Knowledge Discovery in Vehicle Identification Sensor Networks

Pedro Pinto da Silva

Submitted for the degree of Computing at Newcastle University

January 2023

To my parents

Abstract

Intelligent Transport Systems (ITS) are driving innovation in road transport by integrating advances in communication and information systems with traditional engineering practices. Core to ITS development is the collection and analysis of traffic data. One class of traffic sensors, known as Automatic Vehicle Identification, is characterised by the ability to identify vehicles using unique identifiers. Through vehicle re-identification, such sensors can provide reliable estimates of travel time and inform route choices, both at the individual and aggregate levels, across all levels of the road hierarchy. In particular, Automatic Number Plate Recognition (ANPR) video cameras require just a visible number plate instead of specialised devices for vehicle detection. The benefits of ANPR technology for traffic monitoring have led to its adoption in cities across the world, forming complex sensor networks with increased potential to power ITS solutions.

Despite successful application in traffic forecasting, two technical barriers prevent a more widespread and diverse adoption of ANPR networks:

- The lack of technical guidance on pre-processing ANPR data. We address this by developing a data pipeline which documents the various data sources and processing steps required to produce traffic data ready for analysis. In addition, we benchmark the pipeline against a real ANPR network, located in the North East of England.
- The methodological gap in representing and extracting popular travel routes (corridors) from observed data. We develop a mathematical framework for corridor identification, which highlights route importance in connecting and distributing regional road traffic.

The second part of this thesis focuses on two new ITS applications of ANPR networks. They demonstrate how traffic authorities can collect evidence of corridor performance and safety issues in order to prioritise transport improvements:

- Bottleneck detection and impact assessment is a critical traffic monitoring activity largely confined to highways. By developing a detection algorithm for ANPR-monitored corridors, bottleneck detection is scaled to an entire urban network. Bottlenecks are categorised by frequency of occurrence, and their impact quantified against other sources of congestion, indicating that recurring bottlenecks account for over 75% of urban traffic congestion. Our method is the first to use ANPR sensors to automatically identify traffic bottlenecks and quantify their impact across an urban road network.
- Frequent overtaking and lane-changing behaviour can have negative impacts on traffic flow. We investigate the link between overtaking rate and traffic conditions as a proxy to understanding and quantifying corridor safety levels. Our findings

suggest that overtaking rate increases with vehicle concentration and inversely with speed, albeit with a scaling relationship that greatly depends on road characteristics. Our method is the first to be able to quantify the scaling effect of vehicle overtakings for a variety of roads and traffic conditions.

Successful traffic management increasingly relies on continuous data collection and analysis. Using our pipeline for data processing and new methodologies of data analysis, stakeholders can extract added value from their ANPR sensing infrastructure and better position themselves to fully realise the vision of intelligent traffic management systems.

Acknowledgements

I would like to thank Ray King, manager for the Tyne Wear Urban Traffic Management and Control centre, for his openness to collaborate and share the ANPR data which underpin my research and this thesis.

I wish to express my gratitude to my supervisor Professor Phil Blythe who first proposed this theme of research and helped to acquire the data resources that made my work possible. A major thanks to him for supporting my Enrichment application to the Alan Turing Institute, which turned out to be a major highlight of my PhD.

My deep gratitude extends to both of my supervisor from the School of Computing, Dr. Stephen McGough and Dr. Matthew Forshaw, for their immense patience in supervising me and supporting me throughout the entire process. They stood alongside me despite all obstacles.

I want to thank Professor Paul Watson for believing in me and offering me a place in the Cloud Computing CDT to begin with. Without this opportunity I would not be writing these words today. A big thank you to all CDT staff, peers and friends who made this journey unforgettable and a truly tremendous learning experience.

Finally, this achievement would have been possible without the support of my family, partner and friends. To my partner, thank you for listening and always being there for me. To my brother, a big hug – we are more alike than we are different. To my mom and dad, thank you for everything: my education, your support and infinite love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Intelligent Transportation Systems (ITS) have emerged to improve all aspects of transportation systems. ITS technologies and applications combine transport practices and operations with technological advancements, namely in vehicle sensing and information systems, to combat a range of efficiency, safety and environmental issues caused by the saturation of transport infrastructure (Dimitrakopoulos & Demestichas, 2010).

Central to the realisation of ITS is the collection of traffic data, achieved across an ensemble of sensing technologies (Zhang et al., 2011). Automatic Number Plate Recognition (ANPR)[1], is an increasingly used sensing technology that combines video cameras for vehicle detection and embedded software for optical character recognition (Antoniou et al., 2011; Debnath et al., 2014; Patel et al., 2013). ANPR is part of a broader family of ITS technologies, known as Automatic Vehicle Identification (AVI), characterised by the ability to detect specific vehicles across multiple sensors (placed in distinct locations) through the use of unique vehicle identifiers[2] (R. Li & Rose, 2011; Tam & Lam, 2011).

ANPR has several desirable properties. Notably, it provides reliable estimates of vehicle travel times through vehicle re-identification, useful in traffic forecasting (Kazagli et al., 2013; E. I. Vlahogianni et al., 2014). Like other AVI technologies, ANPR is non-intrusive – cameras are mounted on poles or structures above or next to the roadway, as opposed to being installed directly into the road surface (Rhead et al., 2012). Ease of installation allows ANPR to operate in a wide variety of contexts, specifically urban networks composed of arterial and collector roads (A and B-roads in the UK). Although arterials and collectors are of lower categorical rank relative to highways, they serve an equally vital role in fulfilling user trips, especially at the regional level (Department for Transport, 2019a). Sensing in arterial roads has traditionally been difficult to achieve using older technologies, which do not posses the same re-identification and travel time monitoring

[1] Also referred to as ALPR (Automatic License Plate Recognition).

[2] ANPR uniquely identifies vehicles via number plates. Other technologies, such as bluetooth and RFID, use MAC addresses or serial numbers, respectively, as unique vehicle identifiers.

capabilities as ANPR, due to complex traffic dynamics (high-density of road intersections and unrestricted travel) (Chow et al., 2014; F. Zheng et al., 2018).

Deployments of ANPR cameras began in small numbers with limited performance, primarily for purposes of law enforcement, car parking and toll collection (Patel et al., 2013; Rhead et al., 2012). More recently, advances in video processing and Optical Character Recognition (OCR) have led to improvements in detection performance and lower operational costs, sparking broader adoption for traffic monitoring (Arth et al., 2006; Bramberger et al., 2004; Silva & Jung, 2018; Zhai et al., 2012). Consequently, ANPR sensors now play a major role in the ITS strategy and traffic management capabilities of many regional traffic authorities (Watson, 2017). Various cities, such as London (Chow et al., 2014; Haworth & Cheng, 2012) and Newcastle upon Tyne (Pinto da Silva et al., 2018) in the UK; Melborne in Australia (R. Li & Rose, 2011); Forteleza in Brasil (Barroso et al., 2020); and Changsha (F. Zheng et al., 2018) and Kunshan (W. Rao et al., 2018) in China, operate hundreds of ANPR cameras on a daily basis, as part of their traffic management systems.

In large numbers, ANPR cameras form complex networks whose increased interconnectivity and coverage unlocks new research opportunities and ITS applications. The potential of ANPR networks lies in the vast amounts of generated data, which add to a more complete and extensive record of user behaviour, route choice and network performance. In addition, ANPR data is generally owned and managed by local authorities, who do not need to rely on privately owned data and systems in order to innovate and implement their ITS program. Therefore, as the backbone of local traffic sensing infrastructure, ANPR networks can help realise the vision of data-driven ITS (Zhang et al., 2011).

Despite their potential, ANPR deployments can end up operating more as a disconnected collection of cameras than an interconnected network of sensors. Several challenges prevent a holistic treatment of ANPR networks. Computationally, ANPR data exhibits key characteristics of BigData – volume and velocity (Pinto da Silva et al., 2018; W. Rao et al., 2018). Due to the storage, computational and security challenges of large scale data collection, the analysis of ANPR networks is often limited to known components of the network (i.e. manually identified routes) and lacks a methodological framework for expressing and discovering interrelated sequences of travel. Furthermore, before it can be adequately analysed, ANPR data undergoes a series of pre-processing steps, which are poorly-documented as a whole despite having noticeable impact on data quality, e.g. the cleaning of matched number plate data (S. D. Clark et al., 2002). Due to these implementation and analysis challenges, few works have expanded the ITS capabilities of interconnected ANPR networks beyond travel time forecasting. Consequently, the siloed nature of ANPR research and application has prevented the technology from fully maturing and fulfilling its potential for widespread data-driven traffic management and

control.

1.1 Aims

The broader aim of this thesis is to foster participation in ANPR research and application, particularly of large-scale ANPR networks, by making the analysis of ANPR data more accessible and better understood. To achieve this, we tackle two types of challenges faced by researchers and industry practitioners alike:

1. Technical challenges faced when working or wanting to work with ANPR data, expressed through a lack of specialised techniques, documentation and open-source tools with which to process and analyse ANPR data.
2. Limited understanding of how the technology can be used at scale, based on a narrow set of use cases for ANPR networks, which can restrict the business case for investing in new or existing ANPR technology and data analysis systems.

To contribute towards solving each of these challenges, we structure this thesis in two parts. The first part focuses on overcoming technical barriers for wider ANPR adoption, and the second part develops two novel applications of ANPR networks, thus demonstrating the innovative aspect of large-scale ANPR networks for traffic management and control.

1.2 Part 1: Overcoming technical barriers to adoption

The first part of the thesis focuses on two of the aforementioned barriers to widespread ANPR adoption and research: pre-processing of ANPR data for analysis and limited representation models of vehicle movement.

The difficulty in pre-processing ANPR data is addressed by building a data processing pipeline capable of producing data ready for analysis at two different aggregation levels: individualised trip data and traffic stream data. The pipeline also serves to document the various datasets and processes within the system, and clarify their execution order. In addition to surveying existing techniques at each processing stage, we develop new methodology to solve the following tasks: the clustering of camera locations, the mapping of camera locations to map locations, the identification and treatment of outlier travel times and the identification of routes for analysis. Using the Tyne and Wear ANPR

network (UK) as a study case, the data pipeline is benchmarked, producing reference figures that researchers and practitioners can refer to.

The second adoption barrier is related to a methodological gap in the representation and treatment of vehicle movement in ANPR networks. Movement between two locations, origin and destination, is commonly represented via an origin-destination (OD) pair (Barroso et al., 2020). However, when vehicle trips are composed of more than two camera locations, they are characterised not by one pair but a sequence of OD pairs (and their corresponding travel times). The resulting sequences contain information about the intermediary steps of the journey, in addition to its start and end points, which reveal user route choices and augment data analysis. We develop a mathematical framework to formalise the representation and discovery of valid (in a spatial sense) and popular travel sequences, named corridors. Corridors are useful in that they highlight preferred travel paths in the road network and provide a methodological foundation for the development of new applications, namely bottleneck identification. Moreover, we use corridors automatically identified in the Tyne and Wear network, to examine route connectivity, that is, how influential a route is in distributing traffic across the network.

1.3 Part 2: Developing novel applications

The second part of the thesis develops two novel research applications of ANPR networks: (1) the identification and impact assessment of traffic bottlenecks and; (2) the modelling of overtaking behaviour across varying traffic conditions. While these are longstanding traffic congestion and safety problems, their investigation across an entire urban road network is made difficult by measurement limitations of traditional traffic sensors (Klein et al., 2006). Since ANPR proves advantageous outside highway settings, its application can be extended, at least in theory, to the investigation of said traffic phenomena. Our goal is not only to demonstrate the potential and novelty of ANPR networks for large scale analysis of urban road networks, but also to advance the understanding of traffic phenomena in diverse driving environments, especially arterial and collector roads.

Traffic bottlenecks are characterised by a performance differential in traffic conditions before and after their location (Daganzo, 1997) – a distinctive feature that is recognised through reliable measurements of travel performance taken along a corridor, such as those provided by ANPR. Being a leading source of congestion, the identification, assessment and ranking of bottlenecks is hence a crucial precursor to their mitigation (Hale et al., 2016). We perform this analysis for the Tyne and Wear road network, which allow us to empirically quantify the proportion of congestion that is attributed to bottlenecks, as opposed to other sources of congestion. Moreover, the assessment and ranking of bottleneck

impact provides regional authorities with a decision mechanism for the prioritisation of mitigation strategies.

Vehicle overtakings are known to have negative impacts on the performance and safety of traffic flow (Hegeman et al., 2004; Mattes, 2003; Z. Zheng et al., 2011). However, measurement is largely impractical through human observation or unavailable since many traffic sensors are incapable of detecting vehicle overtakings. Using ANPR, overtakings can be identified by direct comparison of vehicle detection times at origin and destination point (a later departure but earlier arrival, in respect to another vehicle). At the aggregate level, the relationship between overtaking rate (overtakings per unit of time) and traffic state, specified via a statistical regression model, can be indicative of road safety levels (Navon, 2003). By estimating the model across a range of ANPR routes, we are able to comment on the universality of the relationship and the effect of road characteristics, such as lane count and speed limit, on its intensity. Results suggest that overtaking rate, as a candidate proxy measure for road safety levels, is potentially useful in the identification of hazardous roads (roads where overtaking rate increases more significantly with changes in traffic conditions).

As a whole, this work advances the methodology of data analysis for ANPR networks. We believe that it can be particularly useful in informing new stakeholders about the analytical capabilities of ANPR networks, and supporting the activities of stakeholders who are already operating an ANPR network. A notable example is the application for regional transport improvements. The application requires stakeholders to attach relevant evidence, measured directly or inferred from traffic data, that is indicative of the congestion, efficiency or environmental issues targeted for improvement. ANPR networks can aid in this and other operational and planning activities by complementing the existing sensing infrastructure. Even though the Tyne and Wear ANPR network is used throughout as study case, we anticipate the proposed techniques to be equally applicable and useful in any region equipped with an ANPR network.

1.4 Thesis structure

The remainder of the thesis is split into six chapters – one background and four research chapters followed by a summary chapter, described as follows:

- Chapter 2 offers a brief overview of traffic stream characteristics and key sensing technologies, and outlines operational characteristics of ANPR systems, as well as common and breakthrough applications of ANPR networks.
- Chapter 3 describes the data pipeline for pre-processing and preparing data for analysis in ANPR networks.

- Chapter 4 defines ANPR corridors, describes an algorithm for corridor identification and conducts corridor and road connectivity analyses for the Tyne and Wear network.
- Chapter 5 specifies an algorithm for bottleneck detection, defines metrics for impact assessment and performs bottleneck and congestion analysis across the Tyne and Wear road network.
- Chapter 6 models the relationship between overtaking rate and traffic stream variables for multi-lane roads.
- Lastly, Chapter 7 concludes the thesis by summarising key results obtained at each of the previous Chapters and highlighting directions for future research.

Chapter 2

Background

This section offers a brief background of traffic engineering and operations, with an emphasis on Automatic Number Plate Recognition (ANPR) systems and their applications. Our goal is to elucidate the measurement capabilities and limitations of ANPR technology compared to other forms of data collection, as well as highlight current research trends. Section 2.1 introduces fundamental concepts from traffic flow theory, primarily for the benefit of the readers less familiar with the field, including main traffic flow problems, theoretical properties of traffic streams and types of measurement devices and their various sensing technologies. Section 2.2 describes Automatic Number Plate Recognition (ANPR) technology in detail and specifies several applications of sensor networks resulting from large scale sensor deployment.

2.1 Traffic flow: theory and measurement

Traffic flow theory is concerned with the study of traffic streams, their properties and the interactions between streams and their environment (infrastructure, people). Models of traffic flow provide a mathematical framework with which to measure and predict the characteristics and behaviour of traffic streams on a road network (a set of roads). Through measurement and mathematical modelling, traffic engineering and operations aim to exert control over traffic streams, for example directly by adapting the timings of a traffic signal or indirectly through design and policy measures, in order to create favourable travel conditions and/or improve adverse travel conditions (Daganzo, 1997).

The field of traffic theory and analysis tackles numerous problems, characterised by observation scale, spatio-temporal scope and method of analysis. Two main observation scales are generally considered: microscopic models, such as car-following models (Newell, 2002), describe the movements and characteristics of individual vehicles, whereas macroscopic models, namely kinematic wave models Daganzo (1995), are purely concerned with the

fluid-like properties and behaviour of traffic streams as a whole. Within these, the scope of traffic design and analysis can extend from a single intersection or road, to the entire network. Furthermore, flow models can relate to a single stationary regime, in particular free flowing or congested traffic, or multiple regimes, where traffic dynamics are used to describe how traffic transitions between states as driving conditions change (Treiber & Kesting, 2013).

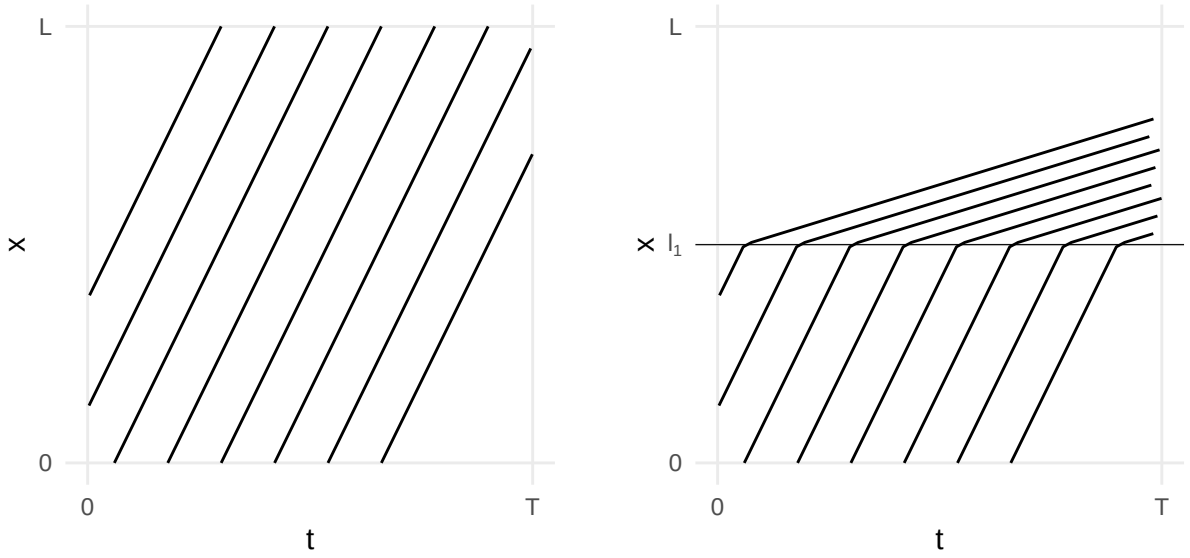
Analysis methods can be predominantly theoretical or empirical. Theory influences road and network design, and allows for simulation of driving behaviour otherwise too complicated for numerical treatment (Barceló & others, 2010). In contrast, recent advances in data collection, computer networking and data storage have led to a proliferation of data-driven methods in transport applications, especially in traffic forecasting (Zhang et al., 2011; L. Zhu et al., 2019). The main attraction of theory-free methods, such as machine learning, lies in the ability to predict very complex traffic dynamics with relatively high accuracy, without the need for strong theoretical support or assumptions. Furthermore, the volume, velocity and variety of available traffic-related data means that empirical methods can operate at a scale previously not possible.

Regardless of analysis scale, scope and methodology, the characterisation and measurement of traffic streams is the fundamental building block of traffic theory and analysis. For a better understanding of the measurement capabilities of ANPR devices compared to other sources of traffic data, it is thus of interest to briefly describe the main properties of traffic flows along with key sensing technologies.

2.1.1 Traffic stream characteristics

A traffic stream is fully characterised by the spatial trajectories of its composing vehicles and their evolution in time. Traffic streams are represented and examined in time-space diagrams – the fundamental graphical tool for transportation analysis. A time-space diagram provides a complete historical description of vehicle trajectories over a transportation link (a connection between two transportation nodes), with each trajectory curve specifying the distance travelled along the link as a function of time. Figure 2.1 illustrates two instances of a time-space diagram and the contrast between stationary traffic (independent of time and space) and non-stationary traffic.

Through the study of time-space diagrams, traffic flow theory emerged to describe traffic streams via measurable quantities without the need to observe complete vehicle trajectories. Under stationary driving conditions, the traffic stream along a link is succinctly characterised, among other variables, by its flow rate (vehicles per unit of time), average speed (distance per unit of time) and density (vehicles per unit of distance). The precise mathematical definition and interpretation of each traffic variable, however, is dependent



(a) Constant speed curves in stationary traffic (parallel and equidistant trajectories).

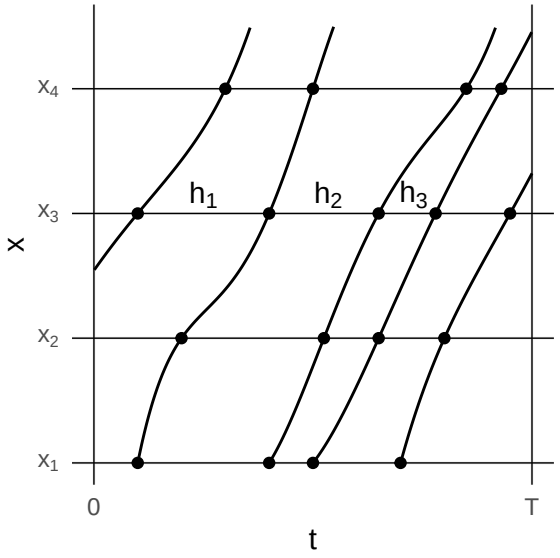
(b) Example of non-stationary traffic: speed decreases significantly after location l_1.

Figure 2.1: Eight vehicle trajectories in the time-space (t, x) plane. Each curve represents a separate vehicle trajectory across time (x axis) and space (y axis).

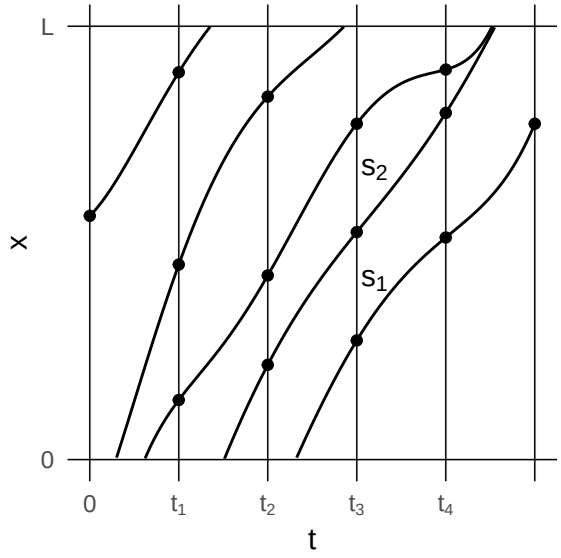
on measurement or estimation method.

The earliest measurement method was that of a stationary roadside observer. By counting the total number of vehicles n passing a location over a period of time T, a static observer or device calculates the traffic flow rate as $q = \frac{n}{T}$. Density, on the other hand, defined as the total number of vehicles m along a length of road L at a given time instant $k = \frac{m}{L}$, could not be measured directly by a static observer but only via an aerial photograph. Figure 2.2 depicts the differences between these two methods of measurement – a series of point detectors placed along a road (stationary observer) or aerial photographs taken over time.

A third main characteristic of a traffic stream is the average speed of its composing vehicles. For any given vehicle, speed is calculated as the ratio of its travel time (the time difference between consecutive vehicle observations over a length of road) and the distance travelled. The speed of a traffic stream is referred to as "time mean speed" (or point speeds) when obtained over infinitesimal/short distances, or as "space mean speed" when calculated over the entire length of road. Measurement of speed over a length of road requires tracking of individual vehicles, e.g. via a moving observer or multiple fixed observers with vehicle re-identification capabilities, whereas point speeds are measured using a detector capable of detecting vehicle passing over a specific point on the road. The use of the space mean speed is traditionally preferred as it incorporates the Lighthill & Whitham (1955) classical steady-state traffic flow model, which relates flow (q), (space) mean speed ($\hat{u}_s$) and density (k) via the relationship



(a) Vehicles are detected by sensors placed at locations $x_1 \dots x_4$. The flow rate at a given location in the interval $(0, T]$ is calculated by counting the number of vehicles observed at that location. The time between consecutive vehicle observations, called headway (h), approximates T as T tends to infinity.



(b) Vehicle detected in photographs taken at times $t_1 \dots t_4$, over a length of road L. Traffic density is calculated by counting the number of vehicles in each photograph. The distance between consecutive vehicles, called spacing (s), approximates L as L tends to infinity.

Figure 2.2: Two ways of obtaining traffic data: through (a) a series of stationary observers and (b) an aerial photographs (Daganzo, 1997).

$$q = \hat{u}_s \cdot k. \tag{2.1}$$

The estimation of space mean speeds from point detectors can be obtained using the method of Rakha & Zhang (2005), with an error rate below 1%, based on the relationship between the time and space mean speeds first derived by Wardrop (1952).

Table 2.1 summarises key variables of interest used to describe traffic streams and the movements of its composing vehicles. Variables are defined in relation to three methods of measurement[1]: at a point and over a length of time (stationary observer), at a time instance over a length of road (aerial photograph) and over a period of time and length of road. Where variables can not be measured directly (e.g. density at a point), the fundamental relationship (Equation 2.1) is often used for estimation. For generalised definitions of flow, density and speed, for any measurable subregion of time-space, see Daganzo (1997).

Note that the variables presented here only describe (on average) a traffic stream for a

[1]Traffic measurement taxonomies may vary slightly. For example, Hall (1996) identifies five measurement procedures: at a point on a road, over a short road section (10 m), over a length of road (a route longer than 0.5 km), via a moving observer and over a wide-area. Antoniou et al. (2011) categorised road sensing technologies into three groups based on measurement capability: point, point-to-point (at a length by vehicle tracking) and wide-area network (aerial photograph).

Table 2.1: Key vehicle and traffic stream variables.

Variable	Symbol	Definition	Measurement[1]		
			Point (interval)	Length (instant)	Length (interval)
Single vehicle					
Travel time	tt_i	Difference between the time at destination t_2 and origin locations t_1			$t_2 - t_1$
Average speed	u_i	Average travel speed of a vehicle over a length of road L			$\frac{L}{tt_i}$
Instantaneous speed	$\hat{u}_i$	Travel speed of a vehicle over a short distance and period of time	$\frac{dx}{dt}$	$\frac{dx}{dt}$	
Headway	h_i	Time interval between the passage of vehicle $i-1$ and vehicle i	$t_i - t_{i-1}$		
Spacing	s_i	Distance between vehicle $i-1$ and vehicle i		$x_i - x_{i-1}$	
Traffic stream					
Flow rate[2]	q	Number of vehicles N observed over a period of time T	$\frac{N}{T}$	*	*
Concentration					
Density	k	Number of vehicles N occupying a length of road L	*	$\frac{N}{L}$	*
Occupancy	ρ	Fraction of time a point in the road is occupied by vehicles	†		
Mean Speed					
Time mean speed	$\bar{u}_t$	Arithmetic mean of the speeds of vehicles passing a point on the road	$\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i$		
Space mean speed	$\bar{u}_s$	Arithmetic mean of the speeds of vehicles taken over a length of road	*	*	$\frac{1}{N}\sum_{i=1}^{N}u_i$
Mean travel time	$\bar{tt}$	Average travel time of vehicles between two locations			$\frac{1}{N}\sum_{i=1}^{N}tt_i$

[1] Three common methods of traffic measurement: at a point on the road and over a time interval (stationary observer), instantaneously over a length of road (aerial photograph) and over a length of road and time interval (moving observer, or two fixed observers capable of vehicle re-identification).
[2] Also sometimes referred to as (traffic) volume.
* Typically estimated using the fundamental relationship: $q = \bar{u}_s \cdot k$.
† Measured using a presence detector. Occupancy is related to density through the relationship $\rho = k\bar{l}$, where $\bar{l}$ is the average vehicle length.

single "slice" of time (stationary traffic) and space (transportation link). Traffic conditions evolve considerably over time and space as a function of road characteristics and traffic supply (existing road infrastructure) and demand (user trips), among other factors (Treiber & Kesting, 2013). Consequently, the temporal and spatial evolution of a traffic stream beyond a steady-state region generally translates into complex changes to their characteristics, which traffic flow models attempt to describe through mathematical relationships between stream variables. Figure 2.3 shows an example from Treiber & Helbing (2002) of traffic dynamics on a highway marked by periods of heavy traffic congestion (low speeds) that propagates backwards in space forming a congestion shockwave (red strips).

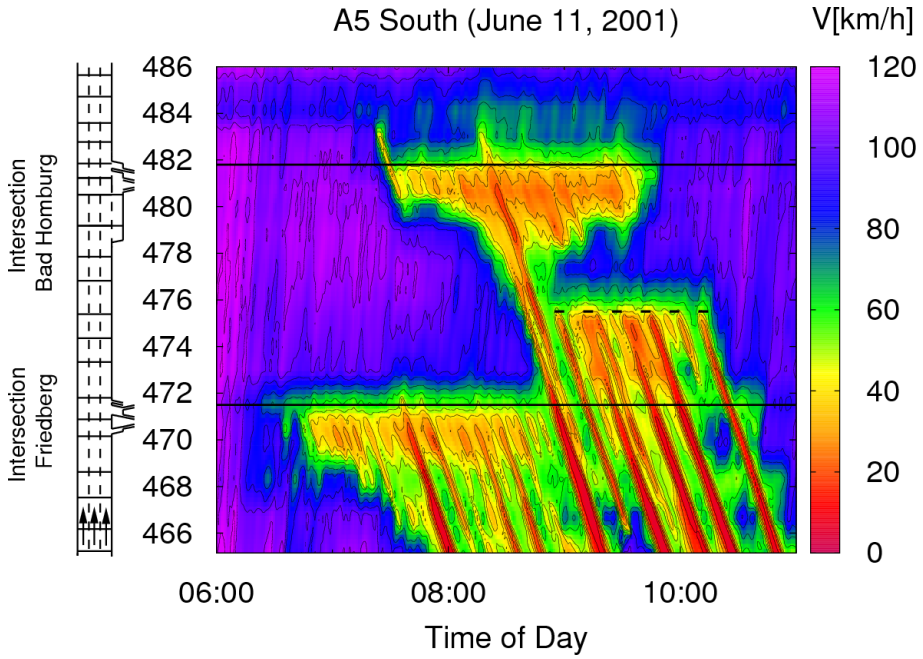


Figure 2.3: Image from Treiber & Helbing (2002) depicting the spatio-temporal complexity of vehicle dynamics on a highway with two merge points. Each merge point causes traffic to breakdown (transition from free flowing to congested conditions) and the formation of a shockwave that travels opposite to the direction of travel (seen by the red strips).

2.1.2 Traffic sensing technologies

The collection of road data has evolved considerably since the first photography-based measurement procedure employed by Greenshields et al. (1935). Currently, road sensing technologies can be divided into four categories[2] by measurement functionality: point (at a point through a fixed sensor), point-to-point (along a length using two fixed sensors), floating car (along a length using a position tracking sensor) and area-wide (along a length via an aerial observer).

[2]Our categorisation of traffic sensing technologies is adapted from Antoniou et al. (2011) with one key difference: floating car data is defined separately to point-to-point data to reflect the moving versus fixed nature of measurement method.

2.1.2.1 Point sensors

Point sensors collect traffic data through a sensor placed at a point on a road or multiple sensors placed over a short road section[3]. Point sensors can be installed below ground, such as inductive loop detectors or magnetic sensors, or mounted on structures over the roadway with a clear view of traffic (spanning one or multiple lanes depending on technology and monitoring goal). A comprehensive list of point sensor technologies, as well as their characteristics and capabilities, can be found in Klein et al. (2006). Primarily, this category includes the following sensing technologies:

- **Loop detectors** are installed below the road surface and detect passing or stopped vehicles using an electrical induction loop. They are the most widely used traffic sensing technology to date. The technology is mature, well understood and capable of supporting a wide range of applications. The primary stream characteristics measured by loop detectors are flow, occupancy and point speeds, which can be estimated either from dual or single detector placement.
- **Magnetic sensors** are intrusive traffic sensors that detect perturbations in Earth's magnetic field produced by a moving ferrous metal vehicle. Magnetic sensors are used in settings where loop detectors can not be adequately employed (e.g. bridge decks), but they are limited to detecting moving vehicles and can not recognise stopped vehicles (can not provide occupancy data).
- **Radar (microwave/infra-red), ultrasonic and acoustic sensors** are non-intrusive point-sensor technologies installed over-roadway. This class of sensors is a cost-competitive alternative to loop detectors and generally capable of multiple lane operation and direct measurements of speed. However, their performance can be affected by adverse weather conditions, such as the presence of fog, rain or snow.
- **CCTV cameras** are video cameras without number plate recognition capabilities, mounted on tall poles over the roadway, which provide television imagery for human analysis and interpretation. They employ image processing techniques to inspect the scene of interest, detect accidents, classify vehicles and retrieve some traffic data like volume.

2.1.2.2 Point-to-point sensors

Point-to-point data characterise traffic streams along a road segment or route (sequence of road segments). Point-to-point measurement is achieved by positioning two sensors with vehicle identification capabalities at the start and end of a road section. When identified

[3]Some authors, for example Hall (1996), distinguish between point and short distance measurements. However, for simplification, these two classes are here considered jointly since measurement is along an approximately infinitesimal length of road.

by both sensors, a vehicle's travel time is calculated as the difference between the times recorded at the end and start locations (Antoniou et al., 2011). Furthermore, if vehicle trajectories are known or estimated, the average speed of each vehicle, and by extension the space mean speed of the traffic stream, are computed by dividing the travelled distance by the observed travel time (as shown in Table 2.1).

The collection of point-to-point data is made possible by Automatic Vehicle Identification (AVI) technology. Two main classes of AVI technology have emerged: radio-frequency (RF) systems, whereby a transmitter device located inside the vehicle communicates with the roadside sensor (responder device) (Kamarulazizi & Ismail, 2010); and Automatic Number Plate Recognition (ANPR) systems, whereby vehicle tags are read directly via image/video analysis and optical character recognition (OCR) algorithms (Ozbay & Ercelebi, 2005).

Several technologies have enabled RF-based AVI systems. RFID (radio-frequency identification), used primarily for electronic toll collection in highways, requires a special transponder, typically mounted on the windshield of a vehicle. Recently, Bluetooth and WiFi technologies, less costly and ubiquitous in smartphones, offer similar AVI capabilities via unique MAC (Media Access Control) addresses (Day et al., 2012; Park et al., 2009). A downside of RF-based technologies is that they are employed only by a fraction of the vehicle population and therefore can not reliably measure traffic volume but only speed instead.

In contrast to RF-based AVI, which rely on communication between two devices for vehicle identification, ANPR systems only require a visible number plate. However, ANPR technology is susceptible to detection errors, influenced by, among other factors, weather and lighting conditions. ANPR systems are further reviewed in Section 2.2.

2.1.2.3 Floating car data

Floating car (FC) data is obtained from a moving vehicle equipped with a location-aware device. There are two prominent sources of location data: satellite positioning systems, which interface with GPS (Global Positioning System) devices, and mobile phone carriers, which achieve cellphone tracking through continuous triangulation of mobile phone signals (Jiang et al., 2013).

FC data acts as fine-grained point-to-point data: location readings are geo-referenced and timestamped. Even at lower sample rates, FC data can retrieve detailed trip information, including trajectory and link-specific travel times (Antoniou et al., 2011). However, because data is collected from a moving rather than fixed sensor, locations are not entirely precise and need to be map matched, that is, an algorithm tries to find the correct road segment that the vehicle is travelling on (M. A. Quddus et al., 2007).

For real-time traffic analysis, each sensor relays information back to a centralised server through an active mobile Internet connection. Google Maps (*Google Maps Platform*, n.d.), Here Maps (*HERE Traffic API V7*, n.d.) and TomTom (*TomTom Traffic*, n.d.) are examples of map services that offer real-time traffic data via an application programmatic interface (API). Users of these platforms provide FC data in exchange for navigation services, for instance GPS-enabled mobile phones provide Google with location data at scale. However, the methodology that underlies such products is proprietary and is not easily replicated by local traffic authorities, who do not have access to the same crowd-sourced data. To complement their static traffic sensor assets, local traffic authorities can sometimes obtain FC data from taxi and bus vehicle populations, though at the cost of increased sampling bias towards these populations (Q. Cao et al., 2020).

2.1.2.4 Area-wide sensors

Area-wide sensors capture traffic conditions across an entire region through aerial photography/video recording (Hayat et al., Fourthquarter 2016) or laser scanning techniques like LIDAR (Light Detection and Ranging) (Reutebuch et al., 2005). Data produced by airborne sensors can be used for surveillance purposes but its unstructured nature (vehicles are not geo-referenced) and volumous quantities make it challenging to process and use for real-time traffic monitoring and control. They contrast with AVI systems and FC data, which can provide link-specific area-wide traffic information when employed at scale. In addition, airborne sensors can be costly to operate, even with unmanned technology.

2.1.2.5 Summary of sensing capabilities

A summary of sensor measurement capabilities is found in Table 2.2. The contrast between point and point-to-point sensors is made clear by the type of traffic measurements available to each. Point sensors can characterise traffic streams along a continuous road link, or intersection of links, under known conditions (Daganzo, 1997; Klein et al., 2006). For a succession of road links in a complex urban network, point sensors fail to measure traffic flow and speed because they are incapable of following individual vehicles or recording vehicle turns (Antoniou et al., 2011). The choice to deploy point or point-to-point sensors is thus application and context dependent, as well as being subject to cost, installation and environmental constraints.

In point-to-point sensor technologies, the advantage of ANPR over RF-based AVI technologies is in its capacity to reach the whole vehicle population instead of just the segment of the population equipped with transponder devices. Although a larger fraction of the population is targeted when using widespread RF technologies in mobile devices like

Table 2.2: Comparison of sensor measurement capabilities.

		Point/Short-length			Point-to-point		Floating car		Area-wide
	Measurement capability	Loop detectors	Radar/infra-red/acoustic	CCTV cameras	RF-AVI	ANPR	GPS	Cellular	Airborne sensor
Point	Flow	✓	✓	✓					
	Speed	✓	✓	✓					
	Occupancy	✓		✓					
Length	Flow (OD)				✓[a]	✓	✓[b]	✓[b]	✓[e]
	Density				✓[a]	✓	✓[b]	✓[b]	✓[e]
	Travel time[1]				✓[a]	✓	✓[b]	✓[b]	✓[e]
	Trajectory				✓[c]	✓[c]	✓	✓	✓[e]
	Vehicle class	✓	✓	✓	✓	✓[d]			✓[e]

(1) Average speed can be calculated from travel time and distance travelled (trajectory).
(a) Only for the subpopulation vehicles equipped with transponder devices.
(b) Only for the sample of vehicles being tracked.
(c) Can be manually annotated by experts, calculated based on assumptions (e.g. shortest path) or inferred from data (typically in denser sensor networks).
(d) Can be obtained from vehicle registry (by merging on number plate).
(e) Not measured directly (requires tracking of individual vehicles in raw image/video data).

bluetooth and WiFi, detection still relies on users having their devices' radio interfaces on during travel as they come into contact with roadside receiver AVI sensors.

Compared to floating car data, ANPR has one main disadvantage: the inability to precisely track the trajectory of vehicles. This limitation is weaker in denser sensor networks, particularly if trajectory inference follows from network design (Antoniou et al., 2011; Klein et al., 2006). In exchange, ANPR offers three desirable properties (Q. Cao et al., 2020): uninterrupted operation (subject to maintenance and sensor failure), known/fixed location of observations (whereas floating car data are subject to tracking errors) and increased sample size.

Lastly, airborne sensors offer a "birds-eye" view over an entire region, but unlike point-to-point and floating car data, vehicle positions are not geo-referenced. Hence, the measurement of travel time and vehicle trajectories requires vehicles to be tracked over a sequence of photographs or video frames. Despite their potential to track the whole road network at once, the operation of airborne sensors is costly and therefore generally impractical for day-to-day traffic monitoring and control operations.

2.1.3 Common traffic flow and engineering problems

We identify three broad categories of traffic flow problems of interest to traffic scientists, engineers and practitioners:

1. Traffic monitoring and prediction – Identifying and forecasting traffic conditions in real-time.
2. Traffic control and optimisation – Exercising control over traffic streams.
3. Transportation planning and scheduling – Preventing and mitigating traffic flow problems through network design.

Traffic monitoring and prediction is concerned with measuring and understanding the state of traffic flow in the road network (Davis, 1997). Informally, the state of traffic flow can be described as free-flowing, when vehicles can travel at or close to the speed limit, or congested, when travel speed is significantly restricted to below the speed limit. One of the main purposes of monitoring is to identify areas of the road network affected by traffic congestion and follow its evolution across space and time (Skabardonis et al., 2003). In addition, traffic monitoring may entail forecasting the state of traffic from current and past observations (E. I. Vlahogianni et al., 2014), estimating traffic flow in unobserved parts of the road network (Carrese et al., 2017), measuring the impact of traffic congestion on a highway or along an urban road corridor (Chen et al., 2004; Wolniak & Mahapatra, 2014), among other goals.

The difficulty with traffic monitoring and prediction lies in the finite measurement capabilities of centralised traffic operations (Elkosantini & Darmoul, 2013) and the complex spatio-temporal dynamics exhibited by traffic networks (Treiber & Kesting, 2013). Whereas traffic flow theory emerged to model and understand traffic phenomena (typically confined in space and time), newer approaches to traffic monitoring focus increasingly on predictive rather than explanatory tasks (L. Zhu et al., 2019). The innovation in approach is enabled by technological advancements in traffic sensing devices (Antoniou et al., 2011) and theory-free computational methods, which allow authors to capture long-distance interactions between traffic flows in large and dense traffic networks (S. Wang et al., 2022).

Traffic control and optimisation seeks to influence traffic flow so as to maintain free-flowing traffic or recover from a state of congestion. Traffic control is traditionally framed in the context of two or more intersecting traffic streams, whose interactions must be regulated by a control device, such as a traffic signal, sign or a road marking (Daganzo, 1997). In the case of a traffic signal, a signalling plan is designed such that vehicle throughput is maximised, while taking into account the technical, physical and operational constraints of the site. Traffic signal control systems such as SCOOT and

SCATS extend this functionality by adapting their signal plan to the traffic conditions observed on the links approaching an intersection (Stevanovic et al., 2009).

The (adaptive) traffic control problem is thus broadly defined any set of control devices (actuators) whose function can be adjusted (automatically or manually) in response to changing traffic conditions, namely to increased user demand or traffic congestion (Papageorgiou et al., 2003). Other common types of traffic control applications include route guidance and driver information systems (such as variable message signs) and highway traffic control strategies like ramp metering (traffic lights at on-ramps or highway interchanges). To allow for adaptive traffic control, actuators receive input information about the current (and potentially future) state of traffic flow from traffic monitoring devices and traffic prediction methods.

Transportation planning and scheduling deals with network design problems. These relate to the distribution of supply (i.e. road infrastructure) and demand (i.e. user trips) on the network, and the deployment of transport services and resources, such as the scheduling of bus routes (Meyer & others, 2016). The primary goal of traffic planning is to determine the operational capabilities and limitations of the road infrastructure – whether it can meet user demand and how – and act as a decision support mechanism for the improvement and expansion of the transport network.

To solve this class of problems, practitioners must understand the needs of transport users and their behaviour across the network, i.e. how they plan and execute trips (McNally, 2007). Compared to traffic monitoring, whose goal is to track and anticipate the evolution of traffic conditions on the road network (so that we can exercise control), this class of problems looks at aggregate trip and flow patterns to understand the broader network usage trends and aid decision-making.

2.2 Automatic number plate recognition

The promise of Automatic Vehicle Identification (AVI) technology for ITS applications was identified early on (Hauslen, 1977). Radio-frequency identification (RFID) technology was first developed for road access control (paid highways) and electronic toll collection (ETC), as a mechanism to automate payment and avoid traffic stops at toll collection points (Blythe, 1999). With increased adoption among private vehicles, it became possible to measure vehicle travel time from time-stamped RFID readings recorded at entry and exit toll stations. The added ability to collect traffic data led to the development of multiple highway traffic monitoring AVI systems, for example the TranStar (TranStar, 2001), TransGuide (SWri, 1998) and Transmit (Mouskos et al., 1998) systems in U.S.

Despite their success, earlier RFID-based AVI systems were limited to highway operation

and required users to opt-in. Outside control access roads, ANPR has proved advantageous since no transmitter device is required for vehicle identification, only a visible license plate (Ozbay & Ercelebi, 2005). Furthermore, ANPR does not face performance issues in highly saturated RF mediums, allowing for its deployment in denser population areas (Omar et al., 2016). Due to a simultaneous decrease in hardware costs and improvements in OCR performance, ANPR has since been increasingly used for traffic monitoring, beyond its original law enforcement purposes (Debnath et al., 2014). In order to better undestand the capabilities and limitations of ANPR systems, we describe their technological and operational characteristics and then outline several applications of the technology at scale.

2.2.1 Operational characteristics

An ANPR system is a video camera whose frames (image captures) serve as input to a number plate recognition algorithm. ANPR cameras are categorised as static, if mounted permanently high on a gantry/bridge or roadside pole, or mobile if handheld or mounted on the windshield of a patrol vehicle. In the UK, a video camera is considered ANPR-compliant if it meets the detection standards set by (Home Office, 2020). This allows cameras not built specifically for number plate recognition, namely CCTV cameras, to be repurposed for ANPR.

Operationally, an ANPR camera has a capture zone, defined as a subregion within its field of view, wherein vehicle number plates are detected. Both of these parameters depend on manufacturer and camera model. For example, Gurney et al. (2013) reports a detection zone between 18 and 23 meters (m), for a static camera with a maximum operating range of 32 m. In addition to correct camera installation, the authors recommend proper calibration of camera settings, like camera type (color, black and white or infrared), resolution, shutter speed and orientation, which can otherwise have adverse effects on system performance.

A number plate recognition (NPR) algorithm is the component which transforms an otherwise “simple” video camera into an ANPR-capable system. A NPR algorithm is generally divided into four main stages: (1) identification of vehicle, (2) extraction of number plate region, (3) character segmentation and (4) character recognition (Du et al., 2013; Ozbay & Ercelebi, 2005; Patel et al., 2013). The first and second stages seek to identify the vehicle and plate regions according to features like boundary, color and character presence (the first stage can be skipped if the camera is handheld or the vehicle encompasses most of the input image). The third stage separates each of the license plate characters through techniques such as color segmentation or position template matching. The fourth and final stage uses a pre-trained machine-learning (ML) classifier, commonly

a neural network, to identify each input character.

A number of approaches have been developed for each stage. Comprehensive surveys are given by Du et al. (2013) and Patel et al. (2013), and more recent advances using deep learning for unconstrained capture scenarios are given, for example, by Silva & Jung (2018) and Weihong & Jiaoyang (2020). Due to plate variations (size, color, font, occlusion, angle, screws, customisations) and environmental factors (illumination, weather, image background), each stage is subject to a binary outcome (success or failure). A failed detection can result in a false positive (wrong number plate), known as type I error, a false negative (failure to identify a passing vehicle), known as type II error, or both. Figure 2.4 compiles several input images from various open datasets, illustrative of the challenges in number plate recognition. The manifestation and treatment of type I and type II errors in ANPR data is explored in Sections 3.3.5 and 3.4.1 of Chapter 3 (ANPR data pipeline).

(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2.4: Sample images from different number plate recognition datasets: (a-d) AOLP dataset (Hsu et al., 2013) (e-f) MediaLab dataset (Anagnostopoulos et al., 2008), (g-h) CD-HARD test dataset (Silva & Jung, 2018). (a) access control category – close range photos taken as the vehicle passes through a gateway at reduced velocity or as it comes to a full stop, (b) road patrol category – taken using handheld cameras at arbitrary angles and distances, (c-d) traffic law enforcement category – photos of vehicles at regular or high speed taken by a roadside camera (c - multiple vehicles, d - at nighttime), (e) instance of blurred image, (f) license plate covered by shadows, (g-h) sharp angle/oblique images.

The system's performance can be measured end-to-end or as the product of each stage's detection rate (Ozbay & Ercelebi, 2005). Rhead et al. (2012) attributed 95.6% of misreads to the location of screw caps (75.6%) and other physical marks on the number plate (23%). In Du et al. (2013), only 11 out of 28 reviewed algorithms report on total performance rate, 9 of which show performance rates above 90% (processing times ranged from tens of milliseconds to a few seconds).

Recently, the development of open NPR datasets such as MediaLab (Hsu et al., 2013), AOLP (Hsu et al., 2013), OpenALPR (*OpenALPR Dataset*, 2014) and CCPD (Z. Xu et al., 2018), have allowed the standardisation of algorithm benchmarking, as well as testing across different number plate types (nationalities) and detection difficulty levels (e.g. variable shooting angles, clear vs skewed and blurred number plate, uniform vs uneven illumination, single vs multiple vehicles per image, close vs distant range). For example, Silva & Jung (2018) achieves performance rates of 93.53% on OpenALPR (EU), 91.28% on OpenALPR (BR), 98.36% on AOLP (road patrol subset) and 75% on their proposed `CP-HARD` dataset, composed mostly of oblique images from various regions and at various vehicle distances. Weihong & Jiaoyang (2020) compares several deep-learning NPR algorithms with reported performance rates generally above 90% across different datasets.

2.2.2 Applications of ANPR networks

In pairs, ANPR sensors measure vehicle travel times and other traffic stream characteristics along a route. At scale, many interconnected ANPR sensors form a network whose vehicle re-identification and data analysis capabilities extend from a series of roads to a whole region or city[4]. As ANPR networks can collect aggregate and disaggregate traffic data at scale, both within short and long time frames, they have been used to solve a range of traffic problems. This section briefly describes four primary research applications of ANPR networks. In each case, we summarise the problem, give examples of how the technology has been used to solve it and its advantages and disadvantages. Our goal is not to provide an extensive list of real ANPR systems and their applications, but to present key research areas where the technology has been demonstratively and successfully applied, with added value in large scale deployments.

2.2.2.1 Short-term forecasting

Short-term traffic forecasting is an important component of ITS systems and an active area of research. Short-term forecasts are required for many real-time ITS applications, for example, to detect and respond to traffic incidents, operate signalised control infrastructure and convey travel time information to drivers via variable-message signs (MVS) or mobile applications (E. I. Vlahogianni et al., 2014). Forecasts can range from seconds

[4]The concept of ANPR sensor networks can be extended to include sensor nodes of any AVI-capable technology. In practice, however, since each technology generates its set of unique vehicle identifiers (id), a heterogeneous AVI network (composed of different technology sensors) requires a mechanism by which multiple vehicle ids are combined into a single unique id. Thus, most AVI networks tend to be composed of a single sensor technology.

(very short forecasting) to a few hours into the future. To make predictions, forecasting models utilise current and past traffic data and, when available, information from nearby traffic sensors and other exogenous factors like weather conditions and special events (Lana et al., 2018). Even though forecasts can be made for different stream characteristics, the emphasis is generally on predicting traffic volume and/or travel time (E. I. Vlahogianni et al., 2014).

Due to their measurement reliability, short-term traffic forecasting is possibly the most common application of ANPR and AVI systems in the domain of traffic monitoring (E. I. Vlahogianni et al., 2014). Many research studies have developed prediction models for AVI/ANPR systems; see for instance R. Li & Rose (2011), Tam & Lam (2011), Kazagli et al. (2013) and E. I. Vlahogianni et al. (2014). Other studies have focused on specific congestion or efficiency problems that require short-term forecasts for prompt response or automated control, namely automatic incident detection (Hellinga & Knapp, 2000; X. Li et al., 2013), estimation of queue length at signalised intersections (Luo et al., 2019) and congestion analysis (Chow et al., 2014).

Together, E. I. Vlahogianni et al. (2014) and Lana et al. (2018) enumerate key challenges and research directions in short-term traffic forecasting. The authors point to the ability to capture the spatio-temporal characteristics of traffic flow as one important feature of forecasting models. At higher sensor densities, spatio-temporal information is incorporated by considering predictions from nearby road segments, whose degree of correlation is typically established based on sensor distance (Cheng et al., 2012). Since ANPR keeps a record of previous road segments in a vehicle's journey, the association between upstream and downstream traffic can be inferred directly from observed vehicle movements, instead of purely from sensor distance. This feature adds to the predictive power of ANPR networks in urban contexts, and is exemplified in Figure 2.5.

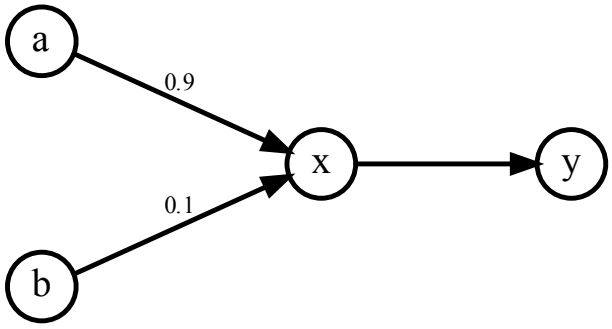


Figure 2.5: The origin of traffic on route `(x,y)` is quantified by virtue of the trip tracking capabilities of ANPR networks: 90% of vehicles arrive from segment `(a,x)` while the remaining 10% arrive from segment `(b,x)`. It provides a basis with which to measure correlation among segments (and monitor their value over time).

2.2.2.2 Estimation of demand matrices

An origin-destination (OD) matrix is a numerical representation of traffic demand and user trip patterns (Barroso et al., 2020). Each cell specifies the aggregate trip count or frequency of a corresponding OD pair (typically representing traffic demand between zones of a larger region). OD matrices serve as input to traffic assignment models, responsible for distributing user trips across the road network to produce traffic volume estimates at the link level. In contrast to short-term traffic forecasting, the resulting estimates, often split into daily and seasonal components, are interpreted as being representative of traffic patterns at equilibrium conditions (Krishnakumari et al., 2020).

The estimation of an OD matrix (or equivalently, OD flows) is the objective of the first two steps (or three steps if modal choice is considered) in the classic four-step model (FSM) for transportation forecasting (McNally, 2007). OD matrices are estimated by matching origin and destination points using trip information obtained, historically, from household surveys. With technological ITS advances, estimation has evolved not only to consider multiple sources of traffic data, but also to the time-varying component of traffic demand, referred to as dynamic flows (Carrese et al., 2017; Krishnakumari et al., 2020).

ANPR systems have been used by several authors to estimate OD matrices and investigate trip patterns in urban environments; see for instance (Barroso et al., 2020; W. Rao et al., 2018; X. Zhou & Mahmassani, 2006). Several properties of ANPR networks are compatible with OD estimation. First, trip OD pairs[5] are easily retrieved following trip identification (the first and last observations of a vehicle trip, respectively). Second, trips composed of more than two observations offer additional information about the route choice of vehicles, by capturing intermediary points along the route, which can better inform traffic assignment methods.

Despite ease of measurement, ANPR networks do not observe the true origin and destination of vehicle trips. In flow estimation, this issue is circumvented by assuming that the true origin/destination points are located near the first/last observed traffic sensors. For example, when estimating dynamic OD flows from ANPR data, Barroso et al. (2020) only considers six major city regions (in Fortaleza, Brazil), wherein camera locations lie. The study suggests that ANPR estimates provide sensible approximations of traffic demand, even though the true OD pairs may remain unknown.

Overall, the utility of ANPR networks for trip OD estimation will greatly depend on camera coverage. Estimates can easily become biased towards regions where camera placement is denser, compared to areas where placement is sparser. Moreover, these may be limited in their ability to provide additional information for traffic assignment models

[5]In this context an OD pair refers to the actual origin and destination of the trip. A trip composed of n observations has an OD pair specified by the first and n-th observations.

if the majority of trips contains no intermediary points (trips of length two or lower). Naturally, these technological limitations can be partly addressed by improving camera coverage and/or combining multiple sources of traffic data, albeit at an added cost.

2.2.2.3 Analysis of traveller behaviour

As previously discussed, OD matrices are used to analyse trip patterns at the population level. While population-level analysis is sufficient for a range of transport applications, there are benefits to examining the behaviour of individual travellers across time (Jones & Clarke, 1988). Traffic demand models can be made more accurate and realistic by incorporating user adaptation and day-to-day variability in route choice. In addition, new policies can be developed and old policies revised, as a consequence of an improved understanding of traveller differences, habits and flexibility in trip making (Cherchi et al., 2017; Chorus et al., 2006; Minnen et al., 2015).

Repeated driver behaviour has been traditionally difficult to measure at large scale. Travel diaries were primarily used to collect research data, but at a high cost and with limited scalability (Crawford et al., 2018). Recently, emerging data sources, namely mobile phone (Järv et al., 2014), smart card (Kieu et al., 2015) and AVI data (Zhao et al., 2019), have been used to collect individual trip data. Individual behaviour is described across a number of dimensions, such as trip frequency (Tarigan & Kitamura, 2009), temporal variability (Chikaraishi et al., 2009), spatial variability (Järv et al., 2014) and mode choice (Cherchi & Cirillo, 2014). ANPR networks, through vehicle re-identification, are particularly suited to capturing the trip patterns of road users across long periods of time. For example, Crawford et al. (2018) identified vehicles using a combination of Bluetooth devices and ANPR cameras and later clustered users in terms of their frequency and diversity of trips. Similarly, Zhao et al. (2019) identified different commuting profiles of private cars, which caused significant variation in day of the week traffic demand, using a network of AVI sensors, in Wuhan, China.

In Pinto da Silva et al. (2018), we used the k-means algorithm to cluster vehicles based on several features, related to trip frequency and their temporal and spatial variability, obtained from ANPR trip data. The resulting vehicle classes were indicative of distinct vehicle types, e.g. bus, lorry, car, and trip-making profiles, e.g. taxi vs privately-owned vehicle. Although we were unable to adequately validate the results (as we did not know the true vehicle classes), future studies can achieve this by merging ANPR data with DVLA[6] information (on key `number plate`), prior to data sharing and anonymisation in order to ensure user privacy. This would add to the richness of ANPR data, particularly for applications where vehicle classification is crucial in model calibration, for instance

[6]The Driver and Vehicle Licensing Agency (DVLA) is an executive governmental agency in the UK, responsible for maintaining the registration and licensing of drivers and vehicle in Great Britain.

in traffic pollution (Berkowicz et al., 2006) or noise pollution models (Subramani et al., 2012).

2.2.2.4 Route choice analysis

Route choice models specify how users plan and execute trips. Not only do they contribute to a better understanding of traveller behaviour (Prato, 2009), but are also an integral component of traffic assignment models (Chiu et al., 2011), route recommendation systems (Su et al., 2014) and dynamic route guidance systems (Liang & Wakahara, 2014).

Despite their widespread use in ITS solutions, route choice models can vary greatly in complexity. The user decision process is made complex by the numerous routing possibilities and the possibility of adaptive behaviour (Prato, 2009). Generally, two main adaptation mechanisms are considered: the day-to-day adjustment resulting from accumulated user experience, and the within-day adaptation in response to current traffic conditions, revealed *en route* by direct observation or indirectly via, for instance, a radio broadcast or route recommendation service (Gao et al., 2010). Adaptive behaviour is particularly relevant to ITS solutions aimed at preventing/mitigating congestion by diverting transit or encouraging users to choose alternative routes, such as in dynamic route guidance systems (Liang & Wakahara, 2014).

As described in the two previous applications, ANPR data can provide details about the route choices of individual users. Even though any trajectories inferred from ANPR data are undoubtedly less fine-grained than their GPS equivalents (which is the norm for trajectory data), ANPR is comparatively more available to local traffic authorities, both in terms of sample size and continuous sensor operation (Q. Cao et al., 2020; Y. Zheng, 2015). In addition to illustrating how route choice can be inferred from a bimodal travel time distribution, shown in Figure 2.6, Q. Cao et al. (2020) develops a semi-supervised route prediction model for sparse AVI data – a key limiting factor in the use of ANPR networks for route choice analysis. Hence, with increased camera density, it is expected that ANPR networks are more successfully employed in route choice analysis.

2.2.2.5 Summary of identified research gaps

Short-term traffic forecasting is the primary application of ANPR technology. Yet, a growing body of research indicates that ANPR networks are well suited to the identification and study of other relevant traffic phenomena, particularly those whose study has been restricted to highway or rural settings. One such phenomenon of interest is the identification and assessment of recurring traffic bottlenecks, characterised by congested traffic and queue formation upstream of the bottlenck location. Extending the identification of bottleneck-behaviour to entire urban networks is a difficult process because of their

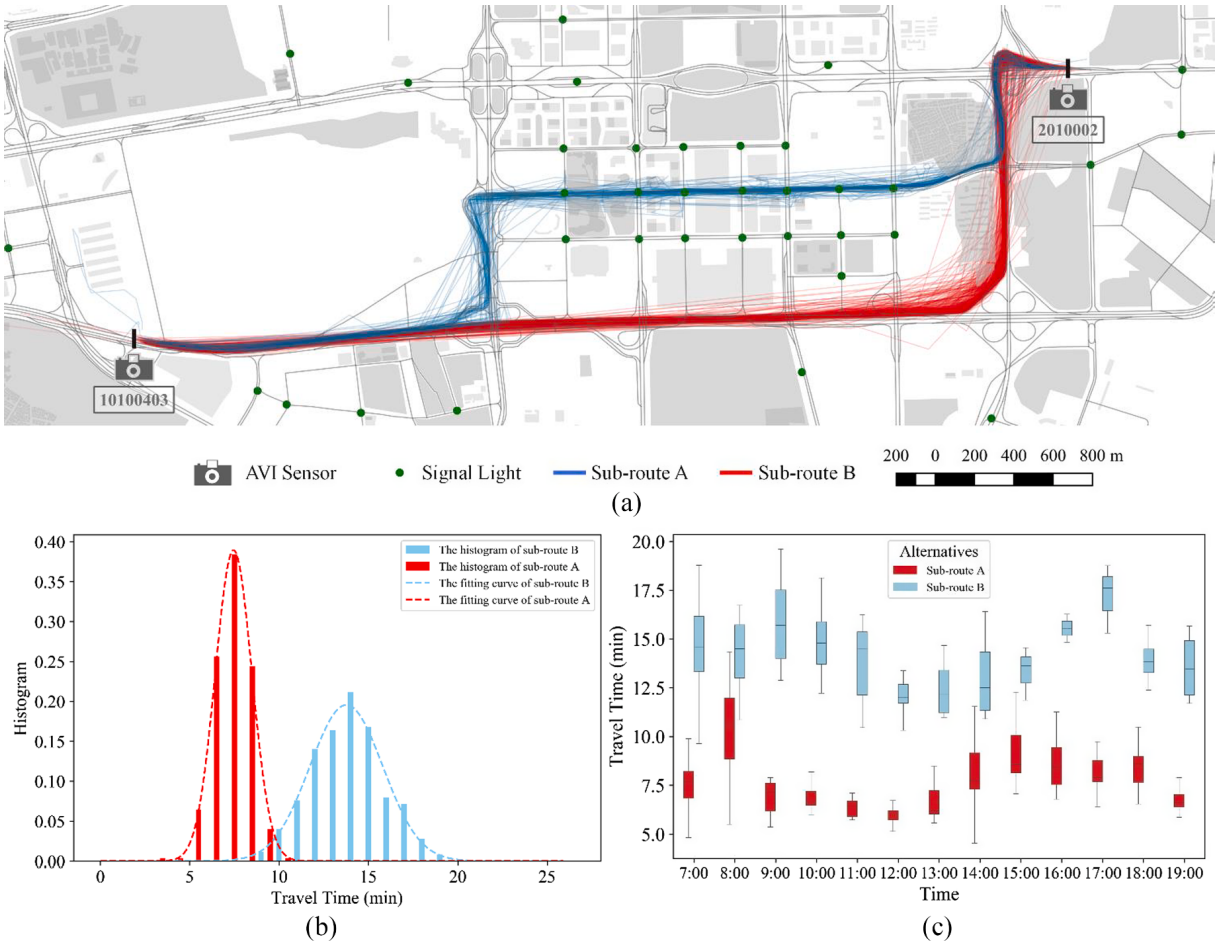


Figure 2.6: Image from Q. Cao et al. (2020) showing two routes (obtained from GPS data) as being frequently used alternatives to travel between two AVI sensors (right to left). One route (A) incurs on average a smaller travel time than the alternative (B).

complexity and scale. However, the analysis of bottleneck behaviour in urban networks is made possible by ANPR networks because they offer trajectory-aware coverage of vehicle travel in urban networks. To the best of our knowledge, ANPR technology is not actively used for the purposes of bottleneck detection. Therefore, developing a methodology for bottleneck identification and impact assessment is one of the main research gaps addressed in this thesis, in Chapter 5.

In developing our methodology of bottleneck analysis using ANPR data, we identified a lack of tooling and guidance about how to prepare ANPR data for analysis. We interpret the lack of tools and documentation as being a major barrier to ANPR research. The treatment of ANPR data is a multi-step process with varying levels of complexity and hence error-prone. Because data processing is a precursor to analysis, we tackle this problem first, in Chapter 3, not only as a means to advance and support our research but also that of other traffic practiotioners and academics.

In addition to the treatment of ANPR data, we recognise a lack of formalism in how other studies treat sequences of vehicle observations that can contain erroneous observations. This problem is relevant since an analysis of bottleneck behaviour requires understanding whether and how any given two routes are related, so that their performance of their performance can be evaluated for bottleneck-like characteristics. Although we identified other relevant trip sequencing approaches in the Literature, we could not find one that eliminates spurious travel sequences, or groups them, on the basis of their spatial qualities. We tackled this gap in Chapter 4, by developing a precise definition of popular and meaningful travel sequences, conceptually known as road corridors, that serves as a mathematical foundation not only for bottleneck identification but also for other applications that wish to analyse vehicle travel patterns over distinct periods of time.

Lastly, we identified overtaking behaviour as one possible application area of ANPR systems. Overtaking data is notouriously difficult to collect at scale, yet valuable to study driver behaviour and understand the safety properties of a road or corridor. As several studies have shown, ANPR can be used to capture vehicle overtakings. ANPR networks are thus good candidate systems for collecting overtaking data reliably and at scale. The study of how overtaking behaviour changes with traffic conditions is a research opportunity that remains unexplored, and constitutes a research gap that we address in this thesis in Chapter 6.

Chapter 3

A Pipeline for ANPR data

3.1 Introduction

ANPR cameras are increasingly employed for traffic monitoring and control (Antoniou et al., 2011). The reliability of measured travel times makes ANPR a rich and appealing source of data for traffic forecasting, particularly in urban cities (Kazagli et al., 2013; F. Zheng et al., 2018). Besides short-term travel time prediction (R. Li & Rose, 2011; E. I. Vlahogianni et al., 2014) ANPR data has been used in a range of traffic problems, namely to estimate origin-destination matrices (W. Rao et al., 2018; X. Zhou & Mahmassani, 2006), detect road incidents (Hellinga & Knapp, 2000; X. Li et al., 2013), and estimate queue length at signalised intersections (Luo et al., 2019; Ma et al., 2017).

Before ANPR data is ready to be analysed, it goes through a series of processing steps. Each step requires the user to make methodological choices that affect the quality of the final dataset and subsequent data analysis. For example, the size of aggregation window determines the amount of noise and variation in aggregated data (E. Vlahogianni & Karlaftis, 2011). To make informed decisions before and during the analysis, the practitioner benefits from understanding, or simply being aware of: (i) the required processing steps and their motivation, so that they are not skipped by accident; (ii) the different methodological choices available at each step and (iii) their impact on data quality and research outcomes.

While reviewing research papers using ANPR data, we have found that many studies fail to report on several of the required processing steps (Table 3.1). Under-reporting is either accidental, if a user fails to recognise the need for a certain processing step, or intentional, if the user deems it too obvious to require reporting. Alternatively, users might be working with data that has already been processed by traffic authorities and whose processing details may be unknown or unavailable. Regardless of the cause, under-reporting of pre-processing steps limits the reproducibility and quality of academic research (Gentleman

& Temple Lang, 2007). Furthermore, it creates an entry barrier for researchers and practitioners new to ANPR data, who may find themselves discovering the necessary processing steps once again.

Table 3.1: Reporting of pre-processing steps in a sample of AVI/ANPR studies executed after 2010.

Authors	Application	Tech	T[1]	T[2]	T[3]	T[4]	T[5]	T[6]	T[7]	T[8]
Tam & Lam (2011)	Traffic forecasting	AVI	✗		✓	✗	✓	✗	m	✗
R. Li & Rose (2011)	Traffic forecasting	AVI			✓	✗	✓	✓	m	✗
Cheng et al. (2012)	Traffic forecasting	ANPR	✗	✗	?	✗	✓	✗	m	o
Haworth & Cheng (2012)	Traffic forecasting	ANPR	✗	✗	✗	✗	✓	✗	m	i
Kazagli et al. (2013)	Traffic forecasting	ANPR	✗	✗	✓	✗	✓	✗	m	✗
F. Zheng et al. (2018)	Reliability analysis	ANPR	✓⁻	✗	✓⁻	✓⁻	✓	✗	m	o
W. Rao et al. (2018)	Demand estimation	ANPR	✗	✗	✓⁻	✗			m	
Crawford et al. (2018)	User clustering	ANPR	✗	✗	✓	✗			n	o
Luo et al. (2019)	Queue estimation	ANPR		✓	✓	✗			m	
Zhao et al. (2019)	Trip analysis	ANPR	✓⁻	✓	✓	✓⁻			n	
Hadavi et al. (2020)	Trip analysis	ANPR	✗	✗	✓	✗	✓	✗	n	✓⁻
Barroso et al. (2020)	Demand estimation	ANPR	✗	✗	✓	✗	✓	✗	n	o
Q. Cao et al. (2020)	Route choice analysis	ANPR	✗	✗	✗	✗			n	

✓= identified the problem and specified the treatment.
✓⁻= identified the problem but not the treatment.
✗= not reported.
(empty) = not applicable.
[1] Camera clustering (Section 3.3.1).
[2] Pre-filtering of number plates (Section 3.3.5).
[3] Outlier detection / Trip identification (Section 3.4.2).
[4] Duplicate observations (Section 3.4.3.2).
[5] Choice of aggregation window (Section 3.5.1).
[6] Sampling method in vehicle aggregation (Section 3.5.2).
[7] Route selection: m = manually/experts, a = algorithm, n = none (Section 3.5.3).
[8] Treatment of missing data: o = omission, i = imputation (Section 3.5.4)

To address these concerns, we develop a curated and comprehensive guide to ANPR data, structured as a data pipeline, that researchers and practitioners can consult when producing research data ready for analysis. The pipeline documents key data sources and processing steps, clearly indicates their execution order and the data flows. While the emphasis is on offline analysis, i.e. batches of collected data, the concepts are equally relevant for online (real-time) analysis. Furthermore, existing techniques at each stage are surveyed. Where existing techniques prove to be inadequate or insufficient, we develop new methodology. New methods are developed to solve the following tasks: the clustering of camera locations, the mapping of camera locations to map locations, the identification and treatment of outlier travel times and the identification of routes for analysis.

