SOME STATISTICAL TECHNIQUES FOR ANALYSING BLUETOOTH TRACKING DATA IN TRAFFIC MODELLING

A thesis presented in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
in
Statistics
at Massey University, Palmerston North,
New Zealand.

Ghazaleh Aslani

2021

Contents

List of Tables

List of Figures

Abstract

The economy and the environment are both affected by traffic congestion. People spend time stuck in traffic, which limits their free time. Every city's road infrastructure is under increased pressure, particularly in large cities, due to population growth and vehicle ownership patterns. Therefore, traffic control and management are crucial to reducing traffic congestion problems and effectively using existing road infrastructure.

Bluetooth is a commonly used wireless technology for short distance data exchange. This technology allows all mobile phones, GPS systems, and in-vehicle applications such as navigation systems to connect with the personal devices of drivers and passengers. A Media Access Control (MAC) address is a unique electronic identifier used by each Bluetooth device. The concept is that, while a Bluetooth-equipped device travels along a road, its MAC address, detection time, and location can be detected anonymously at different locations. Bluetooth technology can be integrated into Intelligent Transportation Systems (ITS) to enable better and more effective traffic monitoring and management, hence reducing traffic congestion.

This thesis aims to develop some statistical methods for analysing Bluetooth tracking data in traffic modelling. One of the challenges of using Bluetooth data, particularly for travel time estimation, is multiple Bluetooth detections, which occur when a Bluetooth sensor records a Bluetooth device several times while it passes through the detection zone. We employ cluster analysis to look at the possibility of extracting meaningful traffic information from multiple detections, and the observed gap distribution, which is the time difference between records when multiple detections occur. We also develop a novel regression method to investigate the relationship between data from Bluetooth and Automatic Traffic Counts (ATCs) through weighted regression analysis, in order to explore potential causes of bias in the representativeness of Bluetooth detections. Finally, we seek the practical objective of recovering ATC from Bluetooth data as a statistical calibration problem, following the development of a new time-varying coefficients Poisson regression model.

Acknowledgements

A famous poem by Persian poet Saadi Shirazi:

> Human beings are members of a whole,
> In creation of one essence and soul.
> If one member is afflicted with pain,
> Other members uneasy will remain.
> If you have no sympathy for human pain,
> The name of human you cannot retain.

This poem, I believe, expresses all of the feelings that we all had throughout this difficult pandemic time. I also believe that appreciating one another is one of the best ways to show respect to all of the important people in our lives. For me, there are so many people I would like to thank.

My sincere thanks to my main supervisor, Professor Geoff Jones, for supervision and all the advice, encouragement, and feedback given to me throughout this research. Your enthusiasm for developing main ideas, defining methodologies, and analyzing results, combined with your wonderful personality, has always provided me with inspiration, positive feelings, and motivation on this journey.

My sincere thanks to my co-supervisor, Dr. Katarina Hannaford, for providing me with the opportunity to do this PhD, and for providing me with amazing guidance and superb friendship all the way through. Your support for me is priceless and very much appreciated.

My sincere thanks to my co-supervisor, Dr. Xun Xiao, I cherish the impact you have made on my research, your invaluable feedbacks, and continuous help to improve the quality of my work. I have learnt a lot from you.

This study could not have been conducted without the data provided by Transport for Greater Manchester (TfGM), and I am grateful that they allowed me access to their data. I would like to express my gratitude to Richard Dolphin and especially Alex Jirat, ITS Engineer at TfGM. I have not had a chance to meet him in person, but he always answered all of my questions kindly and very comprehensively during my research.

Numerous people have provided me with support and input throughout my PhD journey. I would like to offer my special thanks to all my wonderful friends in NZ and the staff members of the School of Fundamental Sciences for friendship and sharing valuable experiences not only around work and science, but also in life. I have enjoyed living every moment of the last few years of my life with all of you.

My dear mother passed away eight years ago. I owe my whole life to her love and energy. I miss her and regret she did not live to see my achievements, but I am sure I have her blessings today and in every moment of my life. I would like to thank my father and my lovely sister who always support me in every decision in my life with their unconditional love.

Finally, to my best friend and my husband, Ahmad, to whom I can not find the words to express my gratitude and love. Your encouragement and positive and hard-working attitude are always my emotional support, and help me to stay motivated. Thank you for being patient and understanding with me while I was stressed. To me, you are everything.

Chapter 1

Introduction

As the world's population and use of private vehicles continues to increase, road congestion creates multiple challenges. In this regard, urban transportation infrastructure has a significant impact on travellers' economic, environmental, psychological, and stress levels. As a result, urban traffic monitoring, strategic management, and transportation planning are essential to reducing traffic congestion. All of them are time-consuming and costly, which also requires close collaboration among traffic administrative agencies, intelligent transport systems (ITS), and traffic experts. For example, traffic management systems reduce traffic congestion by monitoring, optimising subsystems (such as traffic signals), and controlling traffic on road networks (Diebold, 1995; Chen and Miles, 2004). Traffic management requires traffic data collection in order to process and manage various strategies for optimising traffic flow and reducing congestion (Hounsell et al., 2009).

Traditional data collection methods, such as loop detectors, Automatic Number Plate Recognition (ANPR) cameras, Automatic Traffic Counters (ATC) etc., are still developing, but they are costly in terms of both supply and maintenance (Leduc, 2008).

The potential of other technology choices, such as wireless communications, offers further possibilities for the improvement of existing systems with low-power and cost-effective sensors. In early 2000, researchers began to examine Bluetooth for monitoring the movement of vehicles and employing it to ITS (Nusser and Pelz, 2000; Kasten and Langheinrich, 2001; Murphy et al., 2002; Sawant et al., 2004; Friesen and McLeod, 2015). Bluetooth is a wireless communication technology in order to connect many different types of devices such as smartphones, wireless headsets, tablets, heart monitors, medical equipment, in-vehicle navigation, etc. (Haartsen, 1998). Figure 1.1 shows some examples of activated Bluetooth devices used in vehicles, such as in-vehicle navigation, wireless headset, and smartphone.

Figure 1.1: Examples of Bluetooth devices used in vehicles (Source:postoaktraffic.com).

Researchers have been interested in the use of Bluetooth technology as a source of transportation data to monitor traffic conditions, and the Bluetooth sensor is mentioned as having lower-cost installations and maintenance (Puckett and Vickich, 2010). It is also capable of extracting essential traffic data, including travel time, origin-destination (OD) matrices, and speed to be obtained across networks (Barcelö et al., 2010; Puckett and Vickich, 2010; Malinovskiy et al., 2011; Barceló Bugeda et al., 2012a; Michau et al., 2014; Purser, 2016; Michau, 2016; Cotten et al., 2020; Liu et al., 2020). Michau et al. (2014) used data from Bluetooth sensors in Brisbane to create OD matrices. The Bluetooth detector's coverage region is usually divided into smaller geographic zones, with two potential origin and destination locations. By matching the Bluetooth detections between these two locations, the elements of these matrices represent a census of the volume of trips from origin to destination points. After collecting the data from the Bluetooth detectors, it can be used to obtain vehicle travel times between detectors or average speed of vehicles, and the traffic density on the particular road section covered by the Bluetooth detector (Puckett and Vickich, 2010; Malinovskiy et al., 2011; Barceló Bugeda et al., 2012b; Laharotte et al., 2014; Tahmasseby, 2015; Purser, 2016; Zhou et al., 2016).

The majority of research has concentrated on using Bluetooth data to estimate travel time and speed in order to evaluate traffic conditions on a particular road segment (e.g. between two Bluetooth detector locations)(Puckett and Vickich, 2010; Malinovskiy et al., 2011; Díaz et al., 2015; Purser, 2016; Erkan and Hastemoglu, 2016; Zhou et al., 2016; Cotten et al., 2020; Liu et al., 2020). For example, Malinovskiy et al. (2011) explored travel time estimation on a short corridor and compared Bluetooth travel time with travel time estimated by Automated Licence Plate Recognition (ALPR) sensors. The results of this study showed that a larger detection zone is preferable, and so a shorter corridor will result in more travel time errors. Quayle et al. (2010) investigated arterial travel time by comparing Bluetooth and GPS data and concluded that Bluetooth has the ability to accurately measure travel time over long spans of time. Their research was conducted on suburban signalized arterial roads in Portland, Oregon. Bachmann et al. (2013) combined Bluetooth data with loop detector data to estimate freeway traffic speeds and showed that using Bluetooth data and probe

data like GPS can improve estimation. This research was conducted on a freeway rather than on urban roads, which have different characteristics. Díaz et al. (2015) investigated commercial Bluetooth detectors in actual traffic situations on a freeway. The resulting Bluetooth-enabled traffic monitoring system produced highly reliable 5-minute travel time estimations. Liu et al. (2020) also looked into how accurate Bluetooth travel time estimates are in urban arterial areas, considering two major challenges: the multiple detection problem and errors in Bluetooth estimates. When a discoverable Bluetooth device is recorded several times by a Bluetooth sensor while it passes across the detection zone, the multiple detection problem refers to the choice of detections that should be used to calculate travel time estimates. They demonstrated that accurate Bluetooth-based travel time information on signalised arterial roadways can be achieved if a proper matching method is used to smooth out the errors in the travel time estimates, such as average-to-average and last-to-last matching methods.

Since different travel modes (e.g., vehicles, bicycles, and pedestrians) cause different travel times, Araghi et al. (2012) used clustering techniques such as hierarchical clustering, K-Means clustering, and two-step clustering to test the feasibility of using Bluetooth data to estimate mode-specific travel times for different travel modes. Crawford et al. (2018) also looked into the possibility of identifying road user classes based on their frequent travels and classifying them into three categories: infrequent, frequent, and very frequent.

The goal of this study is to look into some statistical methods for analysing Bluetooth tracking data in traffic modelling. In this regard, in the rest of this chapter we will present more details on how Bluetooth detectors operate, introduce properties and some challenges of the data they collect.

1.1 Structure of thesis

The majority of research has concentrated on using Bluetooth data to estimate travel time and speed in order to evaluate traffic conditions on a particular road segment (e.g. between two Bluetooth detector locations). As noted above, multiple detection is one of the challenges that most research, particularly in the area of travel time estimation, tries to handle by choosing one detection and filtering out the rest. There is a time difference between records when multiple detections happen, resulting in a gap distribution between detections. This research first aims to investigate the possibility of obtaining meaningful information (e.g. traffic conditions) from multiple detections and the observed gap distribution in a particular Bluetooth coverage zone.

In Chapter 2, we investigate this possibility in a particular Bluetooth coverage zone. Cluster analysis is a popular unsupervised learning method which can be used

to analyze data and identify underlying patterns or groupings. Therefore, the first approach to investigating the goal is to perform cluster analysis based on multiple detections and the gap distributions to categorize Bluetooth detector sites, MAC addresses, and time intervals of a day. To cluster distributions, we utilize the Kolmogorov-Smirnov statistic.

In Chapter 3, we will investigate the relationship between ATC and Bluetooth detections, which may help us investigate potential causes of bias in Bluetooth detections' representativeness. We will utilize regression analysis for modelling the relationship, taking into account that some observable factors may influence the rate of Bluetooth detection. We develop a methodology incorporating a non-parametric estimate of the variance function to explore some alternative models for the relationship.

In Chapter 4, we will develop a Poisson regression model to describe the rate of Bluetooth detection per vehicle as it varies over time. We will examine the practical goal of recovering ATC from Bluetooth data following the development of an appropriate regression model associated with the statistical calibration problem. The goal will be to predict the unknown ATC value based on the number of Bluetooth counts during a particular time of day. Finally, Chapter 5 reviews the dissertation's outcomes and presents some recommendations for further research.

1.2 Bluetooth technical overview

Bluetooth devices are designed to communicate with and then connect with other Bluetooth devices which are in close proximity. An active Bluetooth device can be detected by a unique Media Access Control Identification address (MAC address), which is a combination of six alphanumeric pairs. The first three pairs are related to the manufacturer and are allocated by the Institute of Electrical and Electronics Engineers (IEEE). The last three pairs are defined by the manufacturer. For example, 7C:6B:9C:39:64:0C is a sample MAC address, where 7C:6B:9C is the manufacturer's Organization Unique Identifier (OUI), and 39:64:0C is the device's unique ID assigned by the manufacturer as the device's series number. Despite the fact that MAC addresses are supposed to be unique, Michau et al. (2014) have discovered that some MAC addresses are shared among cars. For example, some MAC addresses are shared by taxi drivers. In the Manchester Bluetooth network, similar shared MAC addresses also happens. One explanation is the ability to clone Bluetooth device characteristics for fleet-specific requirements (Cherchali et al., 2010).

Bluetooth uses a radio technology called frequency-hopping spread spectrum over short distances, from a minimum of 1 meter to more than 100 meters. Depending on the class radio implementation, the range may be differentiated into three categories (see Table 1.1).

Class	Transmission power	Range
Class 1	100 mW	100 m
Class 2	2.5 mW	10 m
Class 3	1 mW	1 m

Table 1.1: Bluetooth classes, Source:(Frodigh et al., 2000).

In Bluetooth, a piconet is a basic network consisting of two main objects: master unit and slave unit. These networks can vary in size. Figure 1.2 shows sample of three different piconets.

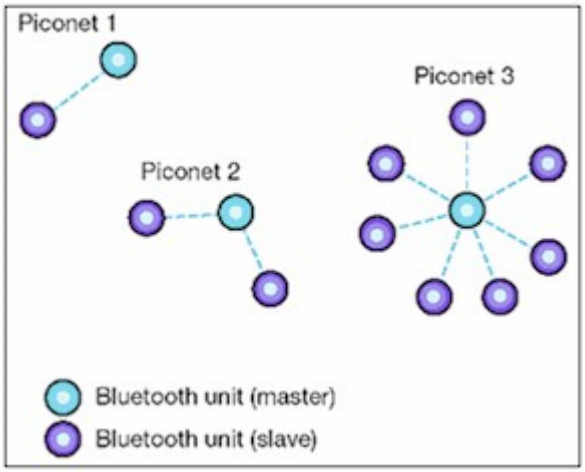


Figure 1.2: Sample of three different piconets.

In the piconet, the master unit acts as the main controlling unit and the other devices that follow the master unit are slave units. The frequency hopping sequences to enable the synchronization between the master and the slave devices are controlled by the master device. The communication is just between master to slave or slave to master, and there is no connection between slaves in a piconet. Note that there can be up to seven active slaves participating in a piconet at the same time, but with only one master. A scatternet is a collection of multiple piconets with overlapping areas that can connect with one another via a shared node. Figure 1.3 depicts an example of a scatternet with one slave unit as a shared node.

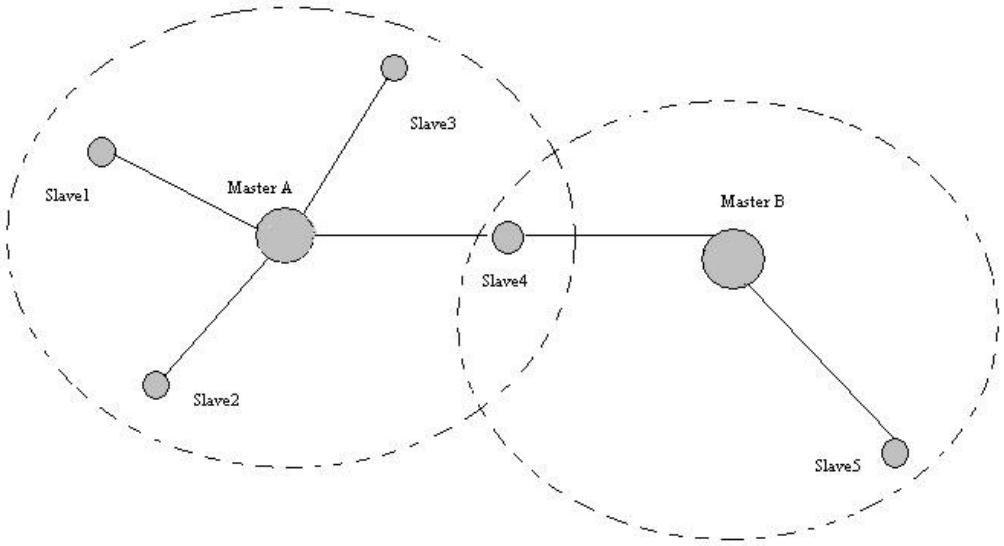


Figure 1.3: Sample of scatternet (Source: sites.google.com/site/securezrp/introduction).

The Bluetooth device has two main states, standby and connect. The default state of the Bluetooth unit is standby and in this state, it is in a low-power mode in which no transmissions occur. The standby state changes to a connected state if there is an active connection to the Bluetooth unit.

For making a connection, Bluetooth follows two procedures: inquiry (or discovering) procedure and paging (or connecting) procedure. The inquiry procedure is designed to scan for other devices within range to discover each other. During the inquiry process, one Bluetooth device (the master) sends out the inquiry request and other Bluetooth devices (the slaves) will respond with their address and possibly their name and other information. In fact, in the inquiry process, the master unit invites the slave units to create a piconet. After completing the inquiry cycle, the paging procedure creates a connection between the master and the slave devices.

When a Bluetooth device is set to the inquiry mode, it continuously sends out inquiry packets called ID (identifier) packets via one of the 32 predefined inquiry channels to detect potential slaves in the neighborhood, and scans for replies. In this phase, after sending an ID packet, the master listens for response packets called frequency hopping sequence (FHS) from the active slave devices. The FHS packet contains information about the slave unit, such as its own address and clock values.

The Bluetooth protocol recommends an inquiry cycle of 10.24 seconds, and during this time period, it is highly likely that an active Bluetooth device within the communication zone of the detector is detected (Peterson et al., 2006; Kasten and Langheinrich, 2001).

The main important factor that may affect the quality of the MAC address data collection process is the characteristics of the scanners used as Bluetooth detectors. For

example, it is important that a Bluetooth detector be able to cover the whole zone even with the different kinds of environmental obstacles, such as trees, buildings, and other physical structures that can interfere with wireless communication. There are two main characteristics of the scanners: the type of antenna and its gain (strength). Basically, directional and omni-directional are two types of antennas. The difference between these two types is that omni-directional antennas send and receive signals from any direction and directional antennas only cover one direction and limited angles. The gain defines the size of the coverage and is called decibels-isotropic (dBi). The scanner range is the maximum distance from the scanner, along with a given direction, over which the scanner can communicate with active Bluetooth devices. Porter et al. (2013) categorized the six different types of Bluetooth antennaes in terms of their capability and suitability for the quality of the data collected, and the results indicated that omni-directional antennas with a gain of 9 to 12 dBi are good choices for Bluetooth data collection.

1.3 Bluetooth for traffic monitoring

In transport applications, Bluetooth detectors sited at locations on the traffic network act as a master to acquire the MAC addresses of the active Bluetooth devices on the road as slaves within their communication zone. Therefore, the Bluetooth detector repeatedly conducts an inquiry process to detect any Bluetooth devices that are within its antenna coverage area. The active Bluetooth devices in the coverage zone will respond to this inquiry by sending a data package containing the MAC address. Note that in transportation applications, only the inquiry process is needed and Bluetooth scanners never make a full connection with an available Bluetooth device.

Figure 1.4 shows the detection area of two detectors (the green and red dot) sited to pick up all vehicles passing through the junction with an activated Bluetooth device.

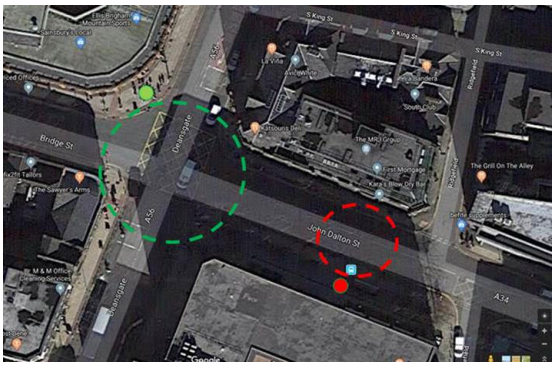
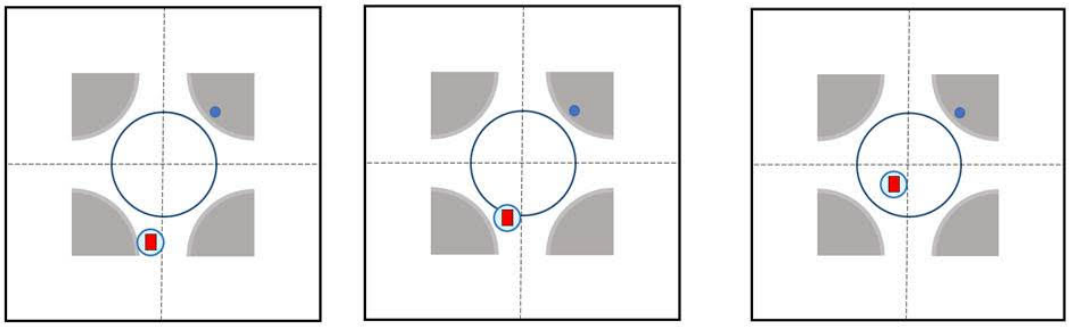


Figure 1.4: Detection zone.

Figure 1.5 demonstrates the detection of a sample car (red rectangle) with an activated Bluetooth device, before entering the detection zone and after leaving. As the car

(red rectangle) is outside of the Bluetooth pick-up area, it would not be recorded. The car will be detected even when it is partially in the coverage zone and will still record until it leaves the coverage zone. Therefore, a device can be detected multiple times if it remains within the coverage zone. In fact, there is a strong correlation between the length of time that a device remains within the detection zone and the number of detection records (Moghaddam and Hellinga, 2014).



(a) **(b)** **(c)**

Figure 1.5: (a) The car is outside of the Bluetooth pick-up area; (b) the car is partially in the pick-up area;(c) the car is now fully in the pick-up area.

Figure 1.6 presents the operational concept of collecting traffic data using Bluetooth detectors. As Figure 1.6 shows, the detectors are deployed on the roadside and they can detect Bluetooth-enabled devices passing within their coverage zone. If the device's MAC address is observed at two consecutive Bluetooth detectors, then travel time and the average speed for this vehicle over the road segment between these two detectors can be calculated. As a result, processing similar data from a larger number of vehicles represents a sample of the vehicle population and provides for estimation of traffic conditions on this road segment.

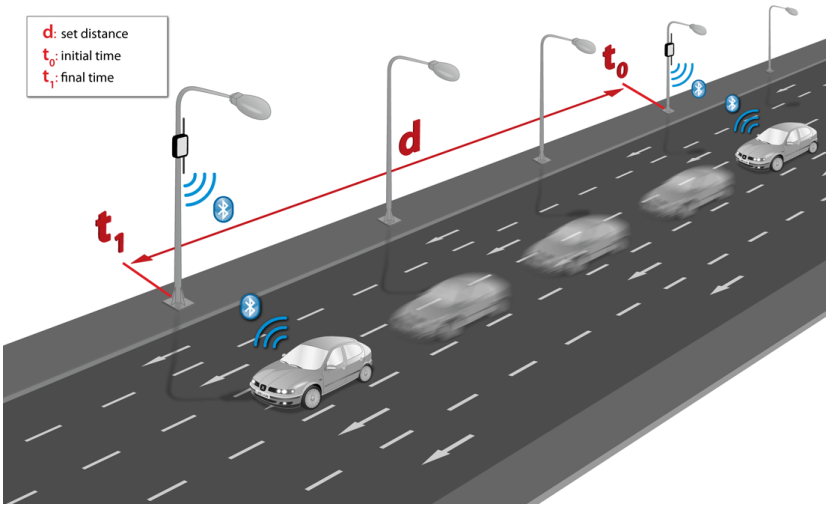


Figure 1.6: The operational concept of collecting traffic data using Bluetooth detectors (Source:www.libelium.com).

1.4 Data description

The data used in this research was collected from Bluetooth MAC scanners installed on urban roadways in Manchester by Transport for Greater Manchester (TfGM). Since 2011, TfGM has been installing fixed Bluetooth detectors on main arterial roads and around key urban centres such as Manchester, Wigan, and Rochdale. Around 525 Bluetooth detectors owned by TfGM are operating on the Greater Manchester network and actively recording data. In the case of the first analysis, we chose Manchester city. Figure 1.7 shows a part of the Manchester Bluetooth sites, where the green circles represent the implemented Bluetooth detectors, and the blue and purple circles represent the permanent and temporary Automatic Traffic Counters (ATC), respectively. Figure 1.8 displays a Bluetooth detector in Manchester.

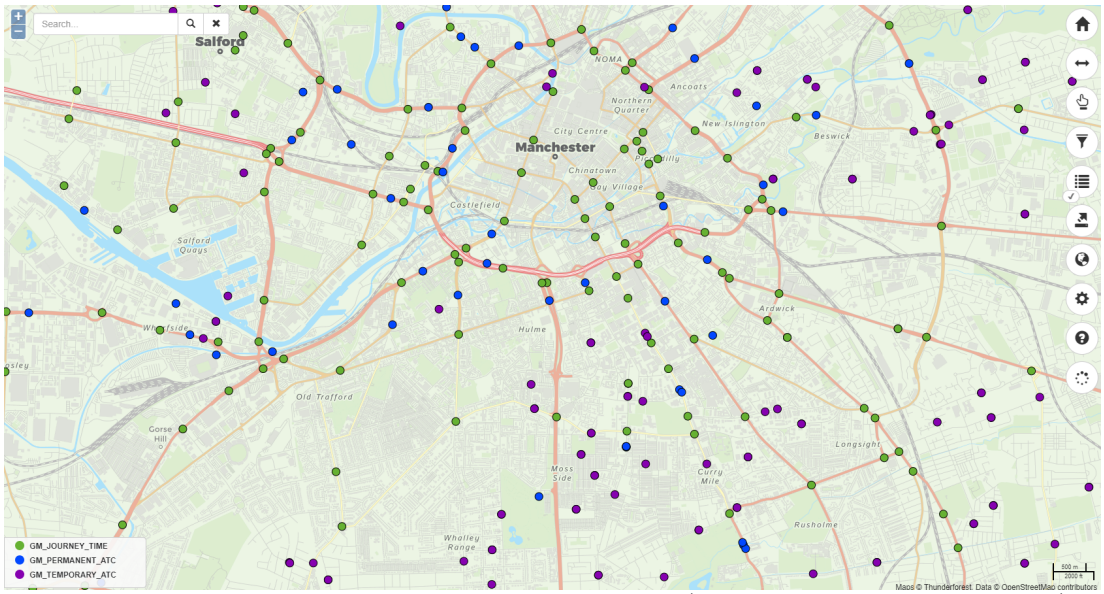


Figure 1.7: A part of Manchester Bluetooth sites (Source:tfgmc2.drakewell.com).

Figure 1.8: A Bluetooth detector in Manchester (Photo taken by Dr. Katharina Hannaford).

The omni-directional antenna with 9dBi gain is used, which provides a range of approximately 100 m (Bhaskar and Chung, 2013). The developer of the detectors claims that Bluetooth detectors can cover up to 6 lanes of traffic travelling at 70 mph. The vehicles with discoverable Bluetooth devices (i.e. smartphones, headsets, navigators, etc.) should be captured by detectors. The settings of the sensor are edited on each device separately and the pick-up zone is changed depending on the location/junction the sensor is covering. This is to avoid picking up other Bluetooth devices from houses.

1.5 Challenges presented by Bluetooth data

Bluetooth data is fraught with a range of issues that have a significant impact on the accuracy of predicted traffic metrics such as journey time, speed, and origin-destination matrix (Araghi et al., 2015; Bhaskar and Chung, 2013; Cragg, 2013). Some of these sources of errors are as follows. As mentioned in Section 1, the multiple detections that happen when a MAC address is likely to be detected multiple times in a short period of time, particularly during periods of congestion, lead to having more than one valid record for an active Bluetooth device. Therefore, before Bluetooth data is used for travel time studies, the strategy for dealing with multiple detections should be specified. There is no general rule for selecting the best detection record amongst the multiple records of MAC addresses and some previous researchers have used the first-to-first and last-to-last records of the devices. In fact, the majority of studies have tried to filter multiple detections in terms of selecting one time stamp amongst the duplicated records of MAC addresses and using it for travel time or speed estimation. Bhaskar and Chung (2013) modelled the theoretical properties of Bluetooth data and analysed the accuracy and reliability of using Bluetooth for travel time estimation in the modelled section of the signalised urban environment based on three different models (exit-to-exit, stop-to-stop and entrance-to-entrance).

Another source of bias is that the trips could be omitted due to missed detections, which is also one of the aspects of Bluetooth data (Michau et al., 2014). Missed detections can happen as some scanners and devices have stronger signals than others, but not all of them are equally powerful. Porter et al. (2013) highlight the antenna's influence on signal strength and detection. Also, Bluetooth devices are not always discoverable. For example, after a few minutes of inactivity, some devices may become undetectable and some devices (such as iPhones) are only discoverable for a short period of time after the user imitates the discovery. For example, in the Manchester network iPhones were non-trackable, as these devices have a low probability of being detected by Bluetooth detectors (Bhaskar et al., 2015; Abbott-Jard et al., 2013). The other main reason is that, as the detection zone becomes more congested with active Bluetooth devices, the rate of missed detection rises. Interference may reduce detection effectiveness as the number of detectable devices increases (Franssens, 2010).

The location of the detectors is important, as physical obstacles (e.g. walls and billboards) decrease Bluetooth signals. For example, Brennan Jr et al. (2010) discovered that the vertical location of the Bluetooth scanner affects the sensor's efficiency. Colberg et al. (2014) also discovered a lane bias effect in Bluetooth data. This bias in Bluetooth units could be due to slower-moving vehicles staying in the Bluetooth detection zone longer than faster-moving vehicles, giving them a better chance of being detected by

Bluetooth detectors. Malinovskiy et al. (2010) discovered that, when compared to automatic licence plate recognition, travel times derived using Bluetooth data are usually overestimated. As a result, it was suggested that a faster moving vehicle has a high possibility of passing through the scanning zone undiscovered. This implies that if the data were collected across a faster-moving traffic zone, there would be a large loss of data in terms of MAC addresses that could have been recorded compared to slower-moving traffic. Therefore, implementing the Bluetooth detector near road intersections with traffic signals and pedestrian crossings, business areas, gas stations, and car parks, might result in inaccurate travel time and speed estimation.

Furthermore, because a recorded MAC address does not indicate the device's type and it can be carried in a vehicle, a bus, by a pedestrian, or a cyclist, distinguishing between modes of transportation may be difficult.

Comparing Bluetooth data with other data sets like loop detectors or ANPR data can be used to estimate the Bluetooth penetration rate as the percentage of vehicles with discoverable Bluetooth devices. For example, TfGM evaluated hourly Bluetooth penetration rates between 16% and 34% by comparing ANPR and Bluetooth data (Crawford et al., 2018). Nicolai and Kenn (2007) reported 2% and 6% as the percentage of people having discoverable devices in Bremen, Germany and San Francisco, US, respectively, where it has approximated 5% in Maryland State, US (Young, 2012). Araghi et al. (2015) conducted a controlled field experiment to check the reliability of travel time estimation using Bluetooth and GPS data, where the GPS formed the ground-truth used to calibrate the Bluetooth detection rate, and reported an estimate of 27% to 29%.

Despite the difficulties mentioned above, data collected from fixed Bluetooth detectors is still thought to have a lot of potential for road data analysis. In a review paper, Friesen and McLeod (2015) motivated the continuing development of non-invasively developed systems using existing communications infrastructure and consumer devices that include short-range communication technologies such as Bluetooth.

Chapter 2

Cluster analysis

2.1 Introduction

This chapter presents some exploratory data analysis techniques in order to investigate whether there might be useful information in the multiple detections of Bluetooth devices at a single site over a short period of time. It starts by developing some candidate variables in terms of multiple detections that have been considered in order to investigate the detection behaviour at a particular detection zone in Section 2.2, and time series plots of these variables is presented in subsection 2.2.2. Following that, hierarchical cluster analysis, a popular unsupervised learning method, is given in Section 2.3. Firstly, the clustering is performed using a subset of the candidate variables defined in Section 2.2 and two applications is presented in subsections (2.3.2–2.3.3). When multiple detections happen, there is a time difference between records, resulting in the gap distribution between detections. As a result of multiple detections, there are gaps between consecutive detections. Based on gap time distributions, the cluster analysis utilising the Kolmogorov-Smirnov statistic is represented in Section 2.4 and, as two applications of this method, the classification of MAC addresses and time intervals of a day are presented in subsections 2.4.1 and 2.4.2, respectively. The discussion for this chapter is provided in Section 2.5.

2.2 Exploratory data analysis

This chapter tries to link between the traffic conditions and detection behaviour in Bluetooth detector coverage zones based on analysing two properties of Bluetooth data; multiple detections and the gap time distribution of multiple detections.

The primary step for achieving this goal is considering some different variables that are related to multiple detections, or the consecutive gap times between the multiple detections of each MAC address. In order to look at how these variables change during

the time of day, the time interval of 15 minutes has been chosen as a reasonable length of time to get enough data. For the purpose of comparing weekdays and weekends, analysis is performed over one week. The variables considered are as follows:

1. **The number of all recorded MAC addresses detected:** It shows how many MAC addresses have been recorded every 15 minutes by each detector, i.e. the number of all recorded detections, including multiple detections, during every 15 minute time interval. This could be regarded as a traffic level indicator for the detector's area. It should be noted that the number of recorded detections will depend on the characteristics of the location of detectors. For example, having a rest area, gas station, toll plaza, or signalised lights will affect the area's congestion and vehicles' behaviour.

2. **The number of all unique MAC addresses detected:** It represents the exact number of detected MAC addresses in a 15-minute period without considering their multiple recorded detections. This shows how many Bluetooth devices come across the Bluetooth detector coverage zone.

3. **The number of MAC addresses with multiple detections:** Due to the wide detection zone of the detector, while a single MAC address goes through the zone, it is likely to be detected more than once. Also, traffic congestion causes the car to stay longer in the coverage zone of the detector, thus the number of MAC addresses with multiple detections would be expected to increase when there is congestion.

4. **The number of MAC addresses with only one detection:** In the ideal state, when the MAC address is detected once, it can be assumed that the vehicle with this MAC address is passing fast enough through the zone of the detector. However, it is wise to think about the possibility that it might be a missed detection, especially during peak times of day, when the chance of a missed detection will be high due to an increasing number of detectable Bluetooth devices in the detector area.

5. **The proportion of MAC addresses with multiple detections:** This variable is considered to check what the proportion of MAC addresses with multiple detections every 15-minutes. As traffic volume increases, it may be expected that this proportion will also increase.

6. **The average number of detections for the MAC addresses with multiple detections:** This variable is considered to test what is the average number of detections for the MAC addresses with multiple detections every 15-minutes.

Again, like the proportion of MAC addresses with multiple detections, it may be expected that this average increases when traffic volume increases.

For the case study area, we have considered a set of Bluetooth detectors and the time series plots of these variables will be presented.

2.2.1 Case study

For this research, we utilized thirteen Bluetooth detectors in a closed loop to create a simple and manageable network for analysis. The considered Bluetooth site network is depicted in Figure 2.1, and their descriptions are given in Table 2.1. The Bluetooth detector, permanent and temporary ATCs are displayed by the green, blue, and purple circles, respectively.

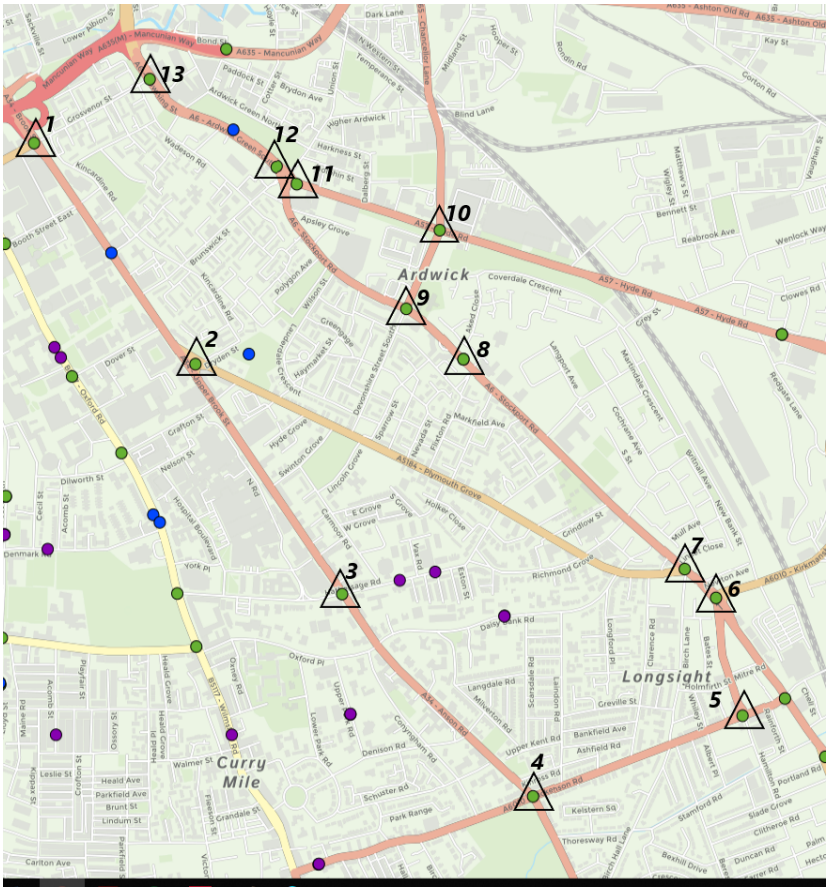


Figure 2.1: A network of Bluetooth sites in Greater Manchester. The Bluetooth detector, permanent and temporary ATCs are displayed by the green, blue, and purple circles, respectively.

Site	Site Name	Description
Site 1	MAC4109MR	Upper Brook Street (A34) / Grosvenor St (18), Manchester (Blu)
Site 2	MAC4108MR	Upper Brook St (A34) / Plymouth Gr (134), Brunswick, Manchester (Blu)
Site 3	MAC1152MR	Upper Brook St (A34) / Hathersage Rd (133), Longsight, Manchester (Blu)
Site 4	MAC4067MR	Birchfields Rd (A6010) / Dickenson Rd (41), Rusholme, Manchester (Blu)
Site 5	MAC4066MR	Dickenson Rd (A6010) / St John's Rd (1049), Longsight, Manchester (Blu)
Site 6	MAC1076MR	Stockport Rd (A6) / Kirkmanshulme Ln (1206), Longsight, Manchester (Blu)
Site 7	MAC1079MR	Stockport Rd (A6) / Plymouth Gr (1205), Longsight, Manchester (Blu)
Site 8	MAC1318	Stockport Rd (A6) / 220m SE of Devonshire St, Longsight, Manchester (Blu)
Site 9	MAC1075MR	Stockport Rd (A6) / Devonshire St (1219), Ardwick, Manchester (Blu)
Site 10	MAC4038MR	Hyde Rd (A57) / Devonshire St (49), Ardwick, Manchester (Blu)
Site 11	MAC4039MR	Hyde Rd (A57) / 200m W of Dalberg St (3/121), Ardwick, Manchester (Blu)
Site 12	MAC1081MR	Ardwick Green S (A6) / 10m N of Brunswick St (3/119), Ardwick, Manchester (Blu)
Site 13	MAC1078MR	Downing St (A6) / Grosvenor St (165), Ardwick, Manchester (Blu)

Table 2.1: Sites description.

Table 2.2 shows the multiple detections for a sample MAC address with its consecutive gap times. This device was detected six times at Site 12, the first at 09:03:13 and the last at 09:03:30 on February 15th 2019.

Date-Time	MAC address	Gap time(sec)
2019-02-15 09:03:13	05140F003F84	-
2019-02-15 09:03:14	05140F003F84	1
2019-02-15 09:03:17	05140F003F84	3
2019-02-15 09:03:18	05140F003F84	1
2019-02-15 09:03:21	05140F003F84	3
2019-02-15 09:03:30	05140F003F84	9

Table 2.2: The multiple detections for a sample MAC address with its consecutive gap times on February 15th 2019 at Site 12.

2.2.2 Time series plots

The traffic conditions vary considerably depending on the time of day. Therefore, it is expected to see an hourly pattern that indicates how the traffic flow varies during the day and night. Normally, an hourly pattern indicating how traffic flow varies over the day and night is expected (Minnen et al., 2015) and normal hourly traffic flow patterns show a variety of distinct peaks, particularly in urban areas. One peak in the morning is often more sharp, reaching its peak over a short duration and quickly dropping to its lowest point. The afternoon peak is characterised by a wider peak and is reached and dispersed over a longer period than the morning peak. The reason for the different dispersion is that people usually start going to work in the morning at around the same time but go back home at different times in the afternoon (Minnen et al., 2015). Except for the morning and evening rush hours, Manchester is known to

have a free flow of traffic. The peak hours in Manchester are usually (7:00-9:30) and (16:00-18:30) on weekdays, with a slight peak in traffic during the lunch break, around (12:30-14:00).

Figure 2.2 shows time series plots of the first four defined variables at 15-minute intervals over a weekday on Site 12 on Monday, February 11th, 2019. It should be noted that Site 12 is located near a roundabout and there is a traffic light and a gas station in its vicinity, and it has a 48 km/h speed limit.

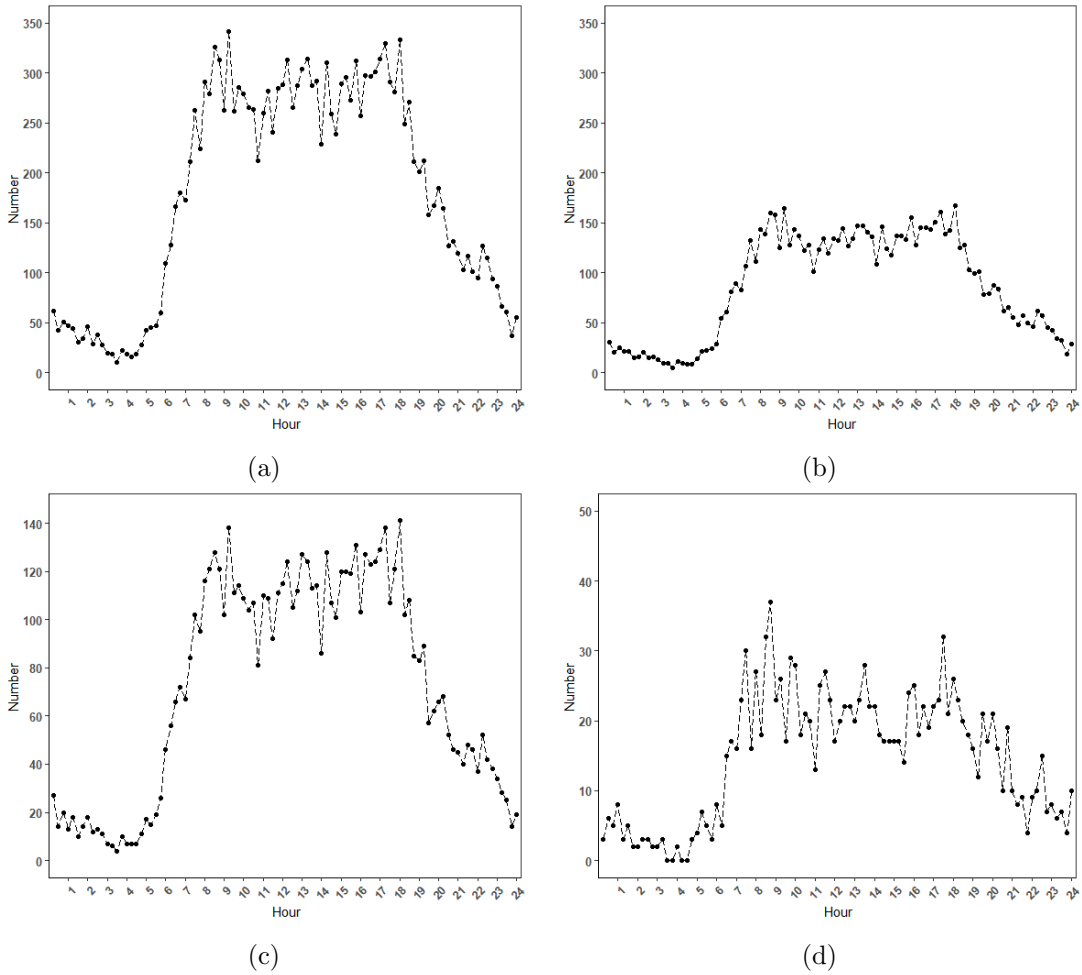
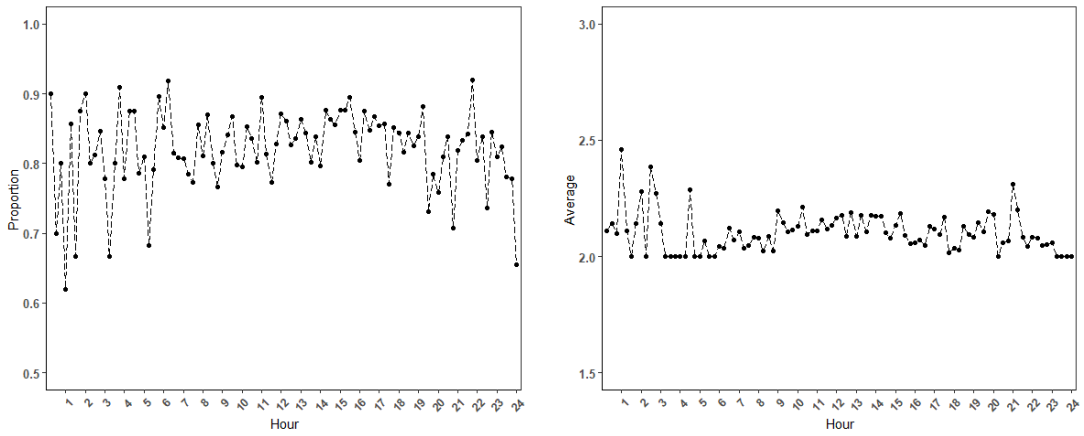


Figure 2.2: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, and (d) the number of MAC addresses with only one detection on Site 12, Monday 11th February, 2019.

