 Indonesian Highway Capacity Manual (MKJI) Volume Delay Function. *Journal of the Eastern Asia Society for Transportation Studies*, 8:350–360, 2010.

[105] D. Nobel and S. Yagi. Network Assignment Calibration of BPR Function: A Case Study of Metro Manila, the Philippines. *Journal of the Eastern Asia Society for Transportation Studies*, 12:598–615, 2017.

[106] M. Bally, A. Alrawi, and L. Leong. Compatibility between delay functions and highway capacity manual on Iraqi highways. *Open Engineering*, 12:359–372, May 2022.

[107] G. Casey, B. Zhao, K. Kumar, and K. Soga. Context-specific volume–delay curves by combining crowd-sourced traffic data with automated traffic counters: A case study for London. *Data-Centric Engineering*, 1(e18), 2020.

[108] R. Kucharski and A. Drabicki. Estimating macroscopic volume delay functions with the traffic density derived from measured speeds and flows. *Journal of Advanced Transportation*, 2017:1–10, 2017.

[109] X. Wu and X. S. Zhou. Characterization and calibration of volume-to-capacity ratio in volume-delay functions on freeways based on a queue analysis approach. In *Transportation Research Board 100th Annual Meeting*, page 23p, Washington, D.C., January 2021. TRB.

[110] G. Gentile, L. Meschini, and N. Papola. Macroscopic arc performance models with capacity constraints for within-day dynamic traffic assignment. *Transportation Research Part B: Methodological*, 39(4):319–338, May 2005.

[111] L. Huntsinger and N. Rouphail. Bottleneck and queuing analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2255:117–124, December 2011.

[112] A. Boardman, D. Greenberg, A. Vining, and D. Weimer. *Cost-Benefit Analysis: Pearson New International Edition*. Pearson Education, Harlow, England, 2013.

[113] Department for Transport. Transport analysis guidance: WebTAG, 2022.

[114] H. Wen and Z. Yang. Removing Links: The Inverse of Braess Paradox in a Complex Network. In *ICTE 2013 - Proceedings of the 4th International Conference on Transportation Engineering*, pages 2977–2981, 2013.

[115] S. A. Bagloee, A. A. Ceder, M. Tavana, and C. Bozic. A heuristic methodology to tackle the Braess Paradox detecting problem tailored for real road networks. *Transportmetrica A: Transport Science*, 2014.

[116] D. Braess, A. Nagurney, and T. Wakolbinger. On a Paradox of Traffic Planning. *Transportation Science*, 39(4):446–450, 2005.

[117] A. Belov, K. Mattas, M. Makridis, M. Menendez, and B. Ciuffo. A microsimulation based analysis of the price of anarchy in traffic

routing: The enhanced braess network case. *Journal of Intelligent Transportation Systems*, 26(4):448–460, 2022.

[118] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *Proceedings of the 16th Annual Conference on Theoretical Aspects of Computer Science*, volume 1563, pages 404–413, Trier, Germany, 1999. Springer-Verlag.

[119] C. Papadimitriou. Algorithms, games, and the internet. In *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing*, pages 749–753, 2001.

[120] T. Roughgarden. Intrinsic Robustness of the Price of Anarchy. *Journal of the ACM*, 62(5):1–42, 2015.

[121] M. Feldman, N. Immorlica, B. Lucier, T. Roughgarden, and V. Syrgkanis. The price of anarchy in large games. *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing - STOC 2016*, pages 963–976, 2016.

[122] T. Roughgarden, V. Syrgkanis, and É. Tardos. The price of anarchy in auctions. *Journal of Artificial Intelligence Research*, 2017.

[123] T. Roughgarden and É. Tardos. How bad is selfish routing? *Journal of the ACM*, 49(2):236–259, March 2002.

[124] J. R. Correa, A. S. Schulz, and N. E. Stier-Moses. A geometric approach to the price of anarchy in nonatomic congestion games. *Games and Economic Behavior*, 64(2):457–469, 2008.