This chapter is structured as follows. Section 3.1.1 gives a broad overview of the pipeline and its main components – objects (datasets) and processes (processing steps). Section

3.2 introduces the primary datasets that interact with the pipeline, as well as the different states each dataset goes through. Sections 3.3 through 3.5 detail issues faced within each of the three major pipeline stages – *Wrangling*, *Trip identification* and *Aggregation* – and solutions available. Section 3.7 concludes by outlining directions for further improvement.

3.1.1 Pipeline overview

The proposed data pipeline is divided into three major stages, depicted in Figure 3.1:

1. **Wrangling** – Stage 1 addresses the collection and treatment of road data from camera descriptors; the filtering of malformed number plate numbers in unprocessed ANPR data and their anonymisation.
2. **Trip identification** – Stage 2 seeks to generate realistic observation sequences for each vehicle, by matching observations to trips and removing observations most likely to originate from detection errors.
3. **Aggregation** – Stage 3 assigns vehicle observations to time intervals; computes sample statistics of traffic flow for each route, such as vehicle counts and mean speed; identifies unlabeled routes likely to be relevant, i.e. routes that were not labelled as such by professionals but whose data suggests so; and provides a method to label and treat missing data.

Figure 3.1: Overview diagram of the ANPR data pipeline: major stages and key data sources.

Data flows linearly through the pipeline – the output of each stage serves as the input to its successor stages. Each stage is thus designed to be self-contained and modular. Users can swap one algorithm for another, at any given stage, without breaking functionality. Additionally, the pipeline specifies well-defined data interfaces between stages. Altogether, these characteristics - modularisation, encapsulation of behaviour and well-defined data interfaces - make this approach distinct of that from previous highlighted studies, shown in Table 3.1, which have not provided any consistent categorisation of tasks or expected processing behaviour.

Sections 3.3, 3.4 and 3.5 describe the role and impact of each stage on data quality, along with its key themes, processing challenges and existing solutions, where applicable. The main outputs of the pipeline are trip data and flow data, produced at Stages 2 and 3 of the pipeline, respectively. Trip data encodes trips made by individual vehicles, whereas

flow data keeps measurements pertaining to traffic streams (groups of vehicles). Both datasets are used in data analysis, albeit generally in different applications (as described in Section 2.2.2)

3.2 Datasets

There are four primary datasets, each serving a specific function:

- **Camera data** encodes the location and characteristics of ANPR cameras.
- **Road data** stores a spatial model of the road network used to compute paths and distances between any two points on the network.
- **Route data** contains spatial features and descriptive attributes of routes formed from camera pairs.
- **Number plate data** encodes the vehicles (number plates) identified by each camera, which are then used to compute vehicle trips and summarise traffic streams over time.

Datasets transition between states as they flow through the pipeline. For instance, camera data is transformed from state *raw* to state *wrangled* in Stage 1 of the pipeline. All possible states of each data source are catalogued next.

3.2.1 Camera data

3.2.1.1 Raw cameras

Camera data is a list of ANPR cameras employed by the stakeholder (e.g. traffic authority). A dataset is labelled *raw* after it has been retrieved from a database and before it is modified. In other words, *raw* data has not yet been checked for problems that may affect the quality of other data that depends on it, as described further ahead in Section 3.3. For every camera on the list, *raw* data must contain the attributes – geographic location, road address, and camera orientation – in one format or another, regardless of other attributes available. Depending on the format chosen by the stakeholder, different steps may be necessary to extract the required information into the format described in Section 3.2.1.2.

Table 3.2 shows a sample of *raw* camera data, as provided by the Tyne and Wear UTMC (Urban Traffic Management and Control) centre. In this instance, camera location is given via latitude-longitude pairs, whereas road address and camera orientation are combined in a text field named *description*. Orientation is specified as a cardinal direction,

i.e. north, east, south and, west. For example, camera ids 1 and 2 are located together but monitor traffic flowing in opposite directions whereas id 17 monitors traffic flowing in both directions. Address serves primarily to identify the road the camera points at, which in some instances, will be one of several nearby links. The address can also indicate whether multiple cameras are employed at one location (ids 1005 and 1003), to cover multiple exit ramps or all traffic lanes (cameras can cover at most two lanes, so a minimum of two cameras are necessary to monitor three lanes). It is desirable to cluster nearby cameras pointing in the same direction, to guarantee that vehicle detections are associated with unique locations. Finally, the description field can be used to identify cameras that are not relevant to analysis. For example, if our goal is to model traffic flow then we can ignore cameras placed inside car parks (id 1081) or registered for testing purposes (id 1119).

3.2.1.2 Wrangled cameras

Wrangled camera data is obtained by treating raw camera data. Table 3.3 depicts a sample of wrangled camera data. Each *wrangled* camera corresponds to a single camera or a cluster of nearby cameras that share the same location/address and orientation. This transformation guarantees that vehicle detections are associated with a unique location and orientation pair. New camera identifiers are issued to reflect camera clusters and camera removal (e.g. test/park cameras). The *address* field is split into several attributes – county, road number and road name – in order to provide a more granular data model and facilitate camera selection for subsequent analysis, for instance confined to a specific area of the road network.

For ease of use, each coordinate pair is instantiated as a (point) geometric object – a data type used to represent and manipulate spatial objects such as points, lines and polygons. To instantiate coordinate pairs as geometric objects, the original coordinate reference system (CRS) in the raw camera data must be known. In Table 3.2, coordinate pairs are provided in `WGS84` (World Geodetic System, 1984), an ellipsoidal 2D coordinate system with coordinates specified in degrees with axes longitude and latitude. In Table 3.3 the coordinate pairs are read into a geometric data object and then projected into `WGS84 / UTM zone 30N`, a cartesian 2D coordinate system with coordinates specified in meters and axes easting and northing.

Table 3.2: Sample of raw camera data.

id	code	description	latitude	longitude	type
1119	CA TEST EVOX	Dual modem test camera	52.0028	-0.9700	Dual Lane
1	CAANPR GHA167DR1NB	Gateshead A167 Durham Road Site 1 Northbound Camera at Whitehall Drive	54.9543	-1.5999	Dual Lane
2	CAANPR GHA167DR1SB	Gateshead A167 Durham Road Site 1 Southbound Camera at Whitehall Drive	54.9543	-1.5998	Dual Lane
17	CAANPR GHA167DR5NS	Gateshead A167 Durham Road Site 5 North/South Camera Barley Mow Birtley	54.8820	-1.5761	Dual Lane
1081	CAANPR NCBANKFTCPIN	Bank Foot Metro Car Park Entry Camera	55.0139	-1.6783	Single Lane
1030	NPANPR GHA695SGEB	Gateshead A695 Stargate Lane Eastbound	54.9642	-1.7485	Single Lane
1005	NPANPR NCA695SBL1NB	Newcastle A695 Scotswood Bridge Eastbound to A1 L1	54.9674	-1.6896	Single Lane
1003	NPANPR NCA695SBL2NB	Newcastle A695 Scotswood Bridge Eastbound to Newcastle City Ctr L2	54.9674	-1.6896	Single Lane

Table 3.3: Sample of wrangled camera data.

id	old id	orientation	county	road number	road name	type	geometry
048	10	N-S	Gateshead	B1296	Old Durham Road	Dual Lane	POINT(591039, 6088021)
208	1005-1003	E	Newcastle	A695	Scotswood Bridge	Single Lane	POINT(583888, 6091954)
056	1009-1002	S	Sunderland	A1018	Wearmouth Bridge	Single Lane	POINT(603657, 6086093)
180	101	N	South Tyneside	B1298	Stanhope Road	Single Lane	POINT(599834, 6093069)
251	96	N	Sunderland	B1522	Ryhope Road	Single Lane	POINT(605119, 6082420)
252	97	S	Sunderland	B1522	Ryhope Road	Dual Lane	POINT(605128, 6082420)

3.2.2 Road data

3.2.2.1 Raw road graph

A road graph is a mathematical abstraction of a road network. The abstraction is useful because it allows a spatial problem to be treated as a graph-theoretical problem. A notable example in transportation is route planning, i.e. to find the shortest or fastest route between two locations on a map. Given a road graph, route planning amounts to solving a shortest path problem (Miller et al., 2001). Figure 3.2 depicts the road network encompassing the admnistrative region of Tyne and Wear, in the Northeast of England – our main case study throughout this thesis.

To compute camera routes, we first need to retrieve the road network that encompasses the set of cameras and build the corresponding road graph. The resulting graph is named *raw* because the cameras are not elements of its node set – thus the problem can not yet be translated to a shortest path one. Section 3.3.2 gives an overview of different graph-based representations of road networks, and outlines different ways to use OpenStreetMap (OSM), currently the largest open bank of geographical data, to retrieve, store and manipulate road data (Haklay & Weber, 2008).

3.2.2.2 Merged road graph

A *merged* road graph extends a *raw* road graph by incorporating the cameras as additional graph nodes. The spatial route that connects two cameras then translates to the shortest path between the two camera nodes on the merged graph. The process of mapping cameras to existing edges in the road graph is explored in Section 3.3.3. Figure 3.3 illustrates the resulting of joining two ANPR cameras with its parent raw graph.

3.2.2.3 Road hierarchy information

Roads are classified according to their function in the road hierarchy. Road classification systems vary with country but a three or four-tiered system is typically employed (Eppell et al., 2001). The category of a road is related to its function in the road network. Highways are high volume and restricted access roads (also called motorways), sit at the top of hierarchy while local streets, which provide residential access and low traffic volumes, sit at the bottom of the hierarchy.

In addition to their primary classification, roads can be associated with additional labels if they serve a specialised role in the network. In England, the Strategic Road Network (SRN) is composed of all motorways and major 'A' roads, while the newly formed Major Road Network (MRN) is composed of the busiest and most economically important local

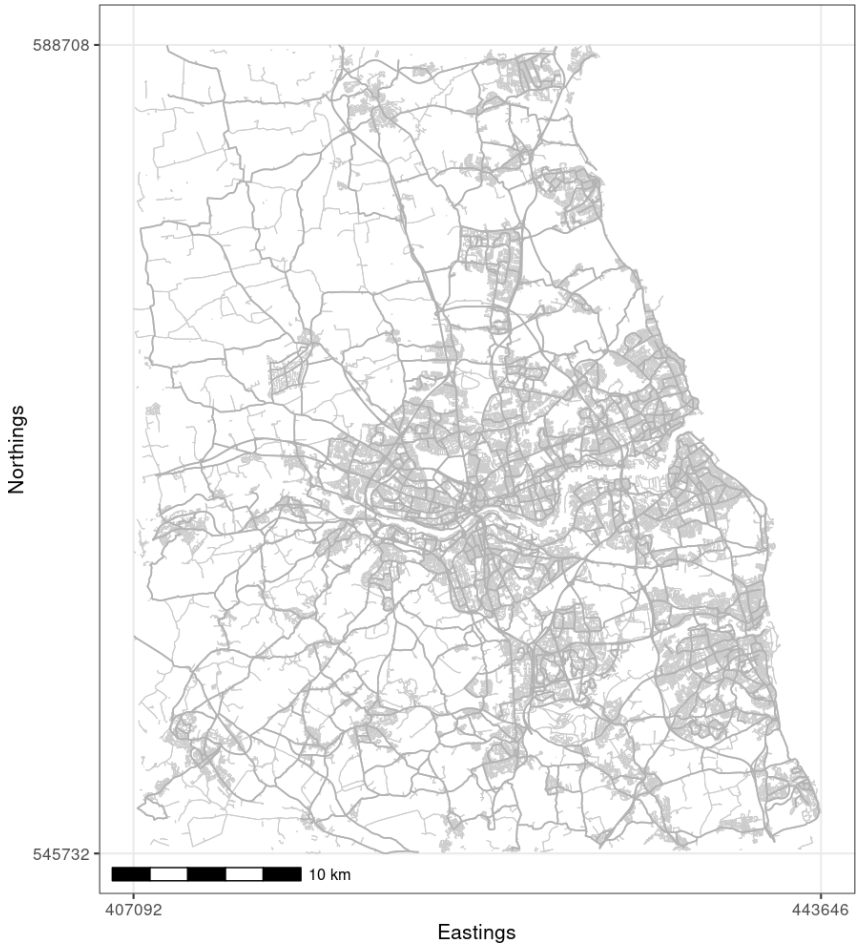


Figure 3.2: Complete road network, obtained from OpenStreetMap, for our case study in the Northeast of England.

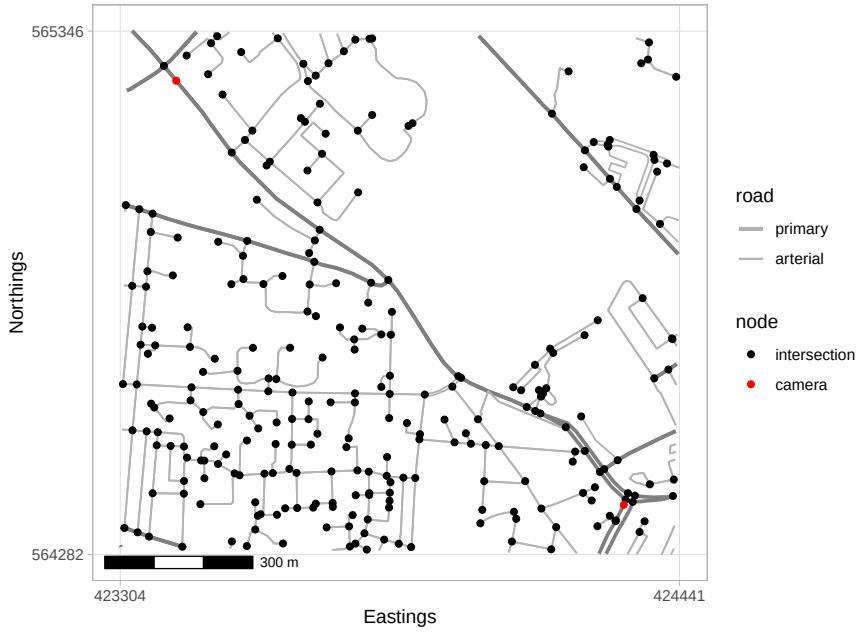


Figure 3.3: Sample of two ANPR cameras merged with the corresponding road graph.

authority 'A' roads (Department for Transport, 2017; Green et al., 2019; Quarmby & Carey, 2016). The MRN ranks between the SRN and the local road networks. Its role is to alleviate congestion and support the SRN by creating a more resilient road network. Spatial data about the SRN and MRN can be obtained online, at the time of writing (*Major Road Network*, 2020; *Ordinance Survey Open Roads*, 2020).

3.2.3 Route data

3.2.3.1 Camera pairs

A *camera pairs* dataset provides a complete enumeration of pairs of cameras and their shortest paths distances. Table 3.4 shows a sample of camera pair data, where x_1 denotes the origin location and x_2 the destination location. Shortest paths are computed as a sequence of nodes in the *merged* road graph (column `path`), with a corresponding line geometry data structure which encodes the spatial mapping of the path (column `geometry`). Note that the shortest path problem requires origin and destination nodes to be distinct, therefore no path is available for pairs whose origin and destination locations are the same.

Table 3.4: Sample of camera-pair data. Columns 'path' (sequence of road graph edges representing the found shortest-distance path) and 'geometry' (corresponding spatial vector data) are included to reflect the process of shortest path computation but their values are truncated because they are not meaningfully represented as text.

x_1	x_2	distance (km)	path	geometry
106	129	7.4	...	LINESTRING(...)
188	204	2.8	...	LINESTRING(...)
139	138	2.5	...	LINESTRING(...)
131	193	2.4	...	LINESTRING(...)
121	236	1.6	...	LINESTRING(...)
073	168	1.5	...	LINESTRING(...)
224	057	0.7	...	LINESTRING(...)
226	190	0.2	...	LINESTRING(...)

3.2.3.2 Routes

A collection of *routes* is a subset of camera pairs selected manually by experts, or automatically according to one or more characteristics, such as average traffic volume. The set of routes will typically be much smaller than the set of camera pairs. Routes not selected may correspond to origin-destination pairs for which there is no meaningful demand, or

an alternative shorter/faster route is available. In addition, *route* data expands the number and quality of features describing each route. Again, the annotation procedure can be performed manually and/or automatically. Route identification (Section 3.5.3) is performed using *flow* data in Stage 3, and is then optionally followed by human/algorithmic annotation.

Table 3.5 depicts a sample of *route* data. Each row represents a different route, as in Table 3.4. Furthermore, note that *route* data is simply a camera pairs dataset that has been cleaned and augmented with additional data columns. A number of characteristics are shown, some manually annotated and others calculated from data. Annual Average Daily Traffic (AADT) and mean free flow speed v_f are calculated from *flow* data. AADT is a measure of average traffic volume commonly used to determine road importance (Davis, 1997). *Dir* represents the orientation of each camera and is obtained from *wrangled* camera data, likewise for county, road number and road name. The remaining features are annotated manually for demonstrative purposes: road function[1] and number of lanes (in addition to type of carriageway and number of traffic lights along the route, not shown). Note that route features are averaged across the route's length, e.g. mix lane denotes routes that are partly single lane and partly multi lane.

Table 3.5: Sample of route data.

x_1	x_2	dir	dist (km)	AADT (K)	county	road number	road name	function	v_f (km/h)	lanes
100	069	W-W	3.0	8.7	Newcastle - North Tyneside	A1058	Coast Road	freeway	92.6	multi
234	003	N-N	1.4	3.4	Gateshead	A167	Durham Road	arterial	66.2	mix
077	036	N-N	4.1	3.4	North Tyneside - Northumberland	A189	Moor Farm	arterial	68.6	multi
075	261	E-N	1.5	1.1	Gateshead	B601	Bensham Road - Team Street	collector	36.3	single
150	080	W-W	1.4	1.5	Gateshead	A1114	Colliery Road - Handy Drive	collector	46.6	single
086	122	S-S	1.6	1.3	Sunderland	A690	Durham Road	collector	44.1	mix

3.2.4 Number plate data

Number plate data exists in four states: *raw*, *wrangled*, *trips* and *flow* data. Raw and wrangled are intermediary data states, not ready for analysis. Trips can be used to analyse individual vehicle behaviour and flows are used to analyse macroscopic traffic flow. Flow data is the more commonly used type of number plate data, as it serves as input for predictive and inferential models of traffic flow. Trip data poses several privacy threats

[1] the categories are given according to Eppell et al. (2001)

and should be stored and shared with care, as it is susceptible to de-anonymisation even after the application of pseudo-anonymisation techniques.

3.2.4.1 Raw detections

Raw ANPR is a stream of data generated by a network of ANPR cameras before any processing takes place. It is obtained by collating the data streams from individual cameras into a larger data stream containing all number plate detection events in the network. Each event represents where and when a given vehicle was observed. A sample of raw ANPR data is show in Table 3.6, where vehicle plate numbers are suppressed to ensure user privacy.

ANPR cameras are uniquely identified by an integer and time is defined according to Unix time (also known as POSIX timestamps Group (2018)). As timestamps are relative to the camera's clock, clock synchronisation is necessary to avoid significant clock drifts that may impact data analysis. To achieve clock synchronisation, the Tyne and Wear traffic authorities connect ANPR cameras to a private Internet Protocol (IP) network, which allows cameras to be synchronised via the Network Time Protocol (NTP) using a central server as the reference clock (Mills, 1991). Our initial analysis of *raw* number plate data indicated that time offsets between the camera and reference clocks are negligible, i.e. below 100 milliseconds. Hence, we assume henceforth that recorded timestamps can be used "as is" – i.e. can be applied directly in the calculation of travel time between two cameras.

Table 3.6: Sample of *raw* ANPR data.

vehicle	camera	timestamp	confidence
————L	29	2018-03-05 08:13:00	78
————B	85	2018-03-05 09:16:35	92
————L	1056	2018-03-05 10:05:29	100
————F	8	2018-03-05 10:34:36	84
————Y	1067	2018-03-05 12:02:13	100
————M	65	2018-03-05 12:22:23	94

Additionally, ANPR cameras may assign a *confidence* value to each number plate detection. The *confidence* score, shown in percentage value, is an optional output provided by the camera's built-in optical character recognition (OCR) algorithm, that is potentially useful to identify and discard plate detection errors associated with low confidence detections.

3.2.4.2 Wrangled detections

Table 3.7: Sample of *wrangled* ANPR data.

vehicle	camera	timestamp
1010680	181	2018-03-05 05:37:02
1201450	078	2018-03-05 09:50:05
1225932	239	2018-03-05 11:07:02
1327660	087	2018-03-05 15:08:55
1192191	253	2018-03-05 16:14:31
1373110	212	2018-03-05 17:51:56

Wrangled ANPR data is the result of passing *raw* ANPR data through the first processing stage of the pipeline: *Data Wrangling*. A sample of *wrangled* ANPR data is shown in Table 3.7. Vehicle number plates are anonymised and replaced by number ids; low confidence observations are optionally discarded (according to the process defined in Section 3.3.5.2); and new camera ids are generated after solving the *camera-clustering* problem (described in Section 3.3).

3.2.4.3 Trips

Trip data is the main output of the second stage of the pipeline – *trip identification* – and results from applying a series of transformations to *wrangled* ANPR data. *Trip* data is of relevance because it groups vehicle observations into distinct trips – a natural representation and partitioning of vehicle behaviour. A sample of *trip* data is shown in Table 3.8.

Table 3.8: Sample of trip data.

vehicle	x_1	x_2	t_1	t_2	tt (sec)	dist (km)	v (km/h)	trip	j	l
1061284	-	048	-	18:21:33	-	-	-	2	0	5
1061284	048	079	18:21:33	18:24:26	173	1.5	31.5	2	1	5
1061284	079	147	18:24:32	18:26:40	128	1.1	29.9	2	2	5
1061284	147	111	18:26:40	18:27:50	70	0.5	26.7	2	3	5
1061284	111	-	18:27:50	-	-	-	-	2	4	5
1306995	-	129	-	00:19:54	-	-	-	2	0	2
1306995	129	-	00:19:54	-	-	-	-	2	1	2
1498214	-	072	-	15:23:19	-	-	-	6	0	3
1498214	072	118	15:23:19	15:25:53	154	2.6	59.9	6	1	3
1498214	118	-	15:25:53	-	-	-	-	6	2	3

Foremost, note that vehicle observations are recast as flows from one location to another, rather than point observations. Vehicle trips, uniquely identified by the attribute pair `(vehicle,trip)` are composed of two or more trip steps, ranging from $j = 1$ to $l \geq 2$. Each trip step specifies an observed vehicle flow from origin location x_1, with departure time t_1, to destination location x_2, with arrival time t_2 and corresponding travel time $tt = t_2 - t_1$. The distance between x_1 and x_2 is obtained by merging from camera pair data, and is then used to calculate the average speed.

As vehicles are observed in transit, the true origin and destination locations of each trip are unknown. This is represented in the first and last steps of each trip – the origin location is missing when $j = 1$ and the destination location is missing when $j = l$. Altogether, a trip of length l is composed of $l - 1$ point (camera) observations, and only $l - 2$ steps have a valid travel time and speed reading. This is shown in Table 3.8: trip 2 of vehicle 1306995 is composed of a single camera observation and hence no travel time measurements are available; trip 2 of vehicle 1061284 is of length 5, meaning that the vehicle was observed at 4 distinct points and resulting in 3 vehicle flows with valid travel time measurements.

3.2.4.4 Flows

Flow data are the main output of the data pipeline. They characterise traffic streams across time and space. ANPR generates two types of flow data: flows at locations (a point on a road) and along routes (the length of road between two cameras). At locations, ANPR is limited to measuring flow rate, i.e. counting number of vehicles per unit of time. Along routes, the measurements of individual vehicles are combined to produce summary statistics of travel time and/or speed, namely the sample mean, median and standard deviation. An excerpt of *flow* data taken from two camera locations along A1231 (Sunderland Highway), and computed over consecutive non-overlapping 15 minute time intervals, is shown in Tables 3.9 and 3.10. Table 3.9 illustrates how vehicle counts can be captured at each location and Table 3.10 shows how origin-destination flow data can be collected by way of aggregation, along the route connecting the two locations (according to the processes described in Section 3.5).

Table 3.9: Sample of *flow* data measured at two locations.

x	t	count
016	2018-03-05 08:00:00	414
016	2018-03-05 08:15:00	412
016	2018-03-05 08:30:00	389
016	2018-03-05 08:45:00	386
255	2018-03-05 08:00:00	645

x	t	count
255	2018-03-05 08:15:00	714
255	2018-03-05 08:30:00	697
255	2018-03-05 08:45:00	684

Table 3.10: Sample of *flow* data measured along a route.

x_1	x_2	t	count	v (median)	v (mean)	v (sd)
016	255	2018-03-05 08:15:00	152	61.0	60.6	14.4
016	255	2018-03-05 08:00:00	153	66.2	64.3	14.6
016	255	2018-03-05 08:30:00	115	62.1	59.3	17.6
016	255	2018-03-05 08:45:00	119	63.1	58.8	20.0
016	255	2018-03-05 09:00:00	85	68.2	62.8	24.1
016	255	2018-03-05 09:15:00	87	69.0	59.8	27.2
016	255	2018-03-05 09:30:00	76	65.8	57.5	25.8
016	255	2018-03-05 09:45:00	89	71.0	65.8	22.3

3.3 Stage 1: Wrangling

The goal of data *Wrangling* is to transform "raw" data into a standardised format with desirable properties. Stage 1 has two primary data inputs – raw number plate data and raw camera data – and three derivative outputs – wrangled number plate data, wrangled camera data and camera pairs data. Primary inputs are provided by the problem owners via an interface to a storage facility, such as a SQL database (referred to as "ANPR database"). In addition, a source of geospatial data is required to build the road network graph, which is subsequently combined with camera data to produce camera pairs.

Figure 3.4 illustrates the complete data workflow for the *Wrangling* Stage. Processes on the same level can be run in parallel provided that parent resources have been made available or executed (e.g. Process "Refactor camera ids" requires wrangled camera data and the output of Process "Eliminate bad numbers plates," which in turn takes raw number plate data as input).

Firstly, cameras are selected for relevance (e.g. discard test/car park cameras) and clustered on the basis of distance, orientation (direction of captured traffic flow) and address. Camera clustering ensures that multiple nearby cameras pointing in the same direction are associated with a single location. Concurrently, a geospatial data provider such as OpenStreetMap, is used to retrieve road network data within the box enclosing the cameras'

geographical point set.

The outputs of these two processes, wrangled camera data and the raw road network graph, are combined to create a merged road network graph – a graph where the cameras are part of its node set. This is achieved through a process that matches camera locations (point geometries) to one of several nearby edges (line geometries representing road segments). Once the cameras are added to the graph, the computation of cameras pairs can be treated as a shortest-path problem.

Lastly, Stage 1 seeks to clean raw number plate data by: (1) eliminating badly formed number plates and low confidence detections, (2) refactoring the ids clustered cameras and (3) anonymising the vehicle detections. Key processes in the Wrangling stage are developed in the following subsections.

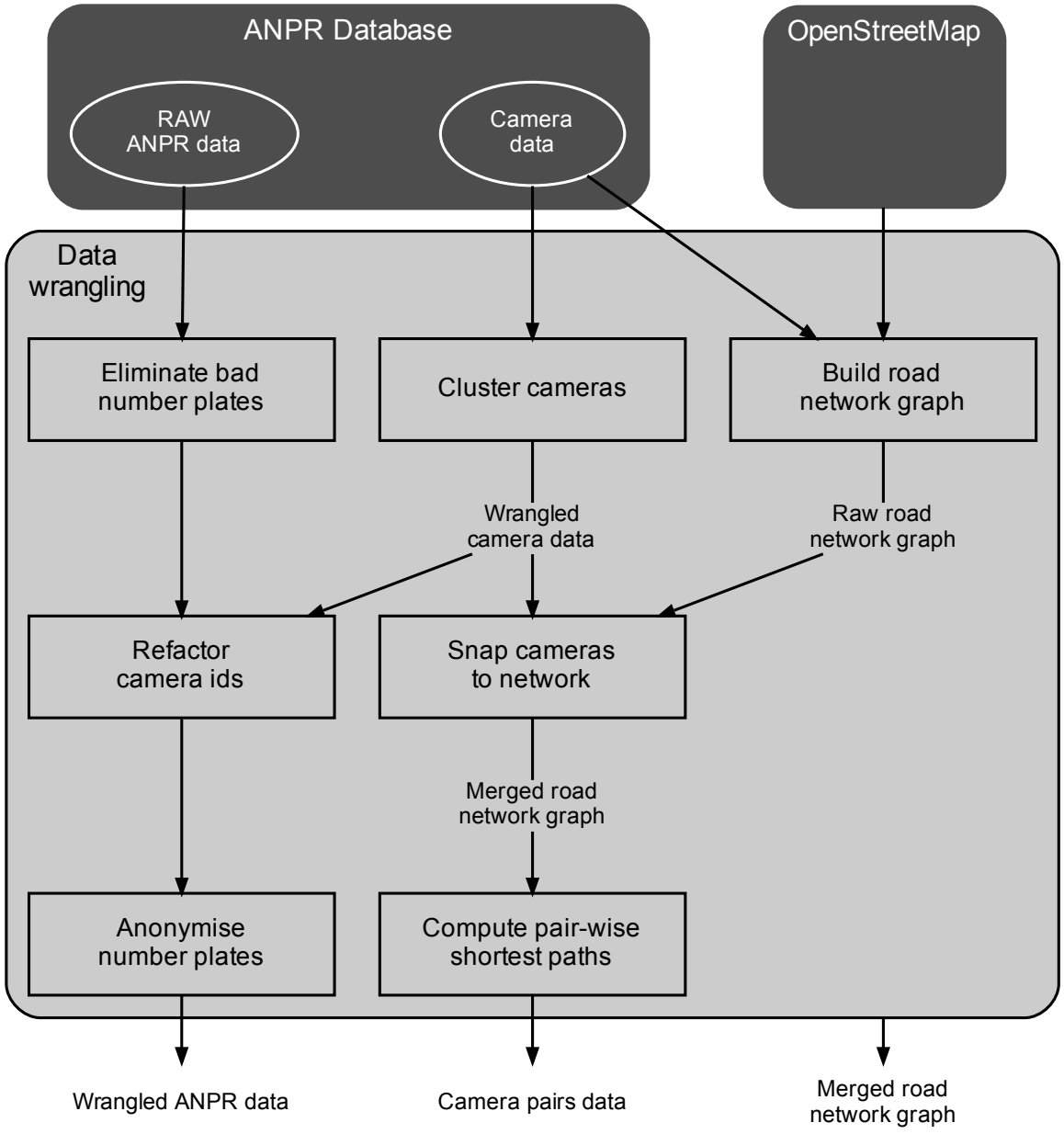


Figure 3.4: Workflow within the data wrangling stage: input data (ellipse), output data (plain text) and processes (box).

3.3.1 Camera clustering

The camera clustering problem arises when multiple nearby cameras are installed on the same road segment, with the same orientation. This is anticipated if a single camera can not provide full road coverage, due to its width (multiple lanes) or geometry (to accommodate merging/diverging traffic at intersections). Because vehicles can be detected by one of several cameras in a cluster, unique locations are not associated with unique camera identifiers. This property is undesirable as it causes vehicles along one route to be registered not by one but multiple camera pairs, effectively disaggregating related traffic flow data.

To address this problem, one creates a mapping of cameras to camera clusters, where applicable, and refactors number plate data to reflect the new camera ids. Generally, traffic operators are expected to be aware of the issue and manually associate cameras to clusters following the installation of ANPR cameras. However, researchers are not guaranteed to be offered "corrected" ANPR data, and thus should be aware of the camera clustering problem and be ready to apply the corrective procedure themselves, if necessary.

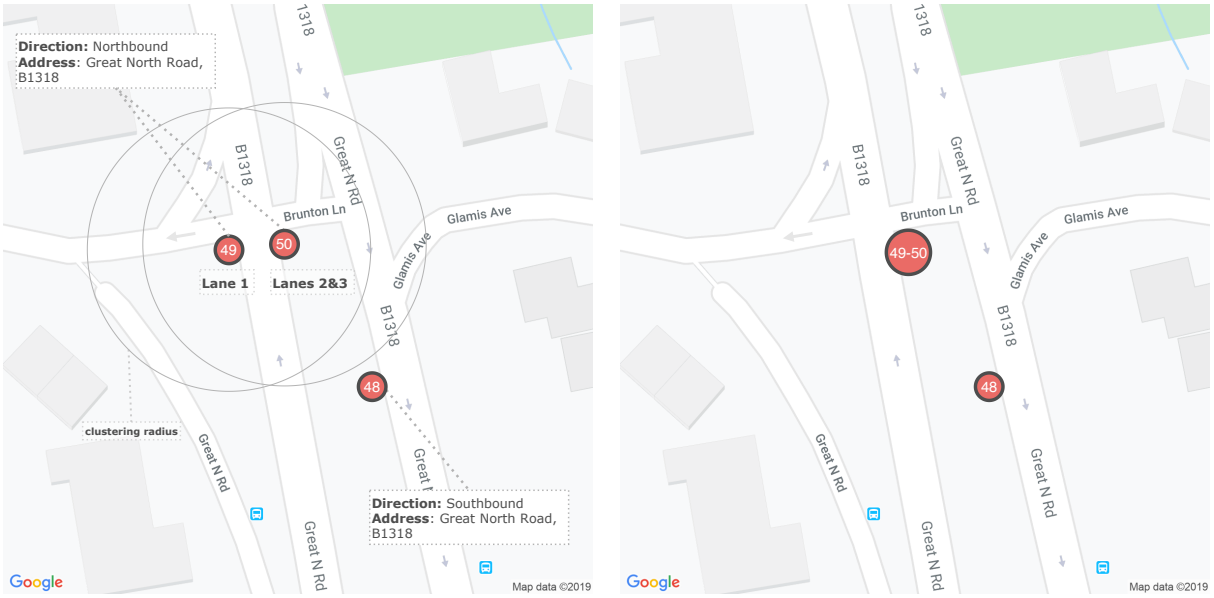
Based on raw camera data for the county of Tyne and Wear, a simple clustering procedure is to group cameras by address and orientation, and merge group elements within $L = 100$ meters of each other (twice the maximum camera operating range, measured in a straight line). It is assumed that camera address and orientation are in standardised formats adequate for string comparison, and geospatial location is provided as a point geometry in a grid reference system (where locations are defined as points in the Cartesian plane) such as UTM (Universal Transverse Mercator) or OSGB (Ordnance Survey National Grid).

Figure 3.5 illustrates the camera clustering problem and its solution: to merge together two neighbouring cameras pointing in the same direction but monitoring different lanes. A total of 30 clusters were found this way, out of 266 input cameras in the Tyne and Wear dataset (after selecting for non test/car park cameras).

3.3.2 Retrieval of road network data

Road network data, like other instances of geospatial data (geodata), is represented and stored in Geographical Information Systems (GIS), from whence it can be queried and analysed (Longley et al., 2015). To obtain accurate and reliable road network data, users depend on map providers capable of collecting and validating geodata at scale, the most notable of which is OpenStreetMap – an open collaborative effort to develop a free and editable map of the world (Haklay & Weber, 2008).

OpenStreetMap (OSM) emerged as an alternative to propretary map data that has since supported public and private research, and services worldwide (OpenStreetMap, 2021b).



(a) **Problem**: Different camera identifiers map to the same location. Cameras `49` and `50` monitor Northbound traffic in the same road but in different lanes (camera radius of 50 meters is shown in grey).

(b) **Solution**: Merge cameras `49` and `50`, but not camera `48` because it monitors *Southbound* traffic instead.

Figure 3.5: An instance of the *camera clustering* problem.

Although the quality of OSM data vastly differs from region to region, it has greatly improved and remains extremely high-quality in countries such as the UK and U.S. (Barrington-Leigh & Millard-Ball, 2017; Barron et al., 2014; Haklay, 2010).

Recently, numerous tools have been developed to simplify the process of querying data from OSM servers. A noteworthy example is OSMnx, a python package that retrieves OSM road data from simple inputs like an address, a latitude-longitude point or bounding box. Moreover, the software builds the corresponding road graph as a python dictionary, including available road attributes such as maximum speed and segment length (Boeing, 2017). OSMnx stands out because the returned graph is (a) a multidigraph, i.e. a weighted directed graph that allows for multiple edges between two nodes, and (b) is topologically corrected and simplified so that all graph nodes actually represent street junctions, as opposed to points where, for example, the street curves. For network analysis, a range of graph-theoretical algorithms, namely Dijkstra's shortest path algorithm, are then directly available via the NetworkX python API (Hagberg et al., 2008).

OSMnx employs a primal representation of road networks – junctions are represented as graph nodes and road segments as edges. Its advantage lies in preserving the network's geographic and spatial characteristics, thus promoting the analysis of networks with spatial features (Porta et al., 2006). The downside of the primal representation, however, is that it fails to properly encode turning restrictions at intersections – limiting its use in applications where vehicle re-routing is key (Añez et al., 1996). This limitation is not

critical, as the goal is to reasonably approximate the distance travelled between any two points in the network (measured along the road rather than "as the crow files").

3.3.3 Map matching of camera locations

Map matching is a problem routinely solved by transport navigation systems. A map matching algorithm is tasked with finding the correct road segment that a vehicle travels on, given inputs from a GPS device or similar location-aware sensor (M. A. Quddus et al., 2007; White et al., 2000).

Map matching of ANPR cameras differs from GPS map matching as there is only a single location to match rather than a sequence of locations. Consequently, the range of applicable map-matching algorithms is limited to approaches that do not consider vehicle trajectory, heading or speed. Of the several types of map-matching algorithms reviewed by M. A. Quddus et al. (2007), only algorithms based on the geometric analysis of spatial road data can be reasonably employed for a single position input.

Geometric algorithms treat map matching as a search problem – their goal is to find the closest road segment to an input location. Point-to-point matching finds the nearest node (a shape point of a road segment), whereas point-to-curve matching finds the nearest curve. The difference between the two methods lies in the class of geometric objects that each is willing to consider. Point-to-point matching considers segments' shape points as specified in the spatial road network data, and point-to-curve matching treats road segments as geometric line objects. In the latter case, the Hausdorff distance finds the distance from the nearest point on the line geometry to the fixed input location (Rockafellar & Wets, 1998).

A limitation of geometric algorithms is that their accuracy is dependent on the quality of the spatial road network data, that is, how well curvilinear roads are approximated by piecewise linear segments (White et al., 2000). Moreover, detection can be challenging in the presence of high road density, as the closest link is not necessarily always the correct link (M. A. Quddus et al., 2007).

Contrary to GPS issued locations, cameras may have annotated attributes that can be useful in map matching. If operators record the camera's orientation[2] along with its location, then an exact matching exists in the nearest road segment across the camera's line of sight. The matching is exact because cameras are installed with a clear view of incoming traffic, not with other roads or infrastructure in between (Gurney et al., 2013). If camera orientation is not supplied, then algorithmic map matching of camera locations

[2]Camera orientation can be recorded as an angle in relation to another object, namely as an absolute bearing. In navigation, the absolute bearing is the horizontal angle between the direction of an object and that of the true north (Hofmann-Wellenhof et al., 2003).

is not guaranteed to be exact. However, the procedure can be avoided altogether if traffic operators manually 'map match' cameras following their installation (and appropriately embed that information in camera data).

In the case of the Tyne and Wear cameras dataset, camera orientation is unavailable, hence the motivation for an algorithmic approach to map matching. The algorithm employed is a variant of point-to-curve matching, wherein only selected candidate links are ranked according to their proximity to the input location. Road segments are considered to be valid candidates if they are within the camera's operating range (set at 50 meters, twice the operating range reported by Gurney et al. (2013), which will depend on the manufacturer), i.e. if its Hausdorff distance is below the defined range, and match its recorded address and cardinal direction. Furthermore, roads whose category is below that of the monitored segment are further disqualified from the set of candidate links.

The comparison between camera address and road name is possible because the OSM road name attribute shows little missing data: over 80% of road segments and 95% of total segment length were found to be associated with a road name. These figures apply to segments whose OSM `highway` key value is one of the following: motorway, trunk, primary or secondary – (OpenStreetMap, 2021a). In road categories below secondary, such as tertiary roads, segments show no recorded road name. To compare camera and road directionality, we calculated road bearings and mapped each bearing to one the four direction quadrants ("NE," "NW," "SE," "SW"). These can then be matched directly, for instance, a Northbound facing camera would match roads with bearing "NE" or "NW."

With the pre-filtering of candidate segments, point-to-curve matching was able to pick the correct road segment for the vast majority of cases (220 out of 234 cases). Obtained map matching results were manually verified using Google Maps' Street View feature. Unsuccessful matches were primarily due to missing data (cameras placed in tertiary roads) or a mismatch of camera direction and the road's bearing. After the correct road network edge is identified, the edge is cut into two at the point where the distance to the camera is minimal, and a new node is added in between. Figure 3.6 demonstrates the map matching procedure for a pair of ANPR cameras.

3.3.4 Computation of spatial routes

In transport planning, route estimation is the primary goal of traffic assignment – the fourth step in the longstanding four-step traffic forecasting model (McNally, 2007). Given user trips in the form of an origin-destination (OD) matrix, traffic assignment models are tasked with assigning each trip to a spatial route in the input road network[3]. To that end,

[3]A spatial route is a geometric object representing a trajectory between two geographical points (Longley et al., 2015).

(a) **Problem**: What is the road segment (network edge) that the camera is observing? Network nodes are represented in blue and edges in grey. Camera point geometries are shown in red.

(b) **Problem**: In many cases, there are multiple roads within range so we need a heuristic to choose between the several candidates. Grey boxes show available edge attributes. Some edges have missing attributes.

(c) **Solution** (step 1): Identify edges whose attributes match the camera's information (address and traffic flow direction). Compute the shortest distance from each candidate edge (line geometry) to the camera (point geometry).

(d) **Solution** (step 2): Rank candidate edges by distance, in ascending order, and select the top candidate. Then, modify the road graph to include the cameras as new graph nodes.

Figure 3.6: An instance of the *camera map matching* problem.

users are generally assumed to pick the route which minimises their travel cost, typically travel time (Ortúzar & Willumsen, 2011).

Traffic assignment models vary in how realistically they model route choice. The simplest method, known as "all or nothing" assignment, has users choose the fastest route (or lowest-cost route) regardless of traffic congestion (Ortúzar & Willumsen, 2011). A widespread method, called static traffic assignment, was built on Wardrop's principles of User and Network Equilibrium – users choose the path of least resistance but in accordance to network conditions (Wardrop, 1952). A computational solution is reached when the conditions for equilibrium are met: when users can not unilaterally improve their travel time. That is, switching to an alternative route would incur an additional cost. More recently, dynamic traffic assignment adds a stochastic component to user choice by considering imperfect knowledge of the route network and its state, as opposed to perfect user knowledge assumed in static traffic assignment (Chiu et al., 2011).

Like OD pairs in transport planning, a camera pair is an abstract representation of vehicle movement between two camera locations. As such, it does not imply a particular spatial route; to a greater extent because ANPR technology is unable to observe true vehicle trajectories (it is incapable of tracking vehicles' positions in between the cameras). Despite the unobservability of vehicle trajectories, their estimation is useful in providing an assessment of travelled distance. Coupled with travel time, distance is used to calculate average vehicle speeds and perform trip identification (as described in Sections 3.2.4.3 and 3.4).

At this stage, the primary goal is to reasonably approximate the spatial route of a camera pair, rather than to perform realistic traffic assignment. A sufficient method is thus to find the shortest path between two nodes in the (merged) road graph (Nicholson, 1966), the equivalent of "all or nothing" assignment. Recall that a shortest path is a sequence of distinct edges between two nodes in a graph, such that the sum of their weights is minimised (Diestel, 2017). If segment length is minimised, the shortest path corresponds to the shortest route. If free flow data or speed limit data are available, the fastest route (under free flowing conditions) can be found instead (by minimising free flow travel time, calculated as the segment length divided by the speed limit). A review of efficient methods for computing shortest paths can be found, for instance, in Dreyfus (1969).

The computation of spatial routes is equally relevant for "expert" and "non-expert" camera pairs. Expert camera pairs are planned by traffic operators via a strategic positioning of the cameras and, as such, imply a specific route of interest. In contrast, non-expert camera pairs occur as a byproduct of increased camera coverage and, unlike expert pairs, their spatial route is not pre-defined. For a sizeable ANPR network, non-expert pairs represent the majority of camera pairs. In the Tyne and Wear dataset, 184 out of 53824 possible camera pairs (0.3%) are classified by the Tyne and Wear UTMC team as "expert"

– where the number of possible camera pairs with different origin and destination nodes in a network with n cameras is $n(n-1)$.

Shortest paths were calculated for all possible pairs, as no reference geodesics for expert cameras were available[4]. Moreover, route length is minimised since we found several OSM road attributes, namely speed limit, to have high degrees of missing data. Therefore, to ensure that spatial routes do not contain residential or service roads, the shortest path algorithm considers only the subset of the road network that is composed of A, B and C category roads (primary, secondary and tertiary highway types in OSM), in accordance with the observed camera positions.

Even though users don't always choose the shortest route from origin to destination (S. Zhu & Levinson, 2015), the shortest path route provides a reference geodesic that is a better approximation of the real spatial route than that obtained if the road network were not considered at all (i.e. if a "straight" line between the two locations was drawn instead).

3.3.5 Treatment of raw number plate data

3.3.5.1 Filtering of malformed number plates

ANPR systems are built on computer vision and character recognition algorithms (Du et al., 2013). As such, they are subject to errors at any stage of the detection algorithm – vehicle image capture, number plate detection, character segmentation and character recognition (Patel et al., 2013). Detection errors can manifest openly, via a malformed number plate, or indirectly if the number plate is valid but the resulting travel pattern is anomalous, e.g. outlier travel time.

A number plate is malformed if it does not conform to valid formats, as defined by the relevant traffic authorities. In raw number plate data, malformed number plates are directly observable if number plates were not previously anonymised. For law enforcement purposes ANPR databases will typically contain non-anonymised number plates (Patel et al., 2013). The treatment for a malformed number plate is to discard the corresponding observation since it fails to represent a vehicle by its correct identifier (no consideration is given to past/future occurrences of the same value).

For data privacy reasons, the UTMC of Tyne and Wear identified and filtered malformed number plates prior to sharing their data with us. Filtering can be achieved by matching registration numbers currently and historically valid in the UK against a regular expres-

[4]If the computation is too expensive, it can be reduced to a subset of all camera pairs, for instance, only those observed in the dataset or those observed above a certain rate.

sion[5] (regex). In total, the local authorities reported close to 0.71% of malformed number plates in the 2018 dataset (approximately 5 million out of 700 million total observations).

3.3.5.2 Filtering of low confidence scores

As described in Section 3.2.4.1, detected number plates may be linked to a confidence score. A confidence score is a secondary output of neural networks – a machine learning algorithm commonly employed in ANPR systems (Hamad & Kaya, 2016) – that represents the probability associated with the predicted label (Hendrycks & Gimpel, 2017). The reporting of calibrated confidence estimates is encouraged to help users gauge the trustworthiness of predicted labels (C. Guo et al., 2017). For instance, decisions can be deferred to a human if the reported confidence is below a desirable threshold.

Thus, in theory, a low confidence score may be indicative of a malformed or misidentified number plate. However, unlike malformed number plates, low-confidence observations do not imply a camera detection error, as it is possible that a low confidence number plate is in fact a valid number plate and accurate detection. In addition, confidence scores can be poor predictors of performance if they are improperly calibrated, resulting, for example, in overly optimistic estimates (Hendrycks & Gimpel, 2017).

Due to the lack of empirical evidence, the relationship between confidence scores and camera detection errors remains unclear (exacerbated by the use of proprietary ANPR technology from different manufacturers). In view of the limited understanding, we recommend that the detection of additional camera errors is relayed to the next stage of the pipeline: trip identification. Nevertheless, we find it useful to highlight several statistics related to confidence scores.

In the Tyne and Wear dataset, malformed number plates were observed to carry, on average, lower confidence scores (measured from 0 to 100) than valid number plates: 73.5 (33.3 s.d.) versus 94.7 (9.86 s.d.), respectively. However, Table 3.11 shows that confidence scores tend to decrease with number plate length (in number of characters) for valid UK number plate lengths (between two and seven characters). Consequently, a sensible approach to filtering of low-confidence observations would require number plate length to be taken into account, e.g. by choosing a different threshold for each length group.

A possible explanation for lower scores in shorter character sequences is the increased number of alternative high-probability labels compared to longer sequences (as shorter sequences can be contained within longer sequences). That would cause non-zero probability scores to be issued to a larger number of possible matches, effectively dilluting the

[5]The regex was defined by experts from the Driver and Vehicle Licensing Agency (DVLA). Alternative example regexes can be found in GitHub, for example in (Bradley, 2013).

Table 3.11: Confidence score statistics by plate number length.

			Confidence Score[§]					
l^*	$p^\dagger$	$p_m{}^\ddagger$	mean	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}
1	7.0e-04	100	19.5	10	15	20	25	25
2	3.4e-03	17.7	15.6	3	10	15	19	25
3	8.3e-02	56.1	75.0	12	48	89	100	100
4	0.39	9.07	69.5	11	37	88	98	99
5	2.96	3.63	83.7	60	74	98	99	99
6	5.36	4.82	91.4	83	92	97	99	99
7	91.2	0.23	95.3	89	93	99	100	100
8	4.0e-02	100	56.7	0	2	84	88	99
9	8.8e-03	100	47.9	0	2	78	95	100
10	2.8e-03	100	58.8	0	2	88	100	100
11	7.5e-04	100	19.2	0	0	0	5	87

[*] Length of number plate (in number of characters).
[†] Proportion of detections in group (percentage) out of a total of approximately 700 million observations.
[‡] Proportion of malformed number plates (percentage).
[§] Mean and n-th percentiles.

maximum probability score issued to the correct number plate. Alternatively, a property of lower-character number plates causes the drop in confidence estimates (e.g. if the neural network was trained primarily on seven-character number plates).

3.3.5.3 Pseudo-anonymisation of number plates

Research ethics dictates that personally identifiable information should be (pseudo) anonymised prior to their use, particularly in the analysis of individual user behaviour (Oliver, 2010). A simple pseudo-anonymisation method is to hash number plates using a cryptographic hash function (Bonneau, 2012). The inclusion of a secret salt (random data passed as input) is encouraged so that reverse hashing by brute force becomes computationally infeasible.

Data hashing of number plates, by definition, does not protect against indirect de-anonymisation: if an outside actor has record of when and where a certain vehicle was, they can potentially infer its hashed identifier from the data and subsequently access other observations of the same vehicle. In ANPR systems, anonymisation is guaranteed only when individual identifiers are removed, that is, when trip data is aggregated to produce flow data.

3.4 Stage 2: Trip identification

Stage 2 seeks to better represent the travel movements of individual vehicles. An instance of vehicle movement between two locations is called a step. Spatially, a step is characterised by an origin-destination (OD) pair and a corresponding spatial route (computed

in Section 3.3.4). Temporally, a step is characterised by its travel time, defined as the difference between the times at the destination and origin locations.

Despite being the distinguishing feature of ANPR, the calculation of travel times is not error-free. Vehicle stops, e.g. for passenger drop-off and refuelling, cause travel times to be measured incorrectly since they do not represent a genuine attempt to drive promptly from origin to destination (Hadavi et al., 2020; Kazagli et al., 2013; Robinson & Polak, 2006). In particular, prolonged vehicle stops result in values that do not represent a true travel time but rather the time between two vehicle trips. If untreated, erroneous measurements produced at the individual level can bias travel time estimates produced at the aggregate level, used to characterise traffic streams (S. D. Clark et al., 2002; Dion & Rakha, 2006; Papagianni, 2003).

The identification and treatment of invalid steps is thus the primary goal of Stage 2. Its purpose is two-fold: (1) to compute the sanitised travel history vector of each vehicle, partitioned by vehicle trip; and (2) reveal the true sample distribution of travel time, given the time of day, for any OD pair. The first objective augments vehicle travel thorugh a more natural representation of vehicle movement (specified between locations rather than at locations), that is not only easier to interpret (steps are grouped on a trip basis) but also simpler to manipulate using the "split-apply-combine" computational strategy for data analysis (Wickham, 2011). The second objective seeks to improve the quality of aggregated traffic data through the removal of outlier values prior to analysis.

This section is structured as follows. Subsection 3.4.1 elaborates on the types of invalid travel steps, how they manifest in data and their treatment. Subsection 3.4.2 documents existing approaches for detection of outlier travel times in ANPR data and evaluates Tukey's box plot method (Tukey, 1977), chosen for its simplicity and speed of computation in handling large batches of input data. Subsection 3.4.3 details how the different treatments are implemented in practice.

3.4.1 Types of invalid travel steps

A step is considered to be invalid if it does not correspond to a direct travel movement between two distinct locations, namely if the observed travel time is significantly different than expected, relative to neighbouring vehicles or free-flow conditions. An invalid step can be caused by different types of events. These can be related to camera detection errors, i.e. a false positive or false negative, or be intrinsic to the data collection process, i.e. originate as a consequence of organic vehicle behaviour such as the end of a journey. Table 3.12 documents the several types of invalid steps, grouped into four categories, how these manifest in a dataset and the required treatment.

Type (i) events correspond to the end or interruption of a vehicle's journey. They manifest as long travel times, significantly longer than those of nearby vehicles or historical summaries captured in similar traffic conditions. In the case of prolonged stops, travel times are generally orders of magnitude greater than expected (i.e. several hours or days compared to minutes), as they do not represent true travel times but the time between two vehicle trips. Treatment seeks to identify these steps and invalidate the corresponding travel time, without discarding any of the observations. Vehicle trips are labelled for intepretability (and to allow for future behavioural analysis) – a trip counter is incremented after a type (i) event is detected. Due to the common occurrence of type (i) events compared to other event types, trip identification is the central theme of this section.

Table 3.12: Types of *invalid* travel steps and treatments applied in Stage 2 of the pipeline: *Trip Identification*.

#	Event	Camera error	Observable	Symptom	Treatment
i	Journey ends or is interrupted (vehicle stops)	No	Yes	Outlier travel time (high value)	Trip labelling
ii	Duplicate vehicle detection	No	Yes	Same origin and destination and low travel time	Remove observation
iii	Wrong vehicle is detected	Yes	When travel time is outlier	Outlier travel time (low or high value)	Remove (low), trip labelling (low,high)
iv	Vehicle is not detected	Yes	No	–	–

Type (ii) events lead to duplicate vehicle observations. These manifest as travel steps whose destination location is the same as the origin location. Duplicates occur because cameras can issue multiple detections of the same vehicle over a short period of time, namely if a vehicle is stopped at an intersection or traffic light. Treatment seeks to keep one of multiple observations and discard the rest. Preferably the earliest observation should be kept so that the travel time recorded immediately after includes any idle time at the original location attributed, for instance, to a signalised intersection.

Type (iii) and (iv) events are caused, respectively, by camera errors of type I (false positives) – number plate mismatch – and type II (false negatives) – missed detection. The two types of camera error are intrinsically correlated, as a type I error can naturally lead to a type II error (not only the wrong vehicle is detected but the original vehicle goes through undetected). Conversely, a type I error may not lead to a type II error if a vehicle is detected twice, once correctly and once incorrectly. To the best of our knowledge, the degree to which the two types of errors are correlated (and how environmental factors affect the relationship) has not been investigated (to that end, the detection of false

negatives has to be done manually or with the assistance of another technology).

Type (iii) events can manifest either as invalid number plates (treatment is discussed in Section 3.3.5), or valid number plates. If a type (iii) event manifests as a valid number plate, then another vehicle will exhibit an extra observation. In that case, the error is only observable if it also manifests as an outlier travel time. Differently to type (i) events, a type (iii) outlier can be either an extremely low or high-value travel time. For high-value outliers, treatment is through trip-labelling regardless of type, as it is not possible to distinguish between a type (i) and (iii) high-value outlier travel time. In contrast, low-value outliers are treated by discarding the observation leading to unlikely or even physically-impossible travel times (if the resulting average speed is above the capabilities of any motorised vehicle).

Type (iv) events are not (individually) observable. They can, however, manifest in groups as missing data if the camera's false negative rate peaks due to, for example, severe traffic congestion (Haworth & Cheng, 2012). The positive/false detection rate of ANPR cameras depends heavily on the built-in Optical Recognition Character (OCR) algorithm, and its ability to capture number plates with different characteristics and across environmental settings (Hamad & Kaya, 2016). Du et al. (2013) provides a comprehensive review of the challenges and OCR algorithms for ANPR. Development of new OCR algorithms is ongoing, and current progress is presented, for instance, in Silva & Jung (2018) and Weihong & Jiaoyang (2020).

3.4.2 Identification of outlier travel times

3.4.2.1 Existing approaches

Table 3.13 lists several authors who have recognised the need to identify and treat outlier travel times in ANPR data (resulting from invalid steps of type i and type iii). It includes only those studies specific to ANPR data as outlier detection is an active area of research (also known as anomaly detection) with a large and diverse range of techniques to choose from. For a complete review of outlier detection methods see, for instance, Chandola et al. (2009) or, more recently, H. Wang et al. (2019).

The approaches listed in Table 3.13 differ significantly in complexity. S. D. Clark et al. (2002) tested three simple statistical tests, based on summary statistics of samples obtained at 5 and 15 minute intervals. Dion & Rakha (2006) observed that the filtering algorithms used by early Automated Vehicle Identification AVI) systems were unable to follow structural fluctuations in travel time at higher sampling rates and low market penetration rates[6]. To address these issues, the authors developed an enhanced filtering

[6]The TranStar (TranStar, 2001), TransGuide(SWri, 1998) and Transmit (Mouskos et al., 1998), were

algorithm that acts as a robust exponential smoothing low-pass filter, and is able to capture both stationary and non-stationary signals. Their approach was later used in several applications, for example to study travel time reliability (Rakha et al., 2010) and build a real-time travel monitoring system (Tam & Lam, 2011). Robinson & Polak (2006) developed an alternative method based on expected distributions of vehicle overtakings, and more recently J. Li et al. (2020) proposed an alternative method based on wavelet decomposition.

Of the listed approaches, Dion & Rakha (2006) stands out for robustness and popularity. However, the algorithm can be costly for processing large batches of ANPR data (hundreds of millions of observations) as it was developed for online settings and therefore functions iteratively – computations at time t rely on previous computations at time $t-1$. Similar computational costs are presented by Papagianni (2003), Robinson & Polak (2006) and J. Li et al. (2020). Alternative outlier detection methods with computational cost $O(n)$ or worse may be generally considered unsatisfactory for large scale processing. Contrary to Dion & Rakha (2006), the statistical-based methods tested by S. D. Clark et al. (2002) are fast and easy to implement due to their simplicity. As shown by the author, although the 10 and 90th percentiles prove too sensitive to outliers, MAD-based tests can be effective in removing the majority of extreme outliers.

Table 3.13: List of proposed methods to filter outlier travel times in ANPR data.

Study	Method
SWri (1998)	Smoothed rolling average
Mouskos et al. (1998)	
TranStar (2001)	
S. D. Clark et al. (2002)	10th and 90th percentiles
	Mean absolute deviation (MAD)
	Quartile deviation student-t test
Papagianni (2003)	Iteration-mean
Robinson & Polak (2006)	Overtaking rule
Dion & Rakha (2006)	Adaptive exponential smoothing filter
Rakha et al. (2010)	
Kazagli et al. (2013)	Mixture model
J. Li et al. (2020)	Wavelet transform

three early Automatic Vehicle Identification (AVI) systems, a former technological iteration of ANPR, deployed in the U.S. to monitor vehicle travel times. Vehicle identification was based on RFID (radio frequency identification) tags pre-installed in vehicles. Hence, the market penetration rate of RFID-enabled vehicles was a critical factor taken into consideration in previous studies. Although market penetration is not a factor in ANPR (instead detection rate is), the procedure of outlier detection is similar in both types of systems.

3.4.2.2 Outlier detection using Tukey's rule

John Tukey introduced the boxplot as a graphical tool for exploratory data analysis (Tukey, 1977). Besides providing robust summary statistics of a data sample, a boxplot also highlights potential univariate outliers. A data point is labelled as an outlier if it lies outside a region delimited by an upper and lower bounds, called "fences," constructed around the median according to the the spread of the distribution.

To describe statistical dispersion, Tukey defined the interquartile range (IQR) to be equal to $q_3 - q_1$, i.e. the difference between the first and third sample quartiles q_1 and q_3[7]. Using the IQR, Tukey specifies two sets of fences: the inner fences $f_1 = q_1 - 1.5 \times \text{IQR}$ and $f_3 = q_3 + 1.5 \times \text{IQR}$ serve to identify *mild* outliers, while the outer fences $F_1 = q_1 + 3 \times \text{IQR}$ and $F_3 = q_3 + 3 \times \text{IQR}$ are used for *extreme* outliers. A more general family of rules is created by varying the constants k_1 and k_3 in:

$$F_1 = q_1 - k_1 \times (q_3 - q_1) \text{ and } F_3 = q_3 - k \times (q_3 - q_1). \tag{3.1}$$

The boxplot fences method, hereafter referred to as Tukey's rule, was not tested by S. D. Clark et al. (2002) but is similar in that it also constructs limits around the median wherein valid data points lie. Aside from simplicity and ease of implementation, Tukey's rule is invariant to change of location and scale[8], and has a breakdown point of roughly 25%, i.e. the fraction of outliers the estimator can tolerate[9] (Carling, 2000).

Figure 3.7 illustrates the problem of outlier detection for an OD pair, in respect to: (a) the signal trend, captured using the median of observations calculated at evenly spaced 15 minute intervals[10]; (b) a constant threshold of 30 km/h used to classify outliers; (c) Tukey's rule, applied to every 15 min interval samples. The sparse cloud of points located above the trend line constitutes roughly the data points to be classified as outliers (note that travel times are shown on the log scale, base 10).

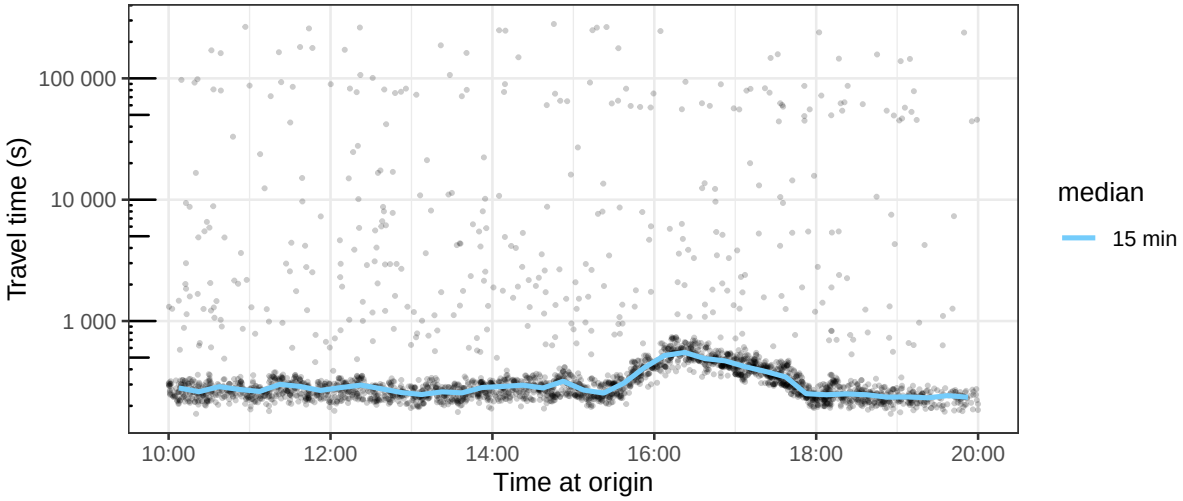
Figure 3.7b demonstrates how the use of a constant threshold for outlier detection can result in valid observations being mistaken for outliers during periods of traffic congestion, indicated by a surge in travel times between 4pm and 6pm, and outliers passing as normal observations during periods of free flowing traffic. A constant threshold may perform satisfactorily if the signal is stationary but, as shown, changes in traffic conditions, i.e. free flow vs congested, lead to structural fluctuations in observed travel times (E. I.

[7] Among others, we assume Tukey's definition of sample quartiles (Frigge et al., 1989; Tukey, 1977).

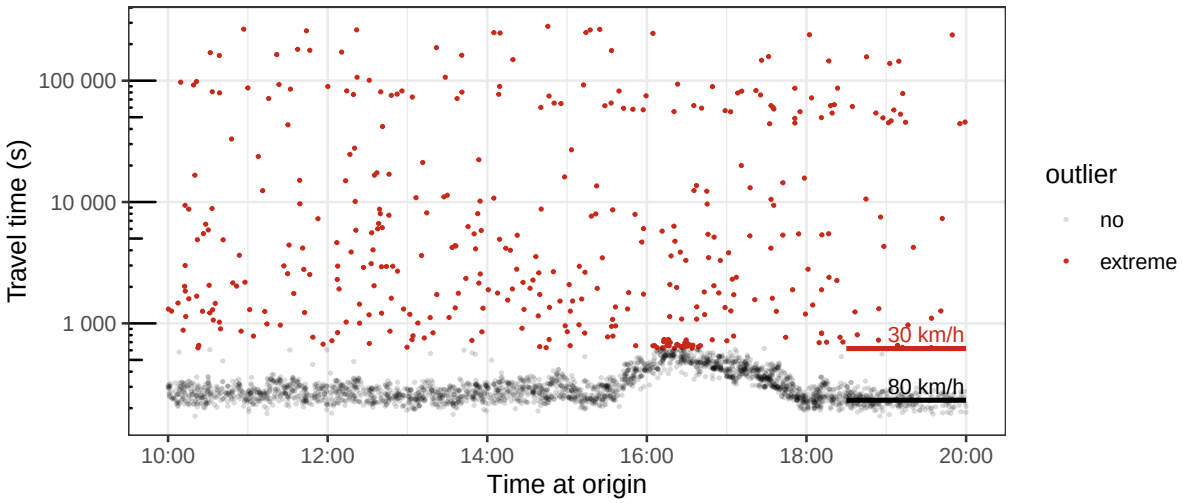
[8] The method "works" regardless of the location and scale parameter values of the underlying data distribution, e.g. mean and standard variation in the case of the Normal Distribution.

[9] Like most outlier detection methods, the boxplot method assumes outlier data points constitute but a fraction of the complete dataset.

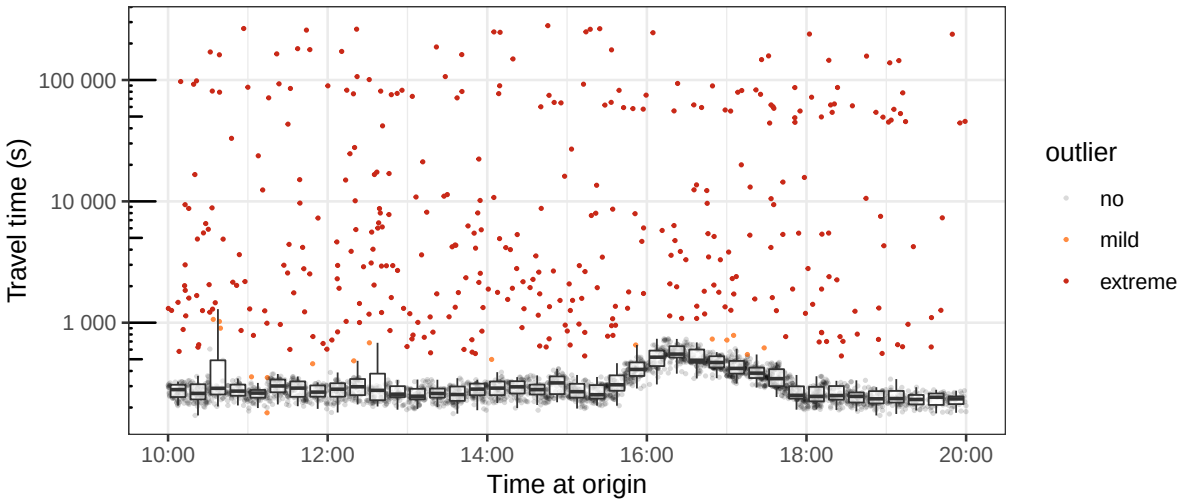
[10] The median is preferred to the mean, because it provides a more robust estimate of the central value of a distribution in the presence of outliers (Leys et al., 2013).



(a) Median travel time, computed every 15 minutes.



(b) Outliers identified using a constant threshold value (30 km/h).



(c) Outliers identified using the boxplot fences method (Tukey, 1977), after grouping data in 15 min time intervals.

Figure 3.7: Travel times recorded on camera pair 10016-100255, between hours 10:00 and 20:00, on 08 March 2018.

Vlahogianni et al., 2014). Figure 3.8 illustrates in detail the evolution of traffic conditions from free flowing at 3pm (characterised by higher travel speeds) to congested traffic during 4 and 5pm (slower travel speeds) and later returning to free flowing conditions at 6pm.