The time series plots of the first four variables show a clear hourly pattern during the day. For example, the time series corresponding to the number of all recorded detections of MAC addresses represents that this variable started to increase from a low number (60 records) at the time interval (5:30-5:45) to the highest number (342

records) at the time interval (9:00-9:15). After that, it shows a slight drop and continues consistently around 250 records, but also has some peaks, for instance, 313 records between (12:00-12:15) due to the lunch break time. The afternoon peak begins at the time interval (17:00-17:15) and falls again because of the end of rush hour.

Figure 2.3 shows time series plots of the last two defined variables at 15-minute intervals over a weekday on Site 12, Monday 11th February 2019.



(a) (b)

Figure 2.3: Hourly patterns for the considered variables: (a) the proportion of MAC addresses with multiple detections, and (b) the average number of detections for the MAC addresses with multiple detections on Site 12, Monday 11th February, 2019.

It was expected that the proportion of MAC addresses with multiple detections and the average number of detections for MAC addresses with multiple detections might change with traffic levels on the street. As noted in Section 2.2, it may be expected that the proportion and the average of MAC addresses with multiple detections are likely to increase as traffic increases throughout the day. However, the rate of missed detection might also be expected to increase as the detection zone becomes more congested with active Bluetooth devices due to detection interference. As can be seen, no special trend is observed in these time series plots over different fifteen minute time intervals of a day. The time series plot of the proportion of MAC addresses with multiple detections shows a fairly constant rate of 0.8 and the time series plot of the average of multiple detections was roughly consistent at around 2.

Figure 2.5 shows time series plots of the first four defined variables at 15-minute intervals over a weekend day on Site 12, Sunday 17th February, 2019.

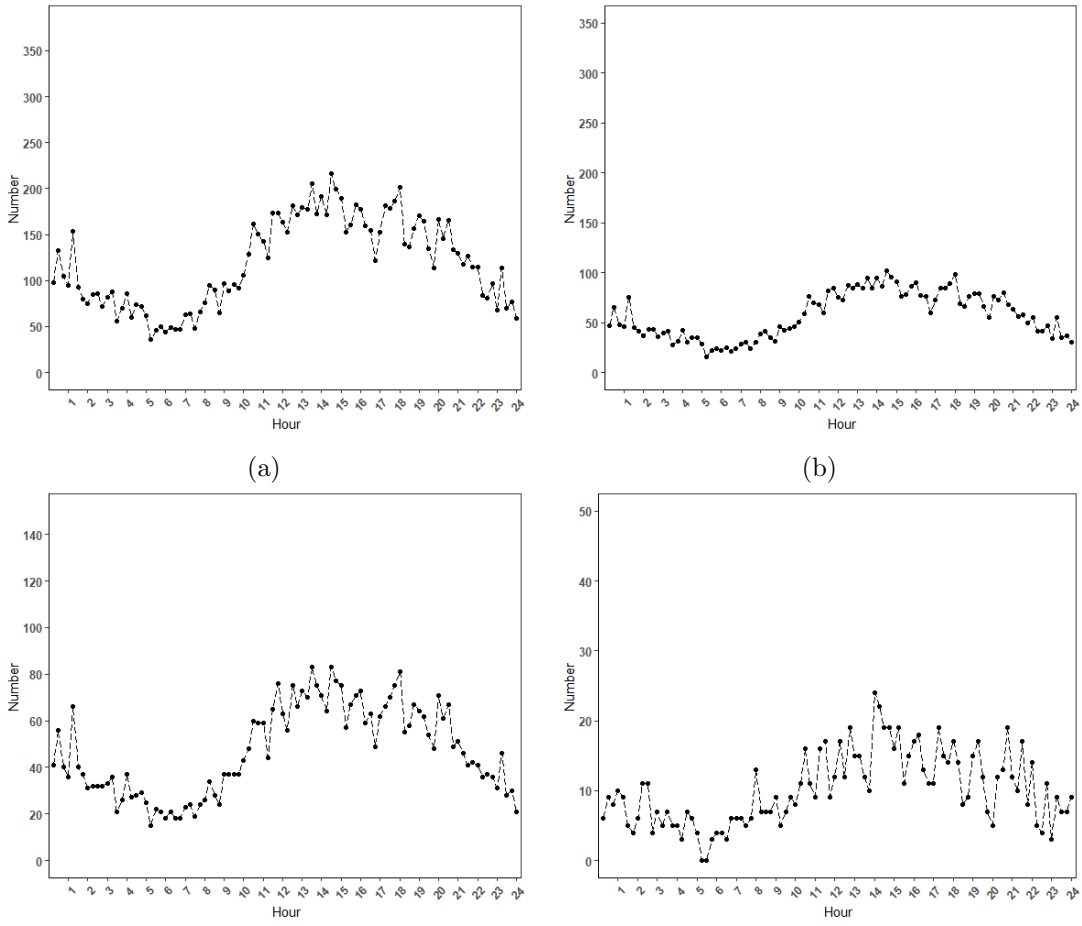
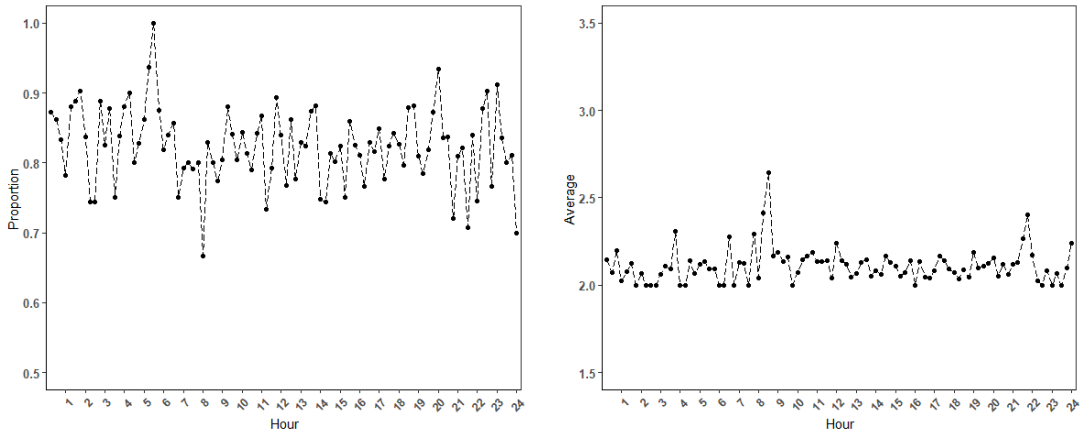


Figure 2.4: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, and (d) the number of MAC addresses with only one detection at Site 12, Sunday 17 February, 2019.

As expected, the traffic conditions differ during the week in terms of working days versus weekends, which have different daily patterns. The traffic during the working days (Monday to Friday) may not vary considerably from day to day, but the traffic volume during the weekend is likely to vary from the working days. The pattern from Monday to Friday is often relatively consistent, apart from Monday morning and Friday afternoon traffic flow. The pattern during the weekend may differ from Saturdays to Sundays. The time series plots of the first four variables reveal a clear hourly pattern during the day that differs from the pattern on weekdays. It emphasises that, according to Bluetooth data, traffic volumes at this site are higher on weekdays than on weekends.

The time series corresponding to the number of all recorded MAC address detections illustrates the difference in behaviour between Monday and Sunday midnight, which

shows 154 records at the time interval (1:00-1:15) on Sunday midnight. This variable starts to increase from 106 records at (9:45-10:00), then remains consistent until around 6 p.m., when it begins to decrease. Figure 2.5 shows time series plots of the last two defined variables at 15-minute intervals over the weekend on Site 12, Sunday 17 February 2019.



(a) (b)

Figure 2.5: Hourly patterns for the considered variables: (a) the proportion of MAC addresses with multiple detections, and (b) the average number of detections for the MAC addresses with multiple detections on Site 12, Sunday 17 February, 2019.

Similarly to the weekday plots, no special trend is observed in the time series plots of the proportion and the average number of detections for the MAC addresses with multiple detections. The time series plot of the proportion shows a fairly constant rate of 0.8 and the time series plot of the average of multiple detections was roughly consistent at around 2 over different fifteen minute time intervals of a day. The interesting result is that these two variables are pretty constant not only during the different times of the day, but also across the two different days (i.e. working and weekend days).

The Appendix A.1 presents the time series plots for the other days of the week for Site 12 (see Figures A.1–A.5) and the time series plots for some of the other sites on one day, Monday 11th February, 2019 (see Figures A.6–A.14). Comparing the time series plots of different sites shows that they all have similar behaviour in terms of the first four variables, but with differences in the calculated numbers for these variables. For example, Site 13 had the highest number of all recorded MAC addresses (over 3000) and the average number of multiple detections (around 10), requiring more investigation into the difference. The proportion of MAC addresses with multiple detections and the average number of detections for MAC addresses with multiple detections have been shown to be roughly similar between the other different sites.

2.3 Cluster analysis

This section discusses the exploratory data analysis of the unfiltered Bluetooth detection data using cluster analysis, one of the most popular unsupervised learning methods. Cluster analysis partitions a set of data points or objects into separate groups called clusters. One of most widely studied clustering algorithms is *hierarchical clustering* that has been used in a wide range of applications (Rokach and Maimon, 2005; Nielsen, 2016).

Hierarchical algorithms produce a dendrogram or tree graph, which depicts the hierarchical grouping structure. The vertical axis of the dendrogram depicts the distance or dissimilarity between clusters, whereas the horizontal axis represents objects or clusters. The number of clusters included in the data is a common goal, and the hierarchical nature of a dendrogram should make it explicit that the number of clusters relies on the particular level of dissimilarity. Choosing a higher or lower level of dissimilarity results in a few large clusters or a large number of little clusters, respectively. The agglomerative or bottom-up clustering and divisive or top-down clustering are two methods for hierarchical clustering. The agglomerative method starts by considering singleton clusters, which means only one data object per cluster at the bottom level, and continues by joining two clusters to create a bottom-up hierarchy of clusters. The divisive approach, on the other hand, starts with all the data objects in one large cluster and cuts them into two groups in a top-down hierarchy of clusters.

In hierarchical clustering, the hierarchy of similar objects is formed based on a pairwise distance matrix. The distance matrix is symmetric (because the distance between object A and object B is the same as the distance between object B and object A) and has zero on the diagonal (because every object is distance zero from itself). The agglomerative hierarchical clustering starts with every data object in a separate cluster, then continues by joining the closest sets of clusters, with the distance matrix updating on each step. This process of agglomerative merging is continued until the final cluster (that includes all the data objects in a single cluster) is achieved.

Hierarchical clustering requires the definition of a distance between clusters, in addition to the distance metric between individual points. The distance between two clusters is called the linkage. The most popular proximity measures which can be used in agglomerative hierarchical clustering are as follows:

- **Single Linkage:** The distance between two clusters is defined as the smallest distance between a pair of the data points within the clusters,
- **Complete Linkage:** The distance between two clusters is defined as the largest distance between a pair of the data points within the clusters,

- **Average Linkage:** The distance between two clusters is defined by the average distance between a pair of the data points within the clusters,
- **Ward's Linkage:** The aim of Ward's method is to minimize the variance within each cluster. This is achieved by defining distance as the difference in cluster variance by fusing the clusters.

There are two different approaches that can be used as a starting point in clustering: (i) using observed variables to define distances between objects or (ii) pairwise distances of data generated directly from other methods. Both approaches have been investigated in order to perform the cluster analysis.

The exploratory data analysis of Bluetooth data will be performed based on observed variables such as the multiple detections and the gap times distribution (i.e. the time difference between two consecutive Bluetooth detections is known as the gap). For the cluster analysis, we will utilize Bluetooth sites, MAC addresses, and the hourly time intervals as the objects.

We briefly describe the clustering using observed variables in Section 2.3.1, with two applications of this approach present in Sections 2.3.2 and 2.3.3. Section 2.4 describes the second approach, with two applications present in Sections 2.4.1 and 2.4.2.

2.3.1 Clustering based on observed variables

Suppose $\boldsymbol{X_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $\boldsymbol{X_j} = (x_{j1}, x_{j2}, \ldots, x_{jp})$ represent two p-dimensional vectors of variables from n vectors in a dataset. The distances between every pair of variables create an $n \times n$ distance matrix, named D, as follows:

$$D = (d_{ij}) \in R \tag{2.1}$$

where d_{ij} denotes the distance between two vectors $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ and can be computed by some commonly used metric such as the Euclidean distance, the squared Euclidean distance, the Manhattan distance, the maximum distance or the Mahalanobis distance. For example, the Euclidean distance between an observation X_i and another X_j is:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}. \tag{2.2}$$

After creating the distance matrix based on observed variables, the agglomerative hierarchical clustering continues as discussed above. The two following sections present two applications using clustering based on variables.

2.3.2 Bluetooth sites clustering

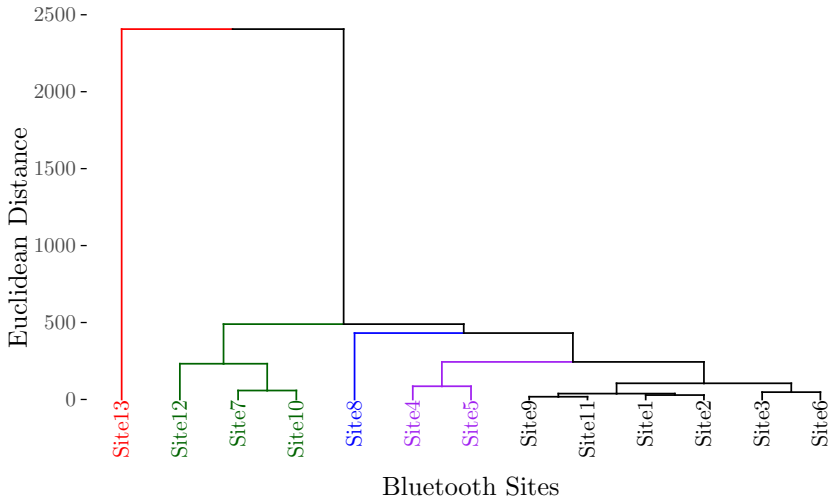
The variability of traffic volumes within each road could depend on its location. For example, a local road located in a high population urban area normally has higher traffic volumes compared to a local road in a lower population area. The Bluetooth sites have different characteristics in terms of their placements, such as proximity to the school area, shopping malls, the gas stations and the traffic lights, etc.(Purser, 2016). Also the orientation and other characteristics of the site might affect the rates of multiple and missed detections.

Clustering the Bluetooth detector sites will be useful for identifying similar sites and possibly detecting unusual sites based on the variables considered. The detection of unusual sites or outliers can be useful in reconsidering where the Bluetooth detector should be placed, as the unusual behaviour could indicate the detector is not in the proper place or that it is malfunctioning. Furthermore, depending on the clustering results, any further study performed on a single Bluetooth site can be extended to all similar sites.

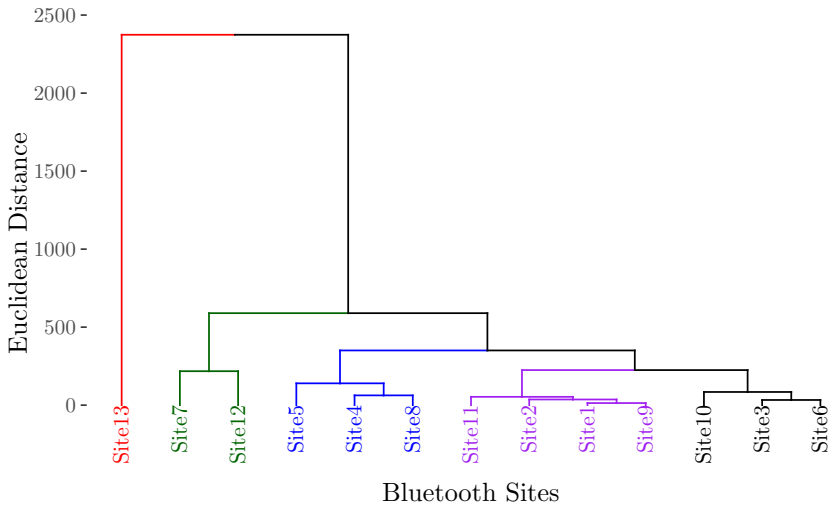
We decided to investigate grouping the Bluetooth detector sites in terms of their patterns of detections to compare the volume of traffic in their coverage zones. This clustering was done as one of the applications of variable-based clustering. When traffic volume increases, the number of all recorded Bluetooth detections increases due to more multiple detections, and the number of unique Bluetooth MAC addresses increases if the Bluetooth site is located in a high population urban area, as shown by time series plots (Figures 2.2 and 2.4). Hence, for clustering the Bluetooth sites in the network, we used these two variables, the number of all recorded Bluetooth detections and the number of unique Bluetooth MAC addresses. Also, it should be noted that the clustering results did not differ by adding the other observed variables, therefore, just these two variables have considered.

The Bluetooth data related to all 13 Bluetooth sites is considered to perform hierarchical clustering on one day (Monday 11th of February, 2019). In a one hour time interval, the number of all recorded data and the number of all unique captured MAC addresses were considered as variables for each Bluetooth site. The Euclidean distance was used to make the similarity matrix containing the pairwise distance for all Bluetooth sites. The dendrogram of Bluetooth site clustering using the average linkage critria and Euclidean distance for two consecutive busy hours in the morning and afternoon is shown in Figures 2.6 and 2.7, respectively. It should be noted that standardization is often used as a preprocessing procedure in cluster analysis. It is important if each variable of data has a different unit or if the scales of each of the variables are very distinct. Standardization prevents variables with greater scales from influencing how clusters are constructed. It enables the algorithm to take into account all variables

equally. We used both un-standardized and standardized data to do the cluster analysis. Because the clustering results were similar, we have presented the un-standardized data to make cluster interpretation easier.

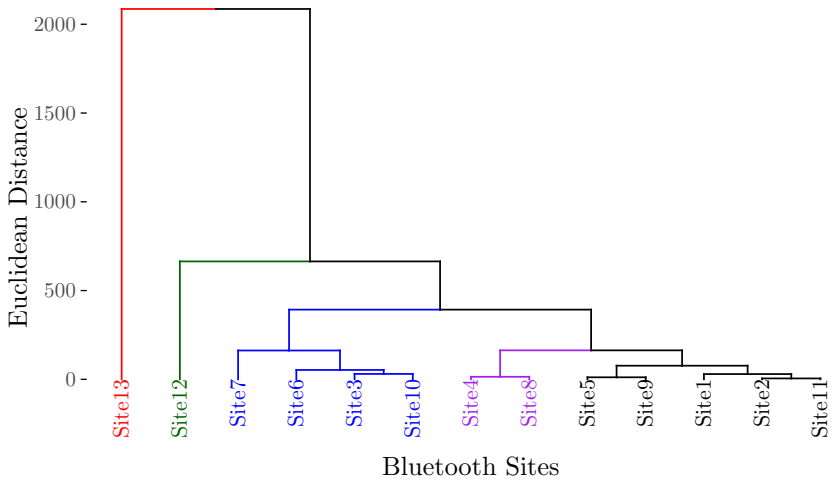


(a)

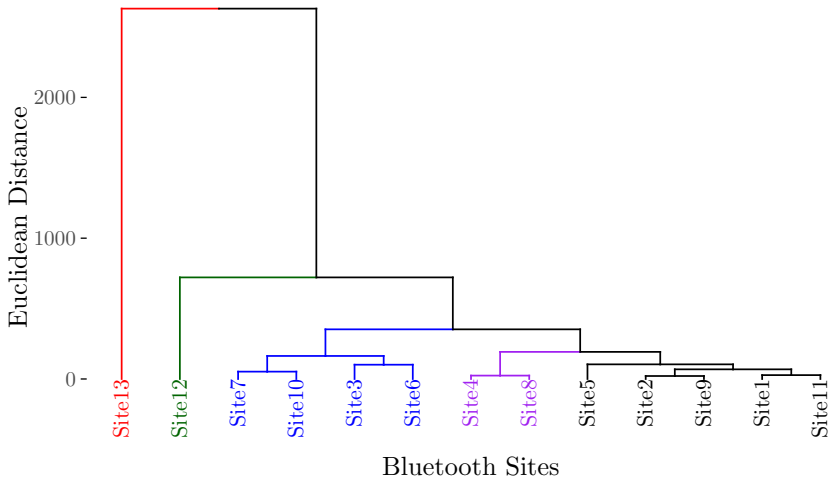


(b)

Figure 2.6: The dendrogram of Bluetooth site clustering based on two variables: i) the total number of Bluetooth detections; and ii) the total number of unique Bluetooth MAC addresses using the average linkage and Euclidean distance between (a) 7:00-8:00 a.m. and (b) 8:00-9:00 a.m., Monday 11th February 2019. The five clusters are represented by different colors.



(a)



(b)

Figure 2.7: The dendrogram of Bluetooth site clustering based on two variables: i) the total number of Bluetooth detections; and ii) the total number of unique Bluetooth MAC addresses using the average linkage and Euclidean distance between (a) 3:00-4:00 p.m. and (b) 4:00-5:00 p.m., Monday 11th February 2019. The five clusters are represented by different colors.

Tables 2.3 and 2.4 present information about the two considered variables for each Bluetooth site within the five sub-clusters, with sites of the same colour clustered together. The number of clusters has been chosen subjectively and based on examining the characteristics of the clusters. Also Figure A.15 in Appendix A shows the dendrogram of Bluetooth site clustering for the time intervals of 9:00-10:00 a.m. and 5:00-6:00 p.m. The obvious result is that Site 13 behaved differently in all four time periods, and

this behaviour has also been observed at other hours of the day. In addition, there was not much of a difference in the Bluetooth site clusters formed between consecutive hours or between morning and afternoon peak hour traffic. For example, clustering using raw data indicates that Sites 1, 2, 9, and 11, which are close to each other in the network, were mereged in all time intervals considered, and the other clustered sites might also have similar characteristics that allow them to be clustered together despite being further apart. The observed variables do not change significantly for

Site	Number of all recorded	Number of unique MAC addresses
Site 13	2914	332
Site 7	897	439
Site 12	1091	537
Site 5	324	173
Site 4	221	137
Site 8	170	101
Site 11	432	232
Site 2	479	285
Site 1	479	251
Site 9	466	249
Site 10	707	391
Site 3	624	345
Site 6	635	376

Table 2.3: Details of two considered variables for time period 8:00-9:00 a.m. Sites with the same colours are clustered together.

Site	Number of all recorded	Number of unique MAC addresses
Site 13	3156	348
Site 12	1133	554
Site 7	754	367
Site 10	743	418
Site 3	646	351
Site 6	552	313
Site 4	223	145
Site 8	247	147
Site 5	338	185
Site 2	393	234
Site 9	400	214
Site 1	454	236
Site 11	463	262

Table 2.4: Details of two considered variables for time period 4:00-5:00 p.m. Sites with the same colours are clustered together.

each Bluetooth site over the sample time interval. In Tables 2.3 and 2.4, the details of the two variables indicate that, for example, Sites 4 and 8 are very similar, as are Sites 1, 2, 9, and 11 at both time intervals. The latter four sites are also located near each

other in the case study network. It shows that the number of all recorded sites, and also the number of unique MAC addresses for these sites, follow a consistent pattern at different times of the day. Also, Site 13 differs in terms of the first variable, since there is not much of a difference in the second variable, this site has proportionately more multiple detections. It is near a small commercial site and is placed at T-junctions with traffic signals, as shown in Figure 2.8.

Figure 2.8: Location of Site 13 (Source: Google Maps)

2.3.3 MAC addresses clustering

To take a closer look at Site 13, we have decided to perform another cluster analysis on the detected MAC addresses for a one hour time interval on a single day (Monday 11th of February, 2019). Therefore, the objects are now the detected MAC addresses and two variables have been considered as follows: i) the number of all detections per MAC address which shows how many times the MAC address has been detected; and ii) the mean of the gap times for each MAC address. If the MAC address is detected multiple times, there are gaps between consecutive recorded detections; if it is detected only once, the mean of gap times is zero. To create the similarity matrix containing the pairwise distance for all, the Euclidean distance was applied. The hierarchical clustering using the average linkage method for clustering the MAC addresses in Site 13 during the time period 3:00-4:00 a.m. is shown in Figure 2.9. For better graphical visualisation, the non-busy time interval of the day is employed, and also the truncated form of MAC addresses.

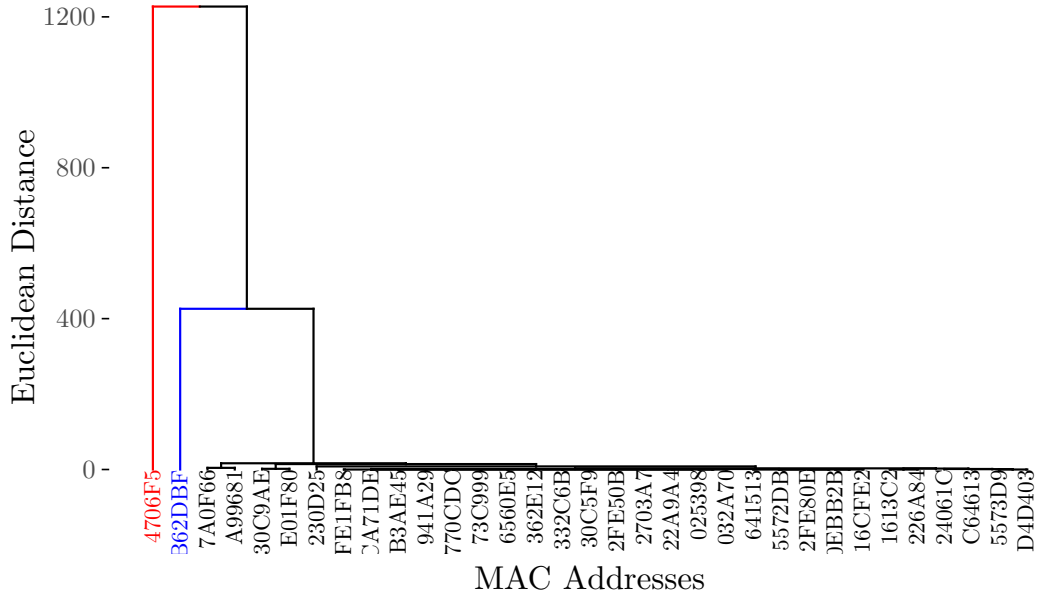


Figure 2.9: The dendrogram of MAC addresses clustering at Site 13 based on two variables: i) the number of all detections per MAC addresses; and ii) the mean of the gap times for each MAC address. It creates using the average linkage and Euclidean distance and displays for time period 3:00-4:00 a.m., Monday 11th February 2019.

Figure 2.9 depicts the classification of MAC addresses at Site 13 within a specific time interval. It can be noticed that one MAC address (4706F5) is an outlier when compared to the other MAC addresses. We also have repeated the clustering for the other time intervals of the day. This MAC address still has different behaviour in terms of the first variable, which is the number of all recorded detections. This MAC address clearly does not belong to an active Bluetooth device passing through the detector zone because it transmits constantly throughout the day. Table 2.5 shows the details of considered variables for the two MAC addresses that have been classified in the two separate clusters. The remaining 31 MAC addresses included 15 MAC addresses that

MAC Address	Number of all detections	Mean gap times (seconds)
4706F5	1228	2.93
B62DBF	2	428

Table 2.5: Details of two considered variables for time period 3:00-4:00 a.m.

were detected only once and 16 MAC addresses that were detected multiple times, with mean gap times ranging from 1 second to 16 seconds.

The MAC address clustering can be helpful for identifying MAC addresses that have an unusually high number of multiple detections, which can be considered as outliers. It may also be useful to classify the behaviour of the MAC address, possibly as a means to

distinguish different types of vehicles. The first MAC address, for example, is an outlier, according to Table 2.5. The second MAC address, which is detected twice with a gap of 428 seconds (about seven minutes), can be interpreted as a device that has gone away and returned. The rest looks to be a sample of MAC addresses that travelled through the detector area during the particular time interval, some of which were detected just once and others which were recorded multiple times, but all of which had similar mean gap times, causing them to cluster together.

The possibility that a device would leave the detection area and then return suggests that MAC address behaviour classification should include more variables. Therefore, gap distributions between multiple detections are suggested as an alternative approach for classifying MAC address behaviours. The following section develops a clustering technique for taking this approach.

2.4 Clustering based on Kolmogorov-Smirnov statistic

Discrete elements have been the most common object of clustering (Celeux and Govaert, 1991; Agrawal et al., 1998; Jain et al., 1999; Bouguila and ElGuebaly, 2009). However, with the complexity of data nowadays, it may be more appropriate to show the data as a series of numbers or functions rather as a single point. Therefore, cluster analysis does not always start with a set of variables that have been observed. As an alternative to discrete elements, probability density functions (pdfs) are considered as items for clustering (Chen and Hung, 2015; Tai et al., 2016; Nguyentrang and Vovan, 2017). All that is required is a way of defining a distance between the objects. Now the objects are distributions, consisting of sets of observations of different sizes. Therefore, the initial step is to construct a similarity matrix based on pair-wise distribution comparisons.

More-López and Mora (2015) suggested an adaptive algorithm for K-means clustering of the cumulative probability distribution functions of a continuous random variable. In the algorithm, they used the Kolmogorov–Smirnov two-sample statistic as a distance function.

The Kolmogorov-Smirnov test (KS statistic) is a non-parametric method for determining if two distributions are different (the two sample KS statistic) or whether an underlying probability distribution differs from a hypothesised distribution (the one sample KS statistic) (Berger and Zhou, 2014). Assume two independent random samples, one of size m with an observed cumulative distribution function of $F(x)$ and the other of size n with an observed cumulative distribution function of $G(x)$. The highest vertical deviation between the two cumulative distribution functions is used as the statistic D for the two sample KS statistic.

$$D = \max_x |F(x) - G(x)| \tag{2.3}$$

Because it satisfies three conditions for defining a metric: identity of indiscernibles, symmetry, and triangle inequality, it can be considered a distance. Figure 2.10 shows an example of the two-sample KS statistic, where the dashed red line is the two-sample KS statistic (i.e. maximum distance D), and the blue and black lines represent the empirical distribution function for two samples.

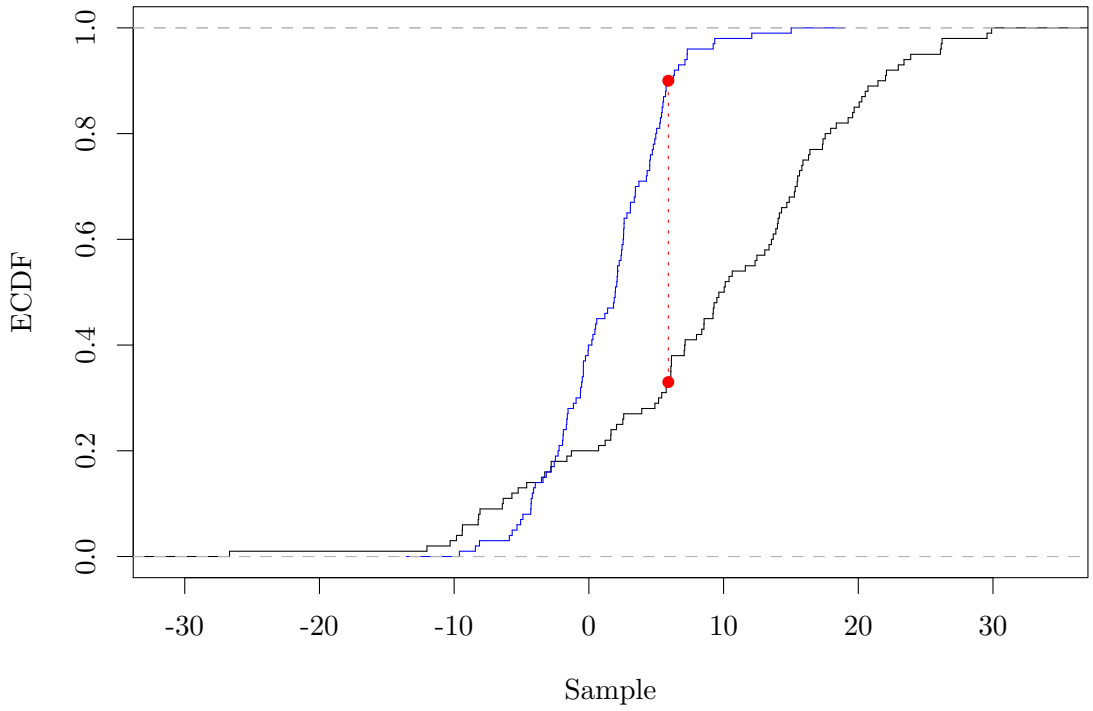


Figure 2.10: An example of the two-sample KS statistic. The dashed red line is the two-sample KS statistic (i.e. maximum distance D), and the blue and black lines represent the empirical distribution function for two samples.

We also decided to use the Kolmogorov-Smirnov statistic for assessing the pair-wise similarity between the gap distributions in the hierarchical clustering. In sub-sections 2.4.1 and 2.4.2, two applications of this approach are shown.

2.4.1 MAC addresses clustering based on gap distribution

This section presents the clustering of the MAC address data in terms of their gap times distributions as an application of the cluster analysis using the KS statistic. The goal is to identify groups of MAC addresses that have similar detection patterns during the tested time interval. We created a set of descriptive labels for the different types

of detected patterns in order to cluster the MAC addresses based on their detection patterns. It is worth mentioning again that when the MAC address has multiple detections in the coverage zone of the Bluetooth detector, it creates a gap distribution with varying lengths of gaps in-between detections for each MAC address.

In order to be able to characterize the nature of the clusters, we first define a *group* as a collection of detections of the same MAC address with a gap time of less than or equal to a fixed threshold between them. The time threshold for assigning a group in this study was set at 10 seconds, based on the assumption that gap times of greater than 10 seconds can be considered a missed detection. Because the enquiry phase of connecting Bluetooth devices is expected to take up to 10.24 seconds, it is highly likely that an active Bluetooth device will be detected within this period; otherwise, it may be missed (Peterson et al., 2006; Abbott-Jard et al., 2013).

Based on its gap pattern, each MAC address in a set of detected MAC addresses displays a different state. Each MAC address was given the following new labelling:

- Unique singleton: A unique singleton is a MAC address that is discovered only once throughout the time interval,
- N singleton group: The MAC address is labelled as N singleton if it is detected N times and the gap between each of the N detections is more than 10 seconds,
- $n_1, n_2, n_3, \ldots$ multiple group : If the MAC address is detected N times and the gap times between consecutive detections are varied, any number of recorded detections with gap times less than 10 seconds forms one group, according to group definition. As a result, it is referred to as the $n_1, n_2, n_3, \ldots$ multiple group, where $n_1, n_2, n_3, \ldots$ represent the number of members in each group, and $\sum_i n_i = N$.
- Unique Group: A unique group is formed when the MAC address is detected many times and all of the detections have a gap time of less than 10 seconds.

Figure 2.11 shows a visual illustration of these new labels. Table 2.6 presents a sample report of MAC address gap distribution for a one-hour time interval 3:00-4:00 a.m.

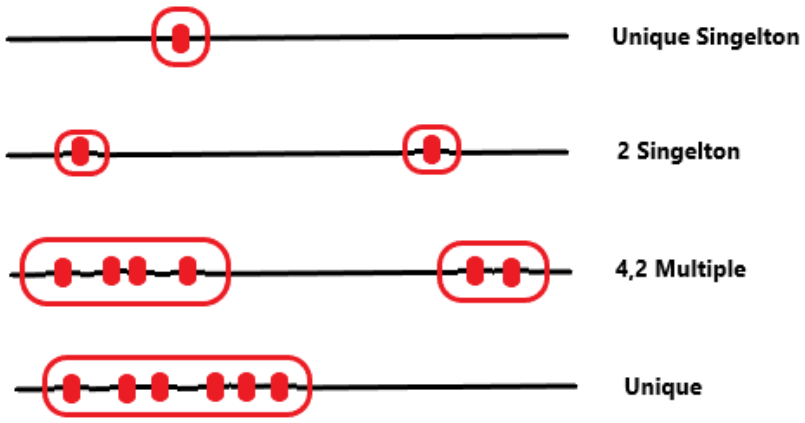


Figure 2.11: Illustration of new labelling for MAC addresses.

MAC address	Time	Status	Gap Distribution	Mean
2B74494DDB0A	3:00-4:00	Unique singelton	-	-
055600673460	3:00-4:00	2 singelton	19	19
30C1F983896F	3:00-4:00	2, 2, 2, multiple	4, 427, 4, 286, 9	146.0
380646000E05	3:00-4:00	Unique	5, 1, 4, 1	2.75

Table 2.6: A sample report of MAC address gap distribution for a one-hour time interval 3:00-4:00 a.m.

The MAC addresses recorded by Bluetooth detectors at one-hour intervals were taken into consideration, and new labelling was assigned depending on the gap times distribution. We used the KS statistic to create a distance matrix comprising the pairwise distance for all pairs of MAC addresses based on the difference of their cumulative distribution functions, and then we used Ward's linkage method to apply the hierarchical clustering approach to this matrix. Also, because there is no gap distribution for unique singleton MAC addresses (i.e. those that have been recoded just once), the gap for those was set to zero, i.e. gap distribution= $\{0\}$. Figure 2.12 displays the dendrogram of MAC addresses clustering based gap distribution using Ward's linkage and KS distance at Site 12 for time periods 3:00-4:00 a.m., Monday 11th February 2019. Note that the KS distance between individual distributions must be between 0 and 1, but the distance between groups defined by Ward's method can exceed 1. This time period was chosen as a non-busy period in order to have fewer MAC addresses that provide a better visualisation of clusters. We also selected Site 12 because it is one of the sites near a roundabout, and being close to a gas station and a park should lead vehicles to stay more within the coverage area. Looking at the details of sub-cluster objects shows that all six MAC addresses in the first cluster from the left are labelled as a unique singelton that has only been recorded once. The next three MAC addresses were discovered twice

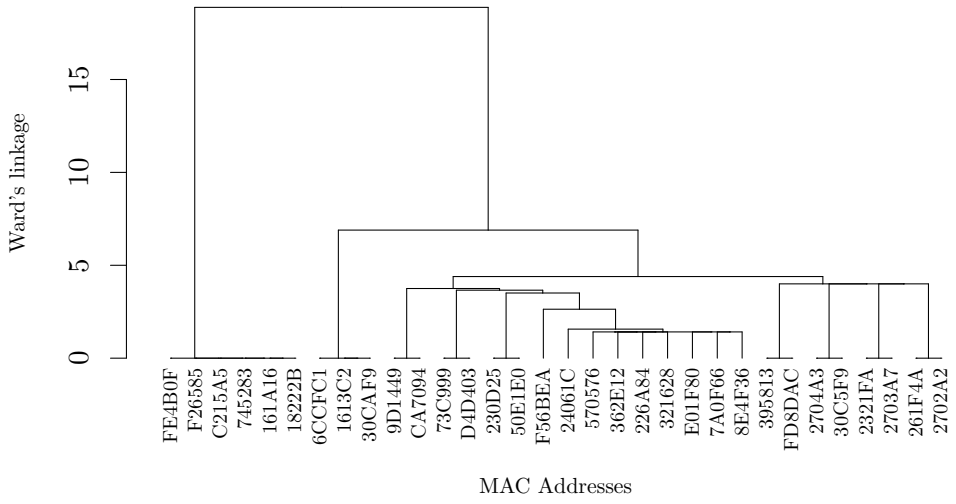


Figure 2.12: The dendrogram of MAC addresses clustering based gap distribution using Ward's linkage and KS distance at Site 12 for time periods 3:00-4:00 a.m., Monday 11th February 2019.

using the same gap distribution and have exactly the same mean gap. The remaining 23 MAC addresses were identified multiple times with different gap distributions and mean gaps.

2.4.2 Time interval clustering based on gap distribution

To give another application of cluster analysis based on KS-statistics, we have investigated using the gap distributions in order to cluster the time intervals of a day. The main aim is to explore if the pattern of gap times between consecutive MAC address detections can be used to divide a day into interpretable clusters, in which time intervals with similar gap distribution patterns, which may indicate similar traffic conditions, merge together. The Bluetooth data related to Site 12 is selected to perform hierarchical clustering. The recorded MAC addresses during every one-hour time interval were considered and corresponding gap times computed. The KS statistic was used to compute the distance matrix between the 24 distributions (24 hours of a day). After computing the distance matrix based on KS statistic, hierarchical clustering was utilized using Ward's linkage method. Figure 2.13 represents the hierarchical clustering dendrogram for Site 12, Monday 11th February, 2019.

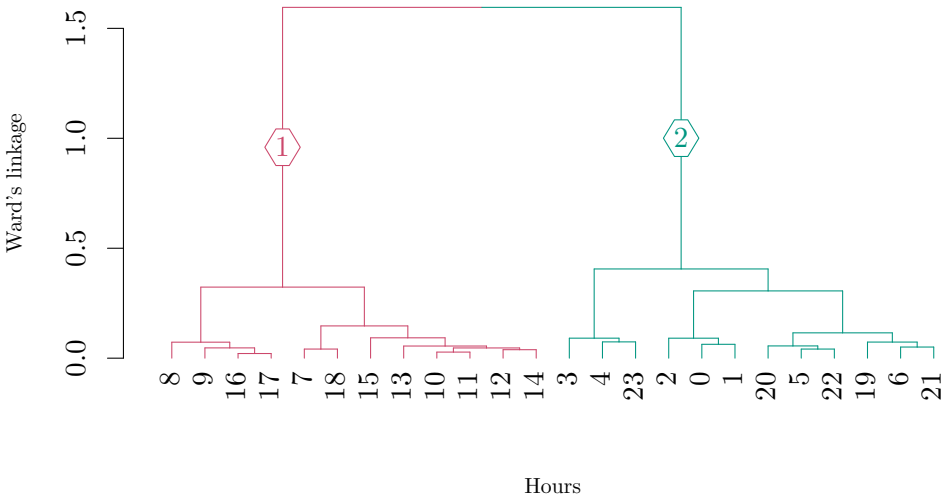


Figure 2.13: The dendrogram of time interval clustering based gap distribution using Ward's linkage and KS distance at Site 12, Monday 11th February 2019.

Looking at the dendogram based on the captured gap times distribution, there are two distinct clusters: the right hand cluster (i.e. cluster 2) seems to include the non-busy hours, while the left hand cluster (i.e. cluster 1) shows to include the most of the busy hours (i.e. cluster 1). As the volume of traffic varies, the amount of time each vehicle spends in the detection area changes, influencing the detection pattern. In congested traffic, we can expect more multiple detections and gap times, whereas in free flow traffic, we can expect fewer multiple detections and gap times. For example, the morning and evening peak periods (i.e. 8, 9 a.m. and 4, 5 p.m.) had a similar gap times distribution and were clustered together. It shows that they have all experienced similar traffic conditions, resulting in a comparable Bluetooth detection gap pattern.

