# Integrated Connected Data for Safer More Efficient Traffic Management Operations (iMOVE-1-060)


Dr. Neema Nassir, Prof. Majid Sarvi, and Dr. Patricia Lavieri
University of Melbourne, Faculty of Engineering and IT

**2025**

**Project Team:**
Mustafa Rezazada, Mohammadreza Ghanbari, Zhi Hern Tom, Bo Wang, Lok Sang Chan, Michael Alexander Leo, Mobin Yazdani, Sina Mohri, Jessica Tong, Zay Aye, Annitta Lin, Stella Han, Keya Roy, Yidian Chen, Ryan Zhang, Jay Song

Contact email: neema.nassir@unimelb.edu.au

## Executive summary

### Objective

This project explored the potential of utilising and integrating *vehicle location and telematics data* from connected vehicles (and bicycles) GPS and gyroscope data, alongside traditional sensors, for modern traffic management systems. The primary goal is to investigate opportunities to enhance traffic safety, efficiency, and environmental sustainability by utilising new and emerging telematics data, developing advanced methodologies, identifying new use cases, and deriving actionable insights. The study also investigated surrogate safety measures (SSMs) based on bicycle movement and risk data for proactive traffic management interventions.

### Methodology

The methodology adopted in this project involved a comprehensive multi-stage framework that included data fusion and integration, exploratory analysis, penetration measurements, and advanced analytics techniques, such as simulation, optimisation, and machine learning models. The project methodology initiates with the integration of high-resolution Connected Vehicle (CV) data with existing traffic sensors (SCATS), demonstrating unprecedented coverage of traffic dynamics and delay information at mid-blocks and intersections. The proposed data fusion boosts the penetration rates of connected vehicle trajectory data which is originally available from 3-5% on average in Melbourne metropolitan area and match it up to all (100%) vehicles registered by the SCATS loop detectors at signalised intersections. This information was then used in prototype models for signal optimisation, priority for public transport and vulnerable road users, and precise estimates of safety risks and emissions.

Advanced analytics methods, including sequential clustering, deep reinforcement learning (DRL), surrogate safety measures, computer vision models, and physics-informed neural networks (PINNs), were employed for various objectives in this project. These methods were used to estimate and classify traffic states, optimise signal control, predict traffic safety conflicts, formulate bicycle safety proxy measures, explore driving style variability, and estimate emissions. The project also included the development of a prototype visualisation dashboard for real-time traffic monitoring and decision-making.

### Key findings

1. **Intersection efficiency, multimodality, and safety:**
    - Deep Reinforcement Learning models, enhanced with high-resolution connected vehicle data, outperformed traditional signal control methods in reducing average travel times for “vehicles” and queue lengths by up to 10-15%.
    - With a focus on throughput of individuals (using all modes of transport) instead of throughputs of cars-only at signalised intersections, the signal priority models developed for pedestrian and public transport vehicles leads to up to 5-10% increased throughput for individuals
    - Safe signal models demonstrated opportunities for including safety objectives in green-time control. An enhanced computer vision model significantly improved conflict prediction accuracy.
2. **Insights into driving behaviour and network performance:**
    - Distinct driving styles and behaviours were identified at intersections, varying based on vehicle type, signal phase dynamics, and time of day. This is unprecedented information offering significant opportunity for interventions.
    - A prototype visualisation dashboard was developed to facilitate real-time traffic monitoring and travel time reliability analysis.
3. **Bicycle safety and Surrogate Safety Measures**:
    - Bicycle-involved crashes are more frequent at intersections (59%) compared to mid-block segments (41%).
    - Harsh braking events were concentrated in central metropolitan areas, particularly around intersections and cycling paths.

- Harsh braking and swerving events recorded by gyroscope sensors were strong indicators of potential crash sites.
- Gyroscopic data revealed significant variations in road surface quality, crucial for targeting maintenance and identifying high-risk locations.
- Surrogate safety measures effectively identified hotspots underrepresented by police-reported crash data.

4. **Bicycle flow efficiency:**
   - Popular cycling routes remain consistent regardless of road type, but dedicated bike lanes increase overall cyclist flow, suggesting infrastructure improvements could boost cycling rates.
   - Cyclists face longer delays and lower speeds at major intersections and in the city centre. Dedicated bike paths improve speed, highlighting the benefits of segregated infrastructure.
   - Roads shared with cars show more speed reductions, especially in central areas, due to traffic signals and lack of dedicated bike lanes. This reduction can be linked to longer waiting times at intersections, car-oriented traffic signals, and a lower prevalence of dedicated bike lanes.
   - Five cyclist clusters are identified using unsupervised machine learning. This includes, Suburban Commuters**,** Daily Cyclists**,** Leisure Cyclists**,** Most Delayed Cyclists, and Fastest Cyclists clusters.

5. **Emission estimation:**
   - Real trajectory data captures emission intensity and variability more accurately than simulation methods.
   - The results show that fluctuating traffic congestion leads to variable emissions. Considering variable congestion and flow dynamics is crucial for estimating emissions accurately.
   - Both emission intensity and variability are higher during peak hours on weekend and weekdays.
   - Peak emissions may start far-distant upstream of intersections, decreasing as vehicles slow down approaching intersections or queue ends.
   - Increasing EV compositions in the vehicle fleet significantly reduces emissions, with higher EV percentages yielding greater emission reductions.
   - Emissions per vehicle are lower during rush hours due to reduced speeds and acceleration but increase during off-peak hours when vehicles travel faster. This suggests that effective speed management can further reduce emissions
   - Despite lower emissions per vehicle during peak hours, overall emissions per meter are higher due to increased traffic volume and congestion.
   - Findings suggest that combining effective demand and speed management strategies with increased EV adoption provides an optimal solution for minimising emissions.

6. **Movement data and insights:**
   - Higher penetration rates of GPS-equipped vehicles provided broader data representativeness for data fusion applications.
   - The proposed Residual Physics-Informed Neural Network (Res-PINN) model significantly improved traffic state estimation using limited connected vehicle data.
   - Transfer learning techniques enhanced the model's performance, resulting in lower error rates and more accurate understanding of traffic conditions.

Finally, the project successfully demonstrated the potential of integrating multi-source big data for advanced traffic management applications. The findings highlight significant improvements in traffic efficiency, safety, and environmental sustainability possible through the use of connected vehicle data, telematic data from bicycles, and other datasets. The methodologies developed provide a robust framework for future research and practical applications, offering substantial benefits for urban mobility and traffic management systems.

# Table of Contents

# 1 Introduction

Current traffic operations and management practices primarily rely on traditional fixed-location sensors, including legacy loop detectors, and costly CCTV camera systems. However, the rapidly emerging transport technologies such as connected vehicles, communication technology, vehicle telematics, and smart signal control systems have created unprecedented opportunities for new methods in traffic management and control. This project aims to explore the potential of connected vehicle location data and telematics in conjunction with other available data sources, for safer and more efficient traffic.

Connected vehicles (CV) have the capability to communicate and share near real-time and real-time location/trajectory information and serve as probe vehicles for traffic data collection at scale. They are expected to play a pivotal role in the future of transport (de Luca et al.) and Intelligent Transportation Systems (Coppola & Morisio, 2016; Milkovits et al.). According to a comprehensive literature review (reported in Stage 1 of this project), the emerging big data from vehicle telematics systems has the potential to revolutionise traffic management in various aspects, including:

- Producing fine resolution vehicle trajectory and delay information at scale (Mehran et al., 2012; Wan et al., 2016)
- Continuous traffic state information and volume estimation (Sunderrajan et al., 2016; Yuan et al., 2021)
- Opportunities for more efficient intersection management (Liang et al., 2023; Y. Zhao et al., 2019; Zheng & Liu, 2017)
- New traffic network insights for freeway management (Wang et al., 2020)
- Signal priority controls and management for emergency vehicles (Goodall et al., 2016; He et al., 2012)
- Better understanding of traffic safety and management (J. Liu & Khattak, 2016).

In this project we investigated the opportunities related to car/truck telematics data with a focus on practical outcomes, and identification of immediate potentials for the Australian major metropolitan traffic operations. However, the scope was not limited to cars/trucks only. Promoting and supporting active modes of transport such as bicycles and scooters while ensuring their safety are bold and emerging transport and mobility goals, globally. We have identified opportunities for improving not only traffic flow efficiency but also safety and efficiency for other modes of transport (pedestrians, public transport and micromobility). In addition to car/truck telematics data, this project also utilises large-scale bicycle GPS and gyroscope data, which hold significant potential for improving both traffic safety and efficiency particularly for vulnerable road users (VRU).

To test all methodological opportunities for using the vehicle location data and identify practical use cases, several methods are developed and tested in this project to leverage the potentials of movement and gyroscopic data from connected vehicles and bicycles in traffic management. These aspects include:

- Trajectory reconstruction, data fusion and automation.
- Intersection safety and efficiency optimisation (including occupancy-based priority and multi-modality).
- Producing insights from network and freeway exploratory analyses.
- Developing proactive road safety measures for vulnerable road users.
- Producing network level insights into bicycle efficiency and movement patterns.
- Traffic state classification and congestion estimation.
- Deducting driving patterns and behaviours on major arterials.
- Emission estimation using reconstructed trajectories.

This report summarises our efforts in testing these methods and provides a comprehensive investigation for effectively utilising and integrating vehicle and bicycle location and gyroscope data into modern traffic management systems and practices.

# 2 Project Scope and Report Outline

While telematics data provides unprecedented spatial and temporal resolutions, accuracy, and scalability, it is currently only available from a limited number of vehicles and bicycles. To fully exploit its potential and unlock novel applications in traffic management, researchers often integrate this data with other sources. However, this integration is challenging due to differences in data characteristics, formats, strengths, and limitations.

This research has developed a structured framework to utilise multi-source data for traffic management. The framework comprises distinct components, steps, and procedures to address project objectives and scope within three traffic management domains: intersection management, road safety management, and network and freeway management.

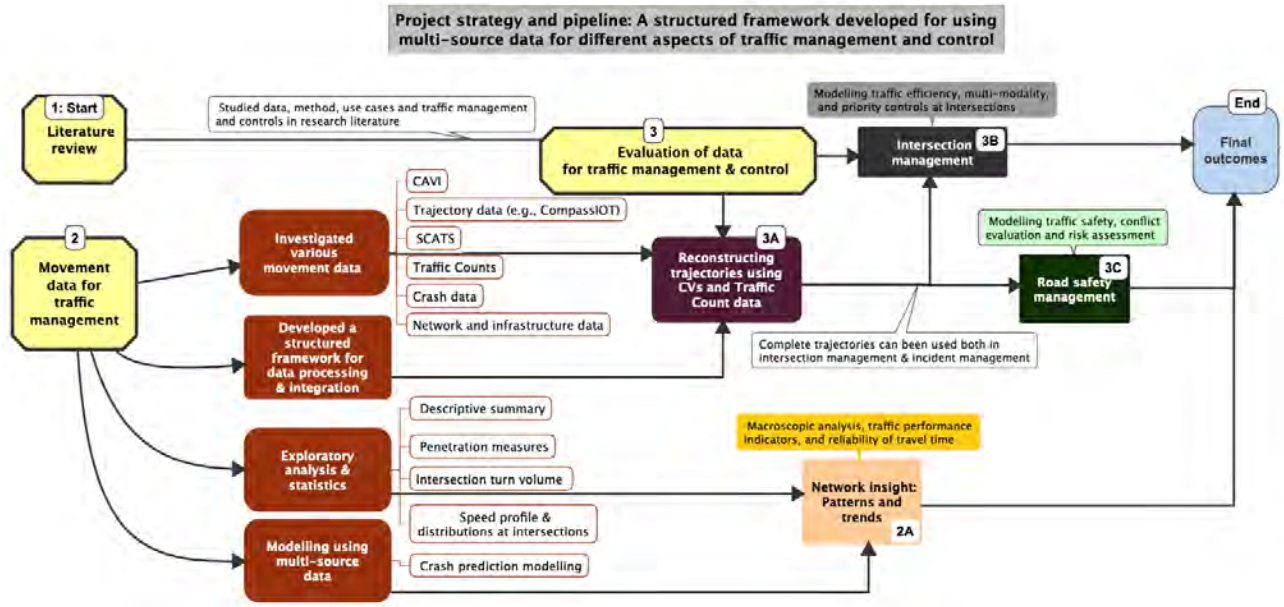



**Figure 2.1: Structured framework for using multi-source data in traffic management and control**

The framework is organized into three main tasks, each consisting of various parts and key elements essential for comprehensive traffic data management and application in real-world traffic systems.

**Stage 1: Literature Review and Stakeholder Interviews**
The foundation of the framework involves an extensive review of existing research literature to examine data, methods, use cases, and traffic controls (refer to literature review report). This step has been vital for gaining a deep understanding of current traffic management practices, identifying gaps in knowledge, and determining the most effective practices and methodologies to be incorporated into this framework.

In addition, we conducted a comprehensive set of interviews and discussions with groups of stakeholders across government and ITS industry in Australia, in order to understand their key areas of focus and interest in using existing or emerging data sources and technology to support existing and emerging network management activities, and the opportunities and challenges associated with this integration. Major findings form these interviews are also reported back in another report from stage 1 (stakeholders interviews report).

**Stage 2: Big Data for Understanding Traffic**
This stage is dedicated to conducting an extensive analysis of movement data and its crucial role in understanding traffic dynamics. This phase starts with a data fusion practice and serves as the foundation for the subsequent.

In this task, a comprehensive investigation is carried out on various movement datasets, such as SCATS signal time data (Sydney Coordinated Adaptive Traffic System), CAVI (Cooperative and Automated Vehicle

Initiative) trajectories, CompassIOT data (connected vehicle data containing location, speed, acceleration, and gyroscopic information), Traffic Counts from loop detectors, bicycle telematic data, and network georeferenced maps and infrastructure data. An automated data processing framework is developed to ensure the readiness of the data for utilisation by traffic management algorithms. This stage encompasses data collection, cleansing, fusion and transformation stages, all of which are essential for enabling the seamless application of traffic management algorithms. The subsequence steps in this stage include:

- Conducting exploratory analysis to understand traffic trends, patterns, and trajectory insights from connected vehicle data and other data sources (presented in Section 3)
- Calculating penetration rates that reveal the proportion of connected vehicles to overall traffic (Section 3.2)
- Computing intersection turn volumes to examine traffic flows, and the volume of traffic traversing intersection turns as well as producing turn speed profiles (Section 3.3)
- Crash prediction modelling using multi-source data (see Section 3.4)
- Data fusion between traffic trajectories and vehicle counts data upstream of intersections (Section 4)

**Stage 3: Multi-source Data for Traffic Management and Control**
This stage focuses on utilising the reconstructed trajectories form connected vehicle data and traffic counts in different use cases of traffic management. The primary objective is to test, evaluate and apply the processed data to enhance traffic management and control.

The reconstructed vehicle trajectories, derived from a combination of limited connected vehicle data and traffic count data, are utilised to model various components of the traffic management system, including:

- Intersection signal control management and multi-modal traffic operations at intersections (Section 5)
- Micromobility management, including bicycle safety and efficiency (Section 6)
- Computer vision for traffic safety (Section 7)
- Network and freeway management (Sections 8-9)
- Emission estimation (Sections 10)

This stage is the main focus of this project and is essential in producing evidence-based knowledge of new opportunities to utilise the reconstructed trajectories for better traffic management systems.

# 3 Exploratory analysis of datasets

This section presents telematic data from various sources, along with insights and descriptive statistics. This is followed by an estimation of the penetration rate of connected vehicles and their trends over time and space.

Furthermore, a linear regression model is formulated to combine the risk measures from connected vehicle data with actual historical crash data in the state of Victoria, allowing for an estimation of their correlation and understanding of the explanatory power of connected data for understanding crash risks.

Finally, the application is extended to analyse the distribution of turn volumes and speed profiles at intersections. The key findings are briefly reviewed in this section; for a detailed report of results, please refer to the interim report.

## 3.1 Vehicle position data

Telematics data, encompassing telecommunications and informatics within the automotive industry, provides a critical lens through which we can analyse and understand driver performance. This type of data, particularly when it includes metrics like acceleration/deceleration and gyroscopic measurements, offers a wealth of information for understanding vehicle dynamics and driver behaviour, albeit not in real-time. The gyroscopic information provides insights into vehicle movements and stability, contributing significantly to areas such as safety analysis and behavioural studies.

Due to the limited market penetration of connected vehicles (CV) and their associated data, researchers have explored data integration methods to leverage the high resolution, wider coverage, and rich information provided by CV data. These integration methods, which combine CV data with traditional roadside sensors, infrastructure/network information, and maps, have proven to be cost-effective, real-time alternatives to costly video systems and fixed/non-mobile sensors. One significant use case of this data integration is the reconstruction of complete trajectories for non-connected vehicles using the rich trajectory data from CV.

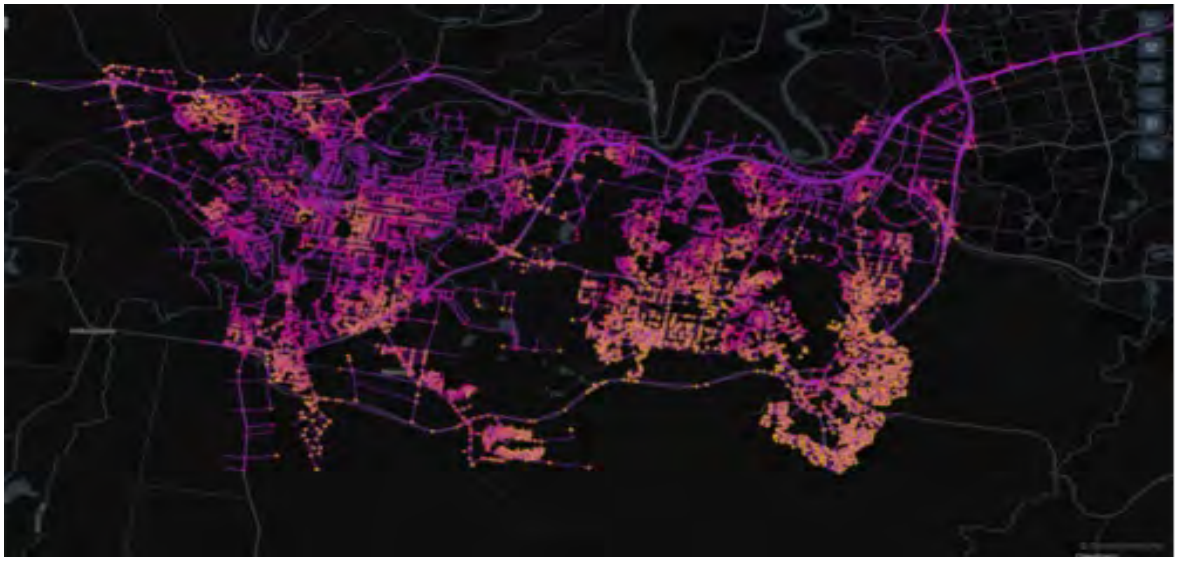


Figure 3.1. Visualisation of CAVI data.


Figure 3.2: CompassIOT coverage in The Greater Melbourne Area

In this project, we extensively examined two datasets from connected vehicles: CAVI data, an initiative undertaken by the Queensland Government (Figure 3.1), and CompassIOT telematics data (Figure 3.2), which comes from a private aggregator of connected vehicle data. Both datasets provide detailed information on vehicle movement and trajectories, including additional features that traditional sensors and data collection methods cannot capture, such as acceleration and gyroscopic information. However, each dataset has its own advantages and limitations, making them suitable for different traffic management use cases and applications.

The CAVI dataset offers precise insights, including speed and altitude values, making it suitable for selection based on these attributes. With relatively higher spatiotemporal resolution, CAVI excels in capturing nuanced vehicle movements, such as deceleration and acceleration when navigating roundabouts, intersections and uphill and downhill roads, providing continuous motion records. However, its coverage is limited to a specific number of vehicles participating in the initiative (see Figure 3.3). Additionally, the CAVI data includes altitude values, which can be integrated into a 3D map, allowing for more detailed analysis of road conditions, including uphill and downhill sections (see Figure 3.4).

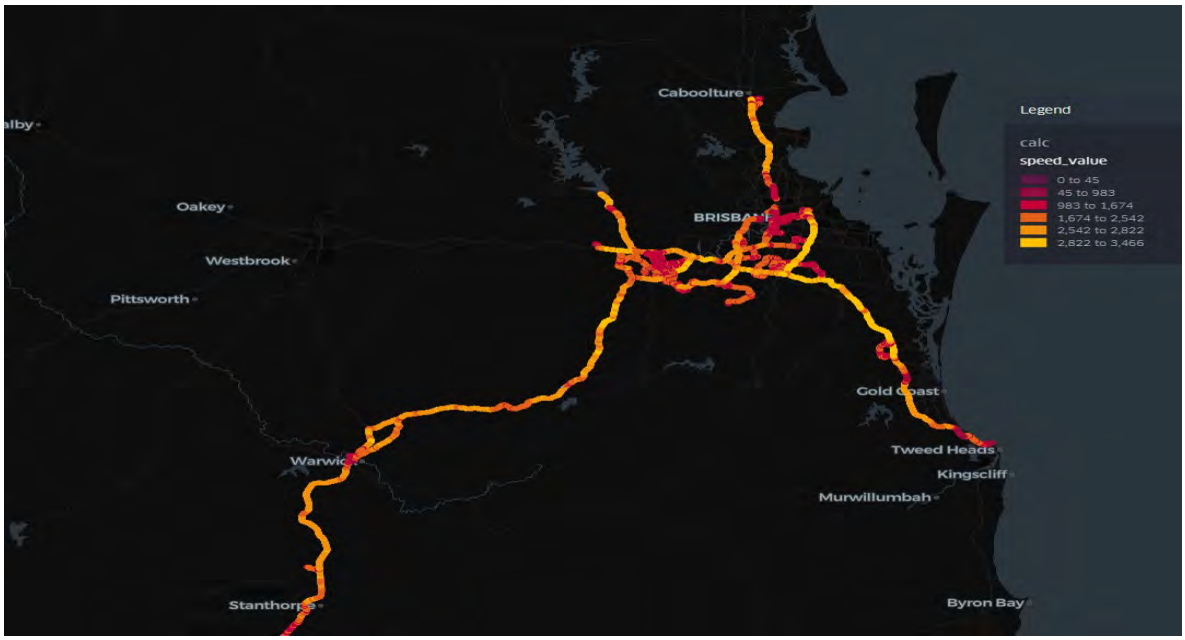



Figure 3.3: The penetration information of station ID QCVCWA240344 which mainly covers the Brisbane area, the speed value is in cm/s.

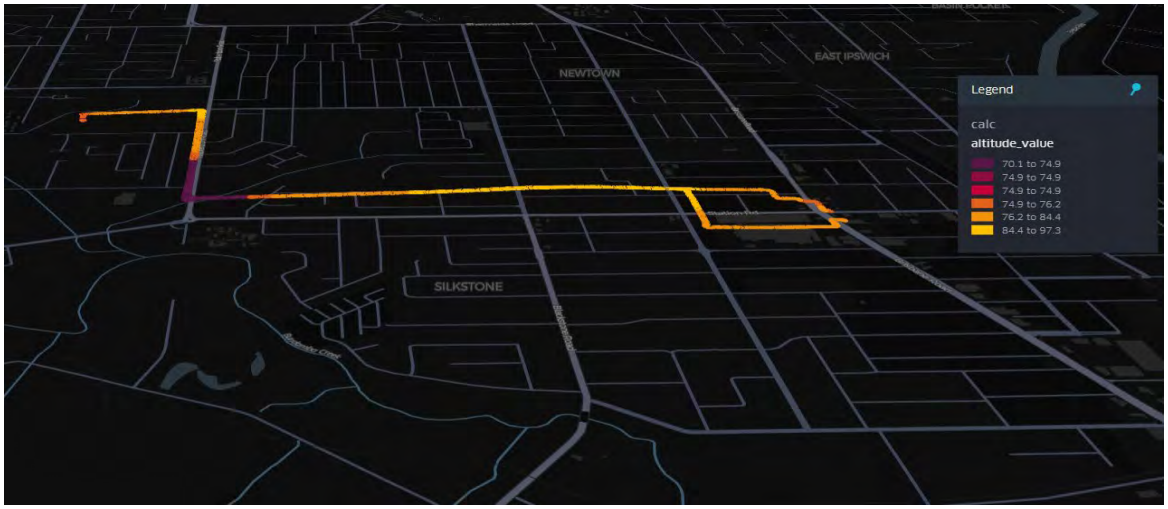
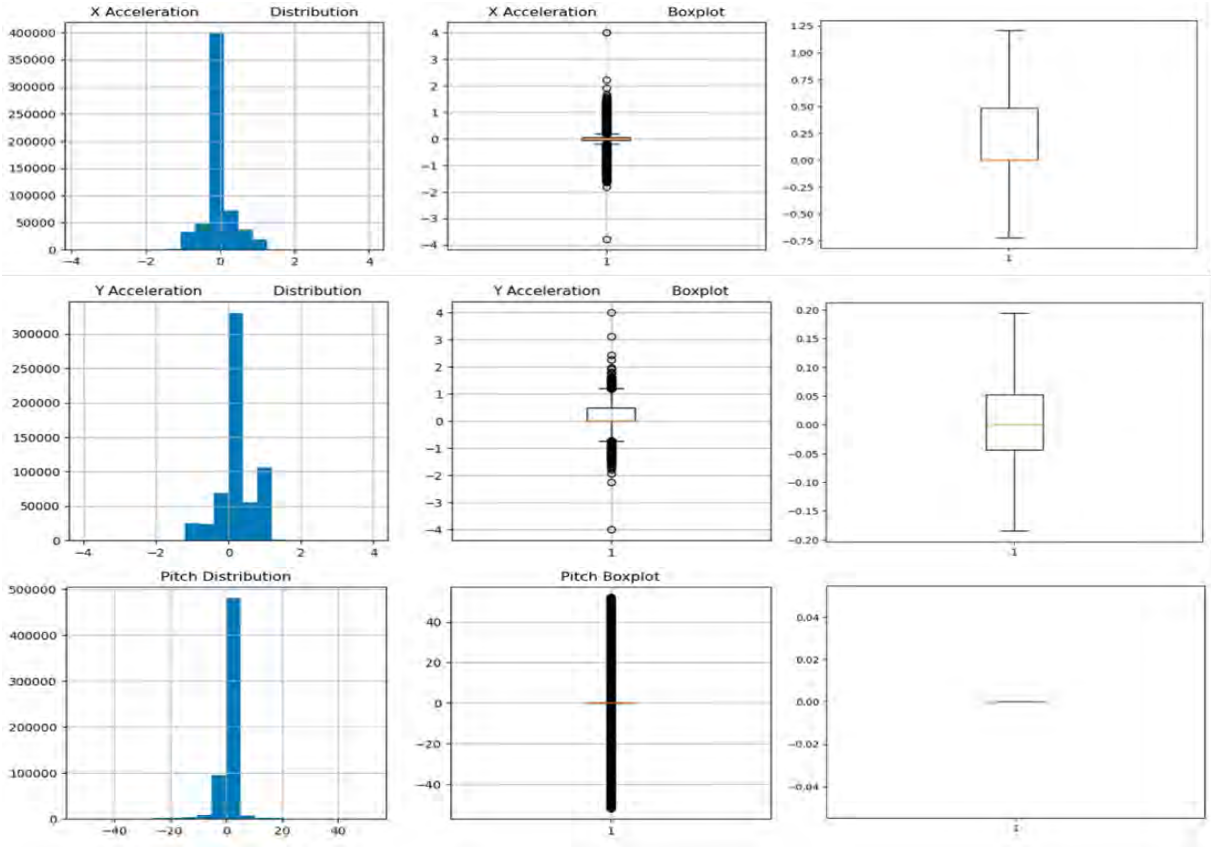



Figure 3.4: 3D visualisation result of altitude variation in one trip from the CAVI data, an obvious uphill and downhill situation is presented

On the other hand, CompassIOT data, as a distinguished source in the field of telematics, provides comprehensive datasets that include gyroscopic data in addition to acceleration information. This data offers valuable insights into vehicle movement, stability, and factors relevant to safety analysis, behavioural studies, traffic flow, and congestion management. The inclusion of acceleration data allows for the analysis of speed variations and driving patterns, which is crucial for driver safety assessments and fleet management efficiency (see Figure 3.5). Similarly, gyroscopic data provides understanding of vehicular orientation and stability, key factors in studying and improving vehicle handling and dynamics (see Table 3.1 for a descriptive summary of the data features and compositions).



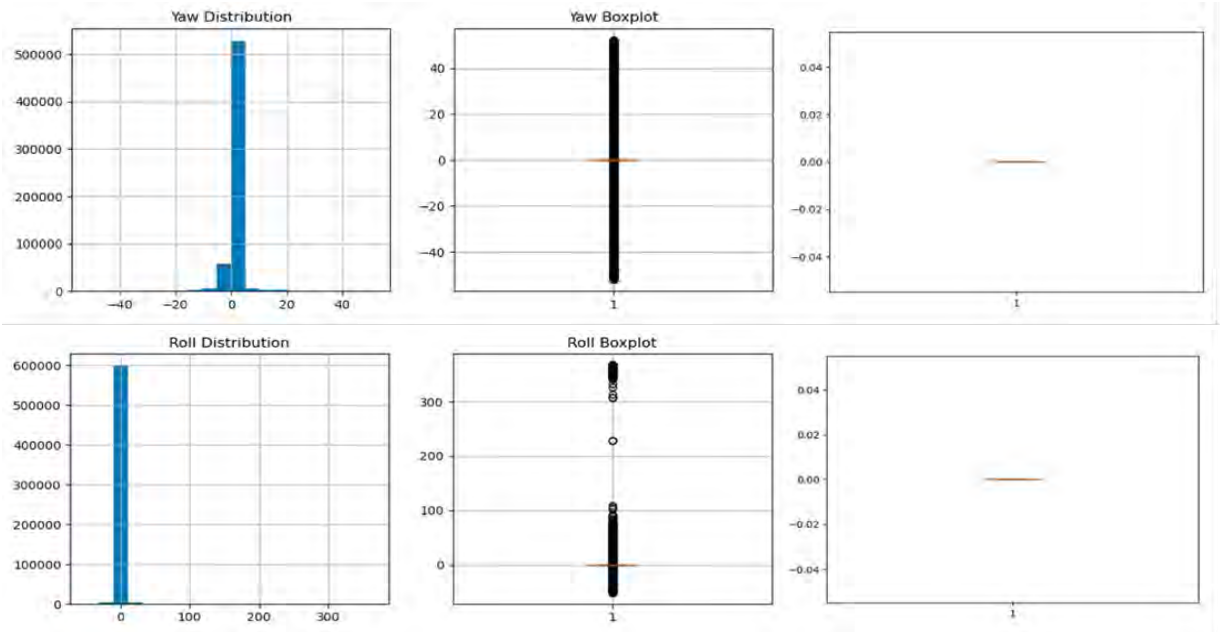



Figure 3.5: Distribution (left), boxplot (middle), Boxplot without outliers (right) for X Acceleration, Y Acceleration, Yaw, Pitch and Roll

**Table 3.1: CompassIOT descriptive summary table**

| | | With Location, Timestamp, Speed | + X Acc and Y Acc | + X Acc and Y Acc + Roll, Yaw, Pitch |
|---|---|---|---|---|
| Number of Trajectories | All | 41,280,155 | 3,411,560 (8.26%) | 1,095,603 (2.65%) |
| | Car | 38,393,085 (93.00%) | 1,674,829 (4.06%) | 1,095,449 (2.65%) |
| | HCV | 2,174,375 (5.27%) | 1,032,798 (2.50%) | 123 (<0.01%) |
| | LCV | 712,695 (1.73%) | 703,933 (1.71%) | 26 (<0.01%) |
| Total Points | | 1,812,9198,028 | 698,674,040 | 60,384,036 |
| Total Travel Distance (Muthugama et al.) | | 292,059,273 | 79,829,803 | 22,622,490 |

While both datasets serve as reliable sources of information on vehicle movement and trajectories, CAVI data provides high spatial and temporal resolutions with a high transmission frequency, enabling the generation of smooth and sufficient data points for individual trajectories. This makes it an unprecedented source for advanced traffic applications in various domains, including road safety intersection management and efficiency, network and freeway congestion management, and more.

CompassIOT data, on the other hand, covers significantly large networks for longer periods. It also provides gyroscopic information that is crucial for traffic safety applications. Additionally, since the CompassIOT data comes from public connected cars, it has much higher penetration and coverage. Therefore, in the subsequent sub-section, we provide insights from penetration tests and other analyses conducted on CompassIOT data.

## 3.2 Penetration measures

Penetration rate, in the context of probe vehicle data, refers to the proportion of vehicles within a given population that are equipped with GPS devices with location information available in our data. This metric is pivotal in assessing the extent to which such technology is adopted and utilised in a specific region or demographic. Calculating the penetration rate for telematics data is crucial for several reasons. Primarily, it aids in evaluating the representativeness of the data collected. A high penetration rate implies that the data encompasses a broader segment of the vehicle population, thereby offering a more comprehensive and accurate depiction of trends, behaviours, and preferences. This, in turn, facilitates more informed decision-making in areas such as traffic management, urban planning, and the development of automotive technologies. Additionally, understanding penetration rates assists stakeholders in identifying potential

markets for expansion and assessing the effectiveness of current telematics solutions. The data from CompassIOT is derived exclusively from public connected cars. In the following sections, we present insights gathered through penetration tests and other analyses conducted specifically on CompassIOT data.

In order to facilitate accurate and meaningful rate calculations, we have undertaken comprehensive data preparation steps, leveraging various sources of information. These preparatory measures ensure the integrity and reliability of the data utilised in our rate calculation process. Below, we provide a summary overview of the data sources and the steps taken to ensure data accuracy:

1. **Vic's Road Corridors Shapefile (City of Melbourne):**
2. **Traffic Signal Volume Data (SCATS) - Site and Detector Information:**
3. **Homogeneous Traffic Volume Network Data (HTVS):**
4. **VKT Data from the Department of Transport:**

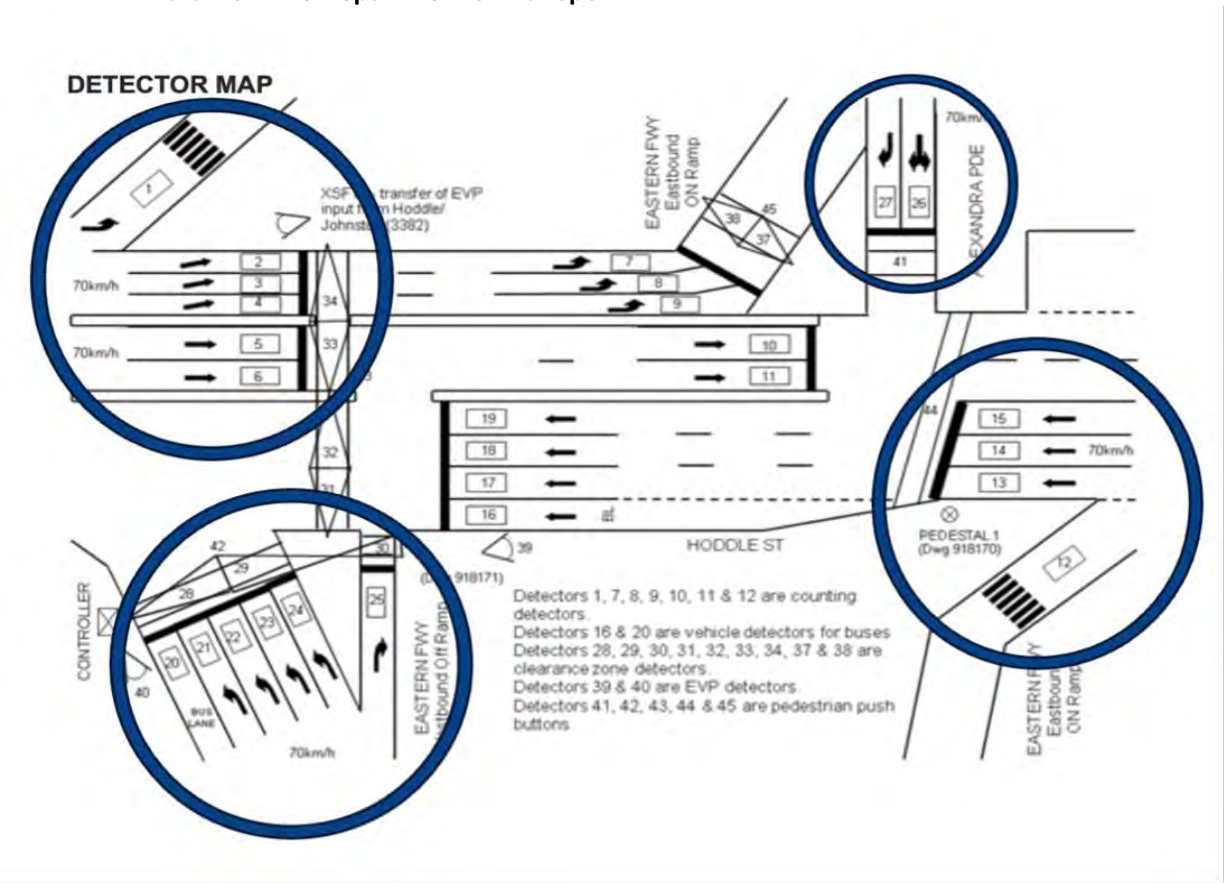



Figure 3.6: Processing loop detector data (only values from the first detector group entering the intersection is selected).

To establish the penetration rate, we have devised a formula to calculate it as follows:

$$\text{Daily Penetration Rate} = \frac{\text{Sum of unique vehicle identifiers in 15-minute intervals for 24 hours}}{\text{Total traffic volume recorded in one day}}$$

This formula quantifies the penetration rate, considering the unique vehicle identifiers within specific time intervals and the overall traffic volume (from loop detectors) over the course of a full day.

We have explored and tested four distinct methods for calculating the penetration rate in our study. Each method has its own set of advantages and drawbacks, which we have carefully considered to ensure the

most accurate and reliable results. Out of these methods (Method 3) demonstrates a notable proficiency in maximising both the capture rate and the accuracy of data (other methods are presented in the interim report). In this method, we manually draw extended polygons for each SCATS (Sydney Coordinated Adaptive Traffic System) site (Figure 3.7). The goal is to maximise both the capture rate and accuracy of data. By strategically expanding the polygons from entry routes to the nearest exit routes, we aim to include all relevant vehicle movement within the defined areas. This approach ensures a high precision penetration rate estimation.

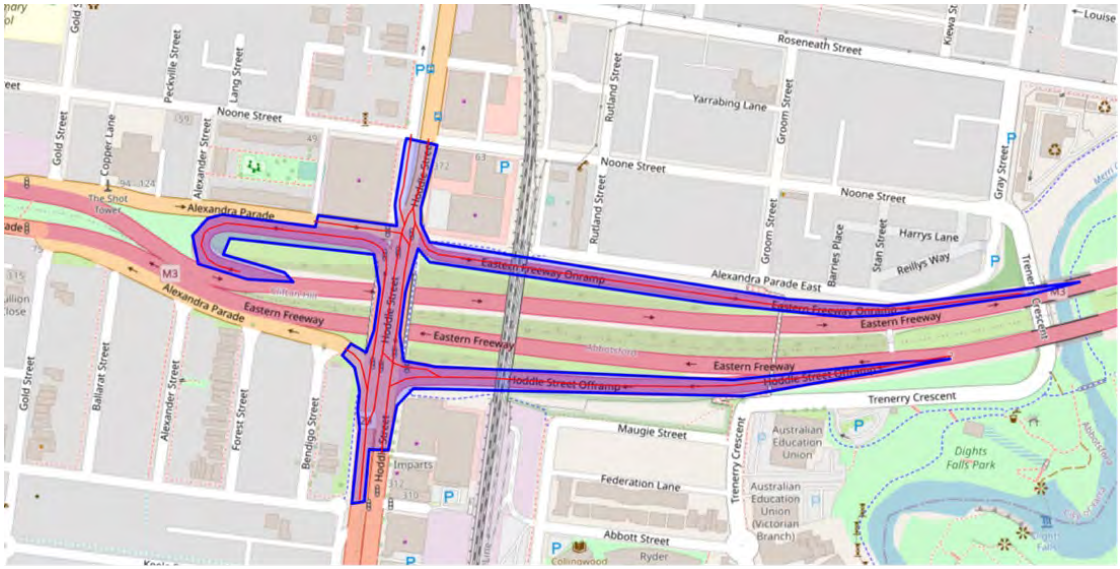


Figure 3.7: Manually drawn polygon at SCATS Site 3380 (HODDLE STREET / EASTERN FREEWAY)

A comprehensive analysis has been conducted to determine the penetration rates of CompassIOT data at different SCATS (Sydney Coordinated Adaptive Traffic System) sites across the greater Melbourne area for the year 2022. This analysis was performed utilising Method 3, and the results are categorised into four distinct color-coded groups for ease of interpretation:

- **Green (< 1.00%):** These sites exhibit penetration rates below 1%.
- **Yellow (1.00 - 2.50%):** This category represents sites with penetration rates between 1.00% and 2.50%.
- **Orange (2.50 – 5.00%):** Sites in this group have penetration rates ranging from 2.50% to 5.00%.
- **Red (> 5.00%):** These locations experience penetration rates exceeding 5%.