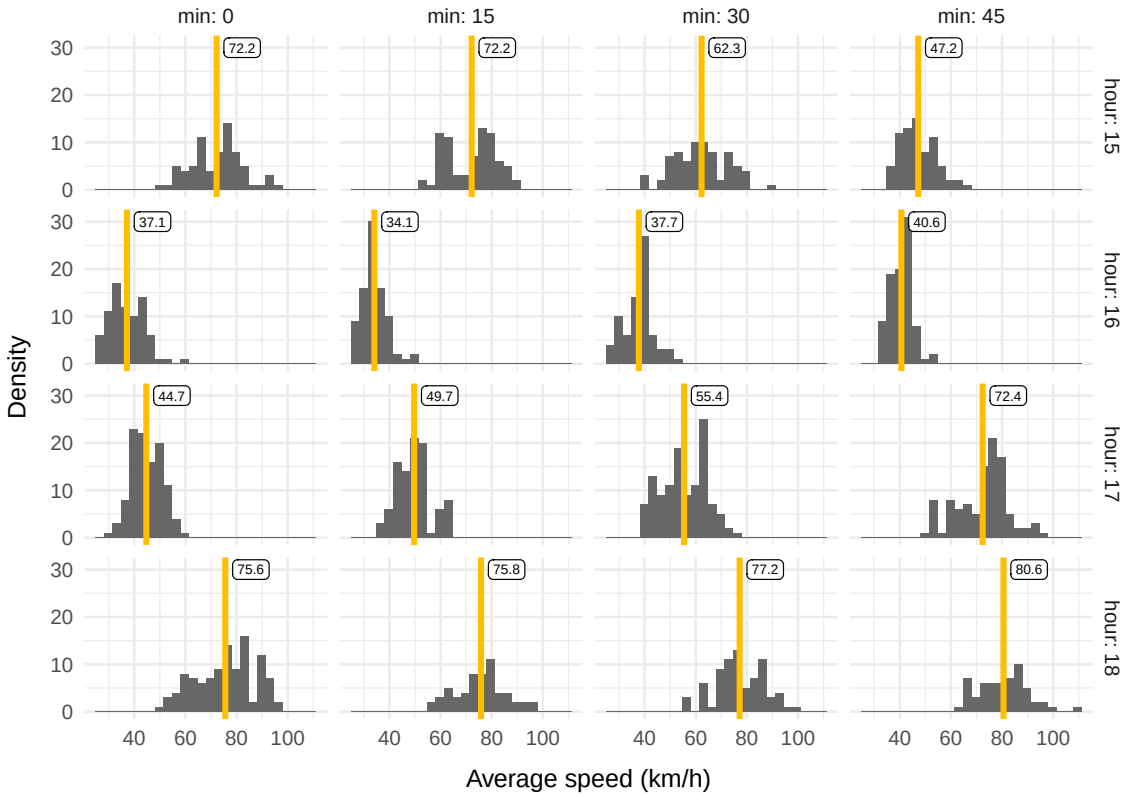


Figure 3.8: Sample distribution of vehicle speeds obtained for the 15-minute intervals (after removal of identified outliers) depicted in the example of Figure 3.7c, between 3 and 6pm. The sample mean is labelled and depicted as a yellow vertical bar in order to show the evolution of traffic conditions from free flowing at 3pm to congested 4 and 5pm and returning to free flowing conditions at 6pm.

To make outlier detection adaptive, time is divided into distinct intervals such that the signal is approximately stationary within each time interval and a suitable threshold is determined for each block. Figure 3.7c shows the result of applying Tukeys's rule to determine detection thresholds for each 15-minute time interval. Time intervals should be wide enough to ensure that most samples have a minimum number of data points and that traffic conditions are approximately stationary within that interval (i.e. that an interval captures either congested or free-flowing conditions but not both simultanously). Time intervals between 5 to 15 minutes are generally recommended (S. D. Clark et al., 2002; J. Guo et al., 2007; R. Li, 2006; E. Vlahogianni & Karlaftis, 2011). Since the application of Tukey's rule functions essentially by way of aggregation, the choice of interval size is elaborated further ahead, in Section 3.5.1.

3.4.2.3 Two-pass outlier detection

As described above, Tukey's rule can tolerate up to 25% of outlier data points in a data sample. For OD pairs with a high volume of observations, this assumption generally holds true. For instance, S. D. Clark et al. (2002) classified less than 5% of data points as outliers in two distinct ANPR routes. However, this figure does not generalise to all OD pairs as some pairs will exhibit low or inexistent user demand (OD pairs which fail to represent any meaningful or frequently chosen route). In such cases, the fraction of outliers may consistently dominate that of valid data points, thus invalidating the application of Tukey's rule (detection is still necessary in order to obtain properly treated travel vectors, i.e. realistic sequences of travel steps). Similarly, high percentages of outliers can occur in small sample sizes arising from periods of low activity, namely during nighttime.

To address these issues, we suggest a two-tiered approach to outlier detection. First, a constant threshold is employed to remove the most extreme outliers, effectively capturing instances of vehicles stopped for prolonged periods of time. Second, Tukey's rule is applied to the remaining data points, provided that the remaining sample contains a minimum of ten data points. A minimum of ten samples is added for robustness as ten is twice the number of data points required to build a boxplot.

Table 3.14 specifies constant thresholds values for the first detection pass. Values are given as average speeds as these can be used across different OD pairs. Travel time tt and average speed v are related via the formula: $tt = \frac{l}{v}$ (excluding any necessary unit conversions), where route length l has been previously calculated in Section 3.3.4. Thresholds are tailored to each OD pair by considering their free flow speed v_f. Free flow speed is estimated as the median of nighttime speed observations (ranging between 30 and 120 km/h), under the assumption that traffic conditions are free flowing during this period, similarly to Gong & Fan (2017). In the event that free flow speed can not be estimated for a given OD pair, a flat threshold of 5 km/h (for lower speed) and 140 km/h (for upper speed) are used instead. Different lower threshold values are used for each time of day so to reflect free flowing conditions during nighttime (the factor 1/3 assumes free flow conditions do not produce average speeds lower than 33% of free flow speed) and allow for periods of heavy congestion during daytime (the factor 1/10 assumes that congestion does not result in average speeds lower than 10% of free flow speed).

The effect of the first pass in reducing a detection threshold affected by a high percentage of outliers is exemplified in boxplots labelled 'before' and 'after' in Figure 3.9a, taken from the 10:45am interval in Figure 3.9a (third boxplot from the left). By applying the boxplot method directly to a sample contaminated by a high percentage of outliers, the detection threshold is artificially increased (boxplot labelled 'before'). In contrast, when the two-step approach is applied (boxplot labelled 'after'), the first step shaves off the

most of the extreme outliers and allows Tukey's Rule to be subsequently executed below its breakdown point.

Another example where the application of a constant detection threshold is required is seen in Figure 3.9b. Graphically, it becomes difficult to distinguish valid from invalid travel times. This example is representative of an OD pair with low demand (387 observations over 13 hours with an average of roughly 7.5 observations per 15-minute time interval) and where the rate of invalid travel steps is approximately half of all observed steps (195 outliers versus 192 valid observations). The application of Tukey's rule in such cases is invalidated because the number of valid samples is often too small to produce meaningful sample quantiles for outlier detection.

Table 3.14: Recommended lower and upper threshold values employed in the first pass of the two-stage trip identification algorithm. If the free flow speed v_f of the route is known (e.g. estimated from night time observations) then it can be used to determine F_1 and F_3 for two different times of day: day and night time, representing busy and free-flowing periods of the day respectively. If v_f is unknown or can not be estimated, a universal threshold can be used instead.

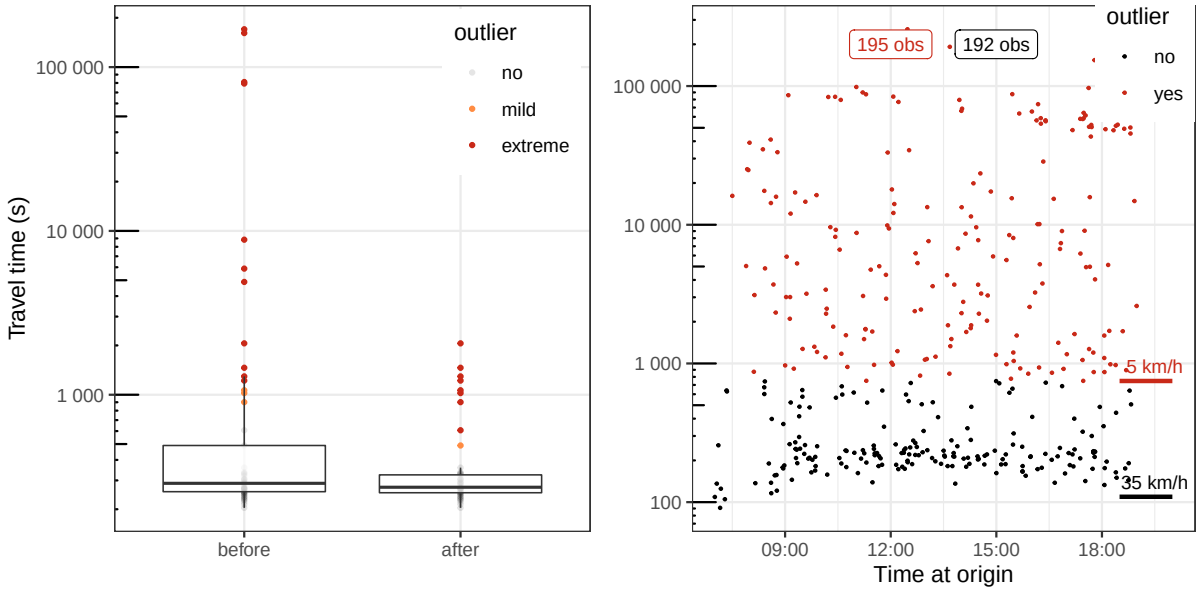
condition	lower threshold F_1	upper threshold F_3
can not estimate v_f	5 km/h	130 km/h
day	$\frac{1}{10} \times v_f$	$2 \times v_f$
night	$\frac{1}{3} \times v_f$	$2 \times v_f$

3.4.2.4 Method evaluation and benchmarking

In proposing the two-step approach for outlier detection, it becomes necessary to gauge its effectiveness. For reference, it is also useful to calculate the final proportion of data points classified as outliers.

As true outlier labels are unknown, evaluation is performed by making statistical assumptions about the distribution of data samples. Two empirical results are relevant: given the time of day, vehicle travel times have been found to be lognormal distributed and vehicle speeds normally distributed (Bauer & Tulic, 2018). Based on these results, we assume that non-outlier speed samples are normally distributed within each temporal window. If the proposed outlier detection method works well, then model fit should generally be good. Otherwise, if model fit is consistently poor, we expect to observe a proportion of extreme values much larger than that predicted by the normal distribution.

For a scalar k, the standard normal distribution (with mean $\mu = 0$ and variance $\sigma^2 = 1$), predicts extreme values below $-k$ or above k to occur at a frequency of



(a) Effect of removing extreme outliers using a constant threshold before applying Tukey's rule for outlier detection.

(b) Low demand OD pair where approximately half of observed travel steps constitute outliers.

Figure 3.9: Two demonstrative cases where performing a first-pass outlier detection has benefits: (a) to bring the proportion of outliers below breaking point (by discarding extreme outliers) and (b) when sample sizes are not large enough to apply Tukey's rule.

$$\begin{aligned}\Pr(|X| > k) &= 1 - \Pr(-k < X < k) \\ &= 1 - (\Pr(X < k) - \Pr(X < -k)) \\ &= 1 - \Pr(X < k) + \Pr(X < -k) \\ &= 2 \times \Pr(X < -k) \\ &= 2\Phi(-k), \end{aligned} \tag{3.2}$$

where $\Phi(x)$ is the (standard) normal cumulative distribution function (CDF). An observation is generally considered extreme-valued if $k \geq 3$. For a sample size of n and any normal distribution with parameters μ, σ^2, the expected number of extreme $k\sigma$ values $E[k\sigma]$ is simply

$$\begin{aligned} E[k\sigma] &= n \times (1 - \Pr(-k\sigma < X < k\sigma)) \\ &= n \times 2\Phi(-k; \mu, \sigma). \end{aligned} \tag{3.3}$$

For example, if $k = 3$, $\mu = 0$, $\sigma = 1$ and $n = 10000$, then $E[3\sigma] = 0.002699 \times 10000 \approx 27$ samples are expected to be extreme values. If $k = 4$ then $E[4\sigma] \approx 0.63$.

The performance of outlier detection methods is thus evaluated on the basis of observed versus expected number of (non-outlier) extreme values that remain as a result of applying the method. To calculate extreme (non-outlier) observations after outlier detection, the mean and variance parameters of the sample distribution are estimated using maximum likelihood (ML) (Casella, 1990). A value x is then considered to be extreme if $\frac{x-\bar{\mu}}{s}$ falls outside the range $[-k, k]$, where $\bar{\mu}$ and s are ML estimates of μ and σ. Again, the estimation process is repeated independently for each time window, wherein traffic conditions are assumed to be stationary.

The evaluation is performed for two types of outlier detection method: threshold and box plot based. Labels "Tukey," "Kimber" and "Schwertman" represent the two-pass method in three variants of the box plot rule: original rule by Tukey (1977), Kimber's rule for asymmetrical distributions (Kimber, 1990) and Schwertman/de Silva's method that accounts for sample size based on an external error rate r (Schwertman et al., 2004). For comparison, two constant threshold methods are included. The first threshold is conservative with lower and upper bound values of 5 and 140 km/h, respectively, and the second threshold is calculated from the free flow speed of that route.

Table 3.15 shows the evaluation results obtained the aforementioned outlier detection methods, across five distinct OD pairs, for $k = 3, 4, 5$. To evaluate the method's effectiveness across various road contexts, the five pairs differ in relation to their daily observation counts and free flow speeds (also shown). Overall, the two-pass method shows a reduction of extreme values to a value which better approximates that expected by the normal distribution. Even when a less conservative constant threshold is applied, the proportion of observed extreme values exceeds the expected proportion by 1 to 2 orders of magnitude.

In comparison, the performance difference between the three variants of the box plot method is small. The Schwertman/de Silva method does not appear to lead to visibly better results in exchange for the added computational cost. Furthermore, using $k_3 \geq 4.0$ for Tukey's and Kimber's rule tends to lead to better approximations than $k_3 = 3$. For Kimber $k_1 = 2$ and $k_3 = 4$, outlier rates for all five OD pairs are [0.9, 8.4, 7.7, 7.7, 10.4] (in percentage). This suggests that the box plot method can be generally applied below its breaking point of 25%, particularly in the two-step method where a significant proportion of outliers is already captured at the first step using the constant threshold. The results also demonstrate the importance of filtering invalid travel steps prior to performing any data aggregation operations.

3.4.3 Treatment of invalid travel steps

As listed in Table 3.12, two types of treatment are applied at this stage. High-valued outliers are treated via trip labelling, while the observations associated with duplicate

Table 3.15: Evaluation of two types of outlier detection method: threshold based and box plot based. Within box plot based methods, Tukey's rule is tested along two variants, Kimber and Schwertman/de Silva, each for three different parameter combinations (k_1 and k_3 are varied in the case of Tukey and Kimber, and r in the case of Schwertman). For each method, we calculate the observed and the expected number of data points within a 3, 4 and 5-sigma distance to the sample mean (obtained within each observation period) – columns $n\sigma$ and $E(n\sigma)$ respectively. Columns $p(n\sigma)$ exhibit the ratio of observed to expected values. Proportion values close to 1 indicate good agreement between the observed expected number of n-sigma observations. Data collected from 28 days, between 7am and 6pm at a 15-minute time resolution (a maximum of 1344 input intervals).

x_1	x_2	dist (km)	v_f (km/h)	AADT (veh/day)	method	count (intervals)	outlier rate	3σ	E(3σ)	p(3σ)	4σ	E(4σ)	p(4σ)	5σ	E(5σ)
016	255	5.2	82.4	4.2K	Threshold $F_1 = 3$, $F_3 = 140$ (km/h)	1325	0.0437	2548	254.9	10.00	482	5.98	80.59	63	0.054
					Threshold $F_1 = \frac{v_f}{5}$, $F_3 = \frac{3\,v_f}{2}$	1325	0.0846	1641	244.0	6.73	465	5.72	81.22	37	0.052
					Tukey $k_1 = 1.5$, $k_3 = 3.0$	1323	0.1068	105	238.1	0.44	1	5.59	0.18	0	0.051
					Tukey $k_1 = 2.0$, $k_3 = 4.0$	1324	0.1044	237	238.7	0.99	4	5.60	0.71	0	0.051
					Tukey $k_1 = 2.0$, $k_3 = 5.0$	1324	0.1023	383	239.3	1.60	9	5.61	1.60	0	0.051
					Kimber $k_1 = 1.5$, $k_3 = 3.0$	1323	0.1100	77	237.2	0.32	1	5.57	0.18	0	0.050
					Kimber $k_1 = 2.0$, $k_3 = 4.0$	1323	0.1062	179	238.3	0.75	5	5.59	0.89	1	0.051
					Kimber $k_1 = 2.0$, $k_3 = 5.0$	1323	0.1044	285	238.7	1.19	10	5.60	1.79	1	0.051
					Schwertman/de Silva $r = 0.05$	1321	0.1180	84	235.1	0.36	3	5.52	0.54	1	0.050
					Schwertman/de Silva $r = 0.01$	1322	0.1144	90	236.1	0.38	5	5.54	0.90	1	0.050
					Schwertman/de Silva $r = 0.001$	1324	0.1112	104	236.9	0.44	6	5.56	1.08	1	0.050

Continued on the following page

x_1	x_2	dist (km)	v_f (km/h)	AADT (veh/day)	method	count (intervals)	outlier rate	3σ	E(3σ)	p(3σ)	4σ	E(4σ)	p(4σ)	5σ	E(5σ)
021	022	1.1	44.1	4.3K	Threshold $F_1 = 3$, $F_3 = 140$ (km/h)	1339	0.0535	1779	258.5	6.88	614	6.07	101.23	83	0.055
					Threshold $F_1 = \frac{v_f}{5}$, $F_3 = \frac{3\,v_f}{2}$	1322	0.0880	835	249.1	3.35	160	5.84	27.38	10	0.053
					Tukey $k_1 = 1.5$, $k_3 = 3.0$	1337	0.0813	156	250.9	0.62	4	5.89	0.68	1	0.053
					Tukey $k_1 = 2.0$, $k_3 = 4.0$	1337	0.0774	276	252.0	1.10	11	5.91	1.86	1	0.054
					Tukey $k_1 = 2.0$, $k_3 = 5.0$	1337	0.0757	369	252.5	1.46	11	5.92	1.86	1	0.054
					Kimber $k_1 = 1.5$, $k_3 = 3.0$	1337	0.0866	139	249.5	0.56	4	5.85	0.68	1	0.053
					Kimber $k_1 = 2.0$, $k_3 = 4.0$	1337	0.0799	246	251.3	0.98	8	5.90	1.36	1	0.053
					Kimber $k_1 = 2.0$, $k_3 = 5.0$	1337	0.0777	343	251.9	1.36	11	5.91	1.86	1	0.053
					Schwertman/de Silva $r = 0.05$	1335	0.0997	186	245.9	0.76	10	5.77	1.73	1	0.052
					Schwertman/de Silva $r = 0.01$	1335	0.0931	193	247.7	0.78	9	5.81	1.55	1	0.053
					Schwertman/de Silva $r = 0.001$	1336	0.0879	205	249.1	0.82	11	5.85	1.88	2	0.053
073	089	1.9	41.9	6.9K	Threshold $F_1 = 3$, $F_3 = 140$ (km/h)	1343	0.0204	3168	419.8	7.55	333	9.85	33.81	21	0.089
					Threshold $F_1 = \frac{v_f}{5}$, $F_3 = \frac{3\,v_f}{2}$	1302	0.0779	2027	395.1	5.13	73	9.27	7.87	6	0.084
					Tukey $k_1 = 1.5$, $k_3 = 3.0$	1343	0.0859	692	391.7	1.77	21	9.19	2.29	0	0.083
					Tukey $k_1 = 2.0$, $k_3 = 4.0$	1343	0.0771	1057	395.5	2.67	48	9.28	5.17	4	0.084
					Tukey $k_1 = 2.0$, $k_3 = 5.0$	1343	0.0704	1396	398.3	3.50	42	9.35	4.49	3	0.085
					Kimber $k_1 = 1.5$, $k_3 = 3.0$	1343	0.0887	594	390.5	1.52	21	9.16	2.29	0	0.083
					Kimber $k_1 = 2.0$, $k_3 = 4.0$	1343	0.0793	966	394.5	2.45	43	9.26	4.65	4	0.084
					Kimber $k_1 = 2.0$, $k_3 = 5.0$	1343	0.0729	1275	397.3	3.21	36	9.32	3.86	2	0.084
					Schwertman/de Silva $r = 0.05$	1343	0.0930	791	388.7	2.04	61	9.12	6.69	6	0.083
					Schwertman/de Silva $r = 0.01$	1343	0.0906	794	389.7	2.04	69	9.14	7.55	8	0.083
					Schwertman/de Silva $r = 0.001$	1343	0.0876	798	391.0	2.04	74	9.17	8.07	11	0.083

Continued on the following page

x_1	x_2	dist (km)	v_f (km/h)	AADT (veh/day)	method	count (intervals)	outlier rate	3σ	E(3σ)	p(3σ)	4σ	E(4σ)	p(4σ)	5σ	E(5σ)
209	054	1.5	64.6	12.6K	Threshold $F_1 = 3$, $F_3 = 140$ (km/h)	1324	0.0033	1801	732.4	2.46	1105	17.18	64.31	734	0.156
					Threshold $F_1 = \frac{v_f}{5}$, $F_3 = \frac{3\,v_f}{2}$	1310	0.0266	1194	715.3	1.67	370	16.78	22.05	177	0.152
					Tukey $k_1 = 1.5$, $k_3 = 3.0$	1324	0.0132	538	725.1	0.74	8	17.01	0.47	1	0.154
					Tukey $k_1 = 2.0$, $k_3 = 4.0$	1324	0.0089	868	728.3	1.19	40	17.09	2.34	1	0.155
					Tukey $k_1 = 2.0$, $k_3 = 5.0$	1324	0.0083	934	728.7	1.28	55	17.10	3.22	2	0.155
					Kimber $k_1 = 1.5$, $k_3 = 3.0$	1324	0.0181	416	721.5	0.58	16	16.93	0.95	0	0.153
					Kimber $k_1 = 2.0$, $k_3 = 4.0$	1324	0.0111	741	726.7	1.02	40	17.05	2.35	4	0.154
					Kimber $k_1 = 2.0$, $k_3 = 5.0$	1324	0.0097	848	727.7	1.17	58	17.07	3.40	5	0.155
					Schwertman/de Silva $r = 0.05$	1324	0.0311	601	712.0	0.84	44	16.70	2.63	5	0.151
					Schwertman/de Silva $r = 0.01$	1324	0.0249	637	716.5	0.89	60	16.81	3.57	7	0.152
					Schwertman/de Silva $r = 0.001$	1324	0.0194	714	720.5	0.99	74	16.91	4.38	7	0.153
221	111	1.9	39.1	1.0K	Threshold $F_1 = 3$, $F_3 = 140$ (km/h)	1155	0.0596	238	64.0	3.72	45	1.50	29.96	1	0.014
					Threshold $F_1 = \frac{v_f}{5}$, $F_3 = \frac{3\,v_f}{2}$	1150	0.0705	129	63.3	2.04	12	1.48	8.08	0	0.013
					Tukey $k_1 = 1.5$, $k_3 = 3.0$	1135	0.0947	23	61.6	0.37	0	1.45	0.00	0	0.013
					Tukey $k_1 = 2.0$, $k_3 = 4.0$	1140	0.0835	44	62.4	0.71	0	1.46	0.00	0	0.013
					Tukey $k_1 = 2.0$, $k_3 = 5.0$	1142	0.0813	56	62.5	0.90	0	1.47	0.00	0	0.013
					Kimber $k_1 = 1.5$, $k_3 = 3.0$	1118	0.1135	26	60.4	0.43	1	1.42	0.71	0	0.013
					Kimber $k_1 = 2.0$, $k_3 = 4.0$	1128	0.0987	44	61.4	0.72	0	1.44	0.00	0	0.013
					Kimber $k_1 = 2.0$, $k_3 = 5.0$	1131	0.0951	50	61.6	0.81	0	1.45	0.00	0	0.013
					Schwertman/de Silva $r = 0.05$	1074	0.1589	34	57.3	0.59	1	1.34	0.74	0	0.012
					Schwertman/de Silva $r = 0.01$	1093	0.1393	40	58.6	0.68	1	1.37	0.73	0	0.012
					Schwertman/de Silva $r = 0.001$	1099	0.1262	49	59.5	0.82	1	1.40	0.72	0	0.013

and low-valued outliers are discarded. Before any of the two treatments is applied, data is transformed from a point-wise representation of movement (observation at location) to a step-wise equivalent (movement between two locations). These three stages are described in the order that they are executed: point to step-wise data transformation, elimination of duplicates and low outlier travel times, and trip labelling (trip labelling comes last to prevent gaps caused by data removal).

3.4.3.1 Point and step-wise data representations

Consider n observations of a vehicle across an ANPR network. The camera locations visited by the vehicle form a time-ordered sequence $(l_1, l_2, \ldots, l_n)$ such that $t_i < t_{i+1} \forall i \in 1 \ldots n-1$, where t_i is the observation time at location l_i. Step i is associated with OD pair (l_i, l_{i+1}), departure and arrival times (t_i, t_{i+1}), and travel time $t_{i+1} - t_i$. The vehicle's first step is from l_1 to l_2, the second step is from l_2 to l_3 and so forth. The complete travel sequence can thus be re-written as a sequence of $n-1$ steps, denoted $\boldsymbol{l} = [(l_1, l_2), (l_2, l_3), \ldots, (l_{n-1}, l_n)]$, with observation times $\boldsymbol{t} = [(t_1, t_2), (t_2, t_3), \ldots, (t_{n-1}, t_n)]$ and travel times $\boldsymbol{tt} = [(t_2 - t_1), (t_3 - t_2), \ldots (t_n - t_{n-1})]$.

The point-wise to step-wise data transformation, depicted in Table 3.16 is computationally advantageous for purposes of data manipulation and interpretation. In a pair-wise

Table 3.16: Point and step-wise data representations of an observation sequence (sampled from one vehicle).

point-wise representation		
vehicle	l	t
287096	107	2018-01-04 14:44:46
	154	2018-01-04 14:48:28
	093	2018-01-04 14:53:02
	093	2018-01-04 16:00:16
	073	2018-01-04 16:08:22
	089	2018-01-04 16:11:38
	057	2018-01-04 16:15:09

step-wise representation							
vehicle	l_1	l_2	t_1	t_2	tt (s)	v (km/h)	distance (km)
287096	107	154	2018-01-04 14:44:46	2018-01-04 14:48:28	222	34.6	2.1
	154	093	2018-01-04 14:48:28	2018-01-04 14:53:02	274	58.6	4.5
	093	093	2018-01-04 14:53:02	2018-01-04 16:00:16	4034	-	-
	093	073	2018-01-04 16:00:16	2018-01-04 16:08:22	486	36.7	5.0
	073	089	2018-01-04 16:08:22	2018-01-04 16:11:38	196	34.7	1.9
	089	057	2018-01-04 16:11:38	2018-01-04 16:15:09	211	32.8	1.9

representation, the three vectors $\boldsymbol{l}$, $\boldsymbol{t}$ and $\boldsymbol{tt}$ all have the same length, making it easier to encode the data in tabular format and implement the "split-apply-combine" computational strategy for data analysis (Wickham, 2011). The transformation can be conveniently implemented by applying a `lead` or `lag` function (a function which shifts a vector forward or backwards by one position) to the time and location columns (Wickham et al., 2021). The distance and speed columns are then obtained by joining against the camera pairs dataset on columns (l_1, l_2). Altogether, the sequences $\boldsymbol{l}$, $\boldsymbol{t}$ and $\boldsymbol{tt}$ are of interest because they encapsulate the travel history of each vehicle: what routes were travelled, when, and how fast.

3.4.3.2 Duplicate observations

Duplicate observations and low travel time outliers are treated by way of elimination. Duplicates occur when two or more successive camera observations of one vehicle are issued within a short period of time. Concretely, a duplicate is a step that meets two conditions: (i) the origin location equals the destination location, and (ii) travel time is below a permissible value. These are formally defined as:

$$\begin{aligned} l_i &= l_{i+1} \\ t_{i+1} - t_i &< T_{\text{dup}} \end{aligned} \tag{3.4}$$

where T_{dup} is a maximum permitted travel time. Condition (ii) is used to to distinguish a duplicate observation from a new vehicle trip. A new vehicle trip occurs if a vehicle visits the same camera twice in a row with significant time in between observations, whereas a duplicate occurs within a few seconds or minutes of the first observation (longer times can originate when the vehicle is stopped at a traffic light or in congested traffic). Treatment of duplicates, illustrated in Table 3.17, is thus achieved by discarding travel steps identified using Equation 3.4, for $T_{\text{dup}} = 300$ s.

The value of T_{dup} should be large enough to accommodate long vehicle stops within the camera's detection region, but not large enough that it would allow a vehicle to circle around to the same location within that time (e.g. due to a routing error). Whether traffic is prone to long stops at a location or can easily loop around to the same point will vary greatly on location, traffic conditions and the surrounding road network. Still, it is not so important to obtain a route-specific precise value of T_{dup}, but simply a sufficiently accurate threshold capable of capturing the vast majority of flagrant duplicate observations. Therefore, we expect T_{dup} to be in the order of tens of seconds or a few minutes. This expectation is confirmed by inspection of the empirical distribution of the time between

Table 3.17: Example of a vehicle trip with two duplicate observations at location *138*: before and after treatment.

l_1	l_2	t_1	t_2	tt (s)	v (km/h)	v_f (km/h)	duplicate	outlier
before treatment								
226	138	09:16:43	09:17:23	40	40.9	42.7		
138	138	09:17:23	09:17:40	17	-	-	x	
138	138	09:17:40	09:17:43	3	-	-	x	
138	009	09:17:43	09:19:00	77	27.1	28.2		
009	092	09:19:00	11:12:44	6824	0.9	-		x
after treatment								
226	138	09:16:43	09:17:23	40	40.9	42.7		
138	009	09:17:43	09:19:00	77	27.1	28.2		
009	092	09:19:00	11:12:44	6824	0.9	-		x

consecutive observations at a location, shown in Figure 3.10 for three distinct locations: the kernel density estimates begin to level off after approximately 90 seconds, only to peak again later at the 24h mark. Based on these examples, a value of T_{dup} between 2 to 10 minutes is suggested.

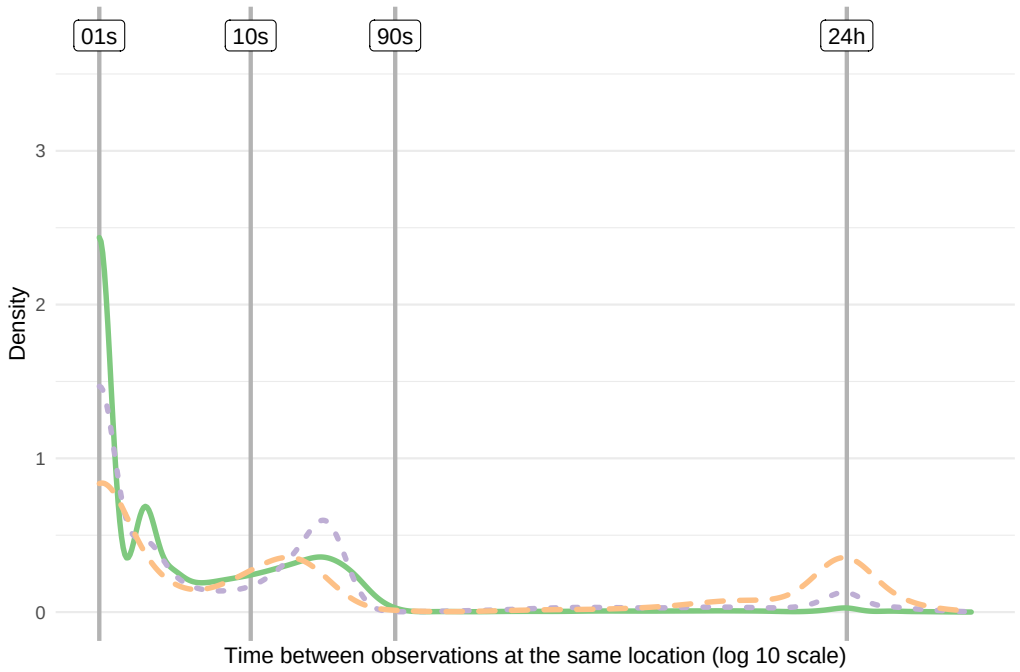


Figure 3.10: Kernel density estimates of the time between two consecutive observations at the same location (shown in log 10 scale), for three distinct locations (chosen randomly amongst the top 50 routes with most such observations).

3.4.3.3 Low value travel time outliers

Low-value travel time outliers (or equivalently high-value speed outliers) originate from camera detection errors: a mistaken vehicle recognition causes an extra observation to appear in its trip history vector. Together with high-value outliers, low-value outliers are identified through Tukey's rule or an alternative outlier detection method. Following identification, treatment is achieved by discarding the steps labelled as low-value outliers. However, unlike duplicate observations, step elimination causes a disconnect in the trip history vector: the destination of the previous step no longer matches the origin of the next step. Treatment should therefore seek to rectify this issue, by updating the row above or below the removed step, in order to guarantee data integrity (otherwise impacting subsequent analysis, namely of trip sequences).

Table 3.18a depicts the treatment of a low-value outlier in a vehicle trip composed of three steps. Note the discrepancy between the two remaining steps after the low-value outlier step is removed: the destination of the first step (value `124`) no longer matches the origin location of the second step (value `88`). Two possible corrections are shown: treatment 1 updates the step above the original outlier step, implying that the vehicle never drove past location `124`; and treatment 2 updates the step below the original outlier step, effectively erasing the observation at location `88`. Because either correction type (up or down) implies updating the location and time of the origin/destination observation, all of the affected step attributes, including travel time, distance and speed, must be recomputed.

A consequence of step updating is that steps previosuly considered valid may no longer be considered so. Therefore, steps updated as per low value outlier treatment must be checked again for validity, i.e. duplicate and outlier, and treated accordingly. This implies that full treatment of low-value outliers can only be achieved iteratively: steps updated at the end of a treatment iteration are re-classified as valid/invalid, and low-value outliers within these treated in the next iteration. In practice, to re-evaluate steps as valid/invalid, it suffices to compare the new value of travel time/speed with the already calculated Tukey fences, as opposed to performing the entire outlier detection process over again. Full low-value treatment is then obtained when no more low-value outliers are found. Note that this does not include high-value outliers, which are treated only once afterwards, via trip labelling.

In the example of Table 3.18a, both corrections lead to an updated step that was considered valid. But this is not necessarily always the case – one correction might lead to a valid step while the other does not, or both might result in invalid steps, as the example in Table 3.18b shows. In the former case, the correction that yields a valid step might be preferred over the one that doesn't, while it is not so clear what correction is preferable in

Table 3.18: Two examples of the treatment of invalid steps characterised by low travel time outliers. Two step updating schemes are shown for each example: update the step (1) above or (2) below.

(a) Example of low-value travel time resulting in very high travel speed. Both correction schemes yield valid steps.

l_1	l_2	t_1	t_2	tt (s)	v (km/h)	v_f (km/h)	dist (km)	duplicate	low outlier	high outlier
before treatment										
073	124	10:59:35	11:04:57	322	25.0	-	2.2			
124	088	11:04:57	11:05:10	13	207.6	84.7	0.7		x	
088	061	11:05:10	11:07:19	129	73.4	77.8	2.6			
after step removal										
073	124	10:59:35	11:04:57	322	25.0	-	2.2			
088	061	11:05:10	11:07:19	129	73.4	77.8	2.6			
(1) after correction by updating the previous step										
073	088	10:59:35	11:05:10	335	22.8	23.4	2.1			
088	061	11:05:10	11:07:19	129	73.4	77.5	2.6			
(2) after correction by updating the next step										
073	124	10:59:35	11:04:57	322	25.0	-	2.2			
124	061	11:04:57	11:07:19	142	85.6	82.8	3.4			

(b) Example of low-value travel time outlier caused by timestamp clash. Neither correction type yields a valid step.

l_1	l_2	t_1	t_2	tt (s)	v (km/h)	v_f (km/h)	dist (km)	duplicate	low outlier	high outlier
before treatment										
196	180	19:52:38	19:56:56	258	0.3	-	0.0			x
180	196	19:56:56	19:56:56	0	Inf	25.1	0.9		x	
196	035	19:56:56	19:57:32	36	47.3	54.3	0.5			
after step removal										
196	180	19:52:38	19:56:56	258	0.3	-	0.0			x
196	035	19:56:56	19:57:32	36	47.3	54.3	0.5			
(1) after correction by updating the previous step										
196	196	19:52:38	19:56:56	258	-	-	-	x		
196	035	19:56:56	19:57:32	36	47.3	52.6	0.5			
(2) after correction by updating the next step										
196	180	19:52:38	19:56:56	258	0.3	60.7	0.0			x
180	035	19:56:56	19:57:32	36	45.0	47.2	0.5			

the latter case. Applying a different correction mechanism to all low-value outliers, however, comes at the cost of extra complexity and computation. A simple correction scheme is hence to perform the correction always in the same direction, either up or down, and then solve any errors by way of iteration, as described above. While such strategy might be suboptimal, it is simple to implement as it is not necessary to compare and choose between multiple possible correction outcomes. Hence, we consider the development of a more complex correction strategy to be outside the scope of this work (in fact, a better strategy should take into account the full trip sequence when identifying and treating outliers).

3.4.3.4 Trip labelling

Trip labelling is the process of grouping all observations pertaining to the same trip under a single label. It is executed after outlier observations have been identified and is implemented using a cumulative trip counter vector for each vehicle. The trip vector, sorted by time, is incremented every time a *valid* step is preceded by an *invalid* step (recall that, at this stage, the remaining invalid steps are of high outlier category since other types of invalid steps have already been treated). Where valid steps are preceded by invalid steps they represent the first step of a new trip.

Treatment of invalid steps is perfomed not by direct elimination but by invalidating its spatial and temporal components. An invalid spatial step with value (l_i, l_{i+1}) is changed to $(l_i, \varnothing)$, where $\varnothing$ is encoded as Not a Number (NaN) to avoid accidental data manipulation after treatment. Equivalently, the temporal step (t_i, t_{i+1}) becomes $(t_i, \varnothing)$ and all measures of travel time, average speed and distance are also set to NaN.

This corrective procedure, exemplified in Table 3.19 , effectively dissassociates the outlier observation from the route – the OD pair (l_i, l_{i+1}) is no longer observed even though the data row is not eliminated. The modified data row represents the final step of a trip whose true destination and route is unknown and unobserved. The utility of keeping unknown destination rows is two fold. First, it facilitates the query "what location was observed last?" via the filter operation `destination == NaN`, without modifying the underlying data model (useful, for instance, in identifying cameras "at the edge," i.e. cameras that are often observed last). Second, it allows trips composed of a single observation to remain in the dataset (which can be useful for certain analyses). In single-observation trips there is no downstream vehicle matching and therefore no observed travel time. As Table 3.20 indicates, almost half of all registered trips contain only a single observation.

Table 3.19: Effect of trip labelling treatment on column and row composition (example for one vehicle). NaN values are represented by character '-'.

before treatment								
l_1	l_2	t_1	t_2	tt (s)	v (km/h)	v_f (km/h)	duplicate	outlier
217	154	12:43:33	12:49:18	345	41.3	43.6		
154	199	12:49:18	12:51:41	143	65.0	62.8		
199	000	12:51:41	15:10:27	8326	0.0	-		x
000	073	15:10:27	15:13:19	172	55.5	60.4		
073	089	15:13:19	15:16:43	204	33.3	43.4		
089	057	15:16:43	15:19:28	165	41.9	53.6		
057	078	15:19:28	15:21:10	102	71.4	77.5		
078	078	15:21:10	15:21:11	1	-	-	x	
after treatment and trip labelling								
l_1	l_2	t_1	t_2	tt	v	v_f	trip	j
-	217	-	12:43:33	-	-	-	1	0
217	154	12:43:33	12:49:18	345	41.3	43.6	1	1
154	199	12:49:18	12:51:41	143	65.0	62.8	1	2
199	-	12:51:41	-	-	-	-	1	3
-	000	-	15:10:27	-	-	-	2	0
000	073	15:10:27	15:13:19	172	55.5	60.4	2	1
073	089	15:13:19	15:16:43	204	33.3	43.4	2	2
089	057	15:16:43	15:19:28	165	41.9	53.6	2	3
057	078	15:19:28	15:21:10	102	71.4	77.5	2	4
078	-	15:21:11	-	-	-	-	2	5

Table 3.20: Trip length frequency of occurrence during the month of March 2018 in the Tyne and Wear dataset.

Trip length	1	2	3	4	5	6	7	8
count (hundreds of thousands)	108.8	52.2	28.4	15.6	8.1	4.2	2.2	1.2
proportion of total	0.489	0.234	0.128	0.070	0.036	0.019	0.010	0.005
cumulative proportion	0.489	0.723	0.851	0.921	0.957	0.976	0.986	0.991

Finally, Algorithm 1 summarises the entire sequence of treatments applied in Stage 2, responsible for converting wrangled number plate detection data into trip data ready for analysis.

3.4.3.5 Observed proportions of invalid steps

For reference, it is of interest to compare the proportion of steps classified as valid and invalid after treatment is applied. Table 3.21 shows the total proportion of invalid steps

Algorithm 1 Trip identification

Input

D_{wrangled} wrangled number plate dataset
D_{pairs} camera pairs dataset with calculated spatial route lengths

Output

D_{trips} trips dataset

1. Transform the input dataset D_{wrangled} from point to point-to-point format (using a lead/lag function).
2. Calculate travel time and merge the camera pairs distances dataset D_{pairs} to obtain route distance and calculate average vehicle speed.
3. Identify duplicates.
4. Identify outliers using a chosen method (e.g. Tukey's rule).
5. Treat duplicates.
6. If there are any low-value outliers, treat them (discard and update the steps above/below) and go to 2.
7. Treat high-value outliers through trip labelling.
8. Optionally augment each vehicle trip with two dummy steps, placed at the start and end each journey, representing travel from and to unknown locations, respectively.

Table 3.21: Proportion of invalid steps (percentage) by type, out of the total number of recorded vehicle steps in the 2018 Tyne and Wear dataset.

duplicates	outliers (low travel time)	outliers (high travel time)	invalid steps (total)	total observation count (millions)
4.6%	0.6%	38.9%	44.0%	619.8

Table 3.22: Proportion of invalid steps (percentage) across OD pairs in the 2018 Tyne and Wear dataset. Summary statistics are shown separately for OD pairs of the form $l_i = l_{i+1}$ (a step is either a duplicate or a new trip observation) and $l_i \neq l_{i+1}$ (a step is either valid or a low/high-valued outlier).

	Cross-group percentiles (OD pairs)					
Proportion of	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	total[4]
duplicates[1]	3.2	14.3	50.0	72.9	87.7	48.6
outliers (all pairs)[2]	32.1	50.0	64.2	75.6	85.6	38.2
outliers (top 500 pairs)[3]	6.1	10.1	17.5	74.5	90.9	31.4

[1] Calculated for $l_i = l_{i+1}$ OD pairs (same origin and destination camera). Non-duplicate steps are high-valued outliers treated through trip labelling.
[2] Calculated for all 50556 $l_i \neq l_{i+1}$ OD pairs.
[3] Calculated for the top 500 $l_i \neq l_{i+1}$ OD pairs with highest observation count.
[4] Calculated for the whole data subset.

(percentage), across the different types of invalid steps, obtained for the 2018 Tyne and Wear dataset. For $T^{\text{dup}} = 10$ min, duplicates comprise roughly 4.6% of all vehicle steps. Applying the kimber rule, with $k_1 = 2$ and $k_3 = 4$, resulted in 0.6% of low-valued travel time outliers and 38.9% of high-valued outliers treated through trip labelling. In total, 44% of almost 620 million observation steps were classified as invalid and treated accordingly.

To understand how these figures vary across groups, Table 3.22 displays percentile statistics of the proportion of invalid steps obtained for each OD pair. In OD pairs of the type $l_i = l_{i+1}$, duplicates occur much more commonly than high-valued outliers in some cameras than in others, resulting in a distribution that is approximately uniform in the range [0,1]. In total, 48.6% of $l_i = l_{i+1}$ type observations were classified as duplicates (and the rest as high-valued outliers). For OD pairs of type $l_i \neq l_{i+1}$, the proportion of invalid steps varies greatly across the whole set of possible OD pairs, indicating that a high proportion of OD pairs issue mostly invalid observations (90% of OD pairs have a proportion of invalid steps above 32.1%). When considering the top 500 ODs based on total observation count, the distribution becomes strongly left-skewed, indicating that it mostly contains OD pairs whose majority of observations are valid (50-th percentile is at 17.5% compared to 64.2% for all OD pairs).

3.5 Stage 3: Aggregation

In the third and final stage of the pipeline, vehicle *trips* are consumed to produce aggregate vehicle data, known as flows (as illustrated in Table 3.10). Besides sensor capability and choice/availability of traffic variables (discussed throughout Section 2.1), researchers are confronted with several problems that may affect the quality of aggregate data prior to analysis. In particular, researchers will generally have to decide on: (a) the size of time intervals; (b) how vehicles are sampled within each time interval (sampling is not straightforwad since observations have a duration); (c) how to build the ANPR network by selecting routes for analysis; (d) how to handle missing data. Cases (a) and (b) are required in order to obtain flow data, while (c) and (d) arise once flow data is obtained and are employed to improve its quality. In this section, each question is examined individually and new solutions are offered if the existing approaches are found to be lacking or would benefit from an alternative method.

3.5.1 Choice of aggregation interval

Traffic streams are commonly modelled as time-discrete stochastic processes (Section 2.2.2.1). Traffic variables are measured at equally spaced time points or intervals, forming

a time-indexed sequence of observations, i.e. a time series (E. Vlahogianni & Karlaftis, 2011). At each time step, measurements are taken by way of aggregation: individual observations are replaced with one or more summary statistics.

One consequence of data aggregation is information loss, the degree of which is largely determined by the frequency at which measurements are taken, i.e. the temporal resolution of aggregated data (W. A. V. Clark & Avery, 1976). The choice of temporal resolution, unless limited by the end application or measurement device, is therefore of practical importance in traffic forecasting (and generally in statistical analysis), as it affects the intrinsic properties of aggregate traffic data (Hurdle et al., 1997; E. I. Vlahogianni et al., 2006).

Several studies have investigated how temporal aggregation affects traffic data and the performance of traffic forecasting models. A recognised result is that longer aggregation intervals (> 10 min) produce smoother time series compared to shorter ones (< 2 min) (E. Vlahogianni & Karlaftis, 2011). For example, J. Guo et al. (2007) found that vehicle flow rates estimated from loop detectors are most noisy at 1 minute collection intervals, but stabilise when the aggregation period is increased above 10 minutes. Consequently, earlier guidelines to practitioners reflect findings that longer aggregation periods improve prediction accuracy (Oh et al., 2005) – e.g. the Highway Capacity Manual (HCM) recommends 15-minute aggregation intervals (Transportation Research Board, 2000).

A concern with lower levels of temporal resolution is that they can eliminate variation in data and alter the temporal structure and statistical properties of time series data, such as non-stationarity and non-linearity (E. Vlahogianni & Karlaftis, 2011). Hurdle et al. (1997) indicates that 15-minute flow data does not necessarily represent uniform traffic conditions throughout the whole interval and that significant variations can occur within a single interval. E. I. Vlahogianni et al. (2006) shows that the loss of time variation in data can produce unrealistic models of traffic flow. The data aggregation problem is not unique to traffic forecasting and extends to other applications of time series data (Rossana & Seater, 1995).

The choice of aggregation interval is sometimes constrained by the application, for example, high adaptive signal control requires high resolution data in order to generate 10 to 30 seconds ahead forecasts (W. Li et al., 2020). E. I. Vlahogianni et al. (2014) reviews numerous traffic forecasting methodologies developed between 2004 and 2013, with prediction steps ranging from below 1 minute to an hour. Despite the variety in data collection (e.g. loop detectors, simulation, bluetooth, GPS), methodological approach (e.g. time series, function aproximation, optimisation, clustering) and model choice (e.g. statistical, neural network), the authors point to the lack of a complete and comprehensive examination of temporal aggregation bias in traffic data and lack of standardised method of determining the level of temporal resolution appropriate for analysis.

Specific to ANPR data, different aggregation intervals have been reported: 5-min for outlier detection (S. D. Clark et al., 2002) and travel time prediction Haworth & Cheng (2012), 2-min for incident detection (X. Li et al., 2013) and 15-min for travel time reliability assessment (Yang et al., 2010). R. Li & Rose (2011) tested different aggregation periods in ANPR data at highways and found that a 10 min window generally provides at least 20 observations of travel time to allow for reliable estimation of travel time average and variability measures, even during less congested periods of the day. It is unclear whether the use of shorter aggregation intervals (1-2 min and < 1 min) is admissible or discouraged in ANPR data due to, for example, summary statistics computed from a small number of vehicle samples.

Ultimately, ANPR offers the practitioner the flexibility to choose the desired level of temporal resolution. Several factors should weigh in the choice of aggregation interval: (i) the level of analysis (how fine-grained predictions/inferences are required to be); (ii) the relationship between model complexity/performance and aggregation interval (more complex models are required for shorter aggregation intervals in order to generate predictions at the same performance level); (iii) sampling bias (shorter intervals implies that summary statistics are based on smaller vehicle samples), (iv) road type (shorter intervals may be suitable for strongly frequented roads like highways but longer ones may be required for less travelled roads such as local arterials).

3.5.2 Vehicle sampling at intervals

To calculate summary statistics, the data samples within each time interval are composed of observations that fall within that interval. While in point data observations fall within a single time interval, in point-to-point data vehicle observations can span multiple time intervals. Hence, if an observation spans multiple time intervals then it can be associated with and be re-used in as many time intervals as the number of intervals it spans across.

Figure 3.11 illustrates the vehicle sampling phenomenom for three vehicle steps spanning three 5-minute intervals. For each vehicle, it is shown its arrival time at the origin (empty square) and destination (filled circle) locations, and the period of time during which the vehicle is travelling along the route (thick horizontal line). Clearly, arrival times fall within a single time interval (point measurement), but travel periods can cross the specified time boundaries (vehicles B and C). Here, one can question whether the first interval T_1 should include the travel time of vehicle B, or whether T_3 should include vehicles B and C. If so, then vehicle B is a sample element not of one but three intervals traffic stream measurements.

As seen in Table 3.1 (Section 3.1), authors do not commonly clarify their approach in vehicle sampling. One possible explanation is that the effect of different sampling strategies

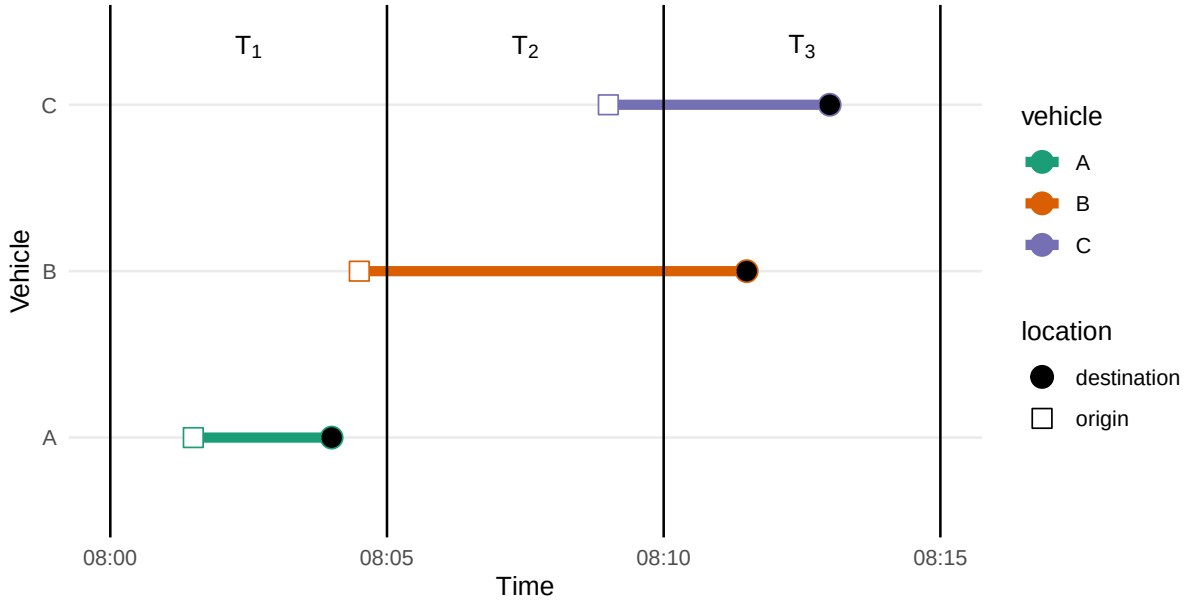


Figure 3.11: Sample of three vehicle travel steps, illustrated for two types of ANPR measurements, point on a road (squares, circles) and along a route (line).

is negligible or, alternatively, not that well understood. In order to better comprehend the effect of vehicle sampling, we define a user-parametrised vehicle *sampling* function, which associates a travel step with one or more time intervals, and investigate its effect in the resulting aggregate flow measurements. Let (t_n, t_{n+1}) denote the n-th temporal step of a vehicle, and let $T_p = [t_p, t_{p+1})$ be the time interval from t_p to t_{p+1}. Then, $g(t_n, t_{n+1}, T_p)$ is a vehicle sampling function that evaluates to zero or one for each of the mutually-exclusive cases shown in Table 3.23.

Table 3.23: List of possible cases describing the relation between a travel step (t_n, t_{n+1}) and a time interval T_p.

Case	Notation	Description	Example (Fig. 3.11)
1	$[t_n, t_{n+1}] \cap T_p = \varnothing$	the travel step happens outside T_p	T_1 and vehicle C
2	$[t_n, t_{n+1}] \subseteq T_p$	the travel step happens inside T_p	T_1 and vehicle A
3	$t_n \in T_p \wedge t_{n+1} \notin T_p$	only a portion of the travel step, that includes t_n, happens inside T_p	T_2 and vehicle C
4	$t_n \notin T_p \wedge t_{n+1} \in T_p$	only a portion of the travel step, that includes t_{n+1}, happens inside T_p	T_3 and vehicle C
5	$T_p \subset [t_n, t_{n+1}]$	the travel step encompasses T_p	T_2 and vehicle B

To accomodate different sampling strategies, the value of g for each condition $k = 1..5$ in Table 3.23 is specified by a user-chosen binary value x_k. Assuming that x_1 and x_2 are always set to 0 and 1, respectively, g can be expressed as

$$g(t_n, t_{n+1}, T_p; x_3, x_4, x_5) = \begin{cases} 0 & \text{if } [t_n, t_{n+1}] \cap T_p = \varnothing \\ 1 & \text{if } [t_n, t_{n+1}] \subseteq T_p \\ \mathrm{x}_3 & \text{if } t_n \in T_p \wedge t_{n+1} \notin T_p \\ \mathrm{x}_4 & \text{if } t_n \notin T_p \wedge t_{n+1} \in T_p \\ \mathrm{x}_5 & \text{if } T_p \subset [t_n, t_{n+1}] \end{cases}$$

subject to $t_{n+1} > t_n$ and $T_{p+1} > T_p$. x_3, x_4, x_5 are set according to one's assumptions or interpretation of vehicle movement along a route. Of the $8(2^3)$ possible combinations of values for x_3, x_4, x_5, three parametrisations of g are arguably more intuitive than the alternatives:

1. $[x_3, x_4, x_5] = [1, 0, 0]$: A travel step is associated with an interval if the vehicle arrives to the destination location within that interval.

2. $[x_3, x_4, x_5] = [0, 1, 0]$: A travel step is associated with an interval if the vehicle departs from the origin location within that interval.

3. $[x_3, x_4, x_5] = [1, 1, 1]$: A travel step is associated with an interval if the vehicle is travelling between the two locations at any point during that interval.

For each proposed parametrisation $i = 1..3$, g_i can be simplified to:

$$g_1 = g(t_n, t_{n+1}, T_p; x_3 = 1, x_4 = 0, x_5 = 0) = \begin{cases} 1 & \text{if } t_n \in T_p, \\ 0 & \text{otherwise} \end{cases}$$

$$g_2 = g(t_n, t_{n+1}, T_p; x_3 = 0, x_4 = 1, x_5 = 0) = \begin{cases} 1 & \text{if } t_{n+1} \in T_p, \\ 0 & \text{otherwise} \end{cases}$$

$$g_3 = g(t_n, t_{n+1}, T_p; x_3 = 1, x_4 = 1, x_5 = 1) = \begin{cases} 1 & \text{if } [t_n, t_{n+1}] \cap T_p \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

The third parametrisation g_3, is perhaps more intuitive than the other two. However, while the first two parametrisations lead to observations being included in only one interval, the third allows observations to be a part of several interval samples. The difference in the definition of g_3 has multiple effects: (1) it leads to increased sample sizes, (2) it increases the correlation of sample statistics taken at consecutive time intervals as they have a number of elements in common; and (3) the original observation count is "lost."

Table 3.24: Application of the proposed vehicle sampling functions to the example depicted in Fig. 3.11.

p	g	t_A	t_B	t_C	N_p
1	g_1	1	1	0	2
	g_2	1	0	0	1
	g_3	1	1	0	2
2	g_1	0	0	0	0
	g_2	0	0	0	0
	g_3	0	1	1	2
3	g_1	0	1	0	1
	g_2	0	1	1	2
	g_3	0	1	1	2

	g_1	g_2	g_3
$\sum_{p=1}^{3} N_p$	3	3	6

The third effect is illustrated in Table 3.24, built on the earlier example shown in Figure 3.11. We evaluate g_i for all $i = 1..3$ for each time period $p = 1..3$ and count the number of observations associated with each interval, and across all intervals. While the final count $\sum_p N_p$ obtained for g_1 and g_2 matches the number of distinct vehicles (3), the same is not true for g_3. On the other hand, as the aggregation window size increases, these effects become negligible because less and less travel steps will cross interval boundaries.

Given a parametrised vehicle sampling function g, we can express the vehicle count between two locations i and j, during interval T_p, denoted by $N_{\mathrm{AB}p}$, as:

$$N_{\mathrm{AB}p} = \sum_{k \in \mathbb{K}} \sum_{i=1}^{n_k - 1} \mathbb{1}[l_i^k = \mathrm{A}] \cdot \mathbb{1}[l_{i+1}^k = \mathrm{B}] \cdot g(t_i^k, t_{i+1}^k, T_p)$$

where t_i^k and l_i^k are the i-th temporal and spatial observations of vehicle k, n_k is the length of the observation vector of k and $\mathbb{K}$ is the set of observed vehicles. $\mathbb{1}[Q]$ is an indicator function that returns 1 if the logical statement Q evaluates to *true* and 0 otherwise. Similarly, we can write the formula for the mean vehicle travel time $TT_{\mathrm{AB}p}$:

$$TT_{\mathrm{AB}p} = \frac{1}{N_{\mathrm{AB}p}} \sum_{k \in \mathbb{K}} \sum_{i=1}^{n_k - 1} \mathbb{1}[l_i^k = \mathrm{A}] \cdot \mathbb{1}[l_{i+1}^k = \mathrm{A}] \cdot g(t_i^k, t_{i+1}^k, T_p) \cdot tt_i^k$$

In practice, the computation is conveniently implemented using the "split-apply-combine" computational strategy for data analysis (Wickham, 2011). Travel steps are grouped on three keys: origin location, destination location and time interval, prior to summarisation.

If g is parametrised as g_1 or g_2, then a travel step is assigned to the appropriate interval via a vectorised date floor function (Grolemund & Wickham, 2011) applied to t_i^k and t_i^k, for all i and k, respectively. If g_3 is used instead, then a travel step assigned to multiple groups is duplicated as many times as the number of groups it is assigned to. For instance, with 5 minute intervals, the travel step `[08:03:30, 08:13:30]` is duplicated three times and assigned once to each of the intervals `[08:00,08:05]`, `[08:05,08:10]`, `[08:10,08:15]`.

3.5.3 Route identification

The use of ANPR to study traffic-related phenomena is often confined to a selection of routes (camera pairs). Routes are often selected manually at the time of the analysis, or based on prior annotations by traffic experts.

We assume that cameras are not deployed at random but intentionally in order to monitor routes of interest. A route may be of interest because it is prone to congestion or plays an important role in connecting different areas of a city. As the size of a ANPR network increases, the number of possible camera pairs grows more rapidly. If we exclude camera pairs whose origin matches the destination, then a network with n cameras has $n(n-1)$ potential camera pairs (approximately quadratic growth). However, only a small fraction of these will typically be actively employed in traffic monitoring. As ANPR networks grow in size, it is therefore of interest to examine whether camera pairs that have not been labelled by humans, may in fact represent routes of interest for analysis.

Formally, we represent a ANPR network as a directed graph $G = (V, E)$, such that its node set V corresponds to the set of ANPR cameras, and E is the set of ANPR routes of interest (camera pairs). The goal of route identification is then to approximate G, or specifically the edge set E from a noisy dataset of ANPR *flows*. The simplest approach is to include a camera pair (u_1, u_2) in E if it occurs at least once in the dataset. However, errors made during trip identification will propagate to the aggregation stage and can later manifest as erroneous camera pairs. Thus, to identify routes of interest, we need criteria to establish whether a candidate edge, i.e. a camera pair observed at least once, is a valid ANPR route or not.

In defining criteria for valid routes, we consider the following volume-related metrics:

- **Traffic volume** - the observation count, summed or averaged over a length of time. A measure of traffic volume commonly used in transportation planning is the annual average daily traffic (AADT), which is simply defined as the total volume of traffic observed in a year, divided by 365 (Fu et al., 2017). In essence, busier routes are associated with higher traffic volumes.

- **Downstream rate** - the proportion of traffic that leaves an origin location l_i towards destination l_j, measured over a period of time, denoted by $r_{\text{dn}}(i, j)$. Let

$$r_{\text{dn}}(i, j) = \frac{N_{ij}}{\sum_{j=1}^{|V|} N_{ij}} \tag{3.5}$$

  where N_{ij} is the observation count between origin l_i and destination l_j, and $|V|$ denotes the cardinality of the node set V (i.e. camera count). $r_{\text{dn}}(i, j)$ can be interpreted as a relative measure of route popularity, relative to the origin location; or as a conditional probability: the probability that a vehicle travels to l_j given that it departs from l_i.

- **Upstream rate** - the proportion of traffic that arrives at a destination location l_j from an origin location l_i, denoted by $r_{\text{up}}(i, j)$. Let

$$r_{\text{up}}(i, j) = \frac{N_{ij}}{\sum_{i=1}^{|V|} N_{ij}} \tag{3.6}$$

  Similarly, $r_{\text{up}}(i, j)$ can be interpreted as a relative measure of route popularity, relative to the destination location.

The last two metrics are useful because they allow us to identify routes that account for a minimum proportion of upstream or downstream traffic, even if they are less frequented than other routes. Using the metrics above, we define four competing criteria. Each criterion $i = 1..4$ that defines an edge to be valid if:

1. It accounts for the top cumulative $\lambda_1 \in [0, 1]$ proportion of total observed traffic volume in the network.
2. Observes an AADT value greater than $\lambda_2 \in [0, 10000]$ vehicles/day.
3. Has either downstream **or** upstream rate greater than $\lambda_3 \in [0, 1]$.
4. Has both downstream **and** upstream rate greater than $\lambda_4 \in [0, 1]$.

where λ_i is a criterion-specific parameter. In short, routes are identified based on observed traffic volume, specified either as an absolute value, or relative to upstream/downstream traffic. The underlying assumption is that traffic volume for routes of interest is comparitively greater than routes that occur only due to erroneous trip identification. Route attributes such as distance, maximum speed and number of lanes, can then be used to further reduce the set of routes to fit a particular application, for instance routes whose length falls within a certain range.

To investigate the effect of different criteria and parameter values on the size and structure of the ANPR graph, we vary λ_i across the range of valid input values and compute the size of the resulting ANPR network, as the number of edges in the graph, separately for each $i = 1..4$. Results are shown in Figure 3.12 (note that graph size is in the log scale). To compare the effect of the different criteria in one graph, we transform criteria 1 and 2 as $1 - \lambda_1$ and $\frac{\lambda_2}{10000}$. For reference, we add the size of the expert annotated ANPR graph, inferred from data annotated by the traffic authorities of Tyne and Wear (Section 3.3.4).

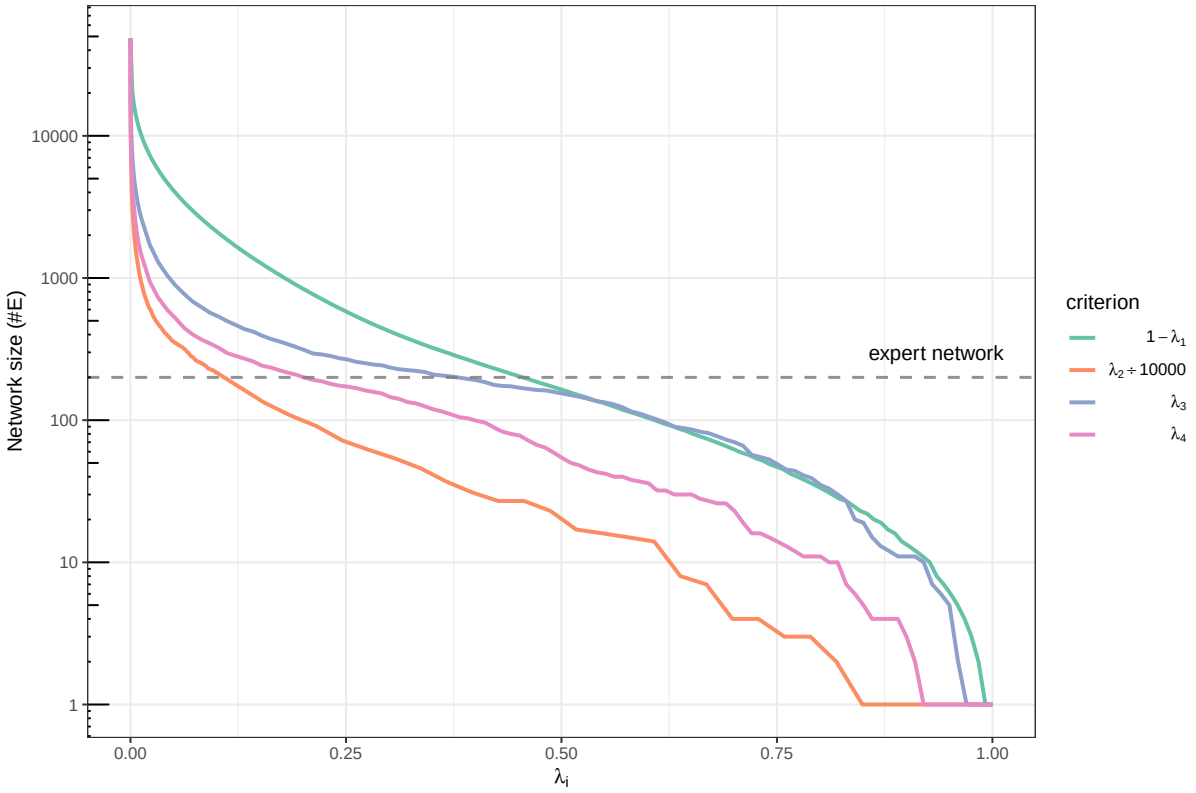


Figure 3.12: Effect of different criteria λ_i on ANPR network size.

In Figure 3.12, network size decreases sharply until about 1000 edges, and then decreases at a slower rate thereafter. Some descriptive numbers are:

- 164, 424 and 842 routes account for 50%, 70% and 80% of all observed traffic, respectively.
- 216 routes have an AADT greater than 1000 vehicles/day.
- 435 routes have a downstream traffic rate greater than 10%.
- 432 routes have an upstream traffic rate greater than 10%.
- 327 routes have simultaneously a downstream and upstream traffic rate greater than 10%.

We conclude that there is potential to increase the number of routes of interest from 200, the size of the expert ANPR network, up to between 300 and 500 routes (for this particular

example). However, the choice of a number within this range is somewhat arbitrary at this stage. Therefore, while the proposed method works well in composing a candidate set of valid routes, ranked according to the specified criteria, it does not guarantee the validity of all of its elements.

3.5.4 Missing data

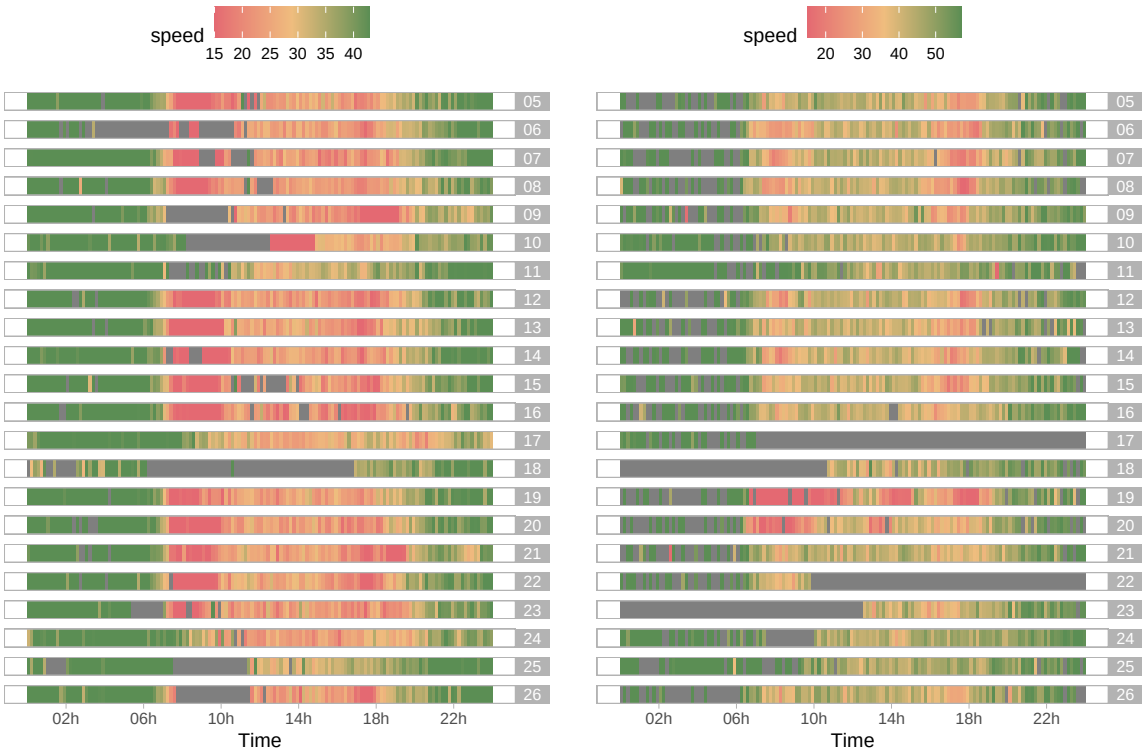
Missing data is a common problem in data analysis. It occurs as a failure or limitation of the data collection process and manifests as an observation that has no recorded value for the variable of interest (Little & Rubin, 2002). Missing values pose challenges to the application of statistical models that do not admit observation gaps or require mostly complete data, such as the multivariate analysis of climate data (Schneider, 2001) hierarchical models (Gelman & Hill, 2006) and vector autogression for traffic forecasting (Ermagun & Levinson, 2018). Traffic data, like other instances of spatio-temporal data, are prone to missing values as collection is distributed across a group of spatial sensors, themselves subject to variable capture and failure rates (Smith et al., 2003)

There are two primary treatments for missing data: imputation, whereby missing values are replaced with estimates; and omission, whereby missing values are removed (Little & Rubin, 2002). As treatment by omission is at times undesirable, namely in real-time analysis, numerous approaches have been developed for the imputation of missing traffic data, among others (Haworth & Cheng, 2012; Qu et al., 2009; Smith et al., 2003; Treiber & Helbing, 2002; Zhong et al., 2004). Developed techniques generally address one of three types of missing values (Little & Rubin, 2002; Qu et al., 2009): (a) missing completely at random (MCR), wherein missing data points occur independently of each other; (b) missing at random (MR), wherein missing points are locally correlated, appearing in clusters or short sequences; (c) not missing at random (NMR), wherein missing data occurs for long periods of time due to sensor malfunction, network communication breakdown or failure in data archival.