Figures 2.14 and 2.15 represent a log transform of gap times for time intervals in cluster 1 and cluster 2, respectively. Looking at these plots, it can be seen that how the time intervals in the same cluster have a similar gap distribution. As seen in Figure 2.15, the first three time periods after midnight, which are merged together as a sub-cluster in cluster 1, had a similar gap distribution.

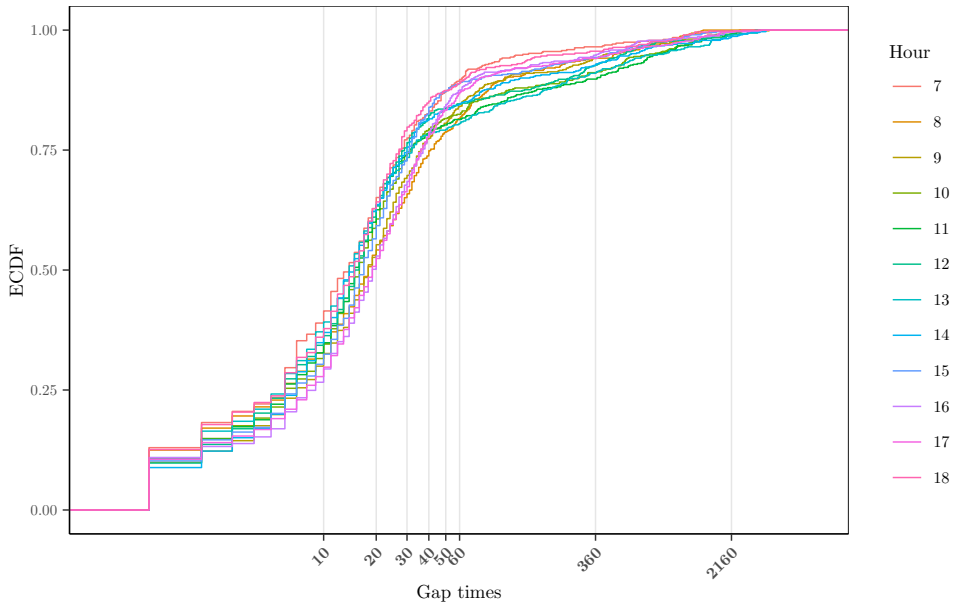


Figure 2.14: The log transform of gap times for time intervals in cluster 1 for Site 12, Monday 11th February, 2019.

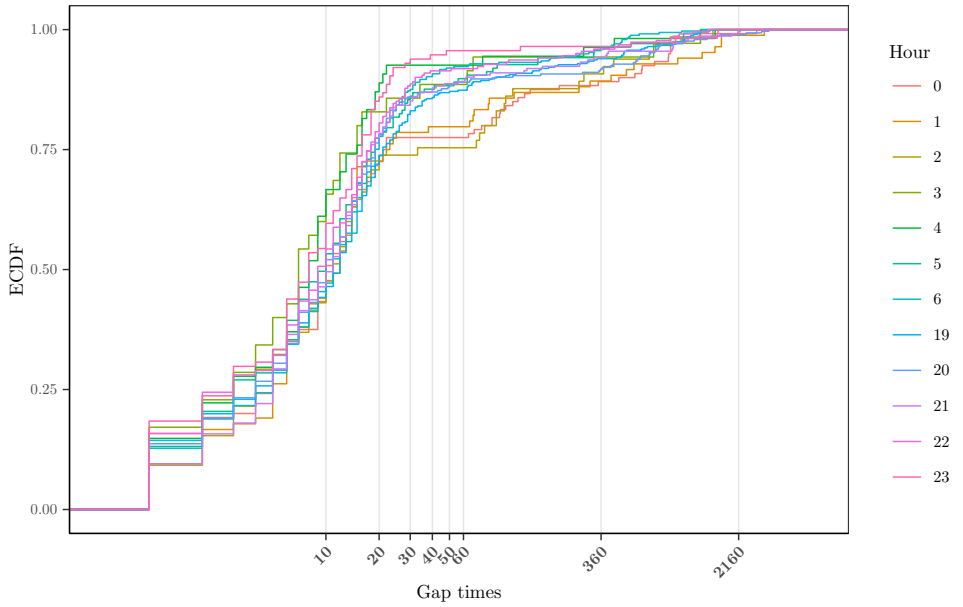


Figure 2.15: The log transform of gap times for time intervals in cluster 2 for Site 12, Monday 11th February, 2019.

Figure 2.16 displays the cumulative distribution function of the log-transform of all gap times in cluster 1 and cluster 2 (i.e. considering the cumulative distribution of all gap times from the sub-clusters in cluster 1 and cluster 2). It shows how the gap times distributions of these two clusters are different. For example, it is going up faster in cluster 2 during non-busy hours than in cluster 1 during busy hours, so the gap times

tend to be smaller at these times. When there is no traffic, it can be said to have smaller gap times because devices can depart the area more quickly than congested traffic. The difference is mostly around 5 and 20 seconds and the biggest difference happens around 15 seconds. The hierarchical clustering time interval for Site 12 for the other days of one week are presented in Appendix A (Figures A.16–A.21).

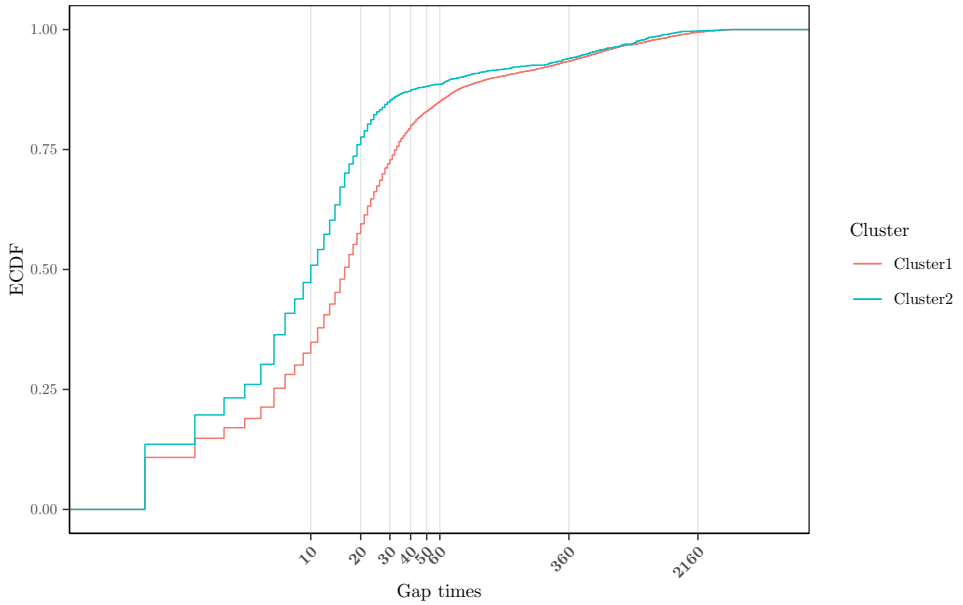


Figure 2.16: The log transform of gap times for time intervals by considering all sub-clusters together in cluster 1 and cluster 2 for Site 12, Monday 11th February, 2019.

2.5 Discussion

In this chapter we have explored whether there might be interpretable information in the complete record of detections at an individual Bluetooth site, including multiple detections rather than filtering them out. We also looked into whether the distribution of observed gap times for multiple detections could provide useful information for traffic inference, because multiple detections result in gaps between consecutive detections. We employed cluster analysis, as an unsupervised learning technique, in two different ways to do this. Firstly, by considering some variables to summarize different aspects of the multiple detection data, and secondly, by directly utilizing the distribution of observed gap times. For the second method, we used KS statistics to compute the distance between the gap times distributions. This KS statistics clustering method could have wide applicability in situations where the clustering objects are sets of data or the distribution of the observed data.

The results of the cluster analysis confirmed that there was information in the multiple detections because the analyses produced meaningful clusters. For example, the

KS clustering of the time intervals of day based on the gap times distributions indicates that the clusters reveal something about the structure of the data. When looking at the dendrogram, it was obvious that there were two different groups of busy and non-busy times of day, with the merging of some of the consecutive time intervals also indicating meaningful clusters. It is important to remember that the clusters results were derived solely from the gap times distribution for multiple detections, therefore there was certainly information since it clustered together adjacent time periods. Also, having meaningful clusters in the outcomes of Bluetooth sites and MAC address clustering demonstrated that there was information in using the complete record of detections. Bluetooth site clustering, for example, showed that sites that were nearby together recorded relatively similar data, and it is possible that sites that were far away in terms of location but clustered over the hours had common characteristics.

Another question that has been raised was if this information could be beneficial for traffic inference. Depending on the considered variables, for example, Bluetooth sites clustering using raw Bluetooth data was useful for identifying unusual Bluetooth sites. After that, outliers can also be examined in order to determine which factors caused the difference between sites. It could, for example, be due to a faulty detector or improper installation. This has been seen in the case of reporting Site 13, where it has shown completely different behaviour in terms of the total number of Bluetooth detections. After reporting to TfGM, it was revealed that there was a printer with an active Bluetooth device in the coverage area which was transmitting all the time. Another advantage that should be mentioned is the generalisation of any future analysis to sites with similar characteristics. Also, it was possible to detect unusual and similar MAC address behaviour using both methods, based on candidate variables and gap distributions, so that any unsual behaviour can be investigated.

It would have been more beneficial if we could have training data with the information about the type of device or type of vehicle, which may also help us categorize different transportation modes, such as pedestrians, bicycles, and vehicles, with respect to their similar gap times distributions. Finally, the time intervals clustering based on the gap distributions using KS statistics can be helpful to distinguish the time intervals with similar traffic patterns. Therefore, such information could be useful for gaining insights into traffic conditions at each Bluetooth location separately and without filtering any data.

However, there are several limitations due to the poor quality of the Bluetooth detection data. In particular, the issue of missed detections will have a big effect, especially on the analysis of the gap times distribution. As noted before, missed detection can happen in both congested and free-flowing traffic. For instance, MAC addresses with one detection are assumed to belong to devices travelling through the area during

a non-busy time interval of the day, but they can be considered as missed detections during a busy time frame. Determining if a MAC address with multiple detection belongs to a device that stayed in the area but the detector lost its detection, or a device that left and returned to the area, can be challenging in some cases.

As mentioned above, KS-clustering can have wider applications in any situation requiring the clustering of distributions. There is an issue of defining the distance between groups for the KS-clustering. We also explored alternative ways of defining the distance between groups, required for iteratively grouping objects into a hierarchical tree. A natural way to do this would be to pool all the data within a group to make a single distribution. Consider the three object distributions of A, B, and C, for which KS statistics compute the pairwise distance as the maximum distance between them. If the KS distance between A and B is less than the KS distance between A and C and between B and C, then the objects A and B will be linked together at a specific level while drawing the dendogram. It is a requirement that the distance between C and the A, B group must be greater than the distance between A and B, so that C would join them in the dendogram at a higher level. However, we found that when the distances between groups are defined as the KS distance between the pooled data in each group, this condition is sometimes violated, so a coherent dendogram cannot be formed. We therefore relied on the standard linkage methods, such as Ward's method. As a counter-example, the black, red, and blue lines in Figure 2.17 correspond to the ecdf for A, B, and C, respectively. The maximum distance between A to B is smaller than the maximum distance between A to C and B to C. However, because the maximum distance from the combination $c(A, B)$ (i.e. green line) to C would be much smaller, C will join them at a lower height in the dendrogram. In this example, the mean of C is in between means of A and B, with a significantly smaller variance.

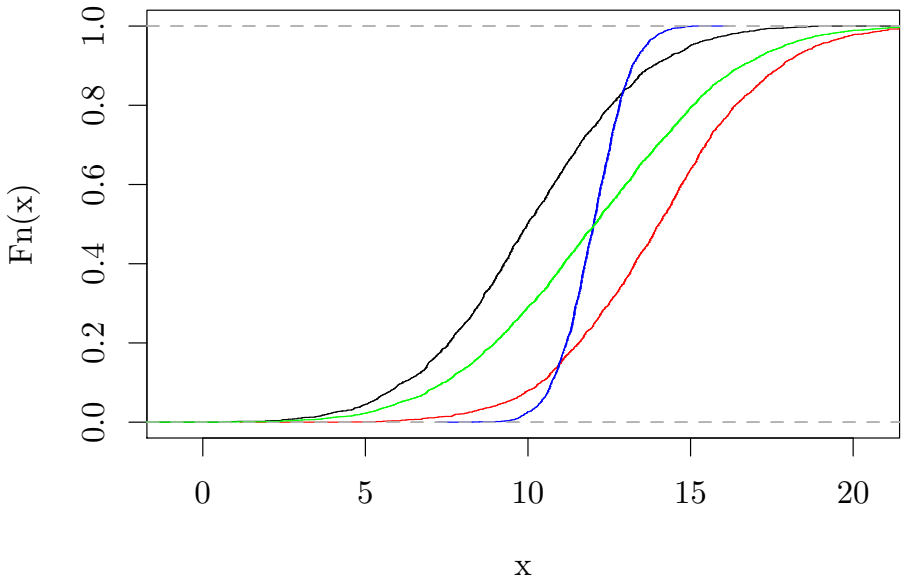


Figure 2.17: An example of using KS statistic as linkage method. The black, red, blue and green lines correspond to the ecdf for A, B,C and the combination $c(A, B)$, respectively.

Chapter 3

Modelling the relationship between Bluetooth and Automatic Traffic Counts

3.1 Introduction

This chapter presents the methodology adopted in this research on modelling the relationship between Bluetooth and Automatic Traffic Counters (ATC). Another method for collecting traffic data is ATC, which is characterized by two rubber tubes spread across the carriageway and attached to a recording box on the roadside. These tubes can operate at any time of day or night for the duration of the specified period. They can record the exact time, the instantaneous speed, and number of vehicle movements, as well as the classification of vehicles passing through, such as cars, buses, or heavy goods vehicles, in contrast to Bluetooth detectors, which are unable to distinguish between vehicles, pedestrians, or their specific location within the detection zone. ATC data can be used to determine the exact number of traffic flows in a particular region, but it is expensive to maintain and install especially if it is needed in numerous places. Table 3.1 shows the ATC classification guide that is utilised in the TfGM database. The Greater Manchester network uses two types of ATC counters. Permanent ATCs are usually installed on major roads and corridors. They measure vehicles via inductive loop detectors which record changes to electromagnetics fields as vehicles pass over them. The other type is the temporary ATC that are deployed for short durations (7-14 days) and they use pneumatic tubes which measures changes in air compression as vehicles pass over them.

Name	Description
Car	Car, van, car + trailer
R2X	Two axle rigid (Heavy Goods Vehicles)
R3X	Three axle rigid (Heavy Goods Vehicles)
R4X	Four axle rigid (Heavy Goods Vehicles)
R2X+T	Two axle rigid + trailer (Heavy Goods Vehicles)
R3X+T	Three axle rigid + trailer (Heavy Goods Vehicles)
A2+1X	Artic, two axle tractor unit + one axle semi-trailer (Heavy Goods Vehicles)
A2+2X	Artic, two axle tractor unit + two axle semi-trailer (Heavy Goods Vehicles)
A2+3X	Artic, two axle tractor unit + three axle semi-trailer (Heavy Goods Vehicles)
A3+1/2X	Artic, three axle tractor unit + one or two axle semi-trailer (Heavy Goods Vehicles)
A3+3X	Artic, three axle tractor unit + three axle semi-trailer (Heavy Goods Vehicles)
Bus	Bus or coach (Public Sevice Vehicles)
Other	Any other vehicle (Not classified)

Table 3.1: ATC classification guide.

Bluetooth data does not provide an exact estimate of all traffic, but instead a proportion of it. Therefore, Bluetooth data only shows a small sample of the total traffic flow. The smaller sample size is assumed because not everyone and all modes in the network have active Bluetooth devices, and even when they are turned on, Bluetooth may not be activated or detected by the detector. The question is whether the sample is biased or representative, and if biased, what factor or factors affect it.

The integration of the ATC data with Bluetooth data could help to describe a possible relationship between the rate of the Bluetooth detections and the number of the vehicles passing through the location. It motivates an investigation into the relationship between ATC and Bluetooth detections.

Before analysing and modelling this relationship, it is necessary to consider the limitations and challenges of employing Bluetooth data. For example, as noted before in Section 1.5, not all vehicles have Bluetooth devices, and not all Bluetooth detections are associated with vehicles. At the detection area, they might be captured from bicyclists and pedestrians. Furthermore, the Bluetooth scanner might lose some MAC address detection. This happens due to some different factors as follows (Michau et al., 2014):

- Signal strength is reduced as a result of physical obstacles such as walls and billboards.
- Due to the increasing number of detectable Bluetooth devices in the detector area interacting with each other, detection quality is decreasing.
- While the sensor is at its back-off time, which is required to detect, truncate, and record MAC addresses, a high-speed car with an active Bluetooth device passes

through the detector coverage area.

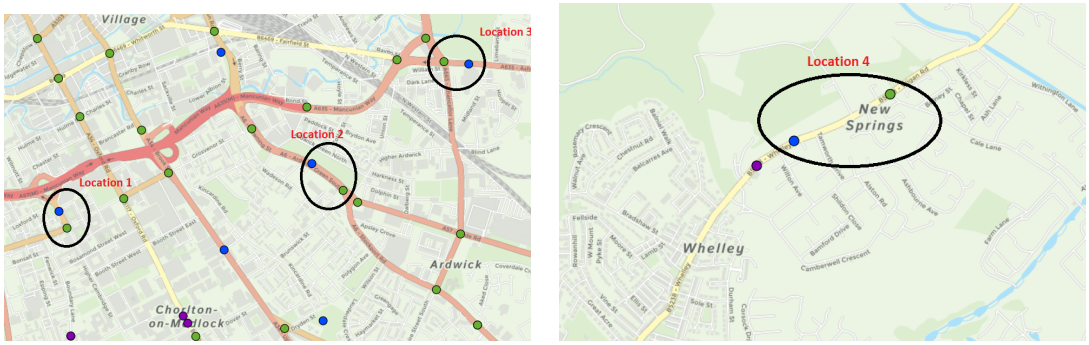
Therefore, the location of the sensor has a significant effect on the detection pattern. For example, having signalised lights or gas stations causes the vehicles to stay more in the coverage zone of the detector, thus the number of Bluetooth detections would be expected to increase. Also, installing the detector near places like shopping malls, parks, etc., where there are lots of pedestrian movements, means higher chances of Bluetooth detection even in the absence of any vehicles (Carpenter et al., 2012). Therefore, the Bluetooth data is not a representative sample of all vehicles passing through the area, and any inferences drawn from the Bluetooth data might be biased.

Crawford et al. (2017) also looked into some of the potential sources of biases in the use of Bluetooth data for traffic monitoring, such as Bluetooth sensor detection rates that vary depending on location, travel direction, and time of year, to assess the feasibility of using Bluetooth data to examine repeated travel behaviour. To do this, they looked at three locations with Bluetooth detectors and ATC in close proximity. A scatterplot of 5 minute ATC data against the number of unique Bluetooth detections at all three sites revealed a non-linear relationship between the Bluetooth and ATC data. The relationship pattern suggested two stages, the first of which is defined by small volume of traffic and an apparently constant rate of Bluetooth detections, and the second of which begins when traffic numbers reach a particular level and the rate of Bluetooth detections rapidly increases. It was hypothesized that the detection rate may depend on some factors. For example, it could be higher during peak times due to more vehicles in the coverage zone or differences in the types of people travelling; some may be more likely to travel with a Bluetooth-enabled device.

To explore this issue further, we also selected four locations such that both the ATC counter and Bluetooth detector are nearby and one year of data was adopted for each location. Table 3.2 presents description of selected locations and Figure 3.1 shows the study locations (black circles) on the Manchester map.

Location	Site Name	Description
Location 1	1421	Cambridge St (A5067) / 70m N of Cavendish St, Manchester (ATC)
	MAC4149MR	Cambridge St / Cavendish St (1090), Manchester (Blu)
Location 2	1165	Ardwick Green (A6) / 35m E of Hamsell Rd, Ardwick, Manchester (ATC)
	MAC1081MR	Ardwick Green S (A6) / 10m N of Brunswick St (3/119), Ardwick, Manchester (Blu)
Location 3	1416	Ashton Old Rd (A635) / 90m E of Chancellor Ln, Ancoats, Manchester (ATC)
	MAC4013MR	Ashton Old Rd (A635) / Chancellor Ln (11), Ardwick, Manchester (Blu)
Location 4	1277	Whelley (B5238) / 60m E of Wilton Ave, Wigan (ATC)
	MAC4313WG	Ashton Old Rd (A635) / Wigan Rd (B5238) / Cale Ln (10/188), Whelley, Wigan (Blu)

Table 3.2: Description of selected locations.



(a) (b)

Figure 3.1: Maps showing the study locations (black circles) in Manchester, the green and blue circles are the Bluetooth detector and permanent ATC: (a) Locations 1,2 and 3 (b) 4.

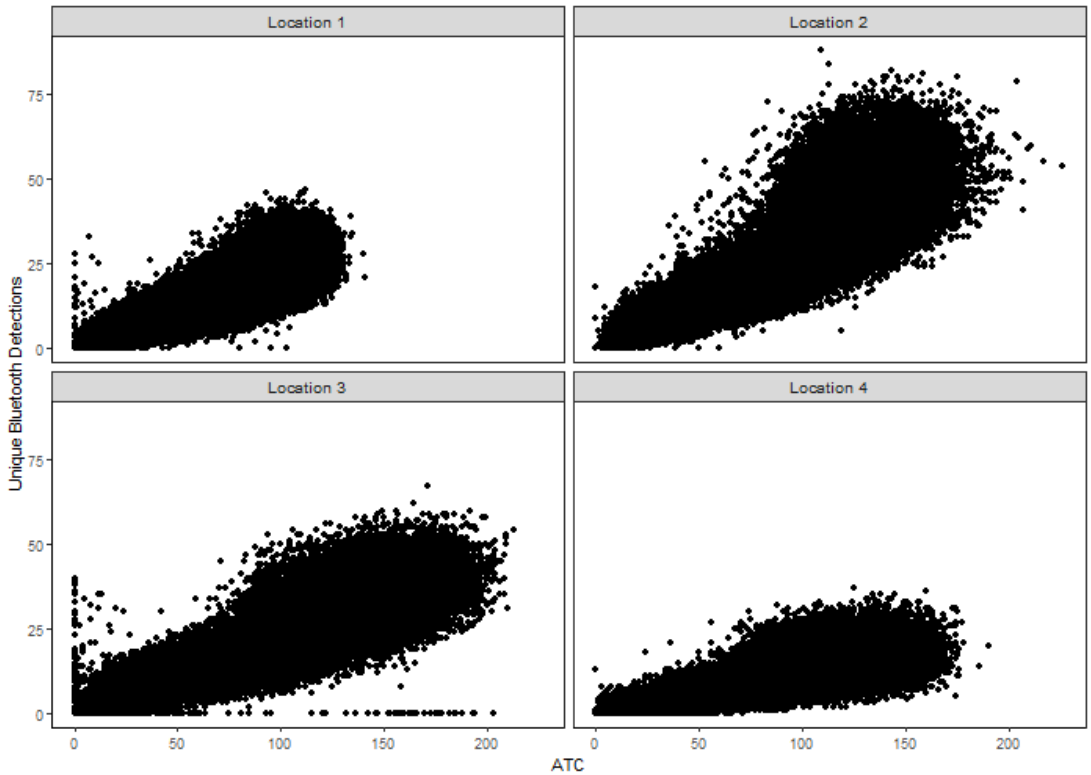


Figure 3.2: Five minute counts of ATC and unique Bluetooth detections at the four considered locations for a one-year period (2018).

Figure 3.2 presents the five-minute counts of ATC against the unique number of Bluetooth detections at all four locations for a one year period (2018). The unique number refers to the fact that we have filtered Bluetooth data by considering only one observation per MAC address, ignoring multiple detections.

A similar non-linear relationship as described in Crawford's study was observed in the four locations (Crawford et al., 2017). It can be seen from Figure 3.2 that there is a positive correlation between the number of Bluetooth detections and ATC counts for all locations but the rate (the slope of the line) does not seem to be constant. Moreover, there is more variation, especially when it moves from lower to higher values of ATC. There are a few outliers, especially in locations 1 and 3, which show that even if the movement of any vehicle is not recorded (i.e. no ATC record), the Bluetooth can be detected from an active device by pedestrians or cyclists. Location 3 shows that there were some five-minute intervals when very few Bluetooth were detected regardless of the higher number of ATC. It may have occurred when there is no traffic and the vehicles are passing fast enough through the zone of the detector, so the Bluetooth sensor loses the chance of Bluetooth detection, or maybe the BT detector was temporarily disabled.

One possible cause of the variation of the changing rate of detections is buses. Our first hypothesis is that the Bluetooth detection rate will be higher as buses drive around, because a bus will be transporting a higher number of passengers with active Bluetooth devices. The initial idea in this chapter is to consider whether the number of buses is one of the factors that explains the variation. Table 3.1 explains how the ATC records are classified by vehicle, so we can include the number of buses in the analysis.

A second hypothesis we explore is whether speed is another factor explaining this variation. It is expected that the rate of detection is affected by the speed of vehicles passing through the detection area. A high-speed car passing through the area, for example, may be missed (Stevanovic et al., 2015). However, if there is a traffic jam that forces vehicles to slow down or stop, several interfering active devices may reduce the transfer of data to the detector.

Accordingly, the first major requirement for this exploration will be to develop a valid statistical model to explain the relationship between the rate of Bluetooth detection per ATC, which then can help to give statistically supported insights into the possible factors that might explain the variation. Table 3.3 summarises the total number of unique Bluetooth detections, ATC, and buses in one year for the selected locations during one year (2018). It can be seen that locations 2 and 3 have the highest and lowest total number of buses, respectively. This also demonstrates that, among the other locations, location 4 has the lowest Bluetooth detection records.

Locations	Bluetooth	ATC	Bus
Location 1	1390729	6144323	74523
Location 2	3015851	9476541	593370
Location 3	2011941	8884268	20430
Location 4	857866	7025772	47930

Table 3.3: Summary of the total number of unique Bluetooth detections, ATC, and buses for the selected locations during one year, 2018.

In this chapter, we first describe the multiple regression models which will be considered to investigate the relationship between the rate of Bluetooth detection and the number of vehicles passing the detection area in Section 3.2. Following that, two methods of variance function estimation are discussed in order to deal with heteroscedasticity, as one of the main assumptions for having a valid regression model. Finally, the implementation and results of regression analysis incorporating the number of buses and speed are presented in Sections 3.3 and 3.3.3, respectively.

3.2 Methodology

3.2.1 Linear regression model

Regression analysis is used for examining the relationship between one dependent variable and one or more other variables, called independent or explanatory variables. This relationship can be explained by the regression equation, which is finally used to forecast or predict the dependent variable. It is called simple linear regression if just one single independent variable is employed to predict the value of a dependent variable, otherwise it is called multiple linear regression. The multiple linear regression is represented as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{3.1}$$

where β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_k$ are called the slopes or the regression coefficients, and ϵ is the error term, also known as the residual.

There are some assumptions for the regression model which if violated, the outcome might be biased and it can affect any meaningful interpretation of the model. These assumptions are as follows:

- the relationship between the dependent variable and the independent variables is linear, called linearity,

- the residuals are normally distributed,
- the residuals are independent,
- the variance of the residuals should be constant for all observations, which is known as homoscedasticity, and if the variance changes; it is called heteroscedasticity (non-constant variance).

Normally distributed errors are considered to be the least important of these assumptions (Lumley et al., 2002; Gelman and Hill, 2006; Knief and Forstmeier, 2021), especially for large samples. Lumley et al. (2002) reviewed the statistics for the t-test and linear regression, as well as what the research literature and textbooks say about those techniques. Following that, simulations based on a large dataset of medical costs are shown. These simulations show that, even with highly non-normal data, linear regression and the t-test can perform well in moderately large samples. Regarding the independence assumption if the errors are positively correlated over time, then standard errors calculated ignoring the correlation will tend to be too small, and the significance of terms in the model might be exaggerated. This assumption can be checked graphically by calculating the autocorrelation function of the residuals sorted in time order.

The initial goal of this chapter was to find a relatively accurate relationship between Bluetooth detections and ATC records, while also taking into account the effect of the number of buses to see if the number of buses can influence the rate of Bluetooth detection (see Section 3.1). The regression model approach can be utilized to explore these relationships, by considering a multiple linear regression model as follows:

$$y = \alpha + \beta x + \gamma z + \epsilon \tag{3.2}$$

where the number of Bluetooth detections, ATC recordings, and buses are represented by y, x, and z in these models, with y as a dependent variable and x and z as independent variables. Here α is the average number of Bluetooth detections if no vehicles pass through the area, β is the average number of Bluetooth detections per ordinary vehicle, and this average would be $\beta + \gamma$ if the bus is travelling through, so γ is the average extra Bluetooth detection per bus compared to other vehicle and finally, ϵ is the error term.

However, looking at Figure 3.2, which is the scatter plots of five minute counts of ATC and unique Bluetooth detections for a one-year period, shows that a straight line does not appear to be a good fit for the data, because there seems to be a change in slope at some point (or points). Table 3.4 presents the linear and other alternative regression models that have been used for this purpose, such as the quadratic, cubic, and segmented (or broken-stick) models. Segmented regression is a type of regression that

allows for the fitting of several linear models to the data for different x ranges, which knots or breakpoints are x values where the linear function's slope changes. Therefore, knots are included in these models. For example, in a segmented model with two knots, c_1 and c_2 represent the knot's values and the indicator function $I(x \geq c_1)$ is defined to be 0 if $x \leq c_1$ and 1 if $x > c_1$ and the slope for each stick is considered by $\beta + \Delta\beta_1$ and $\beta + \Delta\beta_1 + \Delta\beta_2$, respectively.

Model	**Regression equation**
Linear	$y = \alpha + \beta x + \gamma z + \epsilon$
Quadratic	$y = \alpha + \beta x + \beta_1 x^2 + \gamma z + \epsilon$
Cubic	$y = \alpha + \beta x + \beta_1 x^2 + \beta_2 x^3 + \gamma z + \epsilon$
Segmented with one knot	$y = \alpha + [\beta + \Delta\beta_1 I(x \geq c_1)]x + \gamma z + \epsilon$
Segmented with two knots	$y = \alpha + [\beta + \Delta\beta_1 I(x \geq c_1) + \Delta\beta_2 I(x \geq c_2)]x + \gamma z + \epsilon$
Segmented with three knots	$y = \alpha + [\beta + \Delta\beta_1 I(x \geq c_1) + \Delta\beta_2 I(x \geq c_2) + \Delta\beta_3 I(x \geq c_3)]x + \gamma z + \epsilon$

Table 3.4: Some alternative regression models for the effect of ATC records on Bluetooth detections incorporating a different effect for buses. The number of Bluetooth detections, ATC recordings, and buses are represented by y, x, and z in these models. Also, for example, in a segmented model with one knot, c_1 represents the knot's value and the indicator function $I(x \geq c_1)$ is defined to be 0 if $x \leq c_1$ and 1 if $x > c_1$.

When we are working with real traffic data, homoscedasticity may fail because of excess variations resulting from different phenomena, such as different times of day, traffic incidents, weather, or missing data due to technical errors or communication failures in detectors, etc. In the presence of heteroscedasticity, the estimated statistical significance of the independent variables is inaccurate and invalid if the heteroscedasticity is not accounted for in the model. Therefore, the following section will discuss the methodology tested for resolving heteroscedasticity based on the underlying data distribution.

3.2.2 Addressing heteroscedasticity

There are several techniques for resolving heteroscedasticity in regression models, such as the methods based on weighting and the methods based on data transformation (Carroll and Ruppert, 1988).

The weighted regression model, as one of the well-known techniques, has been utilised to correct non-constant variance situations. This method gives a weight based on the variance of the fitted value corresponding to each data point, and this procedure

will assign more weight to the observations with smaller variances as these observations present more reliable knowledge about the regression model than the observations with large variances. This is a very good method for removing heteroscedasticity, if it can use the correct weights.

In order to implement the weighted regression model, it is important to specify a model for the variance of the error term. In the presence of heteroscedasticity, it is expected that the variance of the error term is not constant (i.e., $\text{Var}(\epsilon_i)=\sigma_i^2$, $i = 1, 2, \ldots, n$).

The most common model assumes that the variance is functionally connected to mean. For example, the variance as proportional to the power 2θ of the mean response, i.e. $\text{Var}(\epsilon_i)=\sigma_i^2\mu^{2\theta}$ (or the standard deviation as proportional to the power θ of the mean response) (Carroll and Ruppert, 1988). The gamma or lognormal data follows with θ=1, whilst θ=0.5 would be relevant to the Poisson distribution.

Because the research data is count data, an initial assumption is also that Poisson regression can be used, with $\theta = 0.5$; this will be discussed in the next chapter (Section 4.2). Here, before deciding on the appropriate θ, we decided to let the data provide us with an indication of θ.

In this regard, we first utilised the rolling standard deviation method to visualise how the standard deviation or variance of Bluetooth detections varies in order to get a good idea of θ according to the data.

3.2.3 Rolling variance method

In the rolling variance method, a window of pre-specified length is shifted across the data and the variance is computed for the data in each window. In this method, the data windows can also be created as non-overlapping or overlapping windows. The difference between these two types of windows is that if the non-overlapping can not provide a sufficient number of observations for an accurate evaluation, then the overlapping windows will be an alternative to get a more detailed picture; however the estimated variance are then not independent, because they are based partly on the same data. Finally, the computed data can be plotted in order to see how the standard deviations or variances change in terms of the power of the mean.

To visualise how the variance of Bluetooth detections varies in connection with the mean, we first utilised the rolling variance method. It was performed over the non-overlapping window of length k across amounts of ATC using the following steps:

- Step 1: Assuming window length (e.g. $k = 10$) and stopping criteria (e.g. maximum number of ATC counts).
- Step 2: Considering the slide window based on the assumed length.

- Step 3: Computing the mean and variance (or standard deviation) of Bluetooth detections in the slice window. The bootstrapping method can be utilized here to estimate the confidence intervals by resampling with replacement the data in the window. Bootstrapping is a statistical procedure that uses various simulated samples from a single dataset. For a range of sample statistics, this procedure allows you to calculate standard errors, generate confidence intervals, and do hypothesis testing (Efron and Tibshirani, 1994).
- Step 4: Rolling the window by length k.
- Step 5: Repeat step 3 and 4 until reaching the stopping criteria.

It should be noted that we used non-overlapping windows in order to have independence between samples of each window, because the large sample size would produce sufficient data for an accurate evaluation of variance within quite small windows. Finally, we display the computed variance to show how the variances change in terms of the mean.

The other alternative approach, if this method does not help us to find the an appropriate parametric form of the variance function, is the non-parametric method of variance-function estimation.

3.2.4 Non-parametric variance function estimation

Chiou and Müller (1999) proposed a nonparametric quasi-likelihood regression technique for an unknown variance function. The quasi-likelihood technique is useful in many applications where the exact distribution of the observations is unknown. In this method, the variance function is estimated non-parametrically by smoothing the squared residuals acquired from an initial regression model fit at the estimated mean. For the smoothing method, they used local polynomial fitting (Fan and Gijbels, 1996), also suggesting that any reasonable smoother could be used. They showed that the inference based on the nonparametric quasi-likelihood estimators of the vector of regression parameters is asymptotically equivalent to quasi-likelihood estimation using a known variance function.

This technique utilizes an iterated process in which in each iteration the non-parametric variance function is estimated from the residuals and estimated means using the smoothing method. Then this non-parametric variance function can be used to update and improve the model parameters. Carroll and Ruppert (1988) suggested alternatives to the squared residuals for variance-function estimation, such as weighted absolute residuals and the logarithm of absolute residuals.

We adapted Chiou and Müller's method by smoothing the logarithm of absolute residuals using LOESS (locally estimated scatter plot smoothing), which results in a smooth line capturing the general trend (Cleveland, 1979). The advantage of using the

logarithm of absolute residuals is that they are less skewed than squared residuals, and the relationship will be a straight line if the power of the mean model is correct. The method starts with an unweighted regression model and a variance function estimated by regressing the logarithm of absolute residuals on the fitted values with LOESS. The weights are computed using the estimated variance, and the weighted regression model is fitted until the coefficients do not change significantly. The non-parametric method for estimating the variance-function is performed using the following steps:

- Step 1: Fit an unweighted regression model and obtain the residuals of the considered model.
- Step 2: Estimate the variance function using LOESS by regressing the logarithm of absolute residuals on the fitted values. The exponential of obtained fitted values (i.e. y_i) of this regression is considered as σ_i (here because of logarithm function, exponential of the obtained fitted values are σ_i):

$$e^{y_i} = \sigma_i \tag{3.3}$$

- Step 3: Use this variance function to estimate the weights as $w_i = 1/\sigma_i^2$ and create the weight vector as $\boldsymbol{W} = \{w_1, w_2, \ldots, w_n\}$.
- Step 4: The normalized weight can be computed by dividing the raw weights by its mean and fit the weighted regression model.
- Step 5: Repeat steps 2 to 4 until the coefficients of the weighted regression model converge (i.e. when the coefficients do not change significantly).

3.2.5 Model selection

The process of selecting the best model from a set of models that must be a reasonable approximation of the data-generating process is known as model selection. We investigated the six different regression models given in Table 3.4 to find the best fit for data. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) were utilised to compare the goodness of fit of the different regression models (Burnham and Anderson, 2002; Gideon, 1978). The AIC is calculated as:

$$\text{AIC} = 2k - 2\log(\hat{L}) \tag{3.4}$$

where k is the number of estimated model parameters and $\log(\hat{L})$ is the natural logarithm of the maximum likelihood estimation for the model. Also, BIC is computed as:

$$\text{BIC} = k\log(n) - 2\log(\hat{L}) \tag{3.5}$$

where k and $\log(\hat{L})$ are the same as defined for AIC, and n is the number of observations, or equivalently, the sample size. The better model will be the one with the lower value of AIC or BIC.

3.3 Implementation and results

The selected four locations as case studies are presented in section 3.1. Location 2 has the most Bluetooth detections, ATC records, and buses in total, as seen in Table 3.3, as well as the most obvious change in slope. Therefore, first, we started to fit the linear regression model (3.2) to the data for this location. It needs to be noted that the observations are considered as the number of Bluetooth detections, ATC and bus counts during every five minute time interval. Table 3.5 shows a summary of the parameter estimates for the fitted model related to location 2. The multiple linear regression equation is estimated as follows:

$$y = -1.67 + 0.31x + 0.44z \tag{3.6}$$

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>|t|$)**
α	-1.67	0.052	-32.33	<2e-16 ***
β	0.31	0.001	390.51	<2e-16 ***
γ	0.44	0.008	53.10	<2e-16 ***

Table 3.5: The multiple linear regression coefficients estimation.

It can be seen from Table 2.2 that all the coefficients are statistically significant. There is an average of 0.31 Bluetooth detections per ordinary vehicle and an average of 0.75 (i.e. 0.31+0.44) Bluetooth detections per bus. The negative intercept for this regression model might not be interpretable, because if no vehicles, including buses, pass through the detection area for five minutes, there is still a chance that Bluetooth will be detected by an active device from a pedestrian or cyclist.

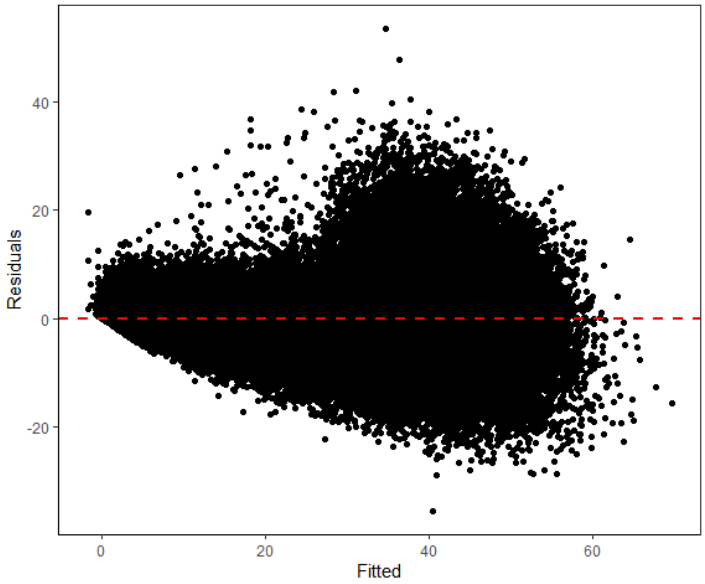


Figure 3.3: Residuals versus fitted value plot for the multiple linear regression model.

As expected, the heteroscedasticity can be seen as a cone shape where the spread of the residuals increases as the fitted value increases (Figure 3.3). Given the above pattern, it will be essential to address heteroscedasticity in order to create a valid statistical model.

In this regard, the rolling variance method and the non-parametric variance function estimation are utilised to specify an appropriate model for the variance function based on data. The results are presented in the following sections.

3.3.1 Results of the rolling variance method

To begin, the rolling variance approach with a window length of $k = 10$ is utilised to specify a variance model based on the data. The result of the variance versus the mean of Bluetooth detections across the rolling windows of ATC counts for location 2 is shown in Figure 3.4. The estimates from each window are shown as circles, with bootstrap 95% confidence limits as dashed lines.

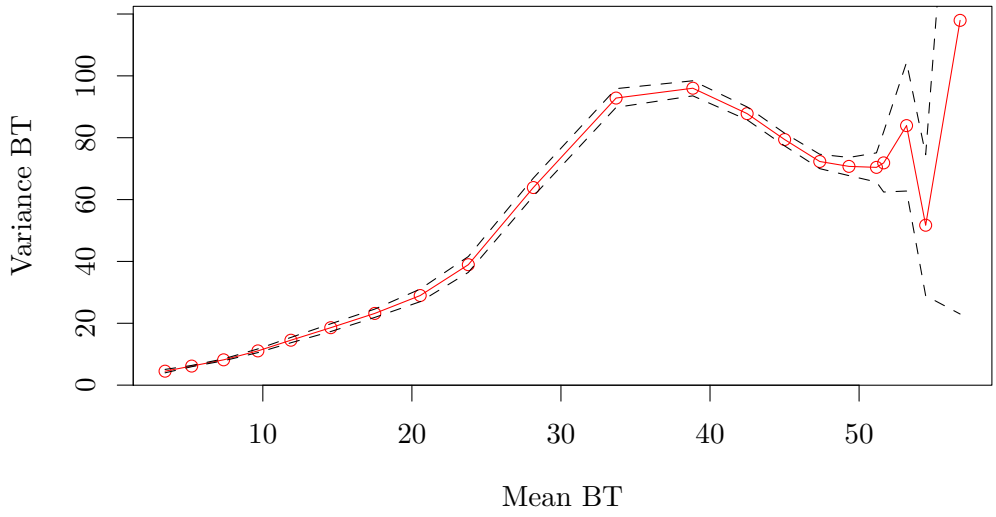


Figure 3.4: The variance versus the mean of Bluetooth detections across the rolling windows of ATC counts for location 2. The estimates from each window are shown as circles, with bootstrap 95% confidence limits as dashed lines.

An initial linear trend suggests that the variance is proportional to the mean for small values of ATC, but this pattern displays that something has happened to cause the variance to first increase sharply and then go down as ATC increases (here when ATC counts are higher than 120). It does not allow us to make a decision about the value of θ. It is difficult to find out what underlying factors cause this extra variation in traffic situations, although it may be that this extra variance is caused by two different situations. Firstly, when there is no traffic jam, a large number of vehicles can pass through the location fast enough to not be detected by the Bluetooth detector. Secondly, when traffic congestion enforces so many vehicles in the area, the interference of detectable Bluetooth devices reduces the effectiveness of detection. In addition, other factors may also affect the hourly, weekday, monthly and seasonal variations. Therefore, as the variance does not appear to be proportional to a power 2θ of the mean, the next step is to employ the non-parametric method as the other alternative approach. The results of this method will be given in the next section.

3.3.2 Results of non-parametric variance function estimation in regression models including bus

We used the non-parametric method of variance-function estimation because the rolling standard deviation method failed to suggest an appropriate parametric variance model

based on the data. Figure 3.5 presents the result of regressing the logarithm of absolute residuals on the fitted values using the LOESS method after the weighted linear regression model is converged for Location 2. The red line shows the LOESS smoother with a span of 0.90, which uses 90% of the points in each window. After ten iterations, convergence was reached, which is considered as no significant change in the coefficients.