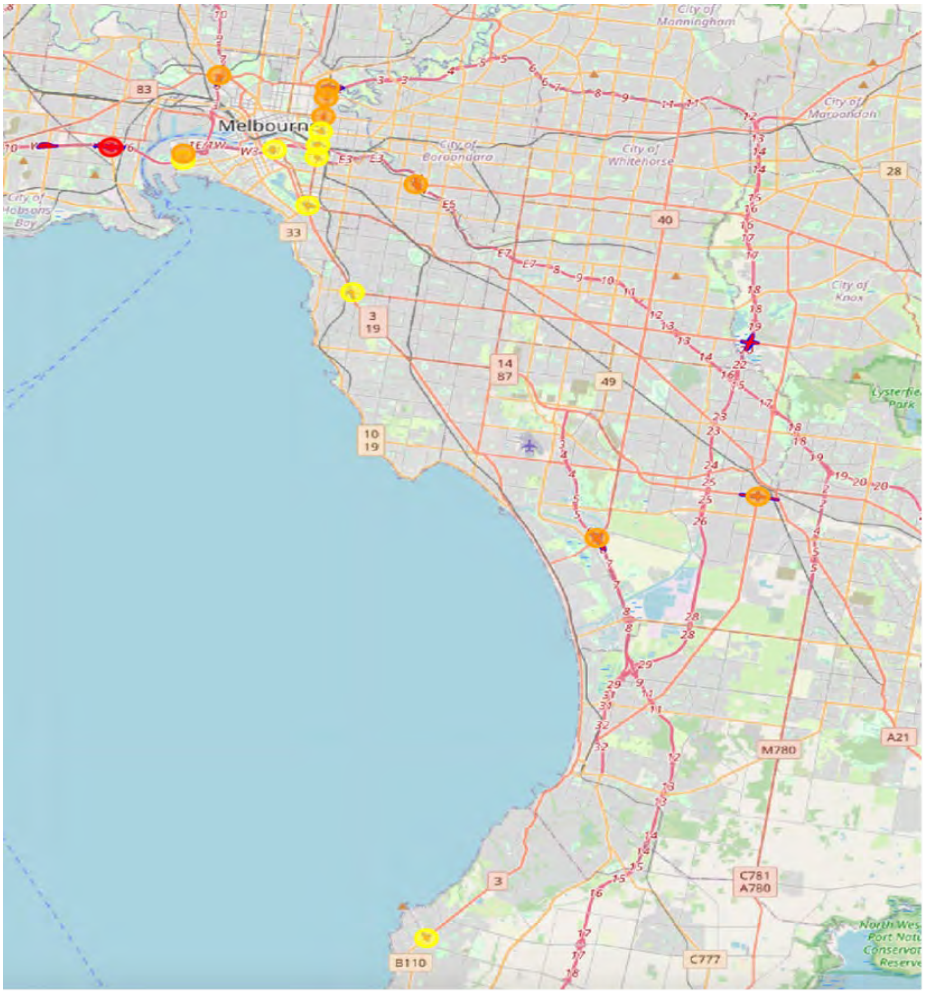


Figure 3.8: Overall Penetration Rate of CompassIOT data in 2022 (Method 3)

On average, the penetration rate across these sites is approximately **2%**. Notably, SCATS site 4489, located at the intersection of Westgate Freeway and Power Street, recorded the lowest penetration rate at **1.38%.** In contrast, the highest penetration rate was observed at SCATS site 2880, situated at Westgate Freeway/Williamstown, with a significant rate of **7.51%.** Furthermore, an examination of major thoroughfares such as Hoddle Street reveals a varying penetration rate, fluctuating between **2.2%** to **4.8%** along the length of the road.

The examination of the monthly penetration rates of CompassIOT data, as illustrated in Figure 3.9, reveals a noteworthy trend. There is a consistent month-over-month increase in the penetration rate. Specifically, for the selected SCATS sites, there is an observable rise in penetration rates ranging from **50%** to **200%** over the span of one year. This trend is a positive indicator, suggesting a significant augmentation in data acquisition over time.

For instance, Figure 3.9 demonstrates this upward trajectory using SCATS Site 3380 (Hoddle Street/Johnston Street) as a case study. The data shows an increase in the penetration rate from approximately 2% in January 2022 to 5% by December 2022. This growth not only highlights the expanding reach of CompassIOT data but also underscores the potential for more robust and comprehensive data analysis in the future.

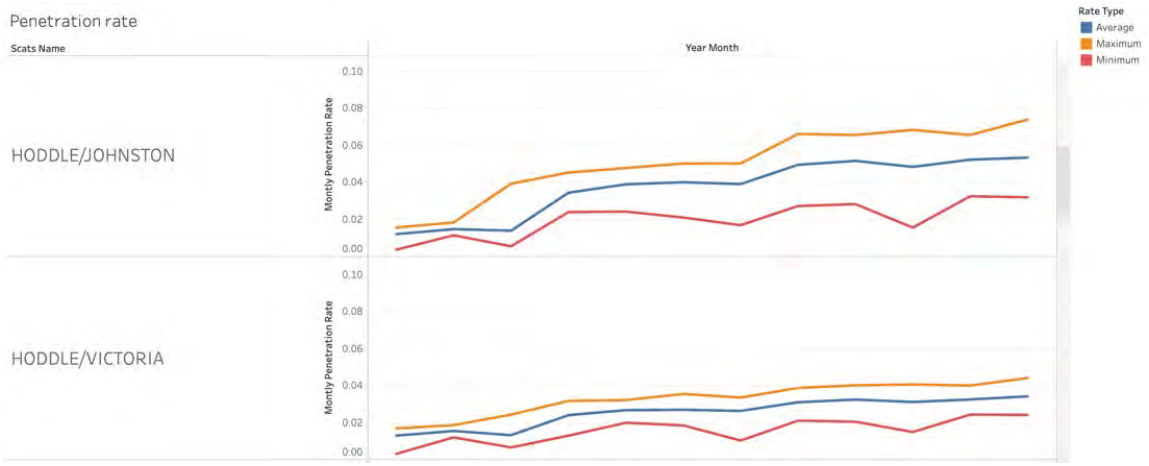



Figure 3.9: Penetration rate for 2 different SCATS sites by month

We have also estimated penetration rates across various suburbs. The dataset spanned a fortnight, from 8th to 23rd May 2022, facilitating the calculation of daily average penetration rates. The illustrative representation in Figure 3.10 elucidates the penetration rates across homogenous traffic flows, while Figure 3.11 distinctly categorises the data by suburbs. The colour-coded scheme in the figures indicates penetration rates as follows: rates between 0 to 0.5% are marked in red, 0.5 to 1.75% in orange, and rates above 1.75% are represented in green.

A noteworthy observation from Figure 3.10 and Figure 3.11 is the pronouncedly higher penetration rates in the south-eastern areas. This phenomenon can be attributed to affluent suburbs such as Toorak and Balwyn. These suburbs are characterised by a higher prevalence of newer and luxurious vehicles, which are more likely to be equipped with telematics data systems. The average penetration rate in these suburbs surpasses that of other areas, typically observed to be below 1%. This disparity underscores the potential influence of socio-economic factors on the adoption and integration of advanced vehicular technologies in different urban areas.

Utilising the VicRoad Corridors Shapefile from Data Vic, SCATS traffic volume data, Homogeneous Traffic Volume Network data, and VKT data from the Department of Transport, we have established a robust framework for estimating penetration rate measures. Our findings reveal a varied penetration rate across SCATS sites, with an average rate of approximately 2%, and a notable month-over-month increase, indicating a growing adoption of telematics systems. Moreover, the analysis of homogenous traffic flows and the association of these with different suburbs highlighted a higher penetration rate in affluent areas, suggesting socio-economic factors may play a significant role in the uptake of advanced vehicular technologies. This study not only provides a detailed understanding of current telematics adoption but also paves the way for enhanced traffic management and road safety strategies, ultimately contributing to the goal of achieving a more sustainable urban transport environment.

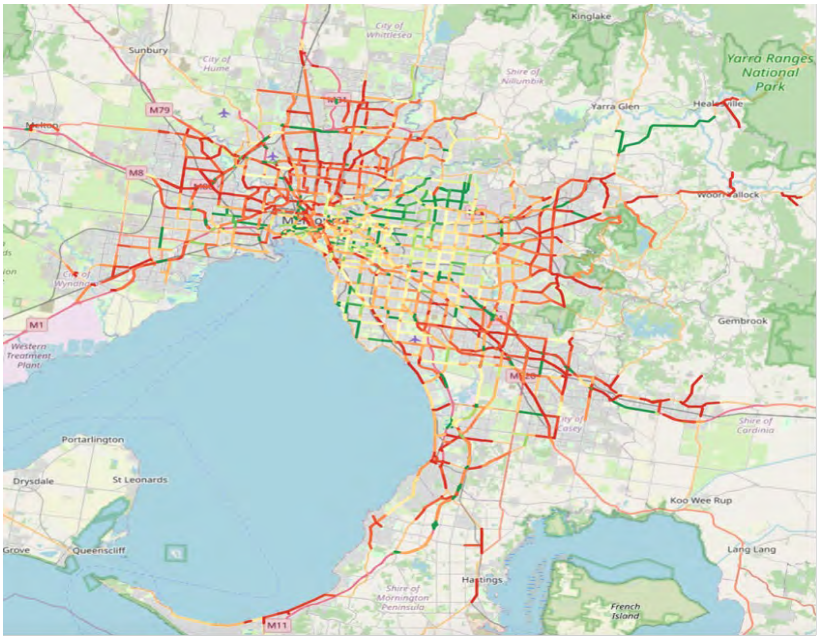



Figure 3.10: Penetration Rate by Homogenous Traffic Flow

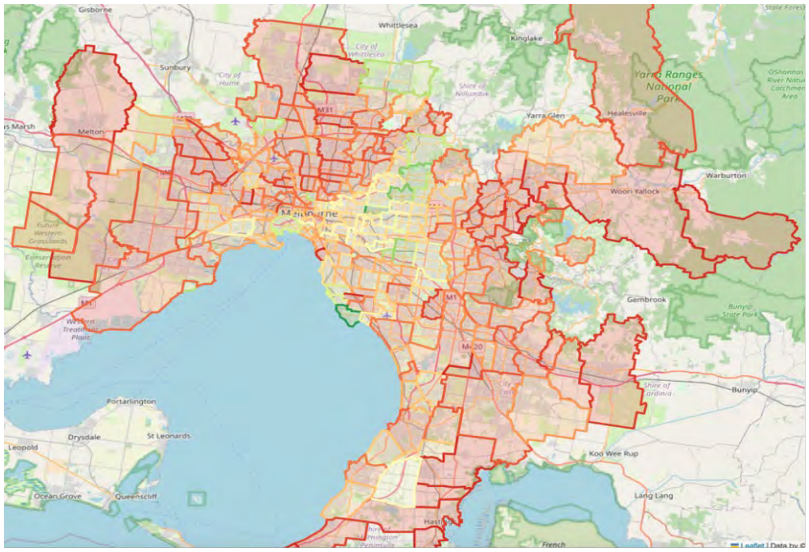



Figure 3.11: Penetration Rate by Suburb

## 3.3 Intersection turns volumes and speed profiles

The analysis of intersection turn volumes and speed profiles constitutes a critical component in the comprehensive evaluation of telematics data. This section aims to elucidate the patterns and characteristics of vehicular movements at intersections, a pivotal aspect in understanding traffic dynamics. The study of turn volumes offers insights into traffic flow distribution, highlighting the predominance of certain manoeuvres over others at specific junctions. Meanwhile, the examination of speed profiles during these turns provides a nuanced understanding of driver behaviour and vehicular performance under varying traffic conditions, please refer to interim report for detailed procedures.

The data presented in Table 3.2 shows the vehicular movement patterns at a specific intersection throughout the year 2022. A thorough analysis of this data reveals a predominant trend of vehicles proceeding straight, followed by a significant number of left turns and a comparatively smaller volume of right turns. Notably, Figure 3.12 confirms the absence of U-turns at this junction.

Table 3.2: Intersection Turn Volume within a 200-Meter Radius at Hoddle Street/Johnston Street

| From/To | S | N | W | E |
|---|---|---|---|---|
| S | 0 | 113,823 | 4,160 | 15 |
| N | 93,188 | 0 | 301 | 9,306 |
| W | 1,354 | 3,437 | 0 | 22,834 |
| E | 1,800 | 1,866 | 20,183 | 0 |

Figure 3.12: Right Turn is not Allowed at Hoddle Street/Johnston Street

The results obtained from our study provide a comprehensive understanding of vehicular movement patterns at intersection throughout the year 2022, as presented in Table 3.2.

Moreover, our findings shed light on instances of non-compliance with traffic rules at some intersection. For instance, the data indicates several occurrences where drivers made right turns despite restrictions. This aspect of the study underscores the importance of monitoring and enforcing traffic regulations to ensure public safety and efficient traffic flow.

In the data presented, a series of figures (Figure 3.13, Figure 3.14, Figure 3.15) portray the speed of vehicles as they approach an intersection (Punt Road / Toorak Road). These figures are helpful in understanding the dynamic behaviour of vehicles in proximity to intersections, a crucial aspect in the study of road safety and traffic flow dynamics.

Upon examination of these speed profiles, a notable pattern emerges, characterised by a distinct 'V' shape in the speed-space profile of each vehicle. This pattern is indicative of a reduction in speed as the vehicles approach the intersection, followed by a subsequent acceleration upon leaving the intersection. This deceleration-acceleration cycle is a clear reflection of the drivers' response to the intersection – typically involving braking and then reaccelerating once the intersection is navigated.

The 'V' shape curve is particularly significant as it encapsulates several critical aspects of vehicular motion near intersections:

- **Braking Phase**: The initial downward slope of the 'V' represents the braking action taken by drivers. This phase is critical for understanding the drivers' perception-reaction time and their response to potential hazards or traffic control devices at the intersection.
- **Lowest Speed Point**: The vertex of the 'V' indicates the lowest speed reached by the vehicles, usually at or near the intersection. This point is significant for assessing the effectiveness of traffic signals, stop signs, and other control measures in ensuring safe speeds at intersections.
- **Acceleration Phase**: The upward slope of the 'V' depicts the acceleration phase, where vehicles increase their speed after passing through the intersection. This phase offers insights into the post-intersection driving behaviour and the efficiency (or lack thereof) of traffic flow downstream.

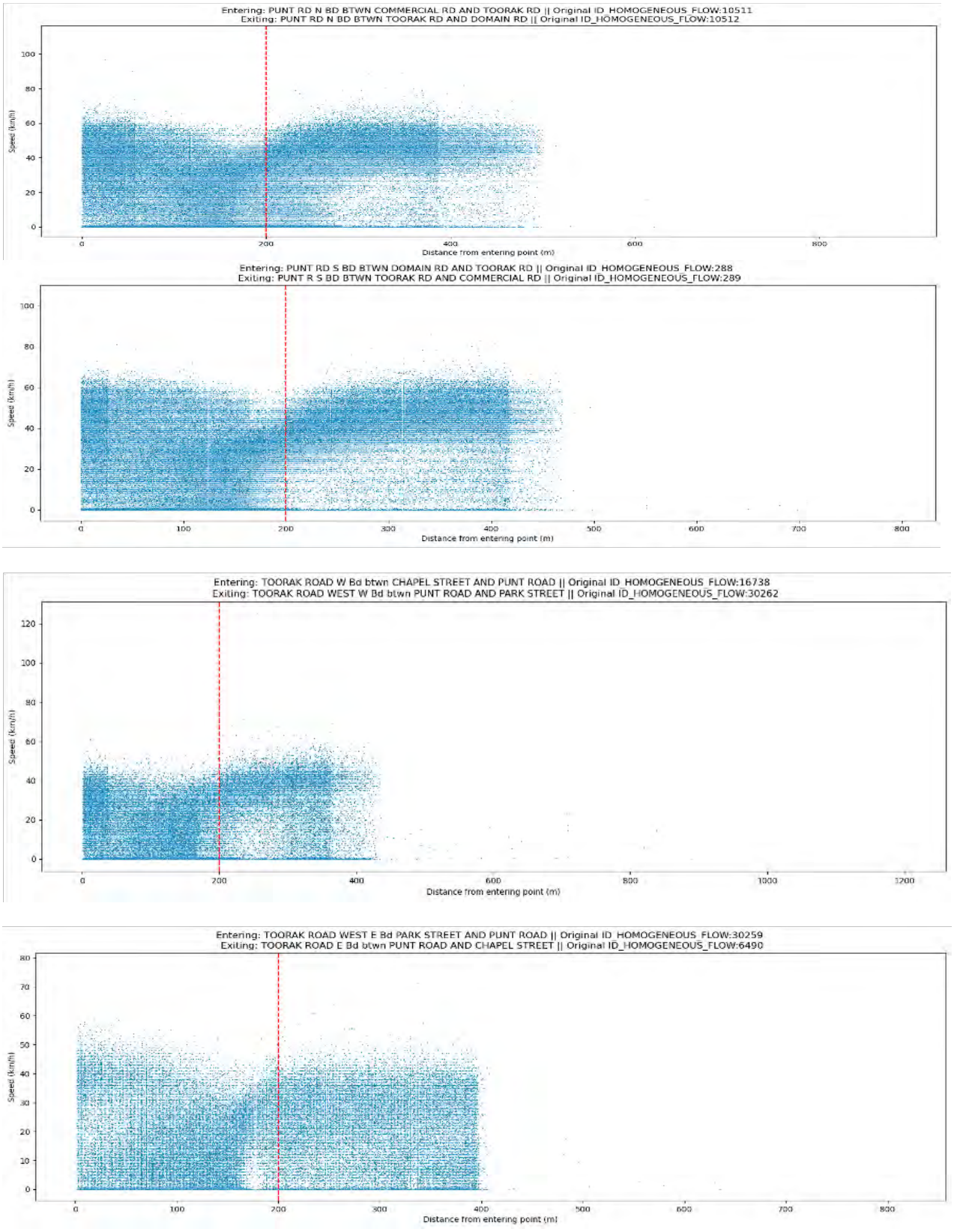



Figure 3.13: Go through speeds at Punt Road / Toorak Road

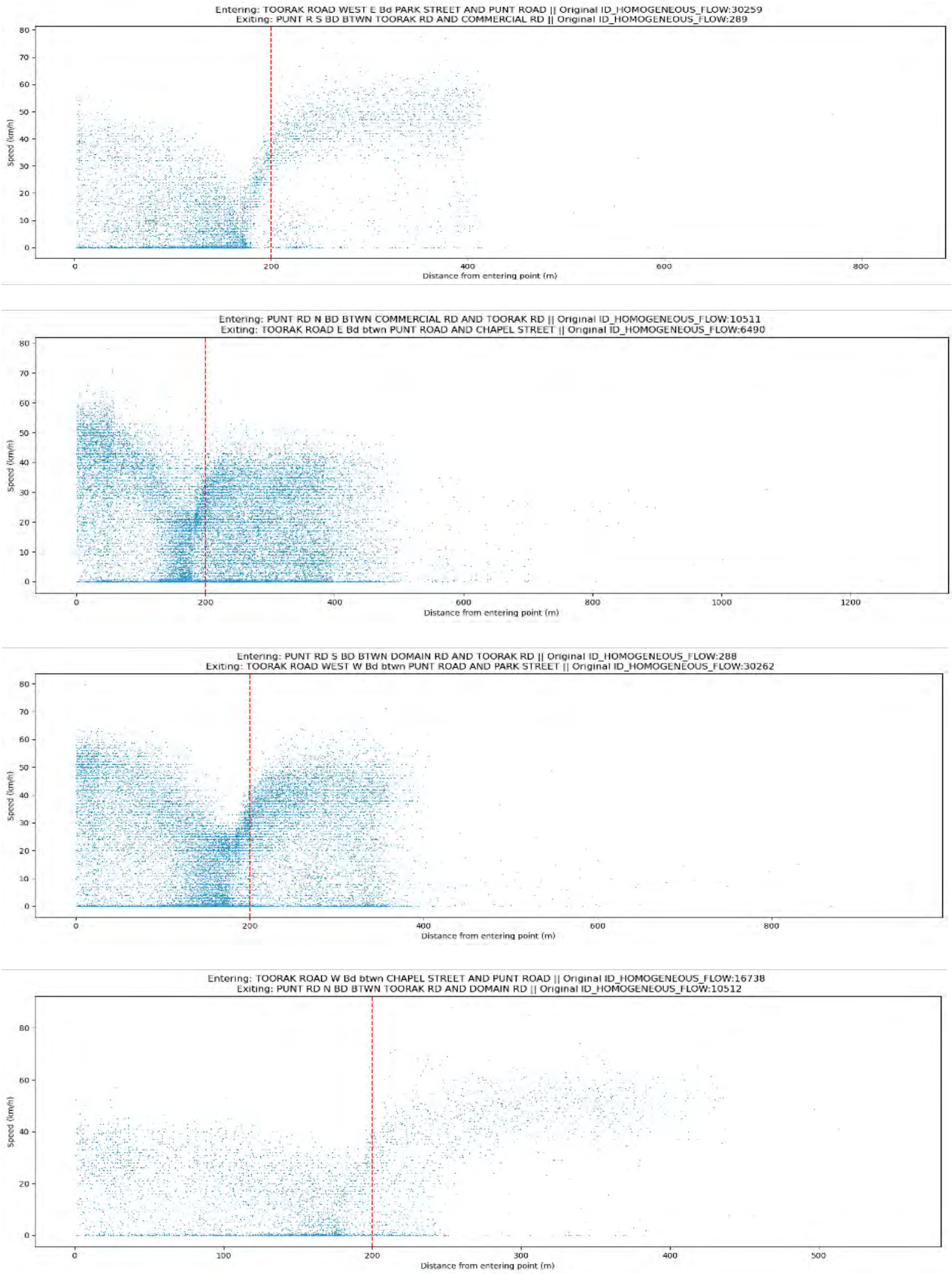



Figure 3.14: Right turn speeds at Punt Road / Toorak Road

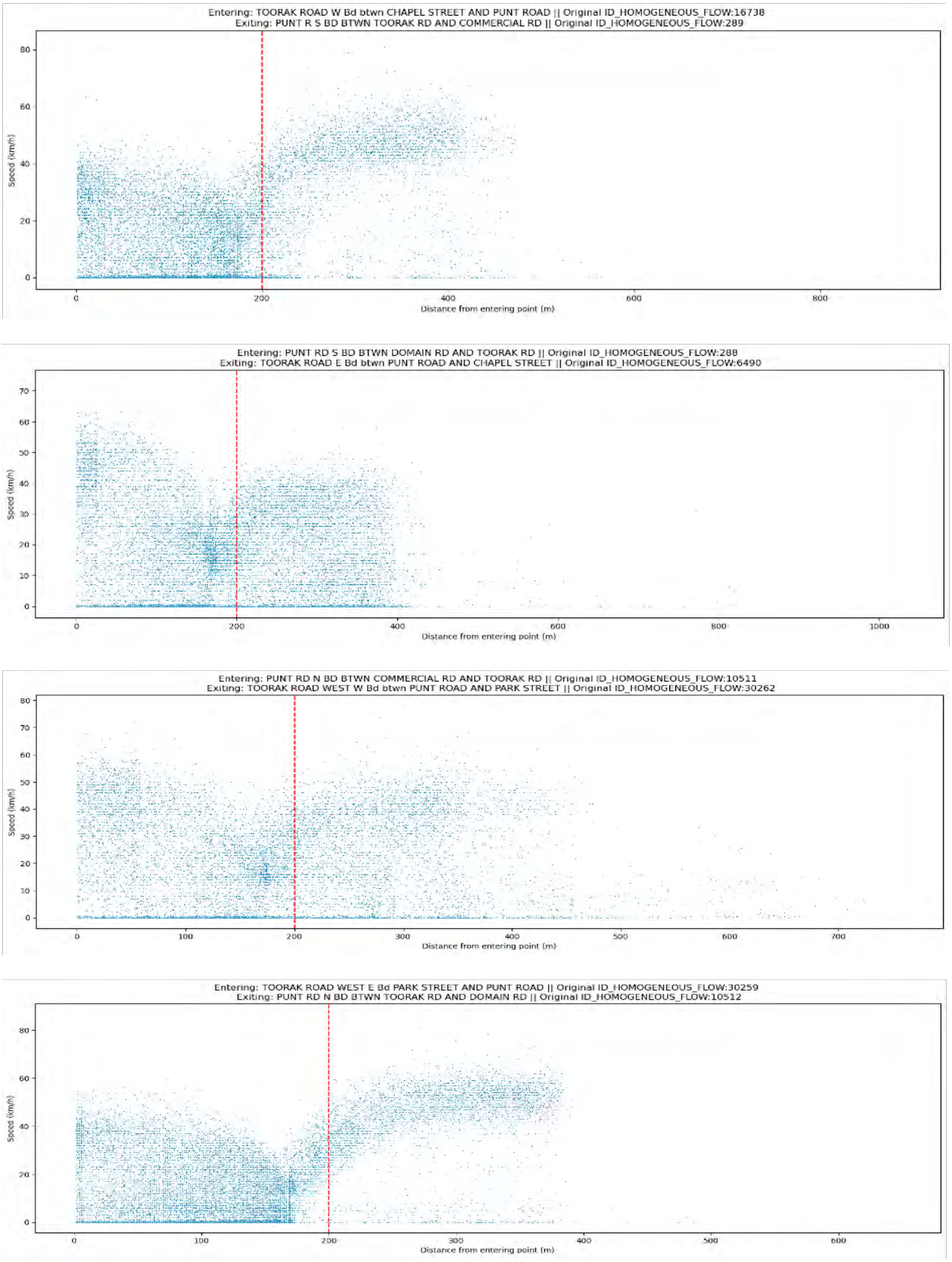



Figure 3.15: Left turn speeds at Punt Road / Toorak Road

As in Figure 3.16, three predominant driving behaviours have been identified when applying a clustering model to speed profiles. The first trend showcases drivers who decelerate upon approaching the

intersection, followed by acceleration upon exiting. This pattern likely indicates a response to traffic signals or pedestrian crossings, reflecting a cautious approach. Specifically, these cases are classified under Cluster 0, which consists of drivers approaching a red light, resulting in a low speed before 200 meters. The second group represents drivers who maintain a consistently lower average speed when approaching the intersection, followed by a noticeable increase in speed after exiting. These cases fall under Cluster 1, which includes drivers approaching a green or yellow light. The third group is characterized by a consistently lower speed average both while approaching and exiting the intersection. These drivers are part of Cluster 2, which consists of drivers seeing queues ahead, with the queue continuing after the intersection, perhaps indicating a downstream bottleneck.

The segmentation of vehicle movements into 40-meter bins, while effective in simplifying the data, may overlook finer nuances in driving behaviour that could occur within these segments. Furthermore, the clustering algorithm, whilst robust, may oversimplify the diversity of driving patterns, particularly in scenarios where driver behaviour does not conform neatly to one of the identified trends. For future work, an exciting avenue for exploration would involve the integration of traffic light conditions into the analysis. Specifically, investigating driver decisions in the 'dilemma zone' the area in which a driver must choose between stopping for a yellow light or proceeding through the intersection could yield valuable insights (Figure 3.17).

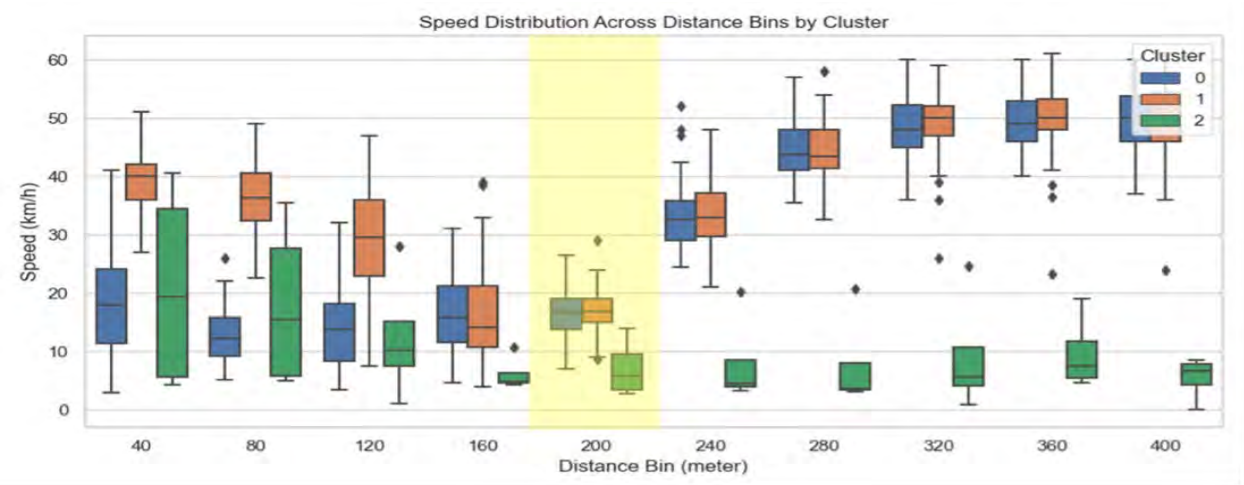



Figure 3.16: Boxplot of 3 different speed clusters

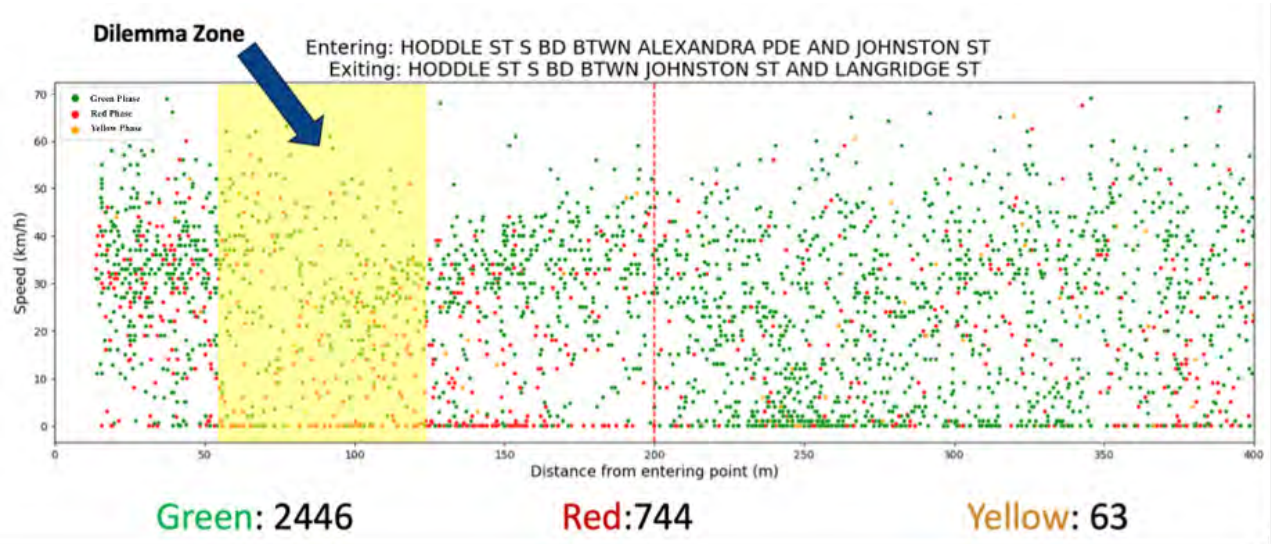



Figure 3.17: Scatterplot of Speed vs Distance (meter) from Intersection Entering Point

## 3.4 Crash Near-miss Data

This section presents an analytical overview of road safety, focusing on telematics data fusion with crash data collected from 2000 to 2020. Utilising advanced statistical techniques, we aim to quantify crash risks and identify key predictors of road accidents. The methodology incorporates three regression models: Linear Regression, Poisson Regression, and Negative Binomial Regression. These models are employed to analyse various predictors such as vehicle speed, acceleration characteristics (X_acc, Y_acc), angular parameters (pitch, roll, yaw), and counts of harsh driving events (harsh deceleration, acceleration, and swerving). Additionally, average speed within different speed zones is considered. The data is further enhanced by incorporating information from CompassIOT for the year 2022.

In examining the relationship between historical crash data (2000-2020) and recent telematics data from CompassIOT (2022), it is important to address the non-overlapping timelines. Despite this, we operate under the assumption that accident hotspots remain consistent over time. To support this hypothesis, a circular plot was generated, averaging the crash data within each unit grid on the map for the period from 2015 to 2020. The rationale for selecting data commencing in 2015, as opposed to 2000, is to mitigate the potential impact of significant changes in road infrastructure or usage over a prolonged period. In Figure 3.20, red points indicate areas where an average of three or more accidents occurred per year within a five-year span. Conversely, green points signify locations with fewer incidents. This visualisation reveals several areas where accidents have consistently occurred over the years. This pattern lends credence to the assumption that historical data can provide valuable insights into current and future accident hotspots. Consequently, this analysis may enhance the predictive accuracy of our model in identifying potential areas of concern based on historical trends.

In the mapping process, we divided the geographical area into grids, adopting an n x n configuration. The dimensions of these grids varied, comprising three distinct sizes: 100 metres by 100 metres, 200 metres by 200 metres, and 250 metres by 250 metres. This detailed segmentation is graphically represented in Figure 3.18. Each grid unit within this spatial arrangement was assigned a specific value, correlating to the number of traffic collisions recorded within that segment from the year 2000 to 2020. This methodological approach allowed for a localised analysis of crash incidences across different areas.

In Figure 3.19, we present a distribution plot which visualises the frequency of accidents in each of these grid units. A notable observation from this plot is the right-skewed nature of the distribution. This skewness predominantly indicates that a significant proportion of the grid units had either no recorded crashes or a very minimal number of incidents.

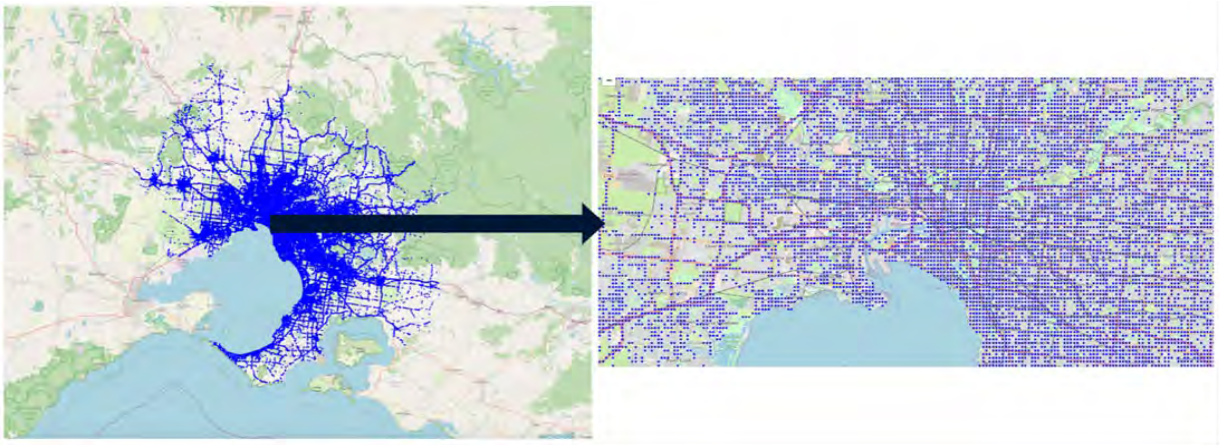


Figure 3.18: Map coded into an n×n meter grid

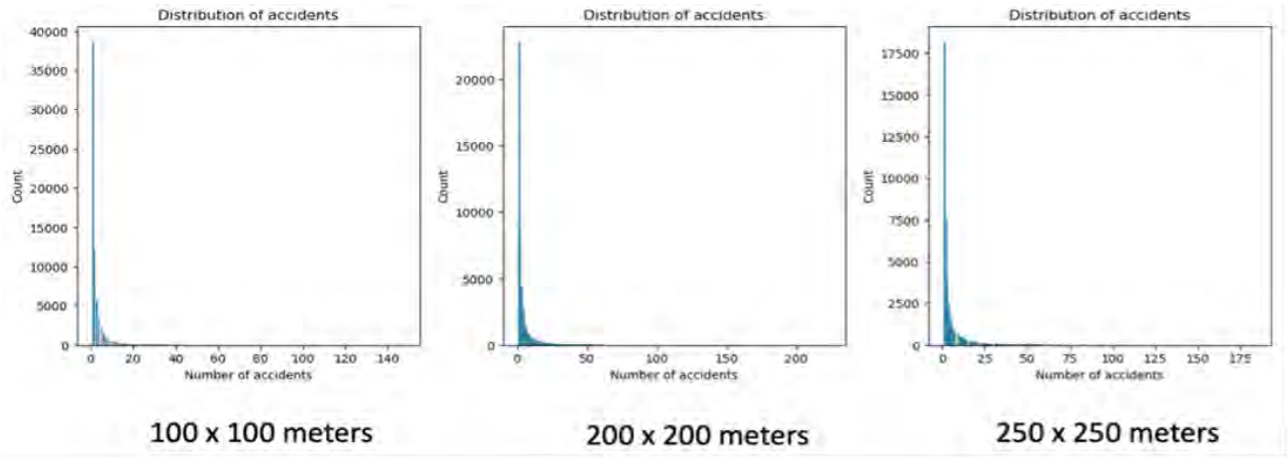



Figure 3.19: The Distribution of accident for different grid sizes

We utilised the Python package *Statsmodels* to develop a comprehensive regression model. This model was precisely constructed with a suite of engineered features, primarily focusing on vehicular dynamics and driving patterns. The predictors encompassed in our model include:

- **Speed Metrics**: These were analysed across various percentiles – average, 85th, 90th, 95th, and 99th – to capture a broad range of speed behaviours.
- **Acceleration and Gyroscopic Features**: We incorporated both X-axis and Y-axis accelerations, as well as pitch, roll, and yaw measurements. Each of these features was evaluated at numerous percentiles (average, 85th, 90th, 95th, 99th, 15th, 10th, 5th, 1st), allowing us to gain insights into both typical and extreme vehicular movements.
- **Event Counts**: The model accounted for the frequency of harsh deceleration events (acceleration < -0.4g), harsh acceleration events (acceleration > 0.4g), and harsh swerving events (swerving > 0.3g).

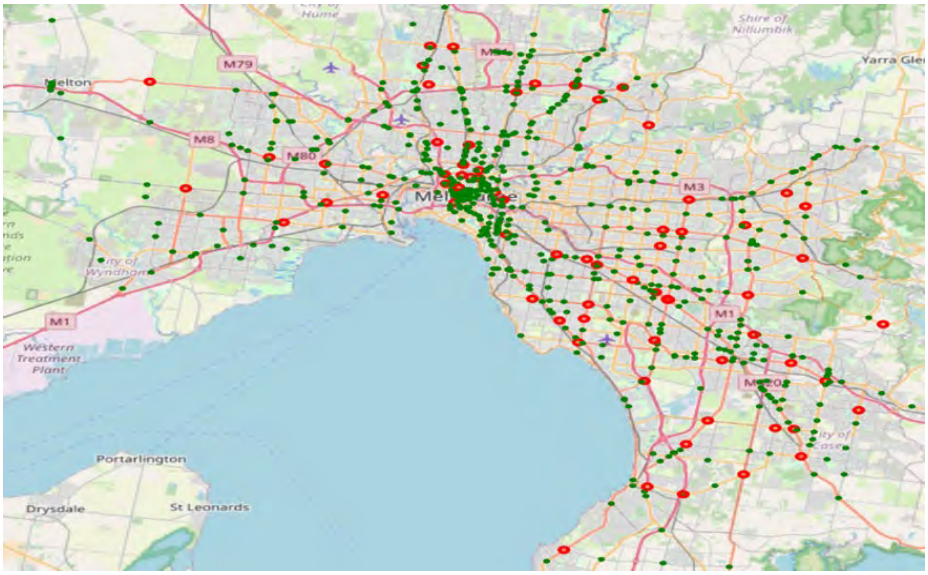



Figure 3.20: Hotspots in The Greater Melbourne Area from 2015 to 2020

We employed a baseline Poisson regression approach within a 100m x 100m grid setting and expanded to other models as illustrated in Table 3.3, see interim report for more details. In this comparative study, conducted within a 250m x 250m grid setting, the Poisson regression model emerged as the most efficacious, evidenced by an R-squared value of 0.9989. This result underscores the suitability of the telematics data to be used as surrogate measures for detecting risky zones and predicting crash risk areas.

**Table 3.3: Comparative summary of performance across different models**

| Grid Size | Model | R-squared | | | AIC | | |
|---|---|---|---|---|---|---|---|
| | | Baseline | With Feature Selection | Stepwise | Baseline | With Feature Selection | Stepwise |
| 100 * 100 meters | Linear Regression | 0.4743 | 0.4390 | 0.4738 | 256846.93 | 258490.68 | 256828.16 |
| | Poisson Regression | 0.9321 | 0.4390 | N/A | 215909.33 | 240191.17 | N/A |
| | Negative-Binomial Regression | 0.3902 | 0.3488 | N/A | 182226.68 | 184924.30 | N/A |
| 200 * 200 meters | Linear Regression | 0.6345 | 0.5996 | 0.6330 | 183668.74 | 184878.10 | 183681.85 |
| | Poisson Regression | 0.9943 | 0.5996 | N/A | 180477.22 | 208783.25 | N/A |
| | Negative-Binomial Regression | 0.5036 | 0.4640 | N/A | 133397.99 | 135409.55 | N/A |
| 250 * 250 meters | Linear Regression | 0.7160 | 0.6858 | 0.7136 | 157534.53 | 158509.06 | 157574.93 |
| | Poisson Regression | 0.9989 | 0.6858 | N/A | 166488.49 | 194071.53 | N/A |
| | Negative-Binomial Regression | 0.5458 | 0.5029 | N/A | 116061.03 | 118010.40 | N/A |

## 3.5 Summary of key findings

### Vehicle position data

1. Telematics data provides critical insights into vehicle dynamics and driver behaviour, particularly through metrics like acceleration and gyroscopic measurements.
2. CV datasets are for limited geographic coverage but store detailed vehicle movement. For example, the CompassIOT dataset includes 38,393,085 car trajectories, 2,174,375 heavy commercial vehicle (HCV) trajectories, and 712,695 light commercial vehicle (LCV) trajectories.
3. CompassIOT covers larger networks for longer periods, with a total of 41,280,155 trajectories, including 1,674,829 car trajectories with acceleration data and 1,032,798 HCV trajectories with acceleration data.
4. It also has higher penetration and coverage, making it valuable for traffic safety applications. The dataset includes 1,095,603 trajectories with complete gyroscopic data (X Acc, Y Acc, Roll, Yaw, Pitch).
5. CAVI data, on the other hand, provides smooth and sufficient data points for individual trajectories due to its higher frequency.