Imputation methods most commonly address MCR and MR type of missing data. Simple heuristic techniques use historical averages, calculated at the same location and time period, or averages of neighbour observations in time or space (Smith et al., 2003). Model-based approaches treat imputation as a short-term traffic forecasting problem. A number of univariate and multivariate models, with varying levels of computational complexity, have been developed in this domain, namely time series models (Autoregressive Integrated Moving Average) (Zhong et al., 2004), state space neural networks (van Lint et al., 2005), probabilistic principal component analysis (L. Li et al., 2013; Qu et al., 2009), tensor decomposition (Tan et al., 2013) and denoising stacked autoencoders (Duan et al., 2016). Whereas the aforementioned imputation methods estimate a missing value once, (Ni et

al., 2005) proposes a multiple imputation scheme aimed at reducing imputation bias and calculating imputation uncertainty, albeit at an added computational cost.

For long intervals of missing data (NMR), treatment is primarily done by omission, that is, by discarding the affected time periods from the analysis (Qu et al., 2009). Alternatively, the problem can be treated as long-term traffic forecasting. Notably, (Haworth & Cheng, 2012) developed a non-parametric kernel regression method, as an alternative to the k-nearest neighbour algorithm proposed by (Liu et al., 2008), that used upstream and downstream sensors to generate weeks-ahead travel time forecasts. The authors applied it to London's ANPR network, known as London Congestion Analysis Network (LCAP), where they reported 60% of observed data to be complete, and 23% of data to be missing in sequences longer than six observations.



(a) (b)

Figure 3.13: Average vehicle speed (10 min flow data) for two distinct routes, observed between the 5th and 26th of March 2018. Missing values are shown in dark gray.

In ANPR systems, type MCR/MR missing data can be the result of limited or degraded detection performance. Performance is influenced by variations in number plates and environmental factors such as luminosity and weather (Du et al., 2013). In addition, vehicle spacing is reduced in congested traffic, leading to increased number plate occlusion (Newell, 2002). Lower detection rates do not necessarily result in missing data but may manifest as anomolous flow observations, that is, a value lower than otherwise expected. Conversely, missing data may not represent an inability to capture vehicle number plates

but simply the lack of traffic demand, primarily during night time.

Figure 3.13 illustrates average speed data for two ANPR routes, determined from 10-min flow data. The different types of missing data (shown in dark gray) are evident: night-time zero flows (seen in both subfigures); isolated occurrences of missing values during daytime (e.g. days 07, 11 and 19 in Subfigure 3.13b); clusters of missing values during intervals of high congestion (e.g. days 06, 07, 08 and 09 in Subfigure 3.13a); and periods of missing data lasting longer than one day (e.g. days 17, 18, 22 and 23 in Subfigure 3.13b).

The chosen treatment of missing data in ANPR flows ultimately depends on the purpose of analysis and research goals. If large amounts of data are available, or only isolated routes/corridors are of interest, then treatment by omission may be adequate. On the other hand, the larger the subset of the ANPR network considered in the analysis, the more likely that missing values are misaligned in time and space, and one or more forms of imputation are required for complete network analysis. For proper identification of missing values, it is recommended that zero flow observations are encoded explicitly in flow data, i.e. by adding zero-valued records to gaps in the time series.

3.6 Summary of recommendations

The following is a summary of the processing steps of each major stage of the pipeline and a list of recommended user actions to ensure greater quality of output data.

Wrangling

- We recommend users to begin by cataloguing the different data assets and understand the state each dataset is in. This will help to determine what processing steps may already have been undertaken and which need addressing.
- Users should be aware of the camera clustering problem (two or more cameras in the same location and direction of traffic flow). Failure to address it will result in undesirable route fragmentation – the presence of more than one camera pairs with the same exact spatial mapping.
- Estimating the spatial mapping and length of a route is not strictly necessary if one simply aims to vehicle travel time. However, trip identification is made easier it .
- Related to the above, the purpose of snapping cameras to the road network is to enable the computation of shortest paths. We provide an automatic method to solve this problem, but users can perform this task manually using software like Booth et al. (2001) or *QGIS* (n.d.).

- Different sources of road map data are available to build the road network graph. We recommend *Ordinance Survey Open Roads* (2020) because it is a high-fidelity data source, though OpenStreetMap (2021b) also exhibits good accuracy levels.

- To match and eliminate bad formed number plate numbers, we recommend using a validated regex string pattern, not only capable of recognising current number plate patterns but also historical ones (in order to detect older vehicles).

- Before performing trip identification, we highly recommend the anonymisation of number plates to ensure user privacy and anonymity. It is possible to prevent against brute force attacks by employing pseudo-anonymisation techniques which combine a hash function with a secret salt (attackers would need to guess the secret salt in addition to the number plate).

Trip identification

- Solving the trip identification problem is a critical step to ensure data quality, both in individual and aggregate vehicle data.
- The use of a two-pass approach is greatly recommended: long-stop outliers are eliminated in the first pass and short-stop outliers eliminated in the second pass.
- A simple threshold works well in the first pass to eliminate the majority of outliers representing long vehicle stops. The use of a conservative threshold is recommended to minimise false negatives, i.e. actual outliers which are not a result of a vehicle stopping but highly congested traffic conditions.
- The second-pass stage should use a dynamic algorithm that is capable of adapting the decision boundary to current traffic conditions.
- Users can choose from a range of available second-pass algorithms, depending on: the amount of input data, the size of the input network (i.e. distinct number of camera pairs), the nature of the analysis (online versus offline), the desired speed of computation and the accuracy of the results (more complex methods are available but at the cost of additional computation).
- In large datasets, the boxplot method provides a positive trade-off between speed of computation (very fast method) and accuracy (reasonably accurate) approach to second-pass outlier detection.
- If the goal is to produce aggregate data for a known set of routes, then trip identification can be performed only for the corresponding camera pairs. However, for producing high-fidelity individual vehicle data (where the resulting travel sequences are more likely to be valid) then it is recommended that trip identification is applied to all registered camera pairs (otherwise travel sequences can contain gaps by omission of certain camera pairs).

Aggregation

- This step produces traffic flow data. Projects using only individual vehicle data may skip this stage.
- Users can choose from different sampling statistics to summarise travel time and average vehicle speed. For instance, the mean and median are typical measures of the location of the distribution, and standard deviation is a common measure of the dispersion of the distribution.
- An estimate of traffic volume can be obtained by counting the number of vehicles in the sample. However, these measurements will be biased and can only be considered lower-bound estimates of traffic volume. Vehicles that do not count towards the total include those that were only detected at the origin camera (they exited midway through the route), only at the destination camera (they merged onto the route) or detected by neither camera (either they travelled the subset of the route that does not include any camera, or passed undetected through both cameras).
- ANPR allows users to choose the level of data aggregation that best matches their final application. For short-term traffic forecasting, or to simply capture the evolution of traffic conditions over time, five to fifteen minute window sizes are recommended as noise has been found to be minimal. For applications requiring lower window sizes, e.g. very short term traffic forecasting for signal control, specialised denoising techniques are recommended to deal with low sample sizes.
- When sampling vehicles, the user must decide how to treat vehicle observations that cross time boundaries. The most natural solution is to sample the same vehicle observation across all time periods that fall in between its origin and destination timestamps (including). However, this solution comes at the cost of computational complexity and added bias in the resulting statistics of travel time/average speed. An alternative solution, easier to implement and with lower computation cost, is to select only the interval containing the timestamp at the origin camera (the equivalent of a date floor operation).
- Route identification is a procedure used to obtain a graph representation of the ANPR network from data, rather than from expert annotations. This step is optional and done post-aggregation, but is nonetheless useful as the starting point for subsequent analysis. Users can choose from different metrics, such as Annual Average Daily Traffic (AADT), when selecting valid routes. For our network, we have found a value of AADT between 300 and 500 routes to provide results that matched and enhanced the route selection provided by experts.
- Missing data is a common issue in ANPR data. Missing data is particularly problematic when heavy congestion causes vehicle number plates to be concealed by reduced vehicle headways and detection rates to drop. Treatment of missing data

is done by omission, by selecting or discarding - or through data imputation techniques. Treatment by omission is only feasible when large amounts of data are available, and recommended only when the integritiy of the analysis is not affected, namely for missing data at random.

- Data imputation is recommended when the integrity of the analysis is affected by the presence of too much missing data, particularly of type not missing at random.
- Users have have a large array of data imputation techniques at their disposal. Careful consideration is recommended when choosing an appropriate technique as methods are often optimised to deal with particular patterns of missing data, i.e. missing completely at random (MCR), missing at random (MR), not missing at random (NMR). For dealing with large amounts of MCR and MR data, we recommend the methods of Smith et al. (2003) or Zhong et al. (2004). For NMR data, we recommend the method of Haworth & Cheng (2012) instead.
- For proper identification of missing values, we recommended zero flow observations to be encoded explicitly in flow data, i.e. by adding zero-valued records to gaps in the time series.

3.7 Discussion and future work

The pre-processing of ANPR data is an essential precursor to analysis. As shown throughout this Chapter, the process of producing aggregate flow data from raw ANPR data is non-trivial. Researchers may accidentally skip certain steps because they are unaware of their need, as shown in Table 3.1. To alleviate this problem, we built a comprehensive processing pipeline for ANPR data. The pipeline is an end-to-end tool for preparing and transforming ANPR data from its starting state, as a unified register of camera detections, to a state ready to be analysed by humans and/or computer programs. Without this tool and the individual and flow data generated by it, our subsequent research would not be possible. Hence, in the context of this thesis, the pipline enables our research of road corridors, traffic bottlenecks and vehicle overtakings, in Chapters 4, Section 5 and Section 6 respectively.

Our pipeline allows users to produce data ready for analysis both at the disaggregate level, in the form of individualised vehicle trips, and aggregate level, in the form of collective OD flows. The simultaneous aggregate and disaggregate elements of data produced by ANPR networks allows their application in variety of traffic problems and ITS applications, from traffic forecasting to the analysis of user behaviour. Furthermore, by using our pipeline and methodology for ANPR data processing, users will be able to reduce the time to analysis and the probability of making errors in pre-processing which can significantly impact the outcome of their analysis. In this way, the pipeline reduces the barrier to entry

for new researchers and practitioners of ANPR data, one of the key aims of this thesis. Although other studies have described the steps involved in pre-processing ANPR, no other work has presented an approach where function is clearly encapsulated and divided into modules. In this case, processing was divided into three major stages, Wrangling, Trip identification and Aggregation, reflecting key functions of the pipeline. This framework makes it easier to model, implement and extend the system in software, with clear and well-defined functions and interfaces.

One other contribution of our pipeline, is the development of new solutions for several processing tasks. We created a camera clustering algorithm that allowed vehicle observations to be associated with unique locations. A point-to-curve algorithm was developed to automatically perform the matching of camera locations to map locations. We developed a framework for the identification and treatment of invalid vehicle steps, including the application of Tukey's rule to identify outlier observations. Lastly, we proposed a directed weighted graph representation of sensor networks that uses volume-based metrics, such as annual average daily volume, to identify a larger set of routes for analysis.

Altogether, the pipeline can be enhanced in multiple ways. First, it should be fully implemented as an open-source software library, so that it can easily shared, modified and adapted. Second, the pipeline can be revised to make a clearer and more detailed distinction between batch and real-time processing techniques, including implementation details relevant to each case. Moreover, it should add practical considerations for dealing with Big Data and ANPR networks above a certain size. Although baseline results were included for most stages, it would be beneficial to augment the benchmark with results from other real ANPR networks. Finally, the pipeline can be extended to consider the retroactive effects of data processing, particularly the effect of route identification and missing data analyses on trip identification.

Chapter 4

Representation and discovery of road corridors

4.1 Introduction

When ANPR networks are sparse, monitoring is often constrained to a pre-determined set of routes (camera pairs). As sensor density increases so does the connectedness of the sensor network, including its ability to capture the routing structure of the underlying road network (Antoniou et al., 2011). Concretely, for a sufficiently dense network, it becomes possible to consider travel sequences composed of three or more locations, which contain information about the intermediary steps of a trip in addition to its origin and destination points. Such travel sequences, when frequently observed, reveal important connections and causal associations between routes (e.g. route B is commonly preceded by route A and followed by route C) and are indicative of users' travel patterns and routing preferences (Q. Cao et al., 2020).

The discovery of core travel sequences in a sensor network can not be achieved purely by inspecting its structure or counting observations. On the one hand, there are theoretically many paths between any two given nodes in a sensor network – approximately $n!$ in a complete graph (Diestel, 2017) – of which only a fraction is realistically expected to be representative of travel patterns (in the same way that a sensor network is composed by a small subset of all camera pairs). On the other hand, trip identification errors can lead to significant numbers in trip sequences whose underlying spatial route is in fact strongly sub-optimal (e.g. a route that loops around needlessly). In addition, different trip sequences may have an identical spatial mapping and should therefore be treated as a singular element.

We tackle these challenges by developing a robust mathematical framework capable of representing and discovering core travel sequences in a sensor network, named corridors,

from observed data. The term corridor is chosen because it often refers to a generally linear path that connects several transport nodes, and which plays an integral role in the diffusion of people and goods across the region, with wider economic impacts for the local economy (Roberts et al., 2020). Concepts from graph theory are employed in corridor representation, allowing for flexible configurations: corridors may be comprised of a single route from origin to destination or offer multiple route choices. For corridor discovery in sensor networks, we develop a two-step algorithm. First, we identify ordinary trip sequences from an input list of observed sequences, which are then used to obtain corridor set, i.e. a collection which contains no duplicate corridors or corridors contained within other corridors.

The chapter is structured as follows. Section 4.2.1 presents formal definitions of corridor and corridor set. Section 4.2.2 describes a corridor labelling algorithm proposed for corridor discovery in an input ANPR sensor network. Section 4.3 explores corridor representation and discovery using a toy example, while Section 4.4 applies the approach to the ANPR sensor network of Tyne and Wear. Within Section 4.4, the frequency of corridor structure and representation is studied in Section 4.4.1 using real-world examples. Sensitivity analysis is performed in Section 4.4.3 to determine how different input parameters affect the output corridor set and Section 4.4.4 describes important corridors in the region.

4.2 Methodology

4.2.1 Definition of corridor and corridor set

This section develops formal definitions of corridor and corridor set (a collection of corridors). To split otherwise long definitions into smaller components, we also defined each of the auxiliary terms: trip sequence, route utility and ordinary trip sequence. A trip sequence is simply a sequence of vehicle steps with a corresponding spatial route, regardless of popularity or validity (an invalid route avoids reaching the destination directly, for instance, by looping around unnecessarily). Conversely, an ordinary trip sequence is a frequently traversed and valid path, where spatial validity is expressed using a proposed measure of route utility (calculated by comparing the route's length with shortest path length between the origin and destination points). A corridor can then correspond directly to an ordinary trip sequence, or be composed of several distinct sequences with the same OD pair (first and last observations). Lastly, a corridor set represents the entire corridor collection of a sensor network, and ensures that there are no duplicates or elements that are completely contained within other elements.

4.2.1.1 Trip sequence

Recall, from Section 3.4, that a vehicle trip is described by a series of movements called steps, each characterised by a camera pair (origin and destination locations), a departure and arrival time and a spatial route (a path in the road network graph consisting of a sequence of geometrical points and the line segments that connect them).

Altogether, the vector of visited locations or camera pairs in a single vehicle trip is called a trip sequence. Formally, a trip sequence P of length n is written either as a sequence of cameras identifiers $P = u_1, u_2, \ldots, u_n$, or a sequence of $n-1$ steps $P = \{s_j\}_{j=1}^{n-1}$, where $s_j = (u_j, u_{j+1})$ represents the camera pair of step j. The two expressions for P are used interchangeably: the point-based expression is useful when referring to visited locations in the trip or succinctly representing trip sequences as n-tuples or comma separated strings (e.g. "A,B,C"), while a step-based expression makes it easier to reference travelled routes.

The spatial route of a trip sequence P is denoted $R(P)$ and is obtained by the concatenation of its spatial sub-routes (route of each step), in order of occurrence:

$$R(P) = R(s_1)^\frown R(s_2)^\frown \ldots^\frown R(s_{n-1}), \tag{4.1}$$

where $R(s_j)$ is the spatial route of the j-th step of P (calculated as a shortest path route in the road network graph in Section 3.3.4) and $X^\frown Y$ specifies the concatenation of two finite sequences X and Y. In turn, the full route length of a trip sequence $\text{dist}(P)$ is given by the sum of its individual route lengths:

$$\text{dist}(P) = \sum_{j=1}^{n-1} \text{dist}(s_j), \tag{4.2}$$

where $\text{dist}(s_j)$ is the length of spatial route $R(s_j)$ (obtained by summing the lengths of its composing line segments).

Lastly, a trip sequence P has origin-destination (OD) pair $P_{\text{od}} = (u_1, u_n)$ with spatial route $R(P_{\text{od}})$, as illustrated in Figure 4.1. Note that $R(P_{\text{od}})$ does not necessarily overlap with $R(P)$ because $R(P)$ has intermediary nodes $u_2, \ldots, u_{n-1}$ where $R(P_{\text{od}})$ is not restricted to pass through the same intermediary nodes. Consequently, the route length of P is always equal or greater than that of its origin-destination counterpart: $\text{dist}(P) \geq \text{dist}(P_{\text{od}})$.

4.2.1.2 Route utility

In traffic assignment theory, the utility of a route is generally related to its length, or more generally, to how quickly it allows users to travel from origin to destination (Chiu et al., 2011). High utility routes are preferred by users compared to low utility ones. The

principle of route utility can be applied to identify trip sequences resulting from failed trip identification or unusual user behaviour. Let $\mathrm{U}(P)$ denote the (normalised) route utility of trip sequence P, calculated as the inverse quotient of the length of P and that of its OD path P_{od}:

$$\mathrm{U}(P) = \left[\frac{\mathrm{dist}(P)}{\mathrm{dist}(P_{\mathrm{od}})}\right]^{-1} = \frac{\mathrm{dist}(P_{\mathrm{od}})}{\mathrm{dist}(P)} = \frac{\mathrm{dist}(P_{\mathrm{od}})}{\sum_{j=1}^{n-1} \mathrm{dist}(s_j)} \,. \tag{4.3}$$

Overall, the utility of a trip sequence represents how much longer its route is compared to the shortest path from origin to destination. A route utility factor of one specifies that $R(P_{\mathrm{od}})$ is equivalent to $R(P)$, while a route utility factor of 0.5 describes a route twice as long as its optimal equivalent. Routes may be suboptimal (low-utility) for a variety of reasons, such as lack of road network knowledge, preferred trip path or a requirement to stop along the route at a particular location (e.g. to drop off a passenger). Figure 4.1a shows a trip sequence whose route does not match its optimal OD, but whose utility is above 0.95. Conversely, Figure 4.1b depicts a trip sequence whose route is sub-optimal.

(a) Trip sequence `230,193,201,199` with origin location `230` and destination `199`, and utility factor of 0.96.

(b) Trip sequence `100,235,011` with origin location `100` and destination `11`, and utility factor of 0.71.

Figure 4.1: Route and utility factor of two distinct trip sequences.

Exceptionally, for trip sequences with the same origin and destination locations, their route utility is not defined. This is in part because shortest path algorithms generally do not admit equal start and end nodes. Furthermore, our interest is in generally linear travel paths (corridors) and not so much in circular travel sequences. Paths containing pairs where $u_{j+1} = u_j$ (two observations at the same location) are also not admissible. While the treatment of vehicle trips ensures that $u_{j+1} \neq u_j$ always holds true, it does not ensure that $u_n \neq u_1$ also holds true. Hence, trip sequences where $u_n = u_1$ are assigned a

utility of zero.

4.2.1.3 Ordinary trip sequence

An ordinary trip sequence represents a route regularly and deliberately chosen by users, as opposed to a route that occurs sporadically due to a mistake, an abnormal event, or erroneous trip labelling. The classification of ordinary trip sequences serves to identify trip sequences that correspond to a road corridor, or a part thereof. Given a graph $G = (V, E)$, denoting an ANPR network (as described in Section 3.5.3) with vertex set V (set of ANPR cameras) and edge set E (set of valid camera pairs), a trip sequence P is classified as *ordinary* if it meets the following criteria:

1. P is a simple directed path[1] in G.
2. Its route utility $\mathrm{U}(P)$ is greater than a minimum utility value ϵ_u.
3. Its traffic volume rate, in vehicles per unit of time, is greater than a minimum rate ϵ_r.

Condition 1 acts to reduce the set of possible trip sequences of length $n \geq 3$, since all steps $s_j = (u_j, u_{j+1})$ in a trip sequence P must be a valid camera pair in G, i.e. an element of the edge set E.

Despite reducing the space of possible ordinary trip sequences, criterion 1 is by itself not sufficient to recognise well-frequented trip sequences, as it implies that every possible path in G has non-zero probability of occurrence. In practice, routes are typically chosen to fulfil a certain trip purpose, that is, to travel between two or more locations known in advance (McNally, 2007). Additionally, a rational user only considers a handful of routes from origin to destination and ignores many other possible routes, perceived as sub-optimal according to some criteria (e.g. slower, more turns without right of way). As our goal is to reasonably approximate the set of *ordinary* trip sequences, rather than studying the underlying process of user mobility and route choice, conditions 2 and 3 are further adopted as heuristics. Condition 2 discards trip sequences representing routes with low route utility (users prefer shorter to longer routes) and condition 3 removes routes seldom travelled.

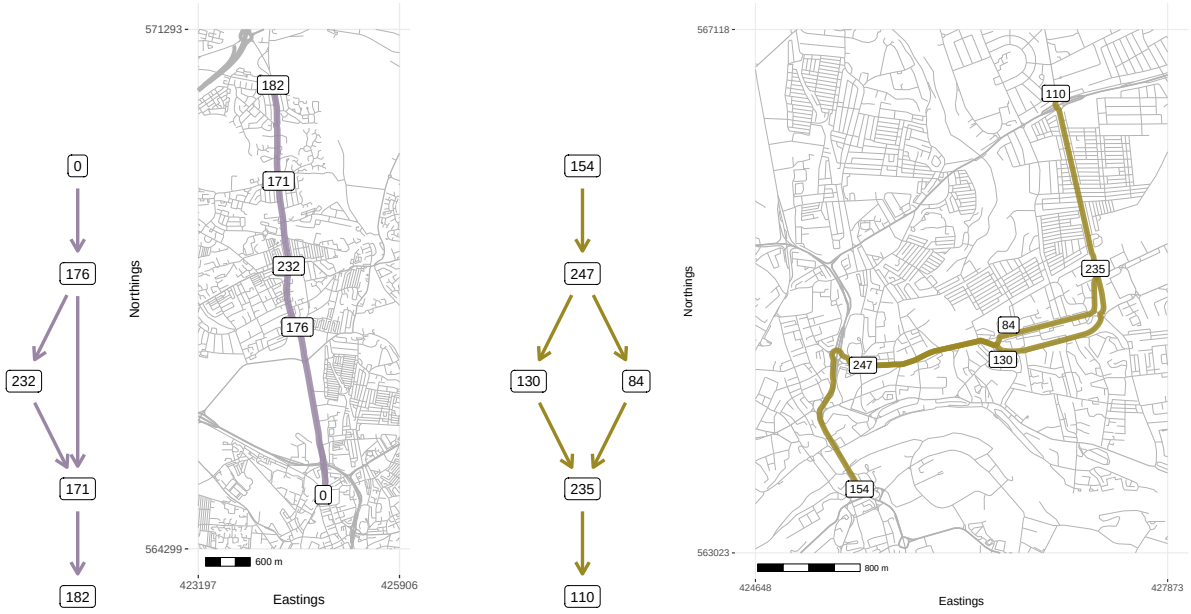
Lastly, a useful property of an ordinary sequence is that all of its substrings are themselves also ordinary. A substring is formed by keeping two or more consecutive elements of an input string and eliminating the rest. In contrast, a subsequence is formed by eliminating elements of another string, occurring consecutively or otherwise (Gusfield, 1997). For example, the sequence `"A,B,C"` (of length three) has two substrings of length two, `"A,B"`

[1] A simple path is a sequence of (directed) edges joining adjacent vertices, whose vertices, and therefore edges, are all distinct (Diestel, 2017).

and "B,C," and not-substring subsequence "A,C". If "A,B,C" is ordinary then "A,B" and "B,C" are ordinary as well, whereas "A,C" is not necessarily so (it may be or not depending on observed traffic rate).

4.2.1.4 Corridor

A corridor is a roughly linear travel path that serves an important role in the diffusion of traffic and goods across two locations in the a road network (Witte et al., 2012). These characteristics are largely present in the definition of ordinary trip sequence: a frequently traversed and valid travel path with distinct origin and destination locations. However, it is possible that two or more ordinary trip sequences represent the same route and hence road corridor. Figure 4.2a shows an example of two distinct ordinary trip sequences, 0,176,232,171,182 and 0,176,171,182, that correspond to the exact same spatial route, thus representing the same road corridor. This particular case occurs because camera 232 can not detect vehicles across all lanes, causing vehicles to pass through that corridor location undetected. Figure 4.2b depicts another example of different sequences representing the same corridor but where the two sequences 154,247,84,235,110 and 154,247,130,235,110 map to different routes that serve the same trip purpose, i.e. to connect origin 154 to destination 110.



(a) Trip sequences *0,176,232,171,182* and *0,176,171,182* with the exact same spatial mapping.

(b) Trip sequences *154,247,130,235,110* and *154,247,84,235,110* representing different routes connecting the same OD locations.

Figure 4.2: Two examples where different trip sequences represent the same road corridor.

In practice, we want a corridor to include all ordinary trip sequences that share the same objective – to reach from a specific origin location to a specific destination. This is

achieved by combining related sequences together in a graph. Specifically, a road corridor D is an induced subgraph[2] of an ANPR network graph that meets all of the following conditions:

1. D has two or more nodes.
2. D is a directed acyclic graph (DAG).
3. D does not contain isolated nodes.
4. D has exactly one source node $s(D)$ and one sink node $t(D)$.
5. Every path in D, from source to sink, forms an ordinary trip sequence.

The conditions above formalise key characteristics of a transport corridor: distinct origin and destination points (conditions 1 and 4), optional intermediate transport nodes (conditions 1 and 3), carrying a significant volume of traffic through an approximately linear travel path (conditions 2 and 5). In short, a corridor generalises the concept of ordinary trip sequence to a group of related sequences. Two sequences are related if one is a subset of the other (a substring), or share the same origin and destination locations (source and sink nodes).

A direct consequence of defining a corridor as a graph is the possibility of variety in corridor structure. A corridor with a single path from source to sink is equivalent to an ordinary trip sequence, possibly offering only one main route from origin to destination. Conversely, a corridor with multiple paths from source to sink may indicate additional routing options, as exemplified in Figures 4.1a and 4.2b, respectively. Figure 4.3 depicts four valid corridor configurations of increasing complexity (edge count), while Figure 4.4 shows several examples of invalid graph configurations: (a) node 3 is isolated, (b) there are two sink nodes: 2 and 3, (c) there are two source nodes: 1 and 2, (d) contains a cycle and is, therefore, not a DAG.

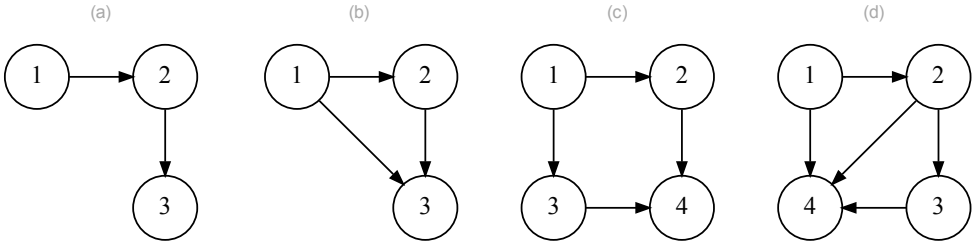


Figure 4.3: Example of valid corridor configurations.

It is expected that the structure of the corridor graph is related to the type of roads it is built on. Concretely, we expect that road corridors with restricted access control (few entry and exit points), such as highways, are more likely to be configured as in Figure 4.3

[2]An induced subgraph G' of a graph G is another graph whose vertex set $V(G')$ is a subset of $V(G)$ and edge set is composed of all edges $u, v \in E(G)$ such that $u \in V(G')$ and $v \in V(G')$ (Diestel, 2017).

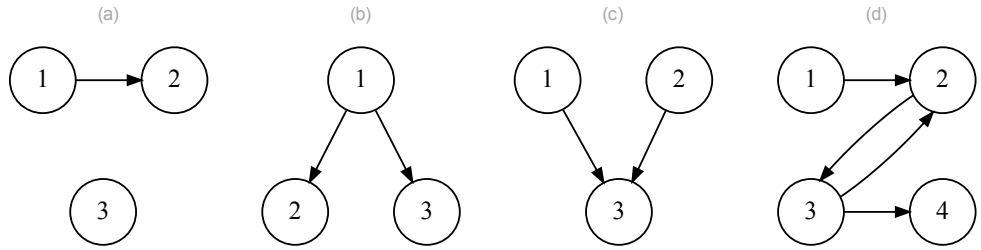


Figure 4.4: Example of invalid corridor configurations.

(a); whereas urban corridors via arterial and collector roads in densely populated areas may offer a range of routes to reach the destination, as in Figures 4.3 (b) and (c).

Finally, note that, for any corridor D, one should not assume that vehicles will always travel the full length of the corridor, that is, produce a trip sequence starting at its source node $s(D)$ and terminating at its sink node $t(D)$. Vehicles may start or terminate their trips at intermediary nodes along the corridor, producing shorter trip sequences which are nonetheless categorised as ordinary.

4.2.1.5 Corridor set

In the same way that an ANPR network consists of all valid and important routes in a road network, a corridor set encompasses all significant route combinations within a region, which can then be incorporated into new levels of analysis. Concretely, a corridor set is a collection of corridors wherein no two corridors are semantically equal, that is, no two corridor graphs are equal[3], no corridor is a subset of another nor two corridors have both source and sink nodes in common with one another. Formally, for any given corridor D in a corridor set $\mathbb{D}$, there is no other corridor $D' \in \mathbb{D}$, such that either of the following conditions is true:

1. $D' \subseteq D$
2. $s(D) = s(D')$ and $t(D) = t(D')$,

where $s(X)$ and $t(X)$ denote the source and sink nodes of corridor X.

Figure 4.5 depicts two collections of corridors (a) $\mathbb{D}_1$ and (b) $\mathbb{D}_2$. $\mathbb{D}_2$ is a valid corridor set but $\mathbb{D}_1$ is not for two reasons: corridor ii is a subgraph of corridor i and corridors i and iii share both the same source and sink nodes despite none being a subset of the other. $\mathbb{D}_1$ can be made into a valid corridor set, with a single corridor, by finding the union of graphs i, ii and iii (resulting in graph i with an additional node `5` and three additional edges: `(1,4)`, `(1,5)` and `(5,4)`).

[3]In the sense of a label-preserving graph isomorphism.

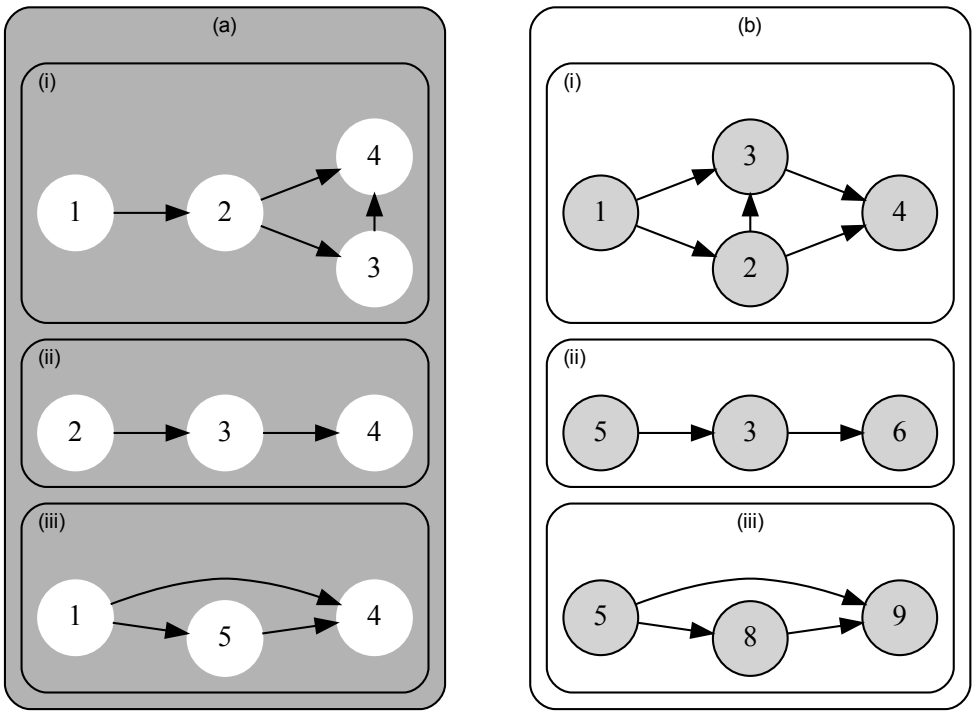


Figure 4.5: Two corridor collections: (a) invalid corridor set (elements i and iii have the same source node 1 and sink node 4) and (b) valid corridor set.

Note that a corridor set is not a partitioning of camera locations $V(G)$ into mutually exclusive groups. In Figure 4.5 (b) nodes 3 and 5 are included in multiple corridors, depicting possible spatial overlap and traffic interactions between corridors. Corridors may thus have one or more locations in common. For example, this can occur when two corridors depart from the same location towards different destinations or have an intermediate node in common. The process of finding a corridor set $\mathbb{D}$ from an ANPR network G is called *corridor labelling* and is elaborated next.

4.2.2 Corridor labelling algorithm

Corridor labelling is the process of obtaining a corridor set from ANPR data. It is a two-step process. First, the set of ordinary trip sequences is determined from observed trip data. Then, the set of ordinary trip sequences is used to compute a valid corridor set. The two steps are distinct insofar as the first depends on several input/parameter values, whereas the second is purely deterministic – the same input generates the same output repeatedly.

4.2.2.1 Phase 1: Ordinary trip sequences

The first phase of the corridor labelling algorithm is the identification of ordinary trip sequences. The algorithm receives as input three distinct datasets and two user-defined parameters, and outputs a list of trip sequences that match all three ordinary criteria, as

specified in Section 4.2.1.3. The input data consists of a trips dataset (illustrated in Table 3.8), all-pairs distance data (as shown in Table 3.4), and an ANPR flow graph – a graph data structure that provides methods for graph traversal and vertex/edge checking, such as the ones implemented in the popular software library `igraph` (Csardi & Nepusz, 2005). In addition, values for parameters ϵ_r (the minimum trip sequence observation rate) and ϵ_u (the minimum route utility factor) must be specified.

Algorithm 2 gives a textual description of phase 1 of the corridor labelling algorithm. Each step is computed sequentially, that is, the output of one operation serves as input to the next (with the exception of the first step). When mentioned, a step also utilises an input dataset or parameter. Trip sequences are represented as comma-separated strings.

Algorithm 2 Get all ordinary trip sequences

Input

D_{trips}	trips dataset
D_{pairs}	camera pairs dataset with calculated spatial route lengths
G	graph data structure representing the sensor network
ϵ_r	Minimum trip sequence observation rate (e.g. vehicles per day).
ϵ_u	Minimum route utility factor (between 0 and 1).

Output

D_{ordinary}	ordinary trip sequences dataset

1. Collapse each vehicle trip in D_{trips} into a comma-separated trip sequence string.
2. Count the number of occurrences of each unique sequence string of length 2 or greater.
3. For all sequence strings of length $l \geq 3$, add a duplicate count entry for each of its consecutive subsequences of length $i = 2 \ldots l - 1$.
4. Calculate the total count of each unique sequence string.
5. Divide the total observed count by the number of elapsed time units in the dataset (e.g. days) to calculate the observation rate of each sequence string.
6. Filter all sequence strings whose average count is above the minimum chosen rate ϵ_r.
7. For each sequence string, compute whether it corresponds to a valid simple path (vertex sequence) in G.
8. Filter all sequence strings that are a valid simple path.
9. For each remaining sequence string, compute its route utility factor using Equation 4.3. Use D_{pairs} to obtain the required spatial route lengths.
10. Filter the sequence strings whose utility factor is above a minimum value ϵ_u.

Step 1 generates a list of trip sequence strings of the form `"A,B,C,D"`, where each letter represents a distinct camera id. Step 2 tallies up string occurrences. Step 3 creates new entries to reflect the fact that sequences of length 3 or greater (in number of non-comma characters) imply the occurrence of shorter subsequences. For example, sequence `"A,B,C,D"` has two substrings of length 3, `"A,B,C"` and `"B,C,D"`, and three substrings of

length 2, `"A,B"`, `"B,C"` and `"C,D"`. Then, if `"A,B,C,D"` has an observed count 100 a new count entry `(substring, 100)` is created for each of its substrings. The partial counts generated this way are then summed up in step 4. Steps 5 through 10 calculate variables and remove entries whose values do not meet the criteria for ordinary trip sequences. The trips dataset is used in steps 1 through 5 to obtain an initial list of trip sequences and calculate the frequency of observed trip sequences (criterion 3). Distance data is used to calculate route utility factors in step 9 (criterion 2) and the flow graph G serves to compute simple paths in step 7 (criterion 1).

The algorithm was implemented using the `tidyverse` R libraries for tabular data operations (Wickham et al., 2019), namely `dplyr` (Wickham et al., 2021) for data manipulation and `stringr` for string operations (Wickham, 2019). Data reduction steps such as frequency counting, summation or string collapsing, are implemented by combining a `group_by` with a `summarise` operation. A `filter` operation is also a data reduction step that retains all rows/entries that satisfy a given logical condition. Computation of variables is achieved using a `mutate` operation, which adds new variables as functions of existing variables in vectorised fashion. Two (tabular) datasets are merged using a mutating join (inner, left, right, full), applied to common variables in the input tables.

In terms of time complexity, step 3 is the most costly operation of Algorithm 2. It involves finding all substrings of length 2 and greater for each trip sequence resulting from step 2. The total cost, in big O notation, is approximately $O(n * m^2)$, where n is the number of input trip sequences and m^2 is the cost of finding all substrings in a string of length m. However, the m^2 factor is amortised given that less than 1% of trip sequences have length greater than 9 (see Table 3.20). Other steps are computed mainly on a single dataset pass (or on a whole vector using a vectorised function), hence their time complexity is approximately $O(n)$: reduction by string collapsing (step 1), reduction by frequency counting (step 2), reduction by summation (step 4), reduction by filtering (steps 5, 7, 9), computation of simple path (step 6) and computation of route utility (step 8). Of these, the computation of simple paths and route utility involve greater overhead, since they are composed of several operations. The calculation of simple paths is not a vectorised function, as it involves several $O(1)$ graph access operations to check for edge presence. The calculation of route utility involves a few vectorised at worst $O(n)$ operations, including joining the current trip sequence data with the camera pairs dataset, substring retrieval and variable calculation.

4.2.2.2 Phase 2: Corridor set

Phase two of the corridor labelling algorithm seeks to generate a valid corridor set from the list of ordinary trip sequences produced by Phase 1. The output dataset consists of a

list of edges, grouped by the corridor graph they belong to. Algorithm 3 gives a textual description of this phase of the labelling algorithm, where each step is again computed sequentially, i.e. the output of each operation serves as input to the next (except for the first step). In contrast to the previous algorithm, Algorithm 3 does not depend on any user-defined parameters and therefore a single output is generated repeatedly for a given input.

Algorithm 3 Get corridor set

Input

D_{ordinary} ordinary trip sequences dataset

Output

$D_{\text{corridors}}$ corridors dataset

1. Implement function $f_{\text{sub}}(x_1, x_2)$ which evaluates to 1 if sequence x_2 is a substring of x_1 and 0 otherwise (evaluates to 0 if $x_1 = x_2$).
2. Compute f_{sub} for all pairs of trip sequence strings (x_i, x_j) in D_{ordinary}.
3. Find all super-sequences: sequences that are not a substring of any other sequence, i.e. for which the condition $\sum_{j=1}^{N} f_{\text{sub}}(x_i, x_j) = 0$ evaluates to true.
4. Find the first and last element of each super-sequence, i.e. its OD pair.
5. Group the super-sequences by OD pair, and assign unique corridor labels to each group of super-sequences.
6. For each resulting corridor, transform each of its trip sequences strings into pairs of consecutive elements and remove any duplicate edges.

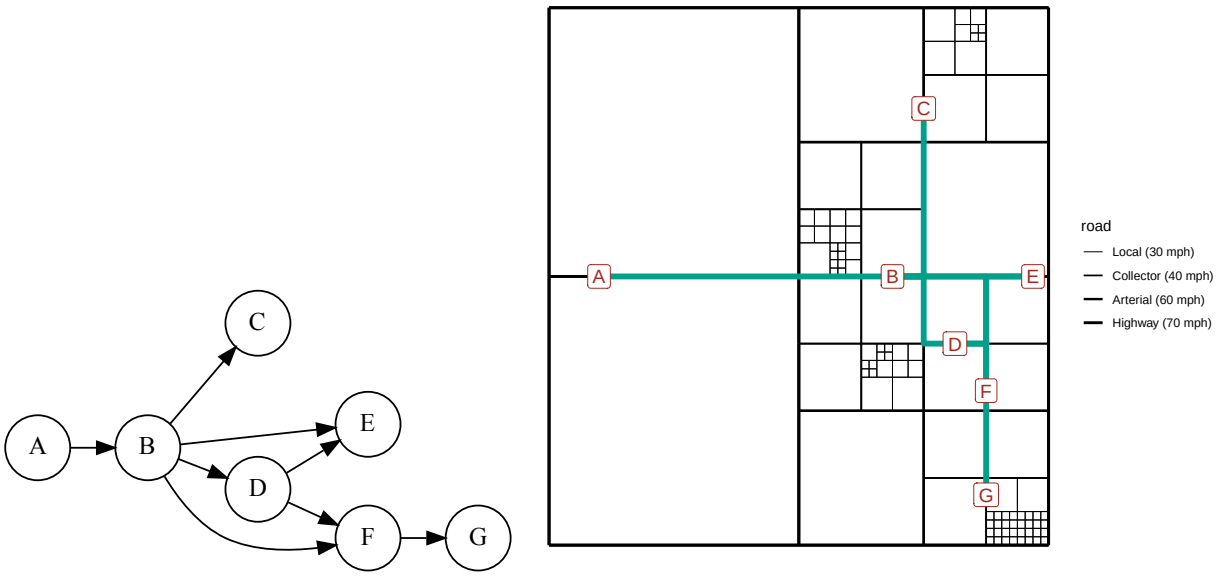
Algorithm 3 is divided into two main parts. First, we find all ordinary trip sequences that are not a sub-sequence of any other sequence, called super-sequences (steps 1 to 3). Second, we combine super-sequences with equal source (origin) and sink (destination) nodes to form unified corridors (steps 4 to 6). Super-sequences reflect the condition that there are no two elements in a corridor set such that one is a subgraph of the other (condition 1 in Section 4.2.1.5). In turn, the merging of super-sequences results in an element set where no two corridors contain the same source and sink nodes (condition 2).

To find super-sequences, step 1 implements function $f_{\text{sub}}(x_i, x_j)$, which detects if sequence x_i is a substring of sequence x_j. Step 2 computes f_{sub} for every pair of sequences (x_i, x_j) in the input dataset, allowing super-sequences to be found when the equality $\sum_{j=1}^{N} f_{\text{sub}}(x_i, x_j) = 0$ is true (step 3). Steps 4 and 5 group super-sequences with equal origin and destination nodes under a single label, for instance `"A,B,C,D"` and `"A,C,D"`. Lastly, step 6 transforms sequence strings into edge lists, e.g. `(A,B),(B,C),(C,D)` and `(A,C),(C,D)`, and removes any repeated edges that arise in corridors composed of multiple trip sequences, `(C,D)` in this example. In terms of time and space-complexity, Algorithm 3 exhibits $O(n^2)$ cost, since function f_{sub} is evaluated for every pair of input sequence

strings.

4.3 Results for toy-worked example

To illustrate the principles of corridor labelling, we work through a simple synthetic example. The following three inputs are given: (1) a sensor network graph G_{toy} depicted in Figure 4.6a, (2) a simplified road network with camera locations shown in Figure 4.6b, and (3) synthetic trip sequence counts listed in Table 4.1.



(a) Sensor network graph.

(b) Road network, camera locations and camera pair routes.

Figure 4.6: Corridor identification toy example: inputs.

For demonstrative purposes, the structures of the sensor and road network graphs are simplified. G_{toy} is a DAG with 7 nodes and 8 edges, whereas the road network is modelled as a quadtree (Eisenstat, 2011). A quadtree is a tree data structure whose non-leaf nodes have exactly four child nodes. Each child node partitions two-dimensional space (typically a square) into four equal regions, called quadrants, according to some defined space-partitioning strategy (de Berg et al., 2008). Eisenstat (2011) used the quadtree as a simple generative model of the road network, where regions of greater road density correspond to areas of increased human activity or population density.

To simulate different levels of the road hierarchy, we assigned each road edge a value of length and maximum speed values according to their depth in the quadtree. Four road types are defined, in decreasing order of importance: highway (70 mph), arterial (60 mph), collector (40 mph) and local road (30 mph). In addition, the segment free flow travel time was calculated from its speed and length (see Table 2.1), and then used as

weighting factor to determine the shortest-path spatial route, for each each of the camera pairs (edges) defined in Figure 4.6a.

Contrary to Eisenstat (2011), the quadtree toy model and respective camera locations in Figure 4.6b were chosen manually in order to create predictable spatial routes for each of the camera pairs defined by G_{toy}. Cameras `A`, `B` and `E` are located on a highway segment connecting different settlement areas in subregions of the top and bottom right quadrants characterised by higher road densities. Cameras `C` to `G` are positioned along second and third tier roads (like 'A' and 'B' roads in the UK) offering access to local roads near the settlement destinations.

Provided with a simple road network model and camera locations, trips are generated along chosen corridors to recreate different travel motifs. Trips taken along corridor `A,B,E` represent vehicles passing through the region along the main highway. Trips taken along corridor `A,B,C` represent vehicles travelling from sparsely populated regions, or outer regions not depicted in the graph, towards the top right settlement. Trips originating in `B` towards cameras `D` to `G` represent movement between different settlements, motivated by activity type, for instance, residential, professional or recreational.

The aforementioned trip patterns are reflected in Table 4.1, which displays trip sequence counts, denoted by N, obtained after steps 2 and 4 of Algorithm 2 (the original synthetic dataset is otherwise too large to be listed here in its entirety). Also shown are attributes calculated in other steps of the algorithm: the sequence length l, its route utility factor, its observation rate (obtained by dividing N by 10 time units), and logical values (T/F) representing whether the sequence is a simple path in G_{toy} and an ordinary trip sequence (for the parameter values $\epsilon_r = 5$ and $\epsilon_u = 0.7$). In total, 14 out of 21 trip sequences are labelled as ordinary. The need for steps 3 and 4 is exemplified by sequence `"(A,B)"`, whose observation count N increases from 100 to 300 after considering longer sequences containing `"A,B"`, namely sequences `"A,B,C"` and `"A,B,E"`.

After the ordinary trip sequences have been identified, Algorithm 3 proceeds to find the super-sequences among these (sequences that are not a substring of any other sequence): `"A,B,C"`, `"A,B,E`, `"B,D,F,G"`, `"B,F,G"`, and `"D,E"`. Next, the super-sequences are grouped by origin and destination (first and last elements of the sequence respectively) – resulting in four groups labelled 1 through 4: `"A,C"`, `"A,E"`, `"B,G"` and `"D,E"`. If the obtained super-sequences were left ungrouped, the resulting corridor set would contain two elements with the same origin and destination, `"B,D,F,G"` and `"B,F,G"`, contradicting the principle that a corridor serves a unique function in connecting two distinct locations in a road network.

The last step of Algorithm 3 splits the super-sequences into pairs, and obtains a list of distinct edges per group. For example, the super-sequences `"B,D,F,G"` and `"B,F,G"`, in

group `"B,G"`, are split into `("B,D","D,F","F,G")` and `("B,F","F,G")` respectively, and combined into the edge list `("B,D","B,F","D,F","F,G")` that defines the corridor graph. If the combination of the two edge lists would not remove duplicate edges, then `"F,G"` would appear twice in the final edge list and misspecify the corridor graph.

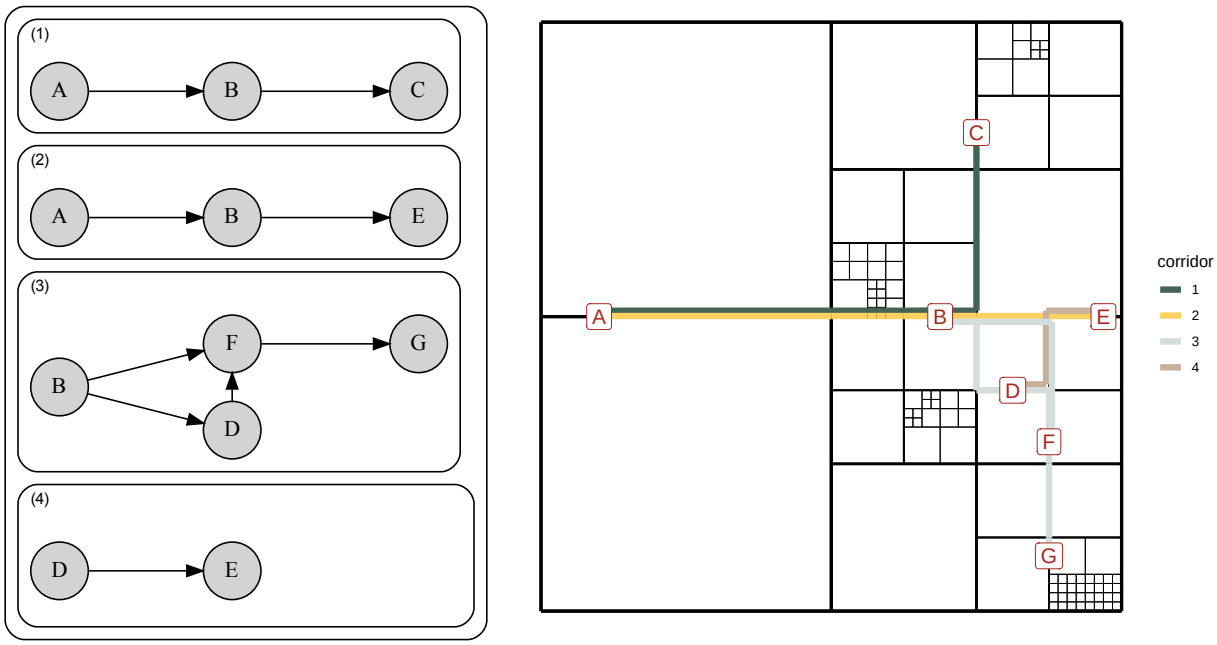
Table 4.1: Synthetic trip sequence counts obtained Algorithm (Phase 1 of the labelling algorithm).

sequence	l	N (step 2)	N (step 4)	utility	rate	simple path	ordinary ($\epsilon_r = 5, \epsilon_u = 0.75$)
A,B	2	100	300	1.00	30	T	T
B,C	2	100	200	1.00	20	T	T
B,D	2	100	250	1.00	25	T	T
B,E	2	0	110	1.00	11	T	T
B,F	2	50	150	1.00	15	T	T
C,B	2	0	10	1.00	1	F	F
D,C	2	0	10	1.00	1	F	F
D,E	2	100	110	1.00	11	T	T
D,F	2	0	150	1.00	15	T	T
E,G	2	0	20	1.00	2	F	F
F,G	2	0	200	1.00	20	T	T
G,E	2	10	10	1.00	1	F	F
A,B,C	3	100	100	1.00	10	T	T
A,B,E	3	100	100	1.00	10	T	T
B,D,F	3	50	150	1.00	15	T	T
B,E,G	3	10	10	0.70	1	F	F
B,F,G	3	100	100	1.00	10	T	T
D,C,B	3	10	10	0.29	1	F	F
D,E,G	3	10	10	0.41	1	F	F
D,F,G	3	0	100	1.00	10	T	T
B,D,F,G	4	100	100	1.00	10	T	T

The resulting corridor set, containing four corridors, is depicted in Figure 4.7. Note how corridors can overlap while serving different purposes: corridors `"A,B,C"` and `"A,B,E"` have segment `"A,B"` in common but then split towards different destinations. Corridor 3, which connects locations `B` and `G`, depicts the special case of a corridor capable of recording multiple routes from origin to destination: via `D` and `F` or solely via `F`. This shows that even though two spatial routes may not be equivalent, they are categorised within the same corridor if they serve the same (trip) function.

Lastly, suppose that `"D,G"` is now considered an ordinary trip sequence. How would that affect the resulting corridor set? In that case, a fifth corridor `"D,G"` would be added to the

set. Furthermore, if sequence `"B,D,G"` was also considered ordinary, then both `"B,D,G"` and `"D,G"` would be integrated into corridor 2 simply by adding the edge `"D,G"` to its edge list. Incorporating `"D,G"` into corridor 2 is not possible if only `"D,G"` is ordinary because the resulting edge addition creates a new non-ordinary path from source to sink `"B,D,G"`, thus breaking condition 5 of the corridor definition (all paths form ordinary trip sequences). Only when `"B,D,G"` is considered an ordinary trip sequence, can `"D,G"` be merged onto corridor 2.



(a) Corridor set.

(b) Corridors and their spatial routes.

Figure 4.7: Corridor identification toy example: outputs.

In short, the toy-worked example demonstrates that corridor discovery can reveal new insights about the travel patterns of vehicles in a road network, specifically how routes are composed together to fulfil user trips. Furthermore, the synthetic example shows that the structure of the input sensor network alone is not sufficient to reveal these patterns.

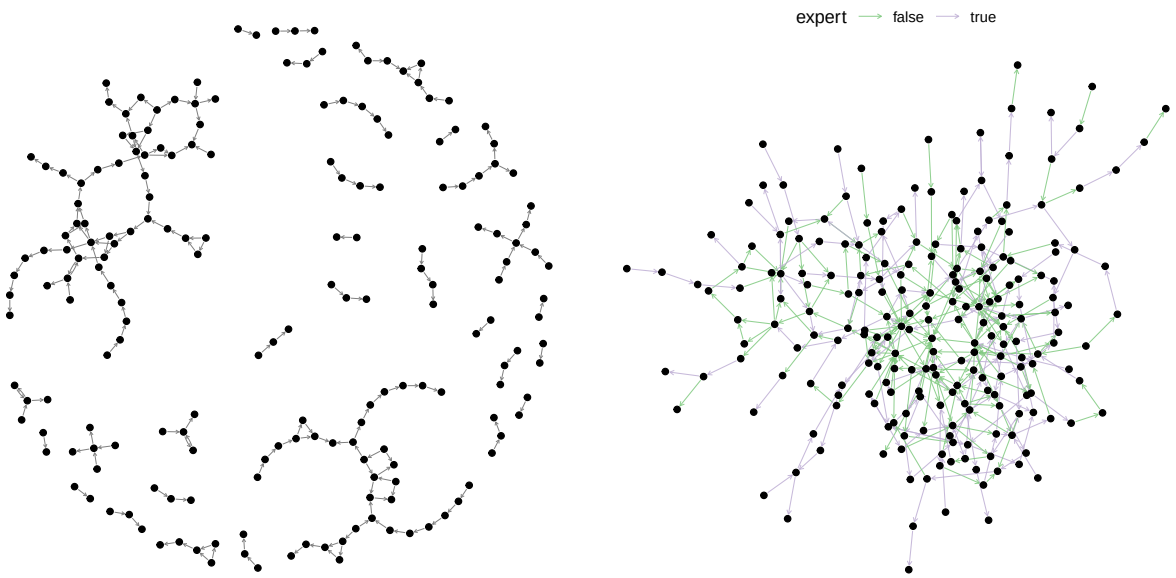
4.4 Results for the Tyne and Wear network

In this Section, we examine obtained corridors in the Tyne and Wear road network. We consider two input ANPR sensor networks with contrasting connection densities, which affect the composition of the resulting corridor set. A detailed comparison of the two input networks is given in Section 4.4.1. Section 4.4.2 explores corridor structural and spatial patterns through the analysis and visualisation of selected corridor examples. Section 4.4.3 performs sensitivity analysis by considering the joint effect of identification parameters and input network on the size and diversity of the resulting corridor set.

4.4.1 Input sensor networks

Two ANPR networks are considered. The first network G_a is built from expert annotations of monitored routes. The second network G_b is determined from data according to the route identification procedure described in Section 3.5.3, using a minimum AADT value of 512 vehicles/day as selection criteria for valid camera pairs. The total node and edge count is 191 nodes and 181 edges for network G_a, and 219 nodes and 391 edges for network G_b. In addition, G_a is a subgraph of network G_b, i.e. all edges of G_a are included in the edge set of G_b.

The structure of each network is shown in Figure 4.8. A clear difference between the two networks is that network G_a is disconnected and divided into smaller communities while network G_b is a fully connected graph. This simple visualisation shows clearly the disjointed nature in which sensor networks may operate: some regions of the road network are monitored separately from others even though there is potential for new camera connections.



(a) G_a, obtained from expert annotations. (b) G_b, obtained from data.

Figure 4.8: Annotated and identified Tyne and Wear sensor graphs.

The increase in camera connectedness is further exhibited by the in and out node degree[4] distributions of the two networks, shown in Figure 4.9. Whereas the maximum node degree is two in network G_a, network G_b exhibits a maximum node degree of ten and a long tailed degree distribution, both for in and out degree metrics. Nodes with higher degree represent cameras placed in higher sensor density areas and centralised traffic locations – a structural feature that the network G_a fails to capture.

[4]Recall that the degree of a node corresponds to the number of edges it connects to. For a directed graph, nodes have in degree equal to the number of incoming edges and out degree equal to the total of outgoing edges (Diestel, 2017).

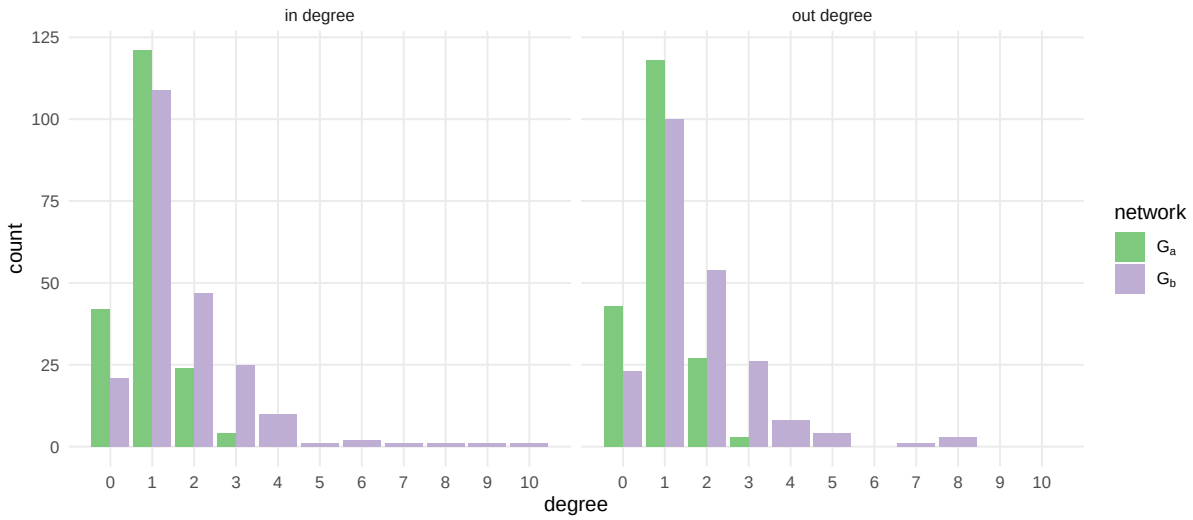


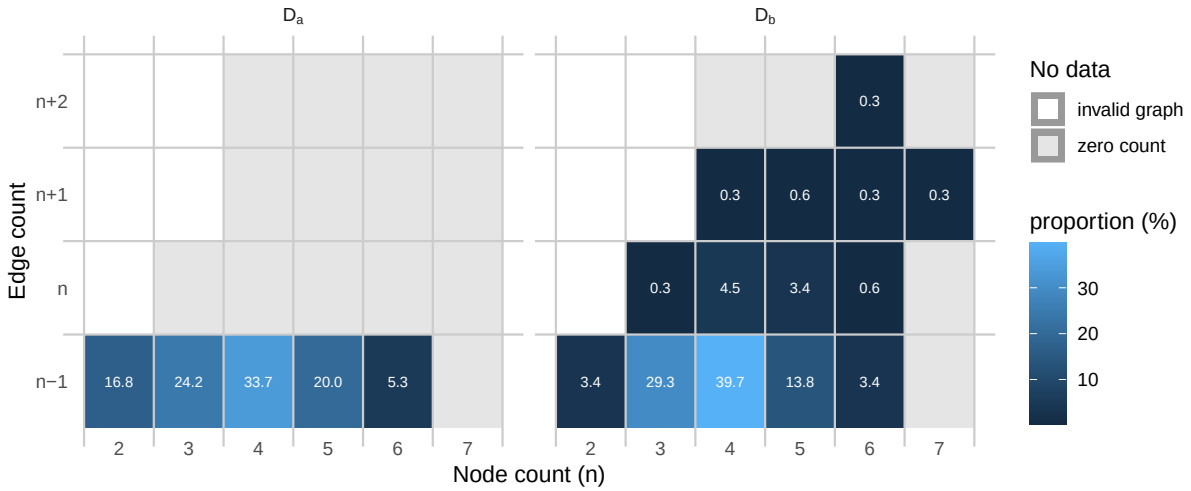
Figure 4.9: In and out degree distribution for the two input sensor graphs.

A limitation of this comparison is that operators did not explicitly reference camera pairs in their annotations, but only specific locations delimiting the start and end of monitored routes. Thus, from the specified locations we had to infer camera pairs based on the distance to nearby camera pairs and direction of traffic flow, similarly to the procedure used to map match camera locations, described in Section 3.3.3. As a result, the subset of expert camera pairs cameras may not reflect the complete set of annotated routes.

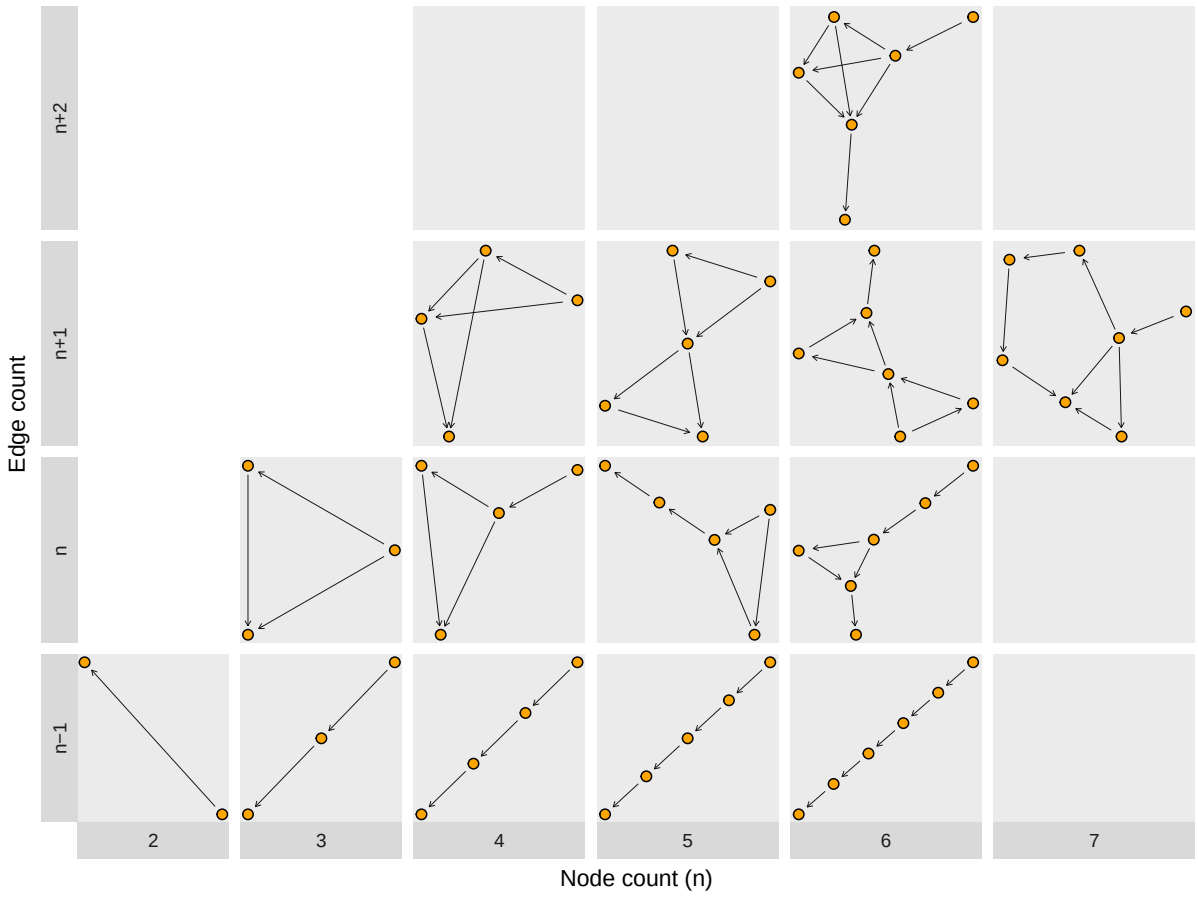
4.4.2 Corridor structure

The size and structural differences of networks G_a and G_b have a strong effect on the outcome of corridor detection. For parameter values $\epsilon_r = 100$ veh/day and $\epsilon_u = 0.75$, the corridor sets of networks G_a and G_b, denoted $\mathbb{D}_a$ and $\mathbb{D}_b$ respectively, yield 108 and 355 elements each. The difference in resulting corridor set size is not only attributed to the lower number of edges in G_a compared to G_b, but also the fact that the set of all simple paths in G_a is strongly limited by its subgraphs (which is not the case in G_b).

Naturally, it is impractical to manually verify each corridor in sets containing hundreds of elements, particularly set $\mathbb{D}_b$. Instead, we examine different corridor graph structures and the frequency with which they occur as a means of describing the structural richness of corridor sets and the capability of the underlying sensor networks in capturing diverse user behaviour. As described in Section 4.2.1.4, corridors can have different graph structures: any corridor or, more generally, any DAG with n nodes has at least $n-1$ edges and at most $\frac{n(n-1)}{2}$ edges (Diestel, 2017). Corridors with $n-1$ edges have "linear" structures matching exact trip sequences, whereas corridors with a greater number of edges can represent more complex spatial patterns, such as the existence of multiple routes from origin to destination.

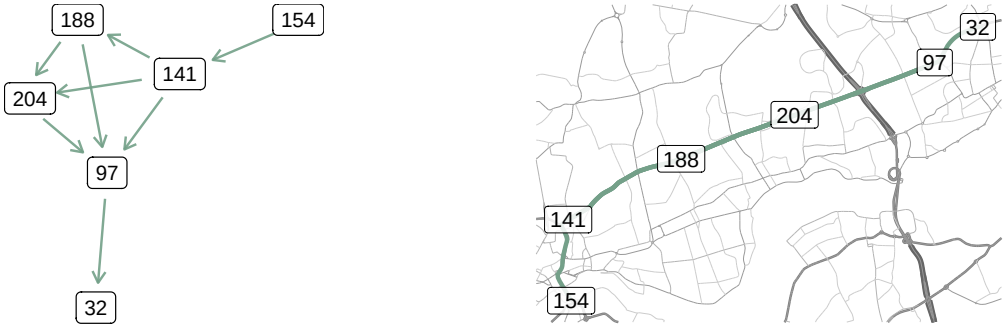


(a) Frequency of corridor occurrence in sets $\mathbb{D}_a$ and $\mathbb{D}_b$.