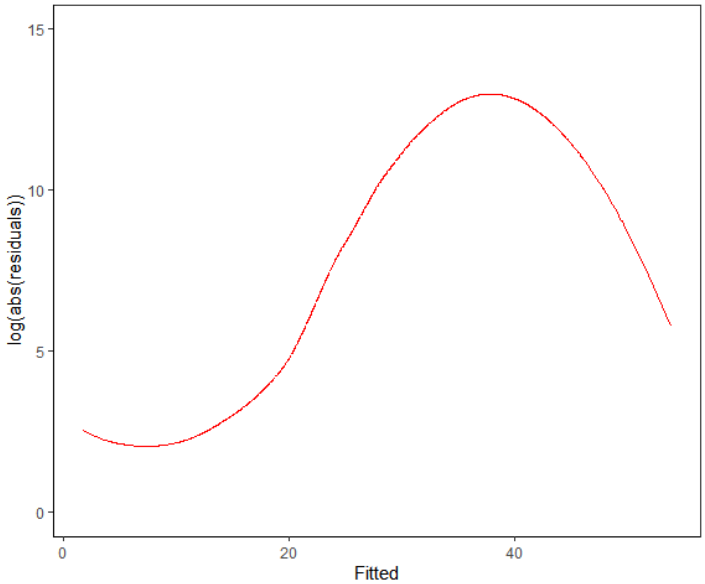


Figure 3.5: The result of regressing the logarithm of absolute residuals on the fitted values using the LOESS method after the weighted linear regression model is converged for location 2.

In this approach, the regression models shown in Table 3.4 are considered as weighted regression models. The multiple linear regression (3.6) after accounting for the non-parametric variance function and as a weighted linear regression is:

$$y = 1.77 + 0.23x + 0.04z \tag{3.7}$$

Table 3.6 presents a summary of the parameter estimates for the fitted this model for location 2. The coefficients are statistically significant, as can be seen, but they have changed and become smaller. The positive, statistically significant, coefficient of buses suggests that buses have an effect on the rate of Bluetooth detection, which increases as the buses pass through the detection area. Although the intercept is not of interest here, as it is influenced by non-vehicle detection, it has also become positive and meaningful.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	1.77	0.022	79.078	<2e-16 ***
β	0.23	0.008	273.131	<2e-16 ***
γ	0.04	0.007	4.881	<1.06e-6 ***

Table 3.6: The estimated coefficients of the weighted multiple linear regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 2.

Figure 3.6 shows homoscedasticity in the residual plot after the weighted regression model is converged, along with a few remaining problems when the average number of Bluetooth detections is low.

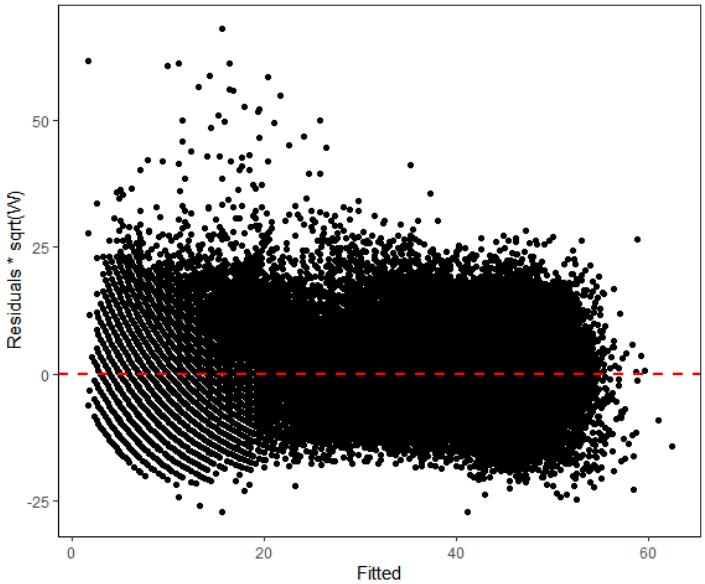


Figure 3.6: Residuals versus fitted value plot after the weighted multiple linear regression model is converged for location 2.

The last step is to compare the alternative regression models given in Table 3.4. Cross-validation is a common approach for model comparison. Cross-validation was used by Rosenberg et al. (2003) to compare models with parametric and non-parametric terms in the mean response model. However, because we have a large dataset and a complex fitting procedure, cross-validation will be too computationally intensive in our situation. Two alternative approaches, AIC and BIC, will be used here to make comparisons. It should be noted that the model for the mean response contains only parametric components, but the variance is modelled non-parametrically. The performance of AIC or BIC in such situations has not been widely studied. Yang (1999) presented a new model complexity criterion called ABC, which differs from AIC in that it includes a model complexity error term for non-parametric components. The method was

demonstrated for adaptive smoothing, in which the amount of smoothing varies between models. We could use the effective number of parameters (enp) for considering the degrees of freedom of the non-parametric variance function aspect and add it as a penalty term to AIC and BIC. However, in our modelling the amount of smoothing in the loess phase is kept constant across all models, so Yang's penalty term for model complexity would change slightly and can be ignored. Examination of the enp for the models confirmed that the values are very close.

The comparison results show that the weighted segmented regression model with three knots or break-points has a lower AIC and BIC than the other models, indicating that it is the best-fitting line for the data (Table 3.7).

Model	**df**	**AIC**	**BIC**
Weighted linear	4	687395.0	687433.3
Weighted quadratic	5	680889.2	680936.4
Weighted cubic	6	678610.8	678668.2
Weighted segmented with one knot	6	678951.0	679008.4
Weighted segmented with two knots	8	676528.1	676604.6
Weighted segmented with three knots	10	**676155.4**	**676342.0**

Table 3.7: Comparison AIC and BIC between the weighted regression models for the effect of ATC records on Bluetooth detection incorporating effect for buses at location 2.

The summary of the coefficient estimations for all other alternative regression models is presented in Tables (3.8–3.12). All of the coefficients are statistically significant, and the results show that buses have an impact on the rate of Bluetooth detections, which as predicted will increase as the buses pass through the detection area. However, if the number of buses was the full explanation for the changing rate, including the buses in the model should have resulted in the linear model being chosen as the best model, which did not happen, indicating that the buses were not the only factor in the rate variation. It should be noted that p-values for $\Delta\beta_1$, $\Delta\beta_2$ and $\Delta\beta_3$ are NA in Tables (3.10–3.12), because the difference between two slopes really comprises two parameters: the size and the location of the change.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	2.06	0.035	58.58	<2e-16 ***
β	0.18	0.001	125.47	<2e-16 ***
β_1	0.001	0.001	81.55	<2e-16 ***
γ	0.35	0.008	44.22	<2e-16 ***

Table 3.8: The estimated coefficients of the weighted multiple quadratic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 2.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	3.88	0.053	73.77	<2e-16 ***
β	0.04	0.003	14.14	<2e-16 ***
β_1	0.003	0.00005	64.09	<2e-16 ***
β_2	-0.00001	0.0000002	-50.37	<2e-16 ***
γ	0.28	0.008	35.45	<2e-16 ***

Table 3.9: The estimated coefficients of the weighted multiple cubic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 2.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	1.94	0.035	55.33	<2e-16 ***
β	0.21	0.001	171.14	<2e-16 ***
$\Delta\beta_1$	0.14	0.001513	93.52	NA
γ	0.31	0.008	38.72	<2e-16 ***

Table 3.10: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with one knot $c_1 = 59$ at location 2.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	1.65	0.031	53.80	<2e-16 ***
β	0.23	0.001	234.01	<2e-16 ***
$\Delta\beta_1$	0.22	0.003	72.36	NA
$\Delta\beta_2$	-0.25	0.005	-47.87	NA
γ	0.23	0.008	28.62	<2e-16 ***

Table 3.11: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with two knots $c_1 = 77$ and $c_2 = 124$ at location 2.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	1.99	0.04	49.91	<2e-16 ***
β	0.21	0.002	136.57	<2e-16 ***
$\Delta\beta_1$	0.07	0.004	21.12	NA
$\Delta\beta_2$	0.21	0.006	33.57	NA
$\Delta\beta_3$	-0.28	0.006	-43.38	NA
γ	0.23	0.008	28.87	<2e-16 ***

Table 3.12: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with three knots $c_1 = 50$, $c_2 = 88$ and $c_3 = 122$ at location 2.

Figure 3.7 shows the predicted values against ATC count, for a fixed number of buses (here equal to 5), obtained from weighted segmented models with different number of knots. It demonstrates that there is only a slight difference between the segmented model with two knots ($c_1 = 77$, $c_2 = 124$) and three knots ($c_1 = 50$, $c_2 = 88$, $c_3 = 122$). The statistically significant improvement from using the third knot could have happened due to the large sample size (i.e. here, the sample size equals 105120; the total number of five-minute periods in a year). Therefore, the slight difference is still statistically (but not practically) significant. Hence, the model with two knots could well be used to explain the relationship.

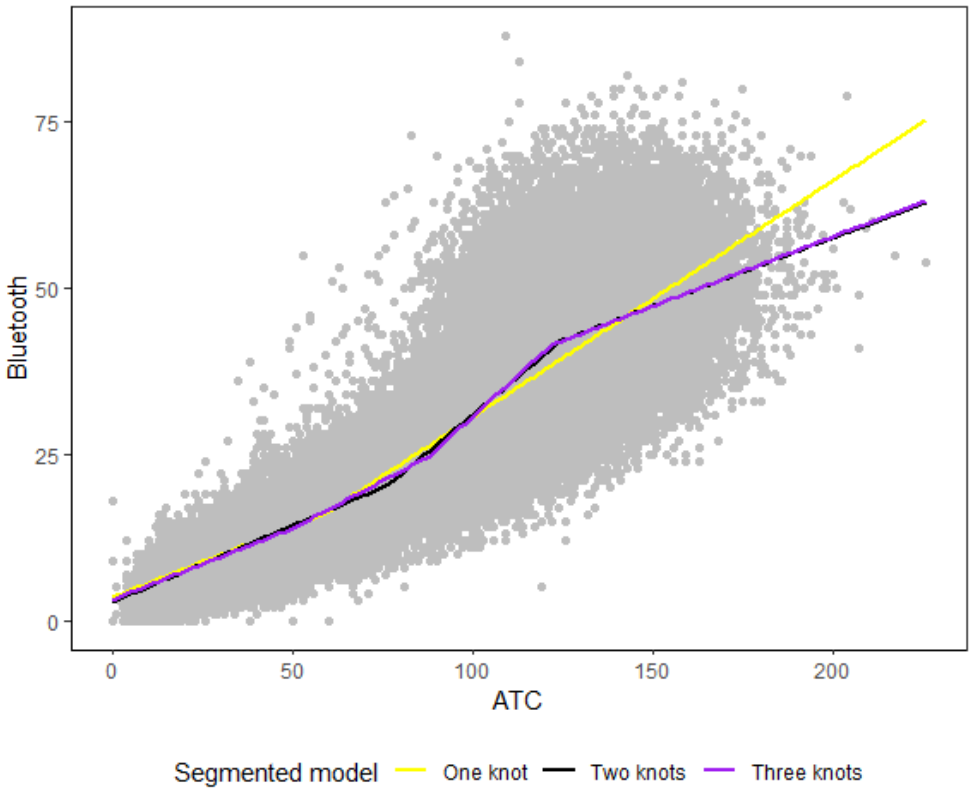


Figure 3.7: The predicted values against ATC count, for a fixed number of buses, obtained from the weighted segmented model by the different number of knots.

Following that, based on the obtained slopes of the model with two knots, the changes in Bluetooth detection can be explained as follows. For all ATC counts below the first knot, $c_1 = 77$, the slope is positive (i.e. $\beta = 0.23$) and that means the rate of Bluetooth detections is constant at 0.23, i.e. approximately 23 % of vehicles are detected. There is a significant increase in the slope between the first knot and the second knot ($\beta + \Delta\beta_1 = 0.45$). This indicates that the rate of Bluetooth detections is expected to increase when ATC counts are between 77 and 124 (i.e. approximately 45 %). It was predicted that when the vehicles stay more in the coverage zone of the detector, the rate of Bluetooth would be expected to increase, so this region may represent traffic congestion. Finally, on the other side of the second knot, the slope is roughly the same as the first segment ($\beta + \Delta\beta_1 + \Delta\beta_2 = 0.2$), which suggests that the rate of Bluetooth detections is predicted drop to 0.2, i.e. approximately 20 % of vehicles detected.

As discussed in Section 3.1, the possibility of missed detections increases as the number of detectable Bluetooth devices in the detector area increases, as well as when there is a fast, free-flowing traffic condition in which vehicles can leave the area without being detected. For this location, the slope of the first segment shows that there

is no traffic (due to the low number of ATC), and the third part reflects the traffic (due to the high number of ATC). Both segments showed the same rate of Bluetooth detection, with the higher rate in the second segment. We performed similar analysis for the other locations, and Tables (A.1–A.21) in Appendix A.4.1 present the estimated coefficients of weighted regression models and comparisons between them. The coefficients of the number of ATC and buses are positive, and statistically significant for all other locations. The rate of Bluetooth detection for the third segment (the congested traffic situation) is lower than the rate for the first segment at these locations (the free-flowing traffic situation). This could imply that in a congested traffic situation, the detector would lose more Bluetooth detections. Also, for locations 1 and 3, the weighted segmented regression model with three knots was shown to provide a better fit. However, the same as location 2, the slight difference between the weighted segmented regression model with three knots and two knots is statistically (but not practically) significant. Hence, the segmented regression model with two knots could be considered as the final model. For location 4, the weighted segmented regression model with two knots was chosen as the best fit in first place. The comparison of locations suggests that location 2, where the number of buses is higher than others (see Table 3.3), also has the highest bus coefficients, and location 1 has the lowest bus coefficients.

Furthermore, there may be some similarities between the traffic patterns for locations 1 and 4, as well as for locations 2 and 3, firstly because the knots of the weighted segmented regression model that indicate the value of the ATC count where the slope of the linear function changed, and hence the rate of Bluetooth detection, seem to be close together. Secondly, the non-parametric variance function estimation also shows a roughly similar trend for these locations.

Finally, the number of buses has an effect on the rate of Bluetooth detections. When the buses are present, the rate will be higher. However, because the segmented regression model with two knots was chosen for the best fit, rather than the linear model , it suggested that this is not the only explanation for the rate variation. In the following section, then, we will incorporate speed to see whether the Bluetooth rate variation is justified.

3.3.3 Results of non-parametric variance function estimation in regression models including speed

As described in Section 3.1, the ATC data also include the instantaneous speed of vehicles passing through the detection area. Here we hypothesized that including speed might be able to differentiate congested versus flee-flowing traffic at higher volumes. To account for the effect of speed in the multiple linear regression, we include speed as an

interaction term as follows:

$$y = \alpha + \beta x + \gamma z + \omega s + \delta(x \times s) + \epsilon \tag{3.8}$$

where the number of Bluetooth detections, ATC recordings, buses and the average speed of vehicles passing through the detection zone are represented by y, x, z and s, respectively. Table 3.13 presents AIC and BIC comparison of the six regression models with the interaction term ATC×speed at location 2, showing that the segmented regression model with three knots is a better fit for data.

Model	**df**	**AIC**	**BIC**
Weighted linear	6	675957.0	676014.4
Weighted quadratic	7	673860.8	673927.8
Weighted cubic	8	672184.4	672260.9
Weighted segmented with one knot	8	675047.8	675124.3
Weighted segmented with two knots	10	669959.3	670054.9
Weighted segmented with three knots	12	**669850.7**	**669965.5**

Table 3.13: Comparison AIC and BIC between the weighted regression models for the effect of ATC records on Bluetooth detection incorporating speed at location 2.

Figure 3.8 shows the predicted values of Bluetooth detection rate versus ATC computed from segmented models with different numbers of knots when a fixed number of buses (here equal to 5) and a fixed speed (here equal to 48 km/h speed) are being used. It demonstrates that there is only a slight difference between the segmented model with two knots ($c_1 = 80$,$c_2 = 123$) and three knots ($c_1 = 50$, $c_2 = 88$, $c_3 = 121$). Therefore, the relationship could be explained using the segmented model with two knots.

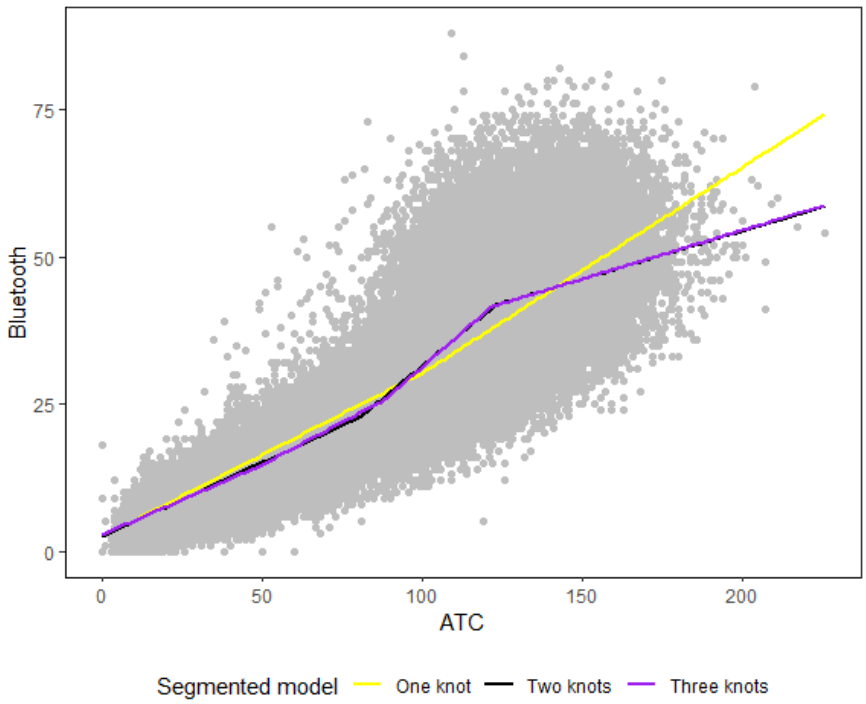


Figure 3.8: The predicted values of Bluetooth detection rate versus ATC computed from segmented models with different numbers of knots when a fixed number of buses (here equal to 5) and a fixed speed (here equal to 48 km/h speed limit) are being used.

The estimated coefficients of the weighted segmented model with two knots for location 2 using the non-parametric variance function estimation approach are summarised in Table 3.14 and other regression model results are presented in Appendix A.4.2 (see Tables A.22 – A.26).

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	1.61	0.254	6.31	<2.82e-10 ***
β	0.43	0.003	119.242	<2e-16 ***
$\Delta\beta_1$	0.19	0.003	56.930	NA
$\Delta\beta_2$	-0.28	0.005	-53.110	NA
γ	0.16	0.007	20.675	<2e-16 ***
ω	0.003	0.004	0.580	0.562
δ	-0.004	0.0001	-57.239	<2e-16 ***

Table 3.14: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporate buses and speed with two knots $c_1 = 80$ and $c_2 = 123$ at location 2.

The results showed that the rate of Bluetooth detection decreases when the vehicle speed increases. The coefficient of ATC $\times$ speed (i.e. δ) can be used to interpret the

effect of speed on rate, which showed that an increase in speed of 10 km/h resulted in a rate reduction of 0.04 (from 0.43 to 0.39). This could relate to free-flowing traffic conditions in which the vehicle can leave the detection area faster and the detector is less likely to detect it. The coefficient of speed (i.e. ω) is difficult to interpret as it represents the influence of speed when ATC = 0, and surely there is no recorded speed if there is no vehicle passing through the area. However, this coefficient is not significant in the best chosen model.

The result of regressing the logarithm of absolute residuals on the fitted values using the LOESS method after the weighted segmented regression model with two knots, which was selected as the best model, is converged for location 2 is shown in Figure 3.9.

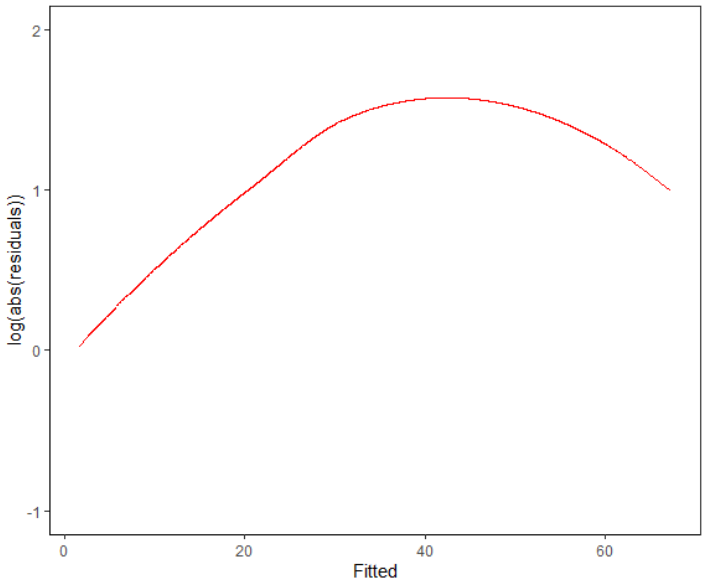


Figure 3.9: The result of regressing the logarithm of absolute residuals on the fitted values using the LOESS method after the weighted segmented regression model with two knots selected as the best model is converged for location 2.

Figure 3.10 shows homoscedasticity in the residual plot after the weighted segmented regression model is converged with a few remaining problems when the average number of Bluetooth detections is low.

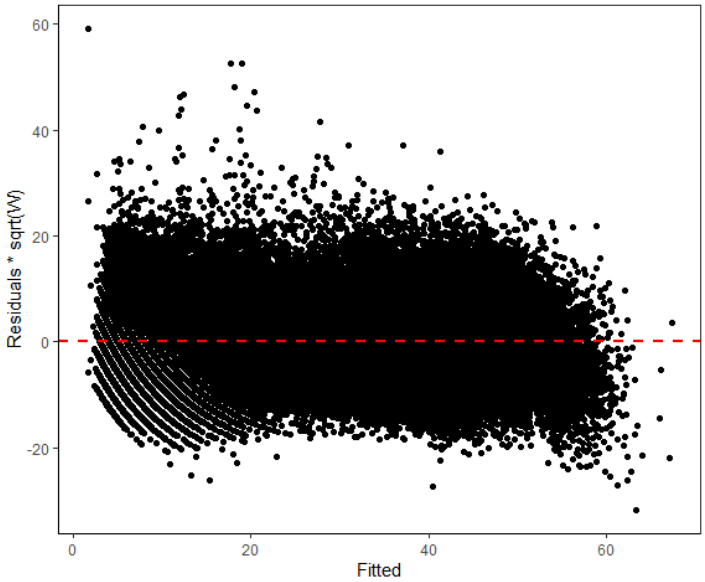


Figure 3.10: Residuals versus fitted value plot after the weighted segmented regression model with two knots is converged for location 2.

Figure 3.11 shows the autocorrelation of the residuals from the weighted segmented regression model with two knots. It shows that there is significant, but not strong, positive autocorrelation in the residuals of the final model. This suggests that other unaccounted sources of variation resulted in some positive autocorrelations, which led to the rate being considered as a time-dependent variable in the following chapter.

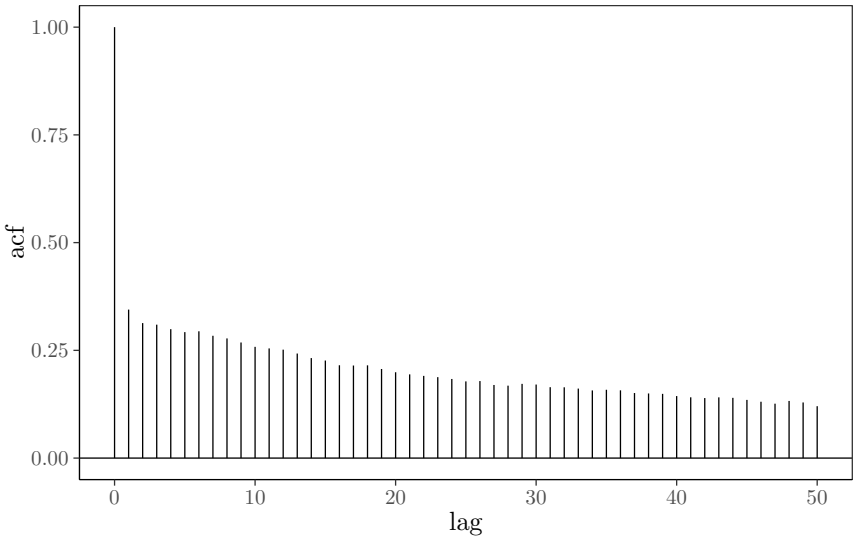


Figure 3.11: The autocorrelation of the residuals from the weighted segmented regression model with two knots for location 2.

We performed similar analysis for the other locations, and Tables (A.27–A.47) in Appendix A.4.2 present the estimated coefficients for weighted regression models and

a comparison between them. The bus and ATC× speed coefficients are statistically significant with the same signs as for the first location positive but different for the bus; and negative but different for ATC ×speed coefficients. In some cases the main effect coefficient for speed (ω) was statistically significant. This is difficult to interpret because the effect of speed on rate is most likely nonlinear, a significant ω coefficient can be interpreted as an attempt to model this nonlinearity. Another alternative explanation is that there are some unobserved confounding factors, and the main effect of speed serves as a channel for expressing those unobserved variables, say the environment and traffic conditions. The weighted segmented regression model with three knots was shown to provide a better fit for locations 1 and 4, while the segmented regression model with two knots was chosen as the best fit for location 3. Given that the weighted segmented regression model's knot values for locations 1 and 4 (as well as locations 2 and 3) appear to be close together, there may be some similarities in traffic patterns and reported speeds for these locations, in particular because their speed coefficients indicate close values.

3.4 Discussion

In this chapter we have explored the modelling of the relationship between the rate of Bluetooth detections and ATC records. Four study locations where both the ATC counter and Bluetooth detectors are deployed nearby were chosen for the analysis. The visualisation suggested a non-linear relationship pattern with two stages between Bluetooth and ATC variables. Firstly, it is characterised by low traffic volumes and lower Bluetooth detections, and the second, which starts when traffic volumes reach a certain level and the rate of Bluetooth detections rapidly increases. We employed regression analysis, as a powerful statistical method for modelling the relationship, by taking into account that some factors might influence the rate of Bluetooth detection. We were able to extract the number of buses and the speed of the vehicles passing through the area using the ATC database. The initial idea was that the number of buses was considered as one of the factors that might explain the variation, because the rate would be higher when buses are likely traveling around with more active Bluetooth devices. The second hypothesis was that the speed of the vehicles might also have an effect on the rate. For example, a high-speed car travelling through the region could be undetected. Also, several interfering active devices may decrease data transfer to the detector if there is a traffic jam that forces vehicles to slow down or stop.

Therefore, the multiple linear regression with some other alternative models was first constructed by considering the Bluetooth detections as a dependent variable, ATC records and the number of buses as two independent variables. After running regression analysis, heteroscedasticity produced an unequal scattering of residuals, which needed

to be addressed before making any inferences from the model. The rolling variance approach was used to specify an appropriate parametric model for the variance function based on data, but it failed, and thus a non-parametric variance function estimation method was successfully used.

The results of regression analysis showed that buses had an effect on the rate of Bluetooth detections, which as we expected, increases as more buses pass through the detection region. However, including the buses in the model should have resulted in the linear model being selected as the best model, which did not happen, showing that the buses were not the complete explanation of the rate variation. The segmented regression model with three knots was considered to be the best model for locations 1, 2, and 3, whereas the weigghted segmented regression model with two knots was prefered for location 4. Here, there was a slight statistically (but not practically) significant difference between the segmented regression model with two and three knots. Therefore, for all locations, the segmented regression model with two knots could be considered as the best final model.

The results of regression analysis that included the interaction speed term demonstrated that as the vehicle speed increases, the rate of Bluetooth detection reduces. That is presumably because of free-flowing traffic, where the car can exit the detection area faster and the detector has a lower chance of detecting it.

Finally, since adding the buses and speed still does not explain all the variations in Bluetooth rates, there must be other factors involved, which are unknown and may not be observable with the data available to us. The only other factor that can be considered based on the data we have is the effect of different times of the day. As a result, in the next chapter, we will investigate how different times of day affect Bluetooth rate variation. Other factors, such as the number of buses and speed, also change during different times of the day. Therefore, even if we do not know what all the influential factors are, we will be able to incorporate them into constructing the relationship between the rate of Bluetooth detection and the number of vehicles passing through the area.

Chapter 4

Calibration based on time-varying coefficients Poisson regression

4.1 Introduction

Consider an area that is being monitored with both a Bluetooth detector and an ATC. When a car with Bluetooth-enabled device(s) passes through this area, ATC will record the vehicle, and Bluetooth items will be recognised by a Bluetooth sensor. As noted in Section 1.5 depending on the detector's location, road traffic condition and the length of time the vehicle spends in the detection area, a Bluetooth device on a vehicle can be recorded multiple times or may not be captured. Also, a vehicle may be carrying more than one active Bluetooth device, however, some devices show a tendency more than others not to being discoverable. For example, smart phones have a lower capture rate due to the fact that the devices must be in discoverable mode in order for the detector to record them (Bhaskar et al., 2013). Therefore, having multiple active Bluetooth devices in one individual car does not necessarily mean that all of them would be detected by the sensor. Another possible source of recorded data is Bluetooth-enabled devices carried by pedestrians and cyclists in the area, although they are expected to form a relatively low percentage of the data set (Araghi et al., 2012).

The purpose of this chapter is to model the rate of unique Bluetooth detections per vehicle, so the research scheme is associated with two forms of data, Bluetooth and ATC, as in the previous chapter. Due to traffic conditions, the rate of Bluetooth per vehicle may vary. For instance, in free-flowing traffic, the detector may not be able to detect all passing Bluetooth-enabled vehicles since they are moving through the region too fast. Hence, it is anticipated that there will be fewer Bluetooth counts per vehicle. As

traffic congestion increases, so will the number of Bluetooth counts per vehicle, and thus the rate of Bluetooth detections, due to the longer time spent in the detection area. Conversely, traffic congestion may cause signal interference, resulting in the device's failure to be recognised (Vo, 2011).

The number of unique Bluetooth detections and ATC record data in sequential five-minute time intervals for one year and for a specified week are shown in Figures 4.1 and 4.2, respectively. It displays a daily pattern for each Bluetooth and ATC data set that is implied by traffic volume change over time. As a result, developing a model that can characterise the rate of Bluetooth detection per car as it fluctuates over time is critical. There may be some random factors that affect the rate of detecting Bluetooth devices, such as weather, detector position, and so on, and it would be impossible to account for them all in the model fitting. Hence, our model will only account for the part of the Bluetooth rate that has constant hourly and daily trends throughout the year. In this chapter same as Chapter 3, we will consider both Bluetooth and ATC records as count data and explore several candidate models for the research goal.

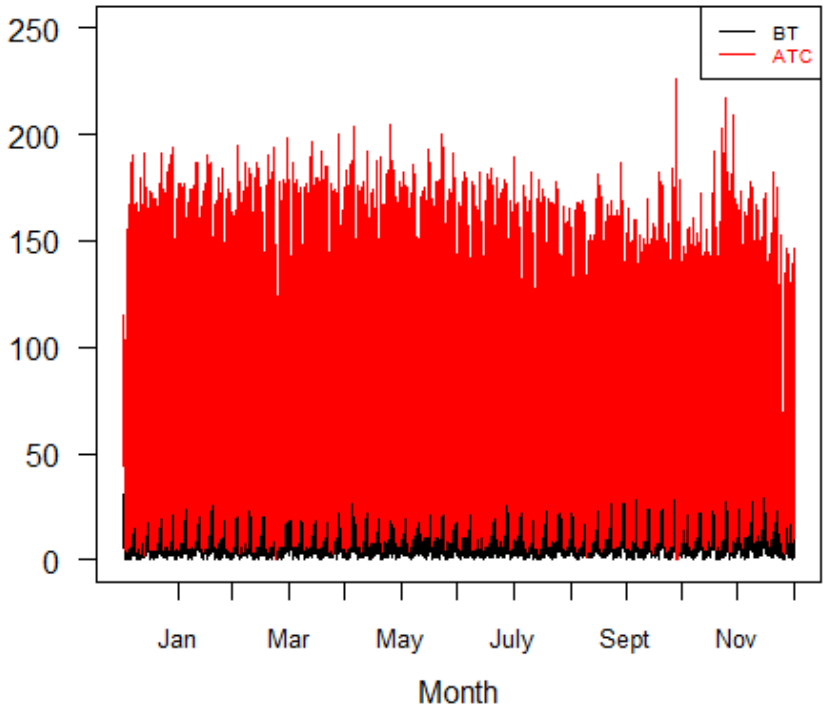


Figure 4.1: Bluetooth and ATC record data in five minutes time interval for one year, 2018.

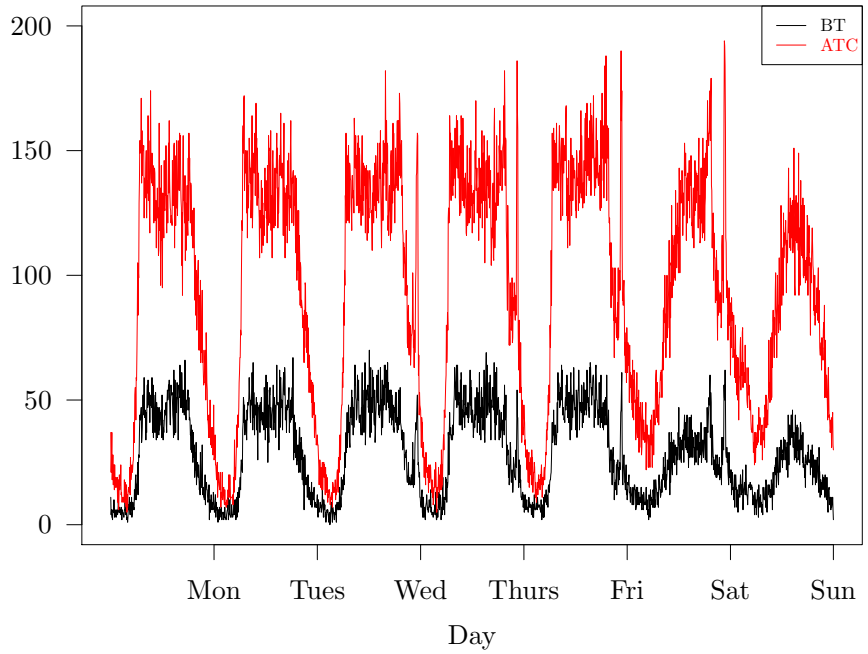


Figure 4.2: Bluetooth and ATC record data in five minutes time interval for one week, 22-28 January, 2018.

A further practical purpose will be to use the model to recover ATC from the count of Bluetooth devices. The fundamental advantage of ATC is that it can work at any time of day or night, capturing the exact time, instantaneous speed, and model of each vehicle passing through the area. Comparing to traditional data collection methods, the Bluetooth detector is known for its low installation and maintenance costs. It also utilizes less power and has high privacy protection when detecting and saving data. However, the Bluetooth counts are very different from the ATC counts. The statistical calibration method will be utilised to predict ATC from the Bluetooth detections after choosing the suitable model to fit the data.

Section 4.2 starts by describing the modelling possibilities, including the counting process, discrete time series Poisson regression, and Poisson regression with smoothly time-varying coefficients using the Fourier series or the periodic B-spline. Section 4.3 describes two different model selection techniques, quasi-likelihood Baysian information criterion and cross-validation. In Section 4.4, the implementation and results of the fitted Poisson regression model are provided. Section 4.5 presents the statistical calibration method using two approches, the classical estimator and the profile log-likelihood approach. Finally, the implementation and results of the calibration are presented in

Section 4.6 and discussion for this chapter is provided in Section 4.7.

4.2 Modelling possibilities

4.2.1 Counting process

Data from both Bluetooth and ATC is collected as count data over time. Therefore, the *counting process* can be used to model them. The counting process is a stochastic process $N(t)$ that displays the total number of events $N(t)$ that occur in the time interval $[0, t]$ under the following assumptions:

- $N(0) = 0$,
- $N(t) \in \{0, 1, 2, \ldots\}$ for all $t \in [0, \infty)$,
- For $0 \leq s < t$, $N(t) - N(s)$ depicts the number of events that happen in the interval $(s, t]$.

The counting process $\{N(t), t \in [0, \infty)\}$, is called a (*homogenous*) *Poisson process*, if $N(t)$ has a constant rate $\lambda > 0$ and **independent Poisson distributed increments**, which means the number of arrivals in non-overlapping intervals is independent, and the number of arrivals in any interval of length $\tau > 0$ has Poisson distribution, $N(t + \tau) - N(t) \sim \text{Pois}(\lambda\tau)$ with the probability mass function as:

$$p(N(t) = x|\lambda\tau) = \frac{e^{-\lambda\tau}(\lambda\tau)^x}{x!}, \qquad x = 0, 1, \ldots \tag{4.1}$$

where x is the possible value from the sample space for the random variable $N(t)$.

Alternatively, a *non-homogeneous Poisson process* has all the characteristics of a Poisson process, except that its rate is a function of time, i.e. $\lambda = \lambda(t)$ and so it does not have stationary increments. Hence, the increments of a non-homogeneous Poisson process are independent but not necessarily stationary. For instance, if $N(t)$ be the number of vehicles arriving at a public car parking by the time t, then the arrival rate of vehicles is larger during working hours compared to off-hours. More specifically, we can write:

$$N(t + \Delta t) - N(t) \sim \text{Pois}\left(\int_t^{t+\Delta t} \lambda(s)ds\right). \tag{4.2}$$

We can assume that ATC and Bluetooth data are generated by two counting processes $N_A(\tau)$ and $N_B(\tau)$, which count the total number of ATC and Bluetooth counts up to the time $\tau > 0$, respectively.

The observations of $N_A(\tau)$ and $N_B(\tau)$ will be recorded in discrete time points, so the realization up to the time $\tau > 0$ for $N_A(\tau)$ can be written as $\{\tau_1, \tau_2, \ldots, \tau_{N_A(\tau)}\}$ which denotes the actual times of events.

As noted before, the main goal is to model the rate of Bluetooth detections per vehicle at different time points by considering the following facts:

- Every vehicle carries a number of active Bluetooth devices, say $\{0, 1, 2, \dots\}$ so not every vehicle has active Bluetooth devices,
- Due to the different characteristics of Bluetooth devices, they may be recorded by the different detection probabilities. For example, some smartphones remain in the discoverable mode only for a limited time, unless the discovery time mode is changed by the user. Therefore, they will have a relatively low probability of being detected by the sensor. In practice, an installed detector may not be able to record all the discoverable Bluetooth devices in the vehicles, so only a certain portion of them are detected,
- There is a possibility that we have non-vehicle Bluetooth detections in the data, however, it is expected to be a small number.
- The number of Bluetooth counts can vary due to traffic conditions. In the free-flow traffic conditions, the detector may fail to capture all of the passing Bluetooth-enabled vehicles as they are passing through the area fast enough before being detected. It is expected to have fewer Bluetooth counts, and therefore, the rate of the Bluetooth detections per vehicle will decrease. Whereas when traffic congestion increases, the number of Bluetooth counts also increases and will raise the rate of the Bluetooth detections due to the longer staying time in the detection area (Michau et al., 2014).

Considering the above facts, suppose the ith vehicle passing through the detection area at time τ_i carrying a number of active Bluetooth devices that follows a Poisson distributed random variable $Z_i \in \{0, 1, 2, \dots\}.$, Each active Bluetooth device has a probability of p being detected. We can construct another random variable X_i that depends on Z_i as:

$$X_i = \sum_{j=1}^{Z_i} B_j \tag{4.3}$$

where B_j are the identical and independently distributed Bernoulli random variables, which represent the detection event, i.e. $B_j = 1$ if the jth Bluetooth device is detected and $B_j = 0$ otherwise.

We were unable to directly model the actual arrival process of Bluetooth devices and the detection event, i.e. Z_i and B_j. We can utilize X_i to explore their joint effects that can be regarded as the number of Bluetooth devices being detected. The equation (4.3) is like thinning operators in Poisson autoregression (Al-Osh and Alzaid, 1987), therefore, X_i is also a Poisson random variable.

If $N_A(\tau)$ follows a non-homogeneous Poisson process, then the Bluetooth count process $N_B(\tau)$ also can be modelled like a non-homogeneous Poisson processes as follows:

$$N_B(\tau) = \int_0^\tau X(s) dN_A(s) = \sum_{i=1}^{N_A(\tau)} X_i \sim \text{Pois}\left(\sum_{i=1}^{n} \lambda(\tau_i)\right) \quad \text{if } N_A(\tau) = n \tag{4.4}$$

where $X(s)$ can be regarded as the number of Bluetooth devices being detected and for the ith ATC count at time τ_i, we assume that it generates $X_i \sim \text{Pois}(\lambda(\tau_i))$ Bluetooth counts. Note that X_i are independent Poisson random variables, then $\sum_i X_i$ is Poisson random variable conditional upon $N_A(\tau)$. The intensity $\lambda(\tau_i)$ characterises the average count of Bluetooth devices detected from a vehicle at time τ_i.

We can also consider another counting process to account for the non-vehicle Bluetooth detections. Therefore, we can further add a noise Poisson process $\epsilon(\tau)$ with arrival rate ρ as follows:

$$N_B(\tau) = \sum_{i=1}^{N_A(\tau)} X_i + \epsilon(\tau) \sim \text{Pois}\left(\rho\tau + \sum_{i=1}^{n} \lambda(\tau_i)\right) \quad \text{if } N_A(\tau) = n \tag{4.5}$$

It can thus reasonably expected that ρ is small due to the low probablility of the non-vehicle Bluetooth detections and we will consider it as a constant in the modeling procedure.

Modeling in a continuous time scale will be difficult due to the large dataset and the computational complexity. Therefore, the continuous-time counting process will not be considered directly in this study. As a discrete time approximation, a model based on time series regression will be examined further in the following section.

4.2.2 Time series regression of counts

Following the property of the Poisson process (4.5), we have the conditional expectation of N_B given N_A over a time interval $(t, t + \Delta t]$ as follows:

$$\mathbb{E}[N_B(t, t+\Delta t] | N_A(t, t+\Delta t]] = \rho\Delta t + \sum_{j=1}^{N_A(t,t+\Delta t]} \lambda(t_j) \approx \rho\Delta t + \lambda_t N_A(t, t+\Delta t] \tag{4.6}$$

where $\lambda_t = \int_t^{t+\Delta t} \lambda(s) ds$. The unit time increments Δt will equal the chosen resolution, for example, if we discretise the data into five-minute slots, $\Delta t = 5$ minutes.

Considering the number of unique Bluetooth detection and ATC for specified time slots Δt (e.g. five minutes interval), Bluetooth and ATC data will be aggregated into two different time series, $\boldsymbol{y} = \{y_t\}$ and $\boldsymbol{x} = \{x_t\}$, $t \in \{1, 2, \ldots, T\}$, where T is the number of Δt over a time period.

We can consider the Poisson time series regression with the identity link function

with the effect of time on the regression coefficient as follows:

$$y_t|x_t = N_B(t, t+\Delta t] \sim \text{Pois}(\mu_t) \tag{4.7}$$

$$\mu_t = \mathbb{E}[y_t|x_t] = \alpha + \beta(t)x_t \tag{4.8}$$

where refers to the equation (4.6) $x_t = N_A(t, t+\Delta t]$, $\alpha = \rho\Delta t$, $\beta(t) = \lambda_t$ and μ_t is a function of α and $\beta(t)$. The likelihood function of the Poisson regression is given by:

$$L(\alpha, \beta(t)|\boldsymbol{y}, \boldsymbol{x}) = \prod_{t=1}^{T} \frac{e^{-\mu_t}(\mu_t)^{y_t}}{y_t!}. \tag{4.9}$$

The log-likelihood function is:

$$l(\alpha, \beta(t)|\boldsymbol{y}, \boldsymbol{x}) = \sum_{t=1}^{T} \left(y_t \log(\mu_t) - \mu_t - \log(y_t!)\right). \tag{4.10}$$

This model considers the effect of time on the regression coefficient, i.e. the relationship between Bluetooth and ATC. Bluetooth counts are expected to be higher during congested time intervals of the day than in free-flow traffic and to be lower on weekends than on weekdays due to changes in traffic conditions. As previously discussed, there should be daily and weekly patterns or cycles in the rate of Bluetooth detections per vehicle which can present a profile related to real-life traffic conditions. Formulating $\beta(t)$ in a suitable way can help us to explore the temporal structure in our data. In this study, two different time resolutions of one hour and five minutes will be considered. Because of the small values of α, i.e. the number of non-vehicle Bluetooth detections, we did not consider the time and day effects on α.