### Penetration measures

1. The penetration rate measures the proportion of vehicles equipped with GPS devices. Higher penetration rates imply broader data representativeness, aiding to more effective traffic management and control.
2. Four methods tested, with Methods 3 and 4 showing superior performance. Method 3: Expanded polygons from entry routes to nearest exit routes, achieving high penetration rate and accuracy. Method 4: Used homogenous traffic flow polygons and VKT data for accurate penetration rate calculation.
3. Average penetration rate across SCATS sites: approximately 2%. Lowest penetration rate: 1.38% at SCATS site 4489 (Westgate Freeway and Power Street). Highest penetration rate: 7.51% at SCATS site 2880 (Westgate Freeway/Williamstown). Penetration rate along Hoddle Street: 2.2% to 4.8%.

4. Consistent month-over-month increase in penetration rates. Example: SCATS Site 3380 (Hoddle Street/Johnston Street) penetration rate increased from ~2% in January 2022 to 5% by December 2022.
5. Higher penetration rates observed in affluent suburbs (e.g., Toorak, Balwyn) with rates above 1.75%. Lower rates typically below 1% in other areas.
6. Growing adoption of telematics systems indicated by increasing penetration rates. Socio-economic factors appear to influence the adoption of advanced vehicular technologies, as evidenced by higher penetration rates in affluent areas.

### Speed profiles at intersection

1. Driver Behaviour at Intersections:
    a. Speed-space profiles of vehicles approaching the intersection reveal a distinct 'V' shape pattern.
    b. This pattern indicates a typical deceleration as vehicles approach the intersection, followed by acceleration upon exiting.
2. Identification of Driving Behaviours:
    a. Cluster 0: Drivers decelerate upon approaching the intersection, resulting in low speed before 200 meters, then speed up after exiting the intersection.
    b. Cluster 1: Drivers approach with variable speeds but maintain a lower average speed (green or yellow light) and then accelerate after exiting, indicating free flow downstream of intersection.
    c. Cluster 2: Drivers exhibit consistently lower speeds both while approaching and exiting, likely due to downstream congestion or navigating through complex intersections.

### Crash prediction modelling

1. Models Used: Linear Regression, Poisson Regression, and Negative Binomial Regression were employed to analyse predictors of road accidents.
2. Key Predictors: Important predictors included vehicle speed, acceleration characteristics, angular parameters (pitch, roll, yaw), and counts of harsh driving events (e.g., harsh deceleration, acceleration, and swerving).
3. Incorporated recent telematics data from CompassIOT (2022) and historical crash data (2000-2020) to quantify crash risks and identify key predictors of road accidents.
4. Hotspot Analysis: Circular plot analysis from 2015 to 2020 supported the assumption that accident hotspots remain consistent over time, aiding predictive accuracy.
5. Poisson Regression Performance: The Poisson regression model within a 250m x 250m grid setting emerged as the most effective, with an R-squared value of 0.9989, indicating its high suitability for capturing crash data nuances.

# 4 Reconstructing unobserved trajectories

## 4.1 Background and Methodology

Vehicle trajectories provide crucial insights into traffic patterns, aiding in various transport applications such as traffic condition estimation, flow modelling, signal optimisation, and emission estimation. However, capturing complete vehicle trajectories using traditional methods like video cameras or fixed sensors is often challenging due to high costs and limited coverage.

To address these limitations, researchers have focused on reconstructing vehicle trajectories from partial data provided by existing sensors. This report builds on the work of Mehran et al. (2012) by exploring the integration of fixed and probe sensor data on signalised urban arterials to improve the accuracy of trajectory reconstruction and enhance the understanding of traffic dynamics in urban environments.

The primary goal is to apply a data fusion framework to reconstruct connected vehicle trajectories on signalised urban arterials by:

1. **Data Integration**: Combining information from fixed sensors and probe vehicles to enrich the dataset for more accurate trajectory reconstruction.
2. **Traffic Engineering Application**: Using fundamental traffic engineering principles, including kinematic wave theory, to analyse shockwaves, queue propagation, and speed variations, and validate the real-world applicability of these models.
3. **Handling Vehicle Entry/Exit**: Managing the complexities of vehicles entering and exiting specific road segments by applying a segment-by-segment approach.
4. **Validation**: Ensuring the robustness of the methodology through validation against real-world data.
5. **Use Cases**: Exploring practical applications of the reconstructed trajectories, such as travel time estimation, signal coordination, and emission monitoring.

The methodology for reconstructing vehicle trajectories on signalised urban arterials is grounded in the kinematic wave theory, specifically the Lighthill-Whitham-Richards (LWR) model. This model describes the relationship between traffic flow, density, and time, providing a robust framework for analysing traffic dynamics. The following steps outline the detailed approach used in this study:

1. **Parameter Initialisation**:
    a. **Forward Wave Speed (u)**: This parameter represents the speed at which traffic flows forward. It is crucial to set this parameter accurately, as it directly influences the estimation of vehicle trajectories.
    b. **Backward Wave Speed (w)**: This speed reflects how quickly traffic disruptions, like congestion, move backward through the traffic stream. Accurate estimation of backward wave speed is essential for identifying shockwaves and understanding queue dynamics.
    c. **Jam Density (kj)**: This is the maximum density of vehicles that a road segment can accommodate. Setting the correct jam density is vital for predicting traffic conditions under congestion.
    d. **Maximum Flow Rate (qmax)**: The highest volume of traffic that a road can handle per unit of time. This parameter is crucial for understanding the capacity of the road and is used to optimise the reconstruction process.
    e. **Horizontal Distance (tstep) and Vertical Distance (sstep)**: These define the spatial and temporal resolution of the time-space matrix, ensuring accurate representation of vehicle movement over time and space.

2. **Matrix Initialisation**:
   A time-space matrix is created to represent the cumulative number of vehicles at each point in time and space. The matrix is populated with data from various sources:
   a. **Traffic Volume Data**: This data populates the initial state of the matrix, providing a baseline for vehicle counts at different points along the road.
   b. **Signal Timing Data**: Incorporating the timing of traffic signals helps to simulate how vehicles are influenced by red and green lights, crucial for realistic trajectory reconstruction, see Figure 4.1. The number of rows and columns are derived from tstep and sstep. Each array represents the cumulative number of vehicles passing through the time and space. if we are studying the trajectories passing a street of length 1000 meters and within 30 minutes, and if we set the tstep equal to 1, then we the matrix will have 1800 columns representing the time steps and the number of rows will be calculated by dividing 1000 by sstep.
   c. **Probe Trajectories**: Data from connected vehicles is integrated into the matrix, enhancing the accuracy of the reconstructed trajectories. These trajectories provide reference points that guide the optimisation process.

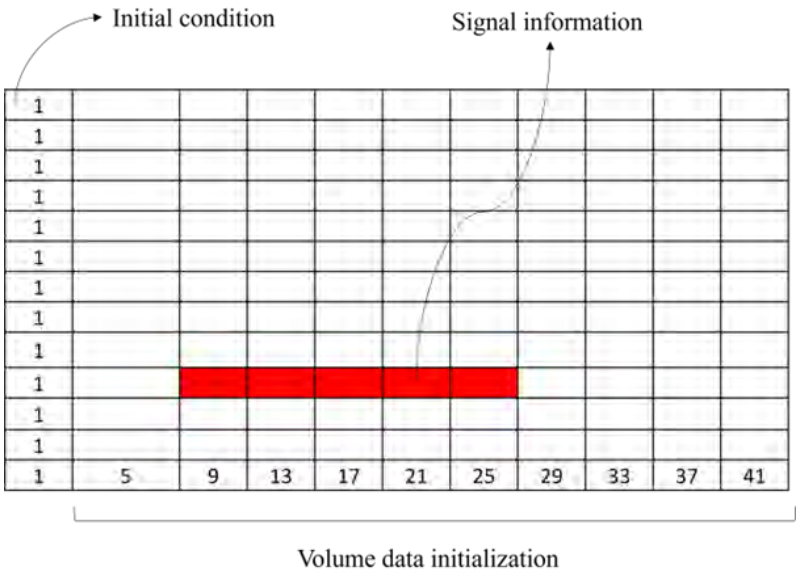



Figure 4.1: Matrix initialisation example

3. Data Interpolation:
   To align the available data with the matrix, interpolation is performed. This step is essential to ensure that the data fits well within the matrix structure, particularly when the frequency of data collection varies. This process refines the data points, improving the accuracy of trajectory mapping. Figure 4.2 represents an example of the interpolation and reference trajectories mapping on the matrix.

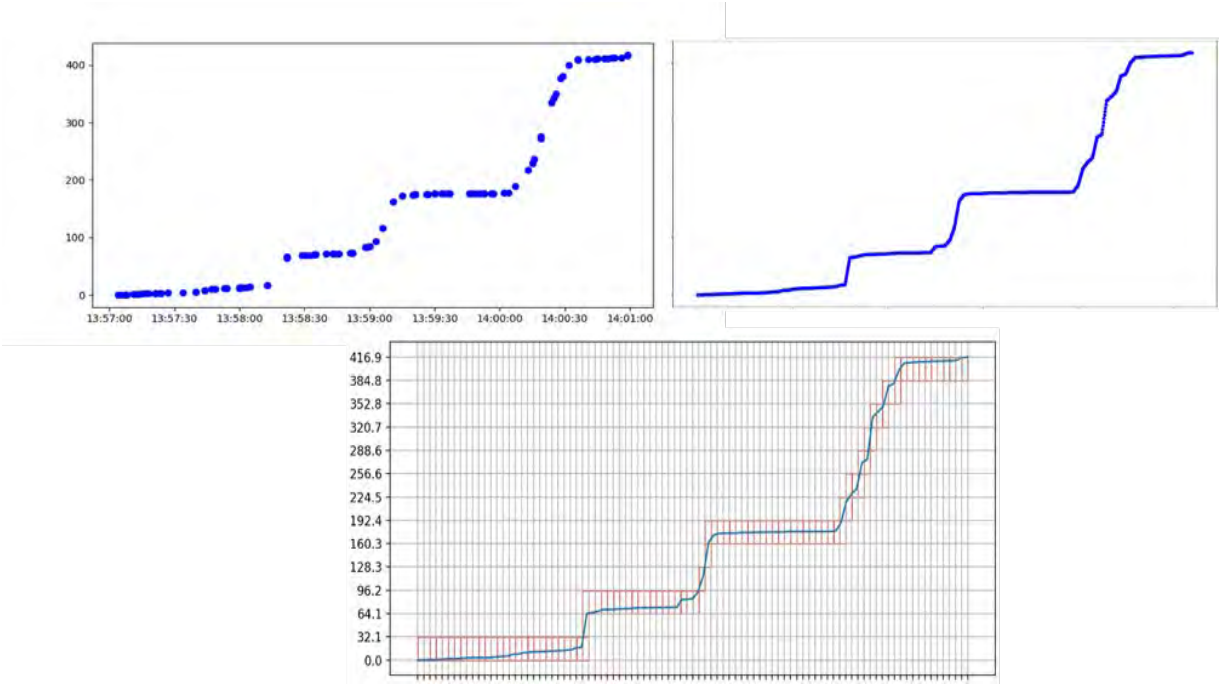



Figure 4.2: Trajectories interpolation and mapping example

4. Optimisation:
   The optimisation process aims to determine the values of empty arrays in the matrix using principles from variational theory. The objective is to find the shortest path for cumulative vehicle numbers, which is achieved by adjusting the values in the matrix to match real-world observations. The optimisation accounts for red light intervals, ensuring that the reconstructed trajectories accurately reflect stop-and-go conditions. Figure. 4.3 represents an example optimisation process without the use of reference trajectories.

5. Trajectory Reconstruction:
   After optimisation, the cumulative values are used to generate contour lines representing individual vehicle trajectories. These trajectories provide a detailed map of vehicle movement, capturing the dynamics of traffic flow across the studied road segments.

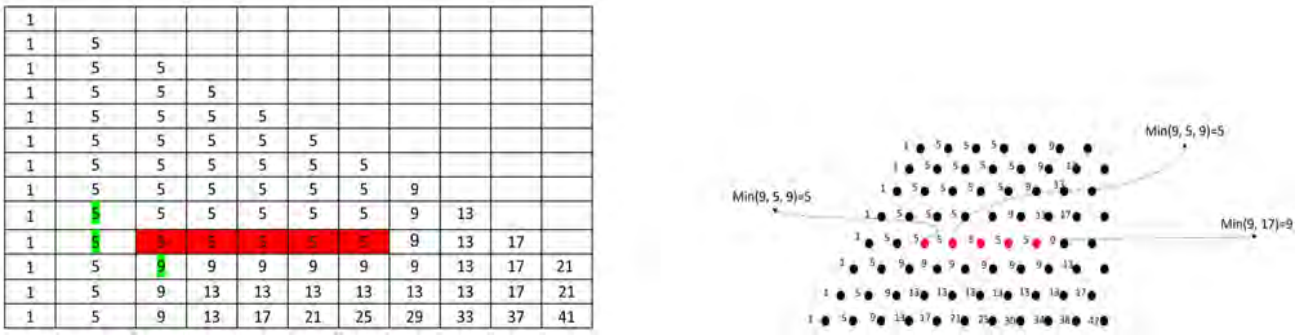



Figure 4.3: Matrix-based and the grid-based optimisation process as an illustration in the absence of reference trajectories.

## 4.2 The implementation

The implementation of the trajectory reconstruction methodology is a multi-step process that integrates various datasets and computational techniques to accurately reconstruct vehicle trajectories on signalised urban arterials. The following outlines the key procedures and steps involved in the implementation:

## Data Collection and Preparation

1. **Connected Vehicle Trajectories (CompassIOT Data)**:

    a. **Source**: Real-world connected vehicle data is collected from CompassIOT, which provides telematics data for vehicles travelling within the study area.

    b. **Scope**: The data includes GPS coordinates, timestamps, and vehicle IDs, which are essential for mapping the actual trajectories of vehicles on the selected road segments.

2. **Signal Timing Data**:

    a. **Source**: Traffic signal data is obtained from the traffic light control system from SCATS, detailing the timing and phases of signals at various intersections.

    b. **Processing**: The raw signal data is processed to extract binary vectors representing green (1) and red (0) intervals for each signalised intersection, refer to Figure 4.4. This data is crucial for understanding how traffic signals impact vehicle movements and for synchronising trajectory reconstruction with actual signal timings. Figure 4.5 provides the outcome we need as a vector of binary values, where 1s represents the green and 0s represents the red time.

3. **Traffic Volume Data**:

    a. **Source**: Traffic volume data is gathered from loop detectors or other traffic count devices installed at key points along the study area. This data is typically recorded at 15-minute intervals.

    b. **Application**: The traffic volume data is used to initialise the time-space matrix, setting the cumulative vehicle counts at specific locations and times as the starting point for the reconstruction process.

| | | | |
|---|---|---|---|
| 121591 | 3381 | Wednesday, 28 September 2022, 8:00:15 AM AEST | Phase demand: D=On |
| 121592 | 3381 | Wednesday, 28 September 2022, 8:01:16 AM AEST | Flags: Flags (1)=[XSF=[1=Off]] |
| 121593 | 3381 | Wednesday, 28 September 2022, 8:01:31 AM AEST | Phase status flags: Max Due=On |
| 121594 | 3381 | Wednesday, 28 September 2022, 8:01:31 AM AEST | Phase termination request: Next phase=B |
| 121595 | 3381 | Wednesday, 28 September 2022, 8:01:31 AM AEST | Phase termination request confirmation: Current phase=A |
| 121596 | 3381 | Wednesday, 28 September 2022, 8:01:32 AM AEST | Phase status flags: Mx Ack=On |
| 121597 | 3381 | Wednesday, 28 September 2022, 8:01:32 AM AEST | Phase termination request: Next phase=B |
| 121598 | 3381 | Wednesday, 28 September 2022, 8:01:32 AM AEST | Phase demand: A=On |
| 121599 | 3381 | Wednesday, 28 September 2022, 8:01:32 AM AEST | Signal group: 1=Not green, 2=Not green |
| 121600 | 3381 | Wednesday, 28 September 2022, 8:01:33 AM AEST | Phase interval: Phase interval=Yellow |
| 121601 | 3381 | Wednesday, 28 September 2022, 8:01:38 AM AEST | Phase interval: Phase interval=All red |
| 121602 | 3381 | Wednesday, 28 September 2022, 8:01:38 AM AEST | Phase demand: D=Off, B=Off |
| 121603 | 3381 | Wednesday, 28 September 2022, 8:01:38 AM AEST | Signal group: 4=Green, 6=Green, 7=Green, 9=Green |
| 121604 | 3381 | Wednesday, 28 September 2022, 8:01:38 AM AEST | Flags: Flags (1)=[MSS=[11=Off, 12=On]] |
| 121605 | 3381 | Wednesday, 28 September 2022, 8:01:39 AM AEST | Phase termination: Terminated phase=A, MX=0, GT=122, CG=71 |
| 121606 | 3381 | Wednesday, 28 September 2022, 8:01:39 AM AEST | Alarm timer: Timer value=0 |
| 121607 | 3381 | Wednesday, 28 September 2022, 8:01:39 AM AEST | Current phase: Current phase=B, Flags=[0] |
| 121608 | 3381 | Wednesday, 28 September 2022, 8:01:39 AM AEST | Phase interval: Phase interval=Late start |
| 121609 | 3381 | Wednesday, 28 September 2022, 8:01:39 AM AEST | Phase status flags: Mx Ack=Off, Max Due=Off, Stretch=Off |
| 121610 | 3381 | Wednesday, 28 September 2022, 8:01:39 AM AEST | Signal group: 5=Green |
| 121611 | 3381 | Wednesday, 28 September 2022, 8:01:39 AM AEST | Pedestrian movement (Region 6.9.4+): Ped 2=[Demand=Off, Interval=Walk], Ped 4=[Demand=Off, Interval=Walk] |

Figure 4.4: A sample of signal timing raw data format.

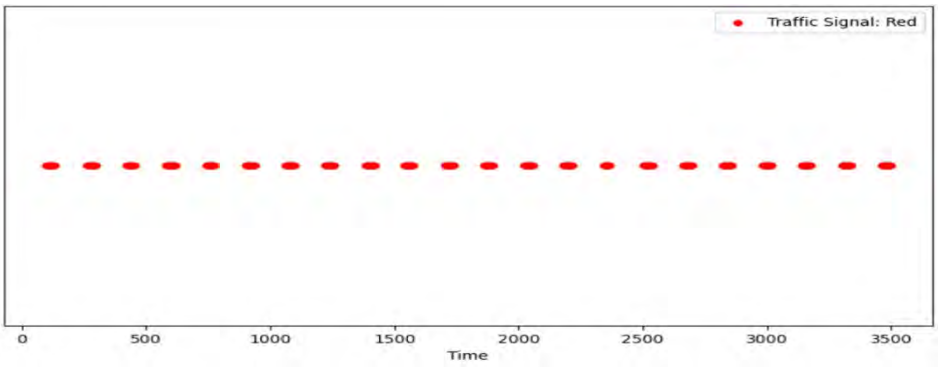



Figure 4.5: Illustration of signal timing vector used for the matrix initialisation.

### Segment-Based Implementation

Given the complexities of urban traffic, especially with vehicles entering and exiting different segments, the implementation is performed on a segment-by-segment basis, see Figure 4.6. This approach enhances accuracy and computational efficiency.

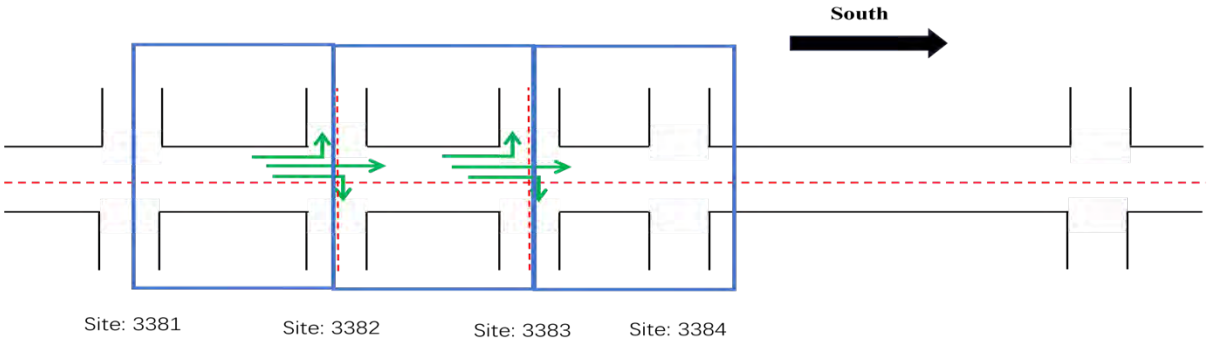



Fig 4.6: Segment implementation. The count values are used for the upstream initialisation

1. **Segment Selection**:
    - The road network is divided into smaller segments, each defined by a pair of SCATS (Sydney Coordinated Adaptive Traffic System) IDs. For instance, one segment might span from SCATS ID 3382 to SCATS ID 3383, covering a specific section of the road between two intersections.
    - **Example**: The study focuses on Hoddle Street in Melbourne, where segments such as (3382, 3383) are analysed individually (Figure 4.7).
2. **Upstream Initialisation**:
    - For each segment, the traffic volume data from the downstream SCATS ID (e.g., 3383) is used to initialise the upstream values (e.g., 3382). This strategy helps to account for incoming and outgoing vehicles by setting the initial conditions based on the observed traffic at the downstream point.
3. **Matrix Initialisation**:
    - A time-space matrix is created for each segment, where the rows represent spatial locations along the segment and the columns represent discrete time intervals. The matrix is populated with:

- **Cumulative Vehicle Counts**: Derived from traffic volume data, this sets the initial values in the matrix.
- **Signal Timing Vectors**: Integrated to reflect the impact of traffic signals on vehicle movements within the segment.
- **Probe Trajectories**: Real vehicle trajectories are mapped onto the matrix to guide the reconstruction process, see Figure 4.8.

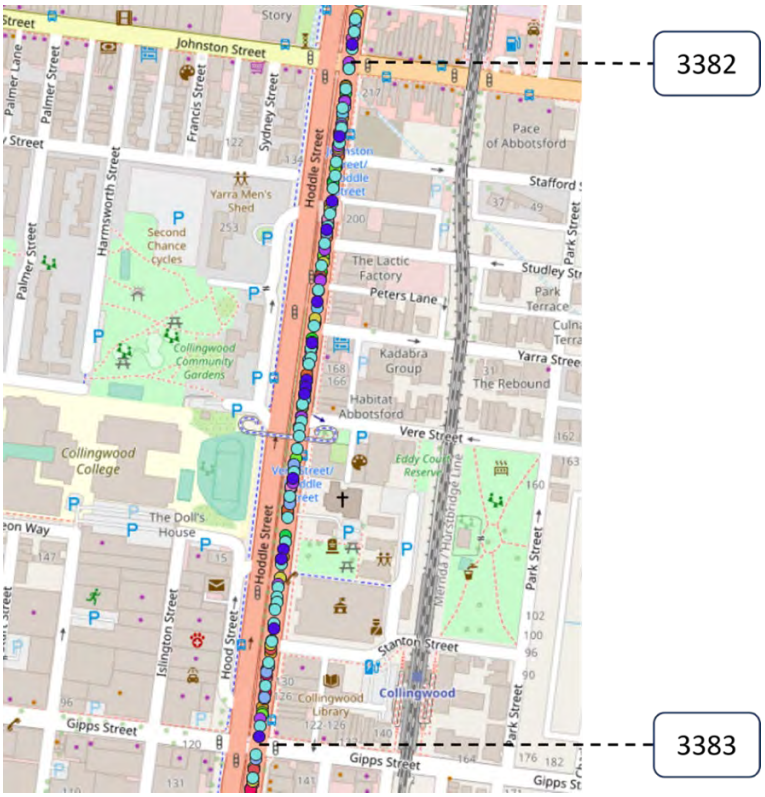



Figure 4.7: The connected vehicles trajectories for the section pair id (3382, 3383) for one hour of Wednesday, 28 September 2022, from 8:00:00 AM AEST to 8:59:59 AM AEST.

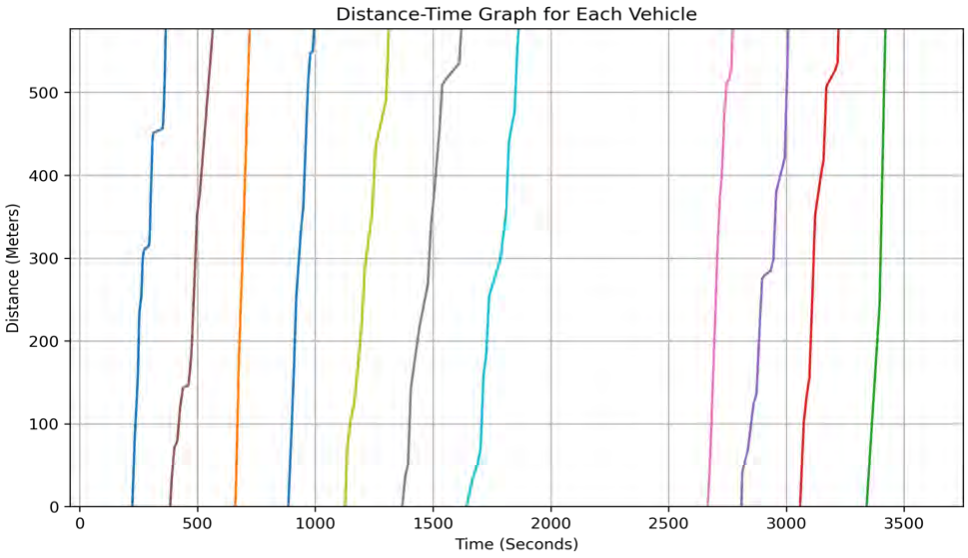



Figure 4.8: The space-time plot of processed trajectories for section pair (3382, 3383) Street for one hour of Wednesday, 28 September 2022, from 8:00:00 AM AEST to 8:59:59 AM AEST.

### Case Study Example: Hoddle Street, Melbourne

The methodology was applied to a real-world case study on Hoddle Street in Melbourne, Australia. Key details include:

1. **Study Period**:
    - The analysis was conducted for a one-hour period on Wednesday, 28 September 2022, from 8:00 AM to 8:59 AM AEST.
2. **Data Integration**:
    - **Connected Vehicles**: 17 connected vehicle trajectories were extracted and used in the analysis.
    - **Signal Timing**: The signal data for intersections along Hoddle Street was processed and integrated into the matrix for each segment.
    - **Traffic Volume**: Count data for vehicles passing through each SCATS ID was used to initialise the matrix.
3. **Segment Analysis**:
    - Each segment along Hoddle Street was analysed individually, with the matrix-based methodology applied to reconstruct the trajectories of vehicles within each segment.
    - **Outcome**: The reconstructed trajectories were visualised as space-time plots, showing the movement of vehicles through each segment over the study period. This visualisation highlighted key traffic dynamics, such as congestion patterns and the impact of traffic signals (Figures 4.7 and 4.8).

### Final Visualisation and Interpretation

1. **Space-Time Plots**:
    - The final output of the reconstruction process is a set of space-time plots that depict the reconstructed trajectories of vehicles across the studied segments. These plots allow for a detailed analysis of traffic flow, highlighting areas of congestion, delays at traffic signals, and overall traffic dynamics within the urban arterial (Figures 4.9 and 4.10).
2. **Application Potential**:
    - The reconstructed data can be used for various practical applications, such as improving traffic signal coordination, estimating travel times, and monitoring emissions, providing valuable insights for traffic management and urban planning.

The methodology has significant potential for improving traffic management, particularly in optimising signal timings. Future work should focus on expanding the model to handle multi-lane roads and refining parameter calibration to enhance model accuracy across different traffic scenarios.

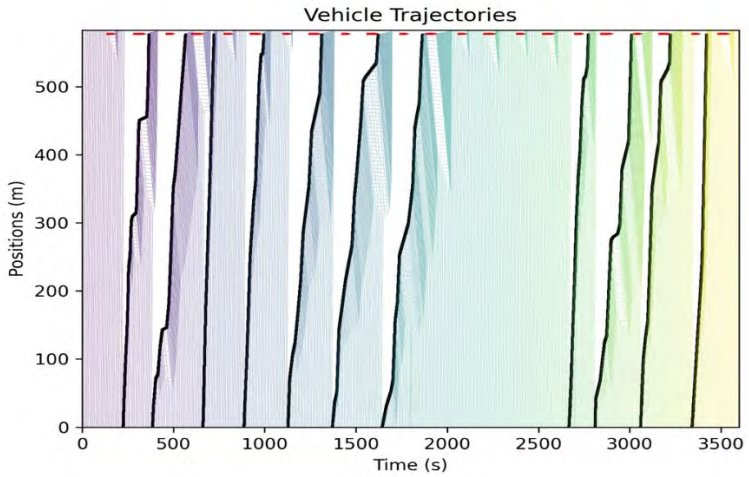



Figure 4.9: The space-time plot of the reconstructed trajectories for the targeted segment with qmax = 1000 and kj = 100. The black lines depict the actual trajectory data, while the coloured lines represent the reconstructed trajectories.

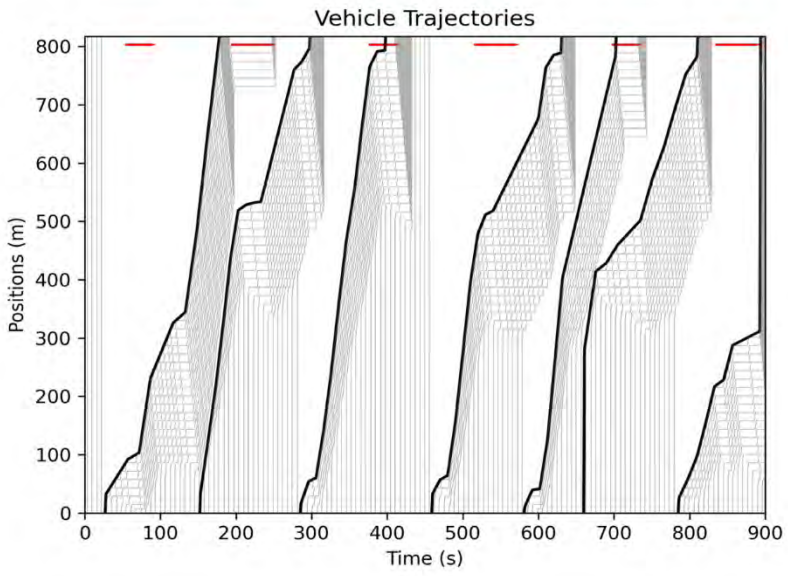



Figure 4.10: The space-time plot of the reconstructed trajectories for the targeted segment.

## 4.3 Summary of key findings

1. The study successfully demonstrated the reconstruction of vehicle trajectories using a data fusion framework combining fixed sensor data and connected vehicle data.
2. The methodology is based on kinematic wave theory, which provides a robust framework for understanding traffic dynamics in urban settings.
3. The integration of signal timing data and traffic volume data with probe vehicle trajectories significantly improved the accuracy of the reconstructed trajectories.
4. The methodology was validated using real-world data from Hoddle Street, Melbourne, confirming its applicability in complex urban environments.
5. Potential applications of the reconstructed trajectories include improved travel time estimation, signal coordination, and emission monitoring, see Section 10.
6. Further research should focus on extending the methodology to multi-lane roads and refining parameter calibration to enhance the model's performance across various traffic conditions.

# 5 Intersection management: traffic efficiency and multi-modality

This section comprises two key parts: Part A focuses on modelling traffic efficiency (cars-only), while Part B addresses occupancy-based fair and equitable priority at signalised intersections. Both methodologies are extensively elaborated in the interim report; however, some key findings are reviewed here.

## 5.1 Improving traffic efficiency at signalised intersections

The optimisation of intersection performance, particularly in terms of travel time reliability and delay, is crucial for effective traffic management. Utilising new data sources such as connected vehicles data, several key aspects such as queue length estimation (Y. Liu et al., 2021; Y. Zhao et al., 2019), delay minimisation (Liang et al., 2023), and public transport priority (He et al., 2012; Park et al., 2023) are considered to enhance intersection optimisation. In this study, we employed deep reinforcement learning (DRL) based signal control to assess the potential of connected vehicles data combined with traditional loop detector data for optimising traffic signals at intersections.

Unlike traditional rule-based systems that primarily rely on vehicle volume and gap time between vehicles for signal optimisation, our study designed a traffic signal that leverages high-resolution connected vehicles data and incorporates embedded intelligence and automation in AI algorithms. DRL methods, capable of automatically extracting features such as queue length from raw data and learning tasks without prior knowledge through state, action, and reward mechanisms, enable the learning of optimal policy and production of improved results. Consequently, these methods are widely utilised in traffic signal control studies.

In summary, we developed and implemented deep reinforcement-based traffic signal control by leveraging multi-source datasets (connected vehicles, loop detector, and public transport demand and operational data) to achieve the following objectives:

- To test and examine possible improvements in signal optimisation through high-resolution connected vehicles location data.
- To evaluate the effectiveness of advanced AI models compared to conventional control systems using both traditional sensor data and connected vehicles data integration.
- To assess the impact of different ranges of traffic data on existing optimisation models.
- To ascertain how data-rich deep learning models can benefit from emerging data sources like connected vehicles data in modelling and controlling occupancy-based priority control at intersections and the extent to which it enhances traffic efficiency and throughput compared to traditional actuated control systems (in Section 5.2).

Three detection technologies are tested in this section as follows:

1. Loop Detectors: Figure 5.1 presents an incoming link of an intersection. Two loop detectors are placed within this link, one at the entrance and the other at stop bar. By having the number of entering and exiting volumes collected from these detectors, the total volume of the link is calculated.
2. Video Cameras: The video cameras can detect the vehicles and provide the volume or density data. With advanced computer vision techniques and the large number of video frames, the queue length data can also be extracted.
3. Connected Vehicles Location Data: The connected vehicle data when fused with loop detectors data can provide all vehicles trajectory information, which include important information on the number of vehicles in the queue and in the residual queue (vehicles stuck behind the red light more than one cycle). In addition to this information, the data can cover the centre of the intersection (i.e., cross-box area). This is also a valuable information as the turning vehicles experience additional delays when seeking a gap to complete their movements. This data cannot be collected from loop detectors and traditional cameras that only monitor the area behind the stop bars.

In this study, trajectory reconstruction methods was employed to provide complete trajectories for training DRL models. The information from above-mentioned sensing technologies is used for the RL's state and, the calculation of reward value during action execution.

Figure 5.2 illustrates the case study for the intersection of interest with 1-hour of simulation duration. The demands are varied across the 1-hour period to ensure the complexity and dynamics of the demands in our experiments. We used Double Deep Q-Network (DDQN) algorithm as a popular deep RL method in the literature. The proposed models are presented in Table 5.1.

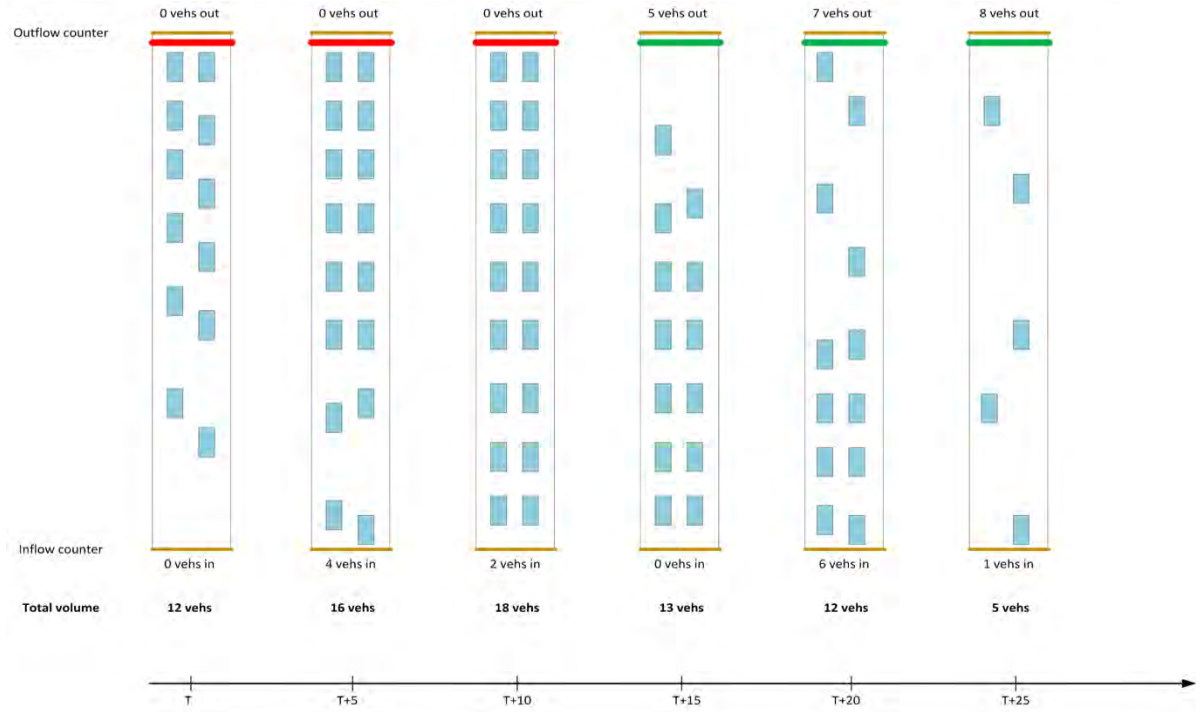



Figure 5.1: Calculation of total volume from inflow and outflow detectors

The first model is Deep Reinforcement Learning Adaptive Traffic Signal Control In-Out Flow (DRL-ATSC InOutFlow) which uses the data collected from inflow/outflow loop detectors. The second model "DRL-ATSC Density R0" is fed with volume collected data from video cameras. It is noted that the volume data in this model is directly collected from video cameras while the volume data in "DRL-ATSC InOutFlow" model is collected from two loop detectors at link entrance and stop bar.

The last two models "DRL-ATSC Density R2" and "DRL-ATSC Queue Density R2" take advantage of CV data to use additional information. The difference is that the former model counts the residual volumes while the latter counts the residual delay/queue data. It is noted that the queue length includes the vehicles with speeds less than 5 km/hr and Density represents the number of vehicles in the area of interest. For all DRL traffic signal models, the actions are 5 different traffic signal times ranging from 10 seconds to 30 seconds with 5 seconds intervals. To provide safe movements, a 5 second protected signal is provided for right turning vehicles after terminating a signal phase.

The DRL models are also evaluated with non-learning baselines as follows:

- Actuated: based on gap times and occupancies
- Advanced actuated: actuated with additional parameters (headway and waste)


| M1 | N | E | S | W |
|---|---|---|---|---|
| N | 0 | 25 | 210 | 25 |
| E | 25 | 0 | 20 | 160 |
| S | 210 | 25 | 0 | 25 |
| W | 25 | 160 | 25 | 0 |

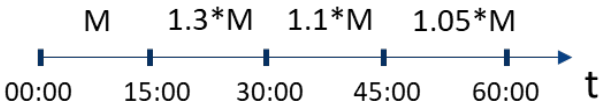



Figure 5.2: Demand variation for 1 hour simulation. Figure on the right and bottom shows volume for different directions, and demand variation across 1-hour period, respectively.

Table 5.1: Descriptions of the proposed model versus baselines

| Model | Reward | Cross-Box Area + Residual Queue data | Sensing technology |
|---|---|---|---|
| DRL-ATSC InOutFlow | Minimise the summation of volumes present in the control volume (within 100 meters from stop bar) calculated from inflow and outflow | N/A | Loop detectors |
| DRL-ATSC Density R0 | Minimise summation of vehicles present in the control volume (i.e., the whole length of the links) | No | Video Camera |
| DRL-ATSC Density R2 | | Yes | Connected Vehicles |
| DRL-ATSC Queue Density R2 | Minimise summation of queue lengths | Yes | Connected Vehicles |

The results demonstrate that the "DRL-ATSC Queue Density R2" model, utilising connected vehicles data, achieves optimal policy learning and outperforms other DRL models in terms of average travel time and queue length reduction, as demonstrated in Table 5.2 and Figure 5.3. On the other hand, the "DRL-ATSC Density R2" model, which lacks queue length data and relies solely on volume data, exhibits reduced training performance; however, it still learns the policy and performs as effectively as the "DRL-ATSC Queue Density R2" model in decreasing delays. Additionally, reducing the data features input to the DRL model results in underperformance compared to the best model, which utilises more features from high-resolution connected vehicles data.