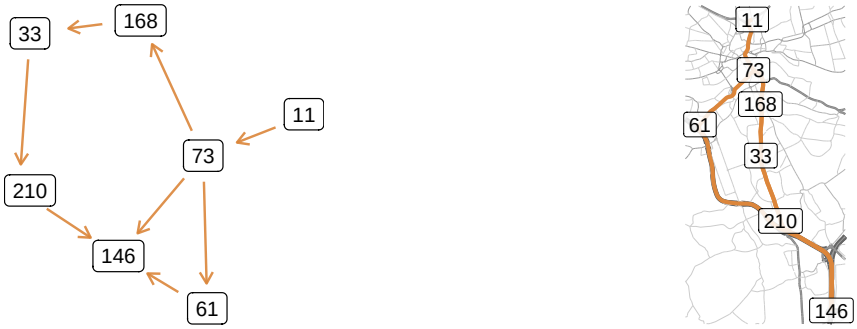


(b) Corridor graph structures in $\mathbb{D}_b$.

Figure 4.10: Frequency of occurrence and structure of corridors, grouped by number of nodes and edges, in sets $\mathbb{D}_a$ and $\mathbb{D}_b$.

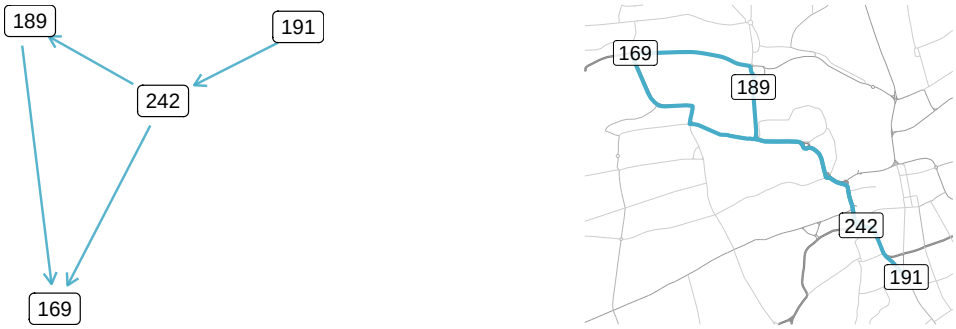


(a)



(b)

(c)



(d)

Figure 4.11: Example of four corridors in $\mathbb{D}_b$ with graph edge count equal or greater to the node count n.

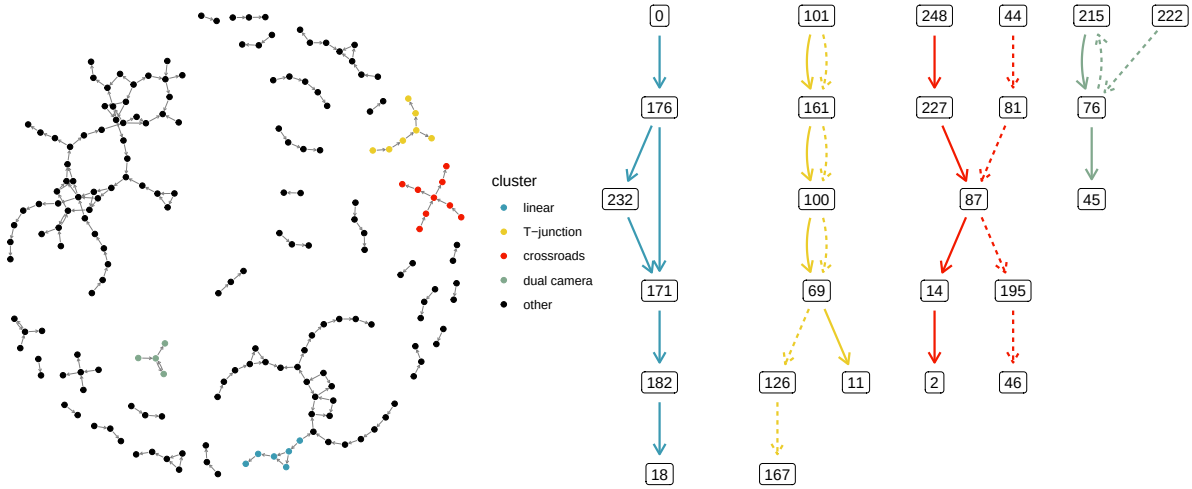
Figure 4.10a depicts the frequency of corridor occurrence by number of nodes and edges, for each of the corridor sets $\mathbb{D}_a$ and $\mathbb{D}_b$. All elements in set $\mathbb{D}_a$ correspond to corridors with $n-1$ edges, that is, simple trip sequences. In contrast, $\mathbb{D}_b$ shows greater diversity of corridor structure: 10.4% of corridors (37 in total) have between n and $n+2$ edges, albeit against a majority of corridors with $n-1$ nodes (89.6%, 318 in total). Corridor structures for each category observed in set $\mathbb{D}_b$ are illustrated in Figure 4.10b: graphs with $n-1$ edges have a linear structure (no other structure is possible); graphs with n edges have an extra edge; and graphs with $n+1$ or $n+2$ edges can have more arbitrary structures within the formal constraints defined in Section 4.2.1.4, possibly linked to higher route choice complexity.

Examples of corridors with edge count equal or greater than node count are shown in Figure 4.11. These are characterised by the existence of multiple paths from the source node (origin) to the sink node (destination). However, as previously discussed, these do not automatically imply the existence of additional routes from origin to destination. For example, corridors 4.11a and 4.11c have several possible paths from source to sink but the resulting shortest-path route is always the same. In contrast, corridors 4.11b and 4.11d exhibit multiple shortest-path routes from origin to destination. Corridor 4.11b has cameras located along each of the two alternative paths, allowing route choice behaviour to be observed more clearly. Corridor 4.11d, on the other hand, does not have cameras located along each of two alternative paths. Corridors exhibiting multiple possible shortest paths are found when their source-to-sink paths present different values of route utility factor (suggestive of different spatial routes).

To complement the examination of corridor structure, it is useful to look at how node clusters and their connections are decomposed into corridors, similar to the toy example in Section 4.3. Furthermore, it is of interest to see how changes in the parameter values affect the outcome of corridor identification. To that end, we select subgraphs from G_a and examine the resulting corridors in $\mathbb{D}_a$. This analysis is more easily performed on network G_a rather than network G_b since G_a is already split into clusters.

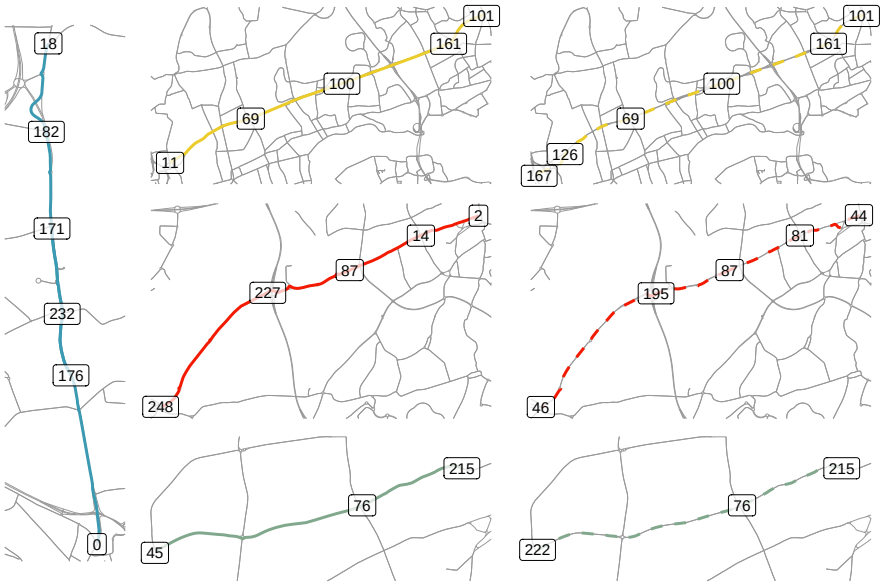
Figure 4.12 depicts the chosen graph clusters and resulting corridors: Subfigure 4.12a highlights the chosen subgraphs in G_a, Subfigure 4.12b visually amplifies the clusters and identifies the respective corridors, and Subfigure 4.12c plots their spatial routes. The selected subgraphs are simple with structures that limit possible corridor outcomes. As before, these smaller examples not only provide insight into the process of corridor identification, but also represent common patterns found in bigger sets like $\mathbb{D}_b$. Each selected graph is labelled according to its geometric structure and can be described as follows:

- Subgraph *linear* has a continuous structure, with single origin and destination nodes,

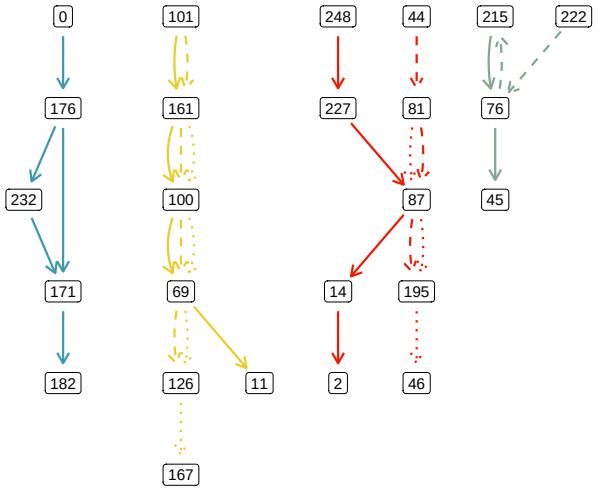


(a) Selected clusters in G_a (total of 4).

(b) Resulting corridors in $\mathbb{D}_a$ (total of 6).



(c) Spatial routes of resulting corridors.



(d) Resulting corridors (total of 9) if ϵ_r is changed from 10 to 50 veh/day.

Figure 4.12: Analysis of selected subgraphs in G_a. Corridors within the same cluster (colour) are differentiated by linetype (solid, dashed, dotted).

that maps to a single corridor. Camera `232` is an "optional" intermediary node – optional not in the sense of providing an alternative route from `176` to `171`, but of detecting passing vehicles.

- Subgraph *T-junction* represents a structure with a common source node `101` but different sink nodes `167` and `11`. The result is two distinct corridors that overlap considerably – they have 3 edges in common (out of 4 and 5 edges in total) with intermediary node `69` acting as a juncture point.
- Subgraph *crossroads* has a distinct X-shape which illustrates the benefit of corridor identification. That is, from the flow graph alone, it is not possible to infer whether vehicles departing from `248` will typically travel towards destination `2` or `48`, or both, since any of those paths is allowed in the graph.
- Subgraph *dual camera* exemplifies a structure wherein one node can act simultaneously as sink and source node, therefore requiring a split of two corridors: one in which node `215` is the source node and another where it is the sink node.

Observe that the subgraph's geometric structure will commonly not translate to an equivalent geographical shape in the road network. For instance, the X-shaped *crossroads* subgraph (red) results in two corridors with seemingly overlapping spatial routes but which relate to traffic flowing in opposite directions. Additionally, corridors overlap when they share one or more edges. For instance, in subgraph *t-junction* overlap is inevitable as there is one source node connected to two sink nodes through a common node trail. At the same time, the corridor overlap is minimal as at least two corridors (the amount observed) are required to connect one source node to two sink nodes (there is only one source node and one sink node in a corridor, as per its definition). Corridor overlap is also minimal in subgraphs *crossroads* and *dual camera*. Figure 4.12d shows how corridor formation changes by updating the value of parameter ϵ_r to 150 veh/day. In this case, corridor overlap is not minimal – clusters *t-junction* and *crossroads* have three instead of two corridors each – which suggests a less "optimal" corridor grouping than previously.

The idea that a corridor set with less overlap is preferable to a set with more overlap for the same input network, suggests that certain parameter choices are preferable to others. This finding motivates a comparison between parameter choices, achieved in the next Section by performing sensitivity analysis.

4.4.3 Sensitivity analysis

The goal of sensitivity analysis is to evaluate more rigorously how sensitive the output is to changes in the input graph and parameter values. To that end, we run the labelling algorithm for each of the two input graphs, G_a and G_b, across all pair-wise combinations

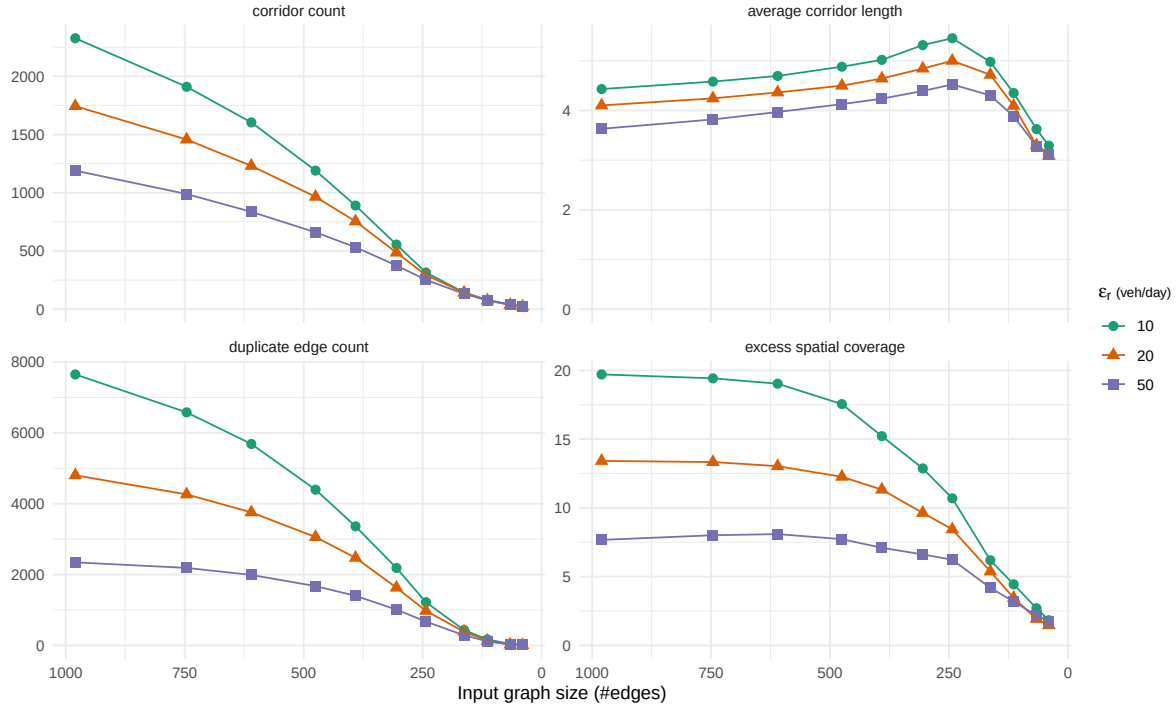
of the parameter values $\epsilon_r = [10, 50, 100, 200]$ veh/day (daily observation rate) and $\epsilon_u = [0.60, 0.75, 0.90]$ (route utility factor) – a total of 24 input/parameter combinations.

We compare the resulting corridor sets on three accounts: (a) total number of corridors, discriminated by length (number of nodes in a corridor graph), (b) total edge count and (c) excess spatial coverage. We include metrics (b) and (c) to measure the degree of corridor overlap in a given set (as shown in Figures 4.12b and 4.12d, two corridors overlap when they share an edge in common). Metric (b) sums the edge count of each corridor across all corridors in the set, while metric (c) sums their physical distances, i.e. the spatial coverage of all corridors, and divides the total by the total coverage of its input route set (edge set of the input flow graph). The calculated factor represents relatively how much excess spatial coverage results from corridor overlap.

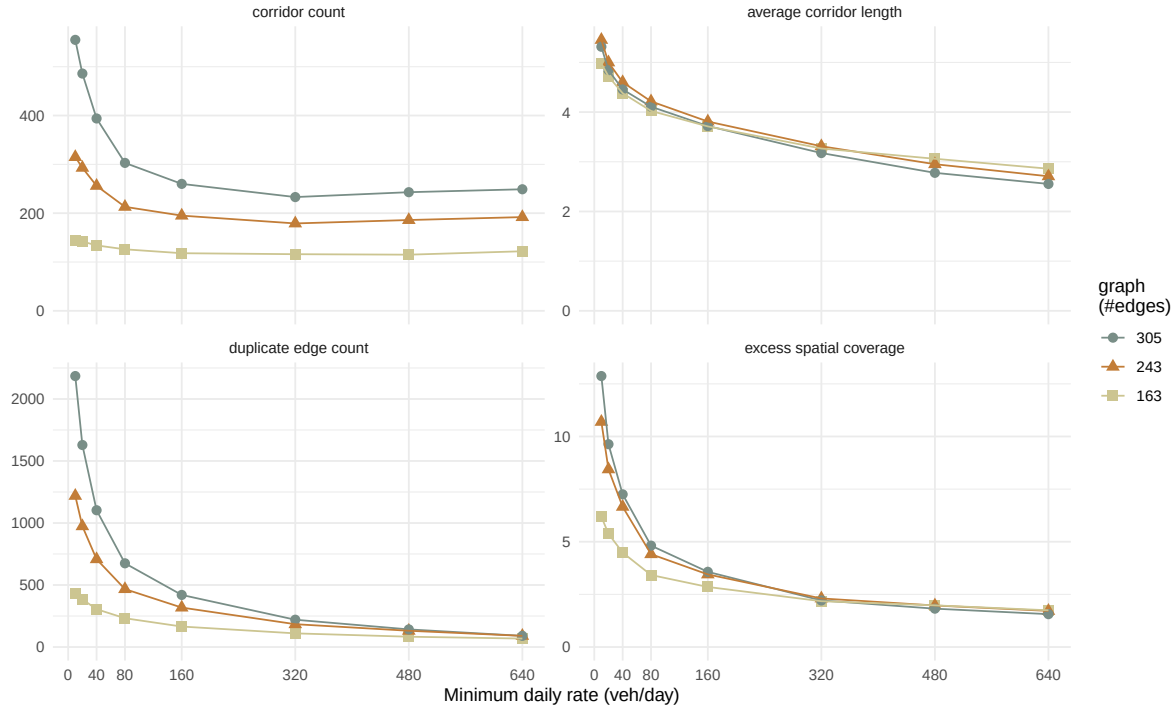
Results are shown in Table 4.2. The two input graphs produce very different results, due to their contrasting graph structures. Of interest is the effect of varying ϵ_r and ϵ_u

Inputs			**Corridor count**							**Overlap**	
G	θ_r	θ_u	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$	$l\geq 7$	total	edges	coverage
G_a	10	0.60	16	20	19	21	15	19	110	212	2.03
G_a	10	0.75	16	20	19	21	15	19	110	212	2.03
G_a	10	0.90	16	20	16	20	16	15	103	176	1.91
G_a	50	0.60	18	26	30	19	14	1	108	133	1.72
G_a	50	0.75	18	25	30	19	14	1	107	131	1.71
G_a	50	0.90	18	25	29	19	10	1	102	108	1.62
G_a	100	0.60	21	31	32	21	5	0	110	108	1.63
G_a	100	0.75	21	31	32	19	5	0	108	100	1.60
G_a	100	0.90	21	31	34	17	5	0	108	98	1.59
G_a	200	0.60	27	47	24	13	1	0	112	69	1.42
G_a	200	0.75	27	47	22	13	1	0	110	63	1.39
G_a	200	0.90	28	46	22	13	1	0	110	62	1.38
G_b	10	0.60	2	104	241	252	188	120	907	3423	8.67
G_b	10	0.75	2	104	229	253	183	119	890	3360	8.56
G_b	10	0.90	5	104	218	220	173	101	821	2998	7.65
G_b	50	0.60	8	123	211	121	39	14	516	1322	3.65
G_b	50	0.75	8	121	210	120	39	14	512	1311	3.63
G_b	50	0.90	13	125	183	109	37	10	477	1156	3.29
G_b	100	0.60	18	123	165	70	16	1	393	784	2.42
G_b	100	0.75	19	119	165	68	16	1	388	769	2.40
G_b	100	0.90	24	120	155	62	17	0	378	716	2.27
G_b	200	0.60	51	158	93	33	5	0	340	435	1.75
G_b	200	0.75	53	156	91	33	5	0	338	427	1.74
G_b	200	0.90	59	154	85	32	4	0	334	400	1.67

Table 4.2: Corridor count, for different corridor lengths l, and corridor overlap, measured as number of duplicate edges and excess spatial coverage, across various combinations of input and parameter values: flow graph G, minimum daily traffic volume θ_r (in veh/day) and minimum route utility θ_u.



(a) The input flow graph (shown as number of edges) is varied for three parameter values of ϵ_r (10, 20 and 50 veh/day).



(b) The minimum flow rate ϵ_r is varied between 10 and 640 veh/day for three input graphs.

Figure 4.13: Effect of varying different input values on corridor count (top left panel), average corridor length (top right panel), duplicate edge count (bottom left panel) and excess spatial coverage (bottom right panel). ϵ_u is set at 0.75 throughout the experiment. Due to different scales, the Y-axis variable is shown on top of each subgraph.

for each graph. In G_b, as ϵ_r increases, the number of corridors decreases rapidly as ϵ_r is increased from 10 to 20 and 100 veh/day, but only slightly when ϵ_r is doubled from 100 to 200 veh/day. Moreover, the effect of ϵ_u is less noticed than that of ϵ_r, suggesting that the output is more sensitive to changes in ϵ_r than ϵ_u. In G_a, the number of corridors remains approximately constant with increases in ϵ_r. In both graphs, the number of duplicate edges and excess coverage decrease inversely to ϵ_r. Additionally, the distribution of corridor lengths shifts progressively towards lower values.

These observations suggest that we can not maximise corridor length and minimise their overlap at the same time. That is expected, since overlap is minimised only when a corridor set equals its input graph, in which case nothing new is learned. This is further corroborated by Figures 4.13a and 4.13b, as it becomes clear that continuously decreasing the size of the input graph or continuously increasing ϵ_r, does not contribute to a "better" corridor set in the sense of longer corridors with less overlap. For input graphs, a desirable input region is found approximately between 400 and 250 edges, where its positive effect on average corridor length is maximised and its effect on corridor count and overlap is minimised. For parameter ϵ_r, the region of preference lies approximately between 80 and 320 veh/day, after which corridor overlap is no longer greatly reduced without also incurring a significant loss in corridor length (or even an increase in corridor count).

4.4.4 Key corridor sections

This Section seeks to highlight key corridor sections in the road network of Tyne and Wear. Corridor sections consist of sequences of one or more corridor routes (edges in a corridor graph), characterised using two metrics: average traffic volume, measured in daily vehicle counts, (calculated previously during corridor discovery) and road connectivity, measured in kilometres (described below). For each metric, sections of the same length are ordered such that the resulting ranking is taken to be indicative of their degree of importance in the road network. Key corridor sections then correspond to top ranked sections representative of significant two-way and higher level interactions between monitored routes, and possibly indicative of route influence in distributing traffic across the network.

The connectivity level of a corridor section is measured by considering its connections to other corridor sections, specifically by calculating the total non-overlapping distance of its spatial route neighbours. The connectivity measurement process is exemplified in Figure 4.14, where the goal is to compute the connectivity of section `117,226,138`, shown in black. In the example, a total of four corridors were found to contain the selected section. Its neighbour routes are found by removing the selected section from each corridor and keeping the remaining corridor components, here categorised as occurring before (dark green) or after (light blue) the selected section. For instance, neighbour sections `138,192`

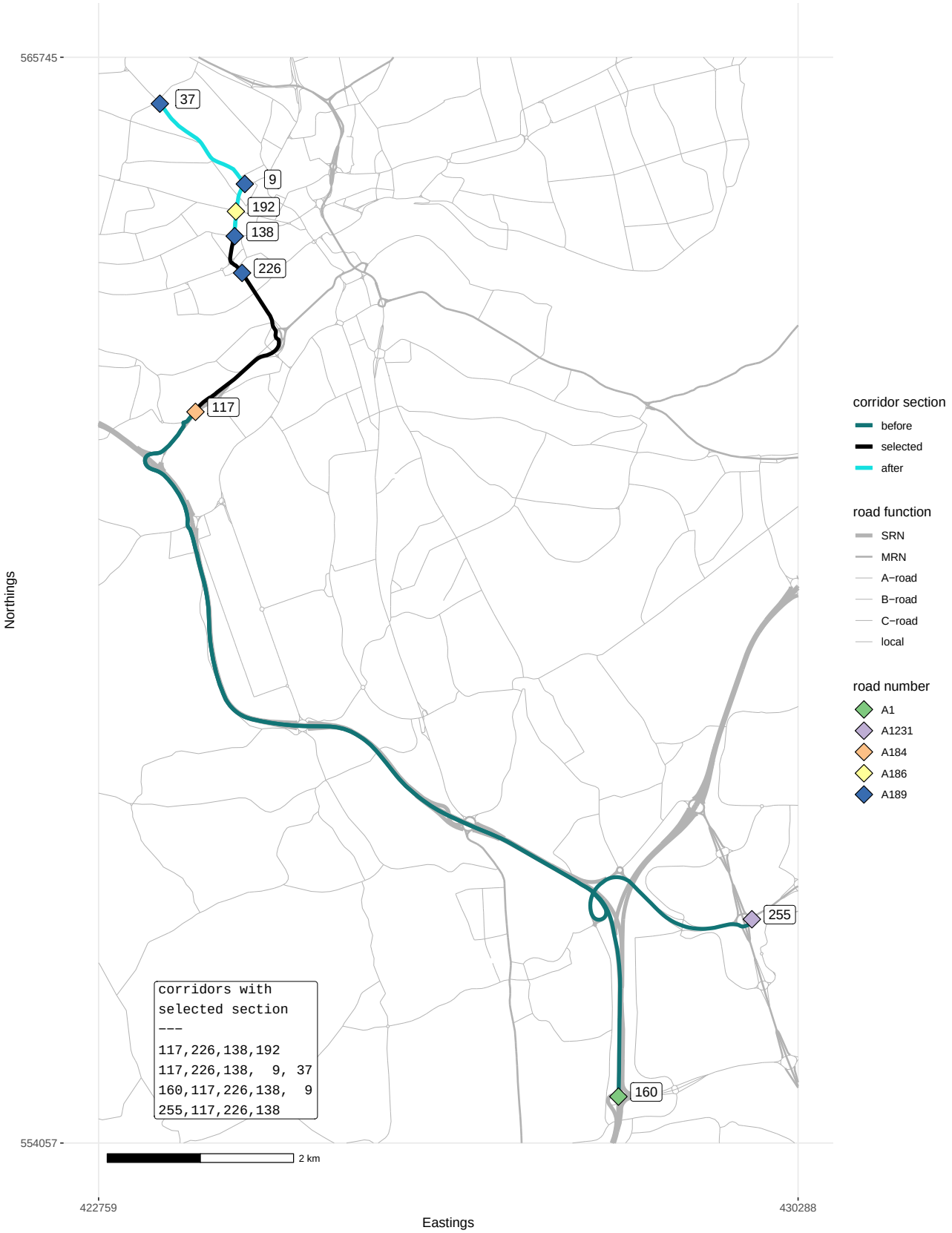
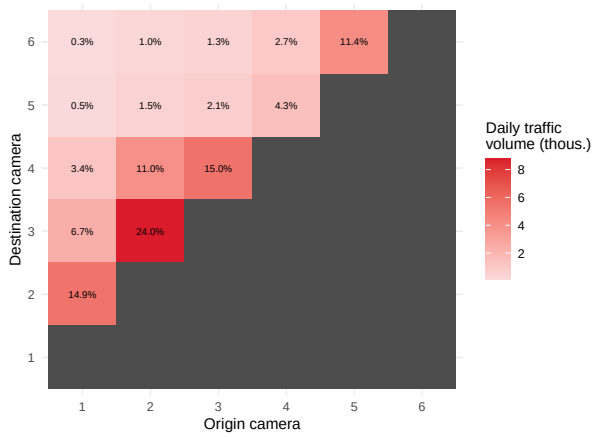


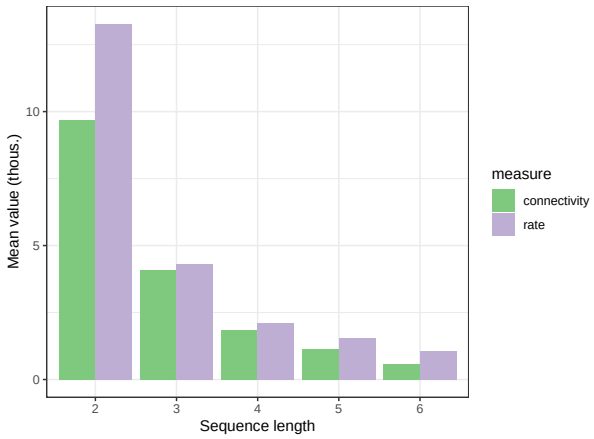
Figure 4.14: Illustration of a corridor section, shown in black and specified by the sequence `117,226,138`, and its spatial neighbour sections, obtained by removing the selected section from all corridors (bottom left panel) that contain it. Neighbours are shown in two different shades of green, depending on whether they appear before or after the selected, and represent meaningful road connections to the selected corridor section, whose connectivity value is thus computed as the non-overlapping distance of its neighbour sections (non-overlapping sum of before and after road segments).

and `138,9,37,34` occur after the selected section, whereas neighbour sections `169,117` or `255,117` occur before. The connectivity value of the selected section then amounts to summing the distance of the unique *before* and *after* spatial segments (which can be obtained through a geometrical union operation). This process is repeated for all unique corridor sections, across all registered corridor lengths.

Ranking of corridor sections is performed only for sections of the same length (equal to the number of routes). Sequences of different lengths are not directly comparable since observation rate and road connectivity decrease with section length, as shown in Figure 4.15. The same rationale applies to the display of corridor sections rather than whole corridors (as these can vary both in length and structure). To reduce visual overlap and allow direction of travel to be identified, sections are categorised as being primarily North/Eastbound or South/Westbound oriented based on the cardinal direction attributes of its composing cameras. Note that for any given corridor there is not necessarily a corresponding similarly ranked corridor in the opposite direction of travel – the existence of an 'opposite' corridor will depend on camera coverage, while rankings can be affected by imbalances in traffic demand (i.e. one direction of travel may be 'preferred' over the other).



(a) Origin-destination matrix for a corridor composed by 6 cameras (5 routes). Shown are daily traffic vehicle counts and respective percentage values for the different sections of the corridor, as vehicles can start and terminate their trips at different points along it. For any given origin location, the longer the section travelled (seen vertically), the lower the volume of traffic (as a proportion of vehicles will always terminate their journey along the way).



(b) Mean sequence rate and connectivity value per sequence length. The observed decrease is explained by the decrease in trip frequency with trip length (as listed in Table Table 3.20).

Figure 4.15: Effect of corridor section length on observed traffic volume.

Figures 4.17, 4.18 and 4.19 depict the top 10 highest ranked corridor sections by daily traffic volume and road connectivity, respectively for sections of length 2, 3 and 4 routes (3, 4 and 5 camera sequences). Of the highlighted corridor sections, several roads consistently

rank highly across section lengths and metrics (labelled individually in Figure 4.16):

- The A1058 (Coast road) is a two-lane dual-carriageway A-road connecting the east part of Newcastle city centre to Tynemouth in North Tyneside. Due to its major role in connecting the two counties and carrying significant volumes of traffic, it is part of the national Major Road Network (MRN).
- A184 Felling Bypass is the primary traffic gateway to (from) Newcastle from (to) South Tyneside and Sunderland. Is is also part of the MRN because of its high connectivity and traffic throughput role across counties.
- The A167 (M) is a short motorway section in Newcastle city centre, which connects traffic from/to Tyne Bridge to all directions, include Eastbound traffic (Coast Rd), Northbound corridors, Southbound. Its central role means that it is a common component in many corridors and a primary source of spatial overlap.
- The A1(M), A184, A169 is a North/Southbound route connecting the West parts of Gateshead and Newcastle City Centre (St James Boulevard) via Redheugh Bridge. It is the only high ranked corridor that includes a section of the Strategic Road Network (the A1 motorway stretch).
- The A167 Durham Road serves North/Southbound traffic in Gateshead, terminating/starting at Newcastle via Tyne Bridge.
- The A189 Newcastle is a North/Southbound corridor that follows after St James Boulevard.
- The B1318 (Great North Road) is an alternative North/Southbound corridor that connects Newcastle city centre with the North region of Newcastle (Jesmond, Gosforth) and links with the A1 towards Northumberland.
- The A189 (North Tyneside) and A1018 (Sunderland) are examples of other highly ranked length two corridor sections.

To evaluate whether a corridor-based ranking agrees with a characterisation of route importance based on individual routes, we characterise and visualise individual ANPR routes according to the above metrics of daily traffic volume and connectivity. Additionally, routes are described using edge betweeness centrality, a graph theoretical measure of road influence and network resilience (Demšar et al., 2008), calculated on the ANPR network graph with different edge weight choices (no weight, shortest path distance, AADT). Figures 4.20 and 4.21 depict ANPR monitored routes coloured by, respectively, daily traffic volume and connectivity, and (weighted) edge betweeness centrality. Figure 4.20c is included to show that connectivity is closely related to route overlap (expectedly, as per the definition of connectivity) but not exactly equal.

A graphical comparison of top corridor sections and routes suggests a degree of agreement between highlighted roads, particularly when both are described using traffic volume and

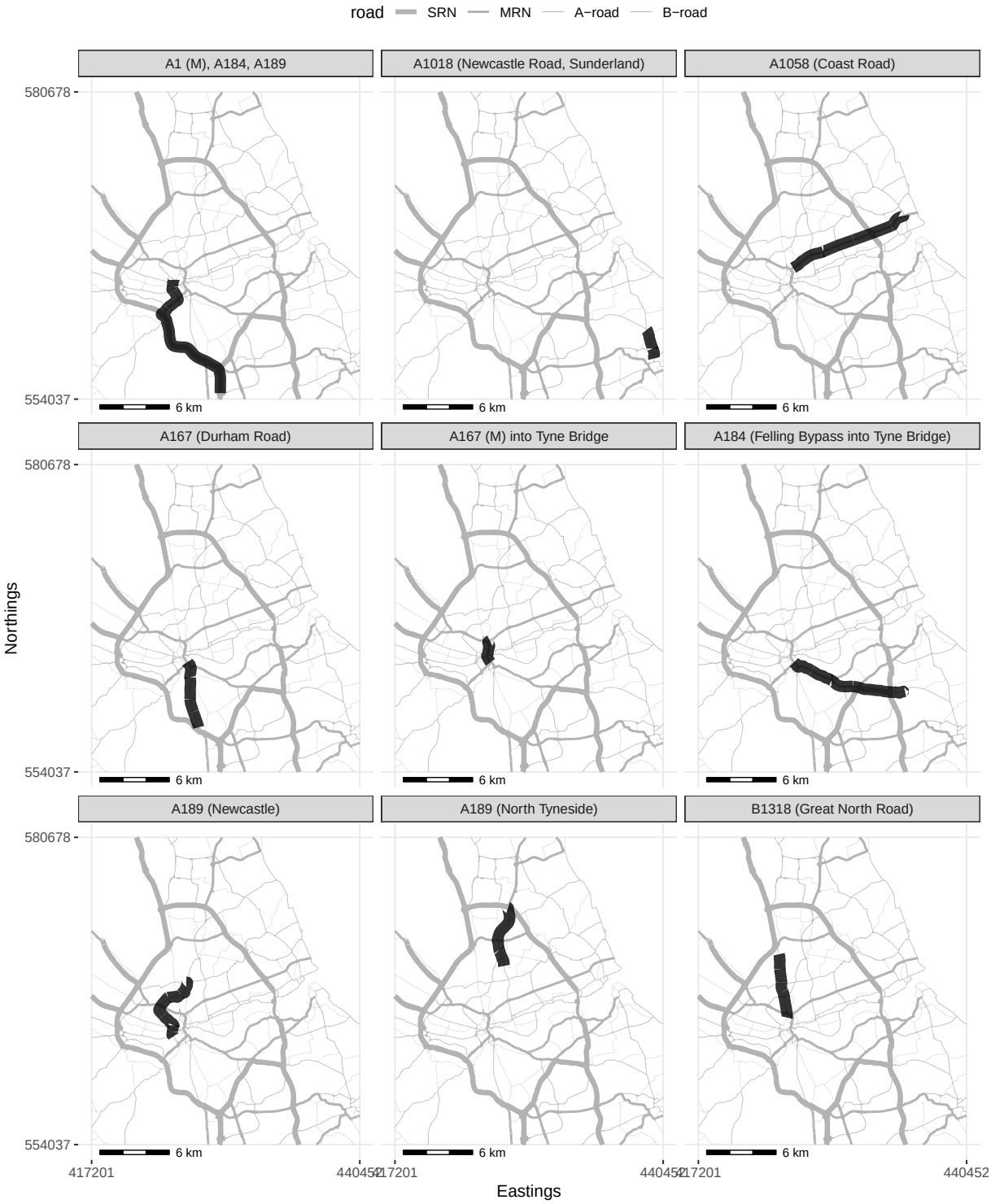
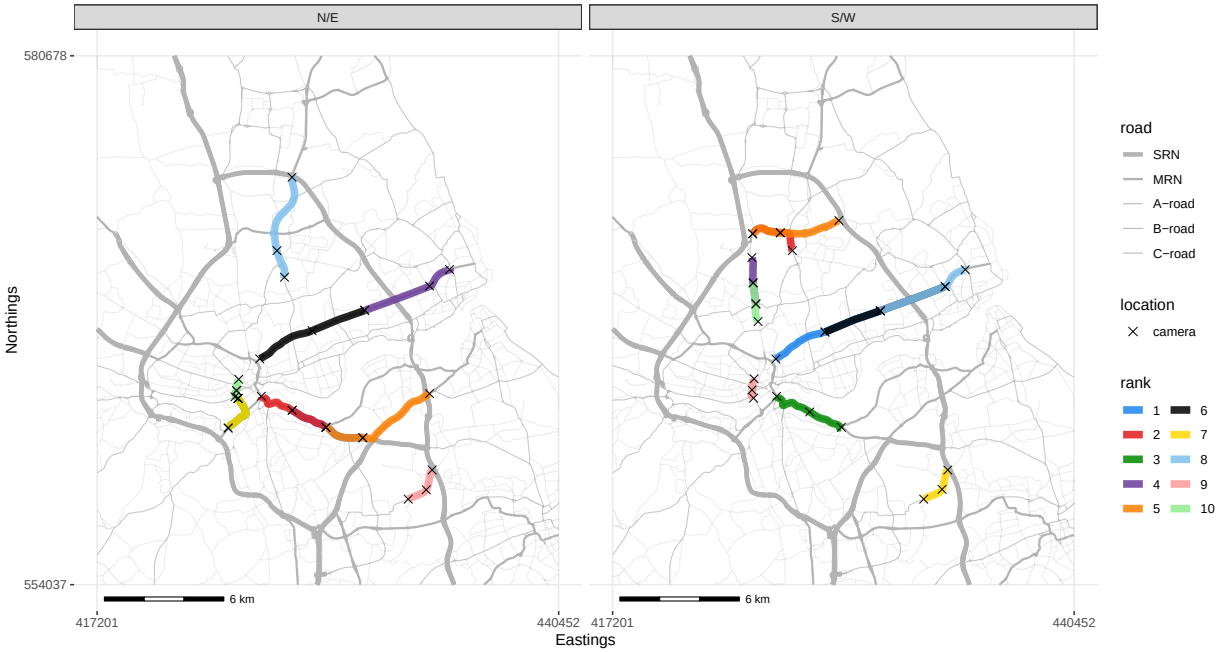
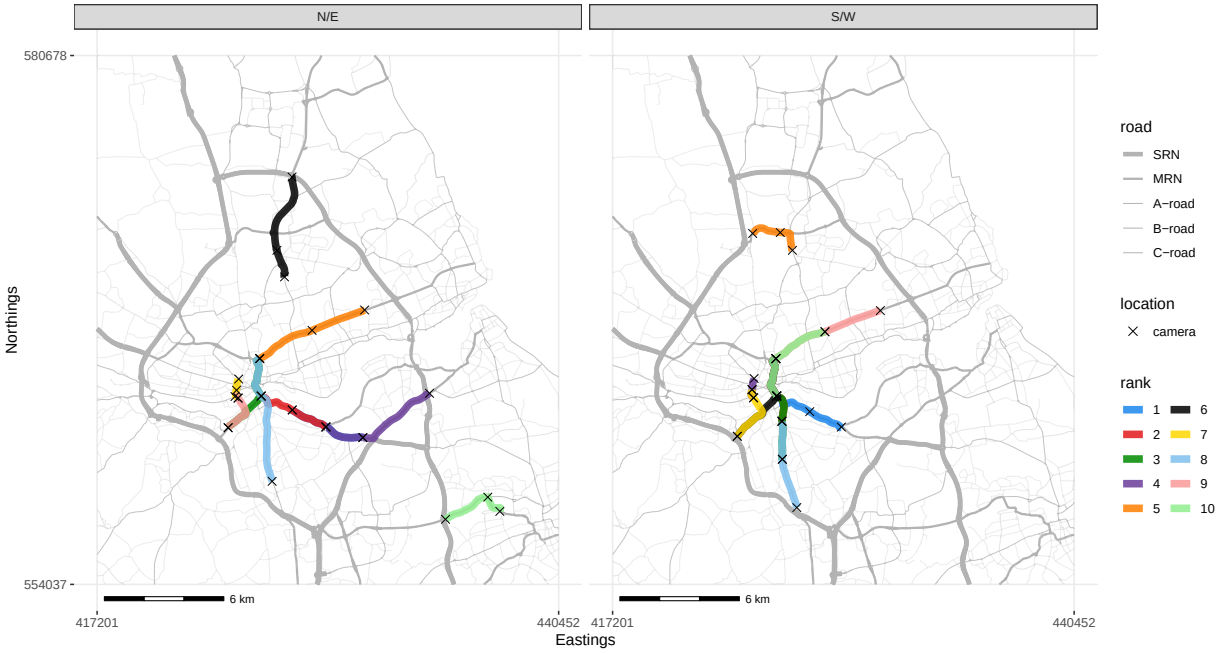


Figure 4.16: A subset of corridor sections that ranked highly across different section lengths and metrics.

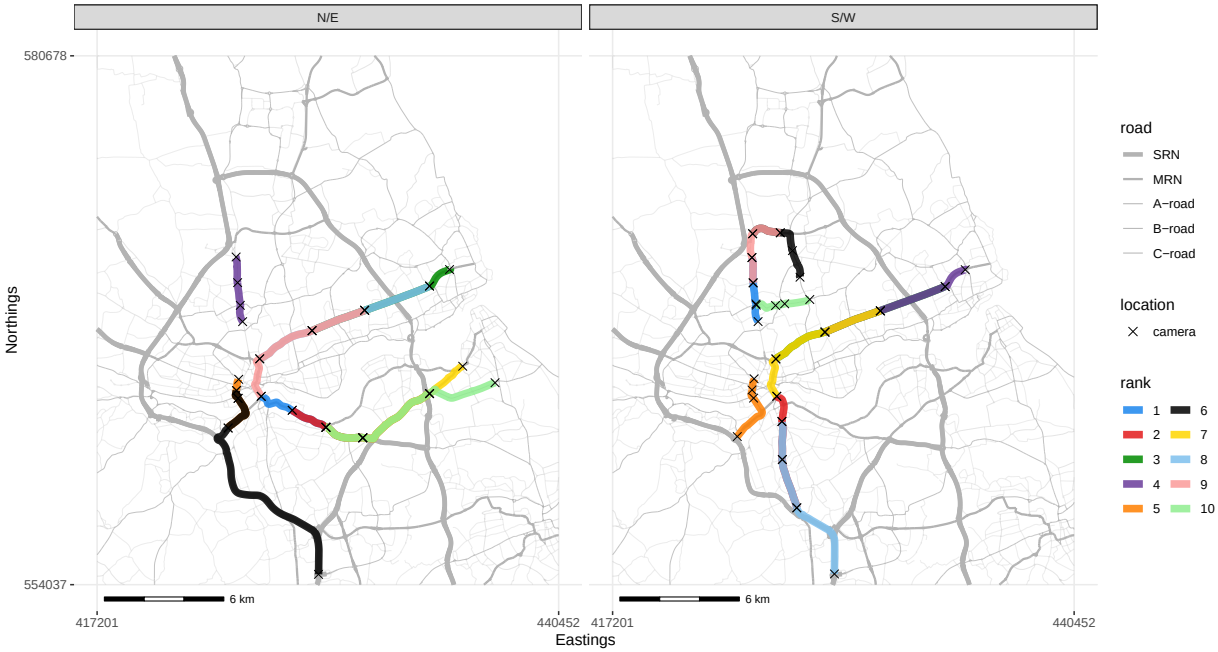


(a) By daily volume of traffic.

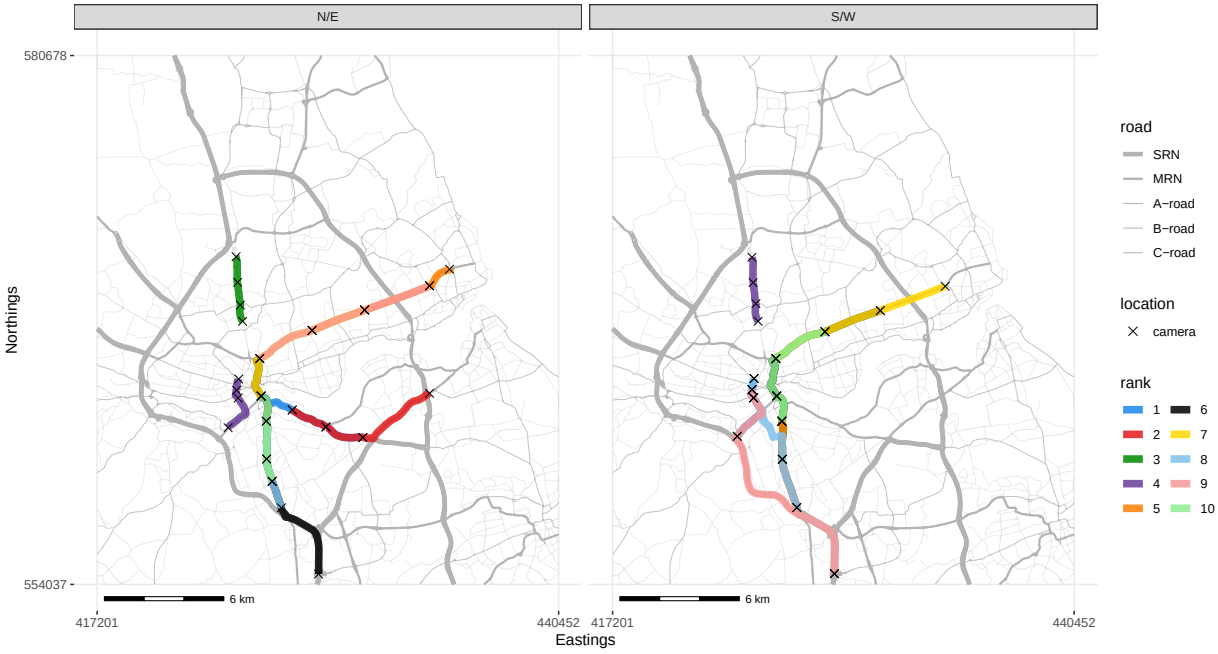


(b) By degree of spatial connectivity.

Figure 4.17: Top 10 highest ranked corridor sections of length 2 routes (3 cameras), faceted across opposite travel directions (North/Eastbound versus South/Westbound traffic).

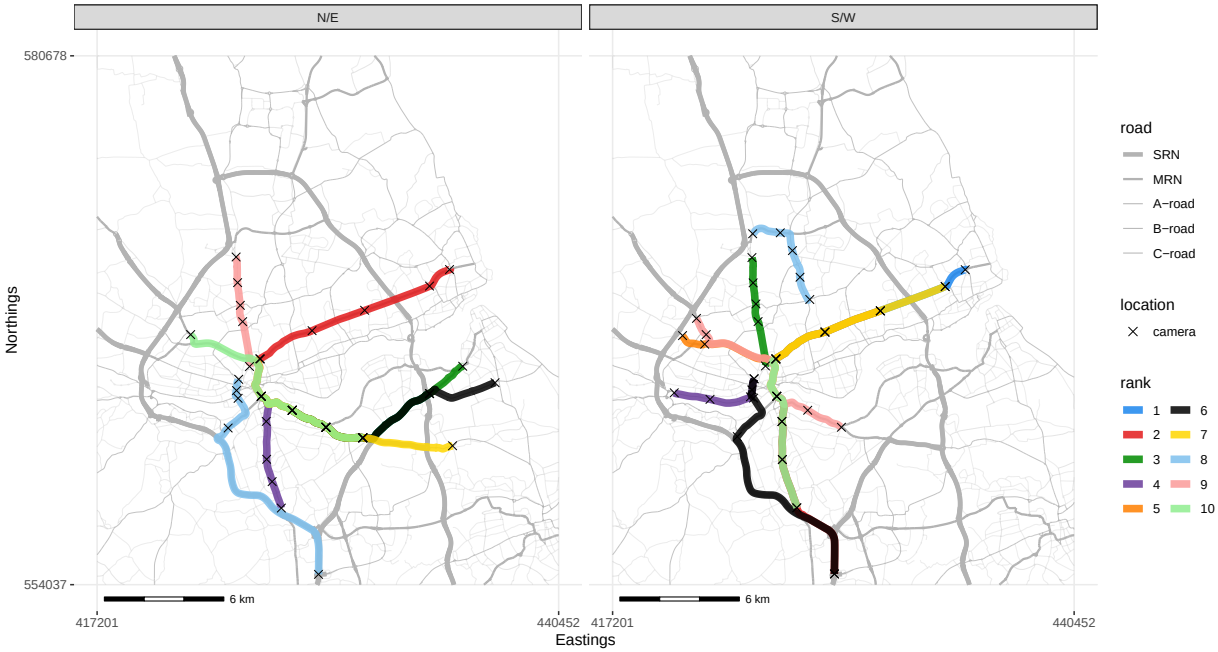


(a) By daily volume of traffic.

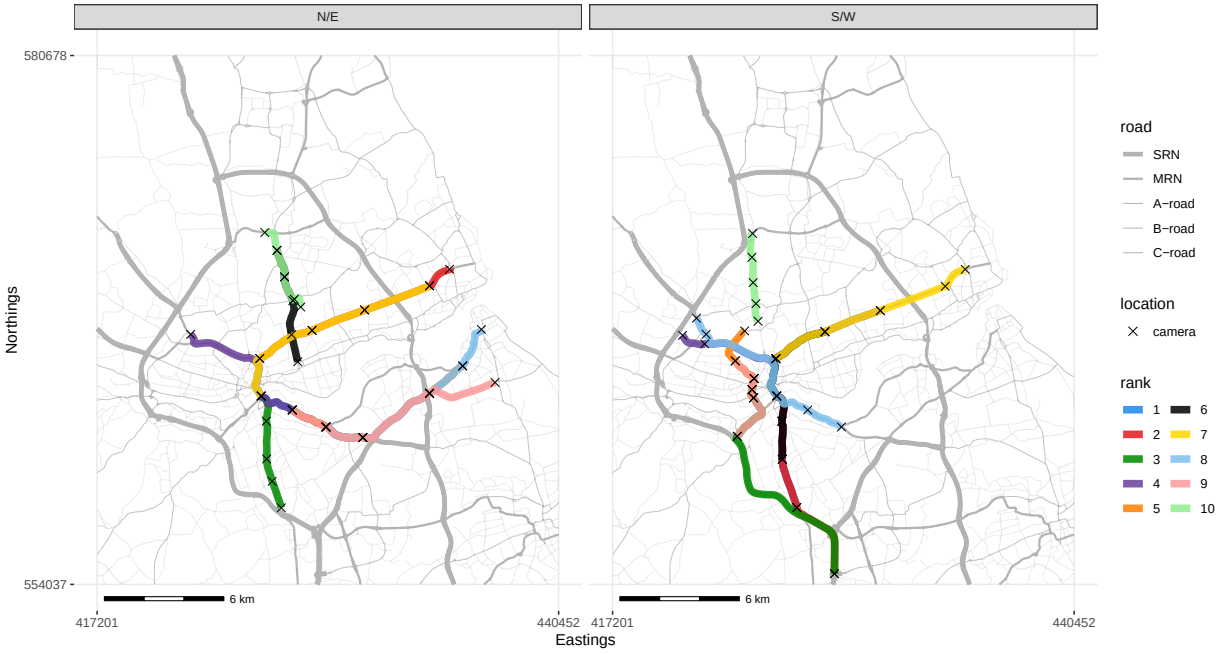


(b) By degree of spatial connectivity.

Figure 4.18: Top 10 highest ranked corridor sections of length 3 routes (4 cameras), faceted across opposite travel directions (North/Eastbound versus South/Westbound traffic).

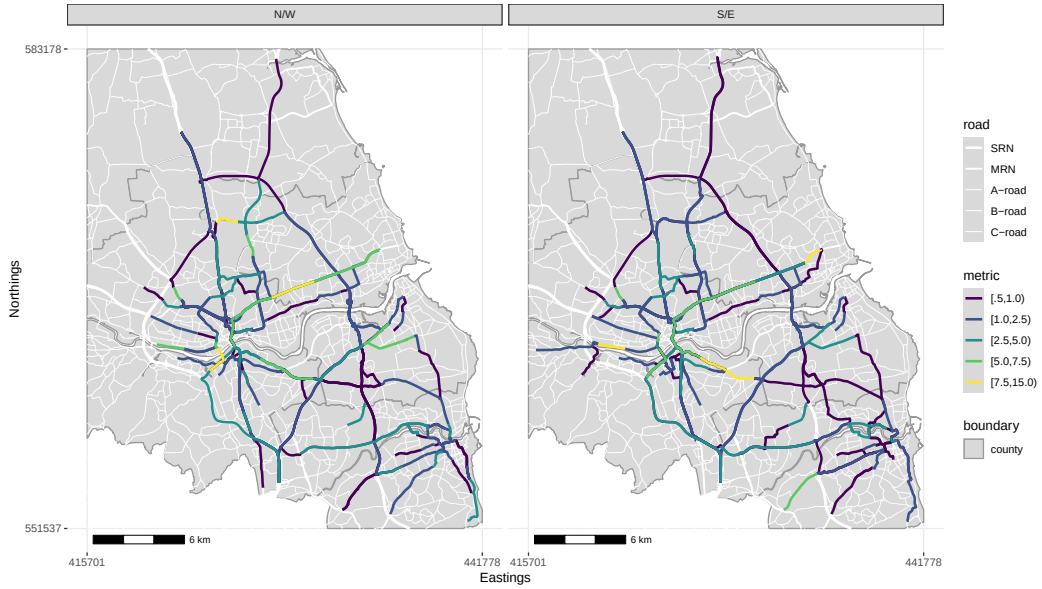


(a) By daily volume of traffic.

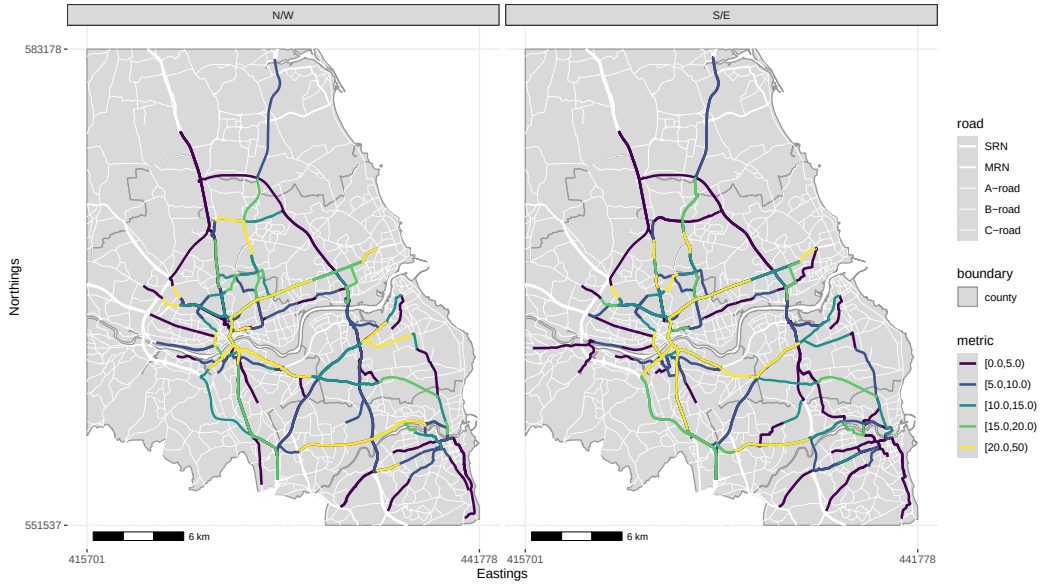


(b) By degree of spatial connectivity.

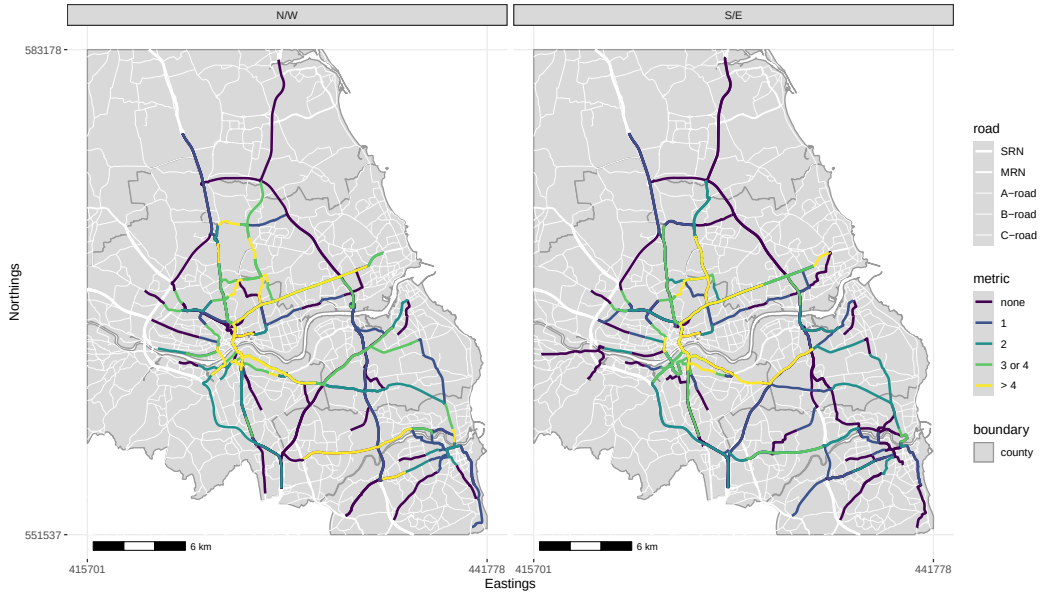
Figure 4.19: Top 10 highest ranked corridor sections of length 4 routes (5 cameras), faceted across opposite travel directions (North/Eastbound versus South/Westbound traffic).



(a) By daily volume of traffic.

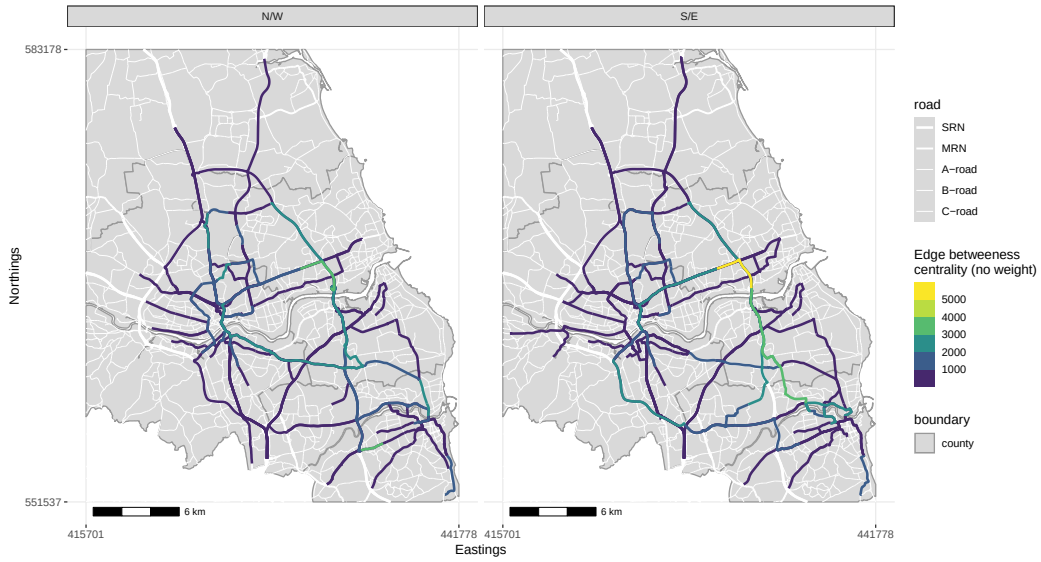


(b) By degree of spatial connectivity.

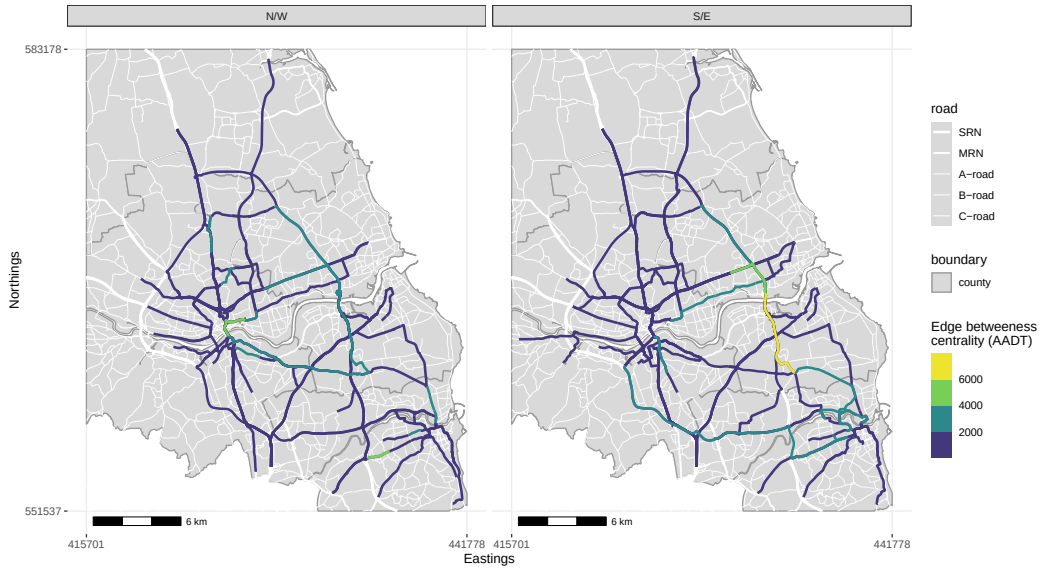


(c) By spatial overlap.

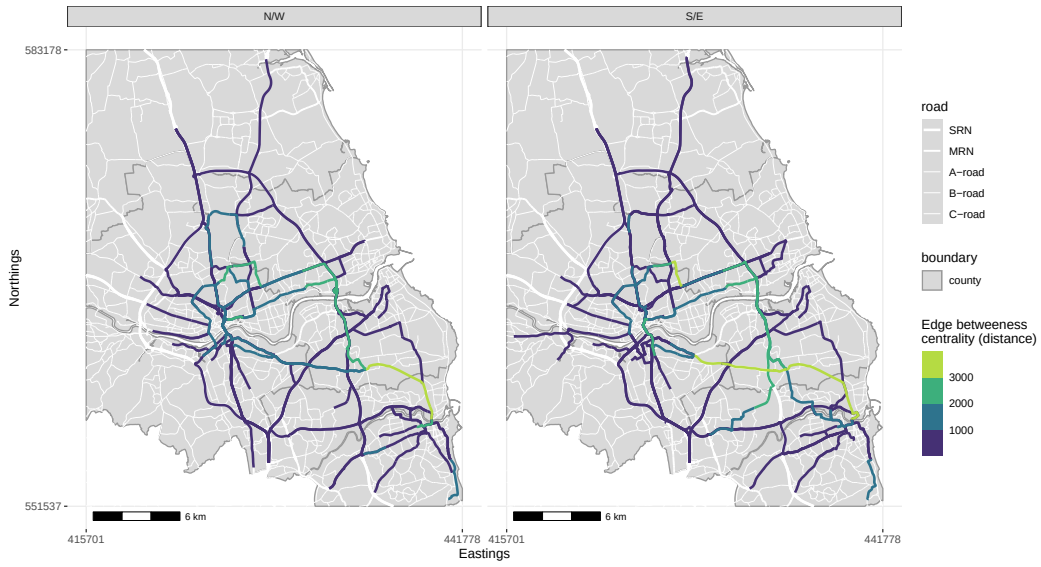
Figure 4.20: ANPR routes according to different measures, faceted by travel direction.



(a) No weighting.



(b) Edges weighted by value of AADT.



(c) Edges weighted by route length.

Figure 4.21: Edge betweeness centrality of ANPR routes, faceted by travel direction.

connectivity. In contrast, centrality based measures seem to be less capable of highlighting differences in corridor importance, particularly across Redheugh and Tyne bridges, well known points of traffic confluence. The difference between a ranking assessment based on corridors versus individual routes is the ability to explicitly identify relationships between routes, which the latter (road/route-based assessment) is incapable of doing. Thus, while methods such as betweeness centrality (Demšar et al., 2008) have been applied successfully to quantify road importance, they are incapable of identifying the causal relationships between different routes and traffic flows.

A limitation of corridor section analysis is the inability to consider whole corridors as opposed to sections of a specific length. This results in significant overlap between top ranked corridor sections that actually belong to the same corridor and is particularly pronounced in sections of length two/three routes. Visible examples are corridors A1058 (Coast Road), A84 (Felling Bypass) and A1-A184, which are split into two or three shorter, overlapping sections in Figures 4.17 and 4.18, but represented wholly in Figure 4.19. On the other hand, the downside of a ranking strategy that chooses to highlight only longer sections is that corridors composed of shorter sequences may be incorrectly ignored. For instance, corridors A189 (North Tyneside) and A1018 (Sunderland) are visible in Figure 4.17 but not in Figure 4.19 since are composed only of two routes. Thus, an ideal corridor ranking mechanism would be able to consider whole corridors regardless of differences in graph length and structure, so to minimise corridor overlap and improve visualisation and interpretation of results.

4.5 Discussion and future work

This chapter develops a mechanism to identify primary travel sequences, called corridors, in ANPR sensor networks. The process of corridor identification is made difficult by the imprecise nature of trip identification, responsible for the occurrence of erroneous trip sequences. To exclude spurious trip sequences from subsequent analysis, the proposed methodology employs a combination of established graph theoretical concepts, empirical observations and spatial data. By representing corridors as directed acyclical graphs it becomes possible to group trip sequences serving the same purpose, that is, of carrying vehicles from one location to another. At the same time, frequency analysis of trip data collected over a period of time (weeks) allowed popular travel sequences to be identified, while spatial data was used to keep only the travel sequences whose routes were indicative of linear and expeditious travel, as is characteristic of road corridors.

By giving a precise definition and implementation of road corridors, we enable a range of different analyses for different traffic management and planning purposes. For example, corridor identification can offer insight into the distribution of sensors (Raines & Rowley,

2008), enable temporal analysis to show where journeys are observed to begin and end at different times of day (Morar & Baber, 2017) and help to evaluate changes in user travel patterns as a result of the Covid pandemic (De Vos, 2020). It can also provide a means to estimate route correlation for short-term traffic forecasting (as discussed in Section 2.2.2), as well as support the process of road categorisation in England, namely that of the Major Road Network (MRN), which requires compelling evidence of the critical role played by a road in connecting local regions.

The advantage of our approach compared to others is a mechanism to exclude unrepresentative and spurious travel sequences. For instance, Crawford et al. (2018) measures the spatial variability of a user's trip history by calculating the spatial differences between any two trip sequences, but does not employ a procedure to validate input trip sequences. In the wider context of this work, corridor discovery allows the identification and impact assessment of traffic bottlenecks, presented in Chapter (Section 5), to be scaled to an entire road network without relying on manual or expert annotation.

Despite its uses, the main shortcoming of the corridor identification algorithm is its inability to generate a succint set of corridors with minimal spatial overlap. Due to the combinatorial nature of the problem, corridors of different lengths can not be compared directly as the probability of occurrence inherently decreases with the length of a trip sequence. Corridor ranking results were produced for sections of different length, depicted in Figures Figure 4.17, Figure 4.18 and Figure 4.19, but could not be combined into a single output, neither visually nor computationally. This restricts the usability of the approach as the different outputs must be combined manually before a substantive analysis of corridor role and importance can be conducted.

Subsequent research work should therefore seek to develop a method of characterising and comparing corridors irrespectively of their length. One potential mechanism for treating variable trip sequence lengths is to consider a trip frequency parameter that is not fixed for all trip sequences but that is rather a function of trip sequence length. This, or a similar method of normalising frequencies by trip sequence length, could allow different trip sequences to be placed on equal footing. However, it is not clear how the algorithm can be adapted to produce a succint visual and computational representation of different yet overlapping corridors. The development of a mechanism to deal with spatial overlap in corridors, in addition to a more principled method of tuning the parameters U and ϵ_u, should also be the focus of future research.

A promising application of this work is to describe corridor function numerically, for instance, a corridor can be 80% A-road, 10% B-road and 10% MRN. Corridors could be then grouped based on their function, e.g. primarily B-road, A-road, MRN or SRN, and their importance assessed within or between groups. This would allow over and underperforming roads to be detected and nominated for categorical re-evaluation (so

that proper road funding and investment is made available to them). However, tracing road category would require accurate and precisely labelled map data (refer to Section 3.3.3 for statistics on missing road data), which were found to be available during the time of analysis.

Chapter 5

Identification and impact assessment of recurring traffic bottlenecks

5.1 Introduction

Bottlenecks are a leading source of traffic congestion in urban areas (Falcocchio & Levinson, 2015). They form choke-points that, under significant user load, cause queues to form upstream of their location, while downstream traffic flows unimpeded (Daganzo, 1997). A bottleneck is classified as recurring, if its nature is predictable (when and where it occurs and its impact on traffic flow), or non-recurring, if the underlying cause is a transient event, such as a traffic incident (Hale et al., 2016). Recurring bottlenecks are of particular interest to traffic authorities as mitigation can often be achieved without major improvements to road infrastructure (Spiller et al., 2012).