The first formulation of $\beta(t)$ will be regarded as a **Poisson regression model with stepwisely time-varying coefficients** and can be rewritten as follows. This model is based on the assumption that there is no hourly and daily interaction effects, i.e. the hourly pattern is the same every day.

$$\mu_t = \alpha + \left[\beta_0 + \sum_{i=2}^{24} \Delta\beta_i^{(h)} \cdot I(t \text{ in the } i\text{th hour}) + \sum_{j=2}^{7} \Delta\beta_j^{(d)} \cdot I(t \text{ in the } j\text{th weekday})\right] x_t \tag{4.11}$$

By considering the one-hour time slot, β_0 represents the baseline rate during 00:00-01:00 AM at Monday, and $\Delta\beta_i^{(h)}$ and $\Delta\beta_j^{(d)}$ adjust the baseline according to the actual time by using the indicator function I(condition) which also convert the model into the discontinuous version. For example, for the time interval t between 02:00-03:00 AM

(i.e. hour 3) at Thursday (i.e. day 4), we will have the following term:

$$\mu_t = \alpha + [\beta_0 + \Delta\beta_3^{(h)} + \Delta\beta_4^{(d)}]x_t \tag{4.12}$$

This model is easy to fit but the assumption that the rate of Bluetooth per vehicle would change suddenly every hour and stay constant within each hour is an unrealistic scenario. The traffic conditions can change dramatically within an hour, especially during the day. As a result, smaller time intervals, such as five minutes, are likely to be a more reasonable interval size for capturing variations throughout the daytime. Choosing the five-minute time slot will increase the complexity of the fitted model as the different parameters should be assigned for every time interval. It would be more realistic to assume that the rate of Bluetooth detection would change smoothly over the different times of the day. Therefore, modifying the step function $\beta(t)$ into a smooth function of time leads to having a less parameterized model. In the next subsections, we will present the Poisson regression model with smoothly time-varying coefficients.

4.2.3 Poisson regression with smoothly time-varying coefficients

Poisson Fourier time series regression

The Poisson time-series regression method to analyse the data can be defined as follows:

$$\begin{aligned} &y_t \sim \text{Pois}(\mu_t) \\ &\mu_t = \alpha + \beta(t)x_t, \quad t \in \{1, 2, \ldots, S\} \end{aligned} \tag{4.13}$$

where $\beta(t)$ is a smooth periodic function to characterise the seasonal effects and S is the seasonal period of this function. One of the most common choices to include the periodic covariates is to utilise Fourier series. The Fourier series expansion theorem indicates that any periodic function can be represented using a summation of sine and cosine functions of various frequencies and amplitudes (Davis and Sampson, 1986). We can use the finite Fourier series approximation to extract the weekly pattern of $\beta(t)$ in (4.13), therefore, the smooth function $\beta(t)$ can be modelled by the summation of finite pairs of sine and cosine functions as follows:

$$\beta(t) = \sum_{k=0}^{m} \left[a_k \cos\left(\frac{2\pi kt}{S}\right) + b_k \sin\left(\frac{2\pi kt}{S}\right) \right] \tag{4.14}$$

where a_k and b_k are the Fourier coefficients, t represents a particular time, S is the seasonal period time and m is the number of the harmonic pairs included in the model that consists of $2m$ number of sine and cosine functions. The constant term a_0 is the average value of the function $\beta(t)$ and will be denoted β_0, i.e. the average rate of the

Bluetooth detections per vehicle. Therefore, $\beta(t)$ also can be written as:

$$\beta(t) = \beta_0 + \sum_{k=1}^{m} \left[a_k \cos\left(\frac{2\pi kt}{S}\right) + b_k \sin\left(\frac{2\pi kt}{S}\right) \right] \tag{4.15}$$

Note if m tends to infinity, the finite Fourier series will converge to $\beta(t)$, However, it cannot be reached in reality. It is necessary to decide a suitable m to approximate the underlying smooth function $\beta(t)$. The more harmonics are employed, the more precise can a function be described. However, we need the optimal number of harmonic functions with a reasonable complexity to capture the actual pattern of the coefficients with a moderate computational burden. Therefore, m will be determined based on certain model selection criteria. Two different model selection criterion, the quasi-likelihood Bayesian information criterion (QBIC) and the cross-validation method will be discussed to estimate the optimal number of the harmonic terms in Section 4.3.

The Poisson regression model with the Fourier coefficients can be characterised as follows:

$$\mu_t = \alpha + \left(\beta_0 + \sum_{k=1}^{m} \left[a_k \cos\left(\frac{2\pi kt}{S}\right) + b_k \sin\left(\frac{2\pi kt}{S}\right) \right] \right) x_t \tag{4.16}$$

The functional-coefficient $\beta(t)$ will tell us how much variability we could expect for the rate of the Bluetooth detections during each time interval of different days.

Figure 4.3 shows the four Fourier basis functions by considering $m = 2$. The highlighted blue and green harmonic functions are corresponding to the cosine and the sine term for $k = 1$ and $k = 2$, respectively. As can be seen, the Fourier basis has global support over the whole domain, which means that any modifications such as changes in the fitted value at the particular time t will affect all the Fourier coefficients in the scope.

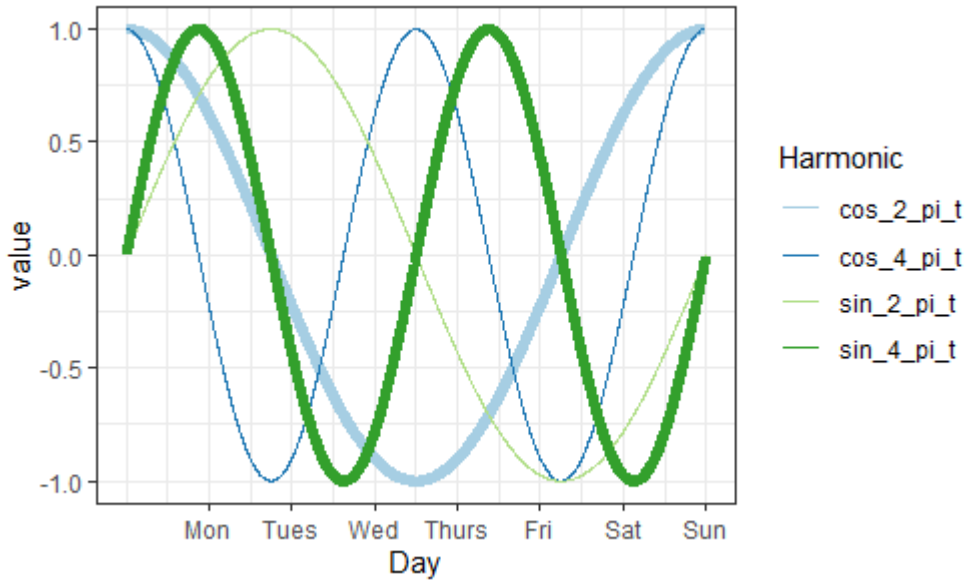


Figure 4.3: The four Fourier harmonic basis functions by considering $m = 2$. The highlighted blue and green harmonic functions are corresponding to the cosine and the sine term for $k = 1$ and $k = 2$, respectively.

Therefore, a Fourier series will be useful in situations such as a function without strong local features. The Fourier series basis would not be the only option in this case as a periodic function and another flexible choice of the periodic covariates that can be considered is a periodic B-spline that we will represent in the next subsection.

Poisson periodic B-spline time series regression

B-Splines (basis-splines) are one of the most popular methods for approximating a function by defining a linear combination of piecewise polynomials, called basis functions (Ramsay, 2004; Perperoglou et al., 2019; Lusa and Ahlin, 2020). A B-spline of order d is a parametric smooth curve constructed from a piecewise polynomial of basis function $\tilde{B}_{i,k-1}(t)$ of degree $k = d - 1$. A sequence of *knots* (usually equally spaced) are needed in order to subdivide the domain on the B-spline curve into a set of knot spans as $[t_i, t_{i+1})$. A periodic B-spline is a B-spline with the property that the first domain knot and the final domain knot produce a closed loop.

Recursive definitions of B-spline functions can be presented in the way that the basis function of degree $k = 1$ have values of unity in a given interval, and zero otherwise (De Boor, 1978). The i-th B-spline basis functions of degree $k = 1$ for the i-th interval $[t_i, t_{i+1})$ defined as:

$$\tilde{B}_{i,1}(t) = \begin{cases} 1 & \text{if} \quad t_i \leq t \leq t_{i+1} \\ 0 & \text{otherwise.} \end{cases} \tag{4.17}$$

and the higher degree of basis functions, $\tilde{B}_{i,k}(t)$ for $k > 1$, will be determined as:

$$\tilde{B}_{i,k}(t) = \frac{t - t_i}{t_{i+k} - t_i}\tilde{B}_{i,k-1}(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}}\tilde{B}_{i+1,k-1}(t) \quad k = 2, \ldots, d. \tag{4.18}$$

The cubic periodic B-splines (i.e. $d = 3$) will be used as the most common choice in this study. The Poisson regression model with the cubic B-spline basis functions as the functional coefficients and also the identity link function will be considered as follows:

$$\begin{aligned} \mu_t &= \alpha + \beta(t)x_t \\ &= \alpha + \beta_0 x_t + \sum_{i=1}^{N} \beta_i B_{i,3}(t)x_t \end{aligned} \tag{4.19}$$

where β_0 is the average rate of the Bluetooth detections per vehicle, as the same with the Fourier model, β_i is the periodic B-spline coefficients, t represents a particular time, and N is the number of periodic B-spline basis included in the model. Also, because of the way they are constructed, periodic B-spline basis functions have local support and each of them will take care of its region and that is the main difference between the Fourier series and the periodic B-spline. The periodic B-spline basis with $N = 4$ knots for a period length of one week is depicted in Figure 4.4. The highlighted blue and green basis functions correspond to the first and last columns of the periodic B-spline basis functions, respectively. As can be seen in Figure 4.4, each periodic B-spline basis has local support and it is nonzero at a certain interval and zero elsewhere, which shows it only affects values within a limited range. It demonstrates that if the coefficients for the specific periodic B-spline basis change, it only locally affects its limited non-zero support domain, unlike the Fourier basis functions that influence the whole domain. The green basis function is the sum of two corresponding basis functions from the ordinary B-spline basis. It also explains periodicity, which occurs when the value of each periodic B-spline basis at time $t = 0$ is the same as the value at time $t = S$, i.e. the periodic B-spline curve matches at the same starting and ending points.

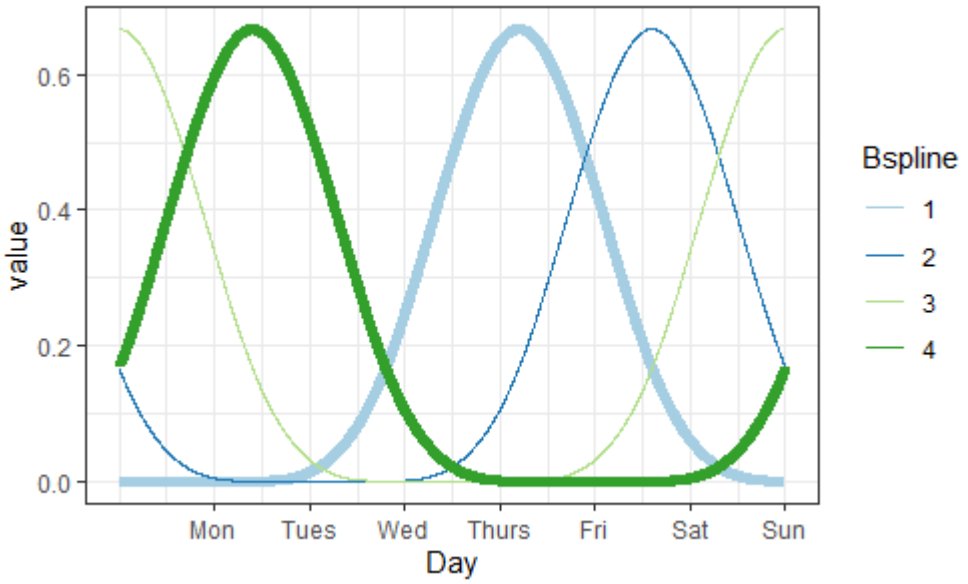


Figure 4.4: The periodic B-spline basis with $N = 4$ knots for a period length of one week. The highlighted blue and green basis functions correspond to the first and last columns of the periodic B-spline basis functions, respectively.

4.3 Model selection

Model selection is about choosing the optimal model that must be a good approximation of the data, but as simple as possible, from a given set of models (Burnham and Anderson, 2002). To incorporate the Fourier series or periodic B-spline functions in the Poisson regression in practice, we need to decide the number of sine and cosine terms or the number of knots, respectively. In order to balance the smoothness and parsimony of fitted models, appropriate model selection techniques are needed to determine the optimal model complexity. This subsection will present two different model selection techniques: one based on the quasi-likelihood Bayesian information criterion and the other based on cross validation.

4.3.1 Quasi-likelihood Bayesian information criterion

We have utilised AIC and BIC as the two most well-known methods for the comparison of the goodness of fit of the different regression models in the previous chapter (see Section 3.2.5). One common problem with the Poisson regression is the *over-dispersion* that happens when the sampling variance is larger than the theoretical variance based on the distribution of the fitted model (Student, 1919; Cox, 1983; Fisher, 1950; Dean, 1992). Therefore, a quasi-likelihood version of the BIC (QBIC) that takes the overdispersion into account is used to choose the optimal number of Fourier terms and the

knots for B-splines. The QBIC statistic is defined as follows:

$$\text{QBIC} = \frac{-2\log(\hat{L})}{\hat{D}} + K\log(n) \tag{4.20}$$

where $\hat{D}$ is the estimated dispersion parameter, K is the number of estimated model parameters and $\log(\hat{L})$ is the log likelihood value of the fitted model and n is the sample size (Pinheiro and Bates, 2000). In model selection, $\hat{D}$ should be estimated as the global overdispersion parameter to use for comparing different models. In fact, the same value of $\hat{D}$ obtained from the global model of the nested models (i.e, the most complex model) will be used in (4.20) to compute the QBICs. Otherwise, if we use different $\hat{D}$ for each model, it will always return the smallest candidate model as the best choice.

The ideal global model, for example, in the Poisson Fourier regression model, would be one with an infinite m (i.e. number of Fourier terms). Therefore, we consider an approximation by choosing a large enough m and running a few Poisson Fourier regression models to see how the overdispersion parameter changes as the model complexity increases. If the dispersion parameter settles down to a stable value, we will consider it as the estimate of the global over-dispersion parameter and utilise it to compute QBIC for a sequence of the Poisson Fourier regression models (Lebreton et al., 1992). Finally, the preferred number of the Fourier basis functions will be chosen as the model with the minimum QBIC value. The same procedure will be applied to select the optimal number of knots for the periodic B-splines.

4.3.2 Cross-validation

Cross-validation is a technique for evaluating model performance where it approximates the mean prediction error to quantify the prediction accuracy of the model. The original sample is divided into two separate data sets in this method, a training and a test set, to train and evaluate the fitted model, respectively.

The cross-validation for the time-series model is performed on a rolling basis that starts with a small subset of data as test data, fits the model based on the training data and finally checks the prediction error for the test data with the fitted model. Note that the test data should be considered based on the assumed periodicity to avoid breaking the internal structure of the time series. For instance, since a weekly periodicity has been built in our model, the procedure is performed by considering one week of test data. The method for the Poisson Fourier regression models is as follows and the same procedure will be applied to the periodic B-spline.

- We use one week of data as test data and Poisson Fourier regression fits on the training data, which is the rest of the data, for a given m (i.e. the number of

Fourier terms). This should be repeated for the total number of weeks in a year, say $w = 52$, for yearly data.

- The prediction error can be calculated by the logarithmic score as follows:

$$s(y, \hat{y}) = -\log(p(y|\hat{y})) \tag{4.21}$$

where y is the real single observation, $\hat{y}$ is the prediction value of the fitted model using the training data and $p(y|\hat{y})$ is the Poisson probability mass function for the observed outcome y given the estimated mean $\hat{y}$ (Good, 1992; Gneiting and Raftery, 2007). The weekly periodicity is used to define the test data in our model. Therefore, the mean logarithmic score would be computed as follows:

$$s_w(\boldsymbol{Y}^{(w)}, \hat{\boldsymbol{Y}}^{(-\boldsymbol{w})}) = \frac{1}{S} \sum_{j=(w-1)S+1}^{wS} s(y_j, \hat{y}_j^{(-w)}) \tag{4.22}$$

where S represents the seasonal period, $\boldsymbol{Y}^{(w)}$ is the vector of real observations of the week that is considered as the test data, $\hat{\boldsymbol{Y}}^{(-\boldsymbol{w})}$ is the vector of the predicted values using the remaining weeks (i.e. training data). The total number of weeks in a year (i.e. $w = 1, \ldots, 52$) will be used to calculate the average score. The final cross-validation error $s_{cv}^{(m)}$ for the number of m Fourier terms is considered as the average over all logarithmic scores.

- The optimal number of m will be chosen by the lowest cross-validation error $s_{cv}^{(m)}$.

4.4 Implementation and results

The methodologies discussed in the previous sections were employed to model Bluetooth and ATC data collected over one year (2018). Both Bluetooth and ATC data were considered as the time series recorded in the one-hour and the five-minute time slots. The implementation and the results of the Poisson regression model with stepwise time-varying coefficients and the Poisson regression with smoothly time-varying coefficients with the Fourier and B-spline basis are presented in the following subsections.

4.4.1 Fitting Poisson regression model with stepwise time-varying coefficients

As the first model, we have been considered Poisson regression model with stepwise time-varying coefficients where $\beta(t)$ have changed according to the time t and defined

as follows:

$$\begin{aligned}\mu_t &= \alpha + [\beta_0 + \beta^{(h)}(t) + \beta^{(d)}(t)]x_t \\ &= \alpha + \beta_0 x_t + \sum_{i=2}^{24} \Delta\beta_i^{(h)} \cdot [I(t \text{ in the } i\text{th hour}) \cdot x_t] \\ &\quad + \sum_{j=2}^{7} \Delta\beta_j^{(w)} \cdot [I(t \text{ in the } j\text{th weekday}) \cdot x_t] \,.\end{aligned} \tag{4.23}$$

where the predictors with interaction terms are generated as $I(t \text{ in the } i\text{th hour}) \cdot x_t$ and $I(t \text{ in the } j\text{th weekday}) \cdot x_t$. In R, the `glm()` function is used to fit the Poisson regression with the identity link function. In addition, we found that if the intercept includes time and day effects, the fitted model can explain the majority of the rate variability with $\alpha(t)$. As a result, the model is developed on the consideration of $\beta(t)$ changing over time, rather than the intercept, which is influenced by non-vehicle detection.

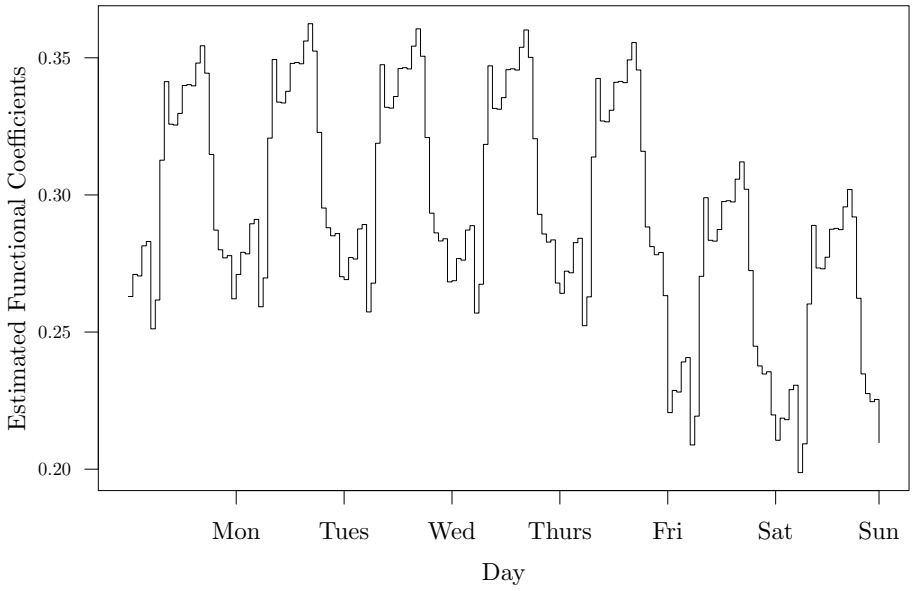


Figure 4.5: The estimated functional coeffcients from the stepwisely time-vary coefficient Poisson regression model.

The estimated functional coefficients from the stepwise time-varying coefficient Poisson regression model over one week are shown in Figure 4.5. The Bluetooth detection rate is varies between 0.2 and 0.36, and the results show a similar hourly pattern over the weekdays and the weekends due to the model assumption of no interaction, but

with lower average rate values over the weekends. It also depicts the assumption that the rate is constant within each hour but changes between the hours of the day. Furthermore, because $\beta(t)$ is defined as a discountinous periodic function, therefore, the beginning and end are not connected. The estimated functional coefficients on Tues-

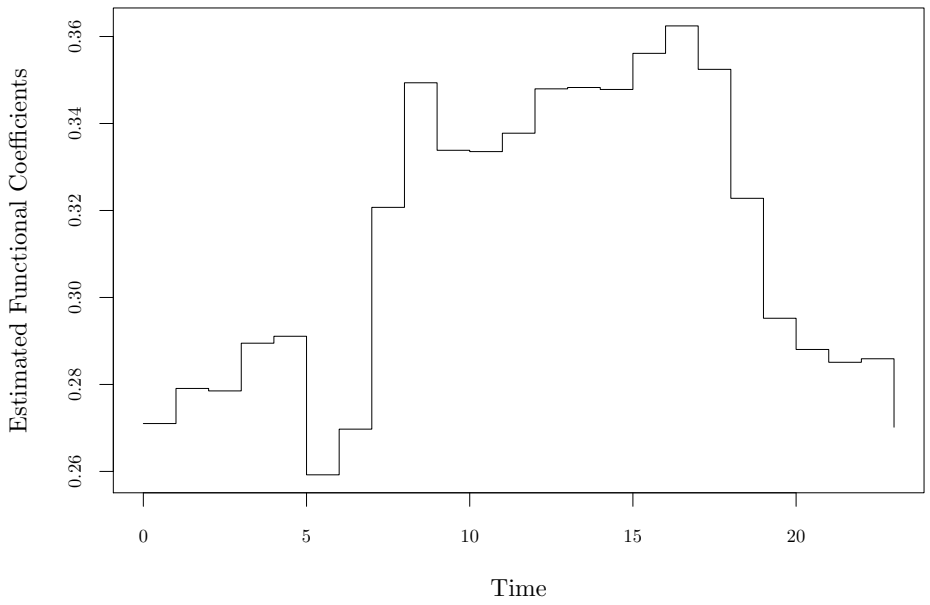


Figure 4.6: The estimated functional coeffcients from the stepwisely time-vary coefficient Poisson regression model for Tuesday.

day are shown in Figure 4.6. It can be seen that the rate increases at 6:00 a.m., with the maximum rate for the morning peak between 8:00-9:00 a.m., then drops, and then begins to increase again at midday 12:00, with the evening peak between 3:00-5:00 p.m.

4.4.2 Fitting the Poisson regression model with Fourier basis

The first choice was using the Fourier series to estimate $\beta(t)$ as smoothly time-varying coefficients. The harmonic functions generated the corresponding cosine/sine values on different time points with the specific frequency for the Bluetooth time-series data. We have discretized the data in five minutes and expected to see the daily and weekly seasonalities as shown in Figure 4.2. Therefore, the frequency is the total number of five minutes slots during one week, say 2016 ($= \frac{60}{5} \times 24 \times 7$). The harmonic functions were generated by the function 'harmonic' in the TSA package in R. The optimal number of sines and cosines terms are unknown and will be determined based on the model

selection criterion. The model with the harmonic basis functions will be considered as:

$$\begin{aligned}\mu_t &= \alpha + \beta(t)x_t \\ &= \alpha + \beta_0 x_t + \sum_{k=1}^{m}\left[a_k \cos\left(\frac{2\pi kt}{T}\right) + b_k \sin\left(\frac{2\pi kt}{T}\right)\right] x_t \\ &= \alpha + \beta_0 x_t + \sum_{k=1}^{m} a_k \left[\cos\left(\frac{2\pi kt}{T}\right) x_t\right] + \sum_{k=1}^{m} b_k \left[\sin\left(\frac{2\pi kt}{T}\right) x_t\right].\end{aligned} \tag{4.24}$$

After constructing the harmonic functions, the next step is to fit the Poisson regression model where the interaction terms of the ATC counts and the Fourier basis variables are the explanatory variables (Algorithm 1). The model is fitted using `glm()`, which includes Poisson regression and the identity link function. It should be noted that to prevent the initialisation problems due to the Poisson identity link function, we ran the linear regression model to obtain the initial values for fitting in `glm()`. Using random initial values would not be useful due to a large number of parameters. The final step is to estimate the optimal number of harmonic functions with the model selection methods as discussed in Section 4.3(Algorithm 2). Figure 4.7 shows the computed over-dispersion parameter $\hat{D}$ for a range number of Fourier terms. It can be seen as the number of Fourier terms m increase, $\hat{D}$ converges to a specified value that will use in the model selection part.

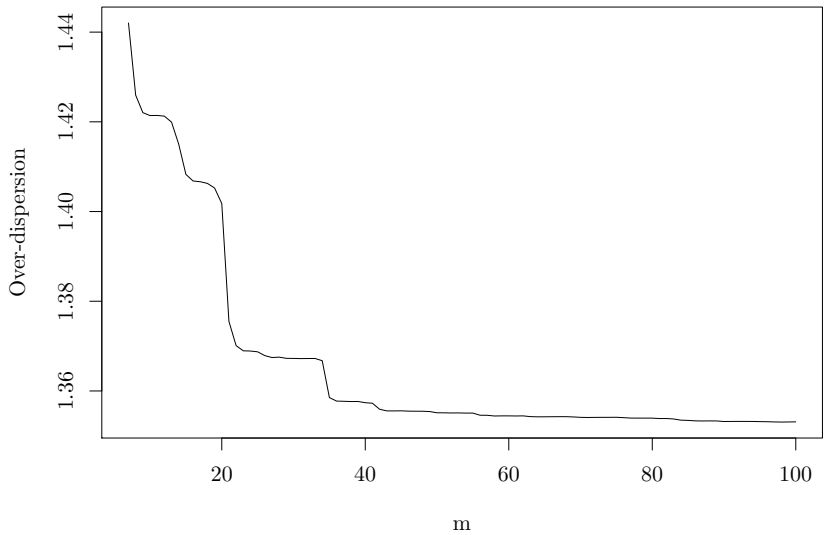


Figure 4.7: The computed over-dispersion parameter $\hat{D}$ for a range number of Fourier terms (i.e. m is the number of Fourier terms).

Algorithm 1 *Poisson regression model with Fourier basis*

Input: $\mathcal{F} = \{(x_t, y_t)|t = 1, \ldots, T\}$, m

1: Generate the m number of harmonic functions basis, $\psi_1, \ldots, \psi_m$, $\phi_1, \ldots, \phi_m$
2: Initialization: Fit the linear regression model with Fourier basis
3: $y_t = \alpha + \beta_0 x_t + \sum_{k=1}^{m} \left[a_k \psi_k + b_k \phi_k\right] x_t$
4: $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}_0, \hat{a}_1, \ldots, \hat{a}_m, \hat{b}_1, \ldots, \hat{b}_m)$
5: Fit `glm()` with identity link function and the initial parameter vector $\hat{\boldsymbol{\theta}}$

Output: Model summary :

6: $\hat{\boldsymbol{\theta}}^{*} = (\hat{\alpha}^*, \hat{\beta}_0^*, \hat{a}_1^*, \ldots, \hat{a}_m^*, \hat{b}_1^*, \ldots, \hat{b}_m^*)$

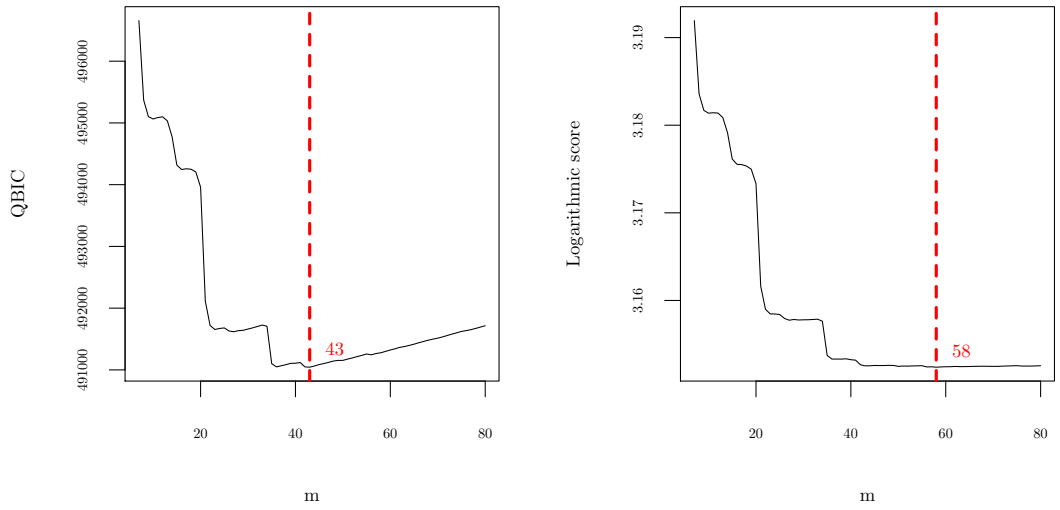
Algorithm 2 *Cross-validation*

Input: $\mathcal{F} = \{(x_t, y_t)|t = 1, \ldots, T\}$, $m^{(min)}$, $m^{(max)}$

1: **for** $i = m^{(min)}$ to $m^{(max)}$ **do**
2: **for** $w = 1$ to 52 **do**
3: Test data: Hold out data from $\mathcal{F}$ for week w
4: Training data: Remainder data
5: Algorithm Poisson regression model with Fourier basis
6: Predict the test data using the fitted `glm()`
7: Calculate the logarithmic score s_w
8: **end for**
9: Calculate the average of the logarithmic scores s_w across all w, $s_{cv}^{(m)}$
10: **end for**
11: Determine the optimal m^* :$\arg\min(s_{cv}^{(m)})$

The results of the QBIC and the cross-validation are represented in Figure 4.8. The step-wise trending in both plots can be interpreted as the number of the basis functions increases (i.e. the frequency increases), the Fourier series can capture the better structure of the periodic function $\beta(t)$ at particular frequencies and it causes a significant drop in the QBIC or the estimated prediction error. Finally, after adding a certain number of the harmonic functions, it will not provide more improvements and more details about the corresponding periodic function. The optimum number of harmonic functions have been evaluated at 43 and 58 based on the QBIC and the cross-validation, respectively. They both show different optimal numbers, however, the QBIC criteria emphasizes that the number of 43 is exactly the best choice as the QBIC begins to increase shortly after it. Whilst based on the cross-validation criteria, the estimated prediction error does not show the significant variations after the number of 43 and remains flat for m greater than 43. Therefore, the number of m equals 43 was chosen to implement the final model. Choosing the number of $m = 43$, the total number of

model parameters is 88, which is the sum of intercept, β_0 and the number of 86 Fourier series coefficients. Figure 4.9 shows the estimated functional coeffcients $\hat{\beta}(t)$ from the Poisson regression with the number of 43 Fourier basis for the one week of 2018.



(a) **(b)**

Figure 4.8: The optimal number of harmonic functions (m) with the model selection methods: (a) QBIC, (b) the cross-validation.

The Bluetooth detection rate is between 0.2 and 0.38 across the whole week for this location. The results showed very similar patterns over the weekdays and the different patterns for the weekends. Figure 4.10 shows the estimated weekly functional coeffcients on Tuesday that is starting to increase and reach the morning peak at 8:15, goes down at 10:00 and finally, 12:10 and 16:00 are the second and third peaks during the daytime, respectively, then starting to decrease after the evening peak. There are slightly different peaks on Tuesday, Wednesday and Thursday compared to the other weekdays. For example, there is an extra second peak for these three days which may appear due to the global properties of the Fourier basis functions and using the periodic B-spline will help to check whether it is a real feature or not. This pattern has been repeated for the other weekdays but has changed over the weekend. A lower rate of Bluetooth detections has been captured on weekend days and also the pattern is different from the weekdays as expected. The first and second main peaks on Saturday are displaying at 12:55 and 18:05, respectively, and roughly the same variability is shown on Sunday with an extra peak at midnight around 01:50.

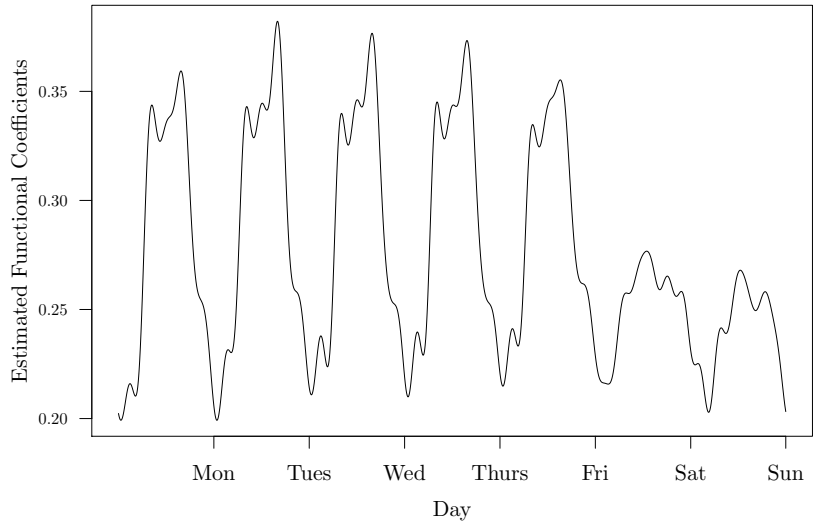


Figure 4.9: The estimated functional coefficients from the Poisson regression with Fourier basis.

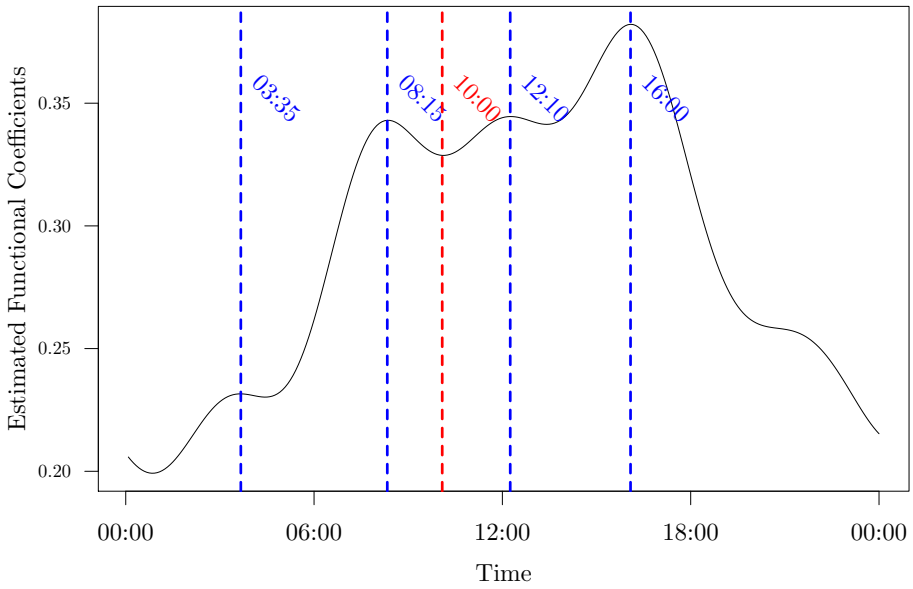


Figure 4.10: The estimated weekly functional coefficients from the Poisson regression with Fourier basis for Tuesday.

Figure 4.11 shows no systematic patterns in the residuals versus the fitted values plot. However, there are a few outliers, especially when the number of ATC records is small. This could be because α is assumed to be constant, or it could be because there

is more non-vehicle detection. In addition, the plot appears to be centred around the horizontal red line at level 0. However, there is a slight tendency to overpredict when the number of ATC records is large.

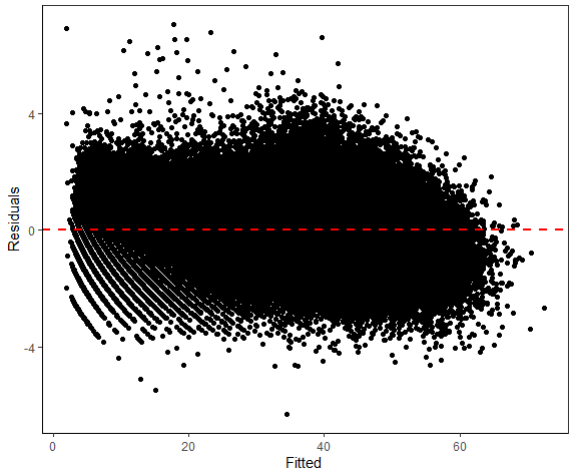


Figure 4.11: The deviance residuals versus the fitted values for the Poisson regression with Fourier basis.

4.4.3 Fitting the Poisson regression model with periodic B-spline

The periodic B-spline was applied as the second method to estimate $\beta(t)$ and the periodic B-spline basis functions were generated by the function `pbs()` in R (Wang, 2013). Again, the optimal number of knots is unknown and will be determined based on the model selection criterion. The Poisson regression model with the periodic B-spline basis functions will be considered as follows, where the interaction terms of the ATC counts and the periodic B-spline basis are the explanatory variables and also the identity function as the link function (Algorithm 3).

$$\begin{aligned} \mu_t &= \alpha + \beta(t) \cdot x_t \\ &= \alpha + \beta_0 x_t + \sum_{i=1}^{N} \beta_i B_{i,3}(t) x_t \end{aligned} \tag{4.25}$$

Figure 4.12 denotes that the periodic B-spline attains the optimal number of knots at 90 based on both criteria, QBIC and cross-validation.

Algorithm 3 *Poisson regression model with the periodic B-spline basis*

Input: $\mathcal{F} = \{(x_t, y_t)|t = 1, \ldots, T\}$, N

1: Generate the N periodic B-Spline basis functions, $B_{1,3}, B_{2,3}, \ldots, B_{N,3}$

2: Initialization: Fit the linear regression model with the periodic B-spline basis

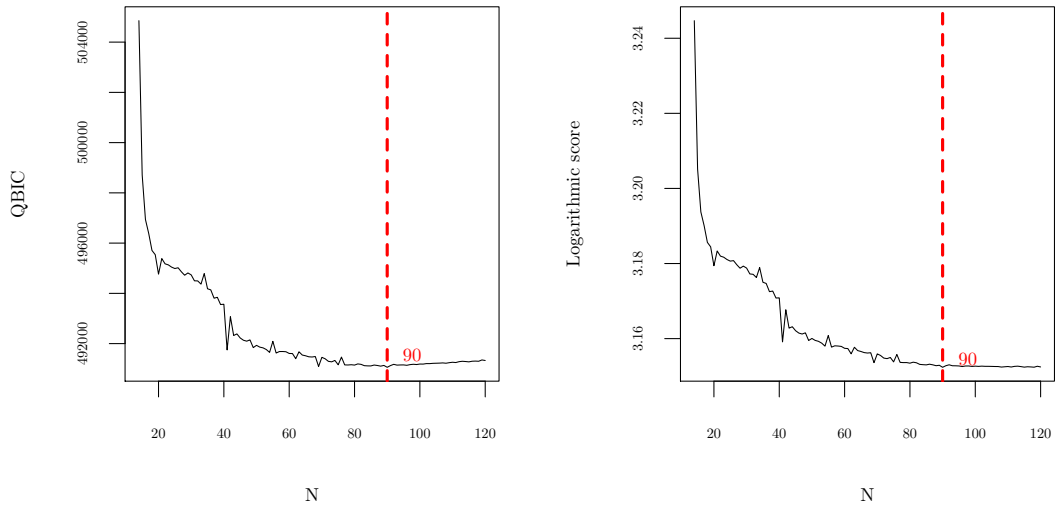
3: $y_t = \alpha + \beta_0 x_t + \sum_{i=1}^{N} \beta_i B_{i,3} x_t$

4: $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_N)$

5: Fit `glm()` with identity link function and the initial parameter vector $\hat{\boldsymbol{\theta}}$

Output: Model summary :

6: $\hat{\boldsymbol{\theta}}^* = (\hat{\alpha}^*, \hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_N^*)$



(a) **(b)**

Figure 4.12: The optimal number of periodic B-spline basis (N) with the model selection methods: (a) QBIC, (b) the cross-validation.

Figure 4.13 shows the estimated functional coefficients from the Poisson regression with the periodic B-spline for the first week of 2018. The periodic B-spline also represents the same range of the Bluetooth detection rate in addition to the similar patterns over the weekdays and the different patterns for the weekends. The second peak observed on weekdays (Tuesday–Thursday) with the Fourier model has not been captured with this model. The total number of the model parameters will be equal to 92 with respect to choosing the number of knots at 90. Figure 4.14 shows the estimated weekly functional coefficients on Tuesday that is increasing to the morning peak at

8:00, reach a bottom at 09:35 and again start to rise to the evening peak at 16:10. The same as the Fourier model, this trend has been repeated for the other weekdays with the changes over the weekend.

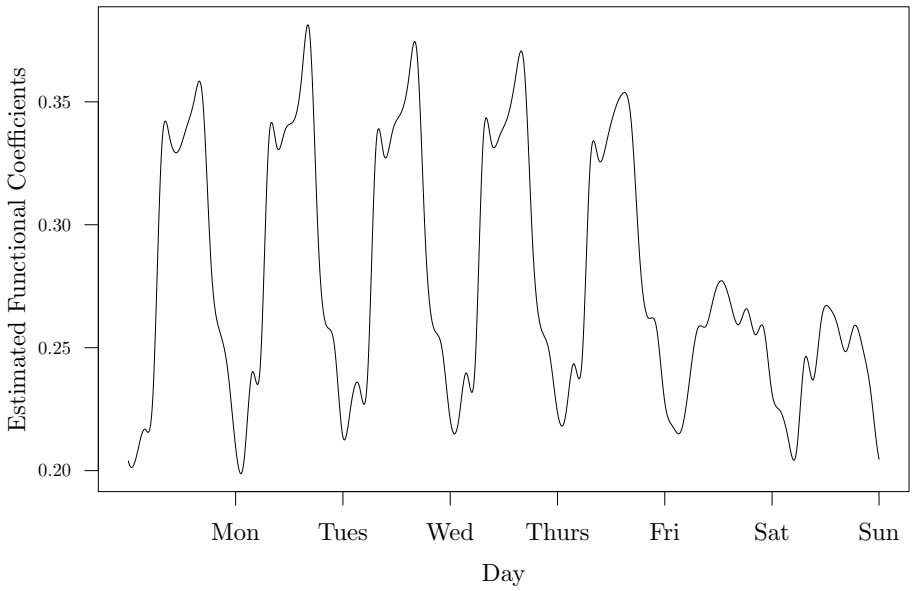


Figure 4.13: The estimated weekly functional coefficients from the Poisson regression with the periodic B-spline.

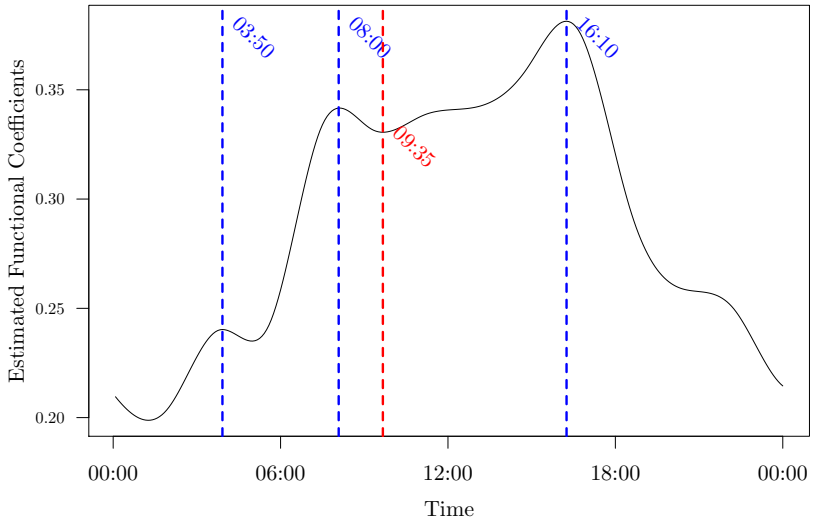


Figure 4.14: The estimated weekly functional coeffcients from the Poisson regression with periodic B-spline basis functions for Tuesday.