Table 5.2: Performance metrics results of first 50 training episodes with 95% confidence interval (trained models)

| Model \ PMs | Travel time (s) | Queue length (m) |
|---|---|---|
| Actuated | 49.2 | 25.27 |
| Advanced Actuated | 42.55 | 18.06 |
| DRL-ATSC InOutFlow | 42.59 | 17.21 |
| DRL-ATSC Density R0 | 43.3 | 17.69 |
| DRL-ATSC Density R2 | 40.48 | 16.12 |
| DRL-ATSC Queue Density R2 | 40.04 | 15.84 |

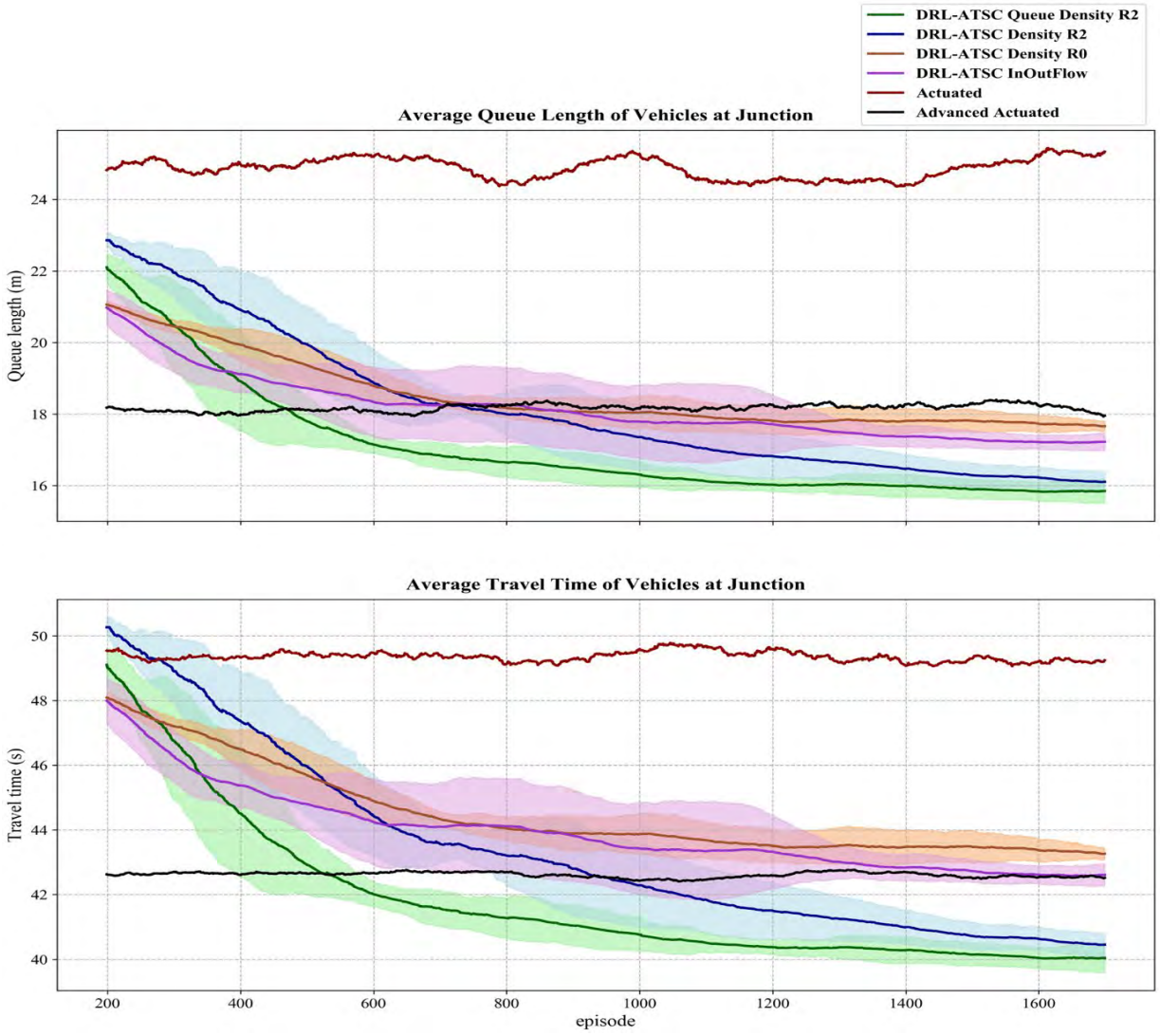



Figure 5.3: Training plot displaying the performance of all models across 1700 episodes, with results presented as the simple moving average over 200 episodes.

Furthermore, it has been found that, when the data is limited to loop detectors only, the DRL model can still learn the policy, but both travel times and queue lengths marginally deteriorate compared to DRL models that use high-resolution data from connected vehicles. Importantly, the inclusion of fine resolution data from connected vehicles, including lane-level volume and residual queue information, facilitates faster learning and more accurate and effective signal control. While all the deep reinforcement learning models significantly outperformed rule-based actuated signal controls, DRL models without in-depth fine resolution connected vehicles data inputs slightly improved average queue length and exhibited similar performance in travel time compared to advanced actuated signal control. This highlights the potential

benefits of fine resolution connected vehicles data for improving data hungry AI models. Our study demonstrates how, with the increasing penetration of connected vehicles, AI and deep learning models can be greatly enhanced in the future.

## 5.2 Occupancy-based priority at signalised intersections

To evaluate the potential of connected vehicles data and advanced machine and deep learning in providing equitable and efficient green time for all modes, we implemented two occupancy-based signal priority deep reinforcement models and tested them in two scenarios, comparing their performance to Actuated Transit Priority Control (ATPC). The first scenario used synthetic occupancy and public transport operational data, while the second scenario incorporated real data from tram operations in Melbourne alongside connected vehicles data and loop detector data. For each scenario, two different deep reinforcement learning models were assessed, one including residual queue information and the other without this additional data from connected vehicles.

Our case study includes a single isolated intersection within AIMES testbed as shown in Figure 5.4. There are two tram stops, one for the Northbound and the other for Southbound direction. It is noted that the tram stops are located at the downstream of the intersection for both directions. Figure 5.5 illustrates the phasing structure of the case study intersection. Each tram line has its own tram signal that allocate an extended green time for the corresponding direction.

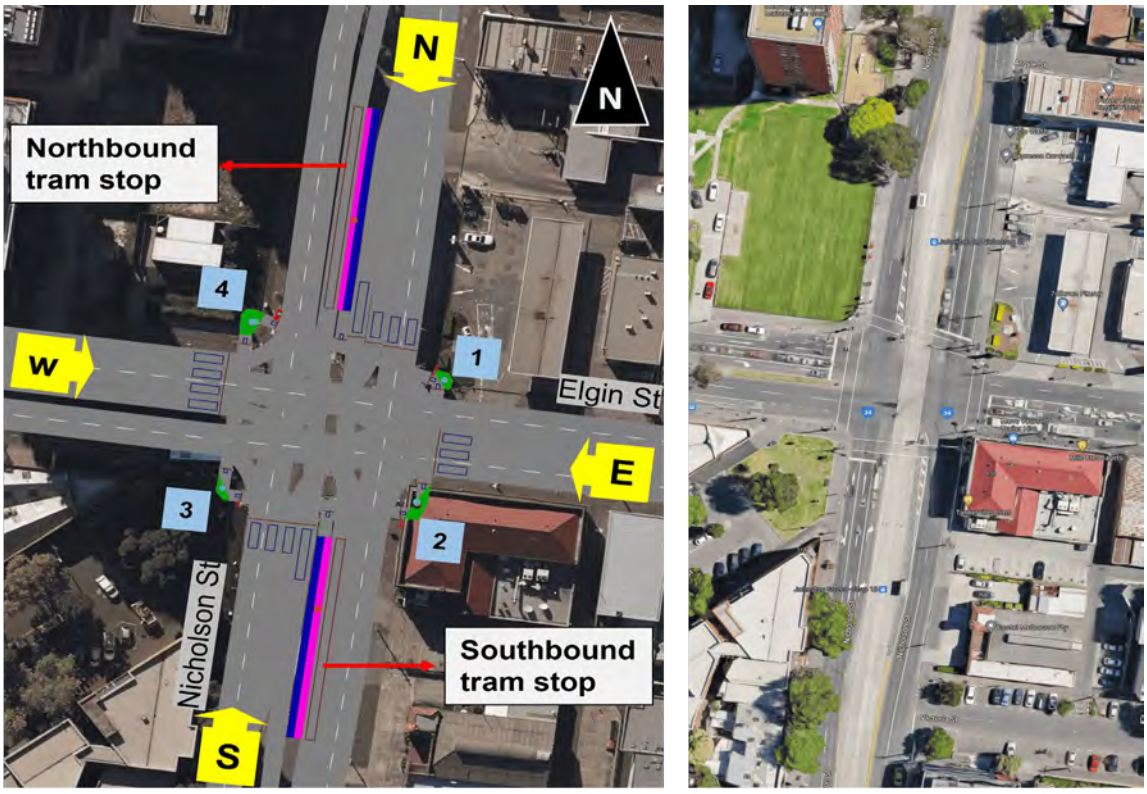



Figure 5.4: Case study on Elgin/Nicholson Intersection

The key components of the proposed DRL model are as follows:

- State: The individual information of vehicles and trams such as speed and location data are fed to deep RL model. For each tram direction, the information of tram occupancy is also included.
- Action: The deep RL method has 6 actions, extending the green time from 0 to 30 seconds for activated phase direction.
- Reward: The vehicle waiting times are calculated by counting the number of vehicles with speed less than 0.02 km/hr in every step of the action direction. For trams, the speed threshold is 5 km/hr to penalise the stopping of trams behind the red light. It is noted that the residual waiting time is

also included for both vehicles and trams. Residual queue occurs when a vehicle/tram get stuck behind the red light for the second time. Hence, a reward values (i.e., waiting time) of the vehicle/tram are the accumulated waiting times behind the red light.

Our proposed model is named Intelligent Multi-modal Signal Priority (IMSP). The IMSP-R2 is the model with extended reward functions (i.e., cross-box data and residual queue) while IMSP-R0 lacks the additional components used in the IMSP-R2. The IMSP models are compared with actuated TSP signals named Adaptive Transit Signal Priority (ATSP) that gives certain level of priorities between 0 to 12 seconds with 3 seconds intervals (e.g., Ttsp=0,3,6,9,12sec) based on the operator need.

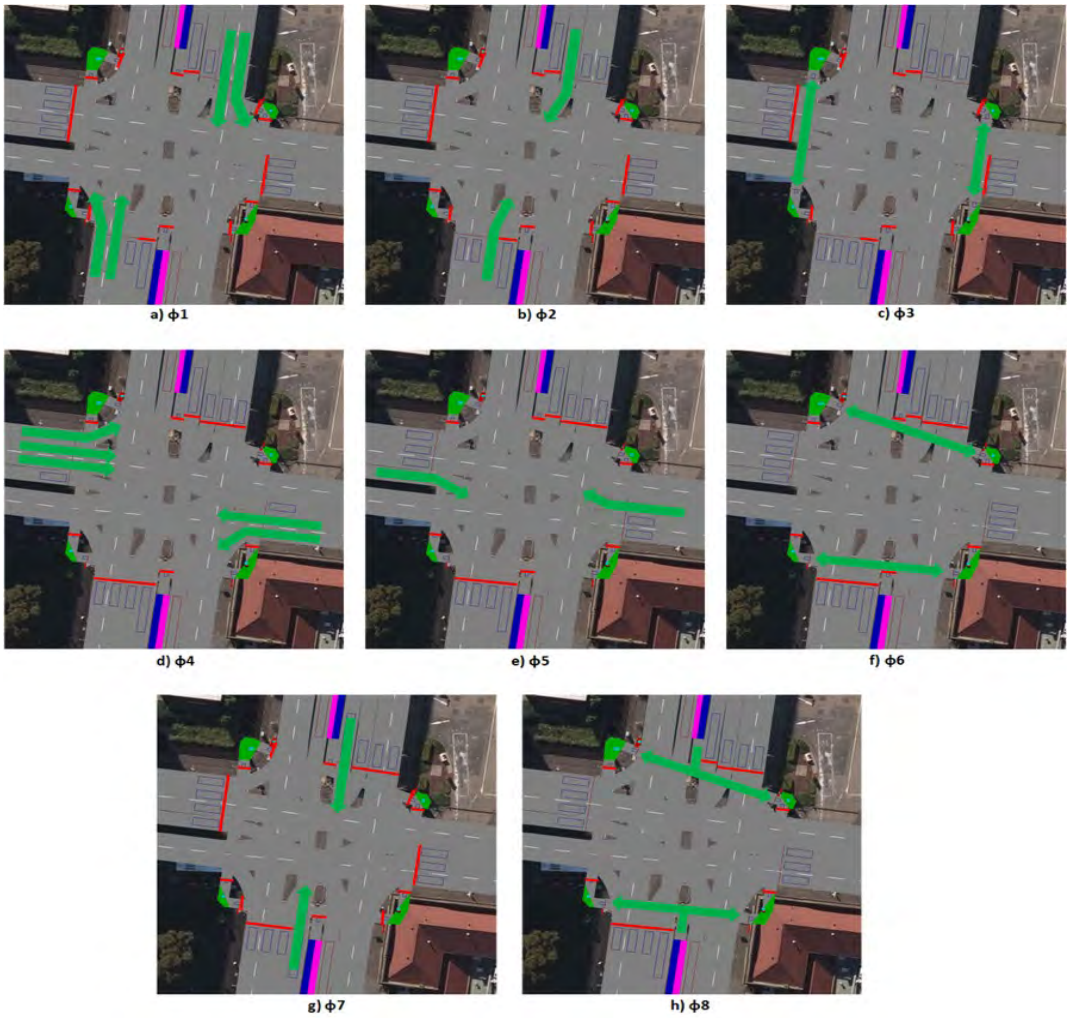



Figure 5.5: Phasing structure of the case study intersection

The proposed model is tested under two scenarios. In the first scenario, a synthetic data is used for the occupancy of trams in three levels 1) low=20, 2) medium=60 and 3) high=100 passengers on-board, assuming a uniform 3-minutes headway for all trams. The second scenario is tested based on the real occupancy and headway data (Table 5.3) to expose the model to more complex real-world scenario. For both scenarios, the vehicles demand is a rush hour scenario.

The evaluation of results is based on following performance metrics:

- Average travel time of vehicles
- Average travel time of trams
- Average user speeds: average speed of vehicle and tram occupants (all users)

Table 5.3: Tram occupancy and headway. A real data collected from the intersection.

| Tram line | Arrival time | Load | Alight | Board |
|---|---|---|---|---|
| NB | 08:07:12 | 38 | 5 | 9 |
| | 08:10:30 | 57 | 7 | 3 |
| | 08:18:06 | 23 | 6 | 10 |
| | 08:21:30 | 44 | 12 | 3 |
| | 08:28:54 | 42 | 6 | 7 |
| | 08:34:24 | 77 | 6 | 7 |
| | 08:39:54 | 42 | 3 | 3 |
| | 08:47:06 | 24 | 4 | 4 |
| | 08:53:00 | 33 | 6 | 6 |
| | 08:58:12 | 38 | 3 | 3 |
| SB | 08:01:54 | 55 | 2 | 0 |
| | 08:03:18 | 33 | 2 | 0 |
| | 08:14:06 | 32 | 2 | 0 |
| | 08:29:12 | 55 | 1 | 0 |
| | 08:31:00 | 28 | 2 | 0 |
| | 08:39:36 | 38 | 2 | 3 |
| | 08:49:00 | 47 | 2 | 1 |
| | 08:51:30 | 28 | 2 | 0 |
| | 08:59:06 | 31 | 4 | 2 |

The findings demonstrate that DRL with queue information data from connected vehicles substantially enhances the average speed for all occupancy levels, as illustrated in Figure 5.6. While the DRL model without additional CV data features can match the performance of the former at lower occupancy levels, it fails to achieve the same level of superiority at higher occupancy levels, particularly with 100 passengers onboard. This emphasises the potential improvement in occupancy-based priorities using advanced deep learning models and leveraging additional high-resolution data from connected vehicles, especially for high onboard loads typical during rush hours.

In the second scenario, the proposed deep reinforcement learning, leveraging additional connected vehicles data, outperforms all benchmarks in every performance metric by a relatively larger margin (13%) compared to ATSP (Figure 5.7). Notably, DRL models using CV data without additional features like residual queue still marginally exceed the benchmarks. This clearly indicates the promising potential of employing connected vehicles data to optimise traffic signals and allocate fairer and more efficient priorities, rather than consistently prioritising public transport while transferring all delays to car users. As expected, DRL models using CV data slightly increase the travel time of public transport users, but the overall average travel time and speed for all users are significantly improved (Figure 5.8). This suggests that the DRL can sense the number of onboard occupancy and prioritise public transport only when necessary.

Overall, this study has revealed that using high spatial and temporal resolution data from connected vehicles can be utilised by data-hungry machine and deep learning models to benefit the comprehensiveness of CV data and the intelligence of AI models, advancing signal control to be more sophisticated, intelligent, and fair, and contributing to efficient, safe, and sustainable intersection management practices.

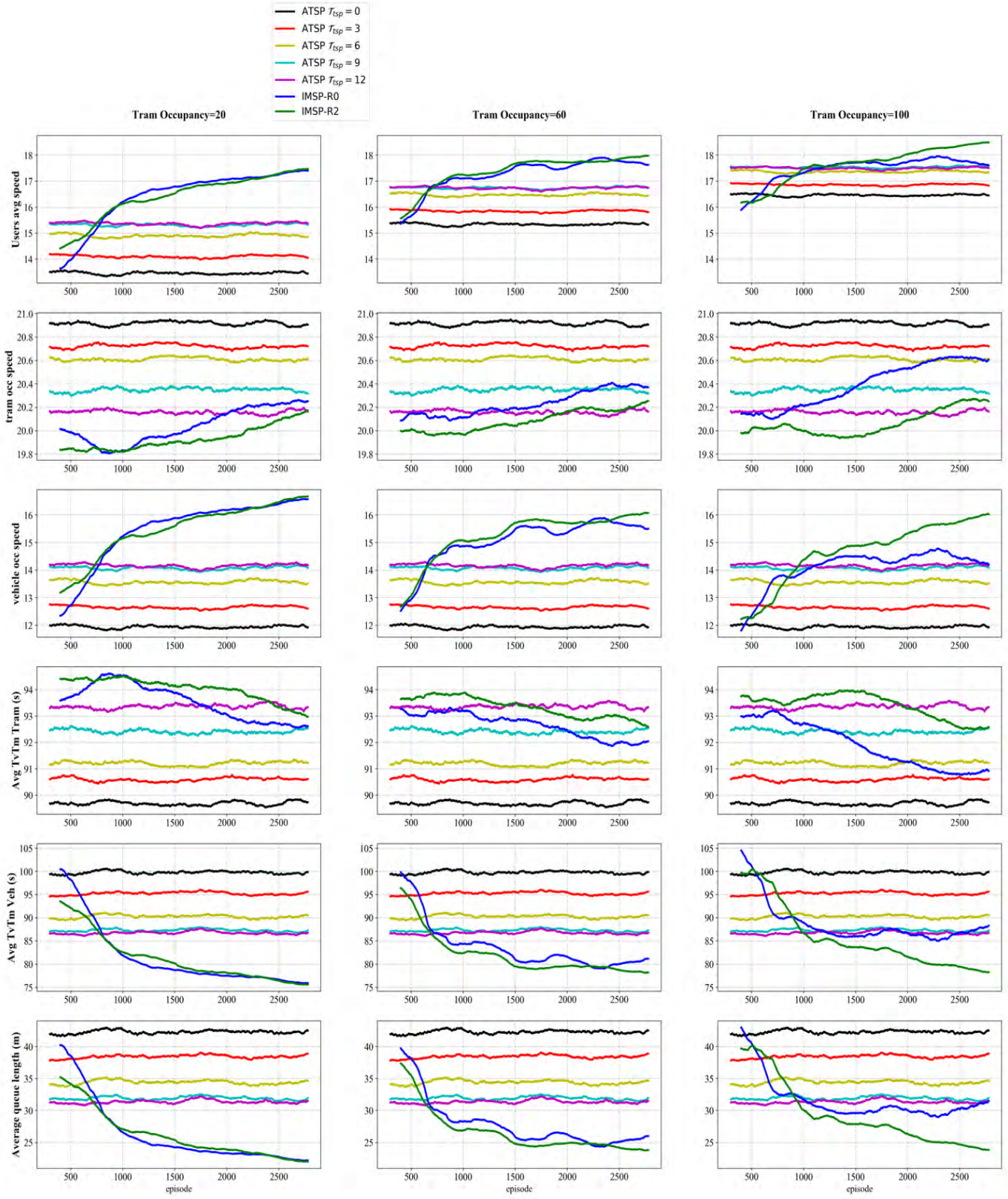



Figure 5.6: Training plots for scenario 1

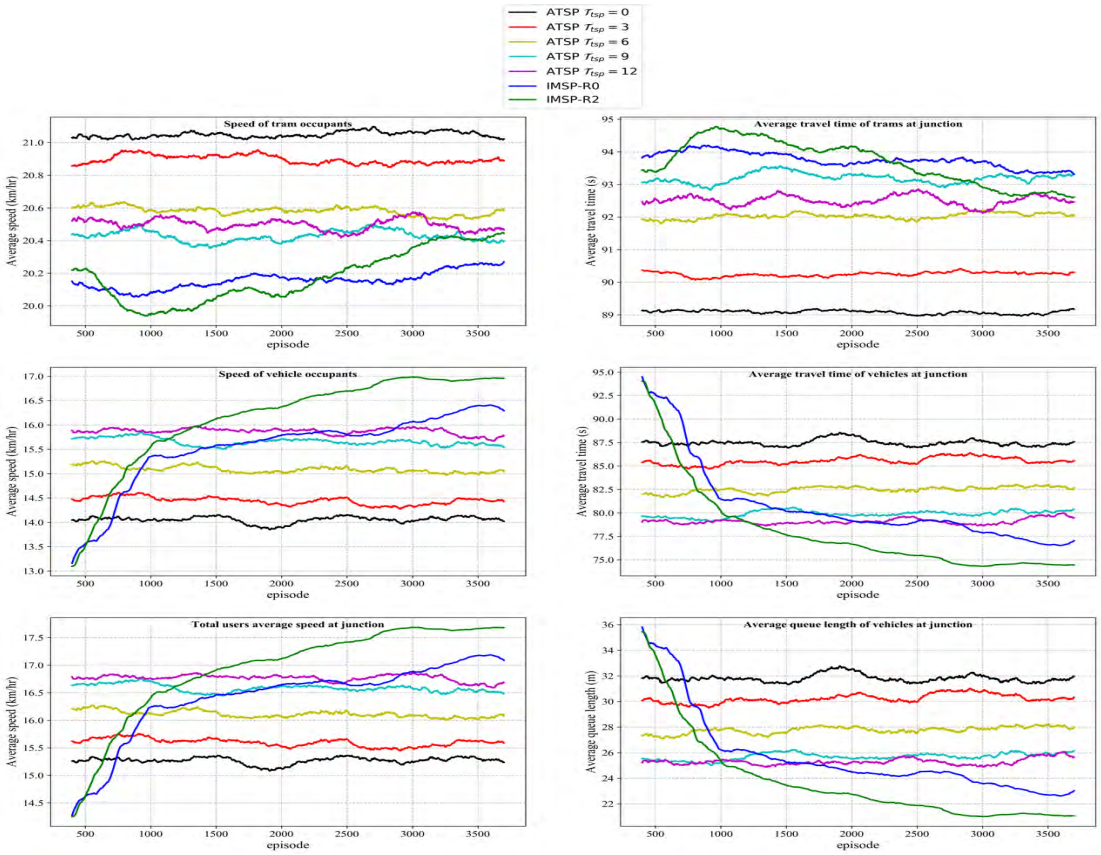



Figure 5.7: Training plots for scenario 2

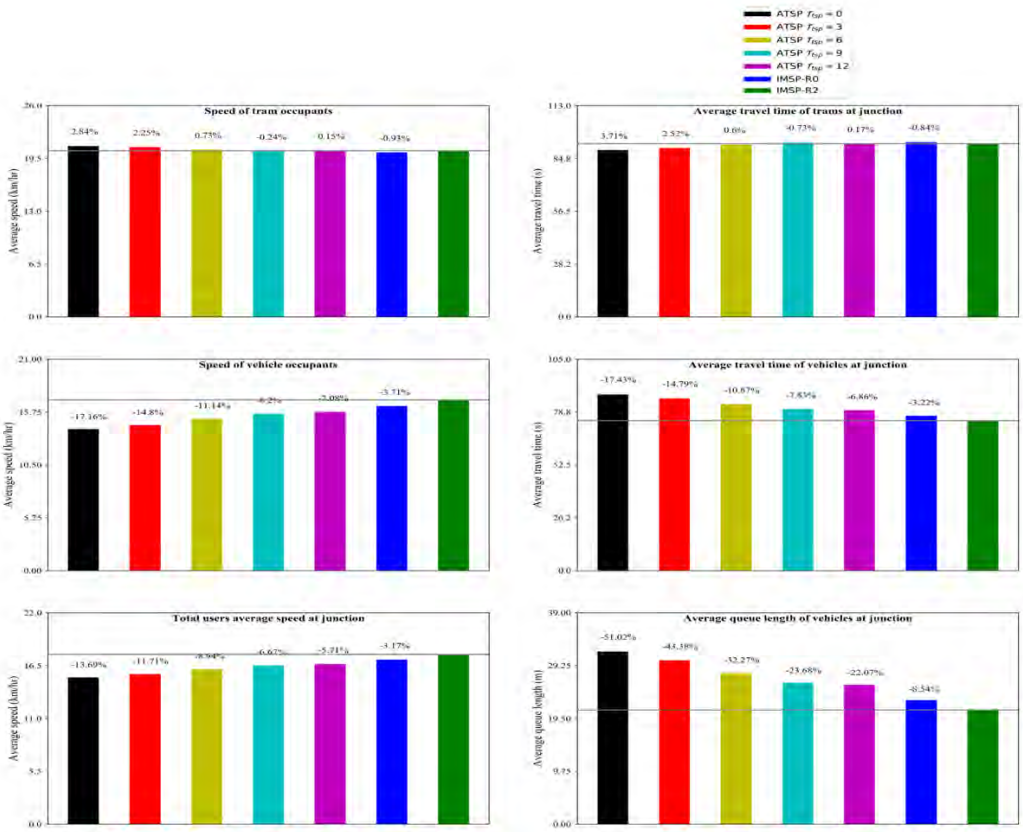



Figure 5.8: Result of performance metrics for scenario 2

## 5.3 Summary of key findings

### Traffic efficiency at intersection

1. Utilising connected vehicle data and deep reinforcement learning (DRL) based signal control can improve traffic signal control and reduce travel times and queue lengths by up to 10-15%.
2. The DRL-ATSC Queue Density R2: model, which uses connected vehicle data and residual queue length information, outperforms other DRL models in terms of average travel time and queue length reduction.
3. The DRL-ATSC Density R2: Without queue length data and relies solely on volume data, exhibits reduced training performance but still learns the policy and performs effectively in decreasing delays.
4. DRL models without high-resolution connected vehicle data inputs slightly improve average queue length and exhibit similar performance in travel time compared to advanced actuated signal control.
5. All DRL models outperformed rule-based actuated signal controls. However, when data is limited to loop detectors only, the DRL model can still learn the policy, but both travel times and queue lengths marginally deteriorate compared to DRL models that use high-resolution data from connected vehicles.
6. Overall, inclusion of fine-resolution data from connected vehicles facilitates faster policy learning and more optimal signal control.

### Occupancy-based priority control at intersection

1. Occupancy-based priority control at intersections can be improved using connected vehicle data and DRL, enabling fair and equitable priority allocation.
2. Implemented two occupancy-based signal priority using deep reinforcement learning:
    a. Intelligent Multi-modal Signal Priority (IMSP), with IMSP-R2 including extended reward functions (cross-box data, residual queue) and
    b. IMSP-R0 without these components and compared them to Actuated Transit Priority Control (ATPC) in two scenarios.
3. DRL with queue information from connected vehicles (IMSP-R2) significantly enhances average speed for all occupancy levels (5-10%).
4. DRL models without additional CV data features (IMSP-R0) perform well at lower occupancy levels but not at higher levels (i.e., 100 passengers).
5. In the second scenario, DRL with connected vehicles data outperforms all benchmarks by a larger margin (13%) compared to ATSP.
6. DRL models using CV data, even without additional features like residual queue, marginally exceed benchmarks.
7. DRL models slightly increase travel time for public transport users but significantly improve overall average travel time and speed for all users.
8. Overall, high-resolution data from connected vehicles enhances DRL models, leading to more sophisticated, intelligent, and fair signal control, contributing to efficient, safe, and sustainable intersection management practices.

# 6 Micromobility management: Bicycle safety and efficiency

## 6.1 Bicycle safety

Bicycling has experienced significant growth in recent years as an environmentally friendly, affordable, and healthy mode of transport. Globally, bicycle usage has surged, with cities investing in cycling infrastructure and promoting bike-sharing schemes (Pucher & Buehler, 2017). However, this growth in bicycling has also highlighted concerns over cyclist safety, particularly on shared roads with motorised vehicles. According to the World Health Organization WHO (2018), among 1.35 million road traffic deaths, more than half are vulnerable road users including bicyclists, pedestrians, and motorcyclists. These alarming figures underscore the urgent need for effective measures to enhance cyclist safety and prevent such tragedies.

A fundamental step towards bicycle safety is understanding unsafe parts of the infrastructure and risky behaviours displayed on the roads. GPS data collected by users' smartphones or diagnostics devices installed on bicycles can provide valuable insights. A combined approach, such as the technology developed by See.sense bicycle lights, which relies on smartphone GPS and onboard sensors, has significant potential due to its low cost and reasonable coverage.

By analysing harsh braking events, abnormal manoeuvres, or near-miss situations experienced by cyclists, researchers can pinpoint high-risk locations, behaviours, or road conditions that contribute to bicycle accidents. This proactive approach allows for the identification of dangerous locations or issues with the current built environment and geometric design, contrasting with the traditional reactive approach. By identifying surrogate safety measures, we can assess and improve site safety without waiting for accidents to occur and cause harm (Strauss et al., 2017).

This study utilises large-scale bicycle telematic data provided by SeeSense, collected from users in Melbourne metropolitan area from 2021 to 2024. The dataset, along with actual crash data from traffic police reports, is used to develop a model that estimates the correlation between harsh braking and abnormal events derived from GPS and gyroscope data with actual crash incidents.

### *6.1.1 Method and data*

A comprehensive framework is developed to estimate the correlation between abnormal events and harsh-braking events recorded from bicycle telematic data with actual bicycle crashes recorded from historical crash data (Figure 6.1).

- Phase 1: Exploratory Analysis
    - Identify patterns and distributions of bicycle crashes, abnormal events, and road surface conditions.
    - Focus on understanding the frequency and locations of crashes and abnormal events.
    - Assess road surface conditions using proxy measures from gyroscope sensor readings.
- Phase 2: Road Network Integration
    - Integrate road network data with bicycle telematic data and historical crash data.
    - Examine how road geometry influences crash patterns and abnormal events.
    - Classify road geometries into intersections and non-intersections using a rule-based algorithm.
- Phase 3: Correlation Analysis
    - Analyse the correlation between identified abnormal events and historical crash data.
    - Distinguish between intersections and non-intersections to provide insights into specific types of road geometries prone to crashes and abnormal events.

This methodological framework offers a robust approach to analysing bicycle crash patterns and road surface conditions. By leveraging multiple data sources and employing a detailed classification of road geometries, the framework provides a comprehensive understanding of the factors contributing to bicycle crashes and abnormal events. This evidence-based approach is instrumental in developing data-driven strategies for improving road safety for cyclists.

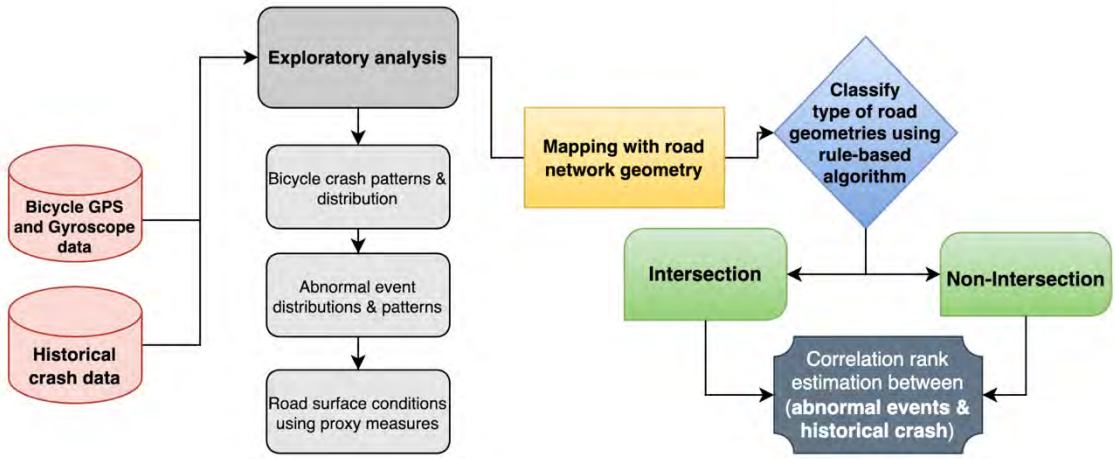



Figure 6.1: Methodological framework pipeline for analysing the correlation between abnormal events from GPS data and historical crashes

**Bicycle GPS and Gyroscope data**

This study utilises GPS and gyroscope data collected by SeeSense (SeeSense, 2024), a company specialising in cycling technology aimed at enhancing cyclist safety and the overall cycling experience. Their innovative approach integrates advanced sensor technology into their flagship products, such as the SeeSense ICON and ACE lights. These smart bicycle lights detect various riding conditions, including road surface quality, harsh braking, and swerving, while also serving as data collection devices.

The data is collected from thousands of public bikers in the Melbourne metropolitan area of Australia who agreed to provide their movement data. The dataset spans three years, from January 1, 2021, to May 1, 2024, and comprises over 20 billion sensor readings and 2 million rows of data per aggregated unit. The data is processed and aggregated per 3x3 meter grids to preserve user privacy while retaining essential information at acceptable resolutions.

The dataset consists of two types:

- Normal Data: Focuses on average cycling dynamics, offering comprehensive details about patterns, road conditions, and behaviour aggregated per spatial unit. It includes features such as latitude and longitude coordinates, average number of cyclists per location, road surface quality, and normalised values for braking and swerving, see Figure 6.2.
- Extreme Data: Captures behaviours considered abnormal for each cyclist, indicating deviations from the normal range, which may correlate with crashes or incidents. Each row represents the number of abnormal events per aggregated 3x3 meter spatial unit, see Figure 6.3.

The extreme dataset captures behaviours considered abnormal for each cyclist. According to SeeSense, patented technology used in bicycle lights profile each cyclist and the bike they ride, determining the normal range of gyroscopic information for each individual. Therefore, the score indicates deviations from this normal range, considered as extreme cases, which may correlate with some crashes or incidents. In total, there are 15 thousand extreme locations with different number of abnormal events per location (3x3 metre grid).

| | lat | lng | average_cyclists | total_datapoints | average_speed_kmh | average_xswerve | average_ysurface | average_zbrake | average_stopped_delay | percent_cyclists_stopped |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -38.009451 | 145.085991 | 23 | 6910.0 | 6.55 | 0.61 | 2.41 | 2.46 | 0 | 16.7 |
| 1 | -37.887943 | 145.132944 | 33 | 27350.0 | 23.08 | 0.96 | 1.66 | 3.57 | 0 | 0.0 |
| 2 | -38.178682 | 144.569724 | 17 | 3850.0 | 37.94 | 2.89 | 4.26 | 9.45 | 0 | 0.0 |
| 3 | -37.798443 | 145.145876 | 33 | 7420.0 | 25.66 | 1.12 | 2.08 | 3.93 | 0 | 0.0 |
| 4 | -38.320147 | 144.335710 | 58 | 25170.0 | 29.87 | 1.67 | 1.47 | 3.23 | 0 | 0.0 |
| 5 | -37.876727 | 145.395021 | 24 | 4650.0 | 11.47 | 0.65 | 2.15 | 2.37 | 0 | 0.0 |
| 6 | -37.828169 | 144.894906 | 15 | 2820.0 | 26.49 | 0.68 | 1.64 | 2.79 | 0 | 0.0 |
| 7 | -38.233217 | 144.297826 | 13 | 1960.0 | 25.47 | 0.58 | 1.58 | 7.26 | 0 | 0.0 |
| 8 | -37.897263 | 145.214025 | 28 | 10120.0 | 14.91 | 0.50 | 1.78 | 3.24 | 0 | 4.3 |
| 9 | -37.970341 | 145.037053 | 14 | 9080.0 | 24.42 | 0.83 | 2.14 | 2.11 | 0 | 0.0 |

Figure 6.2: SeeSense normal data collected from cyclists' smartphone GPS and smart bicycle lights for gyroscopic information, illustrated in a 3x3 metre aggregated format

| | lat | lng | total_abnormal_events | max_swerve_value | max_surface_value | max_brake_value |
|---|---|---|---|---|---|---|
| 0 | -37.794779 | 144.994017 | 8.0 | 34 | 35 | 130 |
| 1 | -37.767817 | 145.199858 | 4.0 | 81 | 17 | 40 |
| 2 | -37.746061 | 144.958118 | 6.0 | 83 | 19 | 78 |
| 3 | -37.770034 | 144.899360 | 7.0 | 80 | 21 | 4 |
| 4 | -37.811619 | 144.983338 | 6.0 | 1 | 0 | 142 |
| 5 | -37.787420 | 145.007217 | 7.0 | 132 | 94 | 3 |
| 6 | -38.163513 | 144.337803 | 2.0 | 0 | 0 | 114 |
| 7 | -38.271813 | 144.629823 | 7.0 | 79 | 30 | 22 |
| 8 | -37.865769 | 144.973411 | 12.0 | 135 | 112 | 0 |
| 9 | -37.749724 | 145.135251 | 10.0 | 159 | 85 | 142 |

Figure 6.3: SeeSense extreme events data collected from cyclists' smartphone GPS and smart bicycle lights for gyroscopic information, illustrated in a 3x3 metre aggregated format (Events outside normal ride profile identified by smart lights for each cyclist).

**Historical cyclist crash data**

Historical road crash data collected and maintained by traffic police and hospital reports are used in this study. The dataset, released by the Department of Transport and Planning in Victoria, Australia, covers records from 2000 to 2024 (DTP, 2024). It contains detailed information on road crashes, providing valuable insights for researchers, policymakers, and transportation authorities.

The Victorian Road Crash Data is organised into several resources, each containing specific information related to crash events, including:

1. Accident: Details about the crash itself, such as date, time, location, severity, and number of vehicles involved.
2. Accident Event: Information about the events leading up to the crash, including factors such as collision type, traffic control measures, and contributing circumstances.
3. Vehicle: Details about the vehicles involved, including type, year of manufacture, and potential defects or issues.
4. Person: Information about individuals involved, including drivers, passengers, and pedestrians, including age, gender, and injuries sustained.
5. Accident Location: Geographic information about the crash location, including road name, suburb, and GPS coordinates.
6. Atmospheric Condition and Road Surface Condition: Information about environmental conditions at the time of the crash, such as weather, lighting, and road surface characteristics.

The dataset is regularly updated and maintained open source, ensuring that researchers and authorities have access to the most current and accurate information available.

### *6.1.2 Exploratory analysis*

This section reports an exploratory analysis using GPS and gyroscope data from bicycles alongside actual crash data from traffic police reports. By integrating high-resolution, comprehensive data from both sources, we aim to gain insights into trends, patterns, distributions, and the severity of crashes, as well as the spatial correlation of abnormal events.

**Historical bicycle crash pattern**

The bar chart in Figure 6.4 illustrates that bicycle-involved crashes are more frequent at intersections (59%) compared to all crashes (41%). This finding aligns with existing literature, which reports that most cyclist crashes occur at intersections (Strauss et al., 2017)(Hirose et al., 2021). Interestingly, crashes involving two bicycles are more frequent on mid-block segments (61%) than at intersections (39%). This trend might be attributed to bike-bike crashes occurring on mid-block corridors or bikeways when cyclists traverse or cross paths, leading to accidents (Figure 6.5).

- **Distribution of Crashes by Severity**: The results show that the number of fatalities and serious injuries in bicycle-involved crashes is slightly lower than in all crashes, with percentages of 0.7% and 32.2% for bicycle-involved crashes compared to 1.6% and 36.4% for all crashes, respectively.
- **Spatial Distribution of Crashes:** Figure 6.6 shows an interesting pattern, with the majority of fatal crashes distributed on the outskirts of the city centre and the Central Business District (CBD). This distribution could be associated with higher speed limits in these outer areas compared to inner city neighbourhoods. Moreover, the majority of injuries fall between non-injury accidents and serious-injury accidents, with serious injury accidents slightly more prevalent in the outer parts of the city compared to the inner areas.
- **Crashes Involving Two or More Bicycles**: For crashes involving two or more bicycles, injuries typically fall between non-injuries and serious injuries (Figure 6.7). This is likely because such crashes often occur between two bicyclists. The spatial distribution of these crashes confirms this hypothesis, indicating that the majority of these incidents are concentrated around sea trails and locations with bike trails. A few fatal crashes are also observed in regional areas, which might be associated with interactions between cars and multiple bicyclists. This emphasises the need for careful design, management, and operation of bike-friendly locations with dedicated bike lanes and trails to reduce the number of crashes and incidents.