The detection and assessment of traffic bottlenecks is thus an integral activity of traffic management. Its goal is to identify bottleneck locations and collect evidence of their impact so that critical interventions can be prioritised (Hale et al., 2016). For large road networks, algorithms can assist in bottleneck assessment by examining available traffic data – known bottlenecks can be monitored continuously and new bottlenecks discovered automatically (Chen et al., 2004; Gong & Fan, 2017; Wieczorek et al., 2010). However, existing methods have been mainly developed for highways, and are not readily applicable to lower category roads despite their importance in fulfilling user trips. A key reason is the limitation of loop detectors, a widespread traffic sensor technology, to gather route-aware traffic data outside controlled-access highways (Klein et al., 2006).

In the UK, several cities now employ a large number of ANPR cameras for traffic monitoring (Debnath et al., 2014; Kitchin, 2016). The technology has two main advantages over other fixed sensors in urban environments: (i) it provides reliable travel time and speed estimates and, (ii) it allows routing choices to be partly inferred (Friedrich et al., 2008;

Hadavi et al., 2020; E. I. Vlahogianni et al., 2014). These factors make ANPR a promising source of traffic data for bottleneck detection and impact assessment in arterial/collector urban roads.

To identify road segments affected by traffic bottlenecks, we develop a time-discrete mathematical model of bottleneck activation[1] that accepts measurements of traffic volume and speed from ANPR flow data. The activation model employs a system of inequalities based on speed differentials, similar to Chen et al. (2004), to determine changes in traffic conditions upstream and downstream of the bottleneck location. To discern between recurring and non-recurring bottlenecks, a recurrence factor is introduced representing daily frequency of activation. Bottleneck impact is then further characterised in terms of intensity, primarily measured as vehicle delay in excess vehicle-hours travelled, spatial extent and temporal metrics, such as expected daily onset and end activation times.

The novelty of the approach lies in its applicability to roads of any category (its use is not constrained to highways), using a fundamentally different source of traffic data (measurements are taken along the length of a road, rather than at a point) not employed before for this purpose. Furthermore, the process is generic and can be applied to any urban centres equipped with ANPR cameras, not exclusively Tyne and Wear.

The methodology is applied to the Tyne and Wear ANPR network, in the North East of England, and used to study congestion patterns in the region. In particular, it is of interest to: (a) analyse the distribution of bottleneck recurrence in monitored road segments; (b) to estimate the proportion of urban traffic congestion due to recurring traffic bottlenecks, against that which comes from other sources, across time of day and county (since no reference figures can be found for non-highway traffic); (c) rank recurring bottlenecks by intensity, highlighting congestion hotspots in the region. Calculated delay is evaluated by comparison to two measures of Travel Time Reliability (TTR), frequency of congestion (FOC) and planning time index (PTI) (Gong & Fan, 2017), which are commonly used to quantify the intensity of traffic congestion.

The current chapter is structured as follows. Section 5.2 briefly describes related work in bottleneck detection and assessment. Section 5.3 defines the proposed model of bottleneck activation and impact metrics. Section 5.4 presents the results of applying the identification and assessment methodology to two known bottleneck locations along corridors of varying length. In turn, section 5.5 presents the bottleneck analysis results for the entire county of Tyne and Wear. Lastly, Section 5.6 summarises the capabilities of ANPR for bottleneck detection and impact assessment, and discusses several methodological limitations and potential enhancements.

[1] A bottleneck is active if its effect is felt upstream of its location, i.e. traffic is congested, while downstream conditions are free flowing (Daganzo, 1997).

5.2 Related work

5.2.1 Bottleneck detection

Traffic flow theory has emerged to predict vehicle and stream behaviour at highway intersections – a common source point of recurring bottlenecks (Daganzo, 1995; Newell, 1993b). Traffic bottlenecks are thus relatively well understood traffic phenomena (Z.-C. Li et al., 2020; Vickrey, 1969). For operational purposes, bottlenecks are divided into two categories based on frequency of occurrence: *recurring* and *non-recurring*. Recurring bottlenecks are predictable sources of traffic congestion, generally occurring at the same location and periods of the day (Spiller et al., 2012). Non-recurring bottlenecks differ from recurring bottlenecks in that they are associated with sporadic events, such as traffic incidents, and begin to dissipate once the source of congestion is treated (Falcocchio & Levinson, 2015). In addition, bottlenecks are considered to be *static* if the underlying cause is stationary like a physical structure or characteristic of the road, or *dynamic* if a slow-moving vehicle or convoy obstructs free flowing traffic (Daganzo, 1997; Spiller et al., 2012).

While theory allows for simulation of realistic traffic behaviour (Long et al., 2008), bottleneck detection is specifically concerned with the identification of bottleneck-related congestion in traffic data. A notable method was established by Chen et al. (2004), who identify bottleneck locations and periods of bottleneck activity from a simple set of inequalities based on downstream-upstream speed differentials. A similar approach is later used by Wieczorek et al. (2010). The authors find the method to perform reasonably well in comparison to earlier methods built on cumulative vehicle arrival curves (Bertini & Myton, 2005; Cassidy & Windover, 1995). These studies are referenced because the authors make a clear effort to distinguish localised from systemic congestion and provide a method for automated detection. This is in contrast to general methods of traffic state prediction and congestion modelling, which are applicable insofar as bottleneck locations are known a priori (see, for example, works by Chow et al. (2014) and He et al. (2016), or A. M. Rao & Rao (2012) for a review on measuring urban traffic congestion).

Previous detection methods are based on the analysis of traffic volume and speed data from a series of loop detectors with known segment lengths, placed along a highway Bertini & Myton (2005). Loop detectors are fixed traffic sensors, installed below the pavement, capable of detecting passing vehicles by way of an electrical induction loop (Klein et al., 2006). Petty et al. (1998) have shown that single-loop detectors can accurately estimate vehicle speeds from observed counts. An array of loop detectors, as used for bottleneck detection, presupposes a linear travel path, i.e. vehicles detected by one sensor will also be detected by the next sensor. The same assumption does not apply to urban environments,

where signalised intersections are commonplace and routing choices are not fully known. Thus, given that loop detectors are incapable of tracking individual vehicles, they fail to provide reliable estimates of upstream-downstream vehicle speed, required for bottleneck detection, in urban environments where vehicle trajectories are unknown.

To circumvent the limitations of fixed traffic sensors in non-highway environments, W. Lee et al. (2011) used a fleet of probe-vehicles (taxis) to pinpoint possible bottleneck locations in a city – their method equally relying on downstream-upstream congestion differentials to reveal bottleneck activations across the network. Similarly, Gong & Fan (2017) demonstrates the potential of floating-car data (FCD) to identify recurring bottlenecks, albeit exclusively in a highway network and through the application of travel time reliability (TTR) measures – whose purpose is to quantify the amount of variability experienced by users in journey times (Carrion & Levinson, 2012). Despite FCD growing, in recent years, to become the standard data source for measuring urban congestion (e.g. Department for Transport (2021) measures average road delay across England using probe vehicle data), ANPR is, in some cases, a more accessible data source for local traffic management, particularly as camera coverage increases (Debnath et al., 2014).

5.2.2 Impact assessment

Delay is the primary metric used to measure the intensity of bottleneck-induced congestion. The DfT reports road delay as the excess travel time per vehicle per mile (Department for Transport, 2021), although it is more commonly defined as a volume-like quantity, expressed in excess vehicle-hours travelled below a reference speed (Chen et al., 2004; Skabardonis et al., 2003). As a source of traffic congestion, a bottleneck is further characterised by its duration – total time it causes vehicles to experience delay – and extent – how far reaching in space its effect is (Skabardonis et al., 2003). Additional metrics have been reported, for example, Wieczorek et al. (2010) calculate the velocity of the backward-propagating congestion shockwave.

Across time, assessment metrics are summarised via a total or daily expected value (oftentimes further discriminated by morning/afternoon period). Bottlenecks can then be ranked based on expected daily intensity and variability, recurrence factor or a combination of measures. For instance, the FHWA (Federal Highway Administration, U.S.) employs a spatio-temporal traffic state matrix (STM) together with measurements of travel time reliability (TTR), to generate an aggregate metric of bottleneck intensity that is then used to compare and rank highway bottlenecks Hale et al. (2021). Gong & Fan (2017) found that different TTR metrics can work equally well in measuring the intensity of bottleneck-induced congestion, but that an ensemble approach better encapsulates the benefits of each individual metric while covering for its shortcomings.

5.3 Methodology

5.3.1 Bottleneck activation

The goal of bottleneck detection is to distinguish periods of bottleneck activity from periods of inactivity. An active bottleneck has three observable features (Daganzo, 1997):

1. congestion present upstream of the bottleneck, characterised by speeds below a critical value,
2. improved traffic flow conditions downstream of the bottleneck, described by a gain in speed (specified in absolute or relative terms), and
3. considerable user demand, defined by traffic volumes within historical medium to high levels.

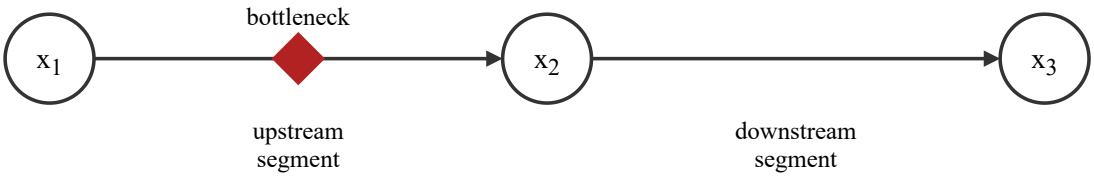


Figure 5.1: Simplified schematic of a road corridor monitored by three sensors.

For loop detectors placed in a highway, Chen et al. (2004) models bottleneck activation as a system of inequalities. A bottleneck located between two sensors, upstream and downstream of its location, is considered to be active if the following four conditions are simultaneously satisfied: (i) speeds slower than 40 mph in the upstream sensor, (ii) a minimum 20 mph speed increase in the downstream sensor; (iii) a continuous speed increase between the upstream and downstream sensors (for any intermediate sensors placed in between); (iv) a maximal sensor separation of 2 miles. If one or more conditions are not satisfied, the bottleneck is said to be inactive.

Due to differences in measurement type and corridor class, the activation model by Chen et al. (2004) is not readily applicable to ANPR data. Figure 5.1 can be used to illustrate differences in measurement type: three sensors divide the corridor into two road segments, and generate traffic measurements at sensor point (x_1, x_2 and x_3), in the case of loop detectors, or along the road segments (upstream and downstream) for ANPR. Thus, to identify an active traffic bottleneck in the upstream segment, loop detectors use traffic measurements in sensors x_1 and x_2, whereas ANPR uses measurements along the two road segments (which in turn requires three cameras).

Differences in corridor type arise because ANPR corridors are not restricted to highways, where driving conditions are uniform and constrained along a single route, but can cut

across different levels of the road hierarchy and intersect with other routes. Thus, unlike Chen et al. (2004), speeds can not be directly compared against a universal threshold provided by experts without first being normalised by a segment-specific reference speed. He et al. (2016) achieve normalisation via a speed performance index r_v, calculated by dividing each speed measurement by the segment's speed limit. Gong & Fan (2017) adopts a similar approach, wherein observed speeds are instead divided by the segment's free flow speed, which is estimated from night-time observations.

The key differences imposed by ANPR are addressed in the adapted bottleneck activation model, formally described as follows. Let C denote a corridor section, expressed as:

$$C = \{s_j\}_{j=1}^{n} \tag{5.1}$$

where s_j designates the j^{th} segment of C, an ANPR route specified by a pair of cameras. A bottleneck activation function $A(s_j, t)$ indicates the presence or absence of an active bottleneck on segment j during time interval t, given sensor readings of vehicle count $q(s_j, t)$ and mean speed $v(s_j, t)$. Based on the three distinctive features of an active bottleneck (described above), function A evaluates to 1 if the following set of inequalities are met and 0 otherwise:

$$r_v(s_j, t) < \theta_u \tag{5.2a}$$

$$r_v(s_{j+1}, t) - r_v(s_j, t) > \theta_d \tag{5.2b}$$

$$q(s_j, t) > q_m(s_j) \tag{5.2c}$$

where $r_v(s_j) = \frac{v(s_j,t)}{v_f(s_j)}$ is the speed performance index of segment j at time t, which normalises speed according to its segment's free flow speed $v_f(s_j)$; $q_m(s_j)$ is the typical-day traffic volume median value; θ_u is a chosen congestion threshold that classifies upstream traffic flow as free flowing or congested; and θ_d is a downstream speed gain factor that reflects substantial differences in upstream and downstream flow conditions.

The terms $r_v(s_j)$ and $r_v(s_{j+1})$ represent the speed performance index of the upstream and downstream segments, respectively. $v_f(s_j)$ and $q_m(s_j)$ are segment specific parameters, estimated from data, while θ_u and θ_d are user defined parameters. For simplicity, $v_f(s_j)$ and $q_m(s_j)$ are assumed to be constant across time and only need to be estimated once for each unique segment (as detailed in Section 5.4.1.1). Instead of hardcoding the value of the normalised thresholds θ_u and θ_d, as in Chen et al. (2004), these are parametrised to allow for sensitivity analysis and user calibration. Condition 5.2c was not proposed

by Chen et al. (2004), likely due to stable traffic in highways, and has been added as a filter against erroneous activations caused by small vehicle samples. It reflects the fact that bottlenecks only manifest themselves under considerable user demand (Spiller et al., 2012).

By definition, the proposed method has two fundamental limitations: it can only be applied to corridors composed of at least two segments (three cameras), and can only detect bottleneck activations in the first $n-1$ segments of the corridor. Additionally, an activation may pass undetected in the event that another bottleneck, located in a segment downstream of the affected segment, is simultaneously active. In that event, the upstream bottleneck is said to be hidden by the downstream bottleneck. If a bottleneck is hidden by another located in a segment that is not the most downstream of the corridor, then the algorithm will identify the downstream bottleneck. However, if the downstream bottleneck coincides with the most downstream segment of the corridor, then both bottlenecks may pass undetected. Refer to Chapter 4, for a complete methodology on ANPR corridors.

Note that an additional distinctive characteristic of a bottleneck is not observable: the specific point along the segment where a queue starts to form upstream of the bottleneck. Thus, for an active bottleneck, it is assumed that an onset location exists somewhere along the upstream segment but that its precise location or nature can not be determined by the algorithm (that task is best subsequently performed by the competent authorities).

5.3.1.1 Sustained activation

Similarly to Chen et al. (2004), the primary interest is in identifying sustained bottleneck activations. A sustained activation treats a sequence of active periods as a single activation as opposed to a series of temporary or spike activations. Sustained activations are easier to interpret and visualise, and simplify measurements such as the identification of the first/last period of day during which the bottleneck is active.

To find sustained activations, we look for a minimum of C active periods every N time periods. For instance, working with 5-minute data, Chen et al. (2004) set $N = 7$ and $C = 5$, so that a bottleneck is sustained if it has at least five active periods (25 min) within every seven consecutive time intervals (35 min).

Let $A_s(s_j, t_1, t_2)$ denote whether a bottleneck is actively sustained in segment j between time periods t_1 and $t_2 = t_1 + N$, and let:

$$A_s(s_j, t_1, t_2) = \mathbb{1}\Big[\sum_{\tau=t_1}^{t_2} A(s_j, \tau) \geq C\Big] \quad \text{subject to } t_2 - t_1 = N \tag{5.3}$$

where C is the minimum number of active periods out of N. Equivalently, the state of segment j during period t, denoted $A_s(s_j, t)$, is equal to $A_s(s_j, t_1, t_2)$, for all $t_1 \leq t \leq t_2$.

Thus, an inactive state $A(s_j, t) = 0$ can be "promoted" to $A_s(s_j, t) = 1$ if there is a reference interval $[t_1, t_2]$ such that $A_s(s_j, t_1, t_2) = 1$. Conversely, $A(s_j, t) = 1$ can be "demoted" to $A_s(s_j, t) = 0$ if it occurs in isolation, i.e. if there is no reference interval $[t_1, t_2]$ such that $A_s(s_j, t_1, t_2) = 1$.

To illustrate the concept of reference interval, consider the simple activation sequence `0001101100`, indexed by $t = 1..10$. Setting $N = 5$, we can form two non-overlapping intervals $I_1 = [1, 5]$ with subsequence `00011` and $I_2 = [6, 10]$ with subsequence `01100`. If we set $C = 3$, then the activation is not considered sustained in any of the two intervals (and therefore in none of the individual t-indexed periods). This happens despite the middle subsequence `11011` clearly matching the condition $C \geq 3$.

To address cases where interval alignment is unfavourable, we introduce a third parameter $1 \leq S \leq N$ that controls the shift of the reference interval across time (in number of periods). In the previous example, S was equal to N, hence there is no overlap between reference intervals. If instead we assign $S = 3$, we obtain the following reference intervals: $I_1 = [1, 5]$ `00011`, $I_2 = [4, 8]$ `11011` and $I_3 = [6, 10]$ `00011`. Because the second interval meets condition 5.3, then activation is sustained for all periods that fall within that interval. For example, $t = 4$ and $t = 5$ both fall within intervals T_1 and T_2 – they are considered non-sustained in relation to T_1 but sustained in relation to T_2. Since a time period only needs to be sustained in relation to one interval, we have $A_s(t) = 1$ for t=4..8, and $A_s(t) = 0$ the remaining values of t.

5.3.2 Impact assessment

5.3.2.1 Induced delay

One of the primary metrics of bottleneck impact is the delay induced to an otherwise free flowing traffic stream. The amount of delay caused by an active bottleneck depends on three aspects: the segment's length, the volume of traffic and the severity of congestion (Skabardonis et al., 2003). Additionally, a bottleneck may cause queue spillback and a congestion shockwave that propagates backwards from its location (Newell, 1993b). Therefore, the delay caused by a bottleneck in segment j should include any delays also caused to its upstream segments.

Formally, the total delay $D(s_j, t)$ caused by a bottleneck in segment j during time interval t is given by:

$$D(s_j, t) = A_s(s_j, t) \sum_{i=1}^{j} d(s_i, t) \cdot A_b(s_i, s_j, t) \tag{5.4}$$

where $A_s(s_j, t)$ specifies whether the bottleneck is actively sustained during t; $d(s_i, t)$ is

a function that calculates the delay associated with segment $i \leq j$; and $A_b(s_i, s_j, t)$ is an indicator function that evaluates to 1 if segment i has been affected by an active bottleneck in j or 0 otherwise. Let

$$A_b(s_i, s_j, t) = \begin{cases} 1 & \text{if } i = j, \\ 1 & \text{if } r_v(s_i, t) < \theta_u \text{ and } r_v(s_i, t) > r_v(s_j, t), \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

where conditions $r_v(s_i, t) < \theta_u$ and $r_v(s_i, t) > r_v(s_j)$ state that a preceding segment i is affected by a bottleneck in j if i is operating under congestion and the congestion of less severity than the congestion in j. We follow the rationale of Chen et al. (2004) that speed performance should deteriorate continuously as a segment approaches the location of the bottleneck, so to eliminate cases where consecutive active bottlenecks create more complicated speed profiles. Lastly, the delay $d(s_i, t)$ occurring in segment i, of length l_i, is the difference between the expected number of vehicle hours traveled in current traffic conditions $\frac{l_i \cdot q(s_i,t)}{v(s_i,t)}$ and the expected vehicle hours traveled in otherwise free flowing conditions $\frac{l_i \cdot q(s_i,t)}{v_f(s_i)}$; which can be written as:

$$d(s_i, t) = l(s_i) \cdot q(s_i, t) \cdot \left(\frac{1}{v(s_i, t)} - \frac{1}{v_f(s_i)} \right) \tag{5.6}$$

where the unit of measurement is vehicle hours (veh-h = km $\times$ veh $\times$ (km/h)$^{-1}$).

5.3.2.2 Daily delay and bottleneck duration

It follows that the total delay induced by a bottleneck across an entire day is simply the sum of the delays occurring in each individual time interval, and its total duration is given by the sum of the periods for which the bottleneck was actively sustained. Assuming one day is broken into $T = \frac{1440}{W}$ consecutive intervals of length W minutes, the total delay by a bottleneck on segment j in a given day is calculated as $\sum_{t=1}^{T} D(s_j, t)$ and its total duration is $W \sum_{t=1}^{T} A_s(s_j, t)$. Observed across multiple days, a bottleneck's impact is characterised by its daily distribution of induced delay and duration, via measures of the central tendency and variability of the distribution, particularly average daily bottleneck delay and duration, plus standard deviation.

5.3.2.3 Recurrence

To complete the impact assessment of a bottleneck, we are interested in measuring its degree of recurrence, as recurring bottlenecks are predictable not only in where and when

they occur, but also how often. Recurrence, denoted by $R(s_j)$ is simply the proportion of days a bottleneck is actively sustained at least once per day:

$$R(s_j) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\Big[\sum_{t=1}^{T} A_s(s_j, t) > 0\Big] \tag{5.7}$$

where n is the number of days in the data sample; and, as per Eq. 5.3, a bottleneck is actively sustained at least once a day if it observes C active periods within any window of N consecutive time periods. Note that n should exclude days that show a considerable amount of missing data affecting one or more corridor segments.

5.4 Results of study-case bottleneck analysis

Bottleneck analysis is performed for two study-cases: a known bottleneck location in the A189/A1056 roundabout along a small corridor section, and along the A189 Southbound corridor, a longer corridor of regional importance.

5.4.1 Bottleneck in the A189/A1056 roundabout

The A189 meets the A1056 at a roundabout in Cramlington, in the North Tyneside district[2]. The roundabout is a well known recurring bottleneck, with queues forming Northbound on the A189 when overdemand is present. Figure 5.2 depicts a schematic of the ANPR corridor and road network around the bottleneck. Three ANPR cameras are located on the A189-A1056 Northbound-Westbound (NW) corridor, forming two road segments with lengths 1.3 km and 1.5 km respectively, roughly located before and after the bottleneck.

The A189/1056 roundabout is a simple example of bottleneck behaviour observed by an ANPR system. It is used to demonstrate the practical details of the bottleneck activation model and inform the choice of parameter values. Unless hidden by another bottleneck located further downstream, the the A189/A1056 bottleneck manifests as congested flow in the upstream segment and free flow conditions in the downstream segment.

5.4.1.1 Parameter choice and conditions for bottleneck activation

To elucidate the mechanism of bottleneck identification and choice of parameters, consider the variables that compose the activation conditions defined in Equation 5.2. The following signals are relevant: the speed performance index of the upstream ($j = 1$) and

[2]The approximate coordinates are 55°02'21.1"N 1°35'33.5"W

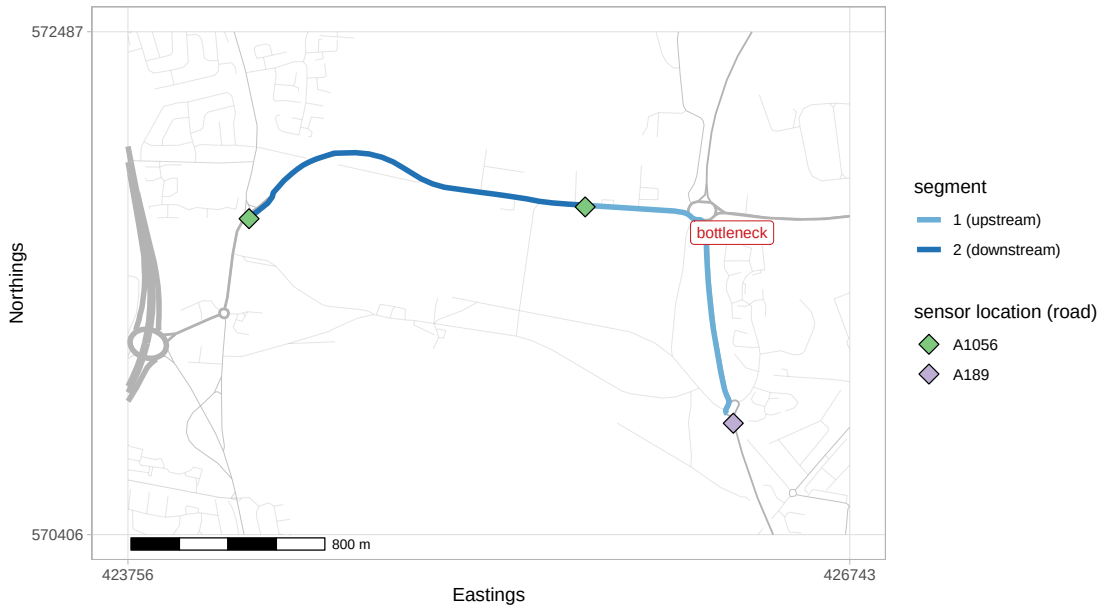


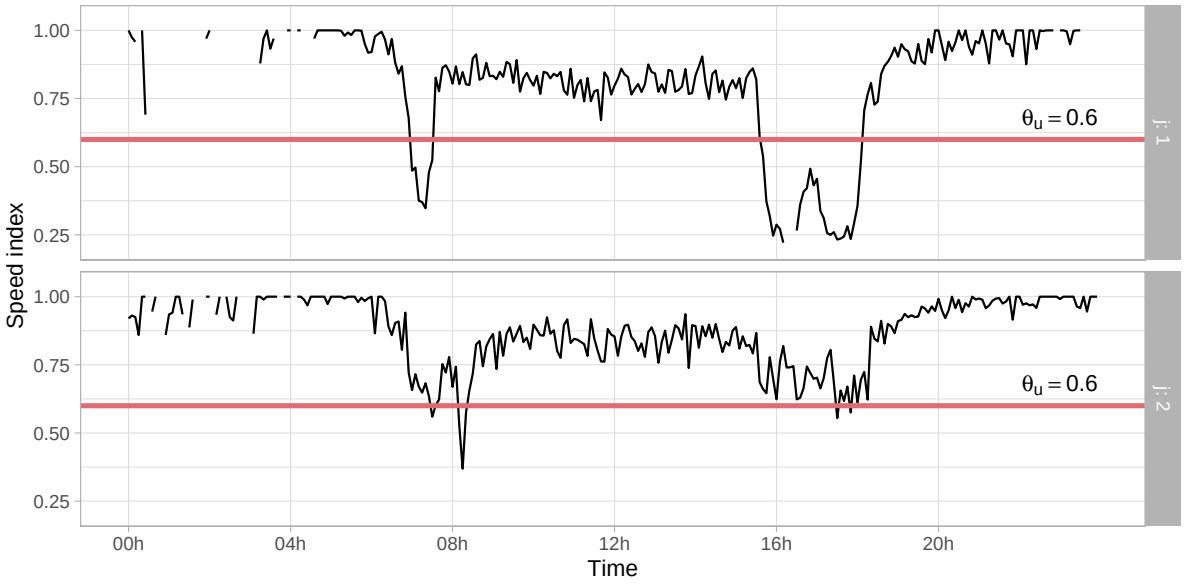
Figure 5.2: A169-A1056 north-westbound ANPR corridor passing through a known bottleneck location.

downstream ($j = 2$) corridor segments, the speed performance differential between the two, and the traffic volume at each of the segments. Figure 5.3 shows the evolution of these five variables, across time of day, recorded for on a random weekday on the A189-A1056 NW corridor.

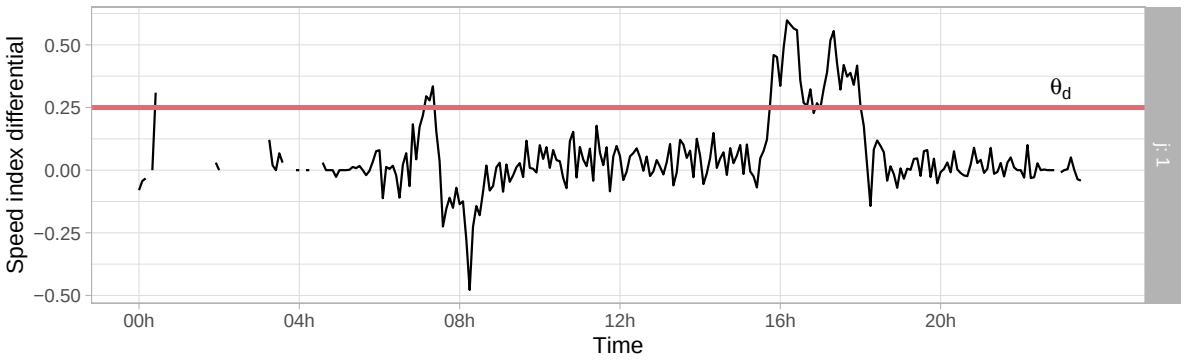
The free flow speed of each segment, necessary to calculate r_v, is determined using night-time speed observations, under the assumption that traffic conditions are free flowing during this period, similarly to Gong & Fan (2017). More specifically, we combine all speed samples taken during the night period (10pm to 6am, inclusive) measuring a minimum of 30 km/h and maximum of 120 km/h. We then estimate free flow speed as the median of the remaining samples, for the entire year of 2018.

A range of values is empirically suitable for θ_u. Reference studies specify this range approximately between 0.50 and 0.75. Gong & Fan (2017) evaluates two reference threshold values, 0.60 and 0.75, but finds no significant difference in the assessment of congestion performed with either parameter value. He et al. (2016) defines the congestion for parameter values below 0.50, however, the authors normalise speeds by the maximum permissible road speed, as opposed to the estimated free flow speed (the first being an upper bound of free flow speed and the second a central value). This range of values is supported by our exploratory analysis of speed performance. Figure 5.3a shows the speed performance index $r_v(s_j, t)$ at each of the upstream and downstream segments. In the example, congestion occurs when the index drops below $\theta_u = 0.60$, chosen in accordance to Gong & Fan (2017) (although other values within the theoretical range $[0.50, 0.75]$ could be reasonably employed as well).

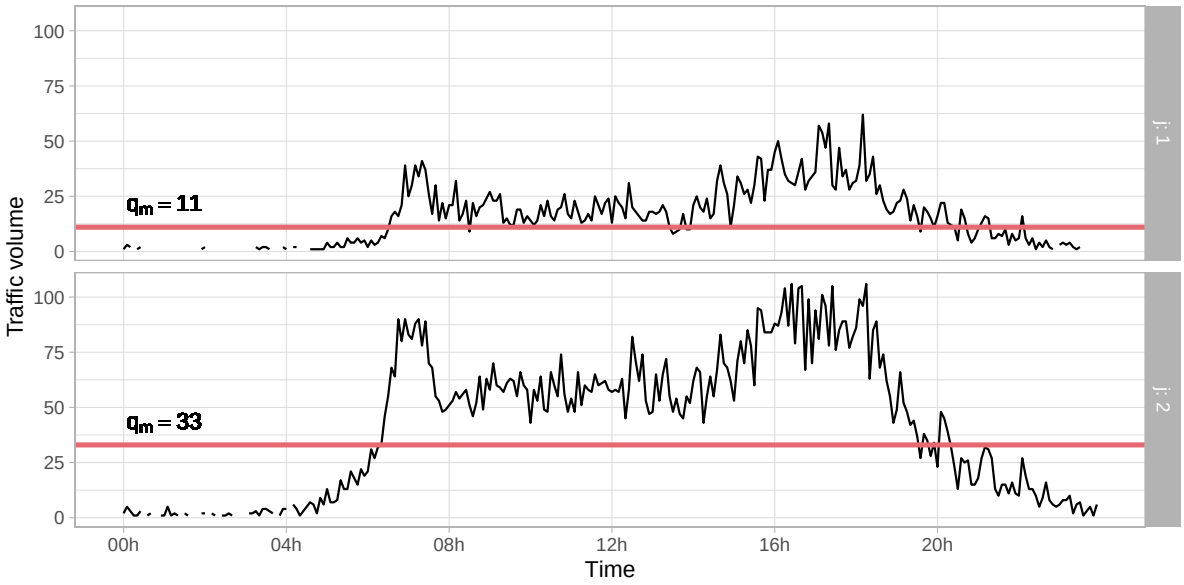
Figure 5.3b displays the difference in speed performance between the downstream and



(a) Speed performance index $r_v(s_j, t)$.



(b) Downstream-upstream speed performance difference $r_v(s_2, t) - r_v(s_1, t)$.



(c) Traffic volume $q(s_j, t)$, in number of vehicles.

Figure 5.3: Time series of the variables used to compute bottleneck activation, at each of the upstream $j = 1$ and downstream $j = 2$ segments of the A169-A1506 ANPR corridor, on 05 March 2018.

upstream segments $r_v(s_2) - r_v(s_1)$. The differential allows for the distinction between localised congestion, characteristic of traffic bottlenecks, and systemic congestion. Since the difference is calculated using speed indices, the resulting quantity reflects a relative change in speed performance instead of an absolute one. The choice of $\theta_d = 0.25$ is motivated by the fact that, at $r_v(s_j) = \theta_u = 0.50$, an increase of 0.25 in performance is roughly necessary to improve traffic conditions to nearly daytime free flowing levels, highlighted by the mid day inter-peak region (9am-3pm).

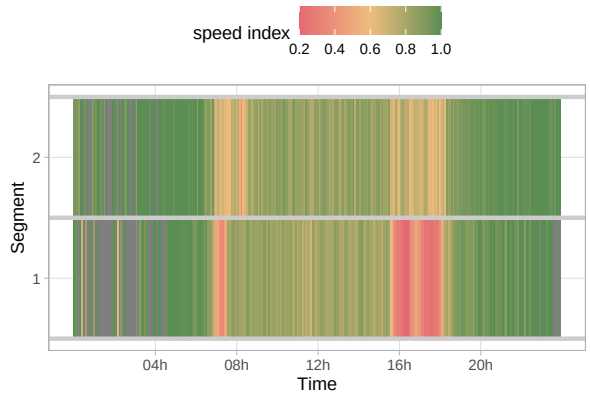
Lastly, Figure 5.3c shows the observed traffic volume across each link. The morning and evening demand peaks are visible, and night-time traffic is a small fraction of day-time traffic. Condition Equation 5.2c is thus employed to prevent spikes that trigger conditions 5.2a and 5.2b from generating dubious bottleneck activations (e.g. the midnight spike in Figure 5.3b). Similar "spikes" may arise from small sample sizes and outlier speed observations passed through undetected from the earlier data pre-processing phases (trip identification).

The segment-specific quantity $q_m(s_j)$ is estimated as the median of traffic volume observations combined across many different days. It acts essentially as a night-day discriminator, while also discriminating against periods of sensor failure. To improve the estimate, care is taken to include only those days that are fully or nearly fully observed, e.g. days that contain at least 75% of time periods with vehicle detections (216 out of the 288 daily 5-minute intervals).

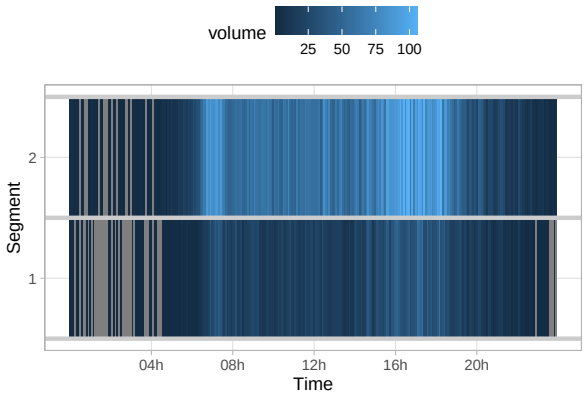
5.4.1.2 Example activation

Figure 5.4 shows an instance of sustained bottleneck activation (Figure 5.4d), during the evening period, made clear by a differential in speed performance between the upstream and downstream segments (Figure 5.4a) while under maximum user demand (Figure 5.4b). Parameter values were set to $\theta_u = 0.5, \theta_d = 0.25$ for bottleneck activation, and $C = 5, W = 7$ for sustained activation, after testing different parameter combinations.

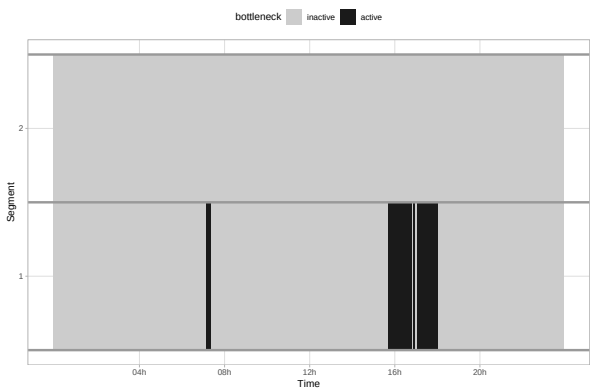
The distinction between simple activation and sustained activation is seen in Figures 5.4c and 5.4d, wherein sporadic activations and gaps between consecutive activations are eliminated. Bottleneck duration is calculated by counting the number of time intervals the bottleneck is sustained, 26×5 minutes = 130 minutes in Figure 5.4d, and the total delay caused during this period is the sum of the delays calculated at each of the affected time intervals: a total of 69.5 veh-hours.



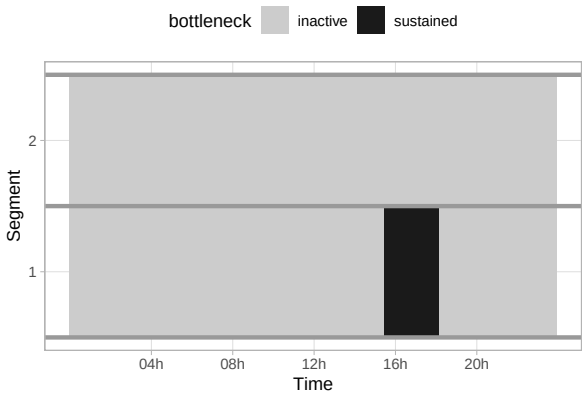
(a) Speed performance (nil volume in gray)



(b) Traffic volume (nil volume in gray)



(c) Bottleneck activations ($\theta_u = 0.60$, $\theta_d = 0.25$)



(d) Sustained activations ($C = 5, N = 7, S = 3$)

Figure 5.4: Space-time raster plots of the A169-A1506 corridor, obtained from flow data at 5 minute resolution, on 05 March 2018.

Table 5.1: Impact assessment of traffic bottleneck on segment 1 of the A169-A1056 ANPR corridor over the 6-week period shown in Figure 5.5. Daytime is split into two periods: *am* (7am-1pm) and *pm* (2pm-7pm). Daily statistics (mean ± standard deviation) are calculated solely based on the days in which the bottleneck recurs.

time	recurrence	duration (min)	delay (v-h)	extent (km)	onset time (min)	cutoff time (min)
am	0	-	-	-	-	-
pm	0.733 ($^{22}/_{30}$)	108.2 (35.8)	60.2 (23.7)	1.3	15:48 (22.8)	17:33 (35.9)

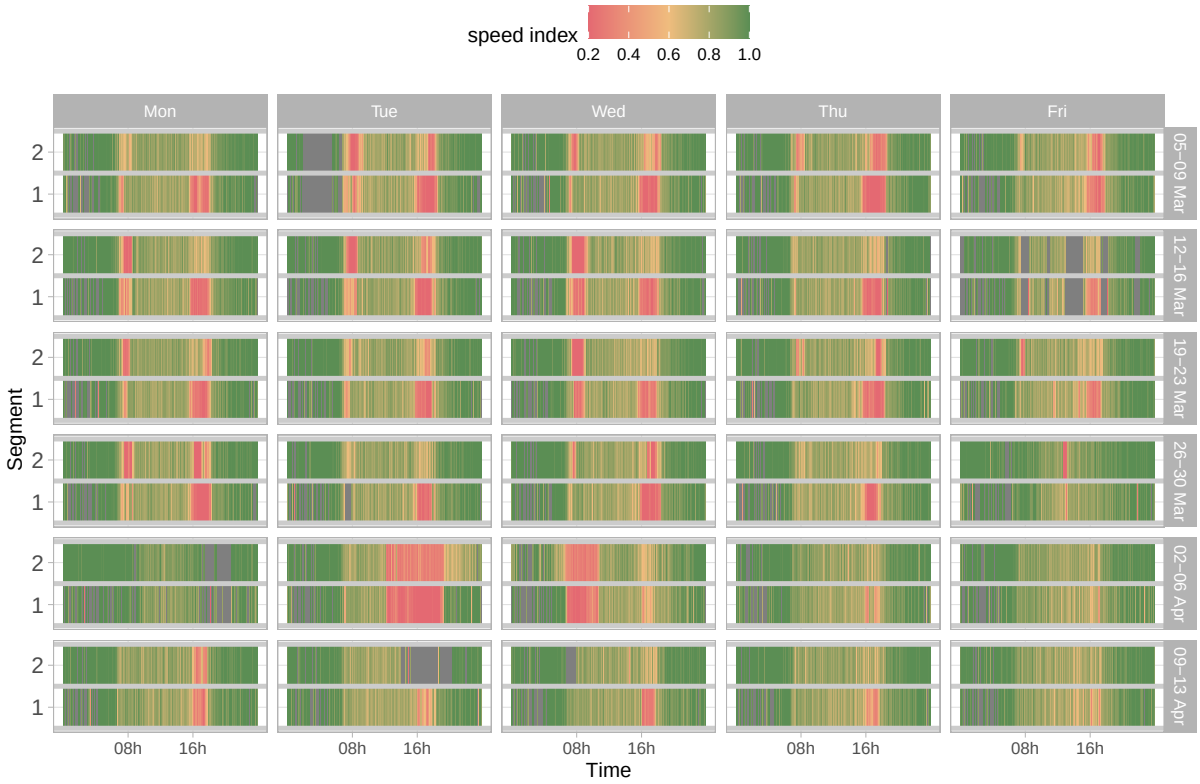
5.4.1.3 Daily impact analysis

The recurring nature of the bottleneck is revealed by predictable onset times and sustained activation periods, observed across its daily speed and activation profiles in Figure 5.5, over a period of six weeks. Only weekdays are shown since no sustained activations were registered on weekends during this period – a trend explained by the significantly lower road usage patterns during weekends.

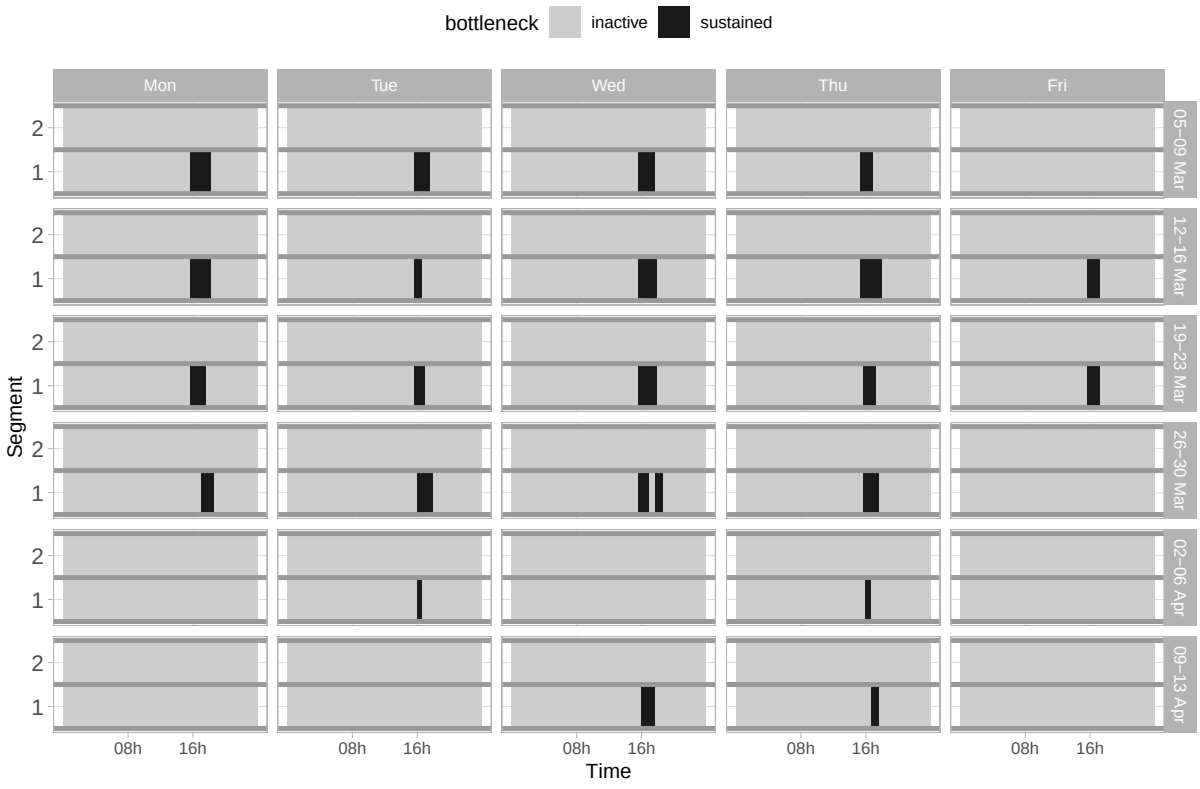
The bottleneck's daily impact on traffic flow is measured via six metrics, calculated separately for the morning and afternoon periods: recurrence (scalar), delay, duration and extent (mean and standard deviation) and first and last sustained activation time periods (mean and standard deviation). The results, displayed in Table 5.1, indicate that the bottleneck recurs mainly during the afternoon period, approximately 75% of week days, often around 4pm, and lasts an average of 1 hour, while causing an average total delay of 49 vehicle-hours.

Note how some days, namely Tue 03 April and Wed 04 April, register no sustained activations despite prevalent and severe traffic congestion. In these instances, the bottleneck is not considered to be active because congestion is systemic rather than localised. This can happen if the bottleneck is hidden by another bottleneck further downstream that is not observed.

Additionally, some days exhibit missing data during time periods which are expected to display considerable user demand. For example, on Fri 16 Mar segments 1 and 2 display multiple afternoon periods with zero vehicle detections when other Fridays show no missing data. Similarly, on Tue April 10, segment 2 shows a long period of missing data when it consistently observes traffic congestion on previous weeks. As described in Section 3.5.4, missing data can be explained by faulty sensors (a broken intermediary ANPR camera, i.e. destination camera in segment 1 and origin camera in segment 2, would explain synchronised missing data in the two segments), or minimal vehicle headway during heavy traffic congestion.



(a) Speed performance index (nil traffic volume shown in gray)



(b) Sustained bottleneck activations ($\theta_u = 0.60, \theta_d = 0.25, C = 5, W = 7, S = 3$)

Figure 5.5: Space-time raster plots of A169-A1506 corridor for a six week period (ranging from 05 Mar to 13 April 2018), obtained from flow data at five minute resolution.

5.4.2 The A189 Southbound corridor

The A189 is an important local A-road connecting Newcastle upon Tyne and Ashington, in Southeast Northumberland. The road acts at times as urban arterial and distributor (e.g. New Redheugh Bridge, St James Boulevard, Grandstand Road), and as a highway in others (major dual carriageway past Cramlington and Blyth). Local traffic authorities have installed a series of cameras along the A189 that monitor most of its urban arterial corridor, beginning at Haddricks Mill junction in South Gosforth (where the A189 meets the A191) and ending at the Redheugh Bridge. This section of the A189, depicted in Figure 5.6, is a key corridor route for commuters, responsible for feeding traffic between urban and suburban areas to the north of the city.

The identification of bottlenecks along the A189 is of interest not only because of its regional importance and multi-segment nature, but also because the road meets other important A roads along the way, the intersection of which constitutes in theory possible locations for the formation of bottlenecks. For example, the A189 meets the Great North Road (B1318) at the Blue House roundabout – a junction used by tens of thousands of commuters daily and currently subject to improvements by the city council (seen as the intersection point at the middle of segment 2 of Figure 5.6) (Newcastle City Council, 2020; Wylie, 2015). Another example is seen on segment 5 along St James Boulevard, where the A189 meets the A186 – a corridor providing access to the western suburbs of the city – at a traffic-light signalised intersection.

Figures 5.7a and 5.7b show space-time raster plots of vehicle speed (input) and estimated sustained activation (output) during a 6-week period for the A189 Southbound corridor. Recurring sustained activations are observed in segments 2 and 4 during the morning period (defined between 7am and 1pm) and in segments 2 and 5 during the afternoon period (specified between 2pm and 7pm). In addition, segments 1 and 4 upstream of the bottleneck location are repeatedly affected by queue spillback during the afternoon period. That is, they operate under congestion but one that is of less severity than that of the downstream bottleneck-affected segments (speed performance decreases continuously as one approaches the bottleneck location, as postulated in Section 5.3.2.1). In these instances, the bottleneck's extent is greater as congestion propagates to the upstream segments. The total delay caused by the bottleneck then also includes the delay caused to the upstream segments, although whether the total delay is more severe in these cases depends also on observed traffic volume and the degree of congestion.

A summary assessment of bottleneck impact across the corridor is given on Table 5.2. The same impact metrics used as in previous (Section 5.4.1), but now on a segment by segment basis. As identified visually, segments 2 and 5 present bottlenecks with more than 50% week-day daily recurrence. Segment 2 sees recurring bottleneck activations

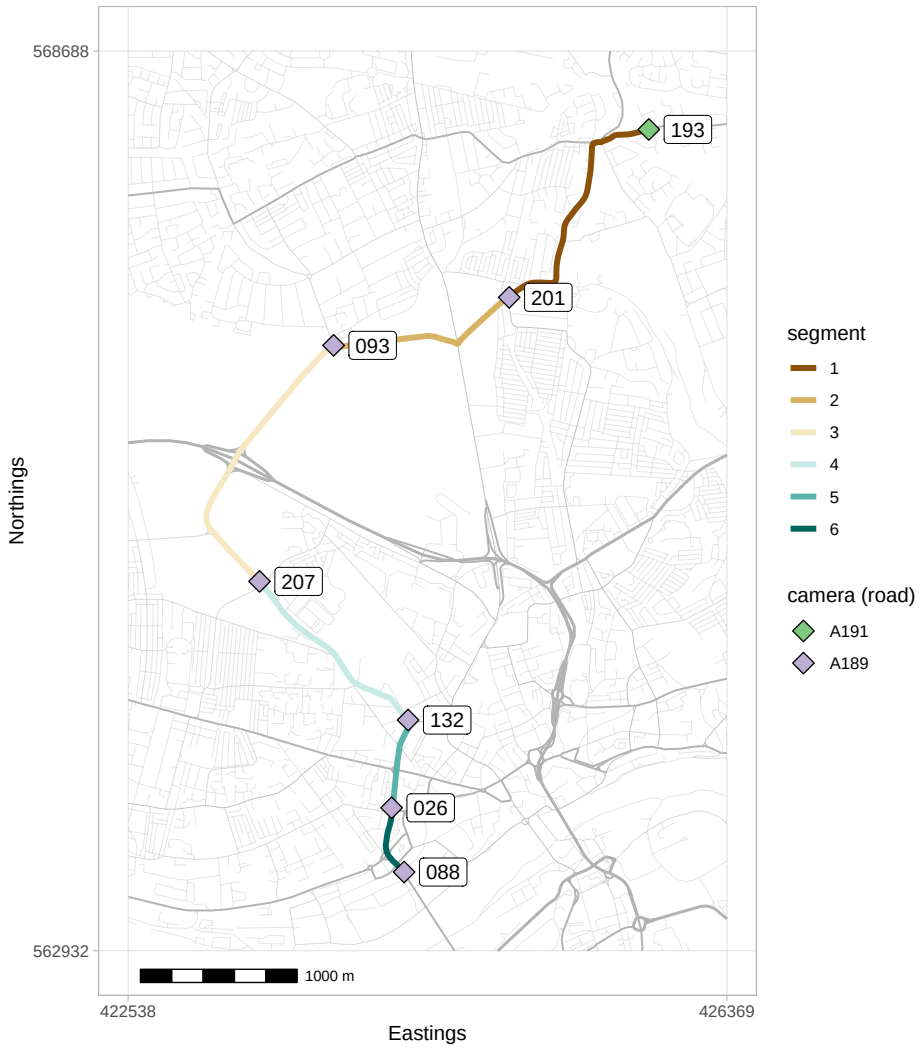
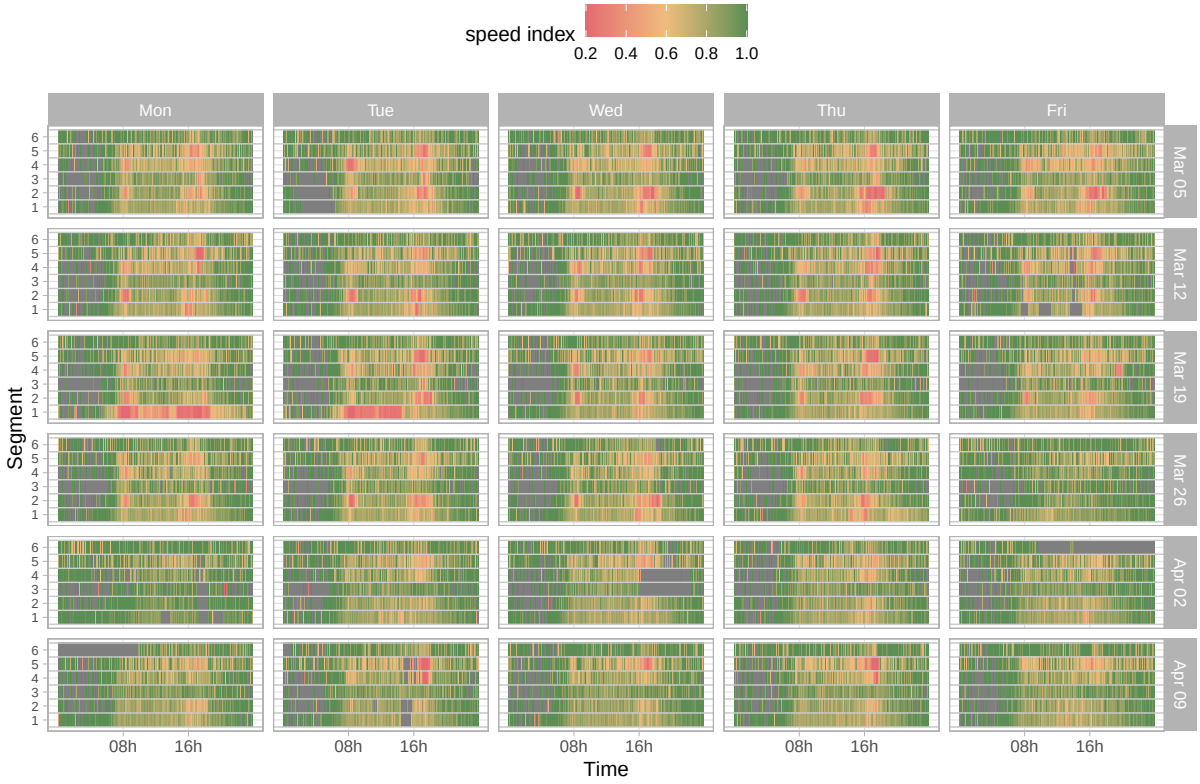
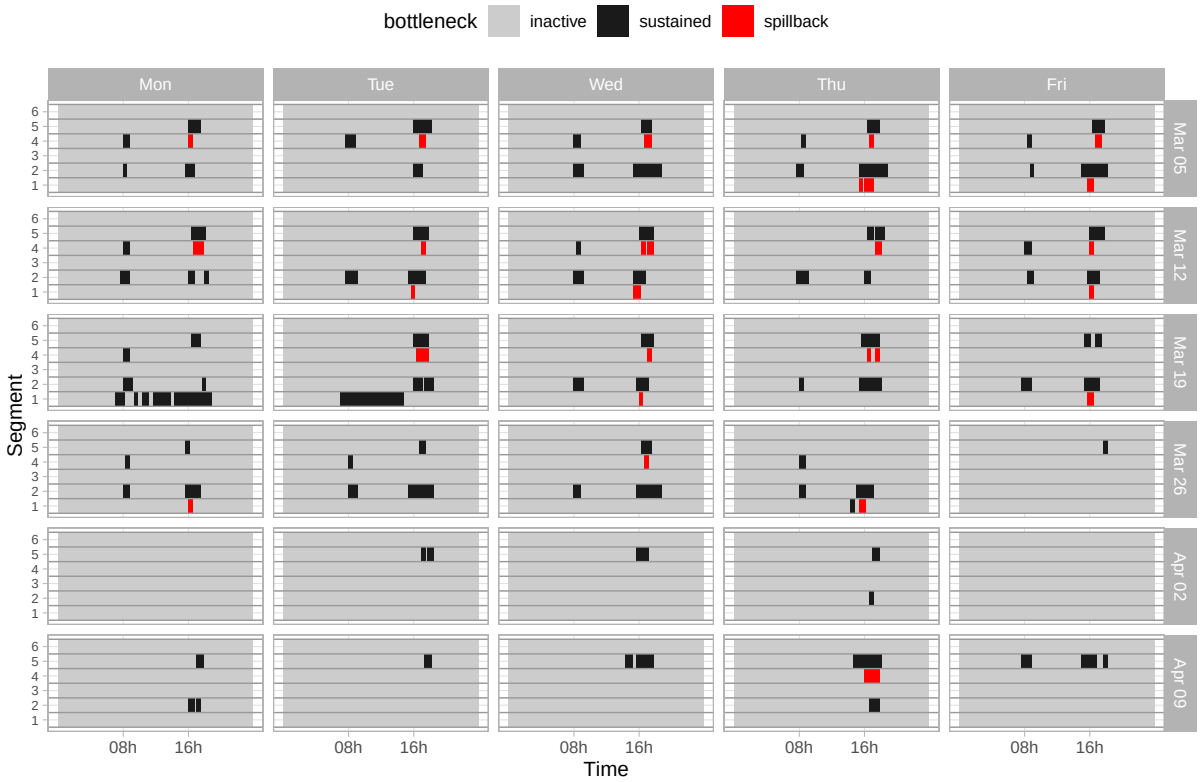


Figure 5.6: A189 Southbound ANPR corridor originating at A191.



(a) Speed performance index



(b) Sustained bottleneck activation and spillback ($\theta_u = 0.60, \theta_d = 0.25, C = 5, N = 7, S = 4$)

Figure 5.7: Space-time raster plots of A189 Southbound corridor for a six week period (ranging from 05 Mar to 13 April 2018), obtained from flow data at five minute resolution (missing data shown in gray).

Table 5.2: Impact assessment of A189 Southbound corridor over the 6-week period shown in Figure 5.7. Daytime is split into two periods: *am* (7am-1pm) and *pm* (2pm-7pm). Daily statistics (mean ± standard deviation) are calculated solely based on the days in which the bottleneck recurs.

j	time	recurrence	duration (min)	delay (v-h)	extent (km)	onset time (min)	cutoff time (min)
6	am	0	-	-	-	-	-
	pm	0	-	-	-	-	-
5	am	0.033 ($^{1}/_{30}$)	75.0	5.4	0.6	07:40	08:50
	pm	0.900 ($^{27}/_{30}$)	133.1 (67.9)	33.0 (22.6)	0.9 (0.4)	16:08 (45.2)	17:48 (27.4)
4	am	0.400 ($^{12}/_{30}$)	48.3 (13.0)	9.2 (5.4)	1.3	08:05 (12.4)	08:48 (5.8)
	pm	0	-	-	-	-	-
3	am	0	-	-	-	-	-
	pm	0	-	-	-	-	-
2	am	0.567 ($^{17}/_{30}$)	64.4 (18.9)	12.4 (5.7)	1.2	07:58 (15.6)	08:57 (15.7)
	pm	0.733 ($^{22}/_{30}$)	142.0 (72.7)	36.4 (26.8)	1.8 (0.7)	15:45 (37.1)	17:45 (42.4)
1	am	0.067 ($^{2}/_{30}$)	360.0 (84.9)	116.5 (40.9)	1.6	07:00	13:52 (3.5)
	pm	0.100 ($^{3}/_{30}$)	120.0 (130.3)	51.0 (68.5)	1.6	14:13 (11.5)	16:10 (138.6)

both in the morning and afternoon periods, the latter causing longer-lasting delays that sometimes extend backwards in space, affecting vehicles in segment 1 (single carriageway road, of approximately 1 mile in length)[3]. Similarly, segment 5 experiences heavy delays, primarily during the afternoon period, that extend backwards to segment 4 on A189 (Barrack and Ponteland Road). Segment 4 also observes bottleneck-induced congestion, but that is less impactful and less recurrent than the afternoon bottlenecks in segments 2 and 5. Segment 1 is an example of an impactful non-recurring bottleneck, since the observed bottleneck-like congestion occurs only sporadically.

The analysis of the A189 Southbound corridor illustrates the mechanics of bottleneck detection on a ANPR corridor composed of more than two segments. In this scenario, the impact assessment of a bottleneck is complicated by the effect of congestion shockwaves that propagate backwards (upstream) from the location of the bottleneck. The results show that the multi-segment nature of the detection model can capture the effects of backwards propagating congestion in non-highway settings, characterised by a continuous decline in speed performance.

[3]Although we can not determine the exact location of the bottleneck, user experience points to the Blue House Roundabout as the most likely source of bottleneck-induced congestion (Newcastle City Council, 2020; Wylie, 2015).

5.5 Results of region-wide bottleneck analysis

To examine the potential of ANPR to identify and assess recurring traffic bottlenecks on a large scale, we apply our methodology to the full network of cameras within the geographic county of Tyne and Wear (comprised of five borough counties: North Tyneside, Newcastle upon Tyne, South Tyneside, Gateshead and Sunderland). Our main goal is to judge the validity of identified segments and estimated impact metrics, while illustrating the usefulness of the approach to highlight urban traffic patterns and network performance issues. To that end, we:

1. compare identified bottlenecks to known congestion and travel patterns in the region;
2. examine features of global vs local congestion;
3. evaluate whether there is agreement between the ranking of bottlenecks according to calculated delay versus proven metrics of congestion.

Finally, we comment on some of the limitations, and extra steps necessary to conduct a complete bottleneck analysis.

5.5.1 Input data

The procedure takes as input the set of ANPR corridors previously determined in the region (as seen in Section 4.4). As developed in Section 4.2, corridors represent frequently traversed route sequences. Any combination of three or more cameras that is not a corridor, or part of one, represents a route sequence that is not experienced by vehicles and is therefore irrelevant for bottleneck detection. For each corridor, we find all paths from source to sink composed of two or more segments (three or more ANPR cameras in succession). The final set of corridor sections is equivalent to the set of ordinary super-sequences, i.e. the set of ordinary trip sequences in which no sequence is a subsequence of another sequence (see Section 4.2.2).

Where corridors overlap, segments in common are evaluated once per corridor. For example, corridor sections `100,11,12,228` and `100,11,265` share segment `(100,11)`, which is thus examined for bottlenecks separately in each section. As a result, a segment may be associated with a bottleneck in one corridor, if congestion is experienced locally, but not in the other, if congestion is instead experienced globally. This reflects the complex nature of urban networks, wherein how one experiences traffic through a junction/segment depends on one's past/future trajectory. A segment's preferred assessment is thus along the corridor for which it observes maximal delay.

As mentioned in Section 5.3.1, Chen et al. (2004) includes only the road segments whose distance is shorter than two miles. In the case of ANPR, the rationale for excluding lengthy

road segments is related to potential measurement limitations: if only the starting section of the segment is affected by congestion (as conceptualised in Figure 5.1), vehicles have an opportunity to compensate for lost time by driving faster in the free flowing portion of the segment. Thus, observed travel time might be somewhat diluted by long distance travel and fail to reflect the congested component of the movement. A distance threshold of 6 km $\approx$ 3.7 miles, was chosen empirically, as it represents roughly 90% of the input road segments.

For an initial input of 369 routes, 45 routes are discarded because they are too long (42 routes) or their spatial route/orientation can not be determined precisely (3 routes), resulting in an input network composed of a total of 324 routes. For the 324 route input network, corridor detection yielded 376 corridors and 413 unique input corridor sections. As a result of corridor identification, 49 routes can not be evaluated for bottleneck-induced congestion as they are not a part of any corridor of length 3 or greater (4 routes), or always show up as the most downstream corridor segment (45 routes). Consequently, out of 324 input routes, 275 cases are bottleneck-identifiable (a proportion of $^{275}/_{324} \approx 0.85$).

To develop a more complete assessment of bottleneck impact than in previous Sections, the input ANPR dataset is extended to the whole year of 2018. Ten minute flow data is used instead of five minute data for ease of computation. As before, only weekdays are considered due to increased user demand relative to weekend days. To address cases where data is missing for long periods of time (as seen on panel "Wed April 04" of Figure 5.7), each route selects days where the total vehicle count is not classified as an outlier according to the Box plot method (Tukey, 1977). The assumption is that long intervals of missing data, particularly during daytime, result in a total daily flow lower than is otherwise expected (predicted by a statistic like AADT). At the corridor level, we select the subset of days wherein none of its segments' total daily count is an outlier.

Lastly, the following parameter values are applied consistently to all corridors: $\theta_u = 0.60$, $\theta_d = 0.25$, $N = 4$, $C = 3$ and $S = 2$. The parameters for sustained activation, N, C and S, are different from those used previously, as aggregation is done over 10 minute intervals instead of 5 minute intervals. In this case, a bottleneck is sustained if it has at least three active periods (30 min) within every four consecutive time intervals (40 min).

5.5.2 Recurring vs non-recurring congestion

First is an examination of the recurring nature of bottleneck-induced congestion, performed across the 275 identifiable road segments. Table 5.3 summarises the distribution of bottleneck recurrence rates by placing segments into one of five bins: no bottleneck ($R = 0$), non-recurring bottleneck ($0 < R \leq 0.10$), sporadic bottleneck ($0.10 < R \leq 0.33$), low recurring bottleneck ($0.33 < R \leq 0.66$) and high recurring bottleneck ($0.66 < R \leq$

Table 5.3: Segment count by bottleneck recurrence rate for different periods of the day.

time	no bottleneck ($R = 0$)	non-recurring ($0 < R \leq 0.10$)	sporadic ($0.10 < R \leq 0.33$)	recurring low ($0.33 < R \leq 0.66$)	recurring high ($0.66 < R \leq 1.0$)	total
am	64	108	45	22	36	275
ip	56	139	40	21	19	275
pm	42	119	38	34	42	275
both	125	89	31	15	15	275
any	14	109	43	40	69	275

1.0). Unequal sized bins were chosen to highlight the lower end of the distribution, specifically the proportion of segments that never display bottleneck-induced congestion ($R = 0$) compared to those that have only displayed it on occasion ($0 < R < 0.33$). The binning process is repeated for five time periods for contrast due to daytime effects: morning peak (am) period (6-9am), inter-peak (ip) period (10am-2pm), afternoon peak (pm) period (3-7pm), both am and pm (how often a bottleneck is sustained during both periods), and any (how often a bottleneck is sustained during any of the am, ip or pm periods).

An analysis of Table 5.3 shows that most segments experience some form of localised congestion at some point throughout the year, albeit only a fraction of these do so a regular basis. During the morning and afternoon periods, about 62% of segments register either no sustained bottleneck activations or only non-recurring activations. On the other hand, between 13 to 15% of segments register high-recurring bottleneck activations ($R \geq 0.66$) in the morning or afternoon peak periods, and approximately 5% during both. Furthermore, several bottlenecks are active nearly every weekday ($R \geq 0.95$): 4 segments during the morning period, 3 segments during the inter-peak period, 16 segments during the evening period, and 3 during both.

To examine the validity of segments classified as high recurring bottlenecks, their traffic performance is compared to that of segments classified otherwise. The comparison is motivated by earlier reports of recurring sources of congestion accounting for a significant piece of the overall network congestion. Chow et al. (2014) reports 85% of urban network congestion due to recurring factors, namely baseload network demand; whereas Falcocchio & Levinson (2015) cites US highway estimates closer to 33-45%. Given these figures, we expect congestion in segments affected by high recurring bottlenecks to represent a major proportion of the overall congestion experienced across the road network.

Traffic performance is measured separately by three congestion metrics: average daily delay (ADD) and two widely-used metrics of travel time reliability (TTR), selected based on the criteria Gong & Fan (2017), the Planning Time Index (PTI) and Frequency Of Congestion (FOC). PTI_{80} is defined as the ratio of the 80-th percentile of travel time to the free flow travel time, whereas FOC is specified as the proportion of time that travel is classified as congested. To calculate PTI, the free flow travel time $tt_f(s_j)$ is obtained

as the ratio of the free flow speed to segment length $tt_f(s_j) = \frac{v_f(s_j)}{l(s_j)}$. The 80-th percentile is used instead of 90-th or 95-th percentiles as it is less sensitive to random events such as weather or traffic accidents (Systematics et al., 2013).

ADD is obtained by taking the average of the daily sums of segment delay across all observed (non-outlier) days. Segment delay is calculated using Equation 5.6 and is applied to time periods classified as congested, i.e. wherein the observed mean speed is below the threshold $v_f(s_j) \cdot \theta_u$. Equation 5.6 is used instead of Equation 5.4 because we are measuring the delay experienced in any given segment (which may be attributed to a bottleneck in the current segment, a bottleneck in a downstream segment, or neither) as opposed to measuring the total delay inflicted by the bottleneck (which may include several segments). The same threshold criteria is used to calculate FOC. FOC and ADD differ in that FOC only captures the duration/frequency of congestion, whereas ADD captures its intensity, by considering traffic volume and the degree of deviation from free flow conditions.

Table 5.4 summarises the results of the performance analysis. Besides their recurring classification (recurrence factor $R > 0.66$), segments are grouped by county and time of day to highlight spatio-temporal variations in performance and underlying monitoring capabilities. Performance metrics PTI and FOC are summarised across each group using the mean and standard deviation, whereas TADD captures the total average daily delay of segments within the group. Additionally, we introduce two measures $\frac{\text{TADD}_\text{b}}{\text{TADD}}$ and $\frac{\text{FOC}_\text{b}}{\text{FOC}}$ to capture the proportion of congestion that is localised, i.e. attributed to bottleneck activation, or queue spillback from a bottleneck downstream of the affected segment.

Overall, the results show that traffic congestion is disproportionately experienced along segments affected by high recurring bottlenecks. On average, congestion not only occurs more frequently in these segments (by a factor of four), but is also more intense than congestion in other segments (almost by a factor of two). These observations are consistent across different metrics, outlined below.