Table 4.1 shows the comparison between the Poisson regression with Fourier and the periodic B-spline basis functions using the QBIC and cross-validation. Also, Figure 4.15 represents the comparison of the estimated weekly functional coefficients from the Poisson regression with Fourier versus periodic B-spline basis functions. The result of the QBIC indicated that the Poisson regression with Fourier basis is slightly better than the model with the periodic B-spline basis functions. However, the cross-validation suggested the periodic B-spline basis model. Because there is only slight difference between these two models, the Fourier will be preferred due to the fewer number of parameters.

Model	QBIC	CV	Number of parameters
Fourier	491043.65	3.15253	88
Periodic B-spline	491056.93	3.15236	92

Table 4.1: The comparison between the Poisson regression with Fourier and the periodic B-spline basis functions using the QBIC and cross-validation.

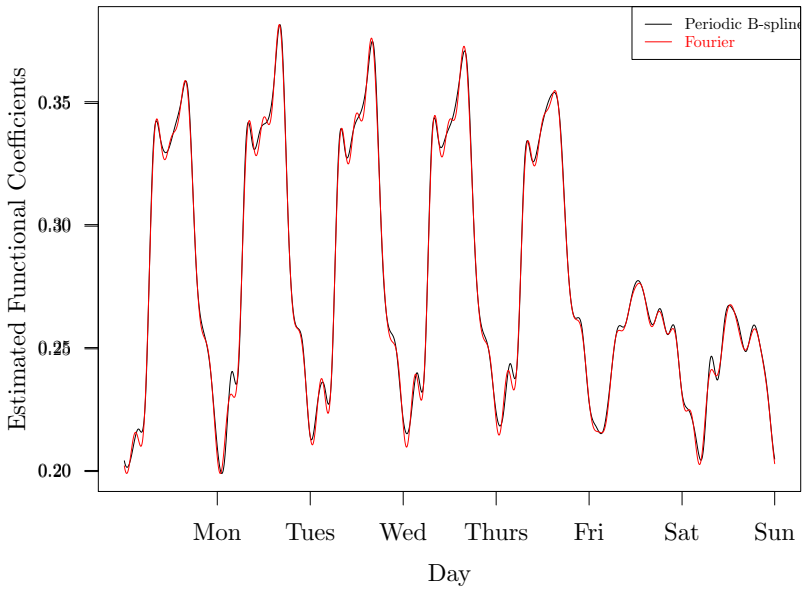


Figure 4.15: The comparison between the estimated weekly functional coeffcients from the Poisson regression with Fourier and the periodic B-spline basis functions.

4.5 Calibration

Bluetooth data represents the advantages compared to traditional data acquisition in terms of low-cost data collection, installation and maintenance of Bluetooth detectors and privacy protection. It has been shown the abilities to deliver some of the important traffic management aspects such as travel time estimation, speed, origin-destination matrix (Barcelö et al., 2010; Puckett and Vickich, 2010; Malinovskiy et al., 2011; Barceló Bugeda et al., 2012a; Michau et al., 2014; Purser, 2016; Michau, 2016; Cotten et al., 2020; Liu et al., 2020).

At this stage, we have built regression models that are more consistent with reality by considering the variability in the relationship between Bluetooth and ATC records at different times and days of the week. Therefore, the practical idea is to recover ATC from the Bluetooth data which means predicting ATC value from the known number of the Bluetooth counts for a particular time of a day. As the ATC is the response and Bluetooth is the predictor in the regression models, the statistical calibration method can be used to achieve this goal (Brown, 1982; Osborne, 1991).

Based on the idea of inverse regression, the *classical estimator* can be considered as the simplest approach of calibration. The inverse solution of the regression model uses the known response value to approximate the unknown regressor. Our final regression

model with the complex combination of the Fourier or B-spline basis functions will make it complicated to use inverse regression. Also, it would be inappropriate to calculate a normal approximation based confidence interval for the parameter of interest when the sampling distribution of the estimate is not normal. Alternatively, there is another approach named *profile likelihood*, which is particularly effective in the non-normal model, can produce the confidence interval by taking into account any possible asymmetry in the shape of the likelihood. This method is profiling the likelihood of the fitted regression model for a range of unknown regressors and known response to choose a point with the maximum likelihood as the best estimation of the regressor with confidence interval.

The profile likelihood method is very computationally intensive, while the classical estimator is more computationally efficient. Therefore, we will first introduce the classical estimator and then profile likelihood. The classical estimator is utilised to acquire a suitable range of unknown regressor in the profile likelihood implementation.

4.5.1 The classical estimator

As a start, assume that the regression model is as follows:

$$y_t = \alpha + \beta(t)x_t + \epsilon_t. \tag{4.26}$$

where α is the intercept, and $\beta(t)$ are the functional coefficients and ϵ_t is the error term. It looks like an ordinary linear regression model, but as y_t is Poisson random variable $y_t \sim \text{Pois}(\alpha + \beta(t)x_t)$, so $\epsilon_t = y_t - E[y_t]$ is no longer normal distributed and we can consider the expected value of ϵ_t is zero as $E(y_t) = \mu_t$. Based on the assumption of the equality of variance and mean in Poisson regression, the variance of ϵ equals to μ_t which is unknown. The inverse regression method can be applied to estimate the unknown value of x_t by the inverse solution of model (4.26) as follows:

$$x_t = \frac{y_t - \alpha - \epsilon_t}{\beta(t)}. \tag{4.27}$$

Finally, with the observed y_t and the estimated values of α and $\beta(t)$ from the model (4.26), the classical estimator of x_t is as follows:

$$\hat{x}_t = \frac{y_t - \hat{\alpha}}{\hat{\beta}(t)} \tag{4.28}$$

given the fact that ϵ_t has mean zero and $\hat{x}_t$ can be obtained by utilising the inverse regression estimator.

As $\beta(t)$ is estimated as a complex combination of the Fourier or the periodic B-spline basis functions, it is slightly complicated to calculate the standard error for the

classical estimator $\hat{x}_t$. The *multivariate Delta method* can be utilised to approximate the standard error for x_t defined in (4.27) and produce a reasonable confidence interval. This method estimates the mean and the variance of a function of random variables using the Taylor series approximation. For example, if $g(\boldsymbol{X})$ is a scalar function of the random vector $\boldsymbol{X}$, using multivariate Delta method, the variance of $g(\boldsymbol{X})$ can be estimated as:

$$Var(g(\boldsymbol{X})) = \Big(\nabla g(\mu)\Big)\, \mathbf{V}\, \Big(\nabla g(\mu)\Big)^T \tag{4.29}$$

where $\mathbf{V}$ is variance–covariance matrix of $\boldsymbol{X}$ and $\nabla g(\mu)$ is the gradient vector of g at $\mu = E(\boldsymbol{X})$.

In our case, the function $g(\boldsymbol{X})$ is the model presented in (4.27), where the vector $\boldsymbol{X}$ consists of three components as follows:

$$\boldsymbol{X} = \begin{pmatrix} \epsilon \\ \alpha \\ \beta(t) \end{pmatrix}. \tag{4.30}$$

Note that it also takes ϵ into account to construct the prediction interval for $\hat{x}_t$.

Suppose Poisson Fourier regression model with the optimal number of m harmonic functions is selected as the final model, the variance-covariance matrix $\mathbf{V}$ can be written in block form as follows:

$$\mathbf{V} = \left(\begin{array}{c|cc} \sigma_\epsilon^2 & 0 & 0 \\ \hline 0 & & \\ 0 & & \tilde{V} \end{array}\right) \tag{4.31}$$

where the error term ϵ is independent of $\hat{\alpha}$ and $\hat{\beta}(t)$ because it will be estimated from the new data set whereas the coefficients are extracted form the model. Therefore, the first row and column of $\mathbf{V}$ include zeros and σ_ϵ^2. Note we do not know about the ϵ, but the variance of ϵ equals to μ_t, therefore, the variance of ϵ is approximated by the known y_t. The sub-matrix $\tilde{V}$ is the covariance matrix of α and $\beta(t)$ for the particular time t that will be extracted from the model.

To compute the sub-matrix $\tilde{V}$, firstly, consider the parameter vector $\boldsymbol{\theta}$ for the Fourier model as:

$$\boldsymbol{\theta} = (\alpha, \beta_0, a_1, \ldots, a_m, b_1, \ldots, b_m)^T \tag{4.32}$$

U is a $2 \times (m+2)$ sub-matrix from the design matrix used to fit the model, which it defines to consider the intercept and the slope and the Fourier coefficients for the particular time t.

$$U = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cos(\frac{2\pi t}{T}) & \cdots & \cos(\frac{2m\pi t}{T}) & \sin(\frac{2\pi t}{T}) & \cdots & \sin(\frac{2m\pi t}{T}) \end{pmatrix} \tag{4.33}$$

An estimation for $\hat{\alpha}$ and $\hat{\beta}(t)$ for the particular time t is given as follows:

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta}(t) \end{pmatrix} = U\hat{\boldsymbol{\theta}} \tag{4.34}$$

where $\hat{\theta}$ is extracted after the fitted Poisson Fourier regression model (see Section 4.4.2). Finally, the variance of the estimations can be obtained as:

$$\tilde{V} = \text{Var}\begin{pmatrix} \hat{\alpha} \\ \hat{\beta}(t) \end{pmatrix} = U\Sigma U^T \tag{4.35}$$

where Σ is a $(m+2)\times(m+2)$ corresponding to the covariance matrix of the parameters obtained from the regression model ($m+2$ is the overall number of sine and cosine in the Fourier series plus the slope and the intercept). Obtaining $\tilde{V}$ from Equation (4.35) will complete the covariance matrix $\mathbf{V}$ in Equation (4.31).

The function g defined in (4.27) is a nonlinear function and the other part of the multivariate Delta method is to estimate the partial derivatives of g with respect to vector $\boldsymbol{X}$ as follows:

$$\nabla g|_{\mathbf{X}} = \begin{pmatrix} \frac{\partial g}{\partial \epsilon} \\ \frac{\partial g}{\partial \alpha} \\ \frac{\partial g}{\partial \beta(t)} \end{pmatrix} = \begin{pmatrix} \frac{-1}{\hat{\beta}(t)} \\ \frac{-1}{\hat{\beta}(t)} \\ \frac{\hat{\alpha}-y_t}{(\hat{\beta}(t))^2} \end{pmatrix} \tag{4.36}$$

Finally, after obtaining the covariance matrix $\mathbf{V}$ and ∇g, the variance and standard error for $\hat{x_t}$ will be computed as shown in equation (4.29). The calibration interval with the confidence level α is given by,

$$\hat{x_t} \pm z_{\alpha/2} \cdot (\text{S.E.}(\hat{x_t})) \tag{4.37}$$

However, there is also the over-dispersion which should be considered as a factor in Poisson regression in the presence of excessive variabilities. As a result, the standard error for $\hat{x_t}$ can be modified by multiplying it by the square root of the estimated over-dispersion parameter derived from the fitted Poisson model.

$$\hat{x_t} \pm z_{\alpha/2} \cdot (\text{S.E.}(\hat{x_t}) \cdot \sqrt{\hat{D}}) \tag{4.38}$$

The calibration interval specifies a range of trial $\hat{x_t}$ values which will be further used as an initial range in the profile likelihood method described in the following subsection.

4.5.2 The profile likelihood

Denoting the original data set by $\mathcal{F} = \{(x_t, y_t)|t = 1, \ldots, T\}$, the profile likelihood method performs by adding the new data (x_τ, y_τ), where τ is a new time point, x_τ is the unknown ATC data and y_τ is a known Bluetooth count. The regression model is fitted for the augmented dataset $\mathcal{F} \bigcup \{(x_\tau, y_\tau)\}$ and the profile log-likelihood function

for x_τ computed as follows:

$$l_p(x_\tau) = \max_{\boldsymbol{\theta}} \left\{ l(\boldsymbol{\theta}|\mathcal{F} \bigcup \{(x_\tau, y_\tau)\} \right\} \tag{4.39}$$

Finally, the optimum estimation of x_τ with the maximum log-likelihood is:

$$x_\tau^* = \underset{x_\tau}{\text{argmax}}(l_p(x_\tau)) \tag{4.40}$$

An approximate (1-α)% confidence interval for x_τ is the set of values satisfying,

$$\left[x_\tau : 2\left\{l_p(x_\tau^*) - l_p(x_\tau)\right\} \leq \chi^2_{1-\alpha}(1)\right] \tag{4.41}$$

where $\chi^2_{1-\alpha}(1)$ denotes the (1-α)th quantile of the chi-squared distribution with the one degree of freedom, i.e. equal to the number of parameters in the profile likelihood (Davidson and MacKinnon, 1993; Murphy and Van der Vaart, 2000; Jones, 2008).

4.6 Calibration implementation and results

The Fourier Poisson regression model with the optimal number of m is used to implement the calibration, but the procedure is the same for the periodic B-spline Poisson regression model. Considering the model,

$$y_t = \alpha + \beta_0 x_t + \sum_{k=1}^{m} a_k \left[\cos\left(\frac{2\pi kt}{S}\right) x_t\right] + \sum_{k=1}^{m} b_k \left[\sin\left(\frac{2\pi kt}{S}\right) x_t\right] \tag{4.42}$$

The `glm()` function estimated α, β_0, and all Fourier coefficients using the Poisson distribution and the identity link function. Assume the number of Bluetooth counts y_τ in the five minute time interval between 17:05-17:10 (i.e. $\tau = 206$) on Monday is 19 (i.e. $y_\tau = 19$) and the goal is to predict how many vehicles (i.e. ATC or x_τ) passed through the area.

As first calibration method, the classical estimator was used to compute an initial value for x_τ, and the multivariate Delta method was used to approximate the standard error to produce a trial range for the profile likelihood method of x_τ. The initial value of $\hat{x}_\tau$ was 51, with estimates of 17 and 1.36 for standard error and overdispersion, respectively (i.e. S.E.$(\hat{x}_t) = 17$, $\hat{D} = 1.36$). Finally, using $z_{\alpha/2} = 3$ (i.e. to have a wide enough range for x_τ), a 99% confidence interval of $(1, 101)$ was derived from Equation (4.38).

To begin, using the classical estimator method, the profile likelihood approach obtained a range trial of $(1, 101)$ for x_τ. Then, the method continued by adding the new point (x_τ, y_τ) to the original data set. The regression model (4.42) refitted for each new

data set and the profile log-likelihood function $l_p(x_\tau)$ is monitored over the trial range. Finally, the technique chose the best estimate as the point with the highest likelihood, $x_\tau^* = 51$, and a 99% confidence interval of $(21, 98)$ retrieved from Equation (4.41).

The same estimation for $\hat{x_t}$ is obtained using both calibration methods; however, the profile log-likelihood technique yielded an asymmetrical (and thus more accurate) confidence calibration interval for the estimation.

The calibration result using the profile log-likelihood is shown in Figure 4.16, where the curve shows the changes in the value of the log-likelihood for the range of x_τ. The vertical red and the dashed blue lines mark the optimum value and the confidence calibration interval, respectively.

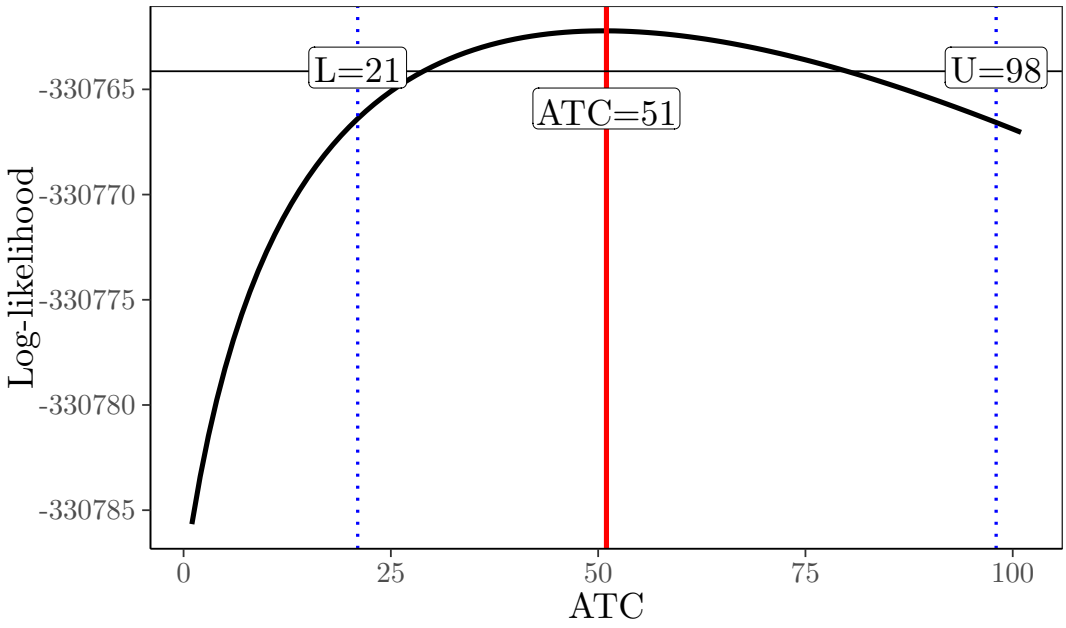


Figure 4.16: The calibration result using the profile log-likelihood where the number of Bluetooth counts is assumed to be 19 in the five minute time interval between 17:05-17:10 on Monday. The vertical red and the dashed blue lines mark the optimum value and the confidence calibration intervals, respectively.

The calibration results using both techniques, the classical estimator and the profile log-likelihood, are also shown in Figure 4.17, and also with actual recorded observations for all five minute time intervals between 8:00 am and 12:00 pm on Monday, February 5, 2018. The blue line represents the actual observations, and the overlapping black and red line shows the prediction results from the profile log-likelihood and classical estimators. The dashed black lines indicate the confidence calibration intervals of the profile log-likelihood. The calibration results indicate that some time intervals were overestimated and others were underestimated, and it captured some but not all of the variability in the data.

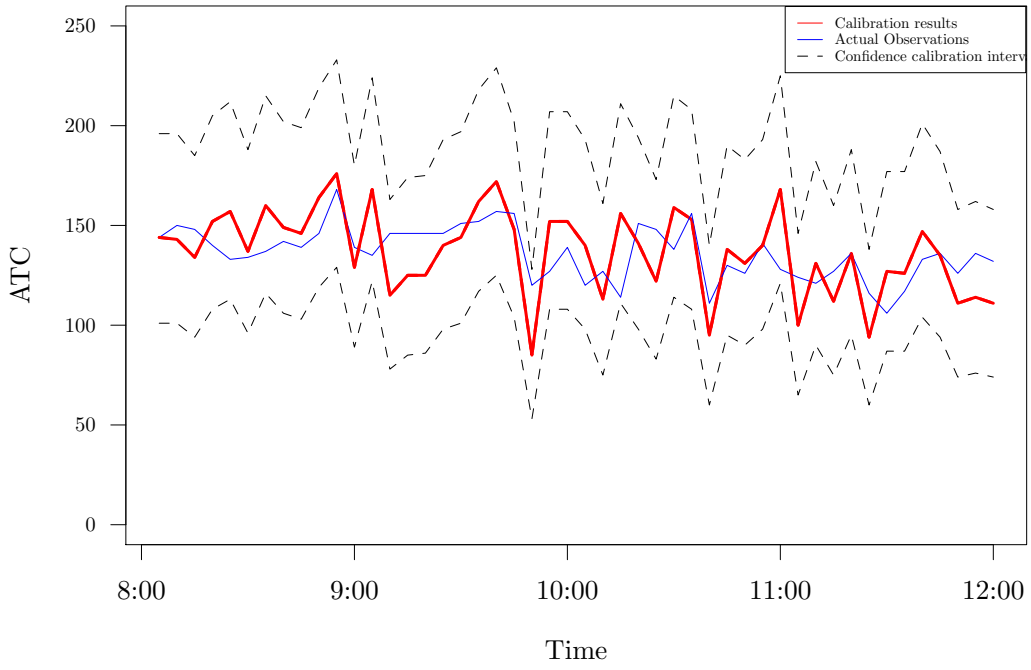


Figure 4.17: The calibration result using the profile log-likelihood method and actual recorded observations for all five minute time intervals between 8:00 am and 12:00 pm, Monday 5th February 2018. The blue line represents the actual observations, and the red line shows the prediction results from the profile log-likelihood and classical estimators. The dashed black lines indicate the confidence calibration intervals of the profile log-likelihood.

4.7 Discussion

In this chapter, we investigate the effect of different times on the days of the week when modelling the relationship between the rate of unique Bluetooth detections and ATC data. A framework based on generalized linear model was presented in this chapter. The Bluetooth and ATC data are considered as time series with two different temporal resolutions, one hour and five minute time slots.

To incorporate the time effects, two different models were used. Firstly, the same hourly pattern was assumed every day, which led to a Poisson regression model with stepwisely time-varying coefficients. This model was constructed based on the one hour time slots and showed the average rate of Bluetooth detection changed between the hours of the day and was constant within each hour. A more realistic assumption, however, was that the rate could vary within an hour on different days. Therefore, the second idea was to construct a Poisson regression model with smoothly time-varying

coefficients. The model was considered to have periodic functional coefficients to determine how the variations in ATC records over time would affect the underlying change in the rate of Bluetooth detections for different times of the day. In this regard, the Fourier series and the periodic B-spline were selected to define the periodicity part of the fitted model. Both methods were implemented, and the comparison revealed a slight difference between them; however, the Poisson regression model with a Fourier basis function can be chosen as the final choice with the fewest parameters. The results obtained from the fitted regression model indicated that the rate of Bluetooth detection per vehicle varied significantly in a consistent pattern weekly. There are some random events happening that will change the traffic conditions (e.g., weather conditions, national holidays and so forth) and the rates of Bluetooth which we cannot model. Thus, this result is only relevant to the part of the Bluetooth rate that is consistent throughout the year.

The potential practical goal was to recover ATC from Bluetooth data after developing an adequate regression model, which predicts ATC value based on the known number of Bluetooth detections for a particular time of day. The motivation was to take advantage of ATC prediction utilising Bluetooth detection so that traffic data could be captured with low-cost Bluetooth detectors. The statistical calibration method was applied using two methods, the classical estimator and the profile log-likelihood. The classical estimator also was utilized as the initial step for the profile log-likelihood implementation. Based on the known Bluetooth quantity, both calibration techniques generated the same estimate for the unknown ATC prediction. The profile log-likelihood procedure creates an asymmetrical calibration interval for the prediction, which is more realistic since x_t is a count from aggregating a Poisson process. A symmetrical interval based on normal approximation is hardly appropriate if considering the skewness in the distribution of x_t.

Chapter 5

Conclusions

5.1 Summary of thesis

Most research in the use of Bluetooth data for traffic monitoring has focused on estimating travel times or origin-destination matrices, with multiple detections at individual cites being treated as irrelevant or a nuisance. The issue of the representativeness of the data as a sample of all vehicles has, with a few exceptions, been ignored. Here we have investigated what, if anything, can be learned by analysing all the detections at a particular site. This has required the adaptation and extension of some sophisticated statistical techniques. These adaptations and extensions may also be useful in other contexts.

The main findings and contributions of this thesis are summarised in this chapter. In Chapter 2, we started with some exploratory data analyse for determining whether the complete record of Bluetooth detections would contain interpretable information at a particular Bluetooth site. The complete record means considering multiple detections rather than filtering them, since multiple detections is one of the problems in computing travel times and extracting meaningful features from the Bluetooth data. As a result of multiple detections, there are gaps between consecutive detections. Therefore, we investigated whether the distribution of observed gap periods for multiple detections could provide useful information for traffic inference. The hierarchical cluster analysis was utilised to categorize Bluetooth sites, MAC addresses, and time intervals of a day. In order to categorise MAC addresses and time intervals of a day, the methodology for clustering based on gap time distribution was proposed, which employed the Kolmogorov-Smirnov statistic. The cluster analysis results confirmed that there was information in the multiple detections because meaningful clusters are generated.

In Chapter 3, we also examined the relationship between ATC and Bluetooth detections, which may contribute to investigating possible sources of bias in the representativeness of Bluetooth detections. By taking into account that some observable

factors may influence the rate of Bluetooth detection, we used regression analysis as a powerful statistical method for modelling the relationship. Using the ATC dataset, we were able to extract the number of buses and the speed of the vehicles travelling through the area. heteroscedasticity caused an unequal scattering of residuals after regression analysis, which required to be resolved before making conclusions from the model. A non-parametric variance function estimate method was successfully applied after the rolling variance approach failed to define an appropriate parametric model for the variance function based on data. Using a non-parametric variance function estimation method, we fitted six alternative regression models for the data of the selected four case study locations, the first of which included the number of buses and the second of which included vehicle speed. The segmented regression model with three knots and two knots in some study locations was chosen as the best fit for the data in both scenarios. The results revealed that the number of buses and vehicle speed were not sufficient for explaining the Bluetooth rate variation. There must be other factors, such as weather, detector position, equipment failure, etc., affecting the rate and it would be impossible to account for them all in the model fitting given the current data.

In Chapter 4, we developed a Poisson regression model that characterises the rate of Bluetooth detection per vehicle as it varies over time. Therefore, this model was considered to have periodic functional coefficients in order to assess how changes in ATC records over time affect the underlying change in the rate of Bluetooth detections at different times of the day. It also took into account the part of the Bluetooth rate that has consistent hourly and daily patterns over the year. The model was designed to assess the feasibility of using the Fourier series or the periodic B-spline to describe the seasonal variations in the Bluetooth rate process by representing the model's parameters throughout the year. Instead of having the parameters constant for each time interval (here, 5 minutes), the parameters are assumed to change with time smoothly over a one-week period. The periodicity part of the fitted model was defined using either the Fourier series or the periodic B-spline. Although there was a slight difference between them, the Poisson regression model with a Fourier basis function was chosen as the final model with the fewest parameters.

We also examined a practical goal of recovering ATC from Bluetooth data after constructing an appropriate regression model in Chapter 4 which is related to the statistical calibration problem. The aim was to predict the unknown ATC value from the known number of Bluetooth counts for a given time of day. The classical estimator and the profile log-likelihood approach were used to apply the statistical calibration method. For the profile log-likelihood implementation, the classical estimator was also used as the initilazation step. Both calibration procedures produced the same prediction for the unknown ATC based on the known Bluetooth quantity. The profile log-likelihood

technique creates an asymmetrical calibration interval for the prediction, which suggests that a symmetrical interval based on approximate normality may not be appropriate.

5.2 Future research

There are a number of obvious next steps to be taken in this research. As noted in Chapter 2, KS-clustering method could be useful in situations where the clustering objects are sets of univarite data observed under different conditions. Simulation studies could be conducted to investigate the performance of KS-clustering, such as the sample size needed to reliably identify clusters, or different ways of defining the distance between groups for KS-clustering beyond the common linkage method.

The Bluetooth data points are assumed to represent a vehicle in most studies (Van Boxel et al., 2011; Bachmann et al., 2013; Remias et al., 2017). However, it can be related to any mode of transportation, such as a car, bus, bicycle, or pedestrian. Furthermore, a traveller may have multiple devices, for example, using a Bluetooth enabled headphone and smartphone, or a group of passengers using the same transportation mode, such as a bus, with an active Bluetooth device. These problems will result in significant bias and errors in travel time estimation or the approximate amount of traffic, especially in urban areas. Therefore, we could also explore ways to distinguish between different vehicle types using Bluetooth detection patterns. However, this would probably require conducting a field experiment. Identifying multiple Bluetooth MAC addresses that are tracked together across consecutive Bluetooth locations, for example, could perhaps be used to distinguish buses. However, it is possible that two MAC addresses in close proximity in traffic are incorrectly considered to belong to the same vehicle.

The Poisson regression model with smoothly time-varying coefficients proposed in Chapter 4 could perhaps be combined with more factors, such as the number of buses and speed. However, the implementation of the calibration method would now require knowledge of speed and the number of buses. The average speed of a vehicle can be determined if the corresponding Bluetooth device is detected by both upstream and downstream Bluetooth sensors. By matching the MAC address, the time required to travel the distance between the two fixed locations is determined and the average speed is calculated. However, this would probably not be a suitable proxy for the instantaneous speed of vehicles at a particular location. The timetable of buses can be used to get an approximation of the number of buses that may pass through that area in a particular time period, but this is unlikely to be very accurate over short periods of time.

We could try to extend the Poisson regression model in Chapter 4 by allowing the constant intercept term in the model to vary over time. The problem with enabling the intercept to change with time was that it posed an identifiability problem, which

led to a functional intercept that explained most of the seasonal variation. Because the number of non-vehicle Bluetooth detections is small, a Bayesian framework approach might be employed by specifying some prior information about the constant to keep it from dominating the model. Also, validating the methodology for more locations and a larger network would be useful.

It would be useful to model the probability of a missed Bluetooth detection using the data available in future research. Modeling this probability is complicated due to the variety of causes for missed detections, such as the signal strength and activation status of different Bluetooth devices, the traffic conditions (congestion or free-flow), and the routes of vehicles leaving and then returning to the detection area. Without conducting a field experiment, it is difficult to understand the linkages between the Bluetooth detection data and the probability of a missed detection. If we could model this probability in a suitable way, especially without undertaking a field experiment, the resulting model could help us discover more realistic connections between the Bluetooth detection data and actual traffic flow.

Finally, a more refined spatial-temporal version of the proposed model in Chapter 4 can be developed as

$$y(s,t) = B(s,t)x(s,t) + \epsilon(s,t),$$

for a transportation network with only Bluetooth sensors. Following the idea of solving the statistical linear inverse problems discussed in Hazelton et al. (2021), the corresponding calibration estimate $\hat{x}_{s,t}$ can be used to recover the actual traffic flow. Although the Bluetooth count $y(s,t)$ may merely yield an inaccurate estimate of the ATC count $x(s,t)$, its low cost and high privacy can help it reach a much higher coverage in the transportation network. In addition, it will be very interesting and challenging to explore how to jointly model ATC and Bluetooth counts and further use a joint model with full information to recover the actual traffic flow.

The main practical question to address is how the analyses performed in this thesis could be used by practicing traffic engineers. Cluster analysis can be useful for finding outliers in the network. The outliers can be regarded as any Bluetooth sites or MAC addresses that show different behaviour. It could be used as a preliminary step before doing other analysis, or for grouping similar Bluetooth sites to consider similar analysis. Also, without any further monitoring traffic tools, it would be possible to classify the different times of the day in terms of the traffic conditions based on the gap time distribution.

In terms of the use of this analysis for traffic engineers, the regression models' results indicated how the rate of Bluetooth detections per vehicle is affected by buses, which could be taken into account when determining the appropriate spacing and positions for detectors in the network.

In terms of how traffic engineers could leverage this analysis, the calibration phase and recovering unknown ATC from the calibrated model could be widely used in traffic applications. Temporary ATC counters are deployed for short periods of time in some Manchester network locations, so by monitoring Bluetooth detections and ATC data in these locations and fitting the appropriate regression model based on the data, the calibrated ATC results could be used to predict the number of vehicles for the time periods when the temporary ATC counter has been removed.

Appendix A

Additional Figures and Statistical Tables

A.1 Time series plots

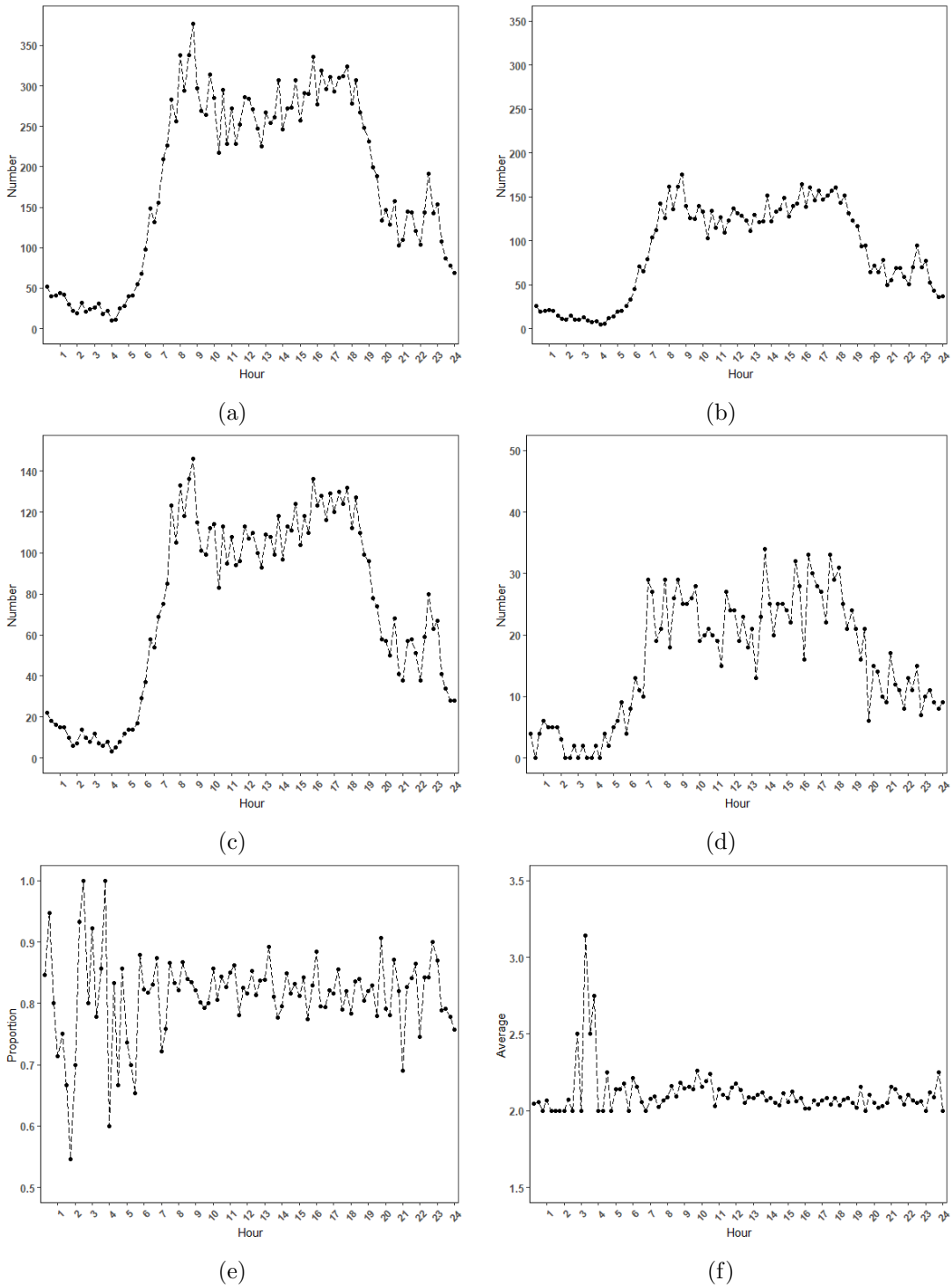


Figure A.1: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 12, Tuesday 12 February, 2019.

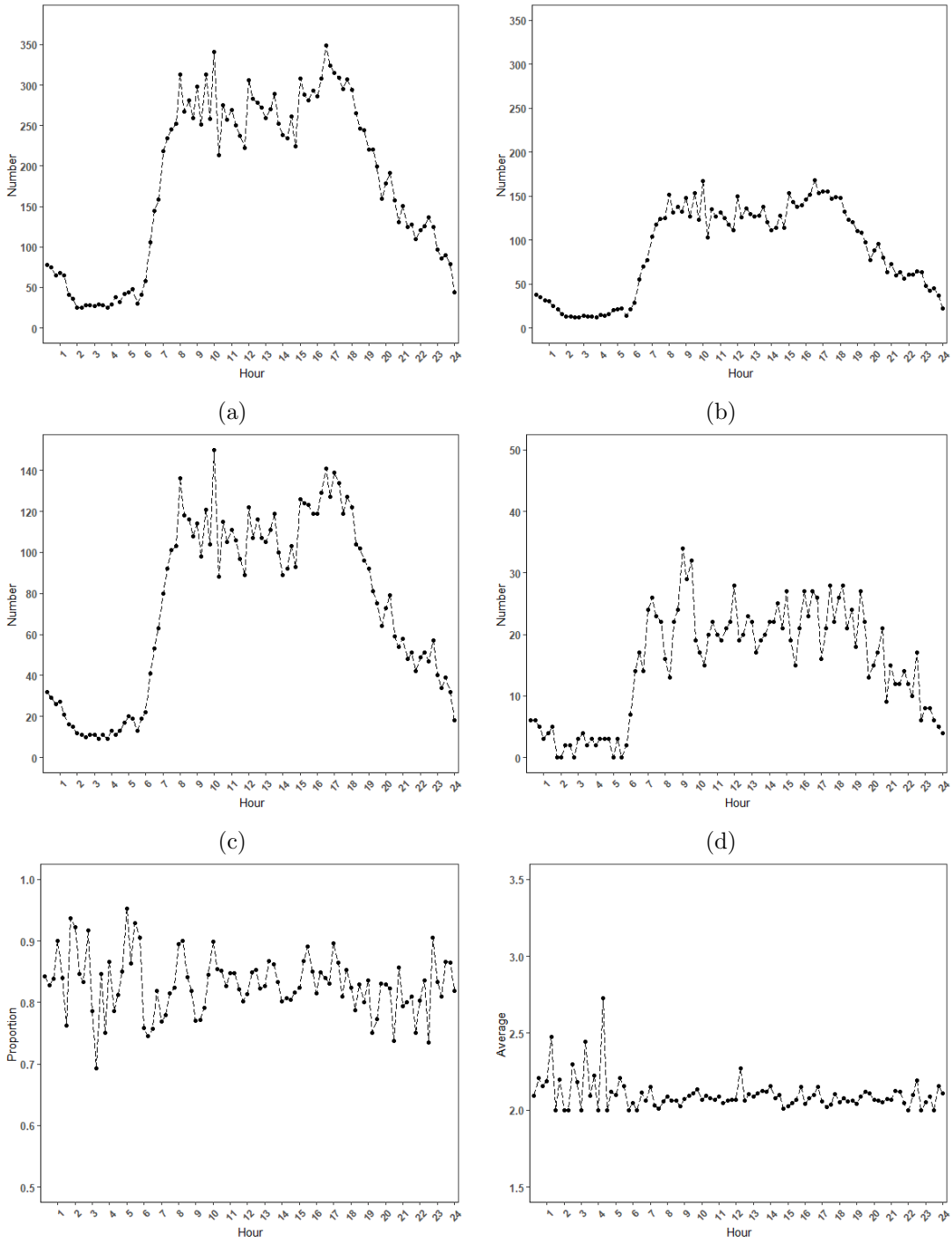


Figure A.2: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 12, Wedenesday 13 February, 2019.

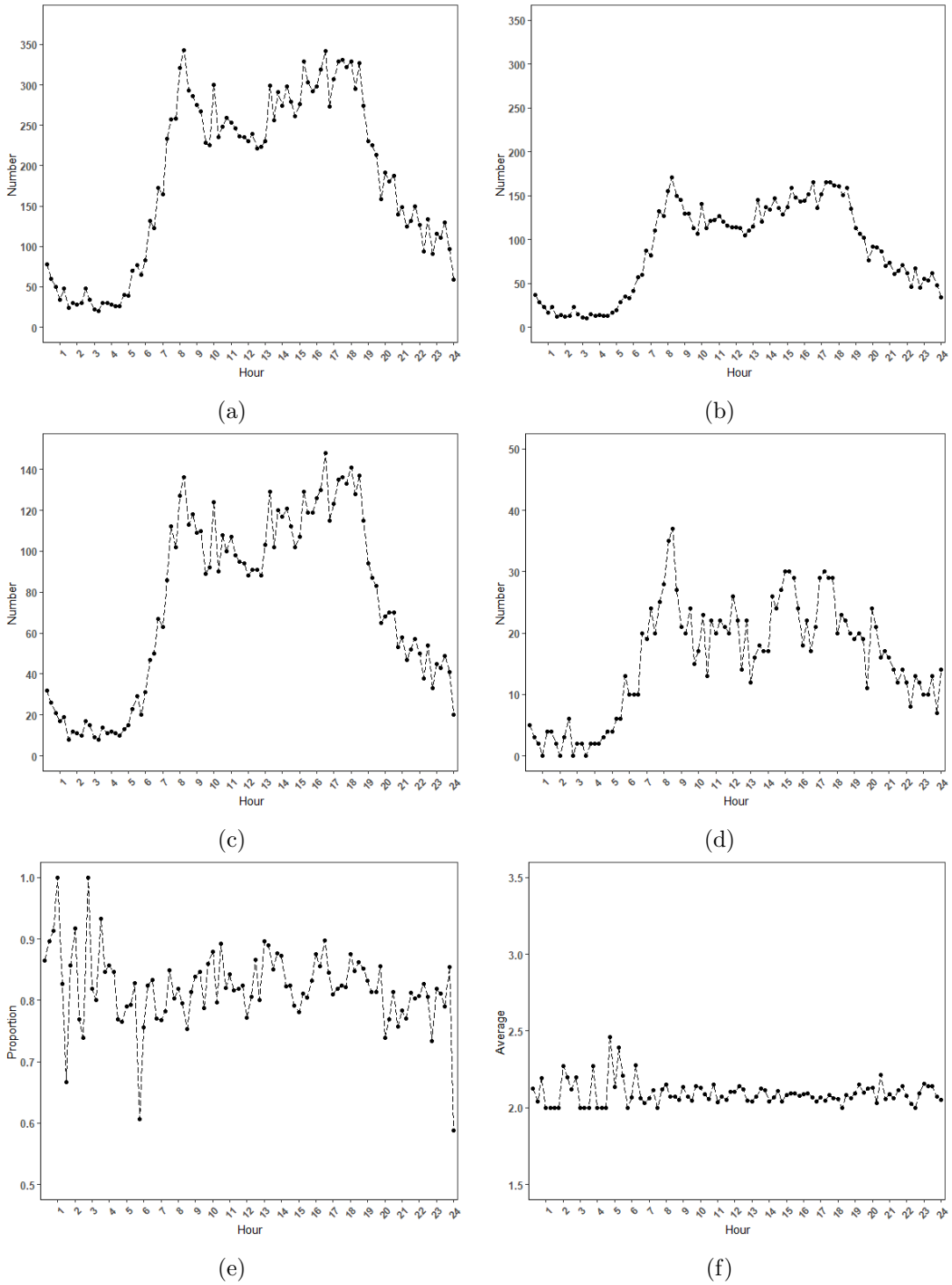


Figure A.3: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 12, Thursday 14 February, 2019.