[125] F. Benita, V. Bilò, B. Monnot, G. Piliouras, and C. Vinci. Data-Driven Models of Selfish Routing: Why Price of Anarchy Does Depend on Network Topology. In X. Chen, N. Gravin, M. Hoefer, and R. Mehta, editors, *Web and Internet Economics*, pages 252–265, Cham, 2020. Springer International Publishing.

[126] S. J. O'Hare, R. D. Connors, and D. P. Watling. Mechanisms that govern how the Price of Anarchy varies with travel demand. *Transportation Research Part B: Methodological*, 84:55–80, 2016.

[127] Z. Wu, R. H. Möhring, Y. Chen, and D. Xu. Selfishness need not be bad. *Operations Research*, 69(2):410–435, 2021.

[128] B. Monnot, F. Benita, and G. Piliouras. Routing games in the wild: Efficiency, equilibration and regret. In N. R. Devanur and P. Lu, editors, *Web and Internet Economics*, pages 340–353, Cham, 2017. Springer International Publishing.

[129] B. Monnot, F. Benita, and G. Piliouras. Routing Games in the Wild: Efficiency, Equilibration, Regret, and a Price of Anarchy Bound via Long Division. *ACM Trans. Econ. Comput.*, 10(1), April 2022.

[130] S. J. O'Hare. *The Influence of Structure in Supply and Demand on the Performance Characteristics of Road Traffic Networks: An exploration of how methodological approaches from network science can be implemented for a transportation research problem*. PhD thesis, University of Leeds, March 2015.

[131] G. Santos, H. Behrendt, L. Maconi, T. Shirvani, and A. Teytelboym. Part I: Externalities and economic policies in road transport. *Research in Transportation Economics*, 28(1):2–45, 2010.

[132] A. Anas and R. Lindsey. Reducing urban road transportation externalities: Road pricing in theory and in practice. In *Review of Environmental Economics and Policy*, volume 5, pages 66–88, 2011.

[133] Z. Gu, Z. Liu, Q. Cheng, and M. Saberi. Congestion pricing practices and public acceptance: A review of evidence. *Case Studies on Transport Policy*, 6(1):94–101, 2018.

[134] M. D. Simoni, K. M. Kockelman, K. M. Gurumurthy, and J. Bischoff. Congestion pricing in a world of self-driving vehicles: An analysis of different strategies in alternative future scenarios. *Transportation Research Part C: Emerging Technologies*, 98:167–185, 2019.

[135] P. Cramton, R. R. Geddes, and A. Ockenfels. Set road charges in real time to ease traffic. *Nature*, 2018.

[136] R. Maggistro and G. Como. Stability and optimality of multi-scale transportation networks with distributed dynamic tolls. In *Proceedings of the IEEE Conference on Decision and Control*, volume 2018-Decem, pages 211–216, 2019.

[137] J. Holguín-veras and M. Cetin. Optimal tolls for multi-class traffic : Analytical formulations and policy implications. *Transportation Research Part A*, 43(4):445–467, 2009.

[138] Z. Liu, S. Wang, B. Zhou, and Q. Cheng. Robust optimization of distance-based tolls in a network considering stochastic day to day dynamics. *Transportation Research Part C: Emerging Technologies*, 79(2017):58–72, 2017.

[139] R. Lindsey and E. Verhoef. Traffic Congestion And Congestion Pricing. In *Handbook of Transport Systems and Traffic Control*, pages 77–105. Emerald Group Publishing Limited, 3 edition, 2001.

[140] N. Gartner, C. Messer, and A. Rathi. *Traffic Flow Theory: A State-of-the-Art Report*. Federal Highway Administration, Washington, D.C., 01 2001.

[141] D. Boyce. Beckmann's transportation network equilibrium model: Its history and relationship to the Kuhn–Tucker conditions. *Economics of Transportation*, 2(1):47–52, 2013.

[142] R. Dowling and A. Skabardonis. Urban arterial speed-flow equations for travel demand models. *Transportation Research Board Conference Proceedings*, 2, January 2008.

[143] G. Dervisoglu, G. Gomes, J. Kwon, A. Muralidharan, P. Varaiya, and R. Horowitz. Automatic Calibration of the Fundamental Diagram and Empirical Observations on Capacity. In *Transportation Research Board 88th Annual Meeting*, pages 1–14, Washington, D.C., January 2009. TRB.