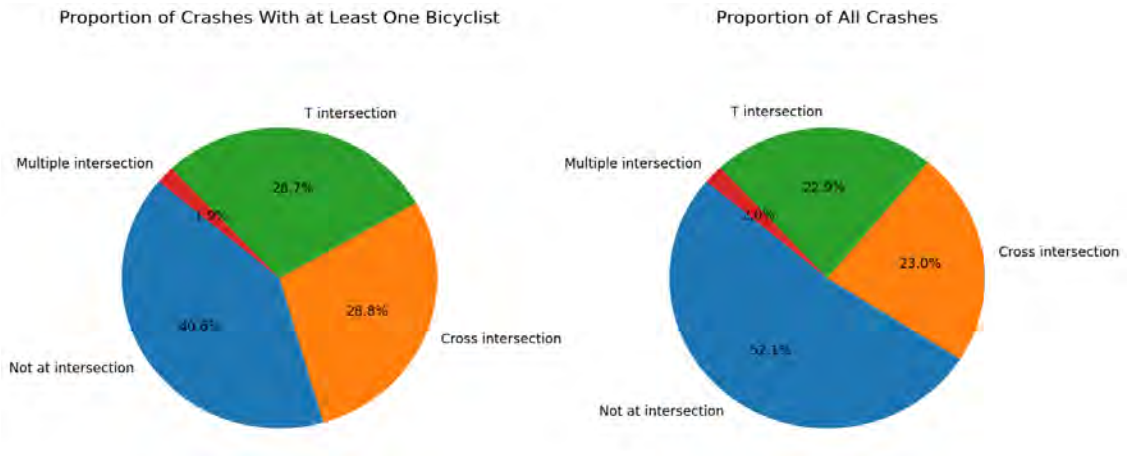



Figure 6.4: Distribution of bicyclist involved crashes versus all crashes at different road geometries

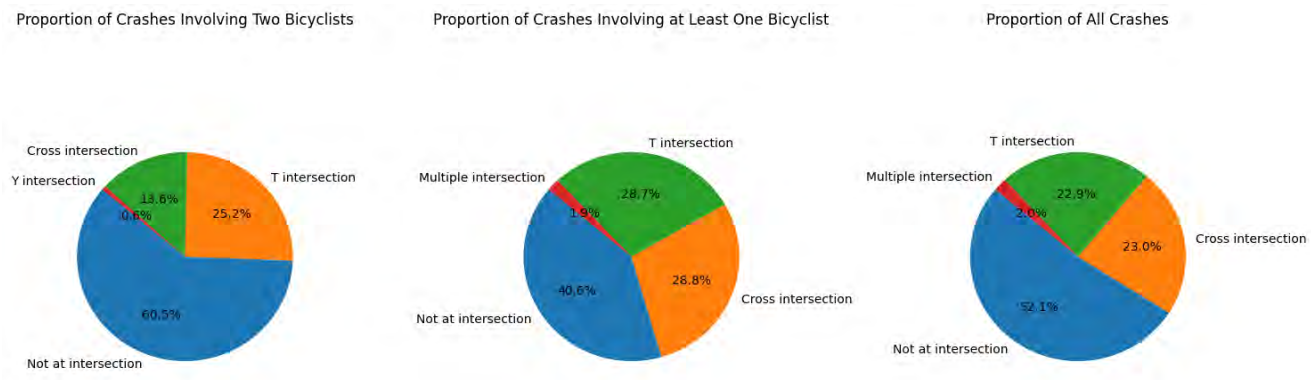



Figure 6.5: Distribution of bicyclists involved crashes, bicycle-bicycle crashes, and all crashes at different road geometries

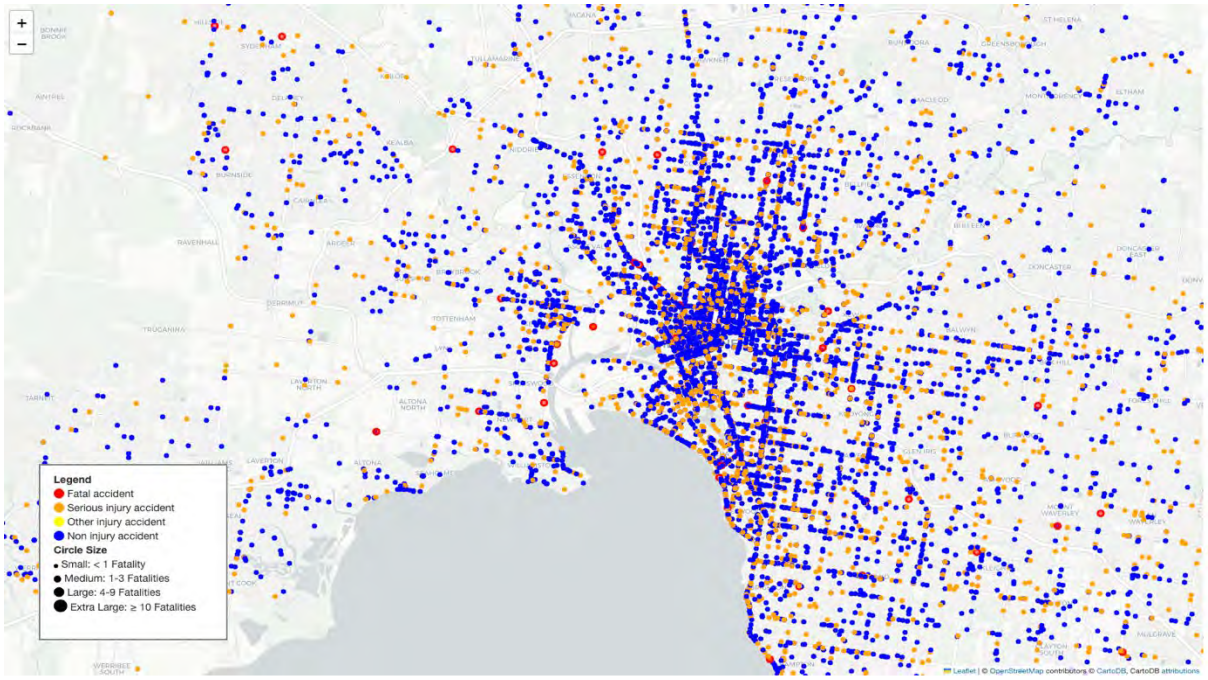



Figure 6.6: Distribution of bicycle-involved crashes based on type of injuries and number of fatalities

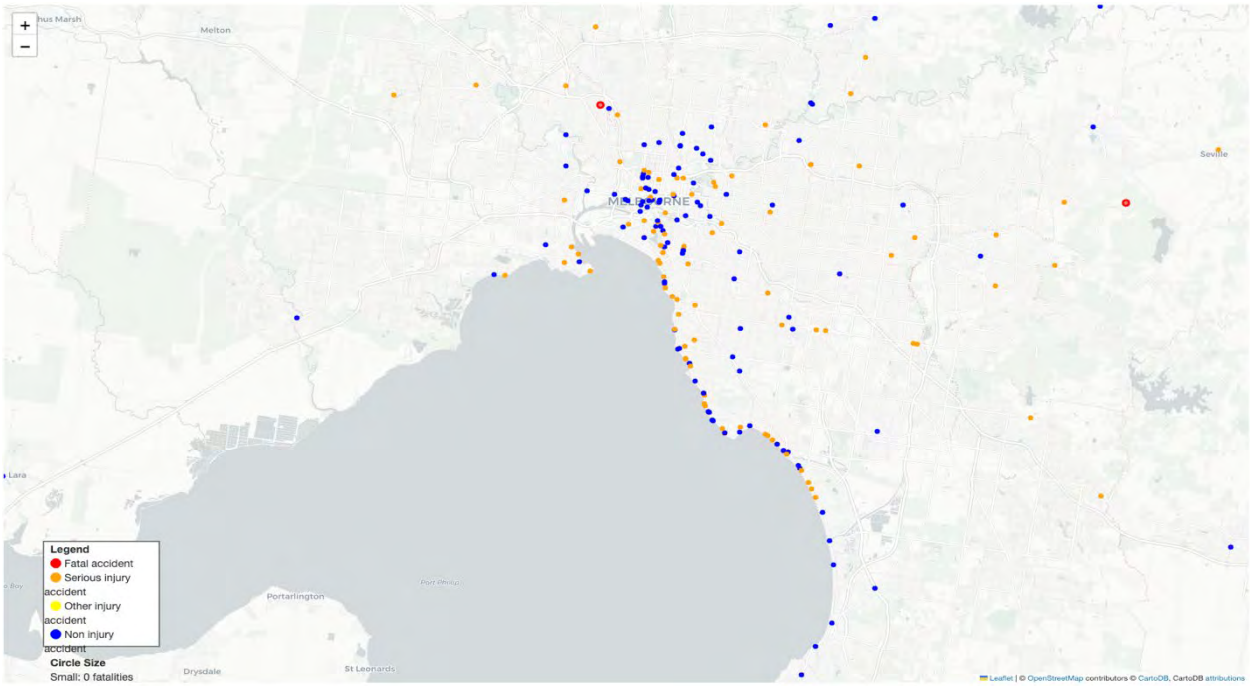



Figure 6.7: Distribution of crashes involving two or more bicycles based on the type of injuries and number of fatalities.

**Historical bicycle abnormal events (harsh braking)**

This section presents an extensive data analysis based on GPS and gyroscope bicycle data. We explored the distribution of features in both normal and abnormal events. Figure 6.8 shows that for every feature in the abnormal data, the mean values are roughly tenfold higher than those in normal conditions, indicating potential crashes or incidents. For example, the distribution of maximum brake value graph shows the distribution of abnormal events per unit grid location. In other words, over 1000 locations recorded more than 10 abnormal events over the past three years. This figure is useful for identifying hotspots where abnormal events occur frequently, helping to highlight areas with potentially higher risks for cyclists.

The spike on the right of the graph is interesting, which may relate to the specific cause of incidents. While it is difficult to definitively attribute to a specific cause without further investigation, the following factors may potentially explain it to some extent. Which included, cyclists traveling downhill and those traveling at higher speeds on roads with higher speed limits. In the case of sudden braking, these two groups indeed generate larger braking values compared to those riding at lower speeds.

**Spatial Distribution of Abnormal Events:**

- Figure 6.9 illustrates the spatial distribution of abnormal events based on different levels of harsh braking at three distinct resolutions. The results indicate that the central part of the metropolitan area exhibits a higher density of extreme and high braking events, suggesting areas with frequent sudden stops or a high likelihood of potential crashes. Clear patterns can be observed, demonstrating a distinct distribution between low braking events and moderate to high braking events.
- There is a high concentration of harsh braking events around the Yarra River at Southbank, where cyclists share paths with pedestrians. This higher concentration could result from the pedalling and braking movements and the higher volume of bicyclists in the area. Additionally, significant number of events are observed at intersections, where the majority of bicycle crashes occur (Hirose et al., 2021).

**Road Surface Quality:**

- Figure 6.10 is a spatial heatmap depicting the quality of road surfaces, derived as a proxy measure using gyroscopic sensor readings from bicycle data. The abnormal values are classified into four different ranges:
    - Extreme (95th percentile - red): 89 to 163
    - High (75th percentile - orange): 16 to 88

- Moderate (50th percentile - yellow): 4 to 15
- Low (<50th percentile - blue): 0 to 3

- A high concentration of red and orange areas in the central part of the map indicates poor and hazardous road surfaces. Overall, y-surface readings from gyroscope sensors can potentially be used as a proxy measure to assess the quality of road pavements and identify hazardous regions requiring maintenance or improvement to enhance road safety and surface quality.
- The higher readings, as indicated by red arrows in the zoomed-in sections, suggest that certain roads and segments may have deteriorated or aged pavements with lower levels of service and potholes.

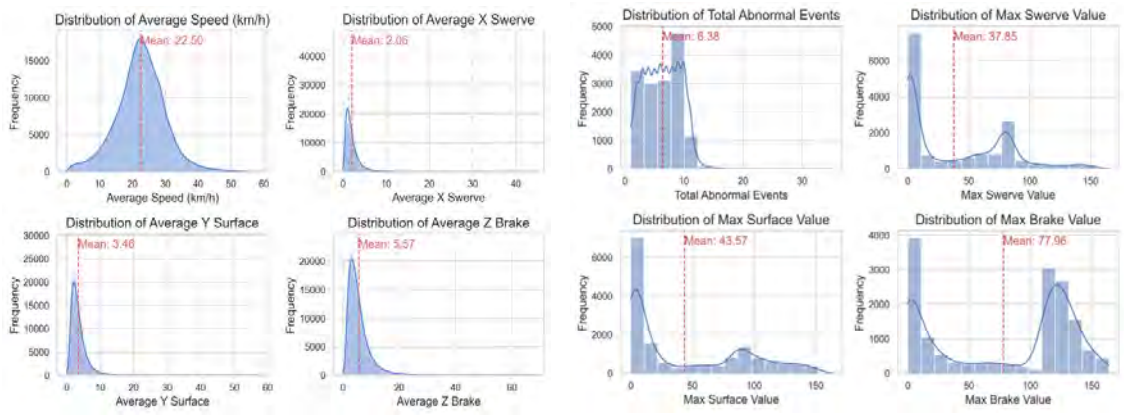



Figure 6.8: Gyroscope scores for normal conditions (left) and extreme cases (right)

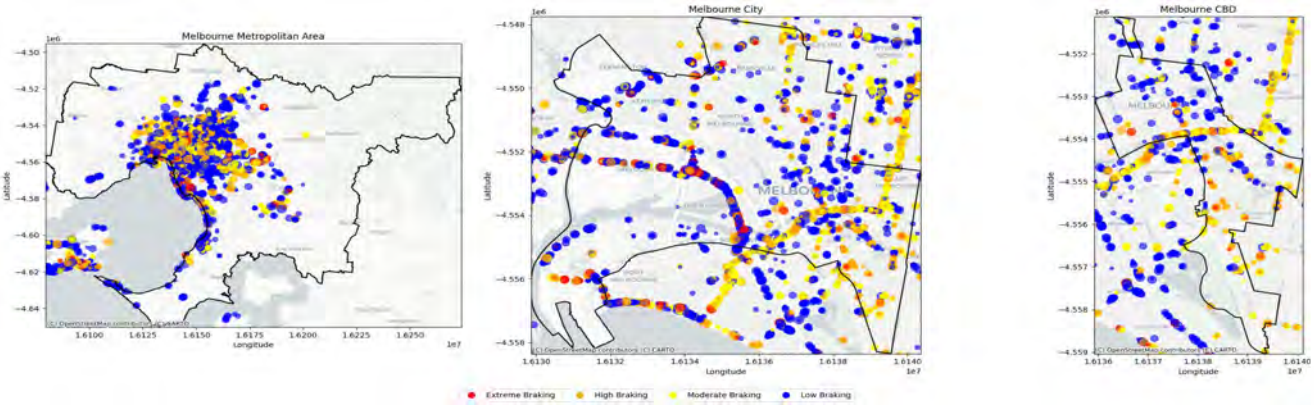



Figure 6.9: Spatial distribution of abnormal events (harsh braking) based on different braking levels. The left panel represents the Melbourne metropolitan area, the middle panel represents Melbourne city, and the right panel represents Melbourne CBD.

**Relationship between Gyroscopic Variables:**

- Figure 6.11 shows a correlation matrix between gyroscopic variables, including swerving, braking, and y-surface score. Correlation matrix is often used to show how each variable is correlated with the others. For example, it suggests that a low-quality road surface is associated with an increased likelihood of swerving and harsh braking, which can indicate a higher risk of accidents for cyclists. For example, the correlation between max swerve and braking is 0.51 which is significant and meaningful.
- Overall, y-surface readings from gyroscope sensors can potentially be used as a proxy measure to assess the quality of road pavements and identify hazardous regions requiring maintenance or improvement to enhance road safety and surface quality.
- However, identifying these abnormal events, potential biases and errors in the design and implementation of the algorithm, profiling and validating these values for different cyclists and types of bicycles, as well as ensuring the accuracy of data collection and treatment, are crucial factors that influence the reliability and potential applications of these datasets.

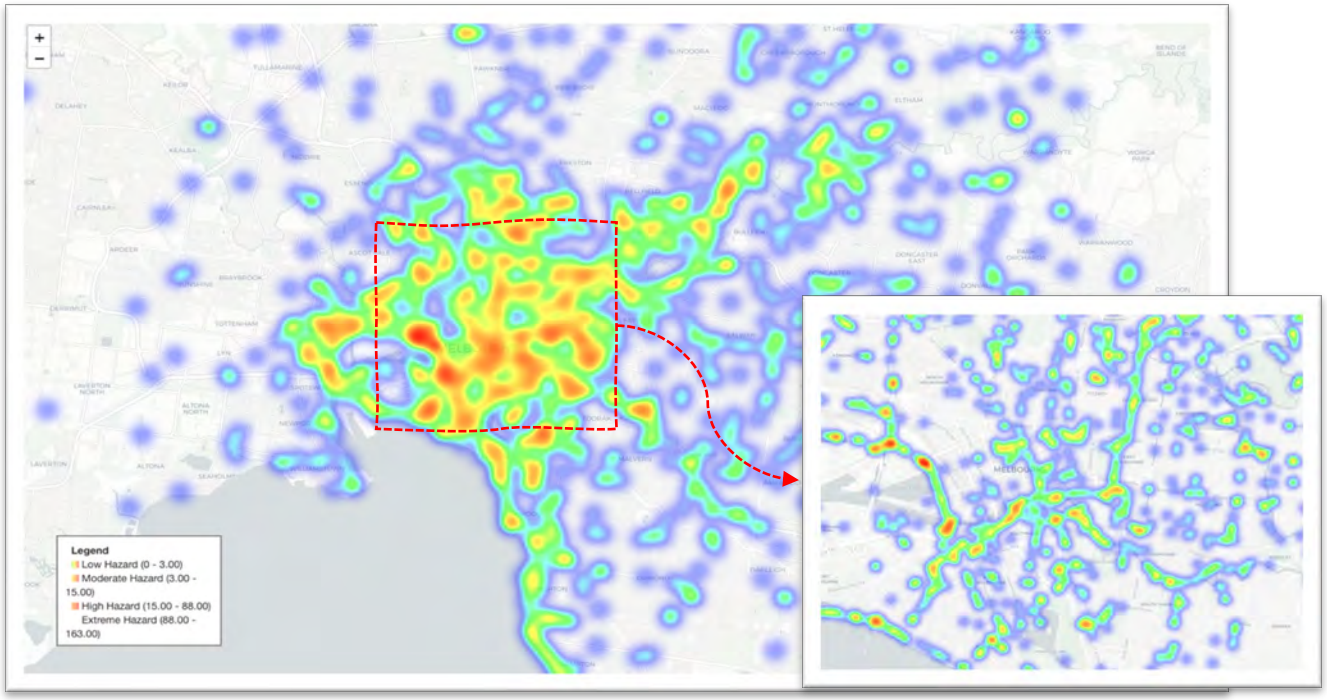



Figure 6.10: Spatial heatmap of the road surface quality (the darker the poorer and hazardous)

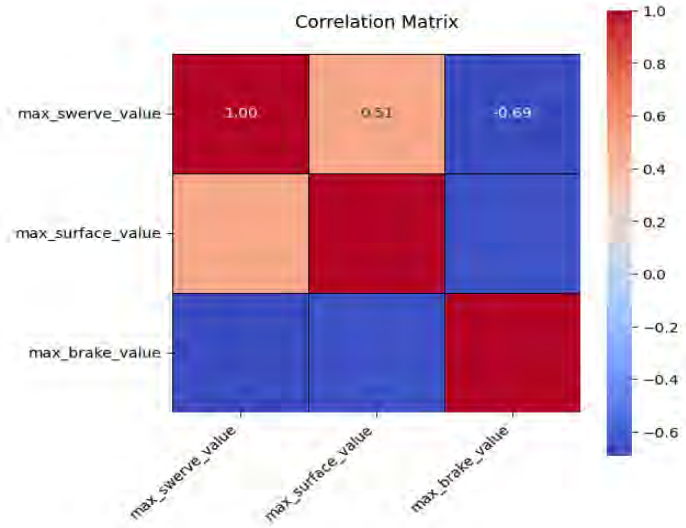



Figure 6.11: Correlation matrix between gyroscopic variables

### *6.1.3 Correlation between near-misses and actual crashes: A surrogate safety measures using bicycle GPS and Gyroscope data at scale*

This method relies on identifying and analysing harsh braking events, abnormal manoeuvres, and near-miss situations recorded by GPS and gyroscope sensors during bicycle journeys. Since the values of these surrogate safety indicators deviate significantly from normal conditions, it is hypothesised that there is a potential correlation with actual crashes. Building on this idea, the research aims to identify high-risk locations, behaviours, and road conditions that may contribute to bicycle accidents. This approach allows for the proactive identification of dangerous locations or issues with the current geometric design, contrasting with traditional reactive methods.

The bike GPS data is collected from a sample of bicyclists over a span of three years (2021 to 2024). To better accuracy, both the crash data and abnormal events data are aggregated into 50x50 meter grids. Additionally, when integrating crash data with SeeSense data to model surrogate safety measures, we

limited the bicycle crash data to the same period as the SeeSense datasets, from January 2021 to May 2024. This alignment ensures a more accurate representation of the correlation estimation. Two different correlation coefficients are utilised: Pearson correlation and Spearman correlation, as shown in Equations 7.1 and 7.2, respectively.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{Equation 7.1}$$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{Equation 7.2}$$

Where, $x_i$ is the number of abnormal events from GPS data at location $i$, $y_i$ is the number of actual accidents from historical crash data at location, $\bar{x}$ and $\bar{y}$, are the mean abnormal events and mean actual accidents across all locations, respectively. $n$ is the total number of locations in the dataset and $d_i$ is the difference between the count of abnormal events and actual accidents.

The spatial heatmaps presented in Figures 6.12 and 6.13 offer a detailed comparison and analysis of road safety dynamics within the Melbourne metropolitan area. Both heatmaps reveal a similar pattern of high incident density in the Central Business District (CBD), indicating areas prone to frequent and severe braking as well as higher accident rates involving bicyclists.

This correlation suggests that the dense urban traffic and complex road networks in the CBD are critical factors contributing to these events. Suburban areas exhibit a more dispersed pattern of harsh-braking events compared to the more localised concentration of bicyclist accidents along major roads and local towns, highlighting different spatial dynamics at play. Peripheral regions show lower incident intensities, reflecting safer conditions likely due to reduced traffic volumes and simpler road layouts. It is worth noting that the GPS data comes from a sample of bicyclists using SeeSense smart bicycle lights who agreed to share their data. Therefore, outer regions and suburbs might be underrepresented in the sample, necessitating further investigation.

The analysis shows a significantly higher correlation between abnormal events and actual crashes at intersections (Pearson correlation rank: 0.69, Spearman rho: 0.36) compared to non-intersection areas (Pearson correlation rank: 0.50, Spearman rho: 0.26). For the non-classified model, which includes both intersections and non-intersections together, the Pearson correlation remains the same as at intersections, 0.69, while the Spearman rho slightly reduces from 0.36 at intersections to 0.32 for the entire network, as shown in Figure 6.14.

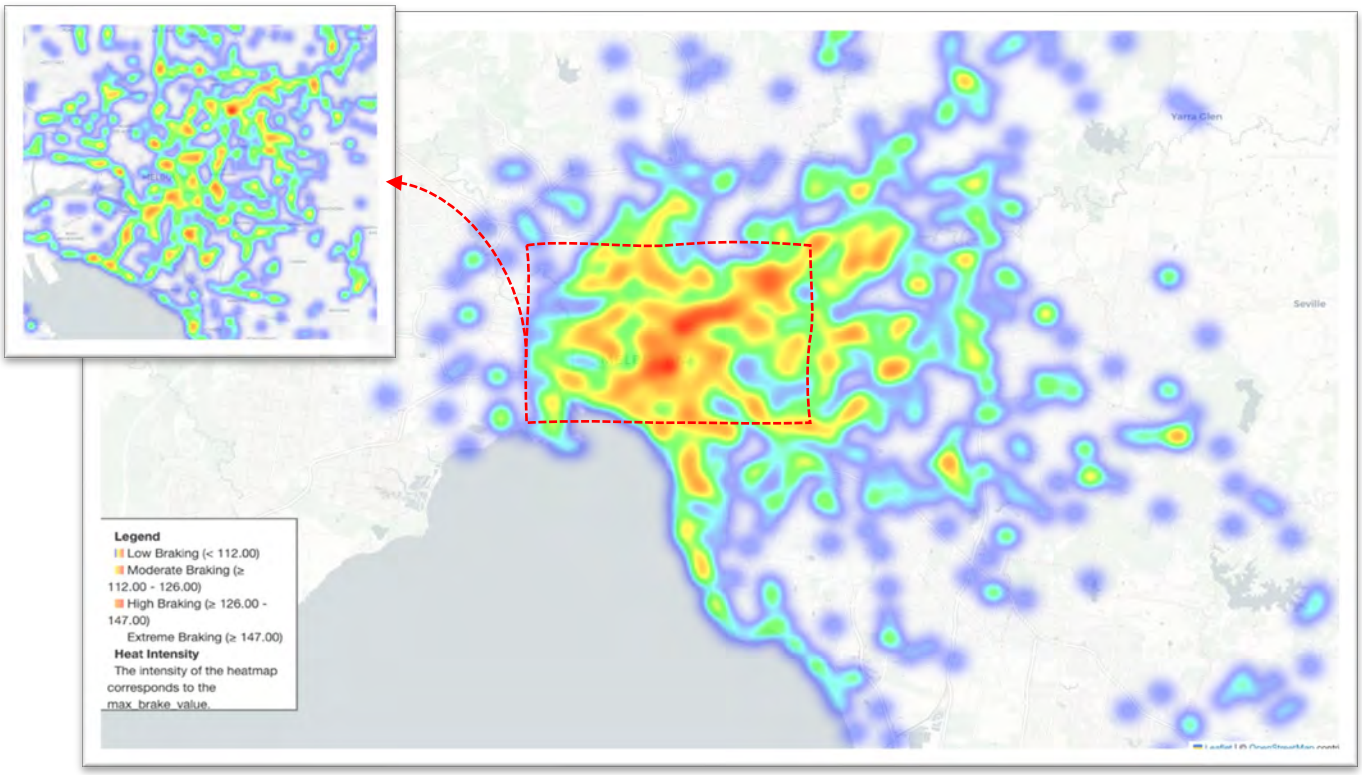



Figure 6.12: Spatial heatmap of the abnormal events based on harsh-braking scores (the darker the more severe instances and/or higher number of abnormal events)

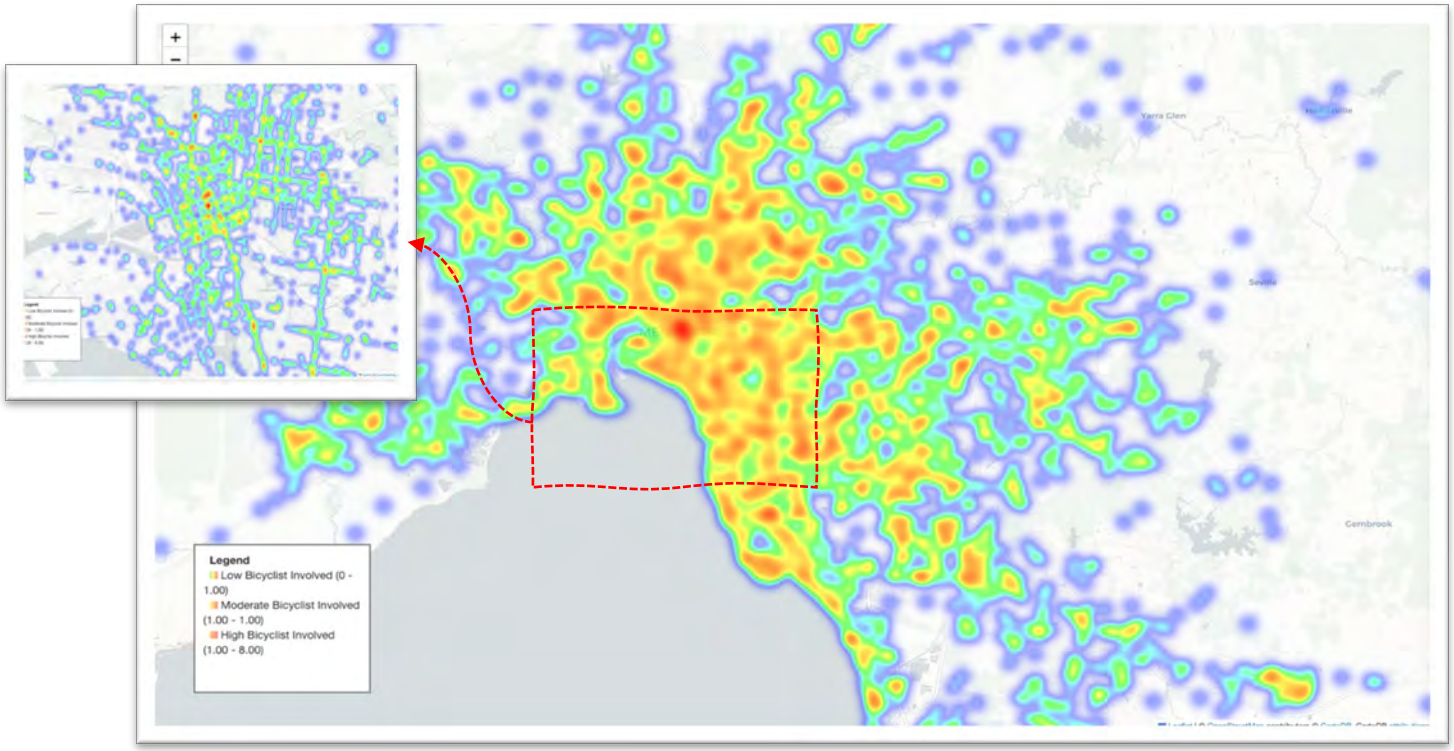



Figure 6.13: Spatial heatmap of the number of bicyclists based on actual crash data (the darker the higher number of bicyclists involved and/or higher accidents)

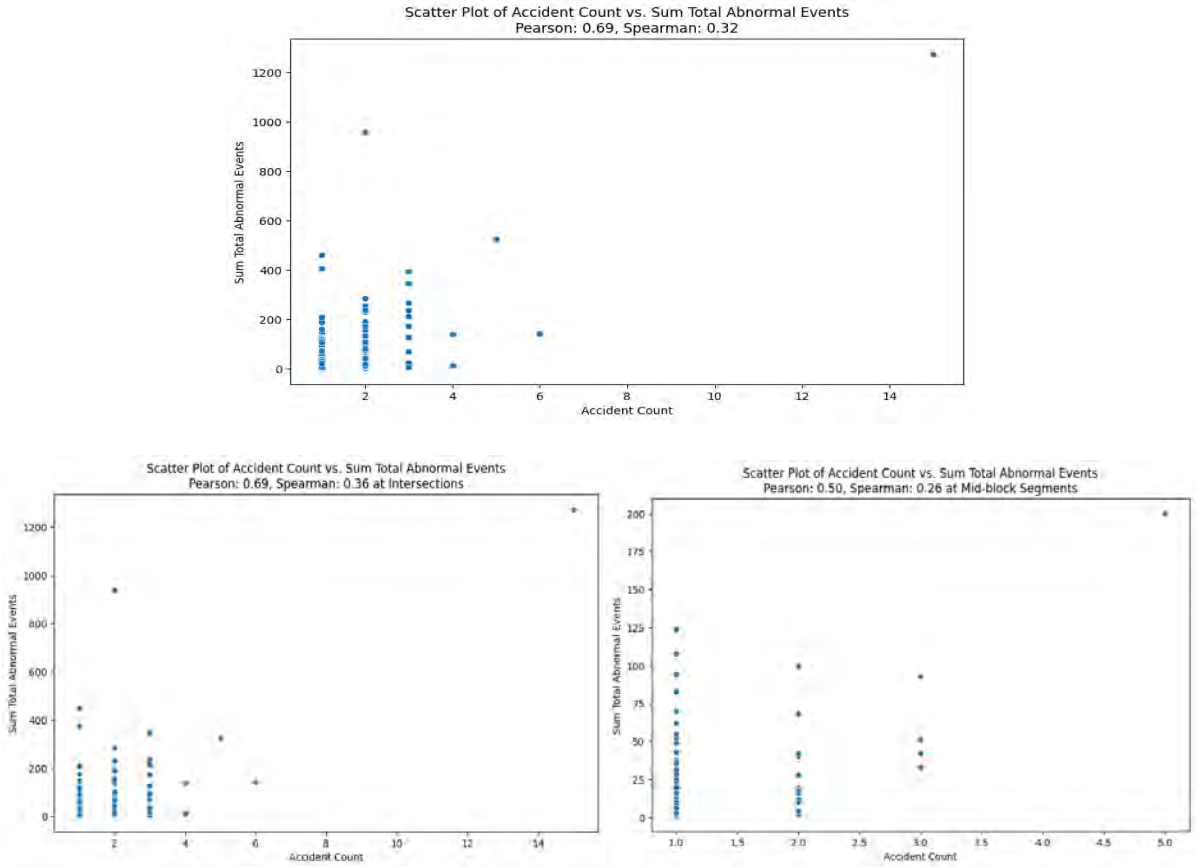



Figure 6.14: Scatter plots and correlation ranks between abnormal events and actual accident counts for the entire network (top), at intersections (bottom left), and for non-intersections (bottom right).

These results demonstrate that utilising surrogate safety measures from bicycle GPS and gyroscope data at scale can be instrumental in developing proactive safety measures, identifying safety hotspots, and using proxy measures to assess and improve the road surface quality, particularly for vulnerable road users.

Most importantly, minor injuries resulting from accidents between pedestrians and cyclists are significantly underestimated in police crash data. Therefore, relying solely on such data falls short in identifying the distribution, patterns, and locations of these hotspots. The high correlation and the significant role of abnormal events in explaining actual crashes, as found in this research, suggest that surrogate safety measures based on GPS and gyroscope data can be valuable for identifying, preventing, and managing safety risks. This approach can substantially improve the quality of roads, biking trails, shared corridors, and intersections, making them safer and more accessible for everyone.

## 6.2 Bicycle flow efficiency

Cycling offers numerous public benefits, including reducing traffic congestion and air pollution, as well as promoting healthier lifestyles. Cities globally are actively encouraging cycling as a sustainable and practical travel option by investing in cycling infrastructure and facilities, and by motivating people to cycle more regularly. Cycling is often a fast and efficient alternative for short-distance trips, particularly in densely populated urban areas (Clarry, 2019 #699). The potential for faster travel makes cycling an attractive option for a broader population. Therefore, it is essential to further study cyclists' flow and speeds across various types of road networks to better understand their needs and improve urban planning.

This section reports a developed framework using large-scale GPS data aiming at our following objectives:

1. Identify the most frequent routes and popular cycleways.
2. Explore where users ride faster and what are the patterns and the frequency of delay.
3. Understand how different road functional classes correlate with speed distributions.
4. Classify different cyclist typology using unsupervised machine learning.

The following framework is proposed to effectively address the designed objectives. This framework illustrates the process for analysing cycling data to provide insights on road network performance, using GPS data from bicycles mapped with road network, see Figure 6.15. The initial step involves mapping this data onto the existing road network infrastructure, which can be classified into different types, such as all network types or specific car network types. This means that we exclusively analysed the distinct behaviour between dedicated bicycle network and overall network shared with other modes including cars. This mapping enables the extraction of segment-based network insights, which include metrics such as speed, bicycle flow and volume, delay distribution, and frequency of delay. These metrics help in understanding the overall performance, hotspots, and usage of different segments of the road network by cyclists.

The processed data is then used to classify roads based on key performance indicators (KPIs) derived from the earlier metrics. This classification helps in identifying which segments of the road network perform better or worse for cyclists, which can guide improvements in infrastructure. Following this, the classified data is fed into an unsupervised machine learning clustering process, which groups similar patterns together to reveal different cyclist topologies and network states. These insights allow for a deeper understanding of how different segments of the road network are used and can inform urban planning and policy decisions to enhance cycling infrastructure and overall urban mobility.

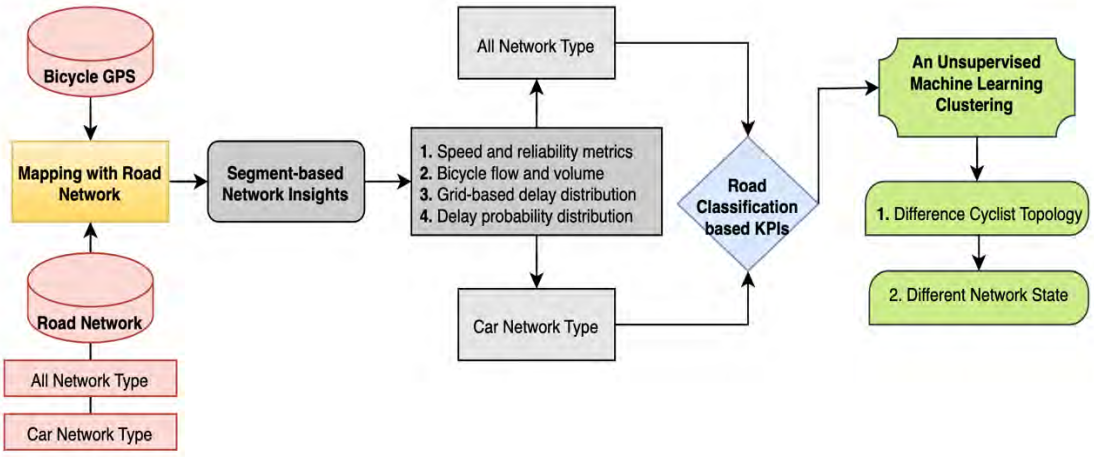



Figure 6.15: The designed framework for bike flow efficiency analysis

### A. Identifying the most frequent routes using average number of cyclists

The Figures 6.16 provides insights into cyclist density across Melbourne City's Road Network. Both maps illustrate the average number of cyclists using different road segments, with a colour gradient from red (low density) to green (high density). In this figure the bicycle telematic data is mapped to two different road network type, the top figure shows average bicycle volume on all type roads that can be used by all other modes and the bottom figure shows road network that is exclusively shared between motorised vehicles and bicycles, other modes such as pedestrians are possibly not permitted to enter. This can help to distinguish whether the popularity of bicycles routes depend on dedicated lanes or the share a similar pattern between dedicated lanes and shared with motorised vehicles.

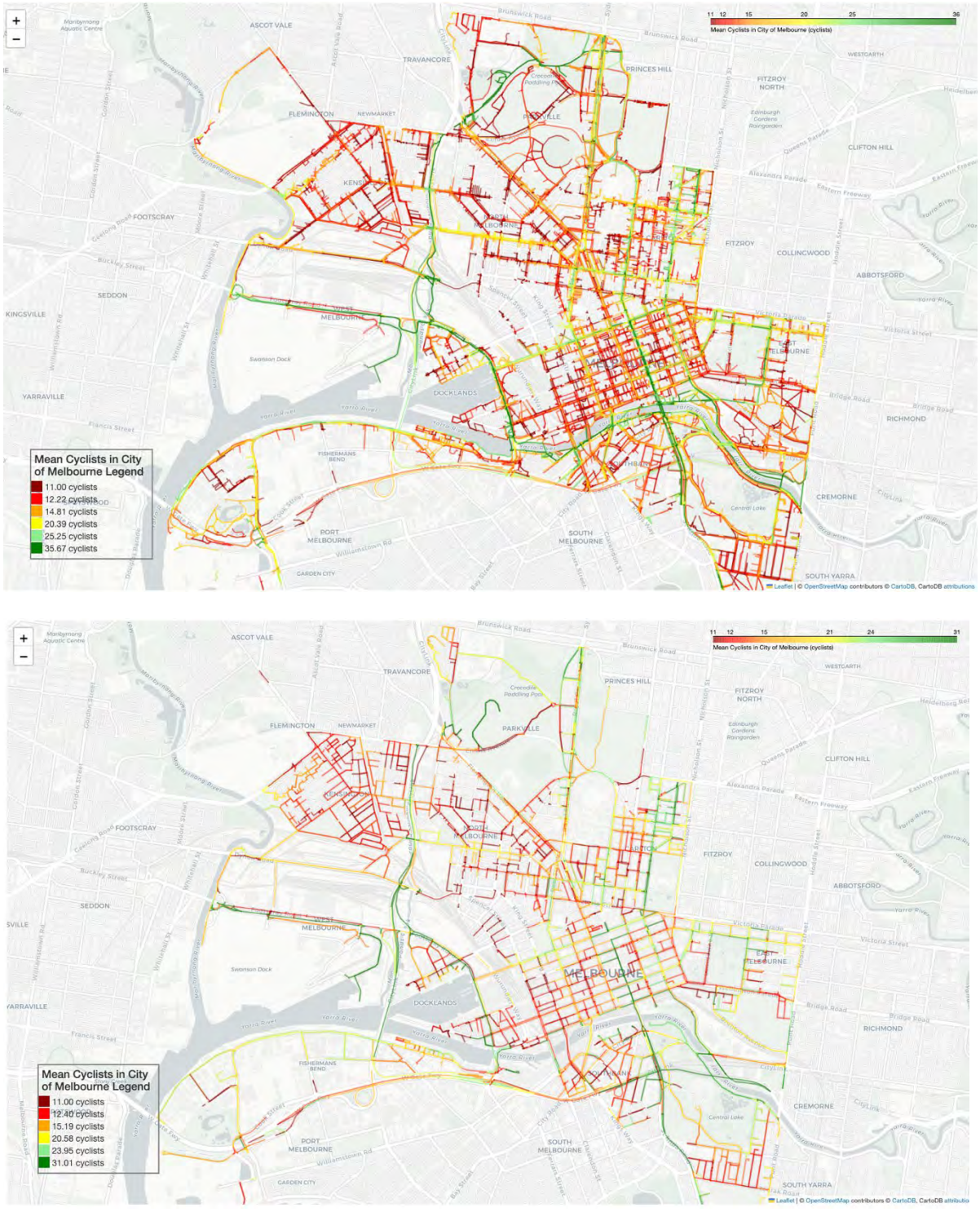



Figure 6.16: Popular route based on average number of bicyclists, all type of road network (top panel) and car exclusive road network (bottom panel).

The key findings and patterns can be summarised as follow:

1. **High Cyclist Density Areas**: In both figures, the highest cyclist density (green lines) is observed around central Melbourne, particularly in the CBD, Docklands, and surrounding inner-city suburbs. This indicates these areas are primary routes for cyclists, likely due to their proximity to business and recreational areas.