The mean PTI value of high recurring segments is always greater than that of non-recurring segments and generally above/close to the 1.75 and 2.50 thresholds employed by Systematics et al. (2013) and Wolniak & Mahapatra (2014) to categorise a roadway segment as extremely unreliable. The mean FOC value of high recurring segments is always greater than that of non-recurring segments and often above 0.50, indicating that these segments experience congestion during at least 50% of the day (6am-19pm). Congestion is generally less intense during the inter-peak period *ip*, but still considerable in some cases, particularly Newcastle (PTI of 2.19 for high recurring segments).

Despite representing a small fraction of total segment length, high recurring segments show a combined TADD greater than that of non-recurring segments during the *am* and

Table 5.4: Congestion analysis of 275 bottleneck-identifiable segments, grouped by origin county and recurring bottleneck classification, for each of the time periods: morning peak *am* (6-9am), inter-peak *ip* (10am-14pm) and evening peak *pm* (15-19pm).

County	Time	Rec*	N†	Length‡	TADD§	$\frac{\text{TADD}_b}{\text{TADD}}$¶	PTI_{80}‖	FOC**	$\frac{\text{FOC}_b}{\text{FOC}}$††
Gateshead	am	no	52	106.6	206	0.478	1.47 (0.38)	0.105 (0.12)	0.410 (0.30)
		yes	11	22.2	450	**0.846**	**3.15** (0.72)	**0.506** (0.10)	0.803 (0.12)
	ip	no	59	123.2	125	0.590	1.33 (0.19)	0.060 (0.08)	0.519 (0.29)
		yes	4	5.5	34	**0.847**	**2.00** (0.10)	0.484 (0.13)	0.796 (0.20)
	pm	no	56	116.4	256	0.448	1.45 (0.26)	0.113 (0.11)	0.402 (0.27)
		yes	7	12.4	304	**0.795**	**2.72** (0.54)	**0.547** (0.10)	0.751 (0.22)
Newcastle	am	no	86	163.1	391	0.403	1.55 (0.36)	0.131 (0.12)	0.340 (0.27)
		yes	15	38.6	288	**0.942**	**2.54** (0.44)	0.384 (0.07)	0.899 (0.04)
	ip	no	90	183.5	163	0.469	1.43 (0.20)	0.087 (0.12)	0.397 (0.26)
		yes	11	18.2	204	**0.851**	**2.19** (0.29)	**0.549** (0.20)	0.802 (0.17)
	pm	no	77	141.1	560	0.331	1.74 (0.45)	0.206 (0.18)	0.276 (0.19)
		yes	24	60.6	865	**0.825**	**3.18** (1.06)	**0.549** (0.16)	0.785 (0.20)
North Tyneside	am	no	23	64.0	150	0.452	1.46 (0.38)	0.115 (0.12)	0.389 (0.32)
		yes	6	15.7	188	**0.884**	**3.05** (0.97)	**0.542** (0.19)	0.840 (0.17)
	ip	no	27	75.8	64	0.582	1.31 (0.22)	0.061 (0.09)	0.527 (0.32)
		yes	2	3.9	29	**0.789**	**1.96** (0.07)	0.411 (0.02)	0.692 (0.09)
	pm	no	23	68.6	144	0.430	1.40 (0.25)	0.093 (0.10)	0.386 (0.32)
		yes	6	11.1	210	**0.771**	**4.10** (1.06)	**0.674** (0.15)	0.676 (0.17)
South Tyneside	am	no	16	38.2	50	0.470	1.38 (0.19)	0.085 (0.09)	0.387 (0.29)
		yes	1	3.8	32	**0.811**	**2.15**	0.314	0.746
	ip	no	17	42.0	54	0.608	1.36 (0.24)	0.069 (0.11)	0.545 (0.28)
		yes	0	-	-	-	-	-	-
	pm	no	17	42.0	53	0.525	1.40 (0.21)	0.084 (0.10)	0.473 (0.26)
		yes	0	-	-	-	-	-	-
Sunderland	am	no	60	138.3	70	0.478	1.34 (0.20)	0.060 (0.07)	0.418 (0.26)
		yes	3	6.9	29	**0.909**	**2.13** (0.62)	0.303 (0.12)	0.877 (0.12)
	ip	no	61	142.9	77	0.526	1.36 (0.19)	0.069 (0.08)	0.477 (0.29)
		yes	2	2.4	14	**0.903**	**1.96** (0.10)	**0.502** (0.09)	0.865 (0.08)
	pm	no	58	137.1	155	0.470	1.43 (0.33)	0.080 (0.09)	0.413 (0.30)
		yes	5	8.1	55	**0.855**	**2.61** (0.62)	**0.500** (0.09)	0.805 (0.13)
Total	am	no	239	520.4	884	0.446	1.46 (0.33)	0.103 (0.11)	0.382 (0.28)
		yes	36	87.2	986	**0.897**	**2.77** (0.71)	0.439 (0.13)	0.854 (0.11)
	ip	no	256	577.6	483	0.530	1.37 (0.20)	0.072 (0.10)	0.466 (0.29)
		yes	19	30.0	282	**0.849**	**2.11** (0.25)	**0.516** (0.17)	0.796 (0.16)
	pm	no	233	515.4	1168	0.415	1.53 (0.38)	0.131 (0.14)	0.364 (0.27)
		yes	42	92.2	1434	**0.816**	**3.16** (1.02)	**0.561** (0.15)	0.766 (0.19)
	all	no	206	452.9	1870	0.434	1.40 (0.21)	0.085 (0.08)	0.375 (0.25)
		yes	69	154.7	3367	**0.780**	**2.11** (0.63)	0.338 (0.15)	0.723 (0.18)

* High recurring traffic bottleneck ($R > 0.66$)
† Segment count within group (County, Time, Recurring).
‡ Sum of segment lengths in km.
§ Total Average Daily Delay (TADD) in vehicle-hours.
¶ Proportion of TADD attributed to localised congestion. Values greater than 0.75 shown in bold.
‖ 80th-percentile Planning Time Index (PTI), averaged over N (± s.d.) Values greater than 1.75 shown in bold.
** Frequency Of Congestion (FOC), averaged over N (± s.d.). Values greater than 0.50 shown in bold.
†† Proportion of FOC attributed to localised congestion (active bottleneck or spillback), averaged over N (± s.d).

pm periods, except for the *ip* period. For example, high recurring segments in Gateshead during the *am* period account for about 69% of the total experienced delay while accounting for less than 17.5% of segment length. The two metrics PTI and FOC are strongly correlated ($\rho \approx 0.90$), whereas TADD is moderately correlated with PTI ($\rho \approx 0.70$) and FOC ($\rho \approx 0.65$). The agreement between ADD and the two widely used metrics of TTR suggests that vehicle delay is being measured correctly, and that ADD can be employed as an alternative or complementary metric of congestion.

The proportion of delay due to localised congestion, i.e. attributed to traffic bottlenecks, is on average greater than 75% for high recurring segments while being generally lower than 50% in non-recurring segments. Thus, localised congestion displays a major piece of traffic congestion sources even in non-recurring segments. These figures are supported by the two indicative metrics $\frac{\text{TADD}_\text{b}}{\text{TADD}}$ and $\frac{\text{FOC}_\text{b}}{\text{FOC}}$, calculated from total delay and temporal frequency, respectively. The two metrics are strongly correlated ($\rho \approx 0.970$) suggesting that either metric can be used to measure the proportion of congestion due to traffic bottlenecks.

Figure 5.8 further illustrates segments affected by high recurring bottlenecks during morning and evening periods. High recurring segments are shown in red (36 in the *am* period and 42 in the *pm* period), and non recurring segments are shown in black (including non-identifiable segments). In addition to time of day, segments are split by orientation (a segment's orientation is predominantly `N/E` or `S/W`) to reduce visual overlap. The numerical labels represent the top 10 bottleneck-affected segments, in each direction, ranked by average daily delay, and are listed on Table 5.5.

A joint analysis of Figure 5.8 and Tables 5.5 and 5.4, highlights additional travel/congestion patterns in the region. Notably, the intensity, regularity and extent of bottleneck-induced congestion affecting Northbound AM traffic into Newcastle and, conversely, Southbound PM traffic leaving Newcastle. The Tyne Bridge acts as a point of confluence that appears to leave behind a trail of congestion across different corridors, affecting commuters to the South of the Tyne river. The Nomis 2011 census statistics report private vehicles to be the main mode of travel to work in these communities (Office for National Statistics, 2011).

The extent of congestion is such that it appears to affect a large component of the system rather than being localised. However, this distinctive visual element is in part due to the nature of the data and demands careful interpretation. In the case of Tyne Bridge, multiple bottleneck-affected segments partially overlap, since they have different origin locations but share one destination point. Consequently, this instance does not reveal as many bottleneck locations as bottleneck-affected segments, but a single bottleneck that causes congestion to spread across distinct urban corridors that converge on the troublesome location (segments labelled 1,5,6,10 AM N/W on the top left panel and

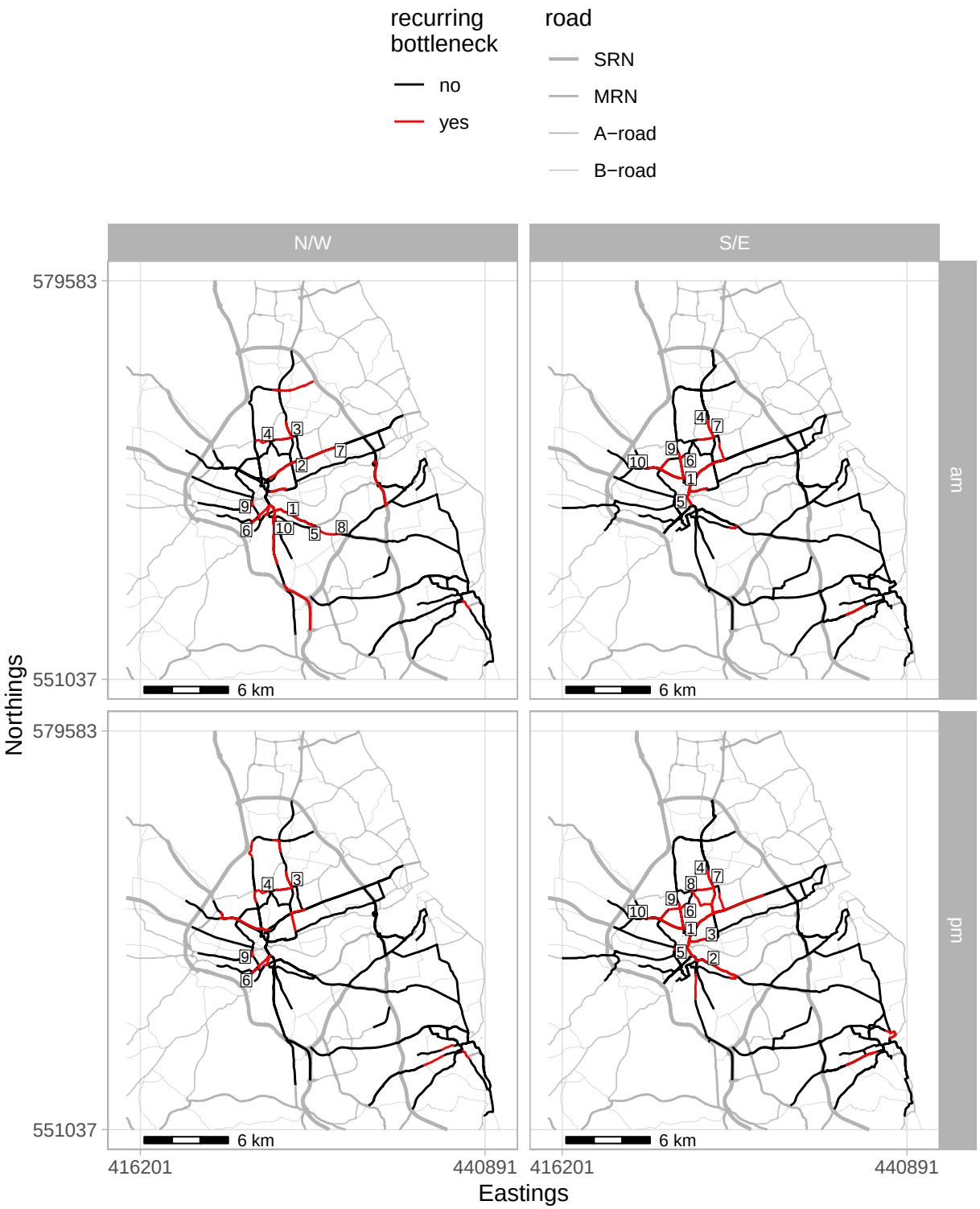


Figure 5.8: Segments affected by high recurring bottlenecks ($R > 0.66$), by time of day and spatial orientation in the county of Tyne and Wear. Labels cross-reference segments by their rank to Table 5.5.

Table 5.5: Ranking of top 10 high recurring segments by average daily delay, for each spatial orientation group N/W and S/E.

dir	rank	R^*	delay† (v-h)	duration†† (h)	extent†† (km)	onset¶ (am)	onset¶ (pm)	PTI§	FOC§	v_f (km/h)	AADT (K)	road	road name
NW	1	0.79	146 (0.9)	2.7 (1.8)	2.2 (0.4)	07:08	-	3.8	0.51	41.9	4.8	A167-A184	Gateshead Stadium - Tyne Bridge
	2	0.91	129 (0.6)	4.3 (2.1)	3.3 (1.0)	07:29	-	2.9	0.45	48.7	5.8	A1058	Coast Road - Jesmond Road
	3	1.00	116 (0.3)	8.0 (2.0)	3.5 (1.2)	07:27	15:03	4.0	0.74	39.1	2.1	A188-A191	Benton Park Road
	4	1.00	104 (0.4)	6.6 (2.2)	1.5 (0.8)	07:48	15:19	5.0	0.79	43.0	3.8	A191	Church Road - Station Road
	5	0.80	104 (0.8)	2.9 (1.2)	5.1 (1.8)	07:25	-	2.5	0.44	43.9	2.8	A167-A184	Newcastle Road - Tyne Bridge
	6	0.96	100 (0.7)	4.0 (2.0)	2.5	07:27	16:20	3.8	0.56	41.1	3.4	A167-A184	Derwentwater Road - Tyne Bridge
	7	0.79	96.8 (0.5)	2.7 (1.4)	5.9	07:29	-	2.2	0.36	63.9	3.5	A1058	Coast Road - Jesmond Road
	8	0.82	84.6 (0.6)	2.7 (1.3)	2.9 (1.2)	07:21	-	3.3	0.51	79.4	5.2	A184	Newcastle Road
	9	0.98	70.6 (0.7)	5.5 (2.8)	0.7 (0.4)	07:28	15:55	2.1	0.64	40.7	7.4	A189	Redheugh Bridge - St James Boulevard
	10	0.81	61.5 (0.5)	3.6 (1.5)	2.7 (0.7)	07:03	-	4.2	0.50	43.7	2.5	A167	Durham Road - Tyne Bridge
SE	1	0.99	272 (0.4)	8.5 (3.2)	4.3 (0.9)	07:23	15:14	5.4	0.73	59.9	5.4	A1058-A167	Jesmond Road - Tyne Bridge
	2	0.90	183 (0.6)	4.0 (2.3)	3.3 (1.2)	-	15:48	3.4	0.46	50.7	10.4	A184	Gateshead Stadium - Newcastle Road
	3	0.89	128 (0.7)	4.4 (2.5)	5.3 (1.4)	-	16:14	2.5	0.38	49.9	6.3	A1058	Coast Road - Jesmond Road
	4	1.00	94.7 (0.4)	6.8 (1.8)	2.5	06:50	15:16	4.0	0.75	40.5	1.8	A188-A191	Benton Park Road
	5	0.99	89.3 (0.4)	10.0 (3.1)	2.5 (0.2)	07:24	15:11	4.4	0.75	53.5	2.0	A167-A193	New Bridge Street - Tyne Bridge
	6	0.98	86.6 (0.5)	5.8 (2.0)	6.3 (2.0)	07:29	15:26	4.2	0.68	58.6	1.4	A167-A189	Jesmond Dene Road - Tyne Bridge
	7	1.00	78.7 (0.5)	7.4 (2.1)	2.0 (0.9)	06:42	15:52	5.6	0.84	41.5	1.9	A188	Benton Road
	8	1.00	74.2 (0.5)	6.1 (2.7)	1.9 (0.7)	-	15:12	4.8	0.71	35.3	2.4	A191	Benton Park Road - Front Street
	9	0.98	65.4 (0.4)	7.6 (2.6)	8.5 (2.1)	07:25	15:25	4.2	0.67	63.5	1.0	A167-B1318	Great North Road - Tyne Bridge
	10	0.98	54.8 (0.5)	4.6 (1.5)	6.6 (2.0)	07:32	15:30	3.6	0.64	71.4	1.2	A167	Ponteland Road - Tyne Bridge

* Recurrence factor.

† Average daily delay (coefficient of variation) under recurring conditions.

‡ Daily average (standard deviation) under recurring conditions.

¶ Mean bottleneck onset time under recurring conditions (symbol '-' represents $R < 0.66$ for said time period).

§ Measured during the time period *am* or *pm* where congestion is worse (i.e. FOC is maximised).

segments 1,5,6,9,10 PM S/W on the bottom right panel of of Figure 5.8).

For top recurring bottlenecks such as those near the Tyne Bridge and Benton/Coast Road accesses, bottleneck impact is thus made worse by the convergence of several traffic corridors or the unavailability of alternative paths to destination due to geographical constraints (i.e. the Tyne River). The analysis of onset times for converging segments can help authorities understand where congestion triggers first and further investigate the mechanisms by which urban congestion propagates backwards due to sustained bottleneck activation (for instance, Table 5.5 shows that N/W ranks 1 and 10 have earlier onset times than N/W ranks 5 and 6, all of which converge on the same destination location).

5.5.3 Ranking of high recurring bottlenecks

For traffic professionals, it is often of interest to rank bottlenecks according to their impact on traffic flow (Hale et al., 2016). Table 5.5 partially demonstrates how this can be done for segments affected by high recurring bottlenecks using vehicle delay – the main proposed measure of bottleneck intensity.

To evaluate the delay-based ranking of high recurring bottlenecks, we compare it to the ranking otherwise obtained by employing a measure of travel time reliability, specifically FOC and PTI. Unlike segment-specific delay (ADD), as shown in Table 5.4, average bottleneck daily delay (ABDD) is measured using Equations 5.4 through 5.6, which not only take into account the delay caused to the bottleneck-affected segment but also any delay caused to segments preceding it. Therefore, we expect TTR-based rankings to differ from delay-based ranking insofar that delay takes into factors that FOC and PTI do not.

Table 5.6: Kendall's rank correlation coefficient for pair-wise comparison of bottleneck rankings obtained using different congestion metrics.

Metric 1	Metric 2	Control	Kendall's tau
ABDD	PTI	no	0.190
ABDD	FOC	no	0.086
ABDD	FOC_b	no	0.211
ABDD	ADD	yes	0.778
PTI	FOC	yes	0.815

For a comparison of rankings obtained from different measures, we use Kendall's rank correlation coefficient – a measure of ordinal association between two variables (Kendall, 1945). The correlation coefficients for 5 pairs of metrics are shown in Table 5.6. Two

control cases are included for reference: it is expected that FOC and PTI produce similar rankings (high correlation), and equally ABDD and ADD.

The resulting values show some, but little agreement between delay-based and TTR-based rankings, and otherwise strong correlation between controls. Note how there is strong but not complete agreement between bottleneck and segment-specific delay, due to the consideration of spillback effects. In addition, while FOC and PTI rankings show strong agreement, they do not agree entirely – different metrics will inevitably offer different assessments (e.g. Gong & Fan (2017) recommends that multiple complementary metrics should be used to provide a fuller assessment). A delay-based ranking is preferable if traffic volume and spillback effects are considered more relevant than other variables, such as travel time reliability. Alternatively, experts may prefer a ranking metric that is the combination of several relevant assessment metrics (Hale et al., 2021).

5.6 Discussion and future work

Overall, it was shown that ANPR networks are capable of highlighting urban roadway segments affected by recurring traffic bottlenecks and measuring their impact on traffic flow. Obtained results support the view that traffic bottlenecks are a major source of urban traffic congestion. In particular, high recurring bottlenecks were found to account for nearly two thirds of the total vehicle delay observed in the region of Tyne and Wear. In addition, high recurring segments register, on average, three to four times more frequent congestion (overall FOC value of 0.338 ± 0.15 against 0.085 ± 0.08 for other segments) and generally of greater intensity than that displayed in other segments (overall PTI of 2.11 ± 0.63 against 1.40 ± 0.21 for other segments). Lastly, it was exemplified how segments can be ranked according to their impact on traffic flow, so that mitigation strategies on problematic sites can be appropriately prioritised given limited funding.

One possible criticism is that the presented approach is largely retrospective and lacks predictive and explanatory elements. However, its main purpose is that of classification – to identify road segments and time periods experiencing bottleneck-like congestion (according to a theoretically principled definition of bottleneck phenomena) – so that city-wide congestion patterns can be understood from historical data. The onset and evolution of congestion depends on the features of traffic, as well as those of the affected location and surrounding roads. The study of congestion and circumstances that give rise to it (the phenomenon known as traffic breakdown) are long-standing areas of research, which are outside the scope of this work (the study of what causes an impactful bottleneck can be done following its detection).

Despite this not being a predictive method, it can be combined with other methods, par-

ticularly machine learning models, to generate predictions of bottleneck activation and queue spillback. In particular, neural network-based models, such as (Lv et al., 2015; Polson & Sokolov, 2017), are capable of learning the complex spatio-temporal nature of traffic flow, at the cost of interpretability. These models can be used to generate accurate predictions of vehicle speed and then fed to our algorithm for predictive bottleneck analysis. Because our approach is fully deterministic and therefore more transparent and easier to interpret, the resulting model ensemble would arguably lead to more interpretable results than a purely neural-network based approach.

An immediate benefit of this approach is that the activation model does not include any region-specific information (spatial data is used only insofar as distance calculations are required). Therefore, we anticipate the approach to be directly applicable to other cities, particularly those where the technology has already been deployed at scale. Within the same city, it can potentially be used to study mobility patterns over time – notably the contrast in travel and congestion patterns before, during and after the several UK lockdowns experienced throughout 2020 and 2021.

Another distinctive feature of this methodology is that it builds on the concept of corridors for generalised bottleneck discovery (i.e. identification across entire urban networks and not constrained solely to highways). This choice was made because urban traffic is not restricted in the same way that highway traffic is. Knowledge of vehicles' trajectory is required to test bottleneck behaviour on urban networks. Otherwise, it is impossible to tell whether a bottleneck-like relationship is real or not: if the speed profile of two roads matches that of bottleneck-like behaviour but few or no vehicles actually traverse the resulting route, then the bottleneck relationship is not meaningful for that pair of routes. Corridors fulfil this role of providing trajectory data to the bottleneck detection algorithm. In addition, the use of corridors greatly restrict the search space of possible bottlenecks (from a combinatorial point of view). By contrast, network-based approaches are limited in the amount or quality of trajectory data available and studies like Gong & Fan (2017) compromise by adopting very broad definitions of bottleneck-behaviour.

Despite the advantages of ANPR in monitoring urban corridors, the technology has two inherent limitations to retain. First, it can not determine the specific location of bottlenecks, i.e. the point where queues start to form, only the roadway segments that are likely to contain them. Thus, traffic authorities will generally need to perform a follow-up analysis in order to ascertain the precise bottleneck location and underlying causes. Second, the procedure requires that segments are mapped within available/known ANPR corridors and is limited to evaluating segments that are not the last corridor segment. Moreover, it can not be applied when two cameras are spaced far enough that the effect of localised congestion is effectively diluted by long distance travel.

Finally, there are aspects of our work that further research can look to improve on. First,

it has not been fully established how impact metrics are affected by changes in parameter values. The choice of parameter values was primarily supported by exploratory analysis and reference values taken from previous studies, (Gong & Fan (2017) reports $\theta_u = 0.6$ as a reasonable value; Chen et al. (2004) uses the equivalent of $\theta_u = 2/3$ and $\theta_d = 1/3$). Although bottleneck analysis seems to accept a range of parameter values, it remains to be studied how sensitive the output is to more significant deviations from these values. Second, treatment of missing data and reduced detection rates was achieved by omission, i.e. by discarding observations. Alternative treatments based on data imputation may be required to avoid developing assessments biased against segments prone to this type event.

An open question is the inclusion of traffic volume in the calculation of vehicle delay. In some cases, measurements of traffic volume may be unreliable because segments will contain junctions where traffic can merge into or exit along the way (other than the origin and destination ANPR camera locations). For example, DfT manual count points in St James Boulevard, and Barrack Road predict AADT values that are about half of those obtained from ANPR data (Department for Transport, 2019c, 2019b). Unless segments are equally affected this way (in which case true vehicle delay is proportional to measured values), assessments of vehicle delay will be biased towards segments with fewer junctions in between the origin and destination points. To correct for this bias, one may look to include more reliable sources of traffic volume data, or calibrate the final assessment results to take into account the proportion of traffic volume unaccounted by ANPR cameras.

Chapter 6

Towards a traffic flow model of overtaking rate

6.1 Introduction

Lane changing (LC) is a fundamental driving task, performed out of necessity or discretion (Gipps, 1986). It has been consistently observed that forced or frequent lane changing can have a significant influence on surrounding vehicles, induce traffic oscillations and negatively impact road safety (Mattes, 2003; Pande & Abdel-Aty, 2006; Srivastava & Geroliminis, 2013; Z. Zheng et al., 2011). In highways, LC has been shown to explain the capacity drop phenomenon, wherein traffic breakdown causes a reduction in maximum traffic flow rate (Ahn & Cassidy, 2007; Cassidy & Rudjanakanoknad, 2005; W.-L. Jin, 2010; Laval & Daganzo, 2006; Sala & Soriguera, 2020; Z. Zheng, 2014). In undivided roads, lane changes (better known as overtakings) are significantly more dangerous manoeuvres as they involve momentarily driving against incoming traffic (Clarke et al., 1999; Hegeman et al., 2004).

Understanding and modelling LC is therefore critical to traffic planning (Toledo et al., 2003), the design of Advanced Driver Assistance Systems (ADAS) (Jula et al., Nov./2000; J. Zhou & Peng, 2005) and, more recently, the development of connected and autonomous vehicles (CADs) (P. Cao et al., 2017). Despite new advances, there is a disconnect between microscopic LC models (LCD) – which capture the decision-making and execution processes of individual vehicles – and macroscopic LC models (LCI) – which capture the impact of LC on surrounding vehicles and the traffic stream (Z. Zheng, 2014). The lack of a comprehensive modelling approach thus limits our ability to replicate LC-induced phenomena realistically across a wide number of traffic contexts (M. Li et al., 2020). Moreover, research is segregated by road type/function, even though the underlying LC decision-making/execution process is unlikely to be fundamentally different (Z. Zheng,

2014) (apart from the added complexity of overtaking manoeuvres on single carriageways (Hegeman et al., 2004; Wilson & Best, 1982)).

The key limiting factor of LCI modelling is the difficulty to collect lane changing data at scale, especially since conventional traffic sensors, namely loop detectors, do not own such capabilities (Laval & Daganzo, 2006; Sala & Soriguera, 2020; Z. Zheng, 2014). Several authors have used human observers to collect LC data, but with severe limits in sample size and length, for example (Hegeman et al., 2004; Marczak et al., 2016). Naturalistic studies and NGSIM data (Alexiadis et al., 2004; S. E. Lee et al., 2004), were designed to fill this gap and have produced numerous insights into LC behaviour (Klauer et al., 2006). This includes empirical distributions for many microscopic variables, in particular gap length, LC duration and time-to-collision (Olsen et al., 2002; Yang et al., 2019). However, these data sources are limited in regards to sample size (naturalistic studies follow individual drivers) or road diversity (NGSIM data are collected using a specialised monitoring system) and have yet to offer a complete and diverse macroscopic analysis of LC frequency (Coifman & Li, 2017; Punzo et al., 2011).

In seeking to contribute to ongoing LCI research, we present a study using data collected from Automatic Number Plate Recognition (ANPR) cameras. ANPR technology is mature and widely used in urban centres, historically with a focus on law enforcement (Patel et al., 2013). Recently, ANPR has been increasingly applied in traffic monitoring and control, because it is a reliable source of journey time data, especially on urban roads where travel times are not easily estimated using loop detectors (Luo et al., 2019; Watson, 2017; F. Zheng et al., 2018). We leverage the reliability of ANPR travel times to compute vehicle overtakings along a route – a quantity assumed to be a (lower-bound) surrogate measure of LC frequency. Simultaneously, we capture relevant flow variables, namely traffic volume and speed, whose relationship to overtaking rate is quantified using a statistical model. Furthermore, by annotating lane composition for a large number of ANPR routes we are able to examine overtaking rate patterns at scale and across road categories, namely single and multi-lane roads.

This chapter is structured as follows. Section 6.2 describes how vehicle overtakings are determined using trip data and combined to measure overtaking rate. Section 6.3 performs exploratory data analysis on a subset of ANPR data across two distinct route classes: single-lane single-carriageway routes and multi-lane dual-carriageway routes. Based on the outcome of exploratory analysis, Section 6.4 develops a statistical model of overtaking rate for multi-lane routes. Lastly, Section 6.5 summarises the contributions and results of this chapter, discusses limitations of this work and lists several directions for future work.

6.2 Measuring vehicle overtakings and overtaking rate

ANPR data can be used to count the number of overtakings performed by individual vehicles because it provides reliable vehicle timestamps at both the origin and destination cameras (F. Zheng et al., 2018). After ANPR raw data is adequately processed and vehicle trips identified, the answer to the question "How many vehicles did vehicle k overtake?" simply amounts to counting all vehicles whose journey started before k but ended after k. To formulate the same query analytically, suppose that we observe n distinct vehicle trips between origin camera l_i and destination camera l_j. Then, let $y(k, l_i, l_j)$ denote the number of overtakings performed by the k-th observed vehicle along route (l_i, l_j), with $k \in 1 \dots n$, and let

$$y(k, l_i, l_j) = \sum_{k'=1}^{n} \mathbb{1}\Big[t(k, l_i) > t(k', l_i)\Big] \cdot \mathbb{1}\Big[t(k, l_j) < t(k', l_j)\Big] \cdot \mathbb{1}\Big[\Delta t(k, k') > \epsilon\Big], \tag{6.1}$$

where $t(k, l_i)$ is the timestamp of the k-th vehicle at the origin camera; $t(k, l_j)$ is the timestamp of the k-th vehicle at the destination camera; $\Delta t(k, k') = \Big(t(k, l_i) - t(k, l_j)\Big) - \Big(t(k', l_i) - t(k', l_j)\Big)$ specifies the travel time difference between vehicles k and k' and ϵ the minimum accepted travel time difference. $\mathbb{1}[Q]$ is an indicator function that returns 1 if the logical statement Q is met (evaluates to *true*) and 0 otherwise. The first two indicator terms inside the sum represent an overtaking event – vehicle k' departed before k but arrived after - and the third term specifies that the event is not a fluke, i.e. that occurred due to rounding errors or other unmeasurable sources of error (once vehicles are within the camera's reach, approximately 50 meters, there is no guarantee that they are detected in the order at which they arrive). Note that n distinct vehicle trips does not imply n distinct vehicles as these can perform multiple trips along the same route over an extended period of time, i.e. hours, days or weeks.

In short, Equation 6.1 counts the number of overtakings performed by a given vehicle travelling between two cameras. Note that the recorded count is only a lower-bound because we are not able to record events of vehicles that take turns overtaking each other. In practice the computation can be significantly sped-up by sorting vehicles by time at the origin location and evaluating Equation 6.1 against the B vehicles preceding k. In that case, the sum $\sum_{k'=1}^{n}$ is written instead $\sum_{k'=k-B}^{k}$. We set $B = 100$ as it is generally unlikely that vehicles register 100 or more overtakings along a route shorter than 5 km (an assumption confirmed by exploratory data analysis).

To examine the overtaking behaviour of a cohort of vehicles, we count the total number of overtakings and normalise it per unit of distance. Let $y(l_i, l_j)$ be the mean number of vehicle overtakings per unit of distance given by

$$y(l_i, l_j) = \frac{\sum_{k=1}^{n} y(k, l_i, l_j)}{d(l_i, l_j)}, \tag{6.2}$$

where $d(l_i, l_j)$ specifies the length of route (l_i, l_j). Normalising the sum $\sum_{k=1}^{n} y(k, l_i, l_j)$ is not strictly necessary when modelling a single route, but becomes useful when considering routes with different lengths as longer routes may allow for more overtakings only because of their length. Equivalently, routes with greater capacity or AADT (Annual Average Daily Traffic) may report higher overtaking counts. Computationally, the window optimisation technique described above can be applied to reduce the time complexity of overtaking calculation from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. Moreover, the computation for multiple routes is embarrasingly parallel and can be further optimised.

6.3 Exploratory data analysis

Exploratory data analysis is an essential precursor to data modelling. The goal of exploratory data analysis is to identify relationships between variables of interest so as to inform future modelling decisions (Gelman, 2004; Tukey, 1977). In our case, exploratory data analysis serves an important role in choosing explanatory variables (predictors) and determining key relationship traits between predictor and response variables (sign, strength, linearity, constant variance, time of day effects and seasonal effects).

The object of our study is overtaking rate, measured in vehicles per km. Observations of overtaking rate are collected for every 10 minute interval, between 6am and 10pm, across two consecutive weeks of 2018, for a total of 127 distinct routes. Our goal is to investigate how macroscopic flow variables influence overtaking rate. Four traffic flow variables are recorded in addition to overtaking rate: traffic density (vehicles/km), mean speed (km/h), standard deviation of speed (km/h) and coefficient of variation of speed, defined as the ratio between the standard deviation of speed and its mean (no unit). During this exploratory phase, we consider both the standard deviation (SD) of speed and its coefficient of variation (CV), absolute and relative measures of speed dispersion, respectively.

The exploratory data analysis is performed for a sample of 127 routes representative of the diversity of route characteristics, namely number of lanes (single vs multi-lane), daily traffic volume and average free flow speed. In addition, three route characteristics were

manually annotated by visual inspection using Google Maps street view feature (Anguelov et al., 2010): number of lanes, type of carriageway and number of traffic light intersections along the route. Because a route is composed by a sequence of streets, possibly with different characteristics, two of these features were annotated using a number between 0 and 1, proportional to the estimated route composition (single lane = 0, two or more lanes = 1, single carriageway = 0, double carriageway = 1).

6.3.1 Single-lane routes

Figure 6.1 depicts a scatter plot of the response variable (overtaking rate) against each of the four explanatory traffic variables previously described. Data are shown for eight single-lane routes (lane < 0.25 and carriageway < 0.25), chosen across a range of free flow speeds, whose characteristics are further listed in Table 6.1. Overall, the plots show the absence of a clear relationship between overtaking rate and the predictor variables.

Table 6.1: Characteristics of eight selected single-lane single-carriageway routes.

route id	O	D	length (km)	AADT (k)	v_f (km/h)	direction	county	road	traffic lights
018	183	254	3.3	2.4	62.7	S-E	Sunderland	A1018	4
041	047	029	4.4	1.1	52.1	N-W	Sunderland	A184	3
042	150	080	1.4	1.4	51.7	W-W	Gateshead	A1114	0
058	040	183	3.0	1.5	47.8	S-S	Sunderland	A1018	1
077	024	079	1.0	1.9	44.0	S-S	Gateshead	B1296	1
103	221	111	1.9	1.0	39.1	N-N	Gateshead	A167	4
117	131	200	1.3	2.4	35.3	E-E	North Tyneside	A191	2
120	003	221	1.2	1.2	34.4	N-N	Gateshead	A167	2

6.3.2 Multi-lane routes

Similarly, we analyse overtaking rate across eight selected multi-lane routes (lane > 0.75 and carriageway > 0.50), whose characteristics are listed on Table 6.2. An analysis of the scatter plots shown in Figure 6.2 reveals several key elements of the relationship between overtaking rate and the measured traffic stream variables:

- Overtaking rate increases with traffic density consistently across cases.
- Overtaking rate increases inversely with traffic speed consistently across cases (except for route 033).
- No visible relationship between overtaking rate and the speed SD. Consequently, this variable is not considered any further.
- Inconsistent relationship between overtaking rate and speed CV – a positive increase is observed in some cases (routes 012, 059 and 069) but not in others.

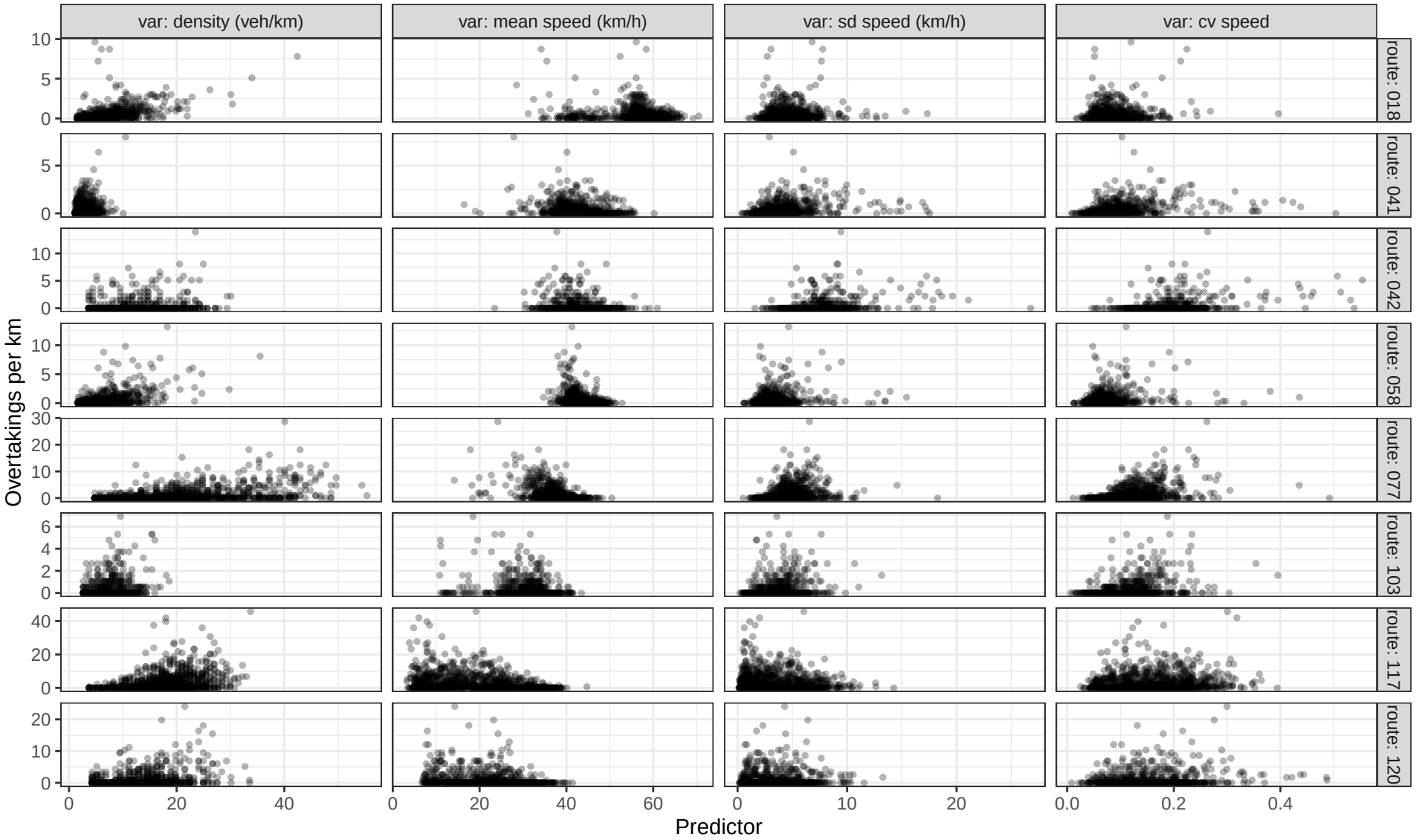


Figure 6.1: Scatter plots of overtaking rate against four traffic flow variables – density, mean speed, standard deviation of speed, coefficient of variation of speed – collected between 21 May and 01 July 2018 (weekdays only), across the eight single-lane routes listed on Table 6.1. Note that the scale of the y axis changes for each route so that the relationship between variables is more clearly visible.

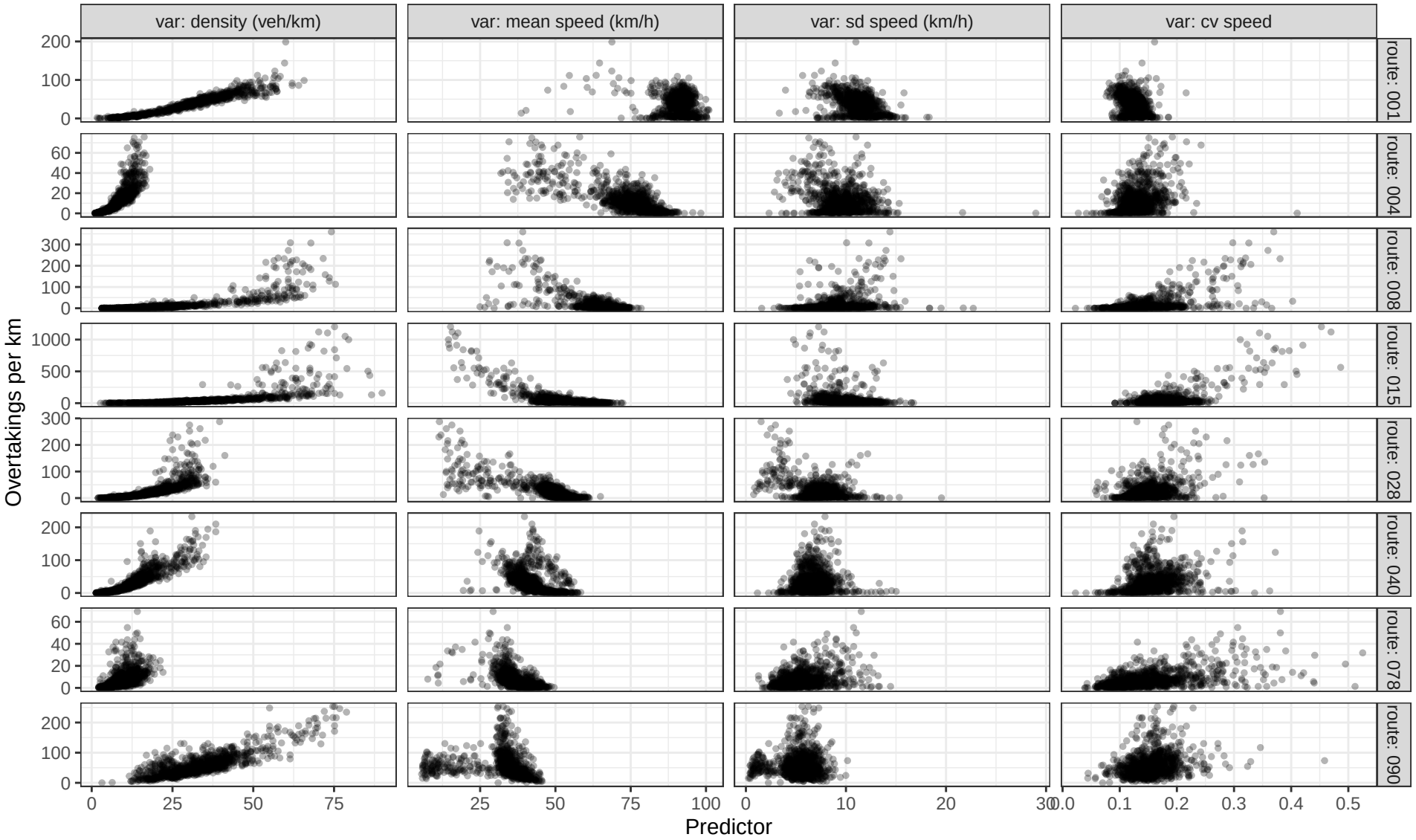


Figure 6.2: Scatter plots of overtaking rate against four traffic flow variables – density, mean speed, standard deviation of speed, coefficient of variation of speed – collected between 21 May and 01 July 2018 (weekdays only), across the eight multi-lane routes listed on Table 6.2. Note that the scale of the y axis changes for each route so that the relationship between variables is more clearly visible.

- Non-linear relationship between the response mean and each predictor.
- Evidence of heteroskedasticity (response variance is not constant).
- The intensity of the relationship varies across routes (note that the scale of the y axis changes for each route).

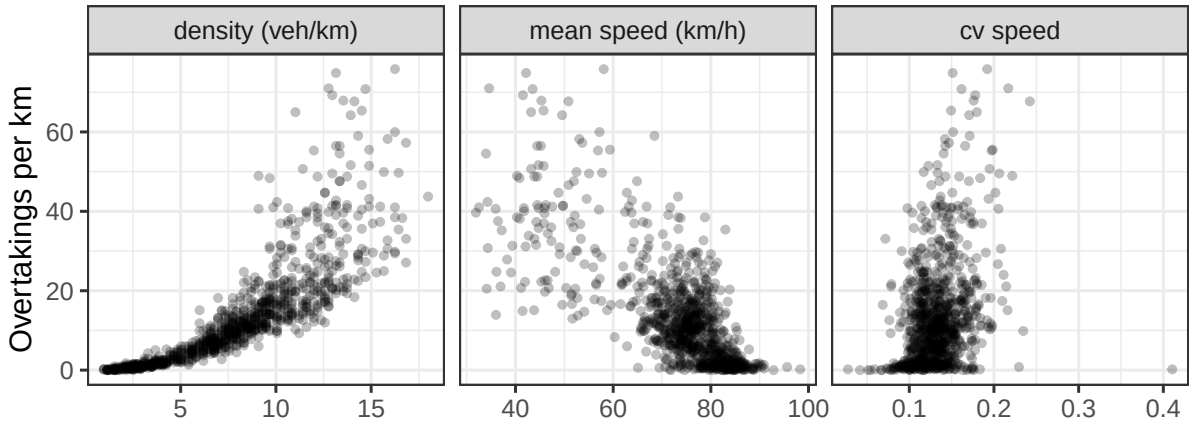
Table 6.2: Characteristics of eight selected multi-lane double-carriageway routes.

route id	O	D	length (km)	AADT (k)	v_f (km/h)	direction	county	road	traffic lights
001	100	069	3.0	9.2	93.3	W-W	Newcastle	A1058	0
004	016	255	5.2	4.2	82.4	W-W	Sunderland	A1231	0
008	116	114	1.6	3.7	71.3	S-S	North Tyneside	A188, A189	3
015	134	070	1.8	6.5	64.0	W-W	Newcastle	A695	5
028	169	016	2.5	3.9	55.7	W-W	Sunderland	A1231	2
040	096	256	3.6	5.3	52.3	W-W	South Tyneside	A1300, A194	0
078	020	198	2.3	2.3	44.0	S-W	South Tyneside	A1018, A194	2
090	073	089	1.9	6.9	41.9	S-E	Gateshead	A167, A184	4

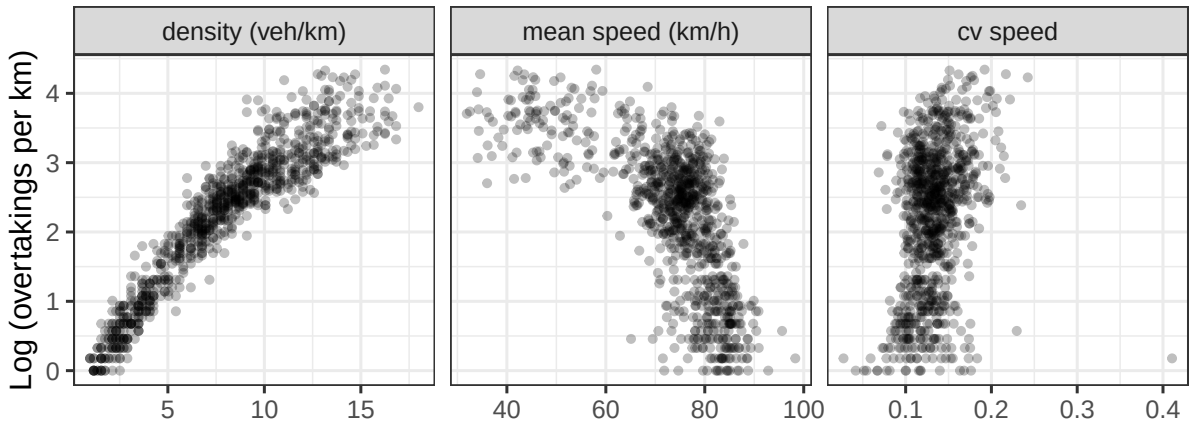
To address non-linearity and heteroskedasticity, the natural logarithm is applied to both the response and predictor variables. Because zeros in the response variable can arise when few or no vehicles/overtakings are observed (the function domain of the logarithm is $\mathbb{R}^+$), these are addressed by adding a small constant c to every output observation y_i so that $y_i + c > 0$. We choose $c = 1$ as $\log(y_i + 1)$ conveniently maps to zero when $y_i = 0$. The resulting transformation corresponds to a two parameter Box-Cox transform with $\delta_1 = 0$ and $\delta_2 = 1$ (Box & Cox, 1964). Conversely, observation samples registering no traffic density/speed (no vehicles are observed) are removed from the dataset as they are not useful to quantify the relationship between the response and predictors.

Figure 6.3 illustrates the effect of the log transformation for a single route: (a) original data, no transformation, (b) log transform applied only to the response variable, and (c) log transform applied to both the response and predictor variables. It can be seen that the linearisation effect of the transformation is greater when applied to both the response and predictor variables. At the same time, the variance of the response better approximates a constant value, suggesting that the transformation serves to treat some of the heteroskedasticity observed in Figure 6.2. However, the linearisation effect is greater in traffic density, than the other two variables.

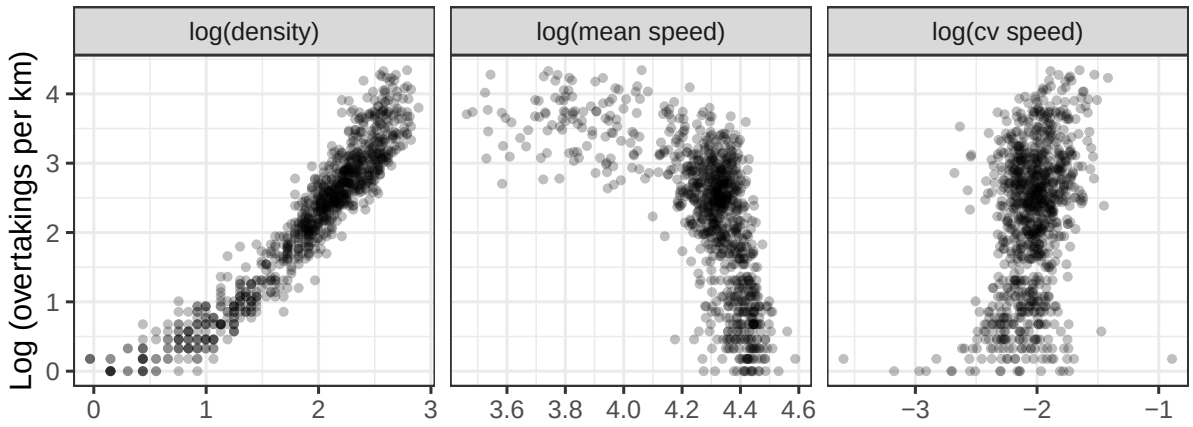
Next, we investigate whether the response-predictor relationship is affected by time of day. The motivation to consider time of day effects stems from observing large variations in overtaking rate for seemingly identical traffic conditions, i.e. similar traffic stream values. One plausible explanation is attributed to unobserved interactions with intersecting roads along the route, as these can feed more or less traffic to the monitored route depending on time of day.



(a) No transformation (original input data).



(b) Log transform applied only to the response variable.



(c) Log transform applied to both the response and predictor variables.

Figure 6.3: Comparison of different data transformation stages for multilane route 004.

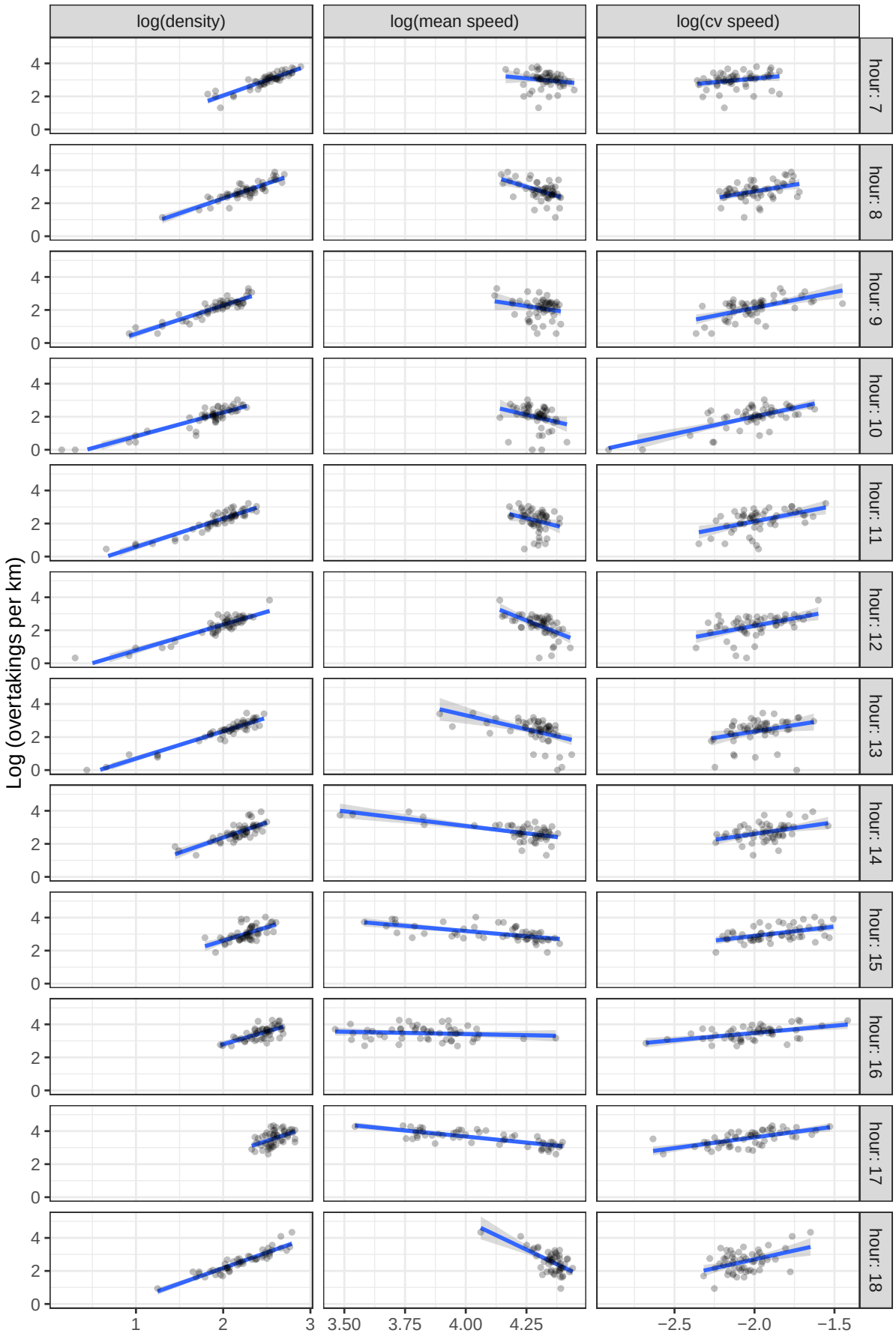


Figure 6.4: Scatter plots of overtaking rate faceted across time of day (hour) for route 104. A linear model fit (blue) is applied to each panel to highlight the trend. Non-parallel lines are indicative of time of day effect.

An analysis of Figure 6.4 suggests that time of day is a determining factor in the relationship between overtaking rate and traffic stream variables. This effect is particularly visible for the mean and CV measures of speed but less so for traffic density (the slope of the linear fits, shown in blue, fluctuates throughout the day for the speed related predictors while remaining mostly parallel for traffic density). Furthermore, within each hour, there is a reduction in the non-linearity and heteroskedasticity features of the relationship, thus better approximating a linear one (in the log domain). Together, these observations suggest that a mixed effects linear model may be capable of adequately quantifying the effect of traffic flow on overtaking behaviour in a multi-lane route. For simplicity, other time-related effects, namely week of year and day of week effects, are not further considered at this point.

6.4 Modelling overtaking rate in multi-lane routes

Based on the exploratory data analysis performed above, we develop a statistical model of overtaking rate for a multi-lane route with the objective of quantifying the effect of the various traffic flow variables on overtaking rate. Two types of statistical model are considered: a fixed effects model, which only captures the global effect of each predictor variable on the response, and a crossed random effects model (a mixture of fixed and random components), which allows for random intercepts and slopes at each hour of the day. The crossed random effects model is introduced to address the time of day effect previously identified and constitutes a compromise between a full random effects model, that would construct a different slope and intercept for each 10-minute period of the day, and a fixed effects model that does not consider time-dependence at all (Singmann & Kellen, 2019).

To define the models, let y_{jk} denote the logarithm of the j-th observation of overtaking rate at hour k, with $j = 1 \dots J$, where $J \leq 60 = 6 \times 10$ (6 ten-min periods, 10 days), and $k = 7 \dots 22$ is a factor variable responsible for the hourly grouping of time-dependent observations. Simultaneously, let $\boldsymbol{X}_{jk} = [1, x_{1jk}, x_{2jk}, x_{3jk}]$ be a vector containing the logarithm of the traffic stream variables of interest. Then, the general specification of the linear model is

$$y_{jk} = \boldsymbol{\beta}_k \boldsymbol{X}_{jk} + e_{jk}\,, \quad e_{jk} \overset{iid}{\sim} N(0, \sigma^2), \tag{6.3}$$

where $\boldsymbol{\beta}$ is the vector of model parameters (regression coefficients), interpreted as the mean percent change in y caused by a unit percentage change in x when all other variables are kept constant. For the fixed effects model (LM), the vector $\boldsymbol{\beta}_k = \boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]$

is constant across k and each coefficient represents the global effect of the corresponding predictor. In turn, for the crossed random effects model (CELM), the vector $\boldsymbol{\beta}_k = [\beta_{0k}, \beta_{1k}, \beta_{2k}, \beta_{3k}]$ represents the intercept and effects of each predictor grouped by hour of the day. In both cases, homoeskedaticity is assumed (constant variance).

To determine whether the effect of the intercept and variable x_1 is better modelled globally or hour-specific, four competing parametrisations of the CELM model are considered: fixed intercept and traffic density effects (β_0, β_1), random intercept and fixed traffic density effects (β_{0k}, β_1), fixed intercept and random traffic density effects (β_0, β_{1k}) and random intercept and random traffic density effects (β_{0k}, β_{1k}). The effects of the other two predictors are assumed to be hour-specific in all four parametrisations (β_{2k}, β_{3k}). The different model parametrisations are motivated by the fact that, in Figure 6.4, the effect of traffic density seems to vary less across hours than the other two predictors.

The model fitting process is exemplified for route `004`. Table 6.3 shows the results of model comparison using different goodness of fit measures, specifically the log likelihood, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). Maximal values of log likelihood and minimal values of AIC and BIC are indicative of improved model performance. In CELM models, computation can lead to singular fits, i.e. estimated parameters on the boundary of the feasible parameter space, which may be indicative of overfitting (Gelman & Hill, 2006). In the example, model singularity is observed only for model number 5 (random intercept and random slopes CELM), but not in other model parametrisations. Therefore, of the non-singular model fits, the preferred CELM parametrisation, according to calculated model comparison metrics, is the fixed intercept and random slopes model with parameter vector $\boldsymbol{\beta} = [\beta_0, \beta_{1k}, \beta_{2k}, \beta_{3k}]$ and a total of $17\times3+4 = 55$ parameters (51 hour-specific slope parameters + 4 fixed effects parameters). This model choice is consistent with the "keep it maximal" principle of Barr et al. (2013), which gives preference to the most complex model compatible with the experimental design, removing only terms necessary to allow for a non-singular fit. All parameter estimates were found to be statistically significant.

Table 6.3: Performance metrics of LM and CELM models for route 004.

	model	singular	Adjusted R^2	s*	logLik	AIC	BIC
1	LM	no	0.946	0.243	-1.3	12.7	36.8
2	CELM (β_0 , β_1)	no	-	0.217	29.2	-44.4	-10.6
3	CELM (β_{0k}, β_1)	no	-	0.216	29.6	-43.2	-4.6
4	CELM (β_0 , β_{1k})	no	-	0.206	66.3	-116.6	-78.1
5	CELM (β_{0k}, β_{1k})	yes	-	0.206	99.5	-171.0	-103.5

* Residual sum of squares.

The overall adequacy of the LM and CELM models is determined by analysis of the

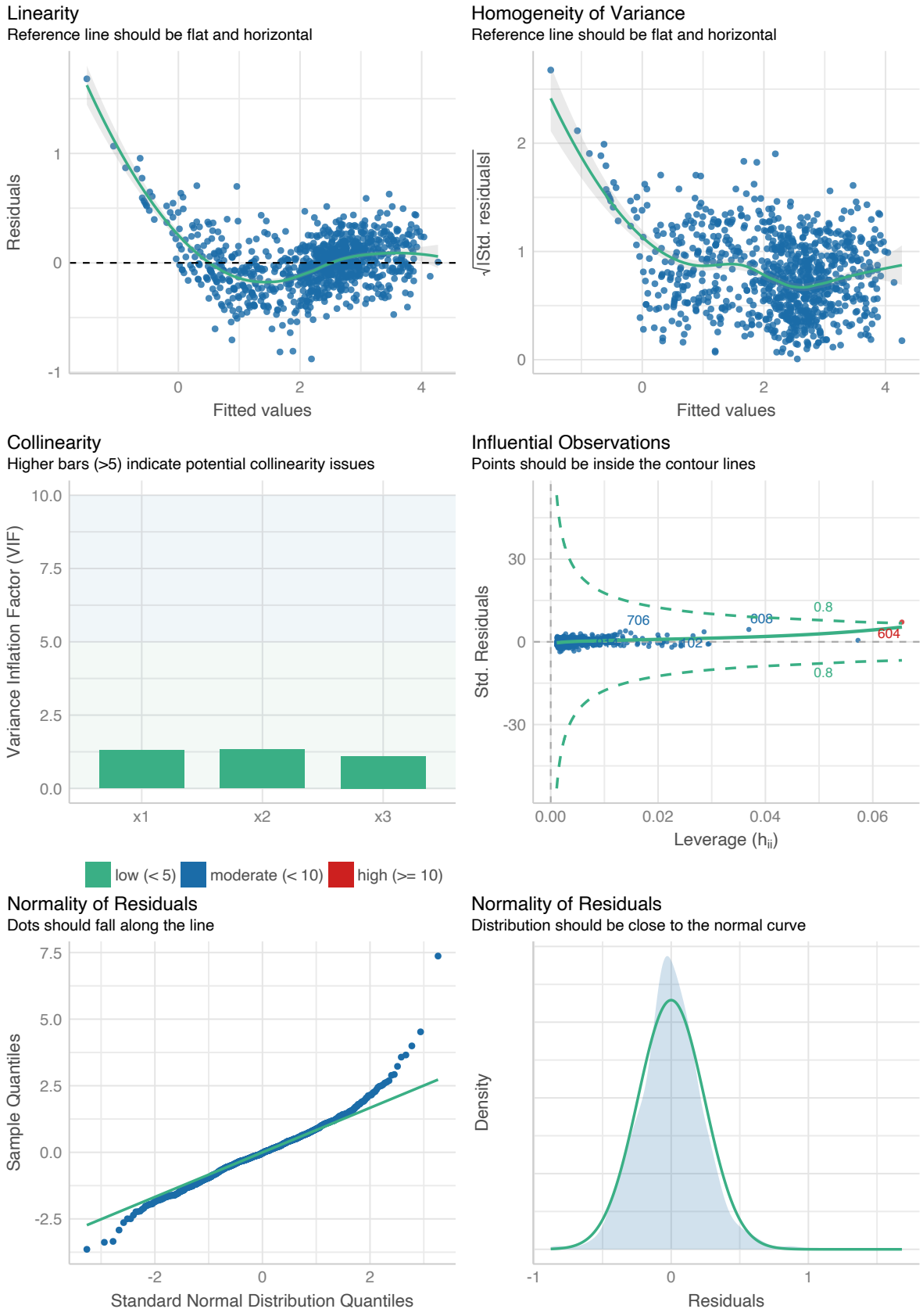


Figure 6.5: Diagnostics analysis of model LM (linear model), fitted to route 004.

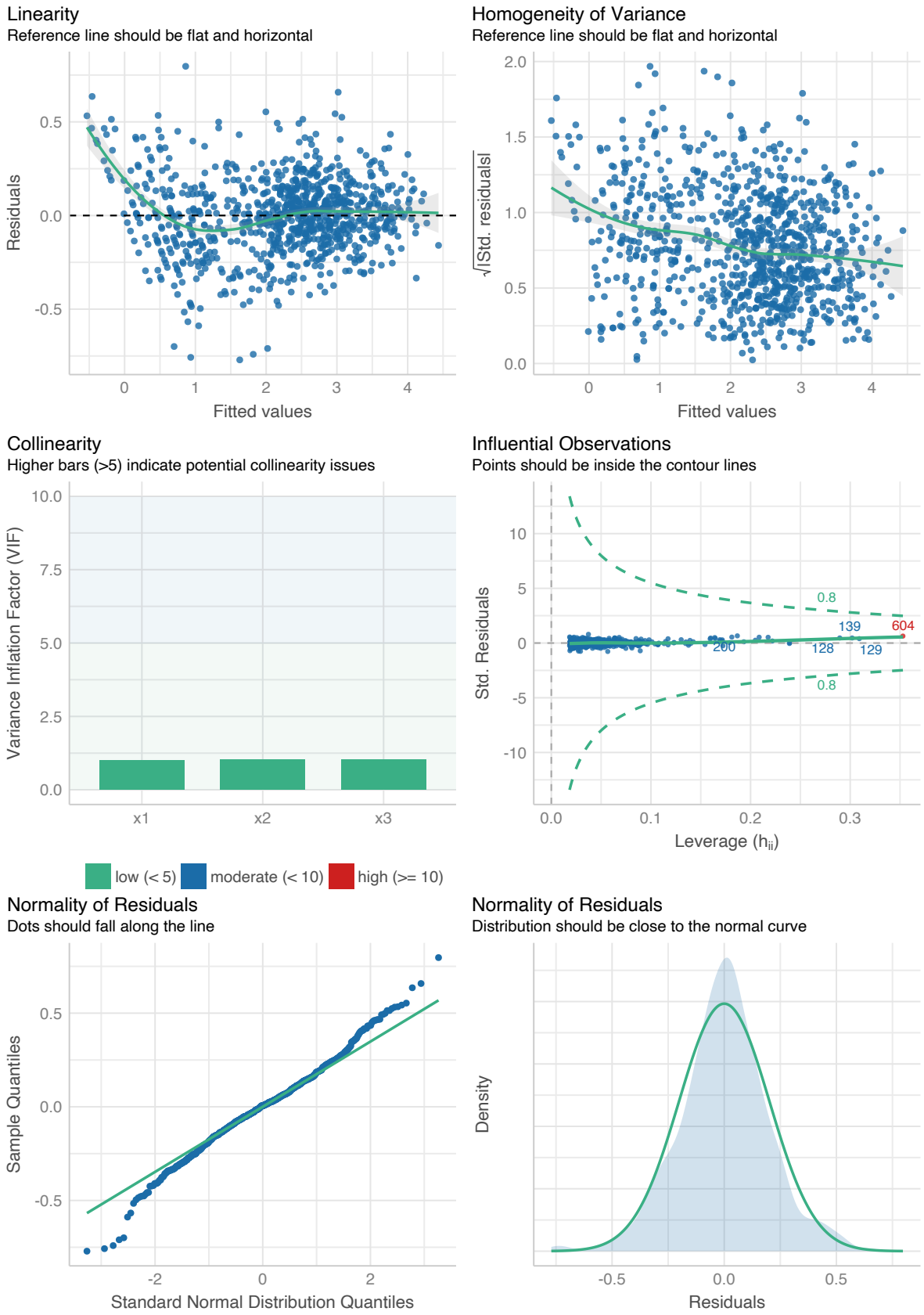


Figure 6.6: Diagnostics analysis of random-slopes-fixed-intercept model CELM, with parameter vector $[\beta_0, \beta_{1k}, \beta_{2k}, \beta_{3k}]$, fitted to route 004.

residuals, shown in Figures 6.5 and 6.6, respectively[1]. While the LM model does not result in a poor fit, it exhibits some features of non-linearity and non-normality of the residuals. In contrast, CELM model 4 leads to a visible improvement in the linearity and normality of residuals (top left, bottom left and right panels) and increase in the homogeneity of variance (top right panel), indicative of an overall meaningful improvement in model fit.

Table 6.4: Fixed effect parameter estimates and random effects (difference between fixed effect and random effect estimates) for the chosen CELM model (route 004).

effects	hour	β_0	β_1	β_2	β_3
fixed	-	5.243	1.495	-1.091	0.642
random	6		-0.044	0.026	0.002
	7		0.259	-0.037	0.192
	8		0.178	0.008	0.164
	9		0.081	0.012	0.100
	10		-0.192	-0.010	-0.199
	11		0.102	0.006	0.116
	12		-0.060	0.056	0.022
	13		0.030	0.036	0.080
	14		0.049	-0.020	-0.009
	15		0.145	0.020	0.166
	16		0.192	0.032	0.261
	17		0.087	0.056	0.122
	18		0.212	0.023	0.229
	19		-0.135	0.073	0.059
	20		-0.347	-0.016	-0.252
	21		-0.224	-0.239	-0.713
	22		-0.334	-0.026	-0.340

Table 6.4 shows parameter estimates for CELM model 4, decomposed into a fixed effects component, representing the overall effect of each variable on overtaking rate, and a random effects component, given as the difference between the fixed and hourly random effect estimates. The estimated fixed effect on overtaking rate is roughly of:

- 1.5% increase with a 1% increase in traffic density,
- 1.1% increase with a 1% decrease in traffic speed,
- 0.6% increase with a 1% increase in relative speed dispersion,

measured for each variable when all other variables are kept constant. This overall effect varies hourly and consistently in certain periods of the day, namely during the morning (7-8am) and afternoon peak periods (15-18pm). The added effect on overtaking rate during these periods, especially visible through increments to parameters β_1 and β_3, is expected since traffic volume is also maximal during these periods. As Figure 6.7 shows, there are

[1]In comparing Figures 6.5 and 6.6, note the re-scaling of the X and Y axes in each case (performed automatically by the graphics model checking engine used).

two or three roads that merge into route 004. During periods of increased traffic activity, more traffic is expected to merge from these roads onto route 004, causing vehicles driving on the left lane to be slowed down and more easily overtaken by vehicles driving on the right lane thus leading to increased overtaking behaviour, as is predicted by theoretical highway merge models (Laval & Daganzo, 2006).