[144] A. P. Silvano, H. N. Koutsopoulos, and H. Farah. Free flow speed estimation: A probabilistic, latent approach. Impact of speed limit changes and road characteristics. *Transportation Research Part A: Policy and Practice*, 138:283–298, August 2020.

[145] P. Bonsall, P. Firmin, M. Anderson, I. Palmer, and P. Balmforth. Validating the results of a route choice simulator. *Transportation Research Part C: Emerging Technologies*, 5(6):371–387, 1997.

[146] A. W. Brander and M. C. Sinclair. A Comparative Study of k-Shortest Path Algorithms. In *Performance Engineering of Computer*

and Telecommunications Systems, pages 370–379. Springer, London, 1996.

[147] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022.

[148] M. Patriksson. Sensitivity analysis of traffic equilibria. *Transportation Science*, 38(3):258–281, 2004.

[149] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1 – 3, 1966.

[150] A. Linero-Bas and D. Nieves-Roldán. A generalization of the monotone convergence theorem. *Mediterranean Journal of Mathematics*, 18(4):158, June 2021.

[151] Google Maps. Map of Central England, 2021. Available online at: `https://www.google.com/maps/place/England,+UK/`, last accessed on 27/10/2021.

[152] K. Mehlhorn and P. Sanders. *Algorithms and data structures: The basic toolbox*. Springer, Berlin, 2008.

[153] V. L. Knoop and W. Daamen. Automatic fitting procedure for the fundamental diagram. *Transportmetrica B: Transport Dynamics*, 5(2):133–148, 2017.

[154] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya. Detecting errors and imputing missing data for single-loop surveillance systems. *Transportation Research Record*, 1855, September 2003.

[155] S. Robinson. *The development and application of an urban link travel time model using data derived from inductive loop detectors*. PhD thesis, Imperial College London, 2006.

[156] Y. Kim and F. L. Hall. Relationships between occupancy and density reflecting average vehicle lengths. *Transportation Research Record*, 1883(1):85–93, 2004.

[157] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69:026113, March 2004.

[158] T. Aynaud. Python-louvain x.y: Louvain algorithm for community detection, 2020. Available online at: `https://github.com/taynaud/python-louvain`, last accessed on 01/09/2022.

[159] TNFR. Transportation Networks for Research, 2022. Available online at: `https://github.com/bstabler/TransportationNetworks`, last accessed on 26/01/2022.

[160] D. Bertsimas, V. Gupta, and I. C. Paschalidis. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming*, 153(2):595–633, 2015.

[161] J. Zhang, S. Pourazarm, C. Cassandras, and I. C. Paschalidis. "InverseVIsTraffic", 2017. Available online at: `https://github.com/jingzbu/InverseVIsTraffic`, last accessed on 10/09/2022.

[162] M. J. Smith. The existence, uniqueness and stability of traffic equilibria. *Transportation Research Part B*, 1979.

[163] S. Dafermos. Traffic Equilibrium and Variational Inequalities. *Transportation Science*, 14(1):42–54, February 1980.

[164] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[165] T. Evgeniou, M. Pontil, and T. Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

[166] J. Zhang, S. Pourazarm, C. G. Cassandras, and I. C. Paschalidis. The price of anarchy in transportation networks by estimating user cost functions from actual traffic data. In *2016 IEEE 55th Conference on Decision and Control, CDC 2016*, pages 789–794, 2016.

[167] J. VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc., 1st edition, 2016.

[168] P. Narváez-Villa, B. Arenas-Ramírez, J. Mira, and F. Aparicio-Izquierdo. Analysis and prediction of vehicle kilometers traveled: A case study in spain. *International Journal of Environmental Research and Public Health*, 18(16), 2021.

[169] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[170] G. Zhang and J. Chen. Solving multi-class traffic assignment problem with genetic algorithm. In *2010 Second International Conference on Computational Intelligence and Natural Computing*, volume 2, pages 229–232, 2010.

[171] P. Kachroo and S. Sastry. Traffic Assignment Using a Density-Based Travel-Time Function for Intelligent Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*, 17(5):1438–1447, May 2016.