2. **Key Cycling Routes**: The figures highlight specific roads and paths with consistently high cyclist numbers. These include key arterials such as Swanston Street and routes along the Yarra River. These routes serve as major corridors for commuting and leisure cycling.
3. **Consistency regardless of road type**: Interestingly, it can be seen that regardless of the type of road network whether it is shared with motorised vehicles and cars or included exclusive lanes for bikes, the popular routes remain the same. However, the average flow increases with existence of the dedicated lanes and pathways for bikes. This shows that policy makers and city planners can boost bicycle usage for both daily commuting and leisure by improving roads and adding protected bikeways.

**B. Where cyclists ride fast**

The Figure 6.17 illustrates the average cycling speeds across Melbourne City's Road Network. The first plot maps average speeds on all road types that can be used by various modes of transport, including bicycles. The second plot focuses specifically on roads that are shared by bicycles and motorised vehicles. In both plots, colour-coding is used to represent speed: red indicates lower average speeds, while green shows higher speeds. It is evident that cyclists experience reduced speeds, particularly in the CBD and along major roads, where interactions with other vehicles are more frequent.

Key findings can be summarised below:

1. **Cyclists experience longer delays at major intersections**, on major roads, and in the city centre, indicated by the lower speeds (red areas) in these locations. This is more evident in Figure 6.18, which shows delay frequency. Generally, 15 to 25 percent of cyclists are delayed at major intersections, compared to only 3 to 8 percent on minor roads and at smaller intersections.
2. **On average, dedicated bike paths improve overall cyclist speed** compared to routes where cyclists share the road with cars, highlighting the benefits of segregated cycling infrastructure.
3. **The car network plot (second figure) shows more areas of reduced speed**, particularly in the city centre and on major roads, suggesting that in multi-modal situations, such as those involving both cyclists and motor vehicles, average cycling speed decreases. This reduction can be linked to longer waiting times at intersections, car-oriented traffic signals, and a lower prevalence of dedicated bike lanes. However, further intersection-specific analysis is needed to confirm these observations.
4. **The significant speed reduction on major roads and in the CBD could imply that cyclists might engage in illegal crossing behaviours at minor intersections and on local roads**, where speed reductions are less evident, see Figure 6.17 and 6.18. This finding is crucial for understanding the safety of vulnerable road users. Additional research, including correlation with traffic incidents and crash data, is necessary to validate these observations and assess road safety implications.

**C. Road functional classes and speed distribution**

The Figure 6.19 displays the distribution of various performance indicators for cyclists across different road functional classes, including walkways and paths, primary roads, and motorways/links. Each road class contains the speed distribution (average speed in km/h), the number of average cyclists, the distribution of delays (time stopped in seconds), and the percentage of cyclists who were experienced delays. These charts provide insights into how cycling conditions vary depending on the type of road and highlight key aspects of cycling performance, such as speed and delay.

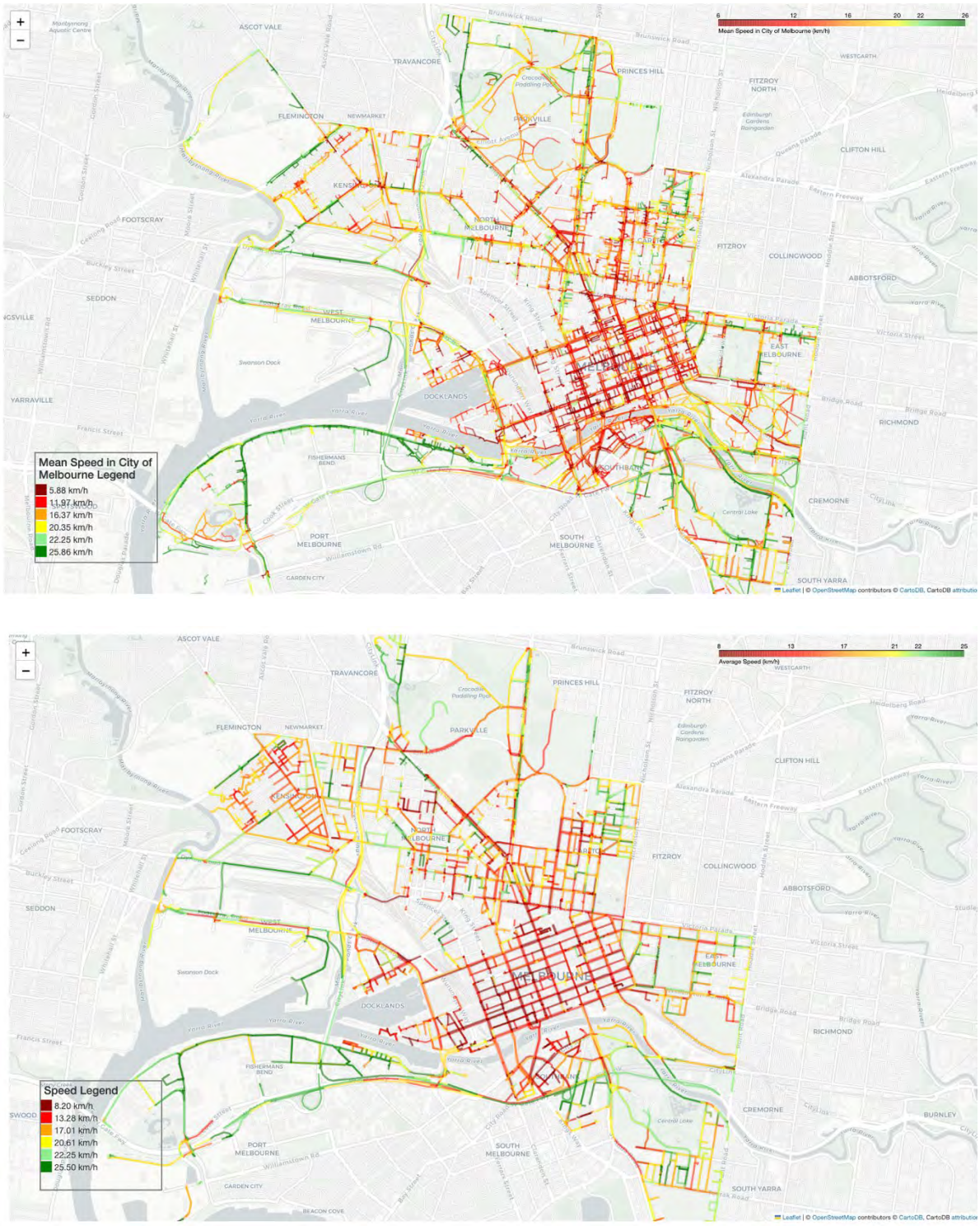



Figure 6.17: Average speed distribution across Melbourne City. All type of roads (top panel), car driving roads (bottom panel).

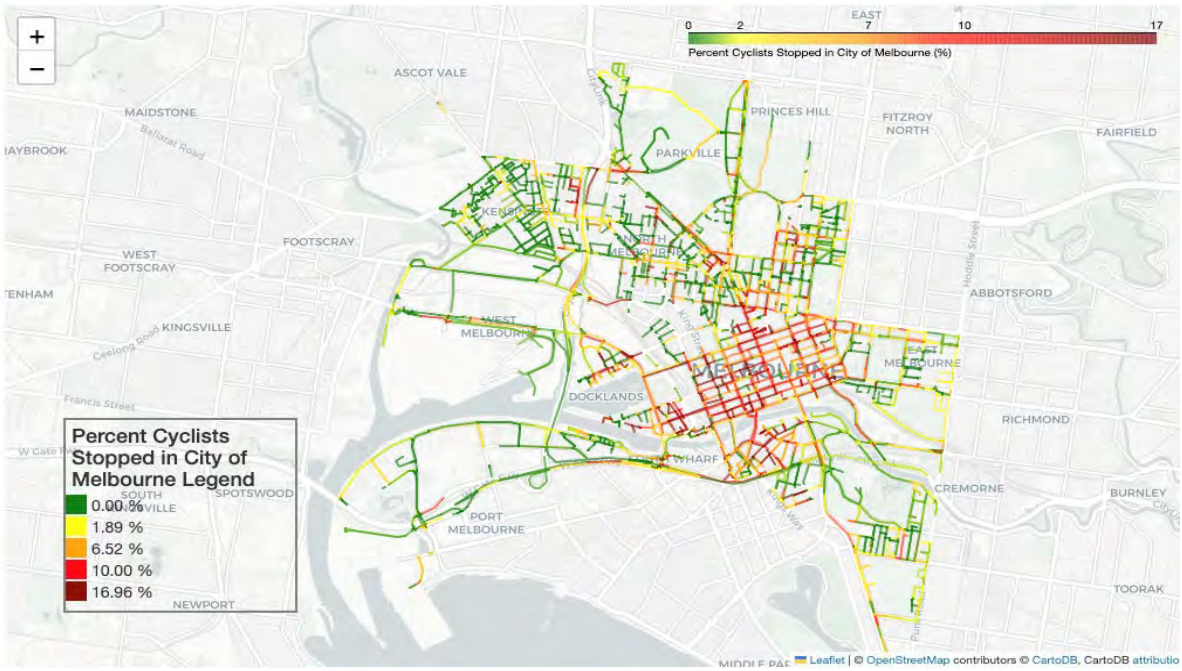



Figure 6.18: The probability of delays in Melbourne City. It shows delayed bicyclists as the percentage of all bicyclists.

Across all road types, the speed distribution shows a peak between 15 and 25 km/h, suggesting that this is the typical cycling speed range. The number of average cyclists decreases sharply as the speed increases, indicating that fewer cyclists maintain higher speeds. Moreover, average cyclists per spatial unit of 3x3 metres, hold the same distributions for different road types with slight increase for bicycle paths as compared to primary roads and motorways, refer to Figure 6.19.

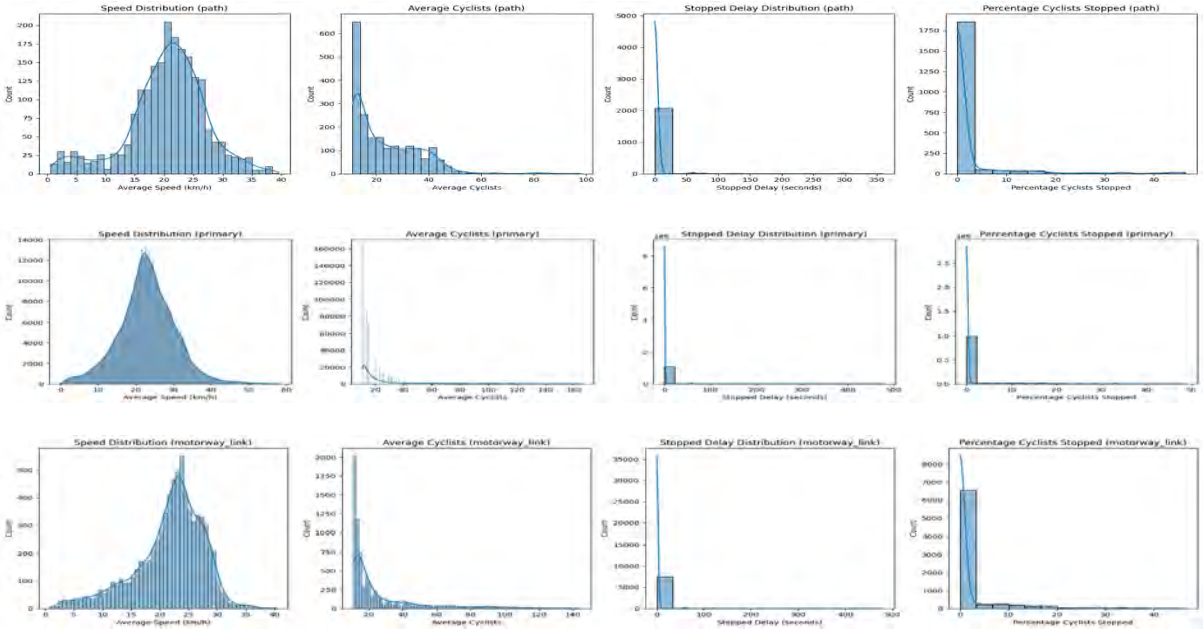



Figure 6.19: The distribution of speed, delay, flow, and delay frequency versus road functional classes.

### D. Cyclist typology classification using unsupervised learning

The methodology involved using unsupervised machine learning, specifically the KMeans clustering algorithm, to analyse cyclist behaviour and identifying cyclist's topology and network state in Melbourne City. Key features considered for clustering were average speed, average cyclists flow, percent of cyclists

stopped and experienced delay, and number of lanes the cyclists use. These features provide insights into cycling speed, volume, delays, and road characteristics. The dataset was pre-processed using standard scaling to normalise these features. The optimal number of clusters was determined using the Elbow Method, which involves plotting the inertia (sum of squared distances of samples to their nearest cluster centre) for different numbers of clusters. Based on findings from the Elbow curve, five clusters were identified as the optimal number of clusters, where the rate of decrease in inertia begins to slow, indicating a balance between underfitting and overfitting.

Figure 6.20 illustrates the spatial distribution of the five identified clusters across Melbourne City. Each cluster is represented by different coloured points on the map, showing the geographical concentration and dispersion of various cycling behaviours. Clusters are distributed across different types of roads and locations, reflecting diverse cycling activities and conditions. For example, certain clusters are concentrated in the city centre and along major roads, indicating areas with higher cycling traffic and possibly more congestion. In contrast, other clusters are spread along recreational areas like beaches and the Yarra River, suggesting leisure cycling activities. This spatial visualisation helps to understand how different cycling behaviours are associated with specific urban areas and infrastructure types.

The identified clusters can be summarised as following:

1. **Cluster 0: Suburban Commuters:**
   Cyclists in this cluster are similar to those in Cluster 1 but exhibit slightly lower speeds and frequencies. They predominantly use 4-5 lane main roads and highways, indicating usage of broader routes. They distributed across outer suburbs and major roads.
2. **Cluster 1: Typical Daily Cyclists:**
   These cyclists are frequent users of the road network, second only to leisure cyclists in number. They primarily use 2-3 lane roads, maintaining relatively high speeds compared to other clusters. This suggests they are commuters or regular users who navigate efficiently through traffic. The distribution in the city centre and along main roads highlights the role of these cyclists in daily urban mobility, see Figure 6.21.
3. **Cluster 2: Leisure Cyclists:**
   Positioned mainly along beaches, the Yarra River, and other dedicated bike lanes, these cyclists maintain the second-highest average speed. They use bike paths and shared corridors, highlighting their preference for safe and scenic routes. This cluster underscores the importance of recreational cycling infrastructure in urban planning.
4. **Cluster 3: Most Delayed Cyclists:**
   Characterised by frequent delays, these cyclists are commonly found around major intersections, showing how car-oriented signal timings affect cycling efficiency. They primarily use 2-4 lane major and primary roads, indicating that traffic signals and intersections are significant bottlenecks for cyclists in these areas. On average over **29%** of these cyclist's experience delays ranging from **30** to **110** seconds.
5. **Cluster 4: Fastest Cyclists:**
   This cluster includes cyclists maintaining speeds above 29 km/h, likely road bikers or those using electric bikes for fast transit. Their distribution outside the city centre and in suburbs and major roads indicates high-speed, long-distance travel.

The findings from this clustering analysis provide valuable insights into the different types of cycling behaviour in Melbourne City. Understanding these patterns allows for better planning and management of urban cycling infrastructure. For example, improving signal timing at major intersections could reduce delays for cyclists, enhancing overall cycling efficiency and safety. Recognising the distribution of leisure cyclists can guide the development of more dedicated bike paths, promoting recreational cycling. Furthermore, acknowledging the needs of fast cyclists can lead to the creation of dedicated high-speed bike-lanes, accommodating different types of cyclists. These insights can help city planners and policymakers create a more cyclist-friendly environment, encourage cycling as a sustainable mode of transport, and improve urban mobility and health outcomes.

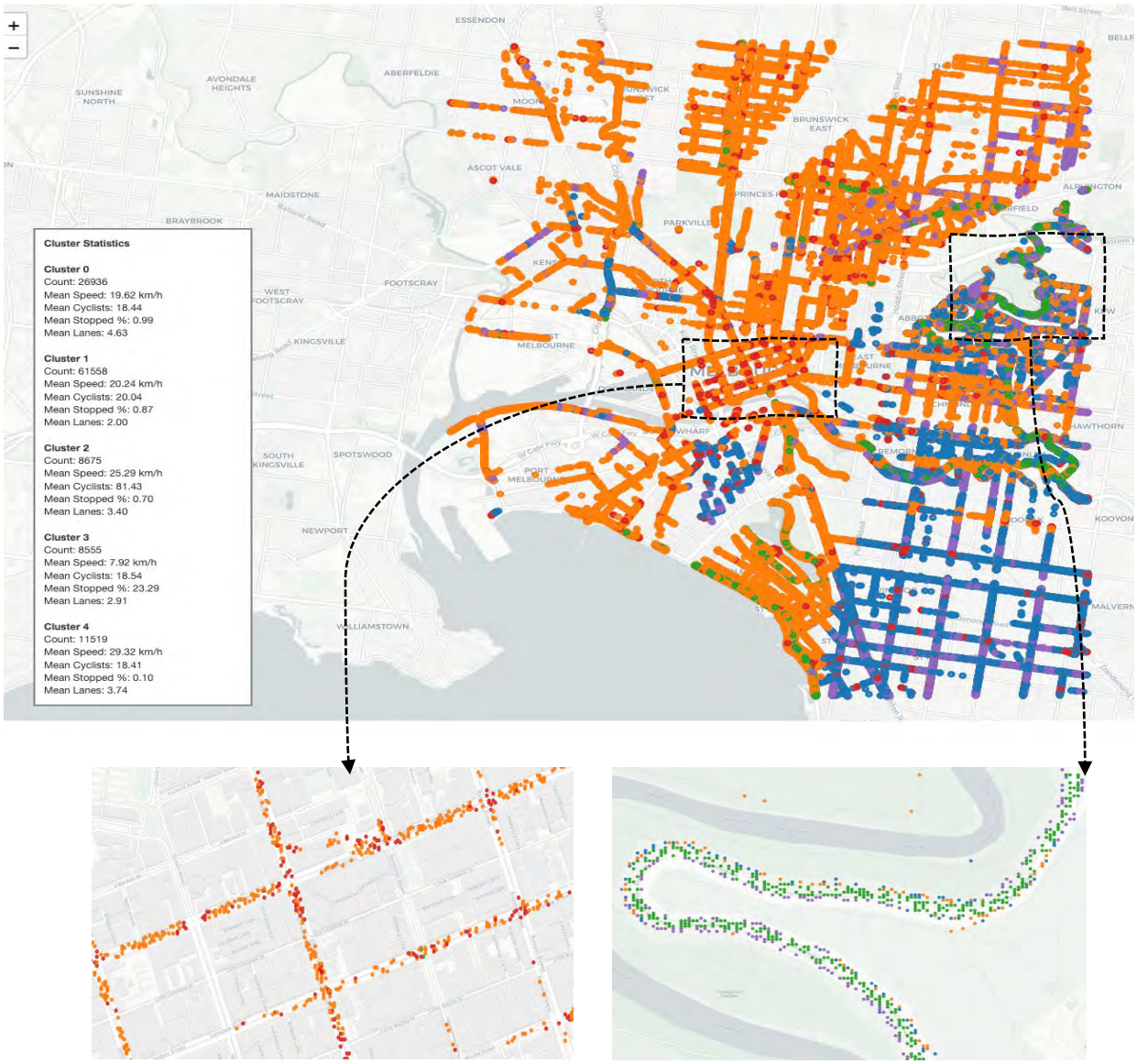



Figure 6.20: Spatial distribution of cyclist clusters in Melbourne City, Australia.

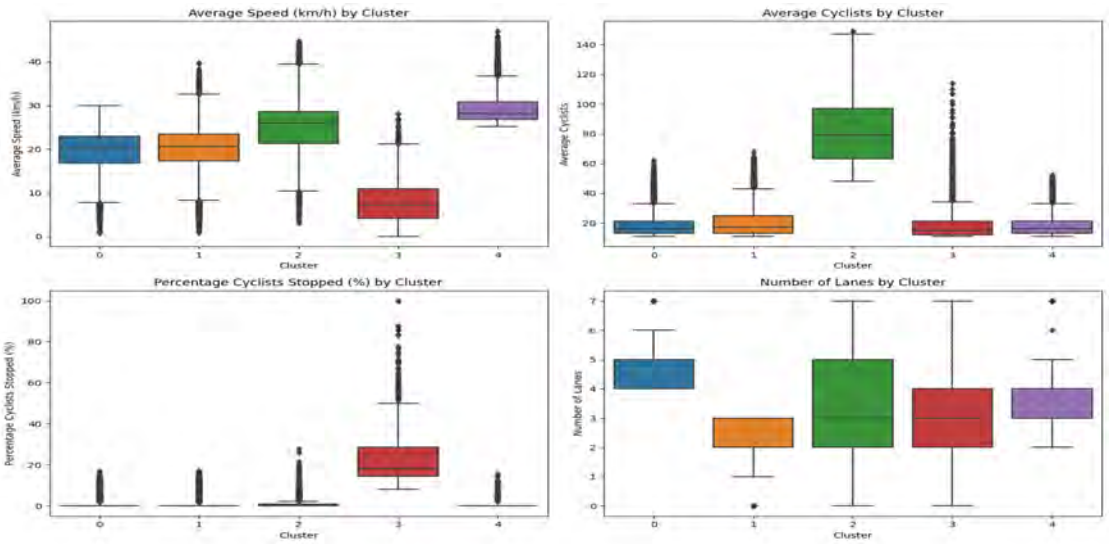



Figure 6.21: Boxplots of key cycling metrics by cluster.

## 6.3 Summary of key findings

### Bicycle Safety

1. **Bicycle crashes are more frequent at intersections**: Analysis of historical crash data shows that 59% of bicycle-involved crashes occur at intersections, compared to 41% of all crashes.
2. **Crashes involving two bicycles are more common on non-intersection:** The results show mid-block road segments accounts for (61%) as compared to (39%) at intersection.
3. **Fatal bicycle crashes are more distributed on the outskirts**: This is possibly due to higher speed limits in these areas.
4. **Higher concentrations of harsh braking events in the city centre**: This pattern seems frequent particularly around shared paths and intersections.
5. **Abnormal events are correlated with actual crashes**: The study finds a significant correlation between abnormal events (such as harsh braking and swerving) identified from bicycle GPS and gyroscope data and actual crashes.
6. **Intersections have a higher correlation between abnormal events and actual crashes**: The correlation is higher at intersections (0.69 Pearson rank, 0.36 Spearman rho) than at non-intersection areas (0.50 Pearson rank, 0.26 Spearman rho).
7. **Surrogate safety measures can identify safety hotspots**: The study demonstrates that using surrogate safety measures from bicycle GPS and gyroscope data can help identify safety hotspots, unsafe infrastructure and risky behaviours, providing a proactive measure to improve road safety without waiting for actual crashes to occur.
8. **Minor injuries between pedestrians and cyclists are underestimated**: The study highlights that minor injuries resulting from accidents between pedestrians and cyclists are significantly underestimated in police crash data.
9. **GPS and gyroscope data can assess road surface quality**: The study finds that y-surface readings from gyroscope sensors can be used as a proxy measure to assess road surface quality and identify hazardous regions.
10. **Targeted interventions can improve road safety**: The study emphasizes the need for targeted traffic management and safety interventions, particularly in high-risk zones like the CBD, popular biking trails, and shared corridors with pedestrians.

### Bicycle flow efficiency

#### Bicycle flow pattern

11. **High Cyclist Density Areas**: Central Melbourne, including the CBD and Docklands, has the highest cyclist density, indicating these areas are key routes for commuting and recreation.
12. **Key Cycling Routes**: High cyclist numbers are seen on major routes like Swanston Street and paths along the Yarra River, serving as main corridors for both commuting and leisure.
13. **Consistency regardless of road type**: Popular cycling routes remain consistent regardless of road type, but dedicated bike lanes increase overall cyclist flow, suggesting infrastructure improvements could boost cycling rates.

#### Where ride faster?

14. **Cyclist Delay and Speed:** Cyclists face longer delays and lower speeds at major intersections and in the city centre. Dedicated bike paths improve speed, highlighting the benefits of segregated infrastructure.
15. **Impact of Multi-modal Roads:** Roads shared with cars show more speed reductions, especially in central areas, due to traffic signals and lack of bike lanes. This reduction can be linked to longer waiting times at intersections, car-oriented traffic signals, and a lower prevalence of dedicated bike lanes. This suggests a need for more dedicated cycling infrastructure to enhance cyclist safety and efficiency.
16. **Safety Risks at Minor Intersections:** The significant speed reduction on major intersections and in the CBD could imply that cyclists might engage in illegal crossing behaviours at minor intersections and on local roads, where speed reductions are less evident.

**Cyclist topology and network state**

17. **Cluster 0: Suburban Commuters:** Similar to daily cyclists but use 4-5 lane roads and highways, showing a pattern of longer-distance commuting from outer suburbs, with slightly lower speed.
18. **Cluster 1: Daily Cyclists:** Frequent users of 2-3 lane roads with relatively high speeds, mainly in the city centre and along main roads, indicating a focus on daily commuting.
19. **Cluster 2: Leisure Cyclists:** Located along recreational routes such as beaches and the Yarra River, these cyclists prefer safe, scenic bike paths and maintain high speeds.
20. **Cluster 3: Most Delayed Cyclists:** Experience frequent delays at major intersections, highlighting the impact of car-oriented signal timings on cycling efficiency. They primarily use 2-4 lane major roads. On average 29% of these cyclist's experience delays ranging from 30 to 110 seconds.
21. **Cluster 4: Fastest Cyclists:** Maintain speeds above 29 km/h, typically found on suburban and major roads, indicating use by road bikers or for fast transit e-bikers.

# 7 An enhanced computer vision model for traffic conflict prediction

This section presents the results from the enhanced computer vision model based on vehicle (car) trajectory data. As the safety model is thoroughly presented in the interim report, we have included only the results from the enhanced computer vision model here. For more detailed information about the safety model, please refer to the interim report.

**Computer Vision Model**

The computer vision model used to perform the traffic safety analysis is an object detection model by Jocher et al. (2023) called YOLOv8. Yolov8 model are pretrained by using COCO dataset to perform object detection and segmentation for common item such as phone, person, chair, etc. The model is chosen because of its speed and accuracy that enables real life object detection application. With the available dataset, it is hypothesised that the model would be able to detect conflict-related trajectories after model retraining is done.

The dataset used to perform the retraining on YOLOv8 model consists of trajectories plot (each plot is referred as a single frame), trajectories data and conflict data. To evaluate the model, the main measure used is the Mean Average precision (MaP). MaP quantifies the model's ability to accurately detect and classify objects within images frame by calculating the average precision for each class based on precision-recall curves. This average precision is then averaged across all classes to provide an overall assessment of the model's performance. Two types of MaP are used for the evaluation of this model: MaP50 and MaP50-95. The numbers after "MaP" refers to the percentage of area that intersects between the ground truth bounding boxes and the predicted bounding box. This intersection is referred to as Intersection of Union (IoU). While the MaP represent the precision of the model in the detection task, the IoU represent how accurate the predicted object location is.

**Severity Model**

The model was initially trained for the identification of conflict based on their severity level. Conflicts are categorised into four categories for both "rear end" and "lane changing" conflicts. The levels are categorised into Property Damage Only (PDO) (level 0), Minor Injury (level 1), Major Injury (level 2) and Fatal (level 3). The training is conducted with a maximum of 100 epochs (iterations) due to the absence of a substantial performance increase beyond this point. This is further supported by the "early stop" feature, which ensures that the model ceases training once it reaches peak performance. Different batch sizes were also tested during training to monitor changes in performance, as illustrated in Figure 7.1. An example of the detection results can be seen in Figure 7.2.

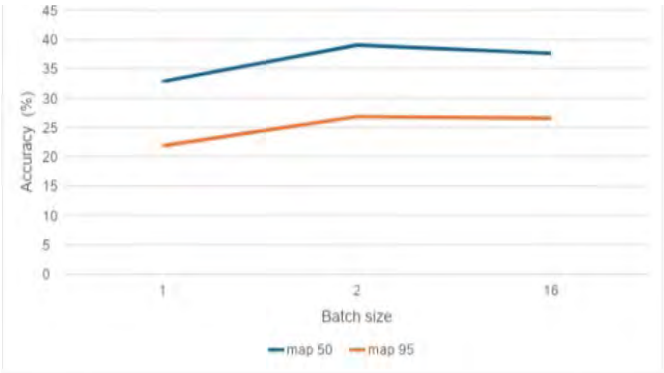



Figure 7.1: Changes in MaP over different batch sizes with a maximum of 100 epochs.

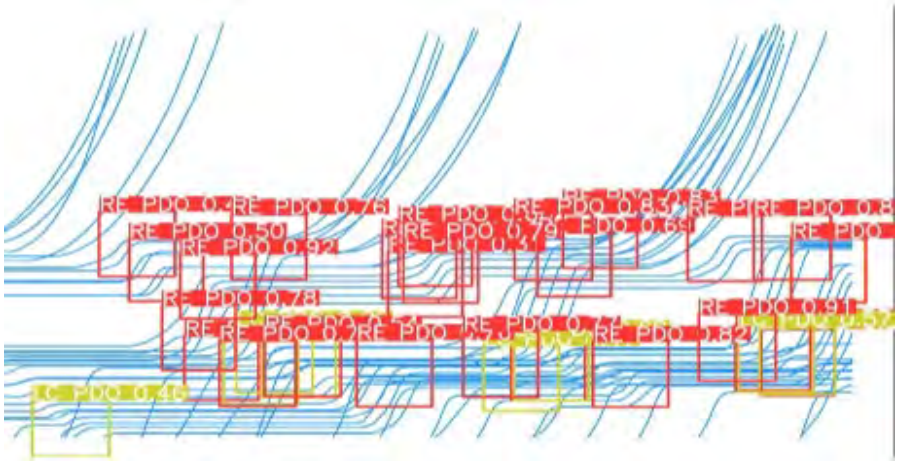


Figure 7.2: Example of the Detection Result of the Severity Model with batch size of 16 and maximum 100 epochs.

Based on the results in Figure 7.1, the MaP of the model remains lower than that of the YOLOv8-based model, which achieves a MaP50-95 of 53.9% (Joecher et al., 2023). This discrepancy is attributed to the limited amount of data available for some conflict levels during training. To address this, an alternative method is proposed to improve the MaP of the model. In our approach, we divided the model into several models, each focusing on predicting a specific conflict class. However, during the development phase, it was determined that the severity class data lacks useful insights for traffic safety analysis. Therefore, it was decided to use the cost level of each conflict instead.

**Cost Model**

The cost model focuses on detecting conflicts that are categorised based on the cost of damages incurred. Conflicts are divided into 9 different levels: rear-end levels zero to four and lane-change levels zero to three, with the corresponding number of observations shown in Table 7.1. Each cost level corresponds to a different range of costs. These cost ranges can be calculated using the formulas below, where ( T ) is the upper cost boundary and ( τ ) is the lower boundary (calculations are in dollars).

$$\tau = 10^{costlevel}$$

$$T = 10^{(costlevel+1)} - 1$$

The model is trained with different bounding box sizes: 100x100, 300x300, and 500x500 pixels. Moreover, the model, which is separated into 9 different sub-models, each trained to detect a single conflict class, is called the Ensemble Model. Ensemble modelling approach where a dedicated sub-model is utilised to each conflict class is used to improve the accuracy and performance.

Table 7.1. Training instances distribution

| Class | Number of Instances |
|---|---|
| Full class | 17799 |
| RE 0 | 6092 |
| RE 1 | 7504 |
| RE 2 | 1538 |
| RE 3 | 89 |
| RE 4 | 8 |
| LC 0 | 986 |
| LC 1 | 1116 |
| LC 2 | 490 |
| LC 3 | 12 |

The training performance progression for the severity models (Full Model and Ensemble Model) is illustrated in Figures 7.3 to 7.5. Some fluctuations are present on the graphs, particularly for Rear-End level 4 and Lane Change level 3. These fluctuations are attributed to the limited number of instances for these two classes, as shown in Table 7.1. The low number of training instances prevented the model from training smoothly, resulting in fluctuations during the early stages of the training process.

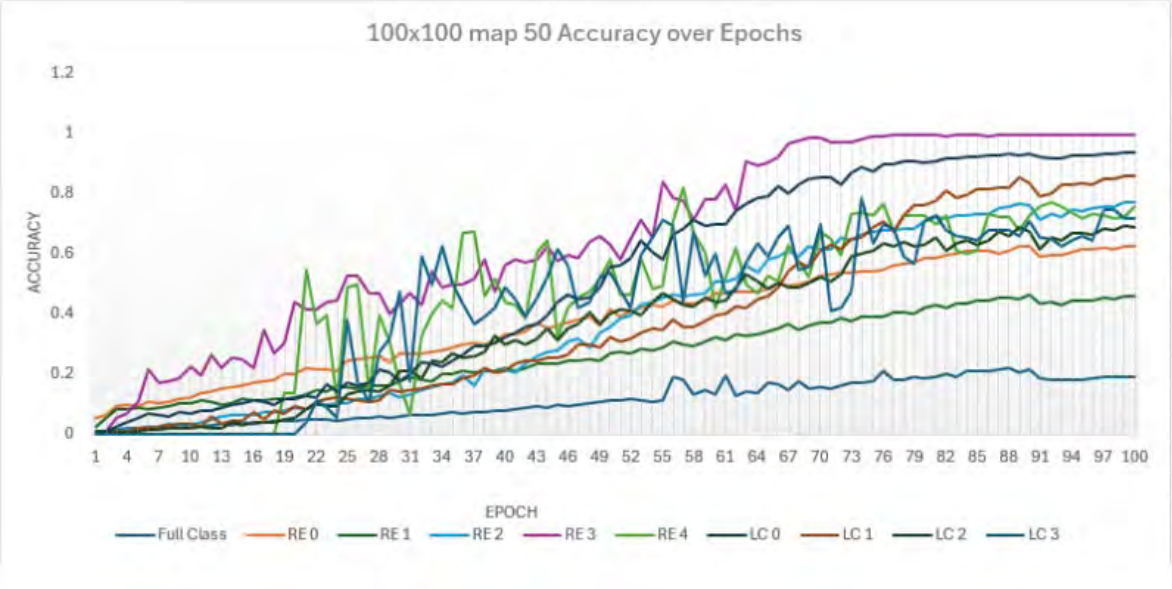



Figure 7.3: Training Progression on Full and Ensemble Model trained with 100x100 bounding box size

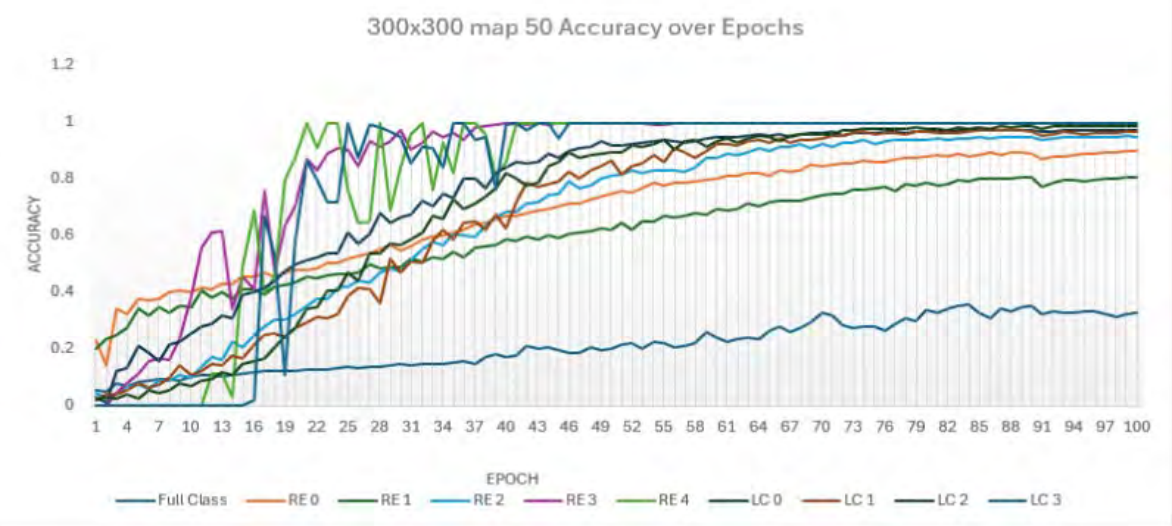



Figure 7.4: Training Progression on Full and Ensemble Model trained with 300x300 bounding box size

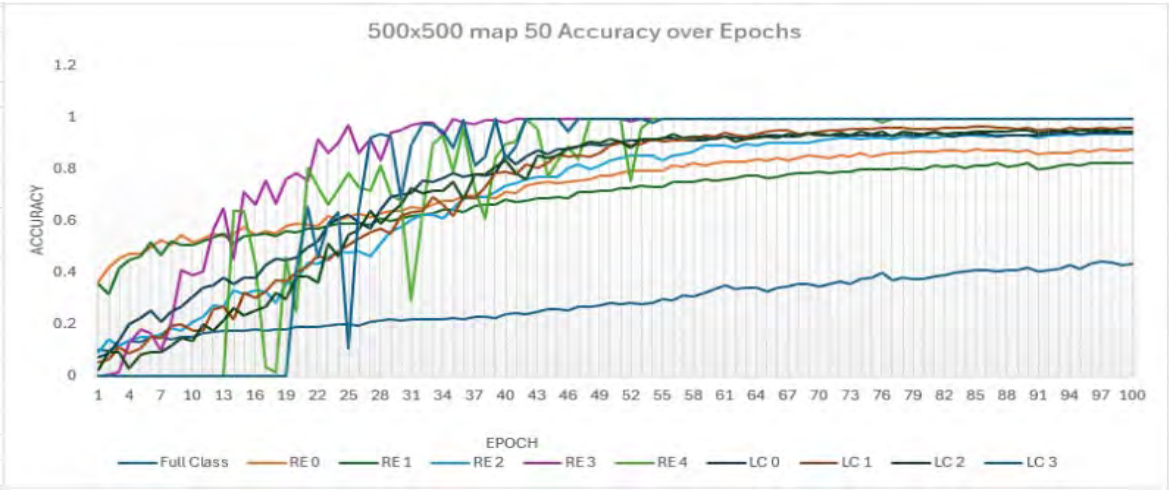



Figure 7.5: Training Progression on Full and Ensemble Model trained with 500x500 bounding box size

Table 7.2. MaP of the models

| Model Training Class | | MaP50 | | | MaP50-95 | | |
|---|---|---|---|---|---|---|---|
| | | *100x100* | *300x300* | *500x500* | *100x100* | *300x300* | *500x500* |
| Full Class | | 0.22 | 0.359 | 0.439 | 0.12 | 0.253 | 0.326 |
| Ensemble | RE 0 | 0.627 | 0.896 | 0.878 | 0.341 | 0.704 | 0.744 |
| | RE 1 | 0.465 | 0.804 | 0.822 | 0.341 | 0.581 | 0.66 |
| | RE 2 | 0.765 | 0.949 | 0.938 | 0.432 | 0.768 | 0.819 |
| | RE 3 | 0.718 | 0.995 | 0.995 | 0.894 | 0.795 | 0.887 |
| | RE 4 | 0.8123 | 0.995 | 0.998 | 0.389 | 0.762 | 0.889 |
| | LC 0 | 0.936 | 0.972 | 0.941 | 0.558 | 0.815 | 0.841 |
| | LC 1 | 0.754 | 0.973 | 0.965 | 0.854 | 0.728 | 0.811 |
| | LC 2 | 0.722 | 0.987 | 0.949 | 0.689 | 0.799 | 0.807 |
| | LC 3 | 0.705 | 0.995 | 0.995 | 0.389 | 0.78 | 0.884 |

The results indicate that the models are able to detect the conflicts with excellent performance, as shown in Table 7.2. Our proposed ensemble modelling approach significantly outperformed the benchmark (full class model), nearly doubling its performance. Figures 7.6 and Figure 7.7 show that the trendline for the 500x500 bounding box predictions is the closest to the ground truth values. This suggests that the 500x500 bounding box size is the most optimal for this model.