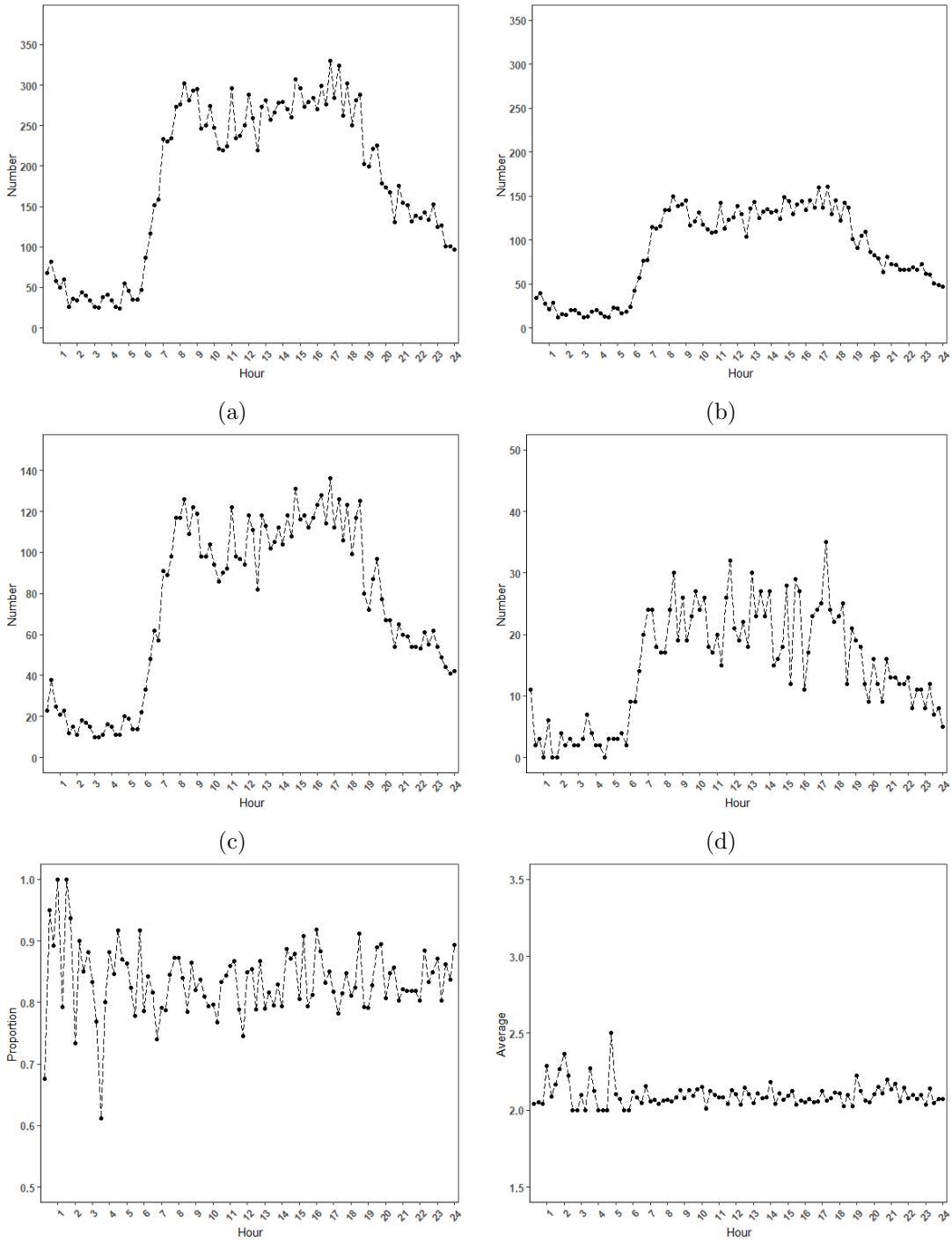


Figure A.4: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 12, Friday 15 February, 2019.

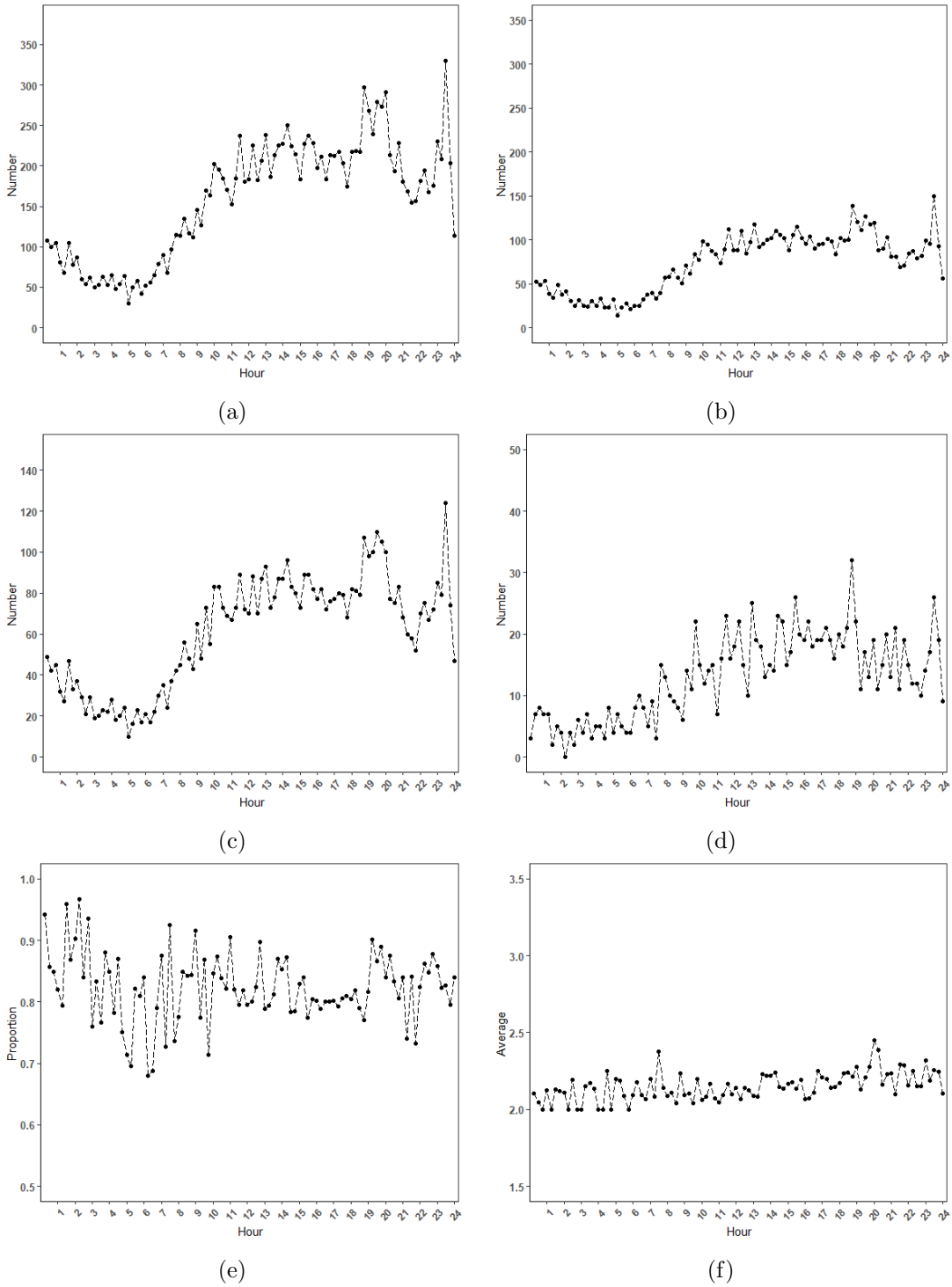


Figure A.5: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 12, Saturday 16 February, 2019.

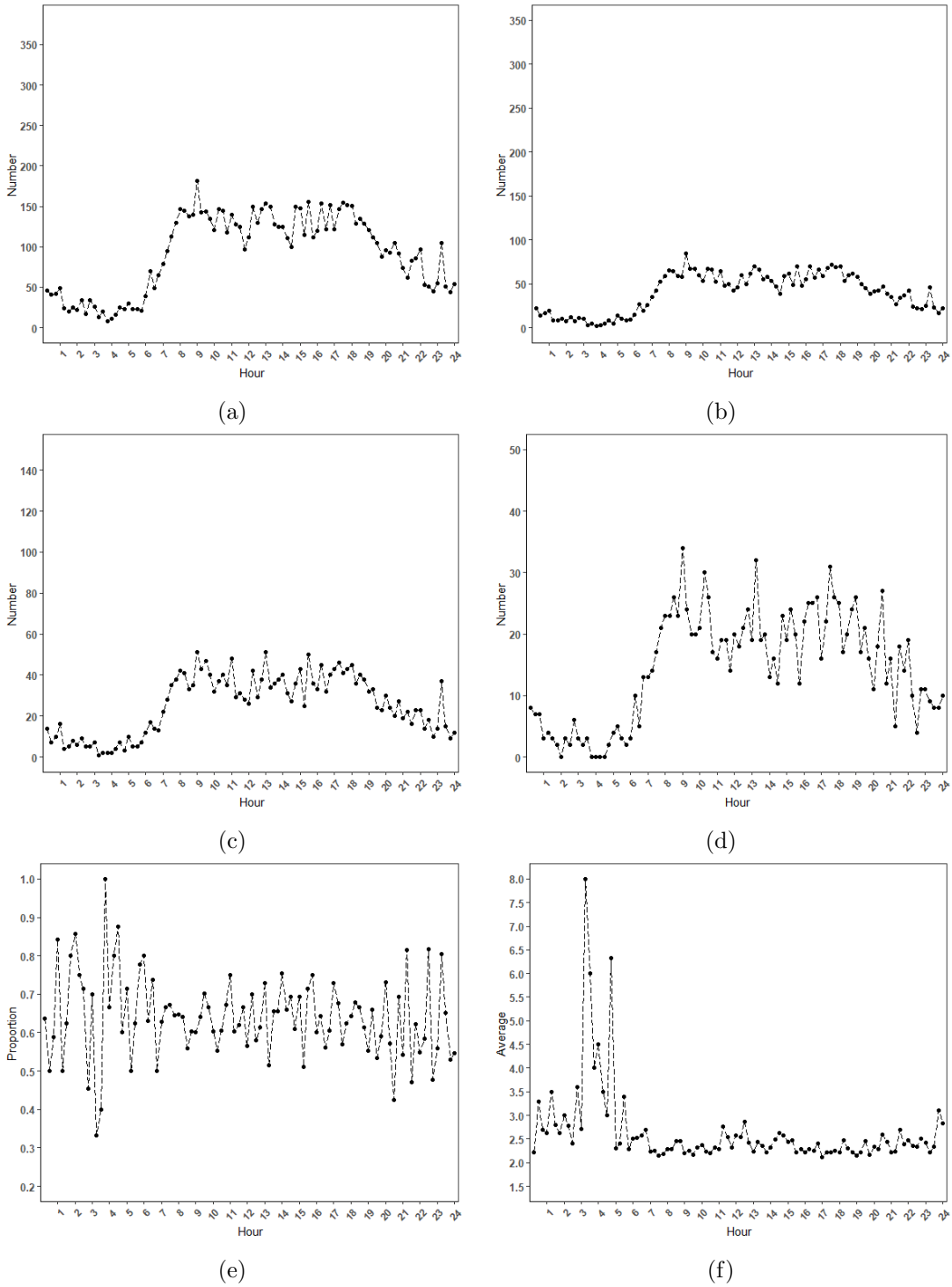


Figure A.6: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 1, Monday 11 February, 2019.

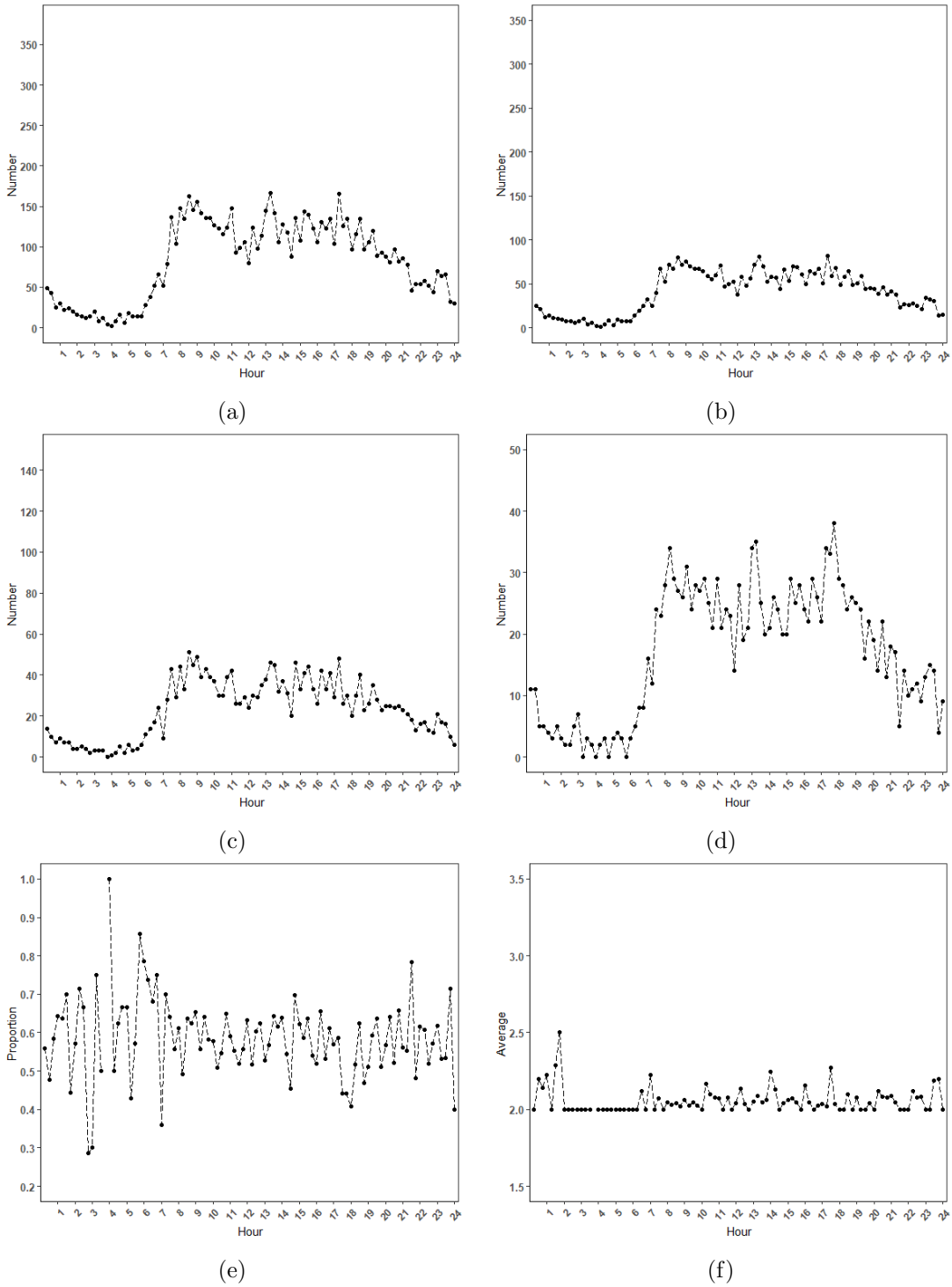


Figure A.7: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 2, Monday 11 February, 2019.

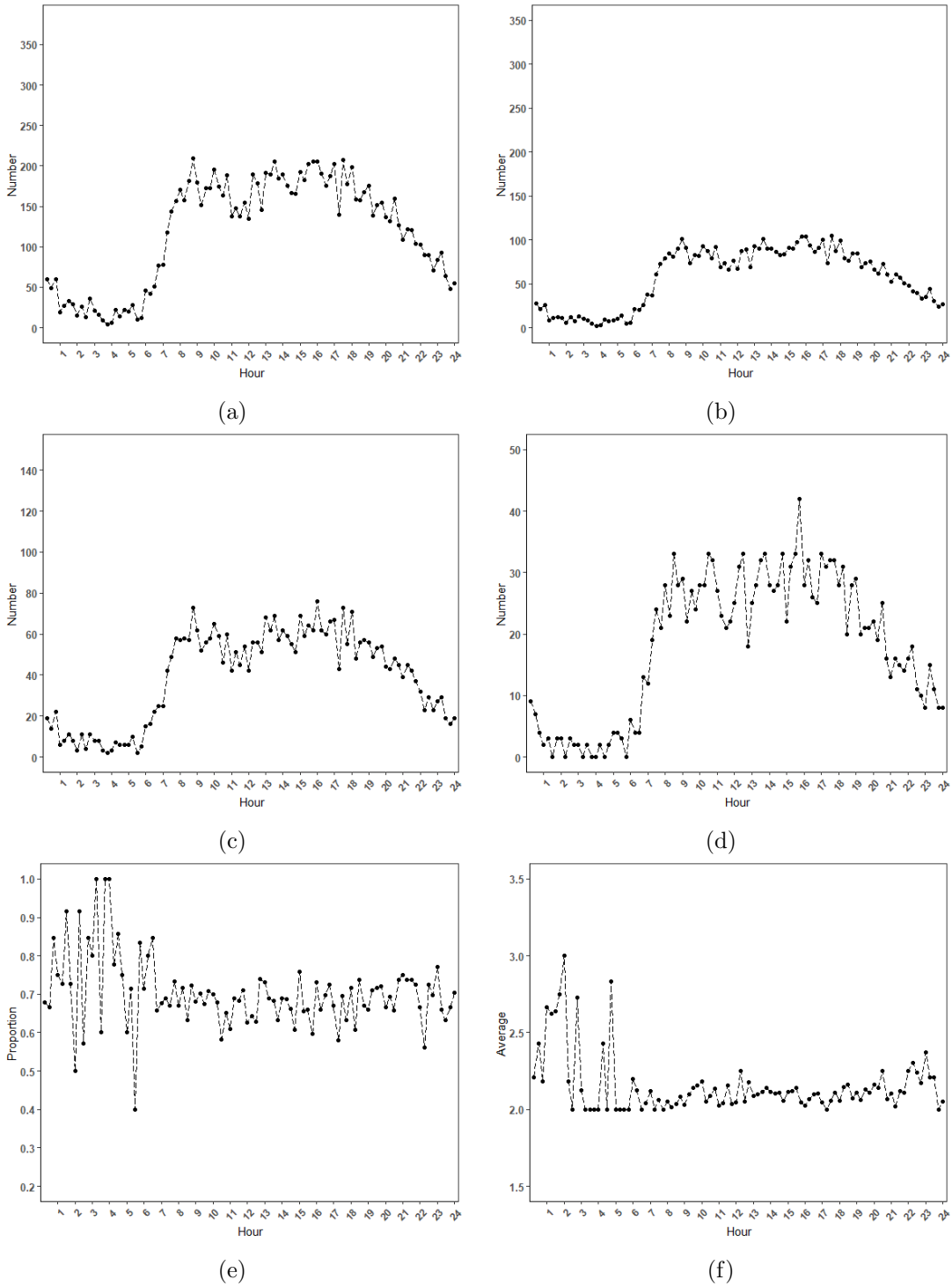


Figure A.8: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 3, Monday 11 February, 2019.

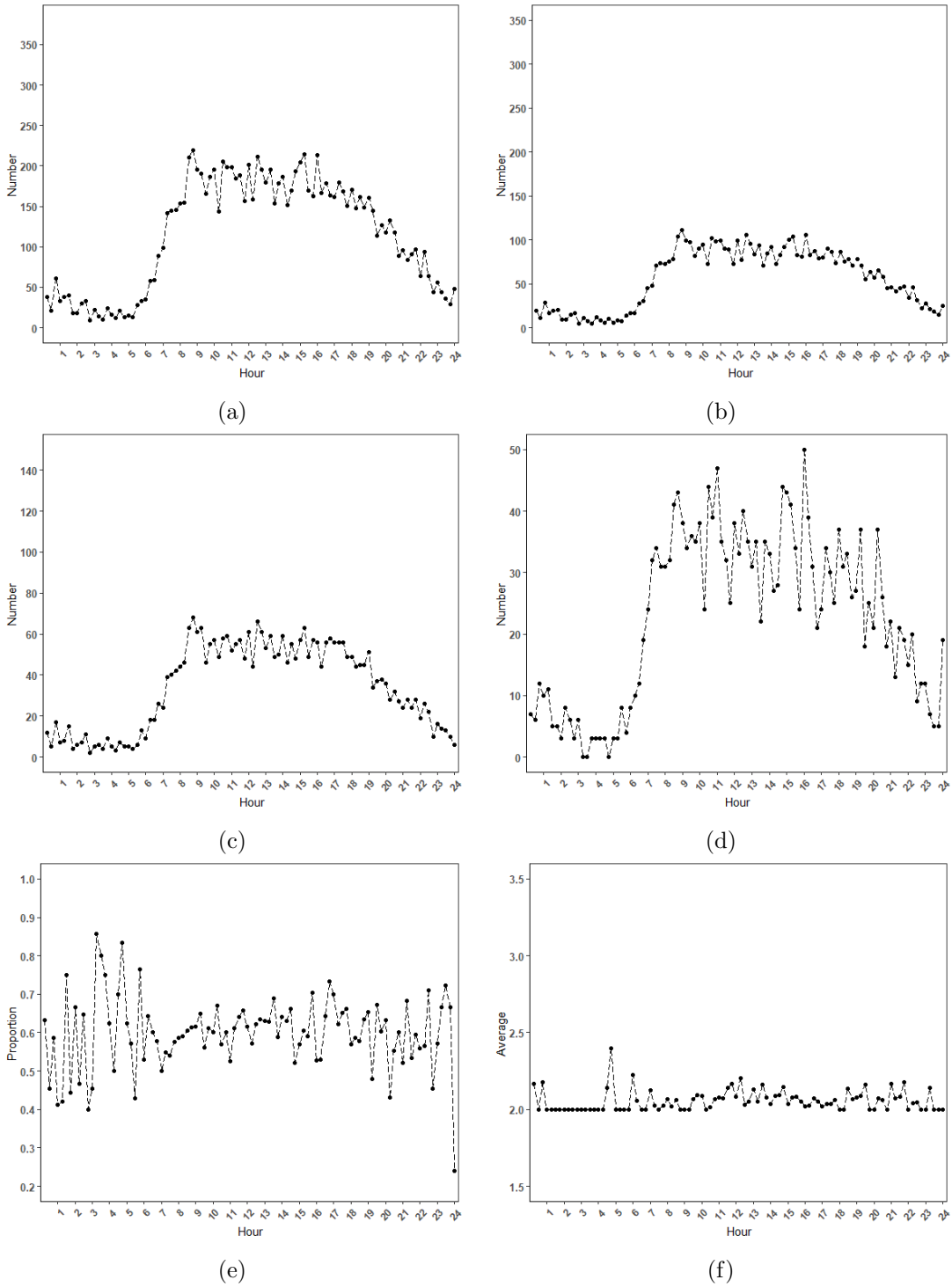


Figure A.9: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 6, Monday 11 February, 2019.

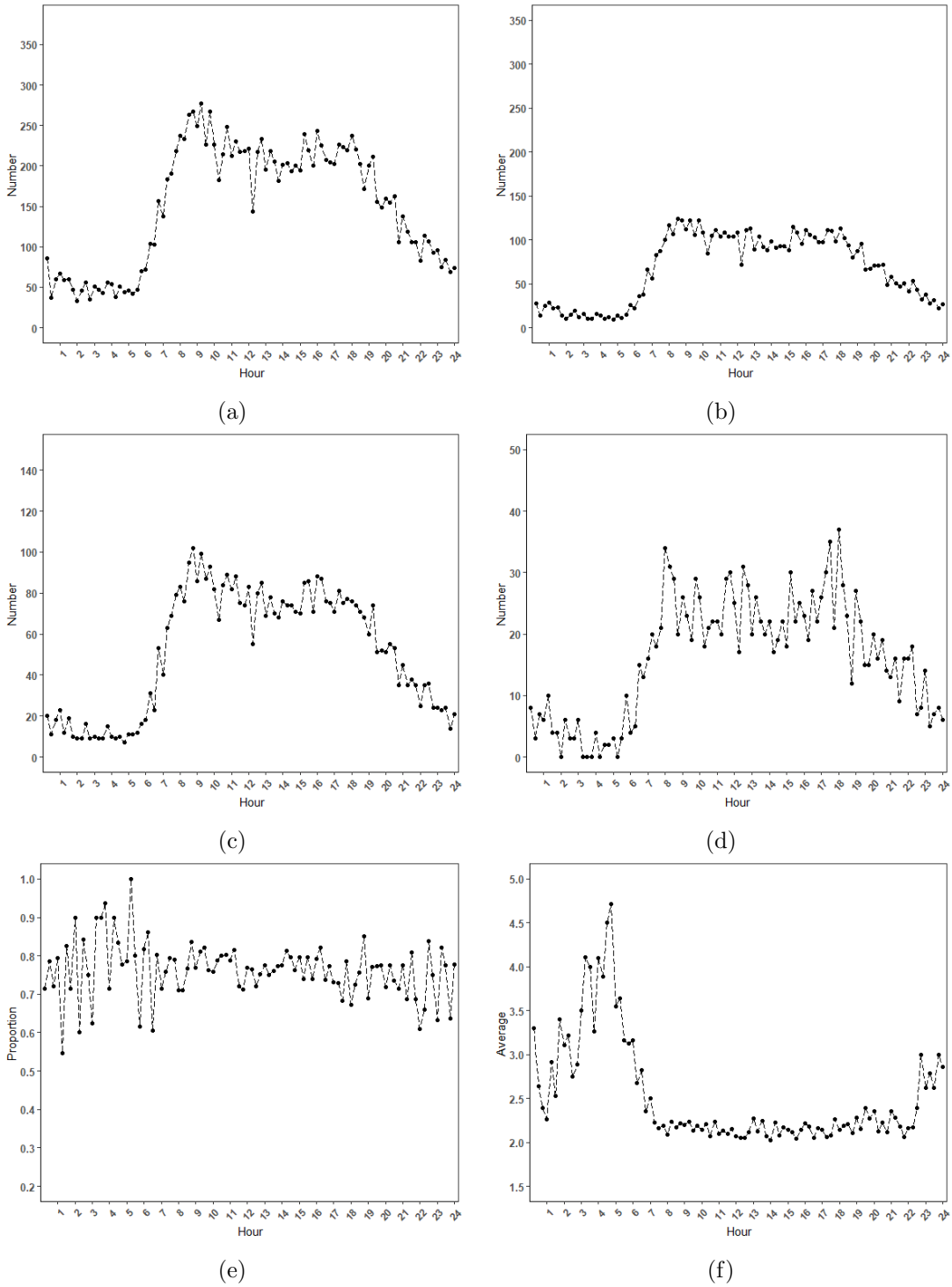


Figure A.10: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 7, Monday 11 February, 2019.

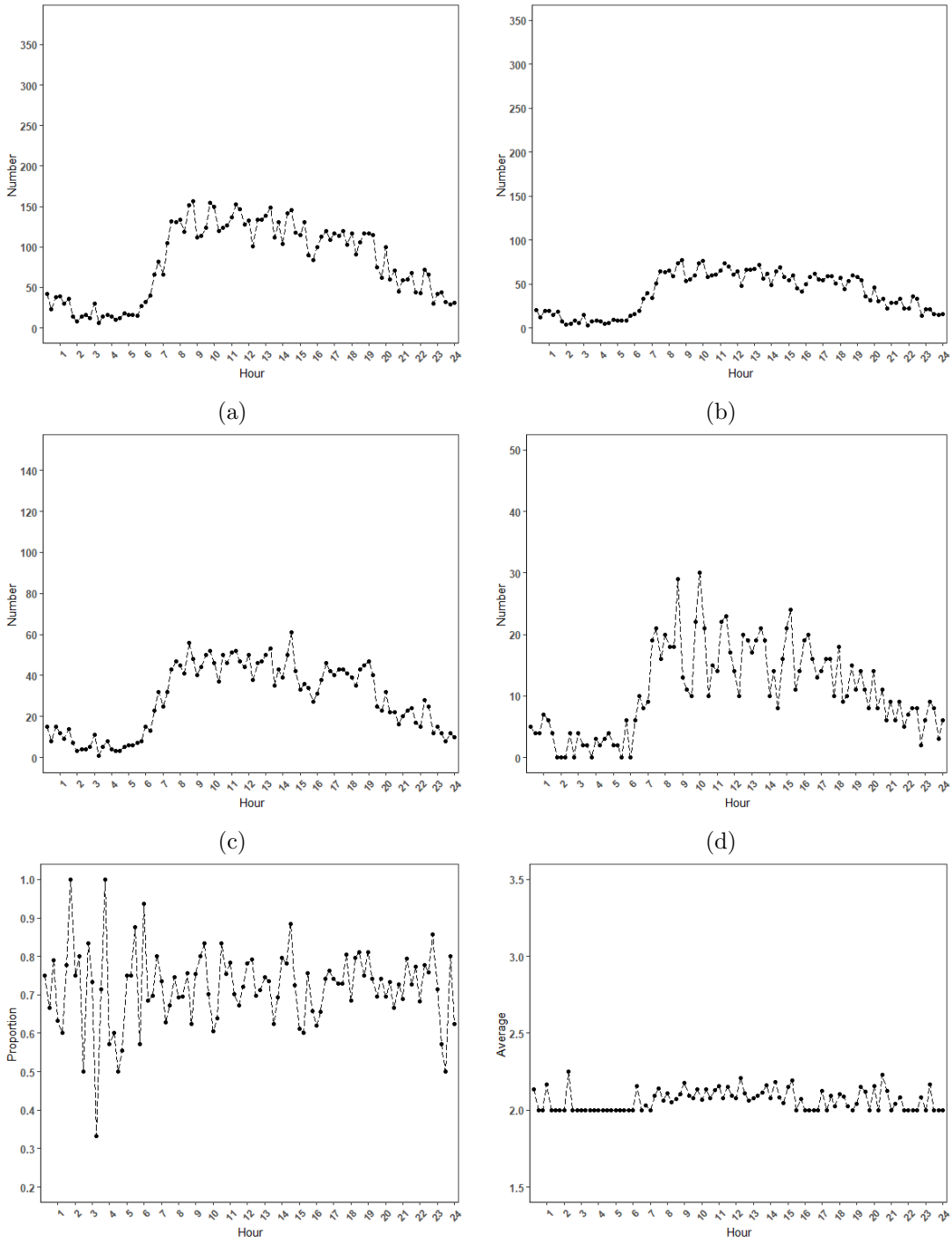


Figure A.11: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 9, Monday 11 February, 2019.

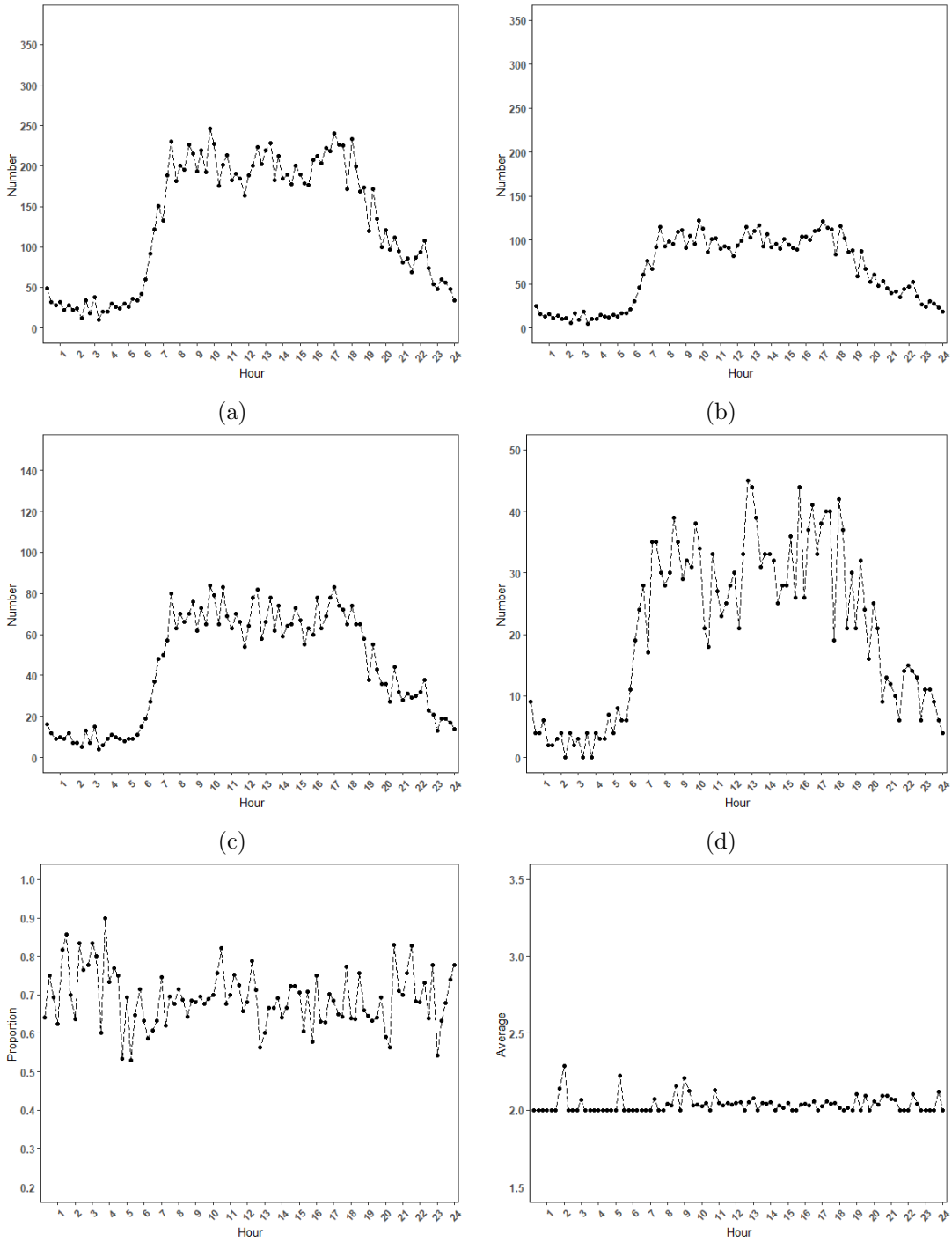


Figure A.12: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 10, Monday 11 February, 2019.

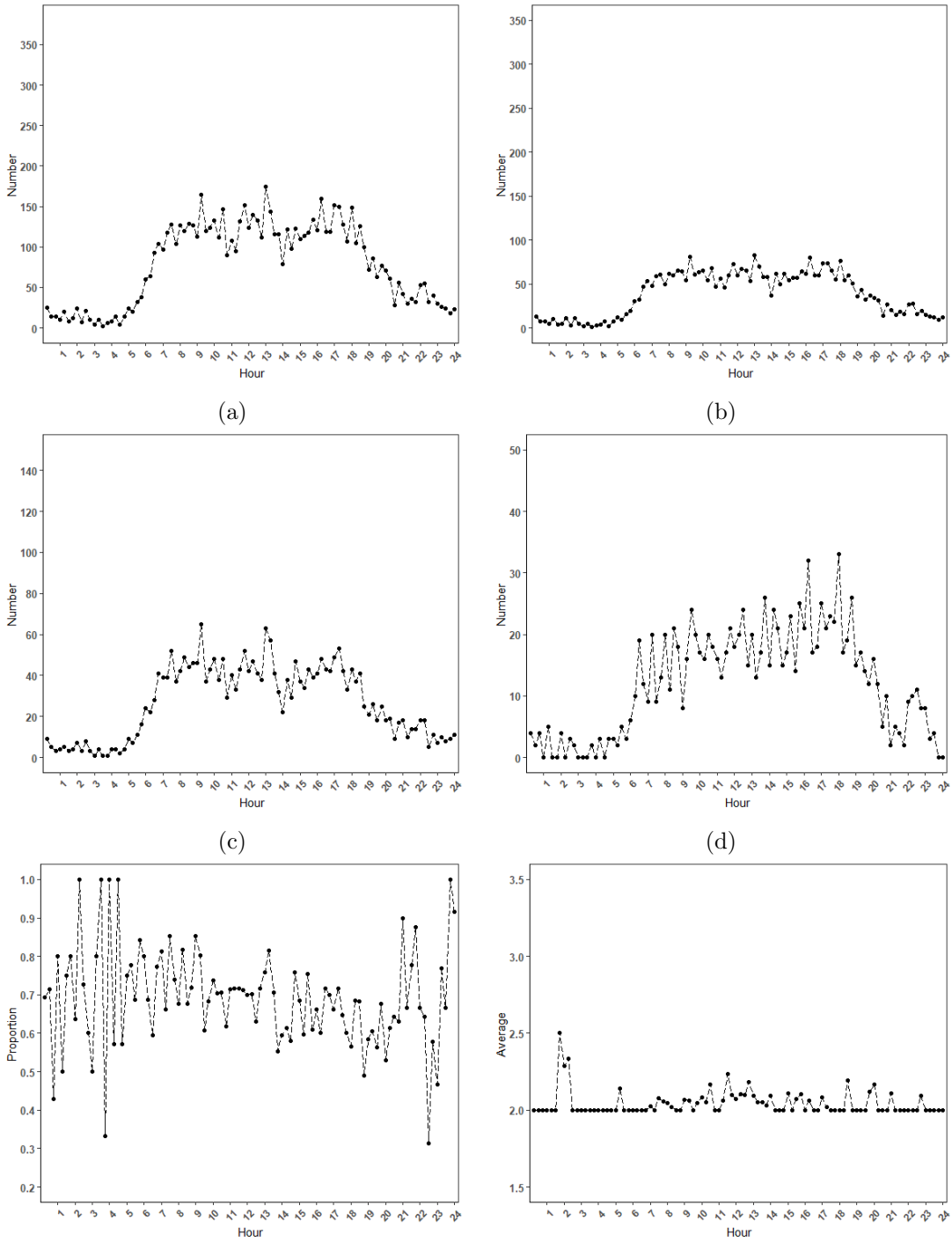


Figure A.13: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 11, Monday 11 February, 2019.

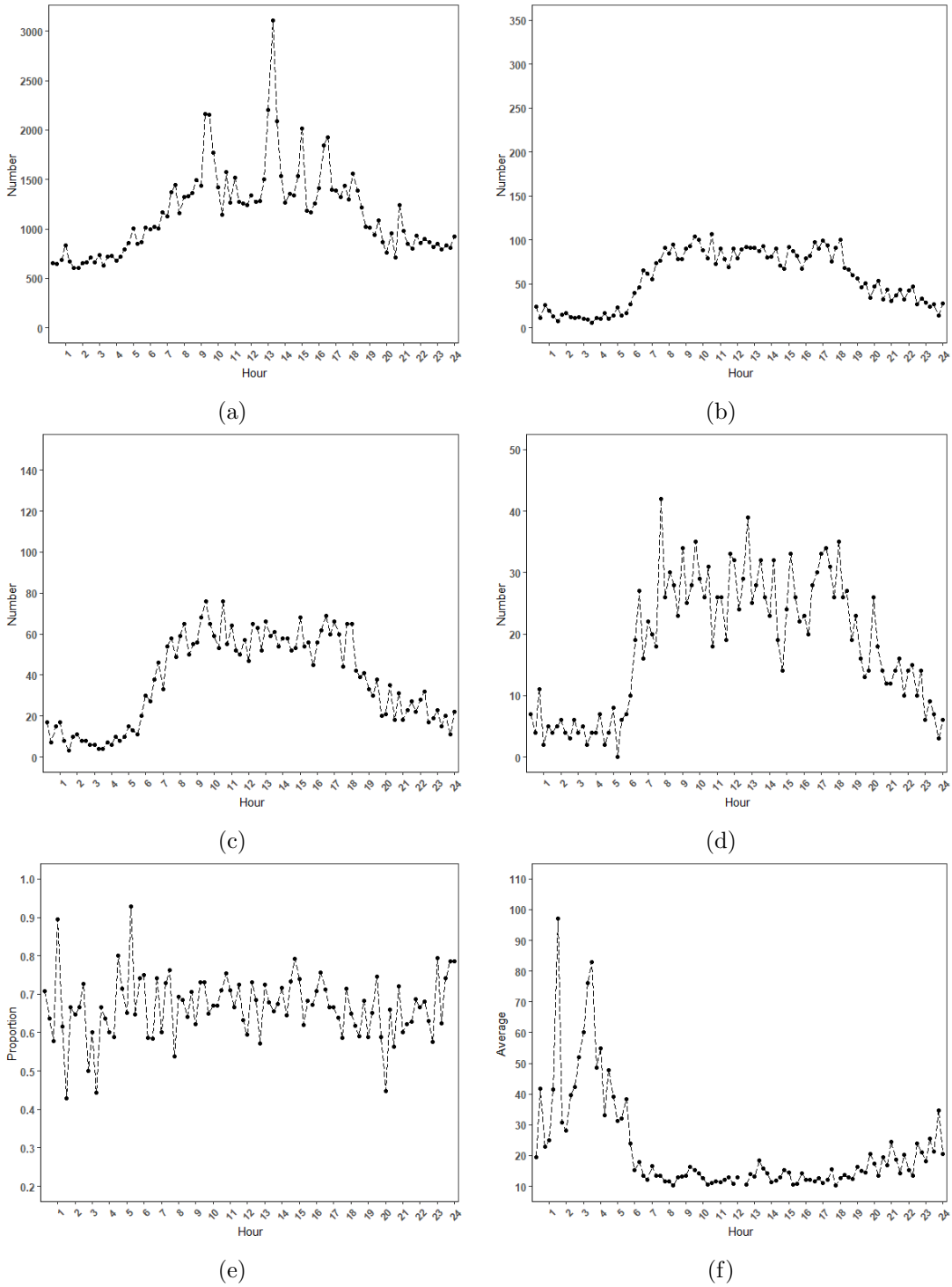
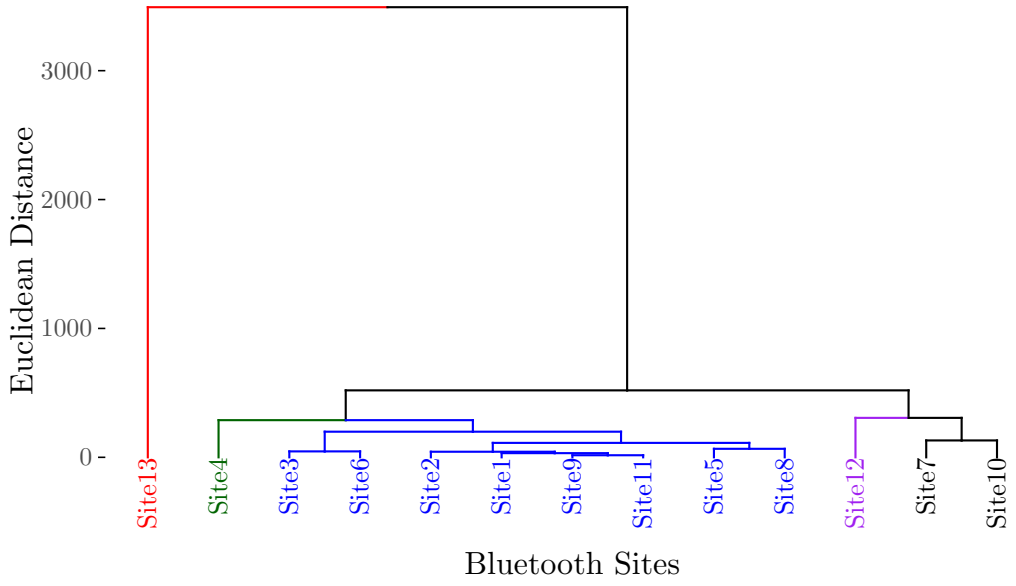
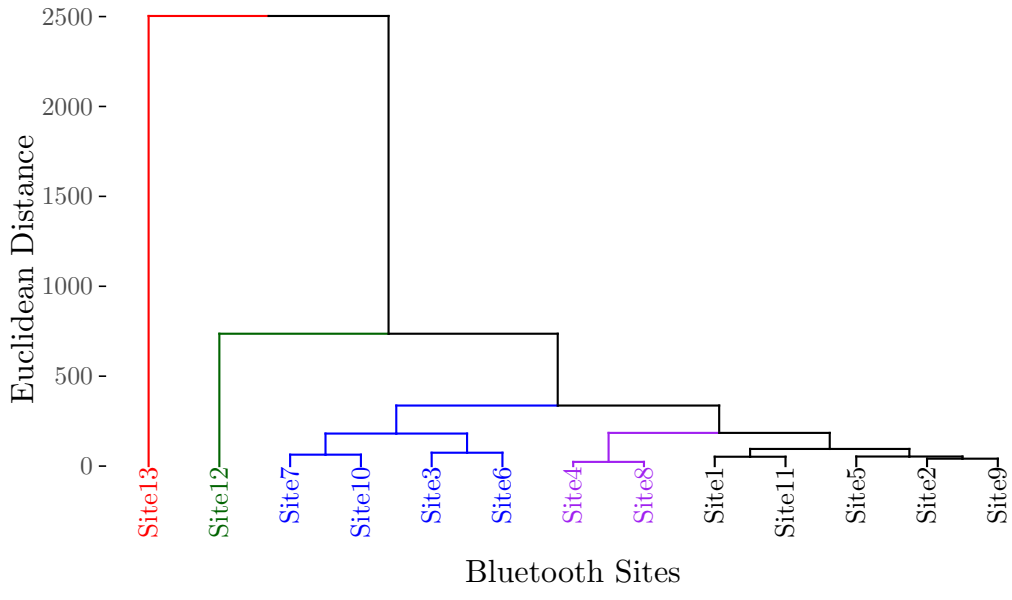


Figure A.14: Hourly patterns for the first four considered variables: (a) the number of all recorded of MAC addresses detected, (b) the number of all unique MAC addresses detected, (c) the number of MAC addresses with multiple detections, (d) the number of MAC addresses with only one detection, (e) the proportion of MAC addresses with multiple detections, and (f) the average number of detections for the MAC addresses with multiple detections at Site 13, Monday 11 February, 2019.