Figure 6.7: Map of spatial route 004 (location 16 to 255).

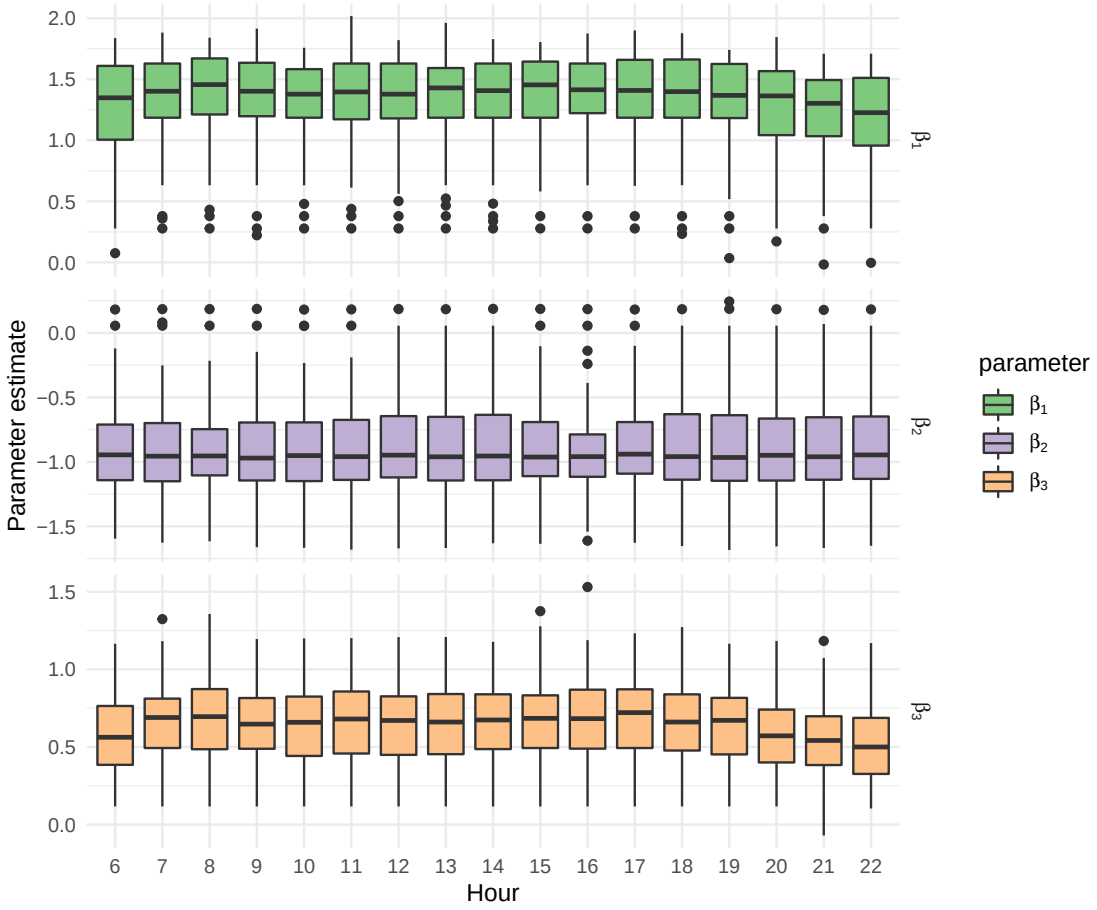


Figure 6.8: Parameter distribution across 50 multi-lane double-carriageway routes.

Lastly, it is of interest to examine how estimated parameter values vary across the whole sample of multi-lane routes. The same model choice process is applied to the subset of 50 routes categorised as multi-lane (lane $>$ 0.75, carriageway $>$ 0.50) from the initial

set of 127 routes. Figure 6.8 depicts the hourly distribution of parameter values across routes. Overall, the relationship between overtaking rate and traffic flow is consistent across routes, albeit with varying intensity (parameter estimates). Intensity variability is possibly due to route characteristics and unobserved traffic flow interactions with intersecting roads, as in the case of route `004`. In addition, hourly variation is consistent across routes as the median of parameter estimates β_1 and β_3 increases during the morning and afternoon peak periods and decreases during the night time period.

6.5 Discussion and future work

In this chapter, we propose a measure of overtaking rate that can be easily computed from existing ANPR infrastructure and data. We were able to investigate and quantify the relationship between overtaking rate and traffic flow variables on multi-lane routes. We found that overtaking rate increases multiplicatively with traffic density and relative speed dispersion, and inversely with mean speed. The observed relationship was valid for 50 routes, with the exception of two instances that showed a small positive increase with mean speed. The intensity of the relationship was seen to vary across routes and hours of the day, with some routes showing a steeper increase in overtaking rate with changing traffic conditions compared to others.

These results agree with the simulations of Navon (2003), who observed that overtaking rate[2] is inversely related to average speed and positively related with increased speed variability, especially at lower average speeds (i.e. increased relative speed variability). W. Jin & Li (2007) also found that vehicle ordering is more significantly violated during congested traffic due to increased numbers of overtaking manoeuvres.

The motivation to model overtaking rate stems from its link to accident risk, as increased lane-changing and overtaking behaviour have been observed to negatively impact road safety (Clarke et al., 1999; Mattes, 2003). However, as Navon (2003) demonstrates, increased numbers of lane-changing/overtaking manoeuvres do not necessarily translate to increased accident risk if the average dangerousness level of each manoeuvre decreases in the same proportion[3]. Nevertheless, the fact that speed dispersion and other traffic stream variables have been consistently linked to accident risk (Navon et al., 2019; M. Quddus, 2013; C. Xu et al., 2019), encourages further work into the use of overtaking rate as a possible proxy measure of accident risk. In that event, the existing ANPR

[2]Navon (2003) does not model overtaking rate directly but type c/d accident-prone interactions, which arise when vehicles travelling on the same road momentarily drive side-by-side (case c) or engage in lane changing manoeuvres (case d).

[3]A complete model of overtaking rate would seek to quantify not only the number of risky manoeuvres per unit of distance, but also the average dangerousness level of each manoeuvre. However, recording the latter variable is much harder in practice.

infrastructure can provide large-scale and automatic monitoring of road safety levels and help to identify low-safety routes, i.e. routes whose overtaking rate increases more rapidly with changing traffic conditions compared to other routes with similar characteristics.

Despite its potential, ANPR-measured overtaking rate has some limitations. As in any analysis of ANPR data, trip identification outliers may be present and contribute to severely inflate measurements of overtaking rate (a slow moving outlier vehicle will be frequently overtaken while a fast moving outlier vehicle will overtake many vehicles). Additionally, the proposed measure is only a lower-bound assessment of overtaking rate. Two vehicles can exchange several overtakings between them without actually recording any overtakings if they arrive in the same order as they departed (at best they will record one overtaking between them).

Two other aspects can contribute to noisy measurements. First, measurement can be unreliable if vehicles are not consistently detected at the same exact point on the road (recall in Section 3.3.1 that a camera cluster is composed of cameras pointing in the same direction within a 50 meter radius). Second, traffic flow and overtaking observations are absent if camera detection rates plummet under heavy traffic congestion. To mitigate this, one could consider applying one of the data imputation techniques recommended in Section 3.5.4, however, one would need to impute vehicle overtakings as well (which is less easy to do).

It is also noteworthy that some types of vehicles may be conferred a natural overtaking advantage on certain routes and contexts, for instance buses and taxis on bus lane fitted routes, and motorcycles under congested traffic. Such vehicles will naturally have inflated measurements of overtaking rate compared to other vehicles. Thus, accounting for vehicle type is a palpable way of improving the model and strengthen the results. Though, as discussed in Section 2.2.2.3, determining vehicle type either requires inference from the data (a task which is subject to errors) or obtaining access to the DVLA database.

Apart from addressing the limitations described above, future work should seek to extend the proposed statistical model to all monitored routes. In particular, routes can be considered together in a hierarchical model that takes into account the effect of route characteristics on overtaking rate. Not only should the already captured route characteristics be incorporated in the model, e.g. free flow speed and intersection count, but also spatio-temporal information from nearby intersecting roads. Finally, future work should seek to establish a link between overtaking rate and accident risk.

Chapter 7

Conclusions

Recently, large scale deployments of automatic vehicle identification systems have resulted in interconnected sensor networks with added potential for intelligent transportation systems, particularly traffic monitoring and control. To fully realise the potential of ANPR sensor networks, two issues must be addressed. The first challenge concerns the technical barriers that limit the technological adoption of ANPR, namely the pre-processing of raw data. The second challenge concerns the development of new applications and uses of large-scale ANPR systems that support traffic management in ways not possible in smaller and sparser networks.

This thesis addresses these two challenges by:

1. Identifying and proposing solutions for two technical barriers to ANPR adoption: the pre-processing of ANPR data, and the discovery of core travel sequences in a road network. These challenges are investigated and resolved in Chapters 3 and 4, respectively.
2. Developing two novel applications of ANPR sensor networks: the identification and impact assessment of traffic bottlenecks and the modelling of overtaking rate using traffic volume and speed. These applications are addressed in Chapters 5 and 6, and support traffic operators in the identification of congestion and road safety issues across the urban road network, regardless of road category.

7.1 Addressing technical barriers to ANPR adoption

Chapter 3 builds a data processing pipeline that facilitates the use of ANPR data for research purposes and supports the development of real traffic monitoring and control applications by:

- Presenting a complete, unified and modular guide to pre-processing ANPR data;

- Cataloguing the sources of data involved in each pre-processing stage;
- Describing the numerous tasks performed in each pre-processing stage, including details of necessary data treatments;
- Surveying existing techniques for each identified pre-processing task and identifying their shortcomings;
- Proposing and evaluating new methodologies for several of the pre-processing tasks: camera clustering, mapping camera locations to map locations, identification of outlier travel times and route identification;
- Benchmarking the proposed pipeline through summary statistics calculated for the Tyne and Wear UTMC dataset, namely the proportion of outliers by outlier type, filtering of malformed number plates and trip length frequency.

Chapter 4 develops a mathematical framework of corridor representation and discovery that allows future data analysis to better utilise sequential travel information in sensor networks and identify causal links between monitored routes. The contributions and benefits of this approach are:

- A formal representation of corridors as directed acyclical graphs that group trip sequences serving the same purpose;
- Use of trip frequency data and spatial information to discard travel sequences possibly contaminated by trip identification errors;
- A two-step corridor discovery algorithm, validated on an illustrative example, and applied to the Tyne and Wear sensor network to provide new insights into vehicle travel patterns in the region;
- Identification of key corridor sections (route sequences of length two or higher) in a road network according to observed traffic volume and degree of connectivity provided to the rest of the network.

The above advances made in data pre-processing and corridor discovery, coupled with the growing size and density of sensor networks, are a facilitator to the development of new traffic applications. The developed data pipeline enables high-quality aggregate/disaggregate traffic data to be produced from raw data sources, allowing it to be used directly in analysis. Corridor discovery allows popular route choices to be detected and easily incorporated in future analysis.

7.2 Novel applications

Chapter 5 describes the development and validation of an algorithm which enables bottlenecks to be identified across any (monitored) road in the network regardless of category,

thus extending traditional bottleneck analysis beyond highways. Furthermore, long-term bottleneck monitoring allows these to be categorised based on frequency of occurrence with *high-recurring* bottlenecks occurring on over two thirds of weekdays. Applied to the road network of Tyne and Wear, the bottleneck assessment procedure reveals that:

- Roads affected by high recurring bottlenecks account for nearly two thirds of 5.2 thousand vehicle-hours – the total average daily vehicle delay registered in the region – despite corresponding to only 25% of the total road length;
- Roads affected by high recurring bottlenecks experience, on average, three to four times more congestion during the day, and generally of greater intensity than the congestion experienced by other segments;
- Bottlenecks can be ranked according to their impact (e.g. total/average delay caused), providing a means with which to prioritise congestion mitigation schemes.

Lastly, Chapter 6 investigates the relationship between overtaking rate and three traffic stream variables – traffic density, mean speed and relative speed dispersion (quantified using the coefficient of variation). Overtaking rate acts as a surrogate measure of lane-changing frequency, which in high numbers is known to negatively impact road safety and cause shockwaves in traffic flow. Exploratory data analysis suggests a clear relationship between overtaking rate and traffic variables in multi-lane routes, consistent with previous findings in the Literature. This effect is formally quantified by a statistical model, with the following key findings:

- Overtaking rate increases multiplicatively with traffic density and relative speed dispersion, and inversely with mean speed;
- The relationship is consistent across a sample of 50 multi-lane routes (with the exception of two routes);
- The intensity of the relationship (slope) varies hourly and is stronger (steeper) when traffic volume is at a maximum, i.e. during the morning and afternoon peak hours;
- The intensity of the relationship also varies across routes, with some routes showing a stronger increase in overtaking rate compared to others.

Based on these preliminary findings, we conclude that ANPR is well suited to measuring the occurrence of vehicle overtakings. The existing ANPR infrastructure can provide large-scale and automatic monitoring of road safety levels and help to identify low-safety routes, i.e. routes whose overtaking rate increases more rapidly with changing traffic conditions compared to other routes with similar characteristics.

7.3 Overall outcomes and impact

Altogether, this thesis advances analytical methods of knowledge discovery in automatic vehicle identification sensor networks. Its main goal is to make ANPR methodology more accessible to different stakeholders and demonstrate the use cases and value of ANPR networks for traffic management and control. This is particularly relevant to regions where ANPR data may already be passively collected at scale, namely for law-enforcement purposes, but not actively used by traffic authorities due to limited understanding of how the technology can be employed to their benefit.

We achieve our goal of making ANPR research and analysis more approachable by building and benchmarking a data processing pipeline for ANPR data. This significantly lowers the entry barrier for new stakeholders and helps to systematize existing approaches to ANPR processing. Similarly, we developed a precise definition and implementation of road corridors, often the focus of analysis and improvement by traffic planners due to their importance in connecting different parts of the road network, which enables a range of new analyses about the state and evolution of the road network, such as the distribution of sensors and the temporal properties of user travel patterns.

We also demonstrate how the scope and use cases of ANPR networks can extend beyond traffic forecasting, adding to their value for traffic management and control. Specifically, we show how ANPR networks can give insight into congestion and overtaking patterns in urban networks, allowing authorities to prioritise interventions in more impacted areas of the network. By increasing the number of feasible applications of ANPR networks, cities are incentivised to invest in creating and growing ANPR networks as a crucial backbone to their smart cities sensing infrastructure and as a means to realise the vision of efficient transportation networks using data-driven intelligent transportation systems.

A major benefit of using our approach is the ability to simultaneously include in one's analysis information about individual vehicles, as well as aggregate information that reflects the conditions of traffic flow over time. Other sources of traffic data are often unable to provide these two levels of information simultaneously (as discussed in Section 2.1.2). This advantageous aspect is reflected in our investigation of bottlenecks. Although bottleneck analysis is done at the aggregate level, it requires knowledge of road corridors, which in turn requires knowledge of vehicle journeys. Similarly, vehicle overtakings are computed using individualised data but the relationship between overtaking rate and traffic conditions is studied at the aggregate level. In addition to the ability to use multiple levels of information, this work also opens up new avenues of research where a whole history of vehicle behaviour, in the form of multiple recorded journeys of the same vehicle, is available for consideration and analysis.

Finally, a vital outcome of this work is the generalisability of the methods here devel-

oped. Because these are not specific to any city or sensor network, we anticipate their applicability to other cities and countries, particularly those where the technology has already been deployed at scale. Overall, by using our pipeline for data processing and new methodologies of data analysis, stakeholders can extract added value from their ANPR sensing infrastructure and intelligent traffic management systems.

7.4 Future work

This thesis establishes a strong foundation that future work can build on. For each research chapter, we identify the key areas of future research.

Data processing pipeline. For many traffic monitoring applications, the analysis of ANPR data must be performed online. Even though many of the identified processing steps apply to both offline and online analysis, offline analysis can make use of past and future data while online can only make use of past data. This methodological difference will limit the type of algorithms applicable to each type of analysis and possibly influence the sequence of steps required to produce high-quality data in either case. Future work would seek to determine the exact differences in approach type and implement solutions particular to each case.

Another contribution to making the pipeline significantly more usable and widespread would be to implement it as an open-source software library. This would allow the pipeline code to be easily distributed and shared for research purposes, as well as adapted and extended to meet specific application goals. Such library could furthermore help to build a community of researchers and practitioners around the topic. A final key contribution would be to extend pipeline benchmarking to other ANPR network instances.

Corridor discovery and analysis. Chapter 4 discusses the difficulty in highlighting and comparing corridor sections of different lengths, and producing a summary set of corridor structures (both visually and computationally) with minimal overlap. Tackling this challenge, for example by adopting a trip frequency parameter or normalising frequencies by trip length, would greatly enhance researcher experience and reduce the time to produce a viable set of corridor outputs - making the method more attractive to stakeholders.

In addition to tackling corridor overlap, future work can seek to group corridors by road function (e.g. SRN, MRN, A-road) and measure the variation of road connectivity within and between-groups, so that over/under-performing corridors can be identified and their role in the network, and corresponding funding, can be adequately re-evaluated and re-prioritised. One other avenue of future research, is to study in more detail how changes in input parameter values, and the structure of the input network, translate to changes in the corresponding corridor set.

Bottleneck identification and impact assessment. To conduct bottleneck detection, one must choose values for the input parameters. Although the choice of parameter values was supported by exploratory analysis and reference values from other studies, results indicate that a range of parameter values can be considered. Consequently, a sensitivity analysis is required to better investigate the effect of parameter value choice on bottleneck activation and impact metrics, and determine the region of the parameter space where robust results can be consistently generated.

In addition to sensitivity analysis, we recommend an evaluation of the traffic volume counts reported by ANPR data. The main motivation being a possible bias towards routes with fewer junctions in between the origin and destination points and therefore less traffic merge/exit points which can affect the reported volume counts. Moreover, we recommend a review of how of missing data is treated. A strategy of treatment by omission was adopted in this work but alternative treatments based on data imputation should be considered to avoid a bias against route segments prone to missing data due to reduced detection rates under heavy congestion (which, consequently, may lead to serious bottlenecks passing undetected).

Modelling overtaking rate. Our modelling results indicate that overtaking rate evolves differently across routes, likely due to distinct characteristics. The next research step is hence to model routes jointly and quantify the effect of route characteristics on overtaking rate. Traffic related features, such as free flow speed of the route and traffic flow conditions in nearby/intersecting roads, as well as spatial features of the route, should be incorporated in the model for this purpose.

The extension of our model should be followed by a close examination of the link between overtaking rate and accident rate. In particular, it is important to understand the relationship between the quantity and quality of overtakings, as overtakings or lane-changing can vary substantially in respect to their dangerousness of the manoeuvre. Lastly, is important to evaluate more closely the limitations of ANPR-measure overtakings, namely the effect of trip identification outliers, the lower-bound nature of the proposed metric and the effect of camera clusters, heavy congestion vehicle type, like buses and motorcycles, on measurement.

Bibliography

Ahn, S., & Cassidy, M. J. (2007). Freeway traffic oscillations and vehicle lane-change maneuvers. *Transportation and Traffic Theory*.

Alexiadis, V., Colyar, J., Halkias, J., Hranac, R., & McHale, G. (2004). The next generation simulation program. *Institute of Transportation Engineers. ITE Journal*, *74*(8), 22.

Anagnostopoulos, C. E., Anagnostopoulos, I. E., Psoroulas, I. D., Loumos, V., & Kayafas, E. (2008). License Plate Recognition From Still Images and Video Sequences: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *9*(3), 377–391. https://doi.org/10.1109/TITS.2008.922938

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., R. Lyon, A. Ogale, L. Vincent, & J. Weaver. (2010). Google Street View: Capturing the World at Street Level. *Computer*, *43*(6), 32–38. https://doi.org/10.1109/MC.2010.170

Antoniou, C., Balakrishna, R., & Koutsopoulos, H. N. (2011). A Synthesis of emerging data collection technologies and their impact on traffic management applications. *European Transport Research Review*, *3*(3), 139–148. https://doi.org/10.1007/s12544-011-0058-1

Añez, J., De La Barra, T., & Pérez, B. (1996). Dual graph representation of transport networks. *Transportation Research Part B: Methodological*, *30*(3), 209–216. https://doi.org/10.1016/0191-2615(95)00024-0

Arth, C., Bischof, H., & Leistner, C. (2006). TRICam - An Embedded Platform for Remote Traffic Surveillance. *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 125–125. https://doi.org/10.1109/CVPRW.2006.208

Barceló, J., & others. (2010). *Fundamentals of traffic simulation* (Vol. 145). Springer.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PLOS ONE*, *12*(8), e0180698. https://doi.org/10.1371/

journal.pone.0180698

Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis: A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, *18*(6), 877–895. https://doi.org/10.1111/tgis.12073

Barroso, J. M. F., Albuquerque-Oliveira, J. L., & Oliveira-Neto, F. M. (2020). Correlation analysis of day-to-day origin-destination flows and traffic volumes in urban networks. *Journal of Transport Geography*, *89*, 102899. https://doi.org/10.1016/j.jtrangeo.2020.102899

Bauer, D., & Tulic, M. (2018). Travel time predictions: Should one model speeds or travel times? *European Transport Research Review*, *10*(2), 46. https://doi.org/10.1186/s12544-018-0315-7

Berkowicz, R., Winther, M., & Ketzel, M. (2006). Traffic pollution modelling and emission data. *Urban Air Quality Modelling*, *21*(4), 454–460. https://doi.org/10.1016/j.envsoft.2004.06.013

Bertini, R. L., & Myton, A. M. (2005). Use of Performance Measurement System Data to Diagnose Freeway Bottleneck Locations Empirically in Orange County, California. *Transportation Research Record*, *1925*(1), 48–57. https://doi.org/10.1177/0361198105192500106

Blythe, P. (1999). RFID for road tolling, road-use pricing and vehicle access control. *IEE Colloquium on RFID Technology (Ref. No. 1999/123)*, 8/1–816. https://doi.org/10.1049/ic:19990681

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, *65*, 126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

Bonneau, J. (2012). The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. *2012 IEEE Symposium on Security and Privacy*, 538–552. https://doi.org/10.1109/SP.2012.49

Booth, B., Mitchell, A., & others. (2001). *Getting started with ArcGIS*. Esri New York.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252. http://www.jstor.org/stable/2984418

Bradley, D. (2013). *Regular expression to validate UK number plate*. https://gist.github.com/danielrbradley/7567269

Bramberger, M., Brunner, J., Rinner, B., & Schwabach, H. (2004). Real-time video

analysis on an embedded smart camera for traffic surveillance. *Proceedings. RTAS 2004. 10th IEEE Real-Time and Embedded Technology and Applications Symposium, 2004.*, 174–181. https://doi.org/10.1109/RTTAS.2004.1317262

Cao, P., Hu, Y., Miwa, T., Wakita, Y., Morikawa, T., & Liu, X. (2017). An optimal mandatory lane change decision model for autonomous vehicles in urban arterials. *Journal of Intelligent Transportation Systems*, *21*(4), 271–284. https://doi.org/10.1080/15472450.2017.1315805

Cao, Q., Ren, G., Li, D., Ma, J., & Li, H. (2020). Semi-supervised route choice modeling with sparse Automatic vehicle identification data. *Transportation Research Part C: Emerging Technologies*, *121*, 102857. https://doi.org/10.1016/j.trc.2020.102857

Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis*, *33*(3), 249–258. https://doi.org/10.1016/S0167-9473(99)00057-2

Carrese, S., Cipriani, E., Mannini, L., & Nigro, M. (2017). Dynamic demand estimation and prediction for traffic urban networks adopting new data sources. *Transportation Research Part C: Emerging Technologies*, *81*, 83–98. https://doi.org/10.1016/j.trc.2017.05.013

Carrion, C., & Levinson, D. (2012). Value of travel time reliability: A review of current evidence. *Transportation Research Part A: Policy and Practice*, *46*(4), 720–741. https://doi.org/10.1016/j.tra.2012.01.003

Casella, G. (1990). *Statistical inference* (R. L. Berger, Ed.). Pacific Grove, Calif. : Brooks/Cole Pub. Co.

Cassidy, M. J., & Rudjanakanoknad, J. (2005). Increasing the capacity of an isolated merge by metering its on-ramp. *Transportation Research Part B: Methodological*, *39*(10), 896–913. https://doi.org/10.1016/j.trb.2004.12.001

Cassidy, M. J., & Windover, J. R. (1995). Methodology for Assessing Dynamics of Freeway Traffic Flow. *Transportation Research Record: Journal of the Transportation Research Board*, *1484*.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3). https://doi.org/10.1145/1541880.1541882

Chen, C., Skabardonis, A., & Varaiya, P. (2004). Systematic Identification of Freeway Bottlenecks. *Transportation Research Record*, *1867*(1), 46–52. https://doi.org/10.3141/1867-06

Cheng, T., Haworth, J., & Wang, J. (2012). Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems*, *14*(4), 389–413. https://doi.org/10.1007/s10109-011-0149-5

Cherchi, E., & Cirillo, C. (2014). Understanding variability, habit and the effect of long period activity plan in modal choices: A day to day, week to week analysis on panel data. *Transportation*, *41*(6), 1245–1262. https://doi.org/10.1007/s11116-014-9549-y

Cherchi, E., Cirillo, C., & Ortúzar, J. de D. (2017). Modelling correlation patterns in mode choice models estimated on multiday travel data. *Transportation Research Part A: Policy and Practice*, *96*, 146–153. https://doi.org/10.1016/j.tra.2016.11.021

Chikaraishi, M., Fujiwara, A., Zhang, J., & Axhausen, K. W. (2009). Exploring Variation Properties of Departure Time Choice Behavior by Using Multilevel Analysis Approach. *Transportation Research Record*, *2134*(1), 10–20. https://doi.org/10.3141/2134-02

Chiu, Y.-C., Bottom, J., Mahut, M., Paz, A., Balakrishna, R., Waller, T., & Hicks, J. (2011). Dynamic traffic assignment: A primer. *Transportation Research Circular, Transportation Research Board, United States of America*, *E-C153*. http://onlinepubs.trb.org/onlinepubs/circulars/ec153.pdf

Chorus, C. G., Molin, E. J. E., & Van Wee, B. (2006). Use and Effects of Advanced Traveller Information Services (ATIS): A Review of the Literature. *Transport Reviews*, *26*(2), 127–149. https://doi.org/10.1080/01441640500333677

Chow, A. H. F., Santacreu, A., Tsapakis, I., Tanasaranond, G., & Cheng, T. (2014). Empirical assessment of urban traffic congestion. *Journal of Advanced Transportation*, *48*(8), 1000–1016. https://doi.org/10.1002/atr.1241

Clark, S. D., Grant-Muller, S., & Chen, H. (2002). Cleaning of Matched License Plate Data. *Transportation Research Record*, *1804*(1), 1–7. https://doi.org/10.3141/1804-01

Clark, W. A. V., & Avery, K. L. (1976). The Effects of Data Aggregation in Statistical Analysis. *Geographical Analysis*, *8*(4), 428–438. https://doi.org/10.1111/j.1538-4632.1976.tb00549.x

Clarke, D. D., Ward, P. J., & Jones, J. (1999). Processes and countermeasures in overtaking road accidents. *Ergonomics*, *42*(6), 846–867. https://doi.org/10.1080/001401399185333

Coifman, B., & Li, L. (2017). A critical evaluation of the next generation simulation (NGSIM) vehicle trajectory dataset. *Transportation Research Part B: Methodological*, *105*, 362–377. https://doi.org/10.1016/j.trb.2017.09.018

Crawford, F., Watling, D. P., & Connors, R. D. (2018). Identifying road user classes based on repeated trip behaviour using Bluetooth data. *Transportation Research Part A: Policy and Practice*, *113*, 55–74. https://doi.org/10.1016/j.tra.2018.03.027

Csardi, G., & Nepusz, T. (2005). The Igraph Software Package for Complex Network Research. *InterJournal*, *Complex Systems*, 1695.

Daganzo, C. F. (1995). The cell transmission model, part II: Network traffic. *Transportation Research Part B: Methodological*, *29*(2), 79–93. https://doi.org/10.1016/0191-2615(94)00022-r

Daganzo, C. F. (1997). *Fundamentals of transportation and traffic operations*. Emerald Group Publishing Limited. https://doi.org/10.1108/9780585475301

Davis, G. A. (1997). Accuracy of Estimates of Mean Daily Traffic: A Review. *Transportation Research Record*, *1593*(1), 12–16. https://doi.org/10.3141/1593-02

Day, C. M., Brennan, T. M., Hainen, A. M., Remias, S. M., & Bullock, D. M. (2012). *Roadway System Assessment Using Bluetooth-Based Automatic Vehicle Identification Travel Time Data*. Purdue University. https://doi.org/10.5703/1288284314988

de Berg, M., Cheong, O., van Kreveld, M. J., & Overmars, M. H. (2008). *Computational geometry: Algorithms and applications, 3rd Edition*. Springer.

De Vos, J. (2020). The effect of COVID-19 and subsequent social distancing on travel behavior. *Transportation Research Interdisciplinary Perspectives*, *5*, 100121. https://doi.org/10.1016/j.trip.2020.100121

Debnath, A. K., Chin, H. C., Haque, Md. M., & Yuen, B. (2014). A methodological framework for benchmarking smart transport cities. *Cities*, *37*, 47–56. https://doi.org/10.1016/j.cities.2013.11.004

Demšar, U., Špatenková, O., & Virrantaus, K. (2008). Identifying Critical Locations in a Spatial Network with Graph Theory. *Transactions in GIS*, *12*(1), 61–82. https://doi.org/10.1111/j.1467-9671.2008.01086.x

Department for Transport. (2021). *Travel time measures for the Strategic Road Network and Local 'A' roads, England 2020*. Department for Transport. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/964503/travel-time-measures-srn-local-a-roads-2020.pdf

Department for Transport. (2017). *Proposals for the creation of a major road network*. Department for Transport. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/670527/major-road-network-consultation.pdf

Department for Transport. (2019a). *Road traffic estimates in Great Britain: 2019 report*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/916749/road-traffic-estimates-in-great-britain-2019.pdf

Department for Transport. (2019b). *Road traffic statistics - Manual count point: 27817*. https://roadtraffic.dft.gov.uk/manualcountpoints/27817

Department for Transport. (2019c). *Road traffic statistics - Manual count point: 47820*. https://roadtraffic.dft.gov.uk/manualcountpoints/47820

Diestel, R. (2017). *Graph theory* (Fifth Edition).

Dimitrakopoulos, G., & Demestichas, P. (2010). Intelligent Transportation Systems. *IEEE Vehicular Technology Magazine*, *5*(1), 77–84. https://doi.org/10.1109/MVT.2009.935537

Dion, F., & Rakha, H. (2006). Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. *Transportation Research Part B: Methodological*, *40*(9), 745–766. https://doi.org/10.1016/j.trb.2005.10.002

Dreyfus, S. E. (1969). An Appraisal of Some Shortest-Path Algorithms. *Operations Research*, *17*(3), 395–412. https://doi.org/10.1287/opre.17.3.395

Du, S., Ibrahim, M., Shehata, M., & Badawy, W. (2013). Automatic license plate recognition (ALPR): A state-of-the-art review. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(2), 311–325. https://doi.org/10.1109/TCSVT.2012.2203741

Duan, Y., Lv, Y., Liu, Y.-L., & Wang, F.-Y. (2016). An efficient realization of deep learning for traffic data imputation. *Transportation Research Part C: Emerging Technologies*, *72*, 168–181. https://doi.org/10.1016/j.trc.2016.09.015

Eisenstat, D. (2011). Random road networks: The quadtree model. In *2011 Proceedings of the Workshop on Analytic Algorithmics and Combinatorics (ANALCO)* (pp. 76–84). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611973013.9

Elkosantini, S., & Darmoul, S. (2013). Intelligent Public Transportation Systems: A review of architectures and enabling technologies. *2013 International Conference on Advanced Logistics and Transport*, 233–238. https://doi.org/10.1109/ICAdLT.2013.6568465

Eppell, V., McClurg, B., & Bunker, J. (2001). A four level road hierarchy for network planning and management. *Proceedings of the 20th ARRB Conference*, 1–7. https://core.ac.uk/download/pdf/10874278.pdf

Ermagun, A., & Levinson, D. (2018). Spatiotemporal traffic forecasting: Review and proposed directions. *Transport Reviews*, *38*(6), 786–814. https://doi.org/10.1080/01441647.2018.1442887

Falcocchio, J. C., & Levinson, H. S. (2015). Bottlenecks. In J. C. Falcocchio & H. S. Levinson (Eds.), *Road Traffic Congestion: A Concise Guide* (pp. 71–91). Springer International Publishing. https://doi.org/10.1007/978-3-319-15165-6_7

Friedrich, M., Jehlicka, P., & Schlaich, J. (2008). Automatic number plate recognition for

the observance of travel behavior. *Proc., The 8th International Conference on Survey Methods in Transport*.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some Implementations of the Boxplot. *The American Statistician*, *43*(1), 50–54. https://doi.org/10.1080/00031305.1989.10475612

Fu, M., Kelly, J. A., & Clinch, J. P. (2017). Estimating annual average daily traffic and transport emissions for a national road network: A bottom-up methodology for both nationally-aggregated and spatially-disaggregated results. *Journal of Transport Geography*, *58*, 186–195. https://doi.org/10.1016/j.jtrangeo.2016.12.002

Gao, S., Frejinger, E., & Ben-Akiva, M. (2010). Adaptive route choices in risky traffic networks: A prospect theory approach. *Applications of Advanced Technologies in Transportation: Selected Papers from the 10th AATT Conference*, *18*(5), 727–740. https://doi.org/10.1016/j.trc.2009.08.001

Gelman, A. (2004). Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics*, *13*(4), 755–779. https://doi.org/10.1198/106186004X11435

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/cbo9780511790942

Gentleman, R., & Temple Lang, D. (2007). Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, *16*(1), 1–23. https://doi.org/10.1198/106186007X178663

Gipps, P. G. (1986). A model for the structure of lane-changing decisions. *Transportation Research Part B: Methodological*, *20*(5), 403–414. https://doi.org/10.1016/0191-2615(86)90012-3

Gong, L., & Fan, W. (David). (2017). Applying Travel-Time Reliability Measures in Identifying and Ranking Recurrent Freeway Bottlenecks at the Network Level. *Journal of Transportation Engineering, Part A: Systems*, *143*(8), 04017042. https://doi.org/10.1061/JTEPBS.0000072

Google Maps Platform: Traffic, Transit, and Bicycling Layers. (n.d.). Retrieved January 17, 2023, from https://developers.google.com/maps/documentation/javascript/trafficlayer

Green, A. S., Wickham, L., Macgarty, M., & Green, S. (2019). *A vision for the governance of the major road network* (February; Vol. v2). WSP, Rees Jeffreys Road Fund. https://www.reesjeffreys.co.uk/wp-content/uploads/2019/02/WSPv2.pdf

Greenshields, B. d., Bibbins, J. r., Channing, W. s., & Miller, H. h. (1935). A study of traffic capacity. *Highway Research Board Proceedings*, *1935*. http://dx.doi.org/

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3). https://doi.org/10.18637/jss.v040.i03

Group, O. (2018). *The Open Group Base Specifications Issue 7, Rationale for Base Definitions, Section A.3 Definitions*. https://pubs.opengroup.org/onlinepubs/9699919799/xrat/V4_xbd_chap03.html

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, 1321–1330.

Guo, J., Williams, B. M., & Smith, B. L. (2007). Data Collection Time Intervals for Stochastic Short-Term Traffic Flow Forecasting. *Transportation Research Record*, *2024*(1), 18–26. https://doi.org/10.3141/2024-03

Gurney, R., Rhead, M., Lyons, V., & Ramalingam, S. (2013). The effect of ANPR camera settings on system performance. *5th International Conference on Imaging for Crime Detection and Prevention (ICDP 2013)*, 1–6. https://doi.org/10.1049/ic.2013.0276

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511574931

Hadavi, S., Rai, H. B., Verlinde, S., Huang, H., Macharis, C., & Guns, T. (2020). Analyzing passenger and freight vehicle movements from automatic-Number plate recognition camera data. *European Transport Research Review*, *12*(1), 37. https://doi.org/10.1186/s12544-020-00405-x

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX*. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. https://doi.org/10.1068/b35097

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/mprv.2008.80

Hale, D., Chrysikopoulos, G., Kondyli, A., & Ghiasi, A. (2021). Evaluation of data-driven performance measures for comparing and ranking traffic bottlenecks. *IET Intelligent Transport Systems*, *15*(4), 504–513. https://doi.org/10.1049/itr2.12040

Hale, D., Jagannathan, R., Xyntarakis, M., Su, P., Jiang, X., Ma, J., Hu, J., & Krause, C. (2016). *Traffic Bottlenecks: Identification and Solutions* (FHWA-HRT-16 -064). Federal Highway Administration, U.S Department of Transportation. https://rosap.ntl.bts.gov/view/dot/39938

Hall, F. L. (1996). Traffic stream characteristics. *Traffic Flow Theory. US Federal Highway Administration*, *36*. http://tft.eng.usf.edu/docs/chap2.pdf

Hamad, K. A., & Kaya, M. (2016). A detailed analysis of optical character recognition technology. *International Journal of Applied Mathematics, Electronics and Computers*, *4*(1), 244–249.

Hauslen, R. A. (1977). The promise of automatic vehicle identification. *IEEE Transactions on Vehicular Technology*, *26*(1), 30–38. https://doi.org/10.1109/T-VT.1977.23653

Haworth, J., & Cheng, T. (2012). Non-parametric regression for spacetime forecasting under missing data. *Special Issue: Advances in Geocomputation*, *36*(6), 538–550. https://doi.org/10.1016/j.compenvurbsys.2012.08.005

Hayat, S., Yanmaz, E., & Muzaffar, R. (Fourthquarter 2016). Survey on Unmanned Aerial Vehicle Networks for Civil Applications: A Communications Viewpoint. *IEEE Communications Surveys & Tutorials*, *18*(4), 2624–2661. https://doi.org/10.1109/COMST.2016.2560343

He, F., Yan, X., Liu, Y., & Ma, L. (2016). A Traffic Congestion Assessment Method for Urban Road Networks Based on Speed Performance Index. *Green Intelligent Transportation System and Safety*, *137*, 425–433. https://doi.org/10.1016/j.proeng.2016.01.277

Hegeman, G., Hoogendoorn, S., & Brookhuis, K. (2004). Observations overtaking manoeuvres on bi-directional roads. *Advanced OR and AI Methods in Transportation*, *1*, 505–510.

Hellinga, B., & Knapp, G. (2000). Automatic Vehicle Identification Technology-Based Freeway Incident Detection. *Transportation Research Record*, *1727*(1), 142–153. https://doi.org/10.3141/1727-18

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*, *abs/1610.02136*. http://arxiv.org/abs/1610.02136

HERE Traffic API V7. (n.d.). Retrieved January 17, 2023, from https://developer.here.com/documentation/traffic-api/dev_guide/index.html

Hofmann-Wellenhof, B., Legat, K., Lichtenegger, H., & Wieser, M. (2003). *Navigation*. Springer Vienna. https://books.google.co.uk/books?id=losWr9UDRasC

Home Office. (2020). *National ANPR Standards for Policing and Law Enforcement*. U.K. Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936912/NASPLE_Version_2.1_November_2020.pdf

Hsu, G., Chen, J., & Chung, Y. (2013). Application-Oriented License Plate Recognition. *IEEE Transactions on Vehicular Technology*, *62*(2), 552–561. https://doi.org/10.1109/TVT.2012.2226218

Hurdle, V. F., Merlo, M. I., & Robertson, D. (1997). Study of Speed-Flow Relationships on Individual Freeway Lanes. *Transportation Research Record*, *1591*(1), 7–13. https://doi.org/10.3141/1591-02

Järv, O., Ahas, R., & Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, *38*, 122–135. https://doi.org/10.1016/j.trc.2013.11.003

Jiang, S., Fiore, G. A., Yang, Y., Ferreira, J., Frazzoli, E., & González, M. C. (2013). A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. https://doi.org/10.1145/2505821.2505828

Jin, W.-L. (2010). A kinematic wave theory of lane-changing traffic flow. *Transportation Research Part B: Methodological*, *44*(8), 1001–1021. https://doi.org/10.1016/j.trb.2009.12.014

Jin, W., & Li, L. (2007). First-in-first-out is violated in real traffic. *Proceedings of Transportation Research Board Annual Meeting*.

Jones, P., & Clarke, M. (1988). The significance and measurement of variability in travel behaviour. *Transportation*, *15*(1), 65–87. https://doi.org/10.1007/BF00167981

Jula, H., Kosmatopoulos, E. B., & Ioannou, P. A. (Nov./2000). Collision avoidance analysis for lane changing and merging. *IEEE Transactions on Vehicular Technology*, *49*(6), 2295–2308. https://doi.org/10.1109/25.901899

Kamarulazizi, K., & Ismail, D. W. (2010). *Electronic Toll Collection System Using Passive RFID Technology*. *22*(2), 7.

Kazagli, E., Koutsopoulos, H. N., & others. (2013). Estimation of Arterial Travel Time from Automatic Number Plate Recognition Data. *Transportation Research Record*, *2391*(1), 22–31. https://doi.org/10.3141/2391-03

Kendall, M. G. (1945). The Treatment of Ties in Ranking Problems. *Biometrika*, *33*(3), 239–251. https://doi.org/10.1093/biomet/33.3.239

Kieu, L. M., Bhaskar, A., & Chung, E. (2015). Passenger Segmentation Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*, *16*(3), 1537–1548. https://doi.org/10.1109/TITS.2014.2368998

Kimber, A. C. (1990). Exploratory Data Analysis for Possibly Censored Data from Skewed

Distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *39*(1), 21–30. https://doi.org/10.2307/2347808

Kitchin, R. (2016). The ethics of smart cities and urban science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160115. https://doi.org/10.1098/rsta.2016.0115

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., Ramsey, D. J., & others. (2006). The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. *United States. National Highway Traffic Safety Administration*.

Klein, L., Mills, M., Gibson, D., & others. (2006). *Traffic detector handbook: Volume I*. Turner-Fairbank Highway Research Center. https://www.fhwa.dot.gov/publications/research/operations/its/06108/06108.pdf

Krishnakumari, P., van Lint, H., Djukic, T., & Cats, O. (2020). A data driven method for OD matrix estimation. *ISTTT 23 TR_C-23rd International Symposium on Transportation and Traffic Theory (ISTTT 23)*, *113*, 38–56. https://doi.org/10.1016/j.trc.2019.05.014

Lana, I., Del Ser, J., Velez, M., & Vlahogianni, E. I. (2018). Road Traffic Forecasting: Recent Advances and New Challenges. *IEEE Intelligent Transportation Systems Magazine*, *10*(2), 93–109. https://doi.org/10.1109/MITS.2018.2806634

Laval, J. A., & Daganzo, C. F. (2006). Lane-changing in traffic streams. *Transportation Research Part B: Methodological*, *40*(3), 251–264. https://doi.org/10.1016/j.trb.2005.04.003

Lee, S. E., Olsen, E. C., Wierwille, W. W., & others. (2004). *A comprehensive examination of naturalistic lane-changes*. United States. National Highway Traffic Safety Administration.

Lee, W., Tseng, S., Shieh, J., & Chen, H. (2011). Discovering Traffic Bottlenecks in an Urban Network by Spatiotemporal Data Mining on Location-Based Services. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1047–1056. https://doi.org/10.1109/TITS.2011.2144586

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Li, J., Zuylen, H. van, Deng, Y., & Zhou, Y. (2020). Urban travel time data cleaning and analysis for Automatic Number Plate Recognition. *22nd EURO Working Group on*

Transportation Meeting, EWGT 2019, 18th 20th September 2019, Barcelona, Spain, *47*, 712–719. https://doi.org/10.1016/j.trpro.2020.03.151

Li, L., Li, Y., & Li, Z. (2013). Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies*, *34*, 108–120. https://doi.org/10.1016/j.trc.2013.05.008

Li, M., Li, Z., Xu, C., & Liu, T. (2020). Short-term prediction of safety and operation impacts of lane changes in oscillations with empirical vehicle trajectories. *Accident Analysis & Prevention*, *135*, 105345. https://doi.org/10.1016/j.aap.2019.105345

Li, R. (2006). *Enhancing motorway travel time prediction models through explicit incorporation of travel time variability* [PhD thesis]. Monash University.

Li, R., & Rose, G. (2011). Incorporating uncertainty into short-term travel time predictions. *Transportation Research Part C: Emerging Technologies*, *19*(6), 1006–1018. https://doi.org/10.1016/j.trc.2011.05.014

Li, W., Yang, C., & Jabari, S. E. (2020). Short-Term Traffic Forecasting Using High-Resolution Traffic Data. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–6. https://doi.org/10.1109/ITSC45102.2020.9294706

Li, X., Lam, W. H. K., & Tam, M. L. (2013). New Automatic Incident Detection Algorithm Based on Traffic Data Collected for Journey Time Estimation. *Journal of Transportation Engineering*, *139*(8), 840–847. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000566

Li, Z.-C., Huang, H.-J., & Yang, H. (2020). Fifty years of the bottleneck model: A bibliometric review and future research directions. *Transportation Research Part B: Methodological*, *139*, 311–342. https://doi.org/10.1016/j.trb.2020.06.009

Liang, Z., & Wakahara, Y. (2014). Real-time urban traffic amount prediction models for dynamic route guidance systems. *EURASIP Journal on Wireless Communications and Networking*, *2014*(1), 85. https://doi.org/10.1186/1687-1499-2014-85

Lighthill, M. J., & Whitham, G. B. (1955). On kinematic waves II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *229*(1178), 317–345. https://doi.org/10.1098/rspa.1955.0089

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. https://doi.org/10.1002/9781119013563

Liu, Z., Sharma, S., & Datla, S. (2008). Imputation of Missing Traffic Data during Holiday Periods. *Transportation Planning and Technology*, *31*(5), 525–544. https://doi.org/10.1080/03081060802364505

Long, J., Gao, Z., Ren, H., & Lian, A. (2008). Urban traffic congestion propagation and bottleneck identification. *Science in China Series F: Information Sciences*, *51*(7), 948–964. https://doi.org/10.1007/s11432-008-0038-9

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic information science and systems*. John Wiley & Sons.

Luo, X., Ma, D., Jin, S., Gong, Y., & Wang, D. (2019). Queue length estimation for signalized intersections using license plate recognition data. *IEEE Intelligent Transportation Systems Magazine*, *11*(3), 209–220. https://doi.org/10.1109/MITS.2019.2919541

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F.-Y. (2015). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 865–873. https://doi.org/10.1109/TITS.2014.2345663

Ma, D., Luo, X., Li, W., Jin, S., Guo, W., & Wang, D. (2017). Traffic demand estimation for lane groups at signal-controlled intersections using travel times from video-imaging detectors. *IET Intelligent Transport Systems*, *11*(4), 222–229. https://doi.org/10.1049/iet-its.2016.0233

Major Road Network. (2020). {{Department for Transport}}. https://data.gov.uk/dataset/95f58bfa-13d6-4657-9d6f-020589498cfd/major-road-network

Marczak, F., Daamen, W., & Buisson, C. (2016). Empirical analysis of lane changing behavior at a freeway weaving section. In *Traffic management* (pp. 139–151). John Wiley & Sons, Ltd. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119307822.ch10

Mattes, S. (2003). The lane-change-task as a tool for driver distraction evaluation. *Quality of Work and Products in Enterprises of the Future*, *57*, 60.

McNally, M. G. (2007). The Four-Step Model. In D. A. Hensher & K. J. Button (Eds.), *Handbook of Transport Modelling* (Vol. 1, pp. 35–53). Emerald Group Publishing Limited. https://doi.org/10.1108/9780857245670-003

Meyer, M. D., & others. (2016). *Transportation planning handbook*. John Wiley & Sons.

Miller, H. J., Shaw, S.-L., & others. (2001). *Geographic information systems for transportation: Principles and applications*. Oxford University Press on Demand.

Mills, D. L. (1991). Internet time synchronization: The network time protocol. *IEEE Transactions on Communications*, *39*(10), 1482–1493. https://doi.org/10.1109/26.103043

Minnen, J., Glorieux, I., & van Tienoven, T. P. (2015). Transportation habits: Evidence from time diary data. *Emerging Data and Methodological Considerations in Time-Use Analysis*, *76*, 25–37. https://doi.org/10.1016/j.tra.2014.12.013

Morar, N., & Baber, C. (2017). Joint Human-Automation Decision Making in Road Traffic Management. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 385–389. https://doi.org/10.1177/1541931213601578

Mouskos, K. C., Niver, E., Pignataro, L. J., & Lee, S. (1998). *Transmit system evaluation. Final Report.* Institute for Transportation, New Jersey Institute of Technology. https://rosap.ntl.bts.gov/view/dot/3814/dot_3814_DS1.pdf

Navon, D. (2003). The paradox of driving speed: Two adverse effects on highway accident rate. *Accident Analysis & Prevention*, *35*(3), 361–367. https://doi.org/10.1016/S0001-4575(02)00011-8

Navon, D., Kasten, R., Pomerants, A., & Erev, I. (2019). A novel cost/benefit approach for reducing frequency of deviant driving speeds in expressways. *Transportation Research Part F: Traffic Psychology and Behaviour*, *62*, 855–869. https://doi.org/10.1016/j.trf.2019.03.009

Newcastle City Council. (2020). *Transport Improvements: Blue House Roundabout.* https://web.archive.org/web/20210607093306/https://www.newcastle.gov.uk/our-city/transport-improvements/outside-city-centre/blue-house-roundabout

Newell, G. F. (2002). A simplified car-following theory: A lower order model. *Transportation Research Part B: Methodological*, *36*(3), 195–205. https://doi.org/10.1016/S0191-2615(00)00044-8

Newell, G. F. (1993a). A simplified theory of kinematic waves in highway traffic, part I: General theory. *Transportation Research Part B: Methodological*, *27*(4), 281–287. https://doi.org/10.1016/0191-2615(93)90038-C

Newell, G. F. (1993b). A simplified theory of kinematic waves in highway traffic, part II: Queueing at freeway bottlenecks. *Transportation Research Part B: Methodological*, *27*(4), 289–303. https://doi.org/10.1016/0191-2615(93)90039-D

Ni, D., Leonard, J. D., Guin, A., & Feng, C. (2005). Multiple Imputation Scheme for Overcoming the Missing Values and Variability Issues in ITS Data. *Journal of Transportation Engineering*, *131*(12), 931–938. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:12(931)

Nicholson, T. A. J. (1966). Finding the Shortest Route between Two Points in a Network. *The Computer Journal*, *9*(3), 275–280. https://doi.org/10.1093/comjnl/9.3.275

Office for National Statistics. (2011). *Nomis Census 2011 Data: Method of travel to work.* https://www.nomisweb.co.uk/census/2011/qs701ew

Oh, C., Ritchie, S. G., & Oh, J.-S. (2005). Exploring the Relationship between Data Aggregation and Predictability to Provide Better Predictive Traf-

fic Information. *Transportation Research Record*, *1935*(1), 28–36. https://doi.org/10.1177/0361198105193500104

Oliver, P. (2010). *The student's guide to research ethics*. McGraw-Hill Education (UK).

Olsen, E. C. B., Lee, S. E., Wierwille, W. W., & Goodman, M. J. (2002). Analysis of distribution, frequency, and duration of naturalistic lane changes. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*(22), 1789–1793. https://doi.org/10.1177/154193120204602203

Omar, H. A., Abboud, K., Cheng, N., Malekshan, K. R., Gamage, A. T., & Zhuang, W. (2016). A Survey on High Efficiency Wireless Local Area Networks: Next Generation WiFi. *IEEE Communications Surveys & Tutorials*, *18*(4), 2315–2344. https://doi.org/10.1109/COMST.2016.2554098

OpenALPR dataset. (2014). https://github.com/openalpr/benchmarks

OpenStreetMap. (2021a). *OpenStreetMap Wiki: Key highway*. https://wiki.openstreetmap.org/wiki/Key:highway

OpenStreetMap. (2021b). *Who uses OpenStreetMap?* https://welcome.openstreetmap.org/about-osm-community/consumers/

Ordinance Survey Open Roads. (2020). https://www.ordnancesurvey.co.uk/business-government/products/open-map-roads

Ortúzar, J. de D., & Willumsen, L. G. (2011). *Modelling Transport*. https://doi.org/10.1002/9781119993308

Ozbay, S., & Ercelebi, E. (2005). Automatic vehicle identification by plate recognition. *World Academy of Science, Engineering and Technology*, *9*(41), 222–225. https://doi.org/10.5281/zenodo.1331865

Pande, A., & Abdel-Aty, M. (2006). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention*, *38*(5), 936–948. https://doi.org/10.1016/j.aap.2006.03.004

Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A., & Wang, Y. (2003). Review of road traffic control strategies. *Proceedings of the IEEE*, *91*(12), 2043–2067. https://doi.org/10.1109/JPROC.2003.819610

Papagianni, S. (2003). *Using automatic number recognition data to estimate vehicle speeds in central London* [PhD thesis]. MSc Dissertation, Centre for Transport Studies, Imperial College London.

Park, H., Yang, S., Oh, D., & Kim, J. (2009). Design and Implementation of WLAN-Based Automatic Vehicle Identification. *2009 International Conference on Computational Science and Engineering*, *2*, 310–317. https://doi.org/10.1109/CSE.2009.371

Patel, C., Shah, D., & Patel, A. (2013). Automatic Number Plate Recognition System (ANPR): A Survey. *International Journal of Computer Applications*, *69*(9), 21–33. https://doi.org/10.5120/11871-7665

Petty, K. F., Bickel, P., Ostland, M., Rice, J., Schoenberg, F., Jiang, J., & Ritov, Y. (1998). Accurate estimation of travel times from single-loop detectors. *Transportation Research Part A: Policy and Practice*, *32*(1), 1–17. https://doi.org/10.1016/S0965-8564(97)00015-3

Pinto da Silva, P., Forshaw, M., & McGough, S. (2018, January). Clustering vehicles based on trips identified from automatic number plate recognition camera scans. *1st International Workshop on Big Traffic Data Analytics*. https://core.ac.uk/download/pdf/327365678.pdf

Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *79*, 1–17. https://doi.org/10.1016/j.trc.2017.02.024

Porta, S., Crucitti, P., & Latora, V. (2006). The Network Analysis of Urban Streets: A Primal Approach. *Environment and Planning B: Planning and Design*, *33*(5), 705–725. https://doi.org/10.1068/b32045

Prato, C. G. (2009). Route choice modeling: Past, present and future research directions. *Journal of Choice Modelling*, *2*(1), 65–100. https://doi.org/10.1016/S1755-5345(13)70005-8

Punzo, V., Borzacchiello, M. T., & Ciuffo, B. (2011). On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data. *Transportation Research Part C: Emerging Technologies*, *19*(6), 1243–1262. https://doi.org/10.1016/j.trc.2010.12.007

QGIS: A Free and Open Source Geographic Information System. (n.d.). Retrieved January 17, 2023, from https://www.qgis.org/en/site/index.html

Qu, L., Li, L., Zhang, Y., & Hu, J. (2009). PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach. *IEEE Transactions on Intelligent Transportation Systems*, *10*(3), 512–522. https://doi.org/10.1109/TITS.2009.2026312

Quarmby, D., & Carey, P. (2016). A Major Road Network for England. In *A rees jeffreys road fund study* (October). Rees Jeffreys Road Fund. https://www.reesjeffreys.co.uk/wp-content/uploads/2016/10/A-Major-Road-Network-for-England-David-Quarmby-and-Phil-Carey-Rees-Jeffreys-Road-Fund-October-2016.pdf

Quddus, M. (2013). Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*, *5*(1), 27–45. https://doi.org/10.1080/19439962.2012.705232

Quddus, M. A., Ochieng, W. Y., & Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, *15*(5), 312–328. https://doi.org/10.1016/j.trc.2007.05.002

Raines, A., & Rowley, P. (2008). Coordinated traffic management through data exchange. *IET Conference Proceedings*, 35–35(1). https://digital-library.theiet.org/content/conferences/10.1049/ic.2008.0780

Rakha, H., El-Shawarby, I., & Arafeh, M. (2010). Trip Travel-Time Reliability: Issues and Proposed Solutions. *Journal of Intelligent Transportation Systems*, *14*(4), 232–250. https://doi.org/10.1080/15472450.2010.517477

Rakha, H., & Zhang, W. (2005). Estimating Traffic Stream Space Mean Speed and Reliability from Dual- and Single-Loop Detectors. *Transportation Research Record*, *1925*(1), 38–47. https://doi.org/10.1177/0361198105192500105

Rao, A. M., & Rao, K. R. (2012). Measuring urban traffic congestion-a review. *International Journal for Traffic & Transport Engineering*, *2*(4).

Rao, W., Wu, Y.-J., Xia, J., Ou, J., & Kluger, R. (2018). Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. *Transportation Research Part C: Emerging Technologies*, *95*, 29–46. https://doi.org/10.1016/j.trc.2018.07.002

Reutebuch, S. E., Andersen, H.-E., & McGaughey, R. J. (2005). Light Detection and Ranging (LIDAR): An Emerging Tool for Multiple Resource Inventory. *Journal of Forestry*, *103*(6), 286–292. https://doi.org/10.1093/jof/103.6.286

Rhead, M., Gurney, R., Ramalingam, S., & Cohen, N. (2012). Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems. *2012 IEEE International Carnahan Conference on Security Technology (ICCST)*, 286–291. https://doi.org/10.1109/CCST.2012.6393574

Roberts, M., Melecky, M., Bougna, T., & Xu, Y. (2020). Transport corridors and their wider economic benefits: A quantitative review of the literature. *Journal of Regional Science*, *60*(2), 207–248. https://doi.org/10.1111/jors.12467

Robinson, S., & Polak, J. (2006). Overtaking Rule Method for the Cleaning of Matched License-Plate Data. *Journal of Transportation Engineering*, *132*(8), 609–617. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:8(609)

Rockafellar, R. T., & Wets, R. J. B. (1998). Variational Analysis. *Grundlehren Der Mathematischen Wissenschaften*. https://doi.org/10.1007/978-3-642-02431-3

Rossana, R. J., & Seater, J. J. (1995). Temporal Aggregation and Economic Time Series.

Journal of Business & Economic Statistics, *13*(4), 441–451. https://doi.org/10.1080/07350015.1995.10524618

Sala, M., & Soriguera, F. (2020). Lane-changing and freeway capacity: A Bayesian inference stochastic model. *Computer-Aided Civil and Infrastructure Engineering*, *35*(7), 719–733. https://doi.org/10.1111/mice.12529

Schneider, T. (2001). Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. *JOURNAL OF CLIMATE*, *14*, 19.

Schwertman, N. C., Owens, M. A., & Adnan, R. (2004). A simple more general boxplot method for identifying outliers. *Computational Statistics & Data Analysis*, *47*(1), 165–174. https://doi.org/10.1016/j.csda.2003.10.012

Silva, S. M., & Jung, C. R. (2018, September). License plate detection and recognition in unconstrained scenarios. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In *New Methods in Cognitive Psychology* (pp. 4–31). Routledge. http://singmann.org/download/publications/singmann_kellen-introduction-mixed-models.pdf

Skabardonis, A., Varaiya, P., & Petty, K. F. (2003). Measuring Recurrent and Nonrecurrent Traffic Congestion. *Transportation Research Record: Journal of the Transportation Research Board*, *1856*(1), 118–124. https://doi.org/10.3141/1856-12

Smith, B. L., Scherer, W. T., & Conklin, J. H. (2003). Exploring Imputation Techniques for Missing Data in Transportation Management Systems. *Transportation Research Record: Journal of the Transportation Research Board*, *1836*(1), 132–142. https://doi.org/10.3141/1836-17

Spiller, N. C., Blizzard, K., Margiotta, R., & others. (2012). *Recurring traffic bottlenecks: A primer focus on low-cost operational improvements*. United States. Federal Highway Administration. Office of Operations.

Srivastava, A., & Geroliminis, N. (2013). Empirical observations of capacity drop in freeway merges with ramp control and integration in a first-order model. *Transportation Research Part C: Emerging Technologies*, *30*, 161–177. https://doi.org/10.1016/j.trc.2013.02.006

Stevanovic, A., Kergaye, C., & Martin, P. T. (2009). SCOOT and SCATS: A closer look into their operations. *88th Annual Meeting of the Transportation Research Board. Washington DC*.

Su, H., Zheng, K., Huang, J., Jeung, H., Chen, L., & Zhou, X. (2014). CrowdPlanner: A crowd-based route recommendation system. *2014 IEEE 30th International Conference on Data Engineering*, 1144–1155. https://doi.org/10.1109/ICDE.2014.6816730

Subramani, T., Kavitha, M., & Sivaraj, K. P. (2012). Modelling Of Traffic Noise Pollution. *International Journal of Engineering Research and Applications*, *2*(3), 8.

SWri. (1998). Automatic vehicle identification model deployment initiativesystem design document. *Report Prepared for Trans-Guide. Texas Department of Transportation, Southwest Research Institute, San Antonio*.

Systematics, C., Council, N. R., & others. (2013). *Analytical procedures for determining the impacts of reliability mitigation strategies*. Transportation Research Board. http://www.trb.org/Main/Blurbs/166935.aspx

Tam, M. L., & Lam, W. H. K. (2011). Application of automatic vehicle identification technology for real-time journey time estimation. *Special Issue on Intelligent Transportation Systems*, *12*(1), 11–19. https://doi.org/10.1016/j.inffus.2010.01.002

Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y.-J., & Li, F. (2013). A tensor-based method for missing traffic data completion. *Euro Transportation: Selected Paper from the EWGT Meeting, Padova, September 2009*, *28*, 15–27. https://doi.org/10.1016/j.trc.2012.12.007

Tarigan, A. K. M., & Kitamura, R. (2009). Week-to-Week Leisure Trip Frequency and Its Variability. *Transportation Research Record*, *2135*(1), 43–51. https://doi.org/10.3141/2135-06

Toledo, T., Koutsopoulos, H. N., & Ben-Akiva, M. E. (2003). Modeling integrated lane-changing behavior. *Transportation Research Record*, *1857*(1), 30–38. https://doi.org/10.3141/1857-04

TomTom Traffic. (n.d.). Retrieved January 17, 2023, from https://developer.tomtom.com/products/traffic-api

Transportation Research Board. (2000). *Highway Capacity Manual*.

TranStar, H. (2001). *TranStar description. 14*. http://traffic.houstontranstar.org/layers/

Treiber, M., & Helbing, D. (2002). *Reconstructing the Spatio-Temporal Traffic Dynamics from Stationary Detector Data*. 24.

Treiber, M., & Kesting, A. (2013). *Traffic Flow Dynamics*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32460-4

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass. : Addison-Wesley Pub. Co.

van Lint, J. W. C., Hoogendoorn, S. P., & van Zuylen, H. J. (2005). Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C: Emerging Technologies*, *13*(5), 347–369. https://doi.org/10.1016/j.trc.2005.03.001

Vickrey, W. S. (1969). Congestion Theory and Transport Investment. *The American Economic Review*, *59*(2), 251–260. http://www.jstor.org/stable/1823678

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Special Issue on Short-Term Traffic Flow Forecasting*, *43*, 3–19. https://doi.org/10.1016/j.trc.2014.01.005

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2006). Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume. *Transportation Research Part C: Emerging Technologies*, *14*(5), 351–367. https://doi.org/10.1016/j.trc.2006.09.002

Vlahogianni, E., & Karlaftis, M. (2011). Temporal aggregation in traffic data: Implications for statistical characteristics and model choice. *Transportation Letters*, *3*(1), 37–49. https://doi.org/10.3328/TL.2011.03.01.37-49

Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, *7*, 107964–108000. https://doi.org/10.1109/ACCESS.2019.2932769

Wang, S., Cao, J., & Yu, P. S. (2022). Deep Learning for Spatio-Temporal Data Mining: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 3681–3700. https://doi.org/10.1109/TKDE.2020.3025580

Wardrop, J. G. (1952). Road paper. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, *1*(3), 325–362.

Watson, B. (2017). Intelligent infrastructure: Automatic number plate recognition for smart cities. *Handbook of Optoelectronics: Applied Optical Electronics (Volume Three)*, 29.

Weihong, W., & Jiaoyang, T. (2020). Research on license plate recognition algorithms based on deep learning in complex environment. *IEEE Access*, *8*, 91661–91675. https://doi.org/10.1109/ACCESS.2020.2994287

White, C. E., Bernstein, D., & Kornhauser, A. L. (2000). Some map matching algorithms for personal navigation assistants. *Transportation Research Part C: Emerging Technologies*, *8*(1-6), 91–108. https://doi.org/10.1016/S0968-090X(00)00026-7

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1). https://doi.org/10.18637/jss.v040.i01

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wieczorek, J., Fernández-Moctezuma, R. J., & Bertini, R. L. (2010). Techniques for Validating an Automatic Bottleneck Detection Tool Using Archived Freeway Sensor Data. *Transportation Research Record*, *2160*(1), 87–95. https://doi.org/10.3141/2160-10

Wilson, T., & Best, W. (1982). Driving strategies in overtaking. *Accident Analysis & Prevention*, *14*(3), 179–185. https://doi.org/10.1016/0001-4575(82)90026-4

Witte, P. A., Wiegmans, B. W., van Oort, F. G., & Spit, T. J. M. (2012). Chokepoints in corridors: Perspectives on bottlenecks in the European transport network. *Research in Transportation Business and Management*, *5*, 57–66. https://doi.org/10.1016/j.rtbm.2012.10.001

Wolniak, M. J., & Mahapatra, S. (2014). Data and Performance Based Congestion Management Approach for Maryland Highways. *Transportation Research Record: Journal of the Transportation Research Board*, *2420*(1), 23–32. https://doi.org/10.3141/2420-03

Wylie, I. (2015). 'Traffic lights are so dictatorial' ... But are roundabouts on the way out? *The Guardian*. https://www.theguardian.com/cities/2015/oct/19/traffic-lights-roundabouts-way-out

Xu, C., Wang, X., Yang, H., Xie, K., & Chen, X. (2019). Exploring the impacts of speed variances on safety performance of urban elevated expressways using GPS data. *Accident Analysis & Prevention*, *123*, 29–38. https://doi.org/10.1016/j.aap.2018.11.012

Xu, Z., Yang, W., Meng, A., Lu, N., Huang, H., Ying, C., & Huang, L. (2018, September). Towards end-to-end license plate detection and recognition: A large dataset and baseline. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yang, M., Liu, Y., & You, Z. (2010). The Reliability of Travel Time Forecasting. *IEEE Transactions on Intelligent Transportation Systems*, *11*(1), 162–171. https://doi.org/10.1109/TITS.2009.2037136

Yang, M., Wang, X., & Quddus, M. (2019). Examining lane change gap acceptance, duration and impact using naturalistic driving data. *Transportation Research Part C: Emerging Technologies*, *104*, 317–331. https://doi.org/10.1016/j.trc.2019.05.024

Zhai, X., Bensaali, F., & Sotudeh, R. (2012). OCR-based neural network for ANPR. *2012 IEEE International Conference on Imaging Systems and Techniques Proceedings*, 393–397. https://doi.org/10.1109/IST.2012.6295581

Zhang, J., Wang, F., Wang, K., Lin, W., Xu, X., & Chen, C. (2011). Data-Driven Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1624–1639. https://doi.org/10.1109/TITS.2011.2158001

Zhao, Y., Zhu, X., Guo, W., She, B., Yue, H., & Li, M. (2019). Exploring the Weekly Travel Patterns of Private Vehicles Using Automatic Vehicle Identification Data: A Case Study of Wuhan, China. *Sustainability*, *11*(21). https://doi.org/10.3390/su11216152

Zheng, F., Li, J., van Zuylen, H., Liu, X., & Yang, H. (2018). Urban travel time reliability at different traffic conditions. *Journal of Intelligent Transportation Systems*, *22*(2), 106–120. https://doi.org/10.1080/15472450.2017.1412829

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology*, *6*(3). https://doi.org/10.1145/2743025

Zheng, Z. (2014). Recent developments and research needs in modeling lane changing. *Transportation Research Part B: Methodological*, *60*, 16–32. https://doi.org/10.1016/j.trb.2013.11.009

Zheng, Z., Ahn, S., Chen, D., & Laval, J. (2011). Freeway traffic oscillations: Microscopic analysis of formations and propagations using Wavelet Transform. *Transportation Research Part B: Methodological*, *45*(9), 1378–1388. https://doi.org/10.1016/j.trb.2011.05.012

Zhong, M., Lingras, P., & Sharma, S. (2004). Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. *Transportation Research Part C: Emerging Technologies*, *12*(2), 139–166. https://doi.org/10.1016/j.trc.2004.07.006

Zhou, J., & Peng, H. (2005). Range policy of adaptive cruise control vehicles for improved flow stability and string stability. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 229–237. https://doi.org/10.1109/TITS.2005.848359

Zhou, X., & Mahmassani, H. S. (2006). Dynamic OriginDestination Demand Estimation Using Automatic Vehicle Identification Data. *IEEE Transactions on Intelligent Transportation Systems*, *7*(1), 105–114. https://doi.org/10.1109/tits.2006.869629

Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (2019). Big Data Analytics in In-

telligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *20*(1), 383–398. https://doi.org/10.1109/TITS.2018.2815678

Zhu, S., & Levinson, D. (2015). Do People Use the Shortest Path? An Empirical Test of Wardrop's First Principle. *PLOS ONE*, *10*(8), e0134322. https://doi.org/10.1371/journal.pone.0134322