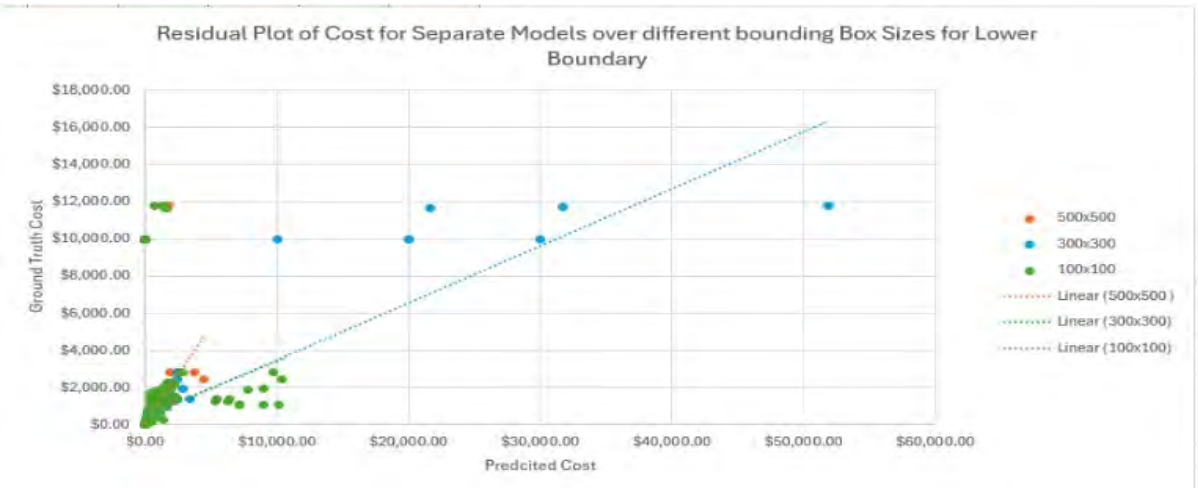



Figure 7.6: Residual Plot of the Ensemble Model over different Bounding box sizes for the Lower Boundary

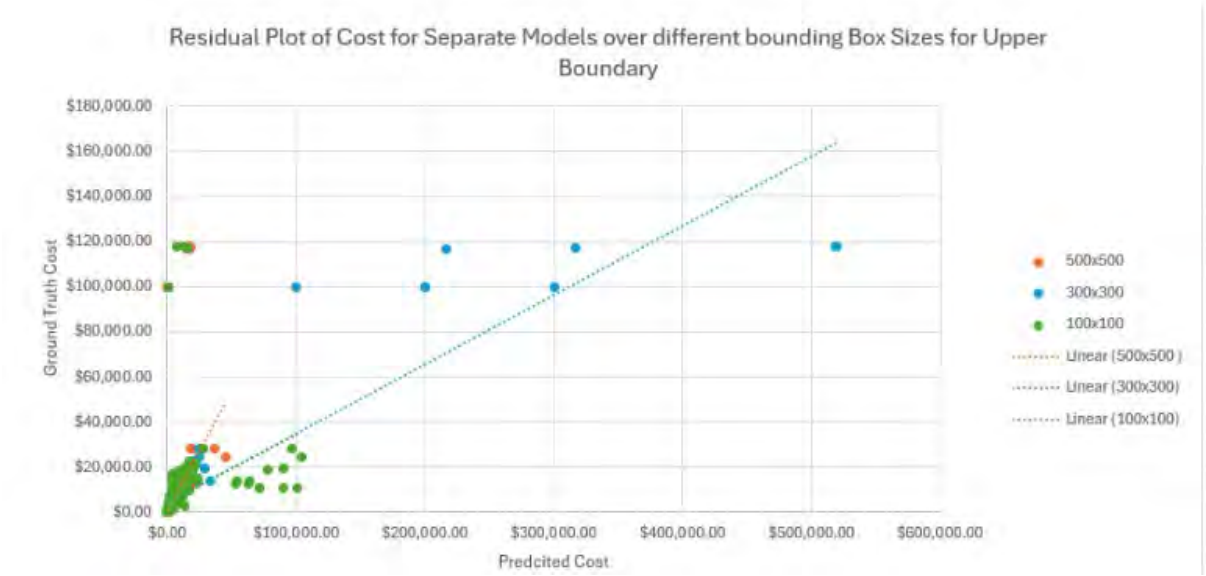


Figure 7.7: Residual Plot of the Ensemble Model over different Bounding box sizes for the Lower Boundary

## 7.1 Summary of key findings

1. **Enhanced computer vision model for conflict prediction**: The study presents an enhanced computer vision model using ensemble modelling approach based on YOLOv8, an object detection model, for traffic safety analysis due to its speed and accuracy.
2. **Model training**: The model was retrained using a dataset of trajectories plots, trajectories data, and conflict data.
3. **Model performance**: The model performance was evaluated using Mean Average Precision (MaP), with MaP50 and MaP50-95 being the primary metrics, and achieves a Mean Average Precision (MaP50-95) of above 80%.
4. **Cost model**: The cost model detects conflicts based on the cost of damages incurred, with 9 different levels and corresponding cost ranges.
5. **Ensemble modelling approach**: The ensemble model, where each sub-model is trained to detect a single conflict class, significantly outperforms the full class model.
6. **Optimal bounding box size**: The 500x500 bounding box size is found to be the most optimal for the model.
7. **Improved performance**: The ensemble model nearly doubles the performance of the benchmark, achieving excellent detection performance.

# 8 Traffic state distribution reconstruction using physics-informed neural network

Accurate traffic state estimation is crucial for effective traffic management in intelligent transportation system. Traditional data-driven approaches often require large amounts of training data and may not fully capture the underlying physical dynamics of traffic flow. Physics-Informed Neural Networks (PINN) have emerged as a promising solution, incorporating the governing physical laws into the neural network training process to predict the traffic state with limited available data. However, PINN models face challenges such as slow convergence and limited data availability in real-world scenarios.

This section presents findings from the manuscript "Investigating Knowledge Transfer in Residual Physics-Informed Neural Networks using Connected Vehicles Traffic Data" by Bo Wang, Neema Nassir, Negin Yousefpour, and Majid Sarvi, submitted to IEEE Transactions on Intelligent Transportation Systems. The study proposes a novel Residual Physics-Informed Neural Network (Res-PINN) architecture for traffic state estimation using connected vehicle data. The Res-PINN model incorporates residual connections to improve convergence speed and performance. Furthermore, a transfer learning approach is explored to leverage knowledge from data-rich scenarios and improve Res-PINN performance in situations with limited data availability.

## 8.1 Data Description

The study utilises real-world connected vehicle data from a major aggregator of vehicle telematics dataset, CompassIOT, focusing on vehicle trips along two key corridors in Melbourne, Australia: the Westgate Bridge on the M1 freeway and a section of the M3 freeway near the city centre, as shown in Figure 8.1. These corridors are selected due to their importance as major arteries in Melbourne and their high traffic volumes and varying traffic patterns across different time. Moreover, the selected sections are city-bound and have no incoming or outgoing traffic, making them suitable for modelling using the proposed physics-based approach.

(a) Case A: M3 Freeway towards city (b) Case B: M1 Freeway towards city

Figure 8.1: The satellite images of two selected traffic corridors with polygons. Source: Imagery Esri.

The vehicle telematics dataset contains historical travel information for individual vehicles, such as location and speed, which is reported with a frequency of every 1 to 12 seconds. By treating these vehicles as probe vehicles, the traffic speed distribution along the selected corridors can be reconstructed. The raw data is processed using interpolation for each trip to give every second a location and speed value. The final speed distribution is a matrix with a temporal resolution of 3 seconds and a spatial resolution of 30 meters.

Two datasets are created based on the processed speed distributions:

- **probe**: the raw data points collected by connected vehicles, which only has 0.62% and 1.69% of the total data points for the M3 and M1 corridors, respectively.

- **random**: randomly selects 5% data points from the entire distribution and has more samples.

These two training data availabilities provide a scenario to test transfer learning from a richer training dataset to one with limited samples.

## 8.2 Method and Results

The proposed Res-PINN model incorporates residual blocks to improve gradient flow and speed up convergence compared to traditional PINN. The residual connections allow the network to learn residual mappings instead of direct mappings, which facilitates the training of deep neural networks. The Res-PINN architecture is designed to effectively extract features from the input data and produce accurate traffic state estimations.

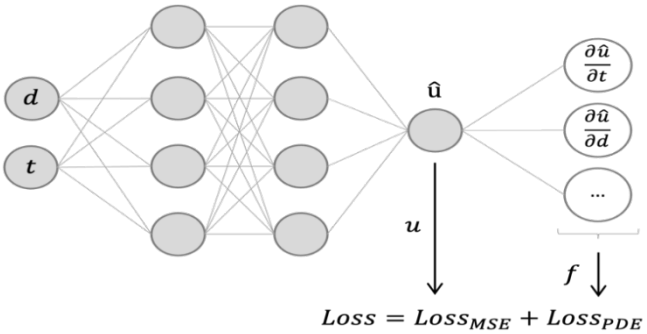



Figure 8.2: Overall workflow of PINN.

The Res-PINN model's loss function includes two parts: the neural network loss (Mean Square Error between the predicted and ground truth speed values) and the PDE loss (residual function of the governing traffic flow dynamics described by the Lighthill-Whitham-Richards model), as illustrated in Figure 8.2. By minimising the total loss, the Res-PINN model learns to estimate traffic speed while satisfying the underlying physical laws.

The Res-PINN model is evaluated on the M1 and M3 datasets using metrics such as Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Feature Similarity Index Measure (FSIM). The results demonstrate that Res-PINN outperforms the baseline models, achieving lower MSE, lower MAPE, and higher FSIM. For instance, on the M3 freeway corridor with the probe dataset (A.1), Res-PINN achieves an MSE of 25.497, MAPE of 0.227, and FSIM of 0.586, compared to the baseline PINN model's MSE of 27.076, MAPE of 0.231, and FSIM of 0.549. This represents improvements of 5.8% in MSE, 1.7% in MAPE, and 6.7% in FSIM.

Transfer learning is employed to initialise a Res-PINN model trained on the limited probe dataset with weights learned from the richer random dataset. This significantly improves the speed estimation performance on the probe dataset compared to training from scratch. For the M3 freeway corridor, transfer learning reduces the MSE from 25.497 to 23.287 (an 8.7% improvement), MAPE from 0.227 to 0.201 (an 11.5% improvement) and increases the FSIM from 0.586 to 0.599 (a 2.2% improvement).

A visual comparison of the predicted speed distributions in Figure 8.3 further highlights the benefits of transfer learning. The models utilising transfer learning (target TL) exhibit a closer resemblance to the ground truth compared to the models trained solely on the limited probe dataset (target). This is particularly evident in the more accurate representation of speed variations and congestion patterns across both the M3 and M1 freeway corridors.

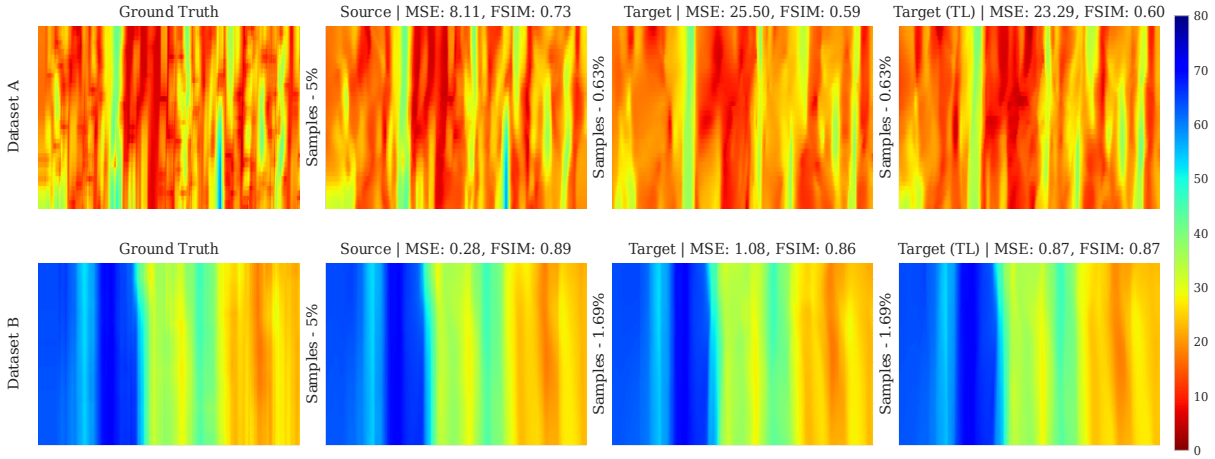



Figure 8.3: Comparison of traffic speed predictions using different Res-PINN models. The source model is trained on a richer dataset (A.2), while the target model is trained on a limited dataset (A.1). The target (TL) model utilises transfer learning from the source model.

## 8.3 Conclusion

The proposed Res-PINN model and transfer learning approach enable accurate traffic speed estimation using limited connected vehicle data by leveraging physical knowledge and learned features from data-rich scenarios. The incorporation of residual blocks improves the model's convergence speed and performance compared to traditional PINN architectures. The transfer learning strategy significantly enhances the estimation accuracy when dealing with limited data availability, as demonstrated by the results on the M1 and M3 freeway corridors in Melbourne, Australia.

The findings presented in this section demonstrate the effectiveness of physics-informed learning and knowledge transfer in enhancing traffic state reconstruction. This approach paves the way for more efficient intelligent transportation systems, enabling better traffic management and control strategies. Future research directions should include extending the Res-PINN model and transfer learning approach to larger road networks, investigating the integration of additional data sources to enhance estimation accuracy, and developing real-time implementations for practical deployment in intelligent transportation systems.

## 8.4 Summary of key findings

1. The study proposes a novel Residual Physics-Informed Neural Network (Res-PINN) architecture for traffic state estimation using connected vehicle data.
2. The Res-PINN model incorporates the governing physical laws of traffic flow into the neural network training process to improve accuracy.
3. The study explores **transfer learning to leverage knowledge from data-rich scenarios** and improve Res-PINN performance in situations with limited data availability.
4. Real-world data from two key corridors in Melbourne, Australia (Westgate Bridge on M1 freeway and a section of M3 freeway) was used. Two datasets were created: "probe" (0.62% and 1.69% of total data points for M3 and M1) and "random" (5% of total data points).
5. The Res-PINN model outperforms baseline models, achieving lower Mean Squared Error (MSE), lower Mean Absolute Percentage Error (MAPE), and higher Feature Similarity Index Measure (FSIM).
6. Transfer learning significantly improves speed estimation performance on the probe dataset, reducing MSE by 8.7%, MAPE by 11.5%, and increasing FSIM by 2.2%.
7. The predicted speed distributions using transfer learning exhibit a closer resemblance to the ground truth compared to models trained solely on the limited probe dataset.

# 9 Driving style variability at intersection

Road **intersections** are pivotal points in urban transport systems where safety risks, emissions, and fuel consumption are notably higher due to frequent fluctuations in vehicle acceleration, deceleration, and queues. Understanding the diverse driving styles and speed distributions at these intersections is essential for accurately forecasting safety risks, emissions, and fuel consumption. Moreover, this understanding can inform educational initiatives aimed at promoting safe and sustainable driving practices. This study aims to uncover spatiotemporal driving behaviour patterns at intersections by leveraging telematics data and intersection signal settings.

A case study in Melbourne was conducted, utilising over 41 million path trajectories from 2022, provided by Compass IoT. Additionally, signal setting data was extracted from the Sydney Coordinated Adaptive Traffic System (SCATS) to enrich the analysis. By examining the interaction between vehicles and traffic signals, this study seeks to identify distinct driving styles and their frequency of occurrence, offering valuable insights for urban planning and traffic management.

## 9.1 Methods

To analyse driving behaviours at intersections, Dynamic Time Warping (DTW) was employed for sequential clustering of path trajectories based on speed attributes. The DTW algorithm allows for the comparison of time-series data that may vary in speed or duration, providing a more meaningful measure of similarity between sequences than traditional metrics like Euclidean distance. This method is particularly suitable for capturing the temporal dynamics of driving behaviours.

The clustering process involved integrating the K-means algorithm with DTW. The initial centroids were selected using the k-means++ initialisation method to ensure a diverse representation of the data. During the assignment step, each sequence was assigned to the nearest centroid based on the DTW distance. The update step utilised the DTW Barycentre Averaging (DBA) technique to find the sequence that minimised the total DTW distance within each cluster, serving as the new centroid. The assignment and update steps were iterated until convergence.

To determine the optimal number of clusters (k), both the Elbow Method and the Silhouette Score were used. The Elbow Method involved plotting the within-cluster sum of squares (WCSS) against the number of clusters to identify the ‘elbow’ point. The Silhouette Score measured how similar an object was to its own cluster compared to other clusters, providing a robust indicator of cluster quality. The integrated approach combined these methods to select the most suitable number of clusters.

Traffic signal data was incorporated into the trajectory records by dividing intersection legs into cells. Each leg was segmented into 40-metre cells, corresponding to the data update frequency of Compass IoT. The signal state (green, red, or yellow) for each trajectory was determined based on the vehicle’s location and the corresponding SCATS data.

## 9.2 Findings and Discussion

**Clustering Results**

The suggested sequential clustering approach is employed for every scenario, determining the optimal number of clusters, the representative trajectory for each cluster, and the cluster sizes. Table 9.1 presents scenario attributes alongside the best cluster number (K), associated inertia, and silhouette score for each scenario's K value. It's notable from the table that each scenario comprises five clusters, indicating five distinct driving behaviours within intersections.

Table 9.1. Clustering Results. C: Car, CV: Commercial vehicle, G: Green, NRH: Non-rush hours, RH: Rush hour, S: Straight, R: Red, Y: Yellow.

| Count | Hour type | vehicle type | Movement direction | Traffic signal | K | Inertia score | Silhouette score |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1885 | NRH | C | S | G | 5 | 0.25 | 0.23 |
| 2049 | RH | C | S | G | 5 | 0.24 | 0.24 |
| 2780 | NRH | CV | S | G | 5 | 0.23 | 0.18 |
| 2766 | RH | CV | S | G | 5 | 0.21 | 0.22 |
| 1172 | NRH | C | S | R | 5 | 0.18 | 0.18 |
| 1122 | RH | C | S | R | 5 | 0.20 | 0.23 |
| 2476 | NRH | CV | S | R | 5 | 0.16 | 0.21 |
| 2346 | RH | CV | S | R | 5 | 0.17 | 0.18 |
| 53 | NRH | C | S | Y | 5 | 0.18 | 0.19 |
| 67 | RH | C | S | Y | 4 | 0.18 | 0.25 |
| 97 | NRH | CV | S | Y | 5 | 0.18 | 0.17 |
| 69 | RH | CV | S | Y | 4 | 0.21 | 0.25 |

The table also highlights an important observation regarding scenarios where drivers encounter yellow signals: a limited number of observations (less than 100) have been identified. This outcome was expected due to the short duration of yellow signals, resulting in fewer vehicles encountering them. However, it's essential to note that because scenarios with yellow signals are based on a smaller sample size compared to scenarios with green or red signals, there might be some limitations in the reliability and representativeness of the clustering results.

**Driving Style**

The analysis revealed sixteen distinct driving scenarios, each characterised by different combinations of vehicle type, time (rush hour or non-rush hour), and traffic signal status (green, yellow, or red). The key findings from the clustering analysis are summarised as follows:

**Green Signal Phases**

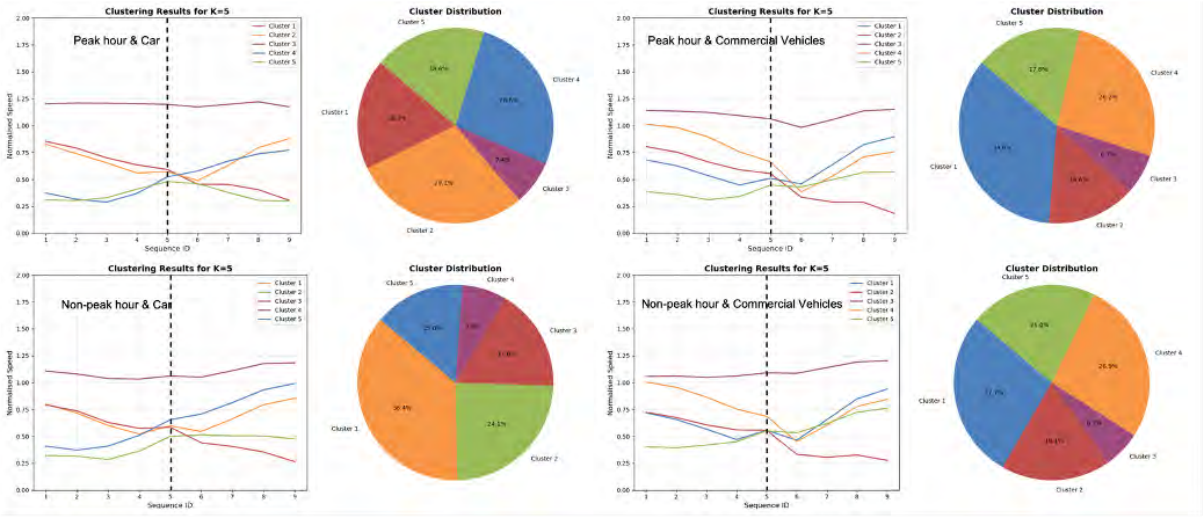



Figure 9.1: Clustering results for scenarios with green signal phases

According to Figure 9.1, During green signal phases, vehicles generally maintained speeds between 50% to 70% of the speed limit. A small cluster (6.2% to 7.4%) of drivers exceeded the speed limit, with this risky behaviour more prevalent among passenger cars (7.4%) compared to commercial vehicles (6.2% to 6.7%). Drivers of commercial vehicles, including both light and heavy commercial vehicles, demonstrated increased caution. This behaviour is likely attributed to the higher perceived consequences of accidents and the challenges associated with sudden braking.

For commercial vehicles, two predominant clusters were identified, reflecting common driving patterns. These drivers decelerated upon approaching the intersection and accelerated upon exiting. Differences between clusters suggested varying levels of caution, with some drivers slowing down earlier and approaching the intersection at lower speeds.

Passenger car drivers exhibit distinct behaviours compared to those observed with commercial vehicle operators. The orange cluster resembled the cautious driving style of commercial vehicles, while the blue cluster represented drivers who maintained very low speeds before accelerating cautiously through the intersection. This variation could be attributed to the diverse socioeconomic characteristics of passenger car drivers.

**Red Signal Phases**

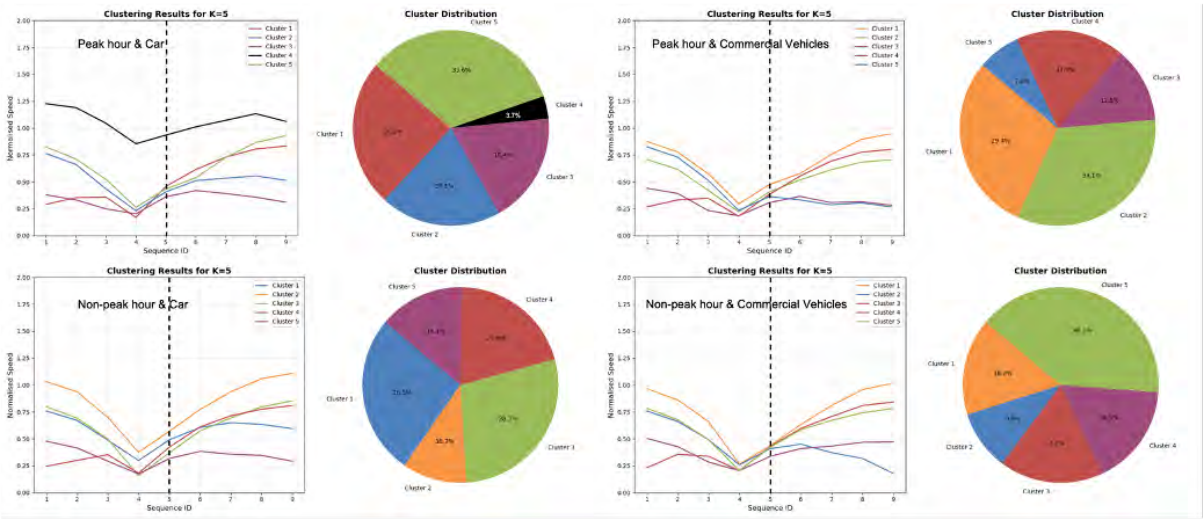



Figure 9.2: Clustering results for scenarios with red signal phases

Figure 9.2 shows that during red signal phases, a significant cluster (3.7%) of passenger cars was observed crossing intersections during red signals, particularly during rush hours. This behaviour poses substantial safety risks and represents a traffic violation. Commercial vehicles, in contrast, did not exhibit this behaviour, indicating greater adherence to traffic rules.

Three major clusters, represented by green, orange, and red colours, encompassed 60% to 80% of driving styles during red signals. These clusters indicated cautious driving, with drivers decelerating to near zero before crossing the intersection and then accelerating gradually. These clusters can be aggregated for inclusion in simulation software and traffic analysis.

Two additional clusters (blue and purple) represented drivers who maintained slow speeds consistently without significant speed reduction before or after the intersection. These behaviours were more prevalent among passenger cars and during rush hours, suggesting that traffic congestion might influence this driving style.

**Yellow Signal Phases**

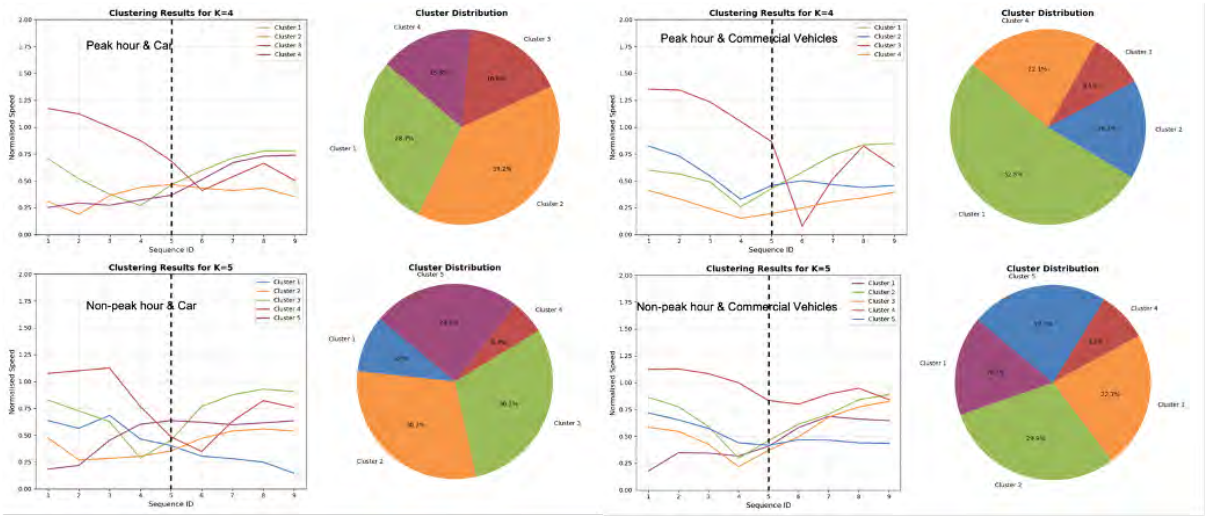



Figure 9.3: Clustering results for scenarios with yellow signal phases

According to Figure 9.3, during yellow signal phases, most drivers adopted cautious behaviours, stopping behind the intersection. This low-risk style was predominant across all scenarios, particularly among commercial vehicles during rush hours, with over 74% of drivers adhering to this approach.

Conversely, a risky driving style was observed, where drivers crossed the intersection at high speeds during the yellow signal and then decelerated afterward. This behaviour was more common during rush hours and among passenger cars, though commercial vehicles showed greater restraint.

## 9.3 Conclusion

Understanding and analysing speed profiles and driving styles within intersections is crucial for enhancing safety and efficiency in urban environments. This research employed telematics data and intersection signal settings, using sequential clustering to reveal distinct driving behaviours at intersections. The study provides valuable insights for urban planning, traffic management, and educational initiatives to promote safe and sustainable driving practices.

## 9.4 Summary of key findings

1. **Distinct driving styles at intersections**: The study identifies 16 distinct driving scenarios, each characterised by different combinations of vehicle type, time, and traffic signal status.
2. **Green signal phases**: During green signal phases, most vehicles maintain speeds between 50% to 70% of the speed limit, while a small cluster of drivers exceed the speed limit, more prevalent among passenger cars.
3. **Red signal phases**: During red signal phases, a small yet significant cluster of passenger cars cross intersections during red signals, posing safety risks, while commercial vehicles show greater adherence to traffic rules.
4. **Yellow signal phases**: During yellow signal phases, most drivers adopt cautious behaviours, stopping behind the intersection, while a risky driving style is observed, more common during rush hours and among passenger cars.
5. **Driving style variability**: The study reveals variability in driving styles, with commercial vehicles showing increased caution and passenger cars exhibiting diverse behaviours.
6. **Rush hour and vehicle type effects**: Rush hour and vehicle type have significant effects on driving behaviours, with passenger cars more likely to engage in risky behaviours during rush hours.
7. **Intersection safety and efficiency**: Understanding driving styles at intersections is crucial for enhancing safety and efficiency in urban environments.

# 10 Estimating emissions using traffic reconstructed trajectories

In this section, we present emission estimations using reconstructed vehicle trajectories (presented in Section 4). The study aims to showcase the value of vehicle trajectory data at scale and assess the impact of varying traffic conditions on emission levels based on real trajectories, as opposed to simulated trajectories. This method is particularly effective in capturing spatio-temporal variability in traffic flow, congestion, and queues, and their subsequent impact on emissions.

Key aspects of the study include:

1. **Impact of Traffic Condition on Emission**: Using real trajectory data allows for a more accurate assessment of traffic flow variations and their impact on emission levels. This research addresses the gap in emission estimation by utilising dense trajectory data for reconstructing traffic states, unlike previous studies that relied on sparse or simulated data.
2. **Spatio-Temporal Variability of Emission**: The approach captures the dynamic nature of traffic, offering insights into how emissions vary over time and space.
3. **Effect of Electric Vehicles (EVs) on Emission**: By modelling different compositions of EVs within the existing fleet, the study examines how varying percentages of EV adoption influence overall emissions.

Understanding traffic dynamics is crucial for optimising transportation systems and enhancing urban mobility. Vehicle trajectory reconstruction plays a significant role in analysing fuel consumption and emissions, particularly at intersections. The primary objective of this study is to provide insights into how effective traffic management and the adoption of sustainable transportation options, such as EVs, can mitigate environmental impacts.

By leveraging real trajectory data, this research offers valuable information for developing strategies aimed at reducing emissions and promoting sustainable urban transport solutions.

## 10.1 Case study with reconstructed trajectories:

The effect of varying traffic conditions on emissions is a critical aspect of our study. This approach significantly improves the accuracy of emission estimations by leveraging dynamic, real-world traffic conditions derived from reconstructed trajectories, thanks to the unprecedented potential provided by connected vehicle (CV) data.

Furthermore, this method assists in understanding how emissions vary over time and across different locations along road segments, considering factors such as traffic volume, congestion, and queues. We used emission profiles based on existing trajectories as the baseline and explored the impact of different compositions of electric vehicles (EVs) within the current fleet. Specifically, we simulated scenarios where **10% and 20% of the existing fleet were randomly transformed into EVs**. Emissions were then estimated and compared to the baseline case to evaluate the effects of these changes.

This approach replicates various levels of traffic congestion and dynamics to analyse their impact on traffic emission. Figure 10.1 showcases a sample of these reconstructed trajectories within the time window of 8:00 am to 8:30 am on a typical weekday. In the figure, red bars indicate periods when the traffic signal is red at the stop line, while the absence of these bars signifies green light phases, allowing vehicles to proceed through the intersection. The black lines represent the actual observed paths from connected vehicles, whereas the grey lines show the reconstructed trajectories.

## 10.2 Emission estimation using optimised trajectories

To evaluate the emission a second-by-second emission function has been used that were developed from the vehicle trajectory data (Panis et al., 2006). The emission functions for each vehicle are formulated using instantaneous speed and acceleration as parameters through non-linear multiple regression techniques. The equation is below:

$$E_n(t) = \max\left[E_0, f_1 + f_2 v_n(t) + f_3 v_n(t)^2 + f_4 a_n(t) + f_5 a_n(t)^2 + f_6 v_n(t) a_n(t)\right] \quad \text{(Eq. 10.1)}$$

where $v_n(t)$ and $a_n(t)$ are the instantaneous speed and acceleration of vehicle $n$ at time $t$ . $E_0$ is a lower limit of emission (g/s) specified for each vehicle and pollutant type, and $f_1$ to $f_6$ are emission constants specific for each vehicle and pollutant type determined by the regression analysis. Table 10.1 gives the insights of the CO2 emission functions.

After reconstructing the vehicle trajectories based on various traffic states at different times of the day and days of the week, we calculated the emissions for each scenario. This analysis helps us examine how different traffic states and conditions affect emissions.

Table 10.1: CO2 Emission functions.

| Pollutant | Vehicle type | $E_0$ | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ |
|---|---|---|---|---|---|---|---|---|
| $CO_2$ | Petrol car | 0 | 5.53e−01 | 1.61e−01 | −2.89e−03 | 2.66e−01 | 5.11e−01 | 1.83e−01 |
| | Diesel car | 0 | 3.24e−01 | 8.59e−02 | 4.96e−03 | −5.86e−02 | 4.48e−01 | 2.30e−01 |
| | LPG car | 0 | 6.00e−01 | 2.19e−01 | −7.74e−03 | 3.57e−01 | 5.14e−01 | 1.70e−01 |
| | HDV | 0 | 1.52e+00 | 1.88e+00 | −6.95e−02 | 4.71e+00 | 5.88e+00 | 2.09e+00 |
| | Bus | 0 | 9.04e−01 | 1.13e+00 | −4.27e−02 | 2.81e+00 | 3.45e+00 | 1.22e+00 |

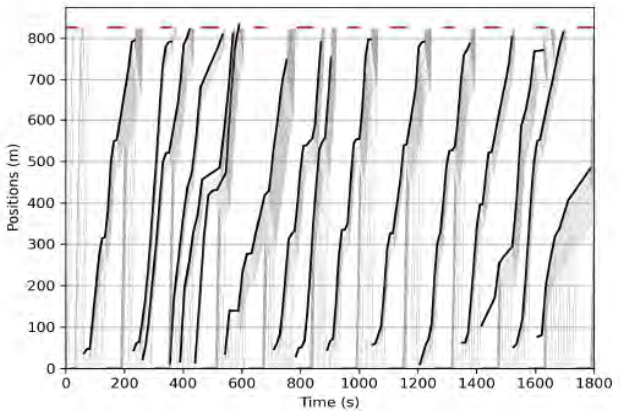



Figure 10.1: Reconstructed Trajectories Based on CV Data, Signal Information, and Traffic Count Data.

### *10.2.1 Impact of Different Traffic Condition on Emission*

Traffic conditions significantly influence vehicle emissions, as demonstrated by the patterns observed in both weekday and weekend data, see Figure 10.2 and Figure 10.3.

During peak hours, particularly on weekdays, emissions exhibit high variability and intensity, especially upstream of intersections. This fluctuation is primarily due to the unpredictable nature of traffic congestion, stop-and-go traffic, and queue formation. As vehicles approach intersections and slow down, emission levels tend to decrease. This pattern highlights the importance of using real trajectory data to capture the complex relationship between traffic dynamics and emissions, as simulation-based methods may not fully represent these nuances.

Conversely, during off-peak hours and in less congested periods, emission profiles show a more uniform distribution across time and space. This uniformity is attributed to smoother traffic flow, with fewer interruptions and less frequent acceleration/deceleration cycles. The contrast between peak and off-peak emissions underscores the significant role that traffic management can play in reducing overall vehicle emissions.

1. Morning peaks on weekdays produce more variable emissions compared to afternoon peaks, which show a more uniform distribution, see Figure 10.2.
2. Weekend peaks, while following similar patterns to weekdays, occur later in the morning and exhibit lower overall emission fluctuation due to reduced traffic variability, refer to Figure 10.3.
3. The relationship between vehicle speed, acceleration, and emissions is evident, with emissions decreasing as vehicles slow down near intersections or at the end of queues.

### *10.2.2 Spatio-temporal Variability of Emission*

The analysis reveals distinct spatio-temporal patterns in vehicle emissions. Emission levels are closely linked to specific times of day and days of the week.

On weekdays, emission hotspots are concentrated during early morning and late afternoon peak hours, corresponding to typical commute times. These hotspots are particularly pronounced far-distant upstream of intersections, where traffic congestion and queue formation are more likely to begin. The spatial distribution of emissions during these peaks is characterised by higher intensities at a distance from intersections, gradually decreasing as vehicles approach and slow down.

Weekends display a different temporal pattern, with peak emissions occurring later in the morning, typically between 10:30 and 11:30 AM. Despite this shift in timing, the spatial distribution of emissions on weekends follows a similar pattern to weekdays, with higher intensities upstream of intersections during peak hours.

1. The spatio-temporal variability of emissions is closely linked to traffic volume and flow conditions.
2. Off-peak periods, both on weekdays and weekends, show more consistent emission patterns across time and space, indicating the potential benefits of encouraging off-peak travel.
3. The observed patterns suggest that targeted measures addressing high-emission zones and peak traffic conditions could be more effective in reducing overall emissions.
4. The use of real trajectory data reveals the complex interplay between traffic dynamics and emissions, capturing localised effects that might be missed in more generalised models.

These findings underscore the importance of considering both spatial and temporal factors in urban planning, traffic management, and emission reduction strategies. They also highlight the potential for data-driven approaches based on real trajectory data to inform more effective and tailored interventions for improving air quality in urban environments.

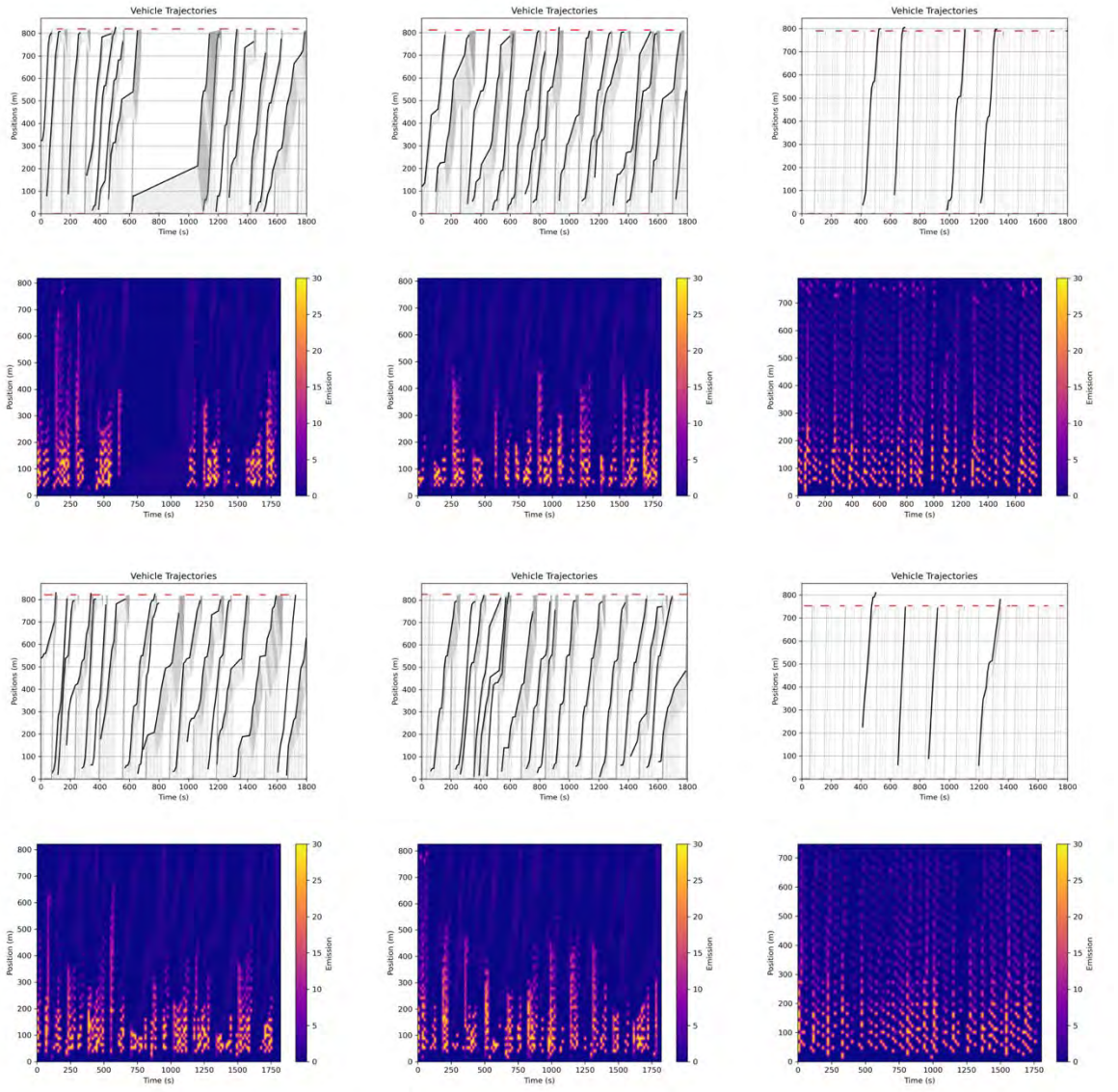



Figure 10.2: Vehicle trajectories and emission heatmaps for weekdays (2022-09-26 and 2022-09-27), showing peak periods from 07:30 to 08:30 (left & centre) and off-peak period from 21:30 to 22:00 (right).

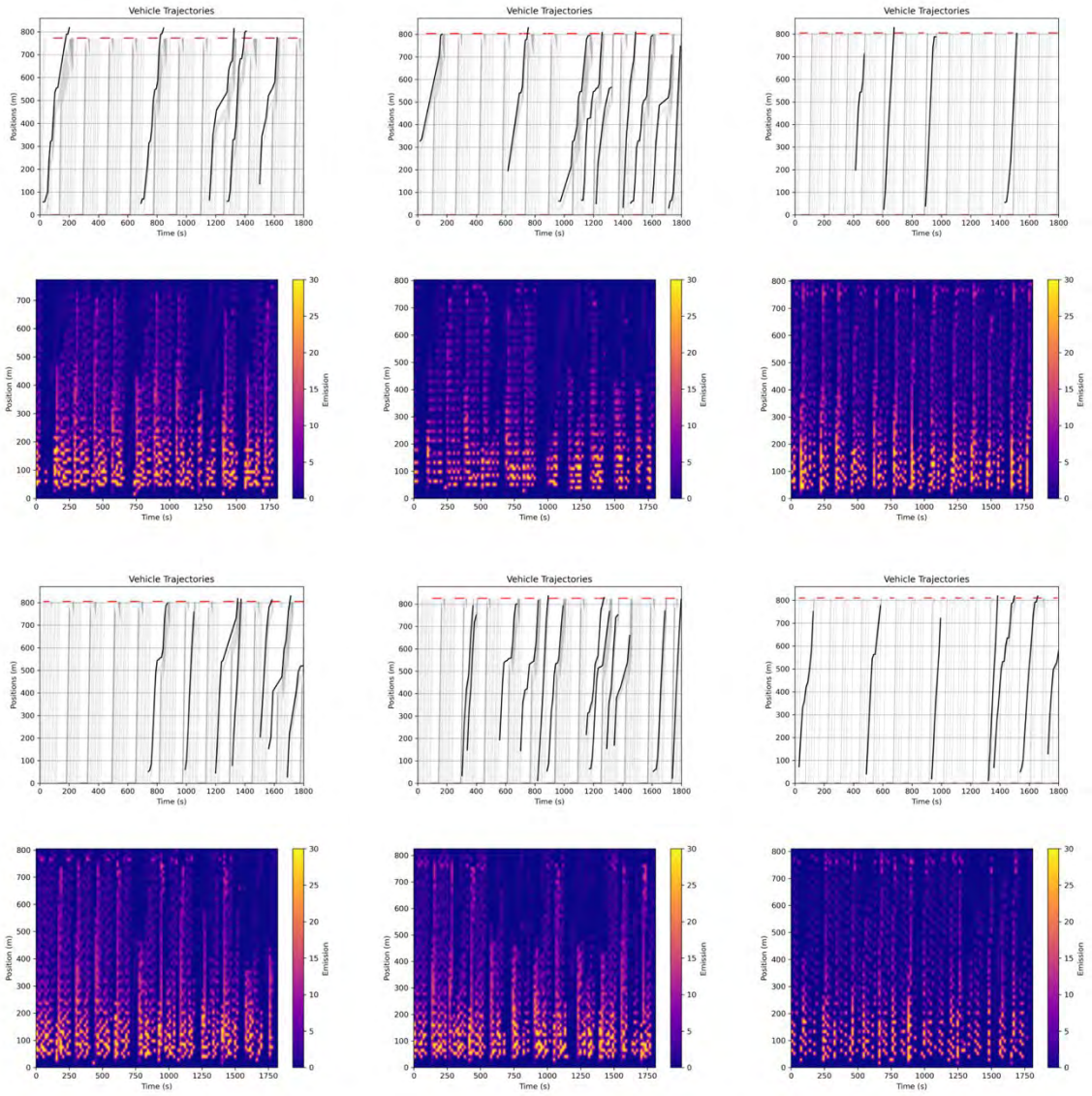



Figure 10.3: Vehicle trajectories and emission heatmaps for weekends (2022-09-24 and 2022-09-25), depicting peak periods from 10:30 to 11:30 (left & centre) and off-peak period from 21:30 to 22:00 (right).

### *10.2.3 Effect of Electric Vehicles (EVs) on Emission*

In this section, we evaluated different scenarios with varying electric vehicle (EV) compositions to assess their impact on emissions within the existing fleet based on real-world trajectories. Two scenarios were developed and compared to a baseline scenario. The baseline represents the current situation with only conventional internal combustion engine vehicles (non-EVs), depicted in black in Figure 10.4. The two alternative scenarios involve a random conversion of 10% and 20% of the existing fleet to EVs. These scenarios are based on a reasonable assumption that EV penetration in the market could reach approximately 10% and 20% over the next 5 to 10 years, respectively. This analysis helps us understand the potential impact on total daily emissions.

As shown in Figure 10.4, the emission reductions roughly correspond to the percentage of EVs in the fleet, as applied to a real-world case study on Hoddle Street in Melbourne, Australia. For instance, a 10% conversion to EVs results in approximately a 10% reduction in emissions. With an increase in the EV composition to 20%, there is a 25.2% reduction in emissions during weekends and a 24.9% reduction during weekdays with a 20% EV composition. This indicates a non-linear improvement, suggesting further research may be required to better understand the exact effect of EV's on emissions.

These findings, however, demonstrate that penetration of EVs in the fleet can significantly reduce emissions, contributing to greener, more sustainable, and lower-carbon transport systems. Additionally, the data reveals no significant difference in emission reductions between weekends and weekdays as EV composition increases. This consistency suggests that the benefits of EV adoption in terms of emission reductions may be robust across different days of the week.

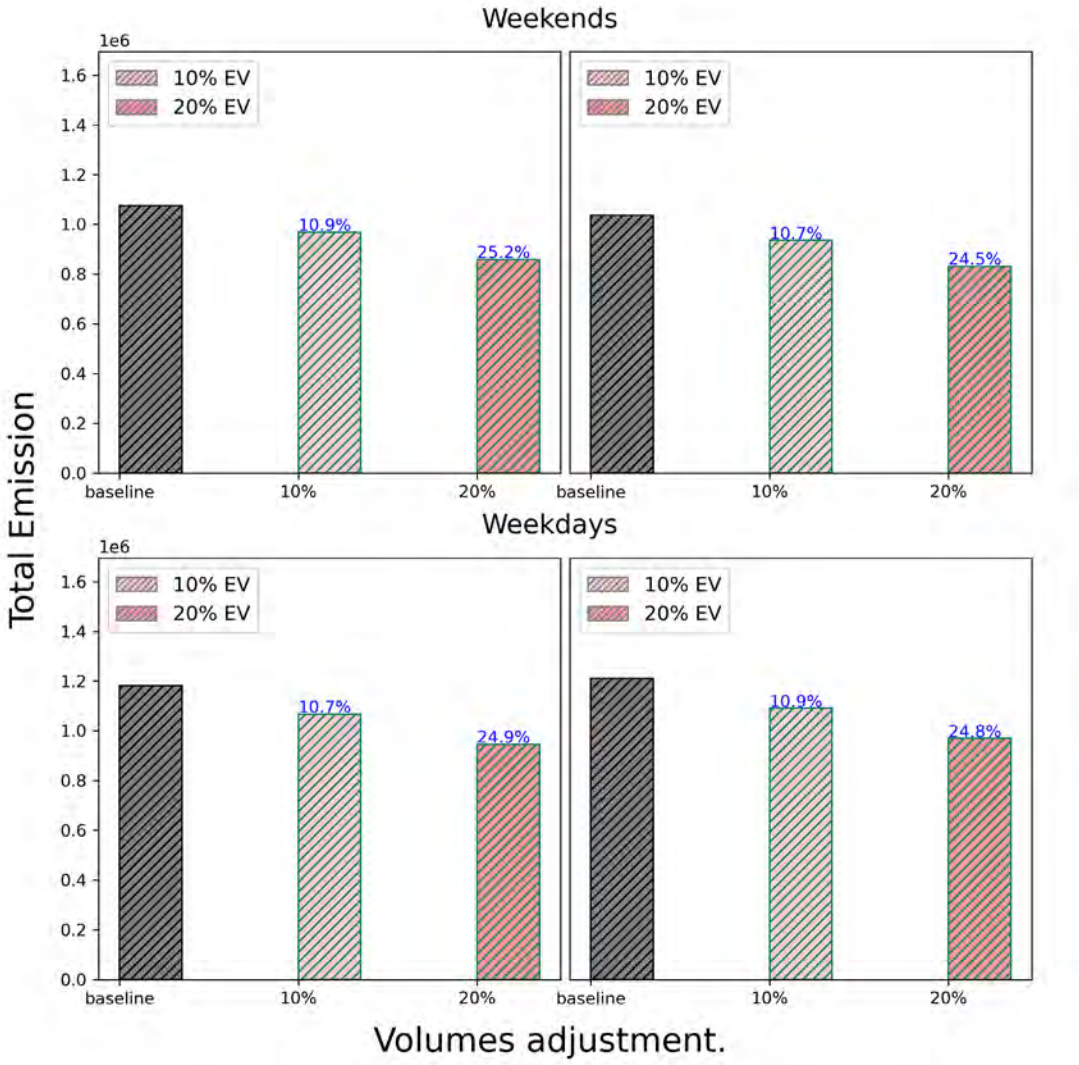



Figure 10.4: Impact of electric vehicle (EV) adoption on total emissions (Metric Ton of CO2) for weekends and weekdays

We also examined emissions with different EV compositions using high-resolution data to better understand the impact of EVs and their spatio-temporal distribution. Two sets of metrics were employed to study this effect in greater detail.

The first metric used is emission (grams of CO2) per vehicle per second, as shown in Figure 10.5. It is important to note that the emission estimates are calculated using Equation 10.1, which is based on vehicle speed and acceleration. In other words, emissions decrease when vehicles move at slower speeds or are decelerating.

From Figure 10.5, it is evident that on average, each vehicle produces less emission per second during rush hours. This trend is observed on weekdays, which exhibit two peaks (morning and afternoon), and on weekends, which show a single peak during late morning. Conversely, during off-peak hours, each vehicle produces more emissions on average, likely due to higher speeds and accelerations when traffic is less

congested. These findings suggest that implementing effective speed limits could help reduce emissions, in addition to increasing the adoption of EVs.

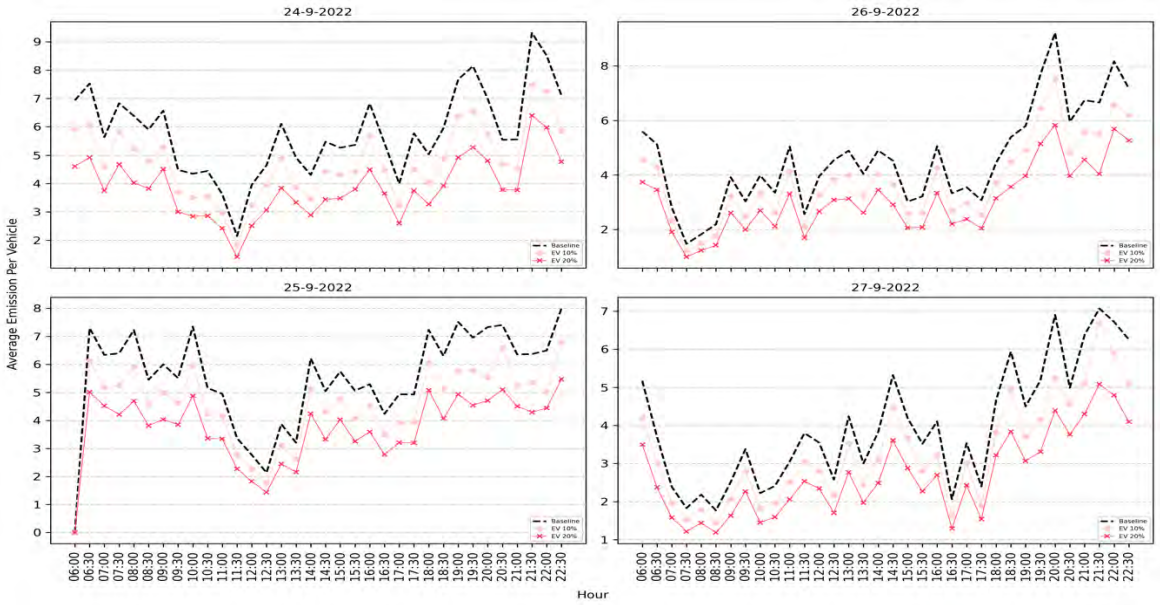



Figure 10.5: Average emission (grams of CO2) per vehicle per second

Given that vehicles produce more emissions during off-peak hours, we were also interested in comparing the impact of additional vehicle volume during peak hours with lower volume but higher emission rates per vehicle during off-peak hours. To explore this, a second metric, emission per meter, was used to show how emissions change at different times of the day.

As shown in Figure 10.6, the overall emission per meter is higher during peak hours compared to off-peak hours. Although vehicles on average produce less emission during peaks due to lower speeds and acceleration, the higher number of vehicles during peak times results in overall higher emissions, despite the lower emission rate per vehicle. The data also show that increasing the proportion of EVs results in smoother and significantly lower emissions across different days, both on weekdays and weekends.

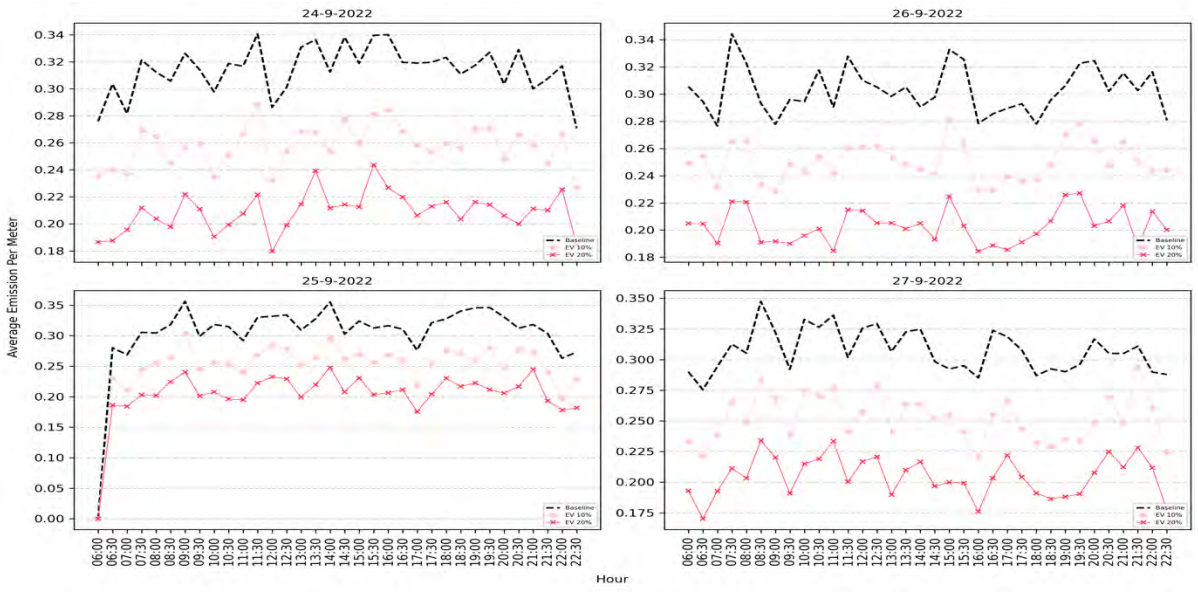



Figure 10.6: Average emission (grams of CO2) per vehicle meter travelled

These findings not only highlight the positive impact of increasing the EV composition in the fleet but also underscore the importance of speed management in controlling emissions. Therefore, a combination of

effective speed management and increased EV adoption could serve as an optimal strategy for reducing emissions.

The most important observation to be made from this analyses is perhaps the opportunity to use vehicle trajectory data and fusion models with count data, for measurement of traffic emissions.

## 10.3 Summary of key findings

The study provides an unprecedented understanding of tailpipe emissions at scale, utilising real-world data and capturing the true dynamics on the road, which offers superior insights compared to traditional simulation models. The key findings emphasise the significant role of traffic volume and congestion in determining CO2 emissions, with a clear distinction between the effects of these two factors.

**Impact of traffic volume and congestion on emissions:**

1. Real trajectory data captures emission intensity and variability more accurately than simulated methods.
2. Unpredictable traffic congestion leads to variable emissions. Reducing congestion and flow uncertainty is crucial for improving emission profiles.
3. Emissions are closely tied to traffic volume and flow, varying by time of day and day of the week. Morning peaks on weekdays produce more variable emissions compared to afternoon peaks, which show a more uniform distribution.
4. Weekend peaks, while following similar patterns to weekdays, occur later in the morning and exhibit lower overall emission fluctuation due to reduced traffic variability.
5. The relationship between vehicle speed, acceleration, and emissions is evident, with emissions decreasing as vehicles slow down near intersections or at the end of queues.

**Spatio-temporal variability of emission:**

6. Peak emissions concentrate at specific times and locations, with highest intensity far-distant upstream of intersections, decreasing as vehicles slow down approaching intersections or queue ends.
7. Both emission intensity and variability are higher during peak hours on weekend and weekdays.
8. Off-peak periods, both on weekdays and weekends, show more consistent emission patterns across time and space, indicating the potential benefits of encouraging off-peak travel.
9. The use of real trajectory data reveals the complex interplay between traffic dynamics and emissions, capturing localised effects that might be missed in more generalised models.

**Effect of Electric Vehicles (EVs) on Emission**

10. **Impact of Increased EV Adoption**: Introducing 10% and 20% EV compositions in the vehicle fleet significantly reduces emissions, highlighting the importance fleet composition in total traffic emissions.
11. **Emission Per Vehicle Per Second**: Emissions per vehicle per second are lower during rush hours due to reduced speeds and acceleration but increase during off-peak hours when vehicles travel (and/or accelerate) faster. This suggests that effective speed management can further reduce emissions alongside EV adoption.
12. **Emission Per Meter**: Despite lower emissions per vehicle during peak hours, overall emissions per meter are higher due to increased traffic volume and congestion. The adoption of EVs smooths and lowers emissions consistently throughout the day, both on weekdays and weekends.
13. In conclusion, combining effective speed management strategies with increased EV adoption provides an optimal solution for minimising emissions and promoting a sustainable, low-carbon transportation system.

# 11 Visualisation and interactive dashboard

Data visualisation plays a crucial role in transforming complex information into easily understandable and actionable insights. In the context of traffic management, visualising traffic data can help identify patterns, trends, and anomalies that may not be apparent in raw data formats. To harness the power of data visualisation and enable traffic managers and researchers to gain valuable insights more effectively, we developed an interactive data visualisation dashboard that focuses on evaluating traffic speed reliability using connected vehicle data.

Although connected vehicle data in this project only represents around 2% of total traffic, they capture important traffic features and provide valuable insights into traffic conditions. By leveraging this data source, the dashboard aims to provide a user-friendly interactive interface for exploring and understanding traffic speed reliability, making it accessible to a wider audience beyond data analysts and experts. The dashboard presents key traffic metrics, patterns, and trends in a visually intuitive manner, facilitating data-driven decision-making and supporting real-time monitoring and analysis of traffic conditions.

The current demo version of the dashboard introduces two main functions: traffic overview and speed reliability evaluation. These functions provide users with a high-level understanding of traffic conditions and performance, as well as the ability to assess the reliability of travel times on specific routes based on the connected vehicle data. By focusing on this representative sample of traffic data, the dashboard enables users to gain actionable insights into traffic speed reliability, which can help inform traffic management strategies and improve overall traffic conditions.

## 11.1 Traffic Overview

The traffic overview page (Figure 11.1) serves as the landing page for the dashboard, providing users with a summary of key traffic metrics and conditions. The page consists of three interactive visualisation subplots that display different aspects of traffic information.

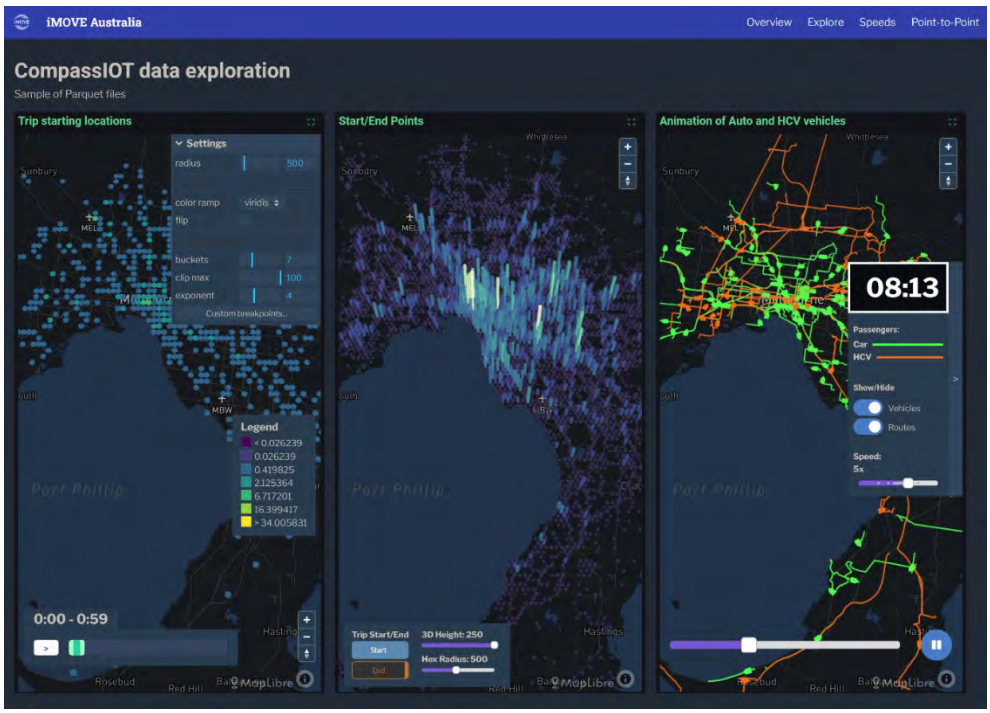



Figure 11.1: Traffic overview page of the iMOVE Australia data exploration dashboard. The page includes three interactive subplots showing (a) trip starting locations, (b) start/end points, and (c) an animation of auto and HCV vehicles.

The first subplot on the left shows the trips starting locations, with the map cantered around the Melbourne metropolitan area. Users can adjust the map view by zooming in or out and moving the map. The subplot includes a settings panel that allows users to customise the visualisation, such as changing the radius of the plotted points, the colour ramp, and the number of buckets for data aggregation.

The central subplot displays the start and end points of trips, with the ability to select a specific trip using the "Trip Start/End" drop-down menu. The selected trip is highlighted on the map, showing the start and end locations and the connecting route. The map height can be adjusted using the "3D Height" slider, providing a 3D perspective of the trips.

The third subplot on the right presents an animation of auto and HCV (heavy commercial vehicle) movements over time. The animation shows the vehicles' trajectories, with different colours representing cars and HCVs. The subplot includes a timeline slider, allowing users to control the animation and view vehicle positions at specific times. The legend indicates the types of vehicles shown (cars and HCVs), and the animation speed is displayed in the lower-right corner.

All three subplots are interactive and synchronised, meaning that when a user changes the position or zoom level of one map, the other two maps will update accordingly. This enables users to maintain a consistent view across all subplots while exploring different aspects of the traffic data.

## 11.2 Travel Speed Reliability Evaluation

The travel speed reliability evaluation feature allows users to assess the consistency and predictability of travel times between specific locations. This feature is available in two modes: single-point evaluation (Figure 11.2) and point-to-point evaluation (Figure 11.3).

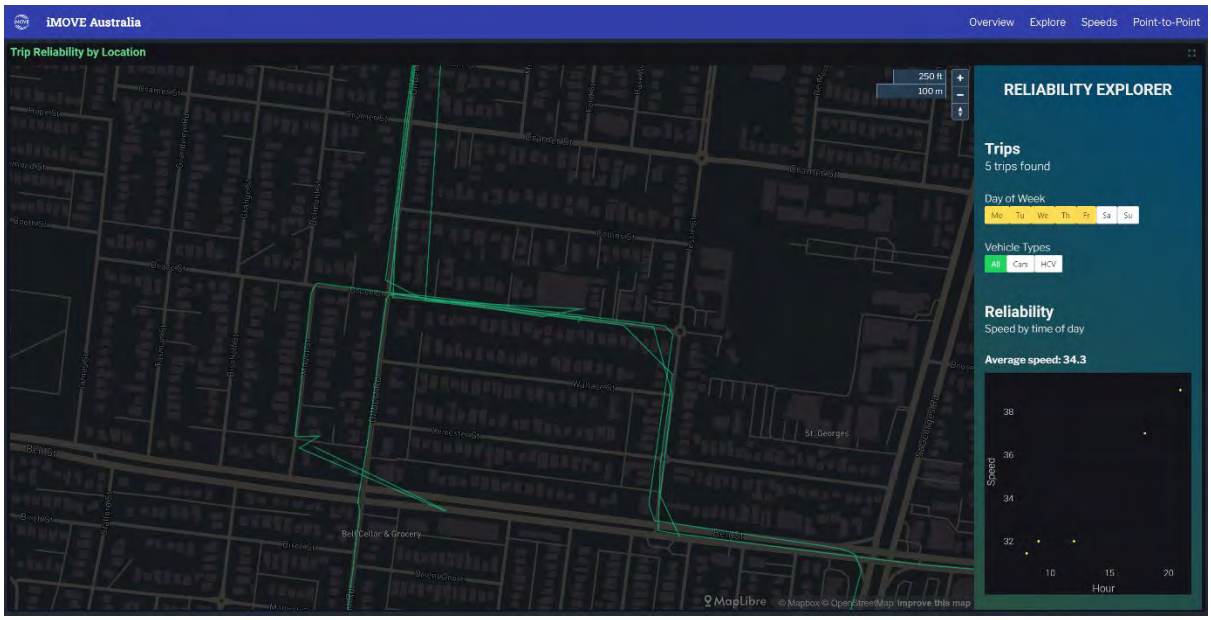



Figure 11.2: Single-point travel speed reliability evaluation. The map displays trips passing through a selected location, with the reliability explorer showing speed statistics for the chosen day of the week and vehicle type.

In the single-point evaluation mode, users can select a location on the map and specify conditions such as the day of the week and vehicle type (e.g., cars or HCVs). The dashboard then queries the database for trips passing through the selected location that match the specified criteria. The resulting trips are visualised on the map, and their travel speed statistics, including the average speed and speed distribution, are displayed in the "Reliability Explorer" panel on the right side of the screen.

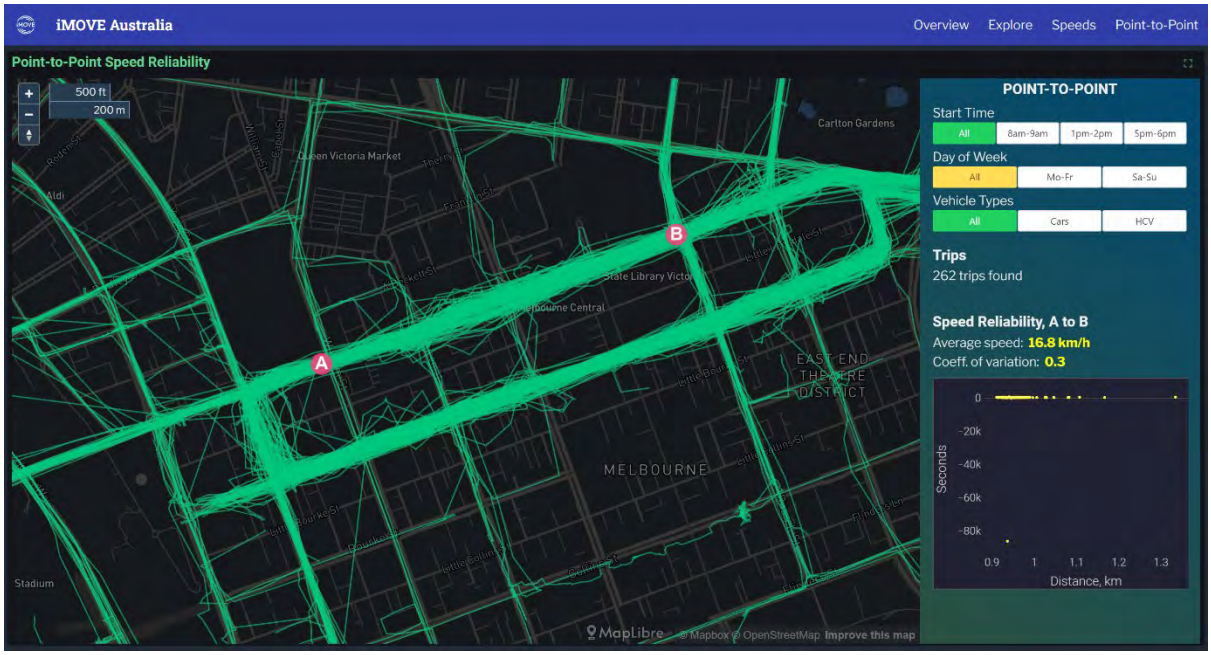


Figure 11.3: Point-to-point travel speed reliability evaluation. The map shows trips between two selected points (A to B) on La Trobe Street in Melbourne CBD, with the speed reliability statistics displayed in the bottom-right corner.

The point-to-point evaluation mode allows users to select two points (A to B) on the map and specify the desired conditions, such as the start time, day of the week, and vehicle type. In the example shown in Figure 11.3, the selected points are located on La Trobe Street in the Melbourne CBD. The dashboard queries the database for trips that match the specified criteria and travel between the selected points. Particularly, Figure 11.3 highlights the trips found between points A and B, while the "Point-to-Point" panel on the right side displays the speed reliability statistics for the selected trips. In this example, the average speed for cars traveling from point A to point B on La Trobe Street is 16.8 km/h, with a coefficient of variation (a measure of speed variability) of 0.3. The speed profile graph in the bottom-right corner shows the distribution of speeds along the selected route, allowing users to identify potential bottlenecks or congestion points.

By providing both single-point and point-to-point evaluation modes, the travel speed reliability feature enables users to assess travel time consistency and variability at specific locations or along specific routes. This information can be valuable for traffic managers and planners in identifying problem areas, evaluating the effectiveness of traffic management strategies, and communicating expected travel times to the public.

# 12 Conclusion

This research project aimed to investigate opportunities for effectively utilising and integrating vehicle and bicycle location and gyroscope data, alongside other emerging technologies and data sources, into advanced traffic management systems and practices. The scope of the project included the analysis of connected vehicle (CV) data, GPS, and gyroscopic data from bicycles to improve safety, efficiency, and environmental sustainability in urban traffic systems. The primary objective was to develop advanced methodologies, identify new use cases, and derive actionable insights for traffic management.

The methodology adopted in this project involved a multi-stage framework, beginning with an extensive exploratory analysis of vehicle position data. This analysis covered penetration measures, intersection turn volumes, and speed profiles at intersections. High-resolution CV data was integrated with traditional sensor data, as well as bicycle telematics and historical crash data. Techniques such as sequential spatial mapping, clustering, deep reinforcement learning (DRL), and physics-informed neural networks (PINNs) were employed to achieve objectives such as trajectory reconstruction, intersection safety and efficiency optimisation, proactive road safety measures for vulnerable road users, traffic state and congestion estimation, driving style variability at intersections, and emission estimation. A visualisation dashboard was also developed to facilitate real-time traffic monitoring.

The applications of integrated data span various aspects of traffic management. For instance, advanced AI models using CV data can optimise traffic signals, reducing delays and improving travel times. Surrogate safety measures derived from bicycle data can proactively identify hazardous locations, underrepresented risky behaviours and hotspots that are not detectable from crash data alone, and can be used proactively to prevent accidents. The visualisation dashboard facilitates real-time traffic monitoring, enhancing decision-making capabilities for traffic managers.

Analysis revealed that bicycle-involved crashes were more frequent at intersections (59%) compared to mid-block segments (41%), with a denser concentration in the CBD and city/town centres. Moreover, high braking events, indicative of potential crashes, were concentrated in the central metropolitan area, particularly around intersections and cycling paths. Gyroscope data identified significant variations in road surface quality, with poorer conditions often found in areas with complex traffic networks such as CBDs, city centres, and popular cycling tracks. This information is crucial for targeting road maintenance and improving cyclist safety.

Spatiotemporal correlation analysis conducted in this project showed that harsh braking and swerving events recorded by gyroscope sensors (for cars and bicycles) were strong indicators of potential crash sites. This surrogate measure can be used to proactively identify and mitigate dangerous locations. The combination of exploratory spatial and correlation analysis indicated a high correlation between gyroscope and GPS data with actual crash events, highlighting their potential as a supplemental tool for analysing safety in regions with numerous abnormal events. These surrogate safety measures can effectively identify hotspots that are underrepresented by police-reported crash data, providing a more comprehensive picture of road safety.

The study found that higher penetration rates of GPS-equipped vehicles provided broader data representativeness, which in turn aided more effective traffic management. The Residual Physics-Informed Neural Network (Res-PINN) model proposed in the study significantly improved traffic state estimation using limited connected vehicle data, particularly when combined with transfer learning techniques. This approach demonstrated lower error rates and a closer resemblance to real-world traffic conditions, making it a powerful tool for accurate traffic state reconstruction and congestion management.

The project also yielded significant insights into intersection efficiency and occupancy-based priority control. DRL models, particularly those enhanced with high-resolution connected vehicle data, outperformed traditional signal control methods, reducing average travel times and queue lengths. The enhanced computer vision models using an ensemble modelling approach, significantly improved conflict prediction accuracy. The study also identified distinct driving styles and behaviours at intersections,

highlighting variability based on vehicle type, signal phases, and time of day. Understanding these driving patterns is crucial for improving intersection safety and traffic flow efficiency.

Reconstructed vehicle trajectories were used to estimate emissions. This exercise showcased a very important use case linked to trajectory data and demonstrated the advantage of this data as compared to traffic simulation for assessment and management of traffic emissions. The study found that while CO2 emissions increased with surge in combustion-engine traffic volumes, the spatiotemporal distribution and intensity of emissions depend directly on traffic dynamics and speed profiles. Access to trajectory data would be crucial for accurate measurement of emissions and strategic planning to optimise traffic management for reduced emissions.

In conclusion, this project has successfully demonstrated the potential of integrating multi-source data for advanced traffic management applications. The findings highlight significant improvements in traffic efficiency, safety, and environmental sustainability through the use of connected vehicle data, telematic data from bicycles and other datasets. The methodologies developed provide a robust framework for future research and practical applications, offering substantial benefits for urban mobility and traffic management systems. This integrated approach paves the way for smarter, safer, and more efficient transportation networks.

# 13 References


Chen, P., Zeng, W., Yu, G., & Wang, Y. (2017). Surrogate Safety Analysis of Pedestrian-Vehicle Conflict at Intersections Using Unmanned Aerial Vehicle Videos. *Journal of Advanced Transportation*, *2017*, 1-12. https://doi.org/10.1155/2017/5202150

Chen, X., Qin, G., Seo, T., Tian, Y., & Sun, J. (2023). A Macro-Micro Approach to Reconstructing Vehicle Trajectories on Multi-Lane Freeways with Lane Changing. *arXiv preprint arXiv:2306.05627*.

Coppola, R., & Morisio, M. (2016). Connected car: technologies, issues, future trends. *ACM Computing Surveys*, *49*(3).

de Luca, S., Di Pace, R., Memoli, S., & Pariota, L. (2020). Sustainable Traffic Management in an Urban Area: An Integrated Framework for Real-Time Traffic Control and Route Guidance Design. *Sustainability*, *12*(2). https://doi.org/10.3390/su12020726

DTP. (2024). *Victoria Road Crash Data*.https://doi.org/https://discover.data.vic.gov.au/dataset/victoria-road-crash-data

Goodall, N. J., Smith, B. L., & Park, B. B. (2016). Microscopic Estimation of Freeway Vehicle Positions from the Behavior of Connected Vehicles. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *20*(1), 45-54. https://doi.org/10.1080/15472450.2014.889926

He, Q., Head, K. L., & Ding, J. (2012). PAMSCOD: Platoon-based arterial multi-modal signal control with online data [Article]. *Transportation Research Part C: Emerging Technologies*, *20*(1), 164-184. https://doi.org/10.1016/j.trc.2011.05.007

Hirose, T., Takada, T., Oikawa, S., & Matsui, Y. (2021). Validation of driver support system based on real-world bicycle and motor vehicle flows. *Accident Analysis & Prevention*, *156*, 106131. https://doi.org/https://doi.org/10.1016/j.aap.2021.106131

Hsueh, C., & Lin, J.-J. (2023). Influential factors of the route choices of scooter riders: A GPS-based data study. *Journal of Transport Geography*, *113*, 103719. https://doi.org/https://doi.org/10.1016/j.jtrangeo.2023.103719

Kapousizis, G., Ulak, M. B., Geurs, K., & Havinga, P. J. M. (2023). A review of state-of-the-art bicycle technologies affecting cycling safety: level of smartness and technology readiness. *Transport Reviews*, *43*(3), 430-452. https://doi.org/10.1080/01441647.2022.2122625

Liang, X., Guler, S. I., & Gayah, V. V. (2023). Decentralized arterial traffic signal optimization with connected vehicle information. *Journal of Intelligent Transportation Systems*, *27*(2), 145-160. https://doi.org/10.1080/15472450.2021.1990762

Liu, J., & Khattak, A. J. (2016). Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles. *Transportation Research Part C: Emerging Technologies*, *68*, 83-100. https://doi.org/https://doi.org/10.1016/j.trc.2016.03.009

Liu, Y., Zhang, Z., Han, L. D., & Brakewood, C. (2021). Automatic Traffic Queue-End Identification using Location-Based Waze User Reports. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(10), 895-906. https://doi.org/10.1177/03611981211013353

Mehran, B., Kuwahara, M., & Naznin, F. (2012). Implementing kinematic wave theory to reconstruct vehicle trajectories from fixed and probe sensor data. *Transportation Research Part C: Emerging Technologies*, *20*(1), 144-163. https://doi.org/https://doi.org/10.1016/j.trc.2011.05.006

Milkovits, M., Huang, E., Antoniou, C., Ben-Akiva, M., & Lopes, J. A. (2010). *DynaMIT 2.0: The Next Generation Real-Time Dynamic Traffic Assignment System*. 2010 Second International Conference on Advances in System Simulation.

Muthugama, L., Xie, H., Tanin, E., & Karunasekera, S. (2022). Real-time road safety optimization through network-level data management. *GeoInformatica*. https://doi.org/10.1007/s10707-022-00473-2

Int Panis, L., Broekx, S., & Liu, R. (2006). Modelling instantaneous traffic emission and the influence of traffic speed limits. *The Science of the total environment*, *371*(1-3), 270–285. https://doi.org/10.1016/j.scitotenv.2006.08.017Park, H. G., Kim, S., & Kim, T. (2023). Traffic-Responsive Signal Control at Intersections Using Real-Time Data of Vehicles Connected via V2X Communication [Article]. *Journal of Advanced Transportation*, *2023*, Article 4025210. https://doi.org/10.1155/2023/4025210

Pucher, J., & Buehler, R. (2017). Cycling towards a more sustainable transport future. *Transport Reviews*, *37*(6), 689-694. https://doi.org/10.1080/01441647.2017.1340234

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016, 27-30 June 2016). You Only Look Once: Unified, Real-Time Object Detection. (Ed.),^(Eds.). 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

SeeSense. (2024). *SeeSense GPS and Gyroscope Data*. Retrieved June 19 from https://seesense.cc/

Seo, T., Tago, Y., Shinkai, N., Nakanishi, M., Tanabe, J., Ushirogochi, D., Kanamori, S., Abe, A., Kodama, T., & Yoshimura, S. (2020). Evaluation of large-scale complete vehicle trajectories dataset on two kilometers highway segment for one hour duration: Zen Traffic Data. (Ed.),^(Eds.). 2020 International Symposium on Transportation Data and Modelling.

Strauss, J., Zangenehpour, S., Miranda-Moreno, L. F., & Saunier, N. (2017). Cyclist deceleration rate as surrogate safety measure in Montreal using smartphone GPS data. *Accident Analysis & Prevention*, *99*, 287-296. https://doi.org/https://doi.org/10.1016/j.aap.2016.11.019

Sunderrajan, A., Viswanathan, V., Cai, W., & Knoll, A. (2016). Traffic state estimation using floating car data. *80*, 2008-2018. https://doi.org/10.1016/j.procs.2016.05.521

Wan, N., Vahidi, A., & Luckow, A. (2016). Reconstructing maximum likelihood trajectory of probe vehicles between sparse updates. *Transportation Research Part C: Emerging Technologies*, *65*, 16-30. https://doi.org/10.1016/j.trc.2016.01.010

Wang, Y., Wei, L., & Chen, P. (2020). Trajectory reconstruction for freeway traffic mixed with human-driven vehicles and connected and automated vehicles. *Transportation Research Part C: Emerging Technologies*, *111*, 135-155. https://doi.org/10.1016/j.trc.2019.12.002

WHO. (2018). *Global status report on road safety 2018* (

Yuan, Y., Zhang, W., Yang, X., Liu, Y., Liu, Z., & Wang, W. (2021). Traffic state classification and prediction based on trajectory data. *Journal of Intelligent Transportation Systems*, 1-15. https://doi.org/10.1080/15472450.2021.1955210

Zakhem, M., & Smith-Colin, J. (2021). Micromobility implementation challenges and opportunities: Analysis of e-scooter parking and high-use corridors. *Transportation Research Part D: Transport and Environment*, *101*. https://doi.org/10.1016/j.trd.2021.103082

Zamboni, S., Kefato, Z. T., Girdzijauskas, S., Norén, C., & Dal Col, L. (2022). Pedestrian trajectory prediction with convolutional neural networks. *Pattern Recognition*, *121*. https://doi.org/10.1016/j.patcog.2021.108252

Zhao, P., Haitao, H., Li, A., & Mansourian, A. (2021). Impact of data processing on deriving micro-mobility patterns from vehicle availability data. *Transportation Research Part D: Transport and Environment*, *97*. https://doi.org/10.1016/j.trd.2021.102913

Zhao, Y., Zheng, J., Wong, W., Wang, X., Meng, Y., & Liu, H. X. (2019). Various methods for queue length and traffic volume estimation using probe vehicle trajectories. *Transportation Research Part C: Emerging Technologies*, *107*, 70-91. https://doi.org/10.1016/j.trc.2019.07.008

Zheng, J., & Liu, H. X. (2017). Estimating traffic volumes for signalized intersections using connected vehicle data. *Transportation Research Part C: Emerging Technologies*, *79*, 347-362. https://doi.org/10.1016/j.trc.2017.03.007

# 14 Appendix

The following papers have been produced from this project and are under peer-review to be published in scientific journals of ITS.

1. Wang, B, Nassir, N, Yousefpour, N, Sarvi, M (2024) *Investigating Knowledge Transfer in Residual Physics-Informed Neural Networks using Connected Vehicles Traffic Data*
2. Tom, Z, Mohri, S, Nassir, N, Sarvi, M (2024) Understanding Driving Styles at Intersections
3. Leo, M, Chan L, Rezazada, M, Nassir, N, Sarvi, M (2024) *Predicting traffic conflicts using vehicle trajectory plots and object detection algorithms*
4. Ghanbari, M, Roy, K, Tom, Z, Nassir, N, Sarvi, M (2024) Emission Evaluation for Connected Vehicles via Traffic State Reconstruction at Signalised Intersection