A.2 Hierarchical clustering Bluetooth sites



(a)



(b)

Figure A.15: The dendrogram of Bluetooth site clustering based on two variables: i) the total number of Bluetooth detections; and ii) the total number of unique Bluetooth MAC addresses using the average linkage and Euclidean distance between (a) 9:00-10:00 a.m. and (b) 5:00-6:00 p.m., Monday 11th February 2019. The five clusters are represented by different colors.

A.3 Time interval clustering based on gap distribution

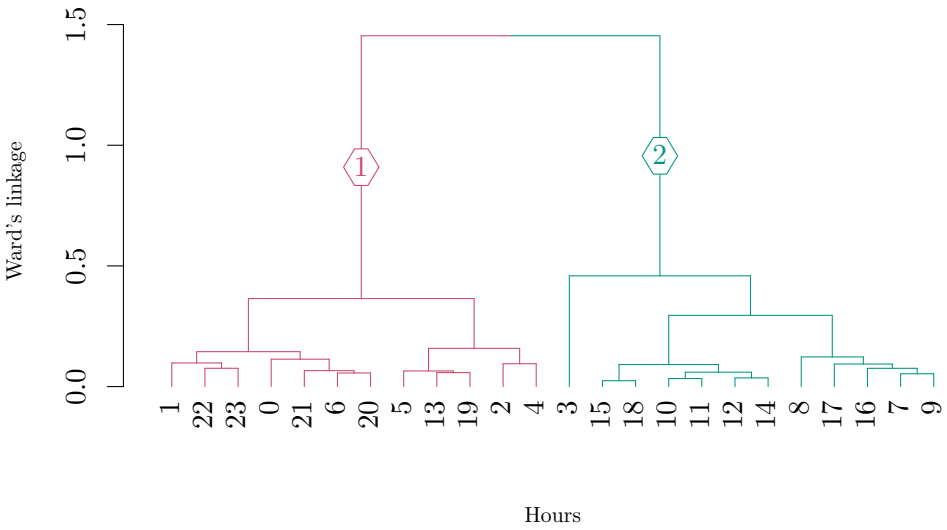


Figure A.16: The dendrogram of time interval clustering based gap distribution using Ward linkage and KS distance at Site 12, Tuesday 12th February 2019.

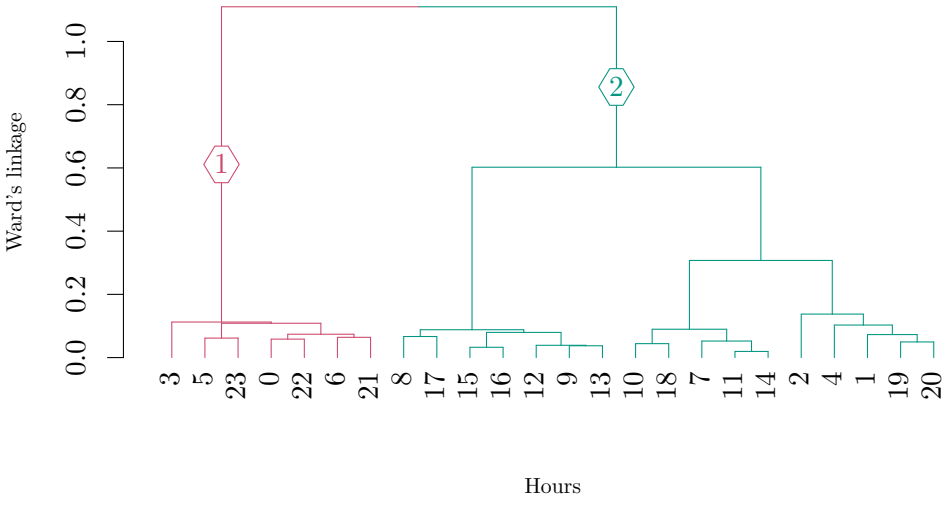


Figure A.17: The dendrogram of time interval clustering based gap distribution using Ward linkage and KS distance at Site 12, Wednesday 13th February 2019.

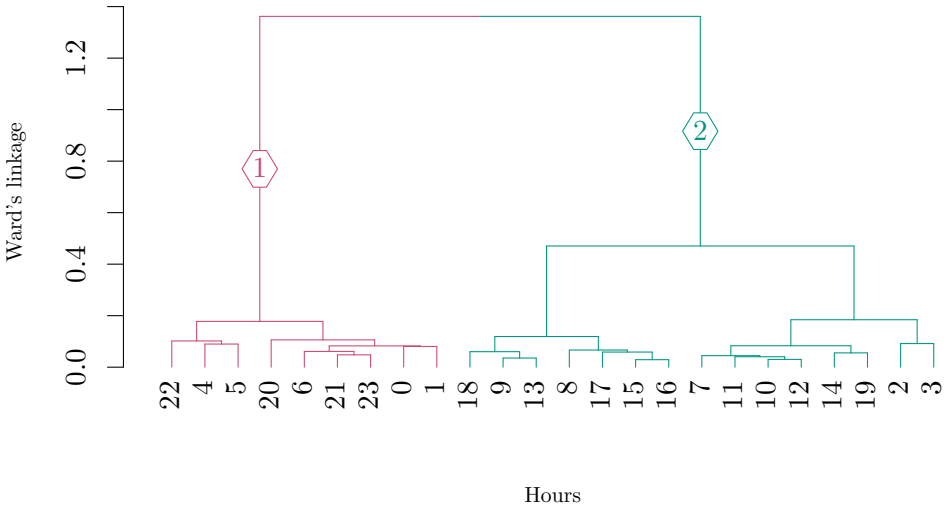


Figure A.18: The dendrogram of time interval clustering based gap distribution using Ward linkage and KS distance at Site 12, Thursday 14th February 2019.

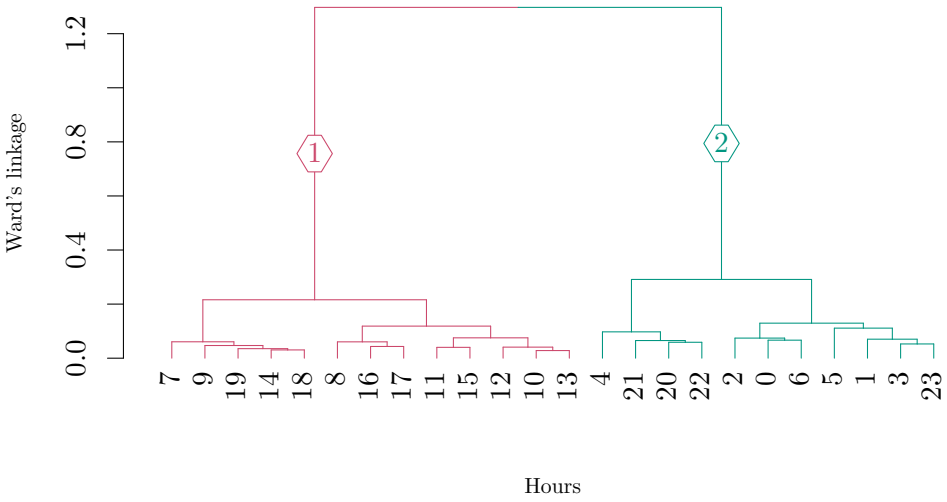


Figure A.19: The dendrogram of time interval clustering based gap distribution using Ward linkage and KS distance at Site 12, Friday 15th February 2019.

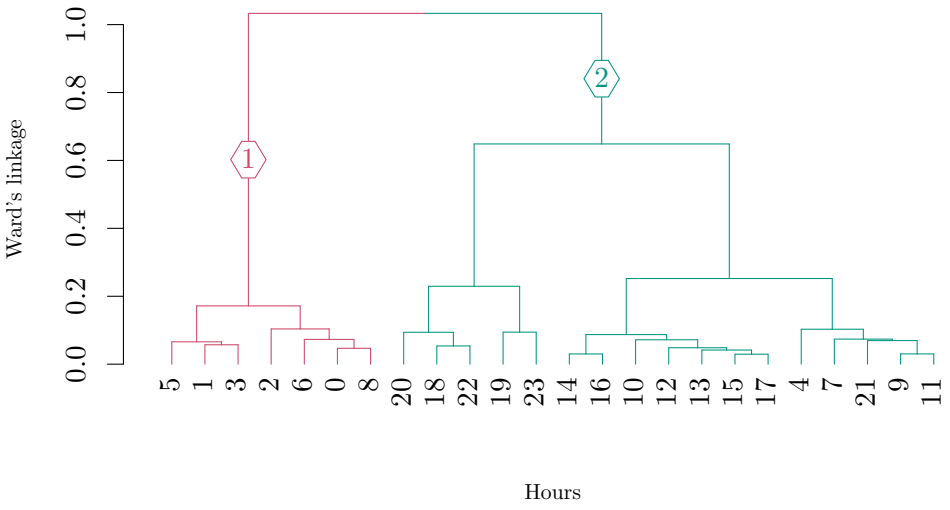


Figure A.20: The dendrogram of time interval clustering based gap distribution using Ward linkage and KS distance at Site 12, Saturday 16th February 2019.

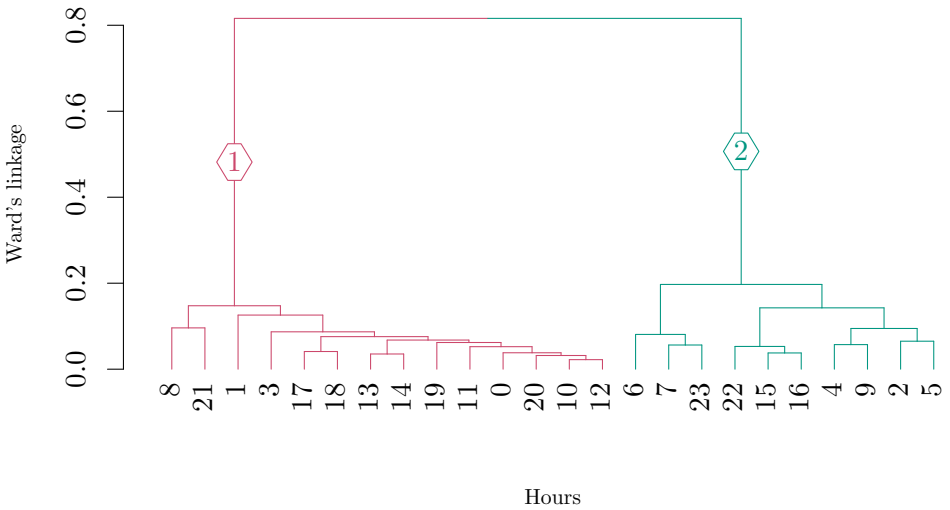


Figure A.21: The dendrogram of time interval clustering based gap distribution using Ward linkage and KS distance at Site 12, Sunday 17th February 2019.

A.4 Results of the weighted regression analysis for the other locations

A.4.1 Results of the weighted regression analysis incorporating buses

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	0.15	0.013	11.595	<2e-16 ***
β	0.22	0.0004	590.648	<2e-16 ***
γ	0.02	0.016	1.017	<3.09e-7 ***

Table A.1: The estimated coefficients of the weighted multiple linear regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 1.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	0.75	0.019	39.457	<2e-16 ***
β	0.17	0.001	148.766	<2e-16 ***
β_1	0.001	0.00001	43.372	<2e-16 ***
γ	0.02	0.015	1.469	<1.42e-7 ***

Table A.2: The estimated coefficients of the weighted multiple quadratic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 1.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	0.99	0.028	35.144	<2e-16 ***
β	0.14	0.003	54.207	<2e-16 ***
β_1	0.001	0.0001	20.678	<2e-16 ***
β_2	-0.000005	0.0000004	-12.776	<2e-16 ***
γ	0.01	0.015	0.674	<5.01e-7 ***

Table A.3: The estimated coefficients of the weighted multiple cubic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 1.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	0.59	0.018	32.80	<2e-16 ***
β	0.2	0.001	235.76	<2e-16 ***
$\Delta\beta_1$	0.06	0.001	45.39	NA
γ	0.02	0.015	1.49	<1.36e-7 ***

Table A.4: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with one knot $c_1 = 52$ at location 1.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	0.56	0.017	33.206	<2e-16 ***
β	0.2	0.00077	278.993	<2e-16 ***
$\Delta\beta_1$	0.087	0.0027	40.263	NA
$\Delta\beta_2$	-0.15	0.0087	-18.776	NA
γ	0.01	0.0157	0.911	<3.62e-7 ***

Table A.5: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with two knots $c_1 = 59$ and $c_2 = 98$ at location 1.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	0.71	0.035	20.521	<2e-16 ***
β	0.18	0.003	58.877	<2e-16 ***
$\Delta\beta_1$	0.02	0.003	6.335	NA
$\Delta\beta_2$	0.08	0.003	32.101	NA
$\Delta\beta_3$	-0.16	0.008	-19.105	NA
γ	0.01	0.015	0.777	<4.37e-7 ***

Table A.6: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with three knots $c_1 = 18$, $c_2 = 62$ and $c_3 = 98$ at location 1.

Model	df	AIC	BIC
Weighted linear	4	565927.3	565965.6
Weighted quadratic	5	564332.4	564380.2
Weighted cubic	6	563512.9	563570.3
Weighted segmented with one knot	6	563844.6	563901.9
Weighted segmented with two knots	8	563279.7	563356.2
Weighted segmented with three knots	10	**562450.1**	**562545.8**

Table A.7: Comparison AIC and BIC between the weighted regression models for the effect of ATC records on Bluetooth detection incorporating buses at location 1.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	0.96	0.019	50.742	<2e-16 ***
β	0.21	0.0003	649.217	<2e-16 ***
γ	0.21	0.043	4.999	<5.76e-7 ***

Table A.8: The estimated coefficients of the weighted multiple linear regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	2.09	0.03	70.080	<2e-16 ***
β	0.16	0.001	138.343	<2e-16 ***
β_1	0.24	0.042	5.739	<9.52e-9 ***
γ	0.0004	0.00001	53.021	<2e-16 ***

Table A.9: The estimated coefficients of the weighted multiple quadratic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	3.55	0.044	80.318	<2e-16 ***
β	0.05	0.002	20.367	<2e-16 ***
β_1	-0.00001	0.0000001	-50.008	<2e-16 ***
β_2	0.19	0.041	4.787	<1.7e-6 ***
γ	0.002	0.00004	58.520	<2e-16 ***

Table A.10: The estimated coefficients of the weighted multiple cubic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	2.15	0.029	73.403	<2e-16 ***
β	0.16	0.0009	176.256	<2e-16 ***
$\Delta\beta_1$	0.08	0.001	65.373	NA
γ	0.23	0.041	5.642	<1.69e-8 ***

Table A.11: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with one knot $c_1 = 63$ at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	2.09	0.026	79.148	<2e-16 ***
β	0.17	0.0007	235.300	<2e-16 ***
$\Delta\beta_1$	0.16	0.003	51.405	NA
$\Delta\beta_2$	-0.18	0.004	-44.124	NA
γ	0.2	0.04	4.902	<9.48e-7 ***

Table A.12: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with two knots $c_1 = 82$ and $c_2 = 124$ at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	2	0.026	76.658	<2e-16 ***
β	0.17	0.0007	241.915	<2e-16 ***
$\Delta\beta_1$	0.16	0.003	46.777	NA
$\Delta\beta_2$	-0.16	0.004	-33.484	NA
$\Delta\beta_3$	-0.08	0.015	-5.014	NA
γ	0.19	0.04	4.728	<2.27e-6 ***

Table A.13: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with three knots $c_1 = 82$, $c_2 = 122$ and $c_3 = 169$ at location 3.

Model	df	AIC	BIC
Weighted linear	4	634788.8	634827.0
Weighted quadratic	5	631568.9	631616.7
Weighted cubic	6	628000.0	628057.4
Weighted segmented with one knot	6	630593.9	630651.2
Weighted segmented with two knots	8	627493.5	627570.0
Weighted segmented with three knots	10	**625009.0**	**625104.6**

Table A.14: Comparison AIC and BIC between the weighted regression models for the effect of ATC records on Bluetooth detection incorporating buses at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	-0.07	0.006	-12.038	<2e-16 ***
β	0.12	0.0002	562.307	<2e-16 ***
γ	0.12	0.015	7.594	<3.13e-14 ***

Table A.15: The estimated coefficients of the weighted multiple linear regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 4.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	0.08	0.008	10.495	<2e-16 ***
β	0.1	0.0006	162.469	<2e-16 ***
β_1	0.14	0.015	8.889	<2e-16 ***
γ	0.0002	0.000005	36.301	<2e-16 ***

Table A.16: The estimated coefficients of the weighted multiple quadratic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 4.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	0.38	0.011	35.173	<2e-16 ***
β	0.05	0.001	46.546	<2e-16 ***
β_1	-0.000005	0.0000001	-49.332	<2e-16 ***
β_2	0.13	0.015	8.746	<2e-16 ***
γ	0.001	0.00002	55.696	<2e-16 ***

Table A.17: The estimated coefficients of the weighted multiple cubic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses at location 4.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	0.19	0.009	19.541	<2e-16 ***
β	0.09	0.0007	121.619	<2e-16 ***
$\Delta\beta_1$	0.04	0.0008	50.488	NA
γ	0.14	0.015	9.632	<2e-16 ***

Table A.18: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with one knot $c_1 = 35$ at location 4.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	0.16	0.0084368	19.099	<2e-16 ***
β	0.1	0.0005223	183.674	<2e-16 ***
$\Delta\beta_1$	0.07	0.0016025	46.208	NA
$\Delta\beta_2$	-0.09	0.0022153	-38.758	NA
γ	0.12	0.014	8.352	<2e-16 ***

Table A.19: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with two knots $c_1 = 54$ and $c_2 = 101$ at location 4.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr($>\|t\|$)**
α	0.19	0.009	19.188	<2e-16 ***
β	0.09	0.0007	118.040	<2e-16 ***
$\Delta\beta_1$	0.02	0.002	8.608	NA
$\Delta\beta_2$	0.06	0.003	18.199	NA
$\Delta\beta_3$	-0.09	0.002	-33.776	NA
γ	0.13	0.014	8.458	<2e-16 ***

Table A.20: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses with three knots $c_1 = 35$, $c_2 = 63$ and $c_3 = 99$ at location 4.

Model	**df**	**AIC**	**BIC**
Weighted linear	4	499965.0	500003.3
Weighted quadratic	5	498148.0	498195.8
Weighted cubic	6	496603.8	496661.2
Weighted segmented with one knot	6	497225.0	497282.4
Weighted segmented with two knots	8	**495413.1**	**495489.6**
Weighted segmented with three knots	10	495481.9	495577.5

Table A.21: Comparison AIC and BIC between the weighted regression models for the effect of ATC records on Bluetooth detection incorporating buses at location 4.

A.4.2 Results of the weighted regression analysis incorporating buses and speed

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	-1.11	0.224	-4.950	<7.45e-7 ***
β	0.56	0.003	175.350	<2e-16 ***
γ	0.28	0.007	35.350	<2e-16 ***
ω	0.03	0.004	8.613	<2e-16 ***
δ	-0.005	0.0001	-89.025	<2e-16 ***

Table A.22: The estimated coefficients of the weighted multiple linear model for the effect of ATC records on Bluetooth detection incorporate buses and speed at location 2.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	1.18	0.252	4.678	< 2.9e-6 ***
β	0.43	0.004	98.374	<2e-16 ***
β_1	0.0005	0.00001	47.551	<2e-16 ***
γ	0.25	0.007	32.785	<2e-16 ***
ω	0.01	0.004	3.337	<8.47e-4 ***
δ	-0.004	0.0001	-61.721	<2e-16 ***

Table A.23: The estimated coefficients of the weighted multiple quadratic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 2.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	2.15	0.266	8.109	<5.14e-16 ***
β	0.28	0.004	57.680	<2e-16 ***
β_1	0.003	0.00005	58.032	<2e-16 ***
β_2	-0.00001	0.0000002	-49.808	<2e-16 ***
γ	0.21	0.007	27.373	<2e-16 ***
ω	0.03	0.004	6.040	<1.54e-9 ***
δ	-0.004	0.0001	-60.233	<2e-16 ***

Table A.24: The estimated coefficients of the weighted multiple cubic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 2.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	0.04	0.236	0.184	0.854
β	0.51	0.003	135.180	<2e-16 ***
$\Delta\beta_1$	0.07	0.002	26.061	NA
γ	0.28	0.008	34.732	<2e-16 ***
ω	0.02	0.004	4.621	<3.82e-6 ***
δ	-0.004	0.00001	-69.864	<2e-16 ***

Table A.25: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with one knot $c_1 = 102$ at location 2.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	1.77	0.258	6.873	<6.31e-12 ***
β	0.41	0.003	107.823	<2e-16 ***
$\Delta\beta_1$	0.05	0.003	15.684	NA
$\Delta\beta_2$	0.18	0.006	30.212	NA
$\Delta\beta_3$	-0.31	0.006	-46.763	NA
γ	0.16	0.007	20.800	<2e-16 ***
ω	0.004	0.004	0.976	0.329
δ	-0.004	0.00001	-56.258	<2e-16 ***

Table A.26: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporate buses and speed with three knots $c_1 = 50$, $c_2 = 88$ and $c_3 = 122$ at location 2.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	1.99	0.113	17.546	<2e-16 ***
β	0.36	0.002	141.463	<2e-16 ***
γ	0.04	0.015	2.951	0.00317 **
ω	-0.031	0.002	-11.842	<2e-16 ***
δ	-0.004	0.00001	-58.217	<2e-16 ***

Table A.27: The estimated coefficients of the weighted multiple linear regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 1.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	2.02	0.122	16.584	<2e-16 ***
β	0.36	0.004	79.265	<2e-16 ***
β_1	0.00001	0.00001	0.870	0.38446
γ	0.04	0.015	2.901	0.00372 **
ω	-0.032	0.002	-11.555	<2e-16 ***
δ	-0.004	0.00001	-42.707	<2e-16 ***

Table A.28: The estimated coefficients of the weighted multiple quadratic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 1.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	2.18	0.125	17.338	<2e-16 ***
β	0.32	0.004	67.804	<2e-16 ***
β_1	0.001	0.00001	17.921	<2e-16 ***
β_2	-0.000006	0.00001	-18.443	<2e-16 ***
γ	0.06	0.015	4.271	<1.95e-5 ***
ω	0.027	0.002	9.630	<2e-16 ***
δ	-0.004	0.00001	-45.221	<2e-16 ***

Table A.29: The estimated coefficients of the weighted multiple cubic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 1.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	1.68	0.113	14.825	<2e-16 ***
β	0.39	0.002	129.612	<2e-16 ***
$\Delta\beta_1$	-0.13	0.008	-16.929	NA
γ	0.06	0.01503873	4.550	<5.38e-6 ***
ω	-0.02	0.002	-9.089	<2e-16 ***
δ	-0.004	0.00001	-59.509	<2e-16 ***

Table A.30: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with one knot $c_1 = 99$ at location 1.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	2.05	0.12	17.023	<2e-16 ***
β	0.35	0.003	93.064	<2e-16 ***
$\Delta\beta_1$	0.03	0.002	13.160	NA
$\Delta\beta_2$	-0.15	0.006	-22.667	NA
γ	0.06	0.015	4.224	<2.4e-5 ***
ω	-0.03	0.002	-11.049	<2e-16 ***
δ	-0.004	0.00001	-43.479	<2e-16 ***

Table A.31: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with two knots $c_1 = 59$ and $c_2 = 95$ at location 1.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	2.05	0.12	16.995	<2e-16 ***
β	0.35	0.003	93.157	<2e-16 ***
$\Delta\beta_1$	0.03	0.003	12.449	NA
$\Delta\beta_2$	-0.11	0.011	-9.144	NA
$\Delta\beta_3$	-0.07	0.018	-3.823	NA
γ	0.06	0.015	4.262	<2.03e-5 ***
ω	-0.03	0.002	-11.013	<2e-16 ***
δ	-0.004	0.00001	-43.458	<2e-16 ***

Table A.32: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with three knots $c_1 = 60$, $c_2 = 92$ and $c_3 = 105$ at location 1.

Model	df	AIC	BIC
Weighted linear	6	559580.9	559638.2
Weighted quadratic	7	559473.5	559540.5
Weighted cubic	8	559065.0	559141.5
Weighted segmented with one knot	8	559073.3	559149.8
Weighted segmented with two knots	10	558691.2	558796.8
Weighted segmented with three knots	12	**558677.3**	**558792.1**

Table A.33: Comparison AIC and BIC between the weighted regression models for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 1.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	5.27	0.136	38.744	<2e-16 ***
β	0.36	0.002	154.552	<2e-16 ***
γ	0.08	0.039	1.969	0.049 *
ω	-0.07	0.002	-26.352	<2e-16 ***
δ	-0.004	0.00001	-68.590	<2e-16 ***

Table A.34: The estimated coefficients of the weighted multiple linear regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 3.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	5.55	0.142	39.017	<2e-16 ***
β	0.34	0.003	99.359	<2e-16 ***
β_1	0.00008	0.000008	10.721	<2e-16 ***
γ	0.09	0.039	2.279	0.0226 *
ω	-0.08	0.003	-26.266	<2e-16 ***
δ	-0.004	0.00001	-55.625	<2e-16 ***

Table A.35: The estimated coefficients of the weighted multiple quadratic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 3.

Coefficients	**Estimate**	**Std. Error**	***t* value**	**Pr(>|t|)**
α	5.75	0.146	39.291	<2e-16 ***
β	0.26	0.004	65.799	<2e-16 ***
β_1	0.001	0.00003	38.601	<2e-16 ***
β_2	-0.000005	0.0000001	-37.519	<2e-16 ***
γ	0.06	0.039	1.675	0.0939
ω	-0.06	0.003	-18.319	<2e-16 ***
δ	-0.004	0.00006	-57.464	<2e-16 ***

Table A.36: The estimated coefficients of the weighted multiple cubic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	5.92	0.143	41.262	<2e-16 ***
β	0.32	0.003	101.409	<2e-16 ***
$\Delta\beta_1$	0.036	0.001	25.919	NA
γ	0.1	0.039	2.573	0.0101 *
ω	-0.08	0.003	-27.059	<2e-16 ***
δ	-0.003	0.00006	-48.147	<2e-16 ***

Table A.37: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with one knot $c_1 = 67$ at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	5.64	0.142	39.575	<2e-16 ***
β	0.32	0.003	106.665	<2e-16 ***
$\Delta\beta_1$	0.11	0.003	31.250	NA
$\Delta\beta_2$	-0.18	0.004	-40.731	NA
γ	0.07	0.038	1.778	0.0754 .
ω	-0.07	0.003	-24.782	<2e-16 ***
δ	-0.003	0.00006	-50.714	<2e-16 ***

Table A.38: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with two knots $c_1 = 87$ and $c_2 = 125$ at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	5.53	0.142	38.803	<2e-16 ***
β	0.34	0.005	66.969	<2e-16 ***
$\Delta\beta_1$	-0.03	0.004	-6.277	NA
$\Delta\beta_2$	0.12	0.003	34.065	NA
$\Delta\beta_3$	-0.18	0.004	-42.462	NA
γ	0.07	0.038	1.830	0.0673 .
ω	-0.08	0.003	-25.846	<2e-16 ***
δ	-0.003	0.00006	-49.255	<2e-16 ***

Table A.39: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with three knots $c_1 = 22$, $c_2 = 85$ and $c_3 = 126$ at location 3.

Model	df	AIC	BIC
Weighted linear	6	620834.6	620892.0
Weighted quadratic	7	620816.3	620883.2
Weighted cubic	8	619440.9	620241.7
Weighted segmented with one knot	8	620165.2	622210.0
Weighted segmented with two knots	10	**617939.9**	**618035.5**
Weighted segmented with three knots	12	617967.8	618082.5

Table A.40: Comparison AIC and BIC between the weighted regression models for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 3.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	0.009	0.043	0.204	0.83817
β	0.12	0.001	157.422	<2e-16 ***
γ	0.078	0.015	5.170	<2.35e-7 ***
ω	0.002	0.0007	2.932	0.00336 **
δ	-0.002	0.00002	-64.264	<2e-16 ***

Table A.41: The estimated coefficients of the weighted multiple linear regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 4.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	-0.05	0.045	-1.019	0.308135
β	0.21	0.002	96.178	<2e-16 ***
β_1	-0.00003	0.000006	-4.630	<3.659e-6 ***
γ	0.07	0.015	4.921	8.62e-7 ***
ω	0.003	0.0007	3.676	2.37e-4***
δ	-0.002	0.00003	-52.585	<2e-16 ***

Table A.42: The estimated coefficients of the weighted multiple quadratic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 4.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	0.16	0.049	3.335	<8.53e-4 ***
β	0.16	0.002	66.220	<2e-16 ***
β_1	0.0009	0.00002	40.330	<2e-16 ***
β_2	-0.000004	0.0000001	-43.615	<2e-16 ***
γ	0.07	0.0147	4.924	<8.48e-7***
ω	0.004	0.0008	4.602	< 4.190e-6***
δ	-0.002	0.00003	-49.765	<2e-16 ***

Table A.43: The estimated coefficients of the weighted multiple cubic regression estimation for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 4.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	-0.19	0.043	-4.364	<1.28e-5 ***
β	0.22	0.001	151.998	<2e-16 ***
$\Delta\beta_1$	-0.078	0.002	-28.629	NA
γ	0.05	0.014	3.339	<8.42e-4 ***
ω	0.005	0.0007	7.311	< 2.68e-13***
δ	-0.002	0.00002	-71.477	<2e-16 ***

Table A.44: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with one knot $c_1 = 114$ at location 4.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	0.09	0.046	1.850	0.06428 .
β	0.18	0.001	97.761	<2e-16 ***
$\Delta\beta_1$	0.05	0.002	25.988	NA
$\Delta\beta_2$	-0.01	0.002	-40.402	NA
γ	0.07	0.014	4.623	<3.79e-6 ***
ω	0.002	0.0007	2.988	< 2.81e-3***
δ	-0.001	0.00003	-48.990	<2e-16 ***

Table A.45: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with two knots $c_1 = 59$ and $c_2 = 99$ at location 4.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
α	0.08	0.046	1.750	0.08014
β	0.19	0.001	98.617	<2e-16 ***
$\Delta\beta_1$	0.06	0.003	20.395	NA
$\Delta\beta_2$	-0.08	0.004	-17.682	NA
$\Delta\beta_3$	-0.04	0.004	-8.270	NA
γ	0.06	0.014	4.506	<6.63e-6 ***
ω	0.002	0.0007	3.073	< 2.12e-3***
δ	-0.001	0.00003	-49.177	<2e-16 ***

Table A.46: The estimated coefficients of the weighted segmented model for the effect of ATC records on Bluetooth detection incorporating buses and speed with three knots $c_1 = 62$, $c_2 = 93$ and $c_3 = 120$ at location 4.

Model	df	AIC	BIC
Weighted linear	6	494644.2	494701.5
Weighted quadratic	7	494515.8	494582.8
Weighted cubic	8	493336.1	493412.6
Weighted segmented with one knot	8	493233.2	493309.7
Weighted segmented with two knots	10	492371.3	492466.9
Weighted segmented with three knots	12	**492298.9**	**492413.7**

Table A.47: Comparison AIC and BIC between the weighted regression models for the effect of ATC records on Bluetooth detection incorporating buses and speed at location 4.

Appendix B

R Codes

B.1 Time interval clustering based on gap distribution by Kolmogorov-Smirnov statistic

```
library(chron)
library(lubridate)
library(scales)
library(dplyr)
library(tidyr)
library(rlist)
library(stringr)
library(tidyverse)
library(anytime)
library(reshape2)
library(zoo)
library(magrittr)
library(reshape)
library(arsenal)
library(imputeTS)
library(ggplot2)
library(sfsmisc)
library(mclust)

# Importing csv files
times_OneHours<-read.csv(file="C:\\TimeInterval24Hours.csv")
for (i in 1:length(filescsv))
{assign(namescsv[i], read.csv(paste(pathcsv,filescsv[i], sep = "\\")))}
```

```
# Using it in order to add extra column to each data set to show
#Site name: example Site 1
Mac2site = sapply(1:length(namescsv), function(x){paste0(b[x])})
names(Mac2site) = namescsv
#************************************************
# Function for Making our data set as desire format
MakingDataSet=function(data, siteCode){
 data <- subset(data, select = c(Date, Vehicle.Id))
 colnames(data) <- c("DayTime", "VehicleID")
 data$DayTime=substr(data$DayTime,start=1,stop=19)
 data$DayTime <- strptime(data$DayTime,"%Y-%m-%d %H:%M:%S")
 data$DayTime <- as.POSIXct(data$DayTime)
 data$Code = siteCode
 return(distinct_data <- dplyr::distinct(data))
}
# Function for spliting data set to desired time interval
time_interval <- function(data,start,end){
data=data[which(data$DayTime>=start& data$DayTime<=end ),]
return(distinct_data <- dplyr::distinct(data))}
#************************************************
# Applying function on each data set
for(k in namescsv){assign(k, MakingDataSet(get(k), Mac2site[k]))}
# Empty list for saving our results
Result_list<-vector("list", length =nrow(times_OneHours) )
ResultAsKS<-vector("list", length =nrow(times_OneHours) )
#************************************************
files="Site12"
check_date<-unique(substr(Site12$DayTime,start=1,stop=10))
date_day<-c("2019-02-11","2019-02-12","2019-02-13","2019-02-14",
            "2019-02-15","2019-02-16","2019-02-17")
date_name<-c("Monday","Tuesday","Wednesday","Thursday","Friday",
             "Saturday","Sunday")
B<-data.frame(date_day,date_name)
# First assign  each data set for computing results
data_set=assign(files, get(files))
data_set_test<-data_set %>%filter(str_detect(data_set$DayTime, check_date[ff]))

for (z in 1:nrow(times_OneHours)){
```

```
temp_data_set<-time_interval(data_set_test,
            paste(check_date[ff],"",times_OneHours[z,1]),
            paste(check_date[ff],"",times_OneHours[z,2]))
#***********************************************
# Order data frame based on VehicleID
temp_data_set= temp_data_set[order(temp_data_set$VehicleID),]
# Computing Gap time diff and adding new column as gap
temp_data_set_gap<-temp_data_set%>% group_by(VehicleID) %>%
mutate(gap=DayTime-lag(DayTime, default = first(DayTime)))
#***********************************************
#  Extracting IDs name from data set
# Number Of All Unique IDs
ID_name=unique(temp_data_set$VehicleID)
Result<-data.frame( VehicleID=character(),
Time=character(),
Number_Group=numeric(),
size=numeric(),
Status=character(),
Gap_dist=numeric(),
Gap_mean=numeric(),
Gap_std=numeric(),
Gap_median=numeric(),
stringsAsFactors=FALSE)
for (j in 1:length(ID_name)){
# storing information for each ID and applying computation
temp=temp_data_set_gap[which(temp_data_set_gap$VehicleID==as.character(ID_name[j] )
# How many detection has an ID: if it has just one detection,
# so definitely it has one group
Size<-nrow(temp) # Size:number of all detection
if (Size!=1) { # N_initial: number of group
N_initial=1
for(k in 2:Size){
if(temp$gap[k]>10){
N_initial=N_initial+1
}
}
} else  {N_initial=1} # if Size=1 then N_initial=1
if (N_initial ==1 & Size!=1 ){ #Unique Group
```

```
temp_one<-data.frame("VehicleID"=as.character(unique(temp$VehicleID)),
"Time"=paste(times_OneHours[d,1], "-", times_OneHours[d,2]),
"Number_Group"=N_initial,
"size"=Size,
"Status"="Unique",
"Gap_dist"=paste(c(temp$gap), collapse = ", "),
"Gap_mean"=mean(as.numeric(temp$gap[2:nrow(temp)])),
"Gap_std"=sd(as.numeric(temp$gap[2:nrow(temp)])),
"Gap_median"=median(as.numeric(temp$gap[2:nrow(temp)]))
)

Result<-rbind(Result,temp_one)
} else if(N_initial ==1 & Size==1){ #Unique Singelton
temp_one<-data.frame("VehicleID"=as.character(unique(temp$VehicleID)),
"Time"=paste(times_OneHours[d,1], "-", times_OneHours[d,2]),
"Number_Group"=N_initial,
"size"=Size,
"Status"="Unique Singelton",
"Gap_dist"="0",
"Gap_mean"=0,
"Gap_std"=0,
"Gap_median"=0)

Result<-rbind(Result,temp_one)
}else if ( N_initial==Size & N_initial==2){
temp_one<-data.frame("VehicleID"=as.character(unique(temp$VehicleID)),
"Time"=paste(times_OneHours[d,1], "-", times_OneHours[d,2]),
"Number_Group"=N_initial,
"size"=Size,
"Status"="2 Singelton ",
"Gap_dist"=paste(c(temp$gap), collapse = ", "),
"Gap_mean"=mean(as.numeric(temp$gap[2:nrow(temp)])),
"Gap_std"=sd(as.numeric(temp$gap[2:nrow(temp)])),
"Gap_median"=median(as.numeric(temp$gap[2:nrow(temp)])))
Result<-rbind(Result,temp_one)
}else if (N_initial==Size){
temp_one<-data.frame("VehicleID"=as.character(unique(temp$VehicleID)),
"Time"=paste(times_OneHours[d,1], "-", times_OneHours[d,2]),
```

```
"Number_Group"=N_initial,
"size"=Size,
"Status"=paste0(N_initial," Singelton "),
"Gap_dist"=paste(c(temp$gap), collapse = ", "),
"Gap_mean"=mean(as.numeric(temp$gap[2:nrow(temp)])),
"Gap_std"=sd(as.numeric(temp$gap[2:nrow(temp)])),
"Gap_median"=median(as.numeric(temp$gap[2:nrow(temp)])))
Result<-rbind(Result,temp_one)
}else if(N_initial!=Size){ # Multiple group

grouped = c(1) # Assigning number to each group
for(t in 2:nrow(temp)){
grouped = c(grouped, ifelse(temp$gap[t]<=10, grouped[t-1], grouped[t-1]+1))
}
# adding one column at end of temp to show group
temp<-temp %>% mutate(grouped=grouped)
groupedlist_1 <- numeric(length = length(unique(grouped)))
for(q in unique(grouped)){
groupedlist_1[q]=nrow(temp[which(temp$grouped==q),])
}
temp_one<-data.frame("VehicleID"=as.character(unique(temp$VehicleID)),
"Time"=paste(times_OneHours[d,1], "-", times_OneHours[d,2]),
"Number_Group"=N_initial,
"size"=Size,
"Status"=paste(c(groupedlist_1, "Multiple"), collapse = ", "),
"Gap_dist"=paste(c(temp$gap), collapse = ", "),
"Gap_mean"=mean(as.numeric(temp$gap[2:nrow(temp)])),
"Gap_std"=sd(as.numeric(temp$gap[2:nrow(temp)])),
"Gap_median"=median(as.numeric(temp$gap[2:nrow(temp)])))
Result<-rbind(Result,temp_one)
}
Result_list[[z]] <- Result
}
test<-Result_list[[z]]
toplot = lapply(1:nrow(test), function(i){
if(test$Status[i] != "Unique Singelton"){
x = strsplit(as.character(test$Gap_dist[i]), ", ", fixed = T)[[1]]
x = as.numeric(x[x!="0"])
```

```
res = data.frame(VehicleID=test$VehicleID[i],Gap_dist = x, Time = test$Time[i])
} else{
x = as.character(test$Gap_dist[i])[[1]]
x = as.numeric(x)
res = data.frame(VehicleID=test$VehicleID[i],Gap_dist = x,
                  Time = test$Time[i])}})
res = do.call(rbind, toplot)
ResultAsKS[[z]]<-res
}
# End of first for loop on z
OverallReport_Site12<-do.call("rbind", Result_list)
#***********************************************
#***********Clustering  Section *****************
S=nrow(times_OneHours)
ss<-times_OneHours %>% unite("Time", start, end, sep=" - ")
time_inter<-data.frame(Time=character())
label<-c("0  ","1 ","2   ","3   ","4   ","5   ","6   ","7   ","8   ","9   ",
"10   ","11   ","12   ","13   ","14   ","15   ","16   ","17   ","18   ","19   ",
"20   ","21   ","22   ","23   ")
time_inter <- data.frame(label,ss)
#Need it if we want details of each cluster
colnames(time_inter)=c("Time","Time1")

# Empty matrix for saving KS results
KS_Result=as.data.frame(matrix(0, ncol = S, nrow = S))
rownames(KS_Result)=time_inter$Time
colnames(KS_Result)=time_inter$Time

for (h in 1:(S-1)){
sample<-ResultAsKS[[h]]$Gap_dist
for(z1 in (h+1):(S)){
sample1<-ResultAsKS[[z1]]$Gap_dist
H=ks.test(sample,sample1)
KS_Result[z1,h]<-H$statistic
KS_Result[h,z1]<-H$statistic
}}
# Turns KS distance matrix into a distance object that R recognises:
ks_dist = as.dist(KS_Result)
```

```
hc=hclust(ks_dist, method="ward.D")
```

B.2 Non-parametric variance function estimation in regression models including bus

```
library(chron)
library(lubridate)
library(scales)
library(dplyr)
library(tidyr)
library(rlist)
library(stringr)
library(tidyverse)
library(anytime)
library(reshape2)
library(zoo)
library(magrittr)
library(arsenal)
library(data.table)
library(imputeTS)
library(ggplot2)
library(sfsmisc)
library(e1071)
library(nlme)
library(segmented)
All_Data_BUS<-read.csv(file="C:\\Users\\LocationTwo\\BT_BUS.csv")
All_Data_BUS<-All_Data_BUS %>% select(-c(X))
colnames(All_Data_BUS)<-c("BT", "Bus.count","ATC","Bus.prop","day","time")
test<-All_Data_BUS
new_data<-data.frame(residual=numeric,fit=numeric())
stop<-10
t<-1
#************ Linear model **********************
# Weight vector
W<-rep(1,length = 105120)
while(t<=stop){
```

```
# linear model
Mod_lm <- lm(BT ~ ATC+Bus.count, data =test,weights = W)
new_data<-as.data.frame(cbind(abs(residuals(Mod_lm)),fitted(Mod_lm)))
colnames(new_data)<-c("residual","fit")
# loess smoothing
sp<-0.9
lowess_values <- loess(residual ~ fit, data=new_data.lin, span=sp)
phat <- predict(lowess_values)
new_data<- as.data.frame(cbind(new_data,phat))
std_est<-exp(new_data.lin$phat)
var_est<-exp(2*new_data.lin$phat)
Normal.W<-var_est/mean(var_est)
W<-1/Normal.W
t<-t+1
}

#************ Quadratic model ***********************
# Weight vector
W1<-rep(1,length = 105120)
while(t<=stop){
#Quadratic model
Mod_Qua <- lm(BT ~ ATC+Bus.count+I(ATC^2), data =test,weights = W1)
summary(Mod_Qua)
new_data<-as.data.frame(cbind(log(abs(residuals(Mod_Qua))),fitted(Mod_Qua)))
colnames(new_data)<-c("residual","fit")
# loess smoothing
sp<-0.9
lowess_values <- loess(residual ~ fit, data=new_data, span=sp)
phat <- predict(lowess_values)
new_data <- as.data.frame(cbind(new_data,phat))
std_est<-exp(new_data$phat)
var_est1<-exp(2*new_data$phat)
Normal.W<-var_est1/mean(var_est1)
W1<-1/Normal.W
t<-t+1
}
#************ Cubic Model ***********************
# Weight vector
```

```
W2<-rep(1,length = 105120)
while(t<=stop){
# Cubic Model
Mod_cub <- lm(BT ~ ATC+Bus.count+I(ATC^2)+I(ATC^3), data =test,weights = W2)
summary(Mod_cub)
new_data<-as.data.frame(cbind(log(abs(residuals(Mod_cub))),fitted(Mod_cub)))
colnames(new_data)<-c("residual","fit")
# loess smoothing
sp<-0.9
lowess_values <- loess(residual ~ fit, data=new_data, span=sp)
phat <- predict(lowess_values)
new_data <- as.data.frame(cbind(new_data,phat))
std_est<-exp(new_data$phat)
var_est2<-exp(2*new_data$phat)
Normal.W<-var_est2/mean(var_est2)
W2<-1/Normal.W
t<-t+1
}
############ Segmented  Model with one knot **********************
# Weight vector
W3<-rep(1,length = 105120)
while(t<=stop){
# Segmented  Model with one knot
Mod_lin <- lm(BT ~ ATC+Bus.count, data =test,weights = W3)
# Building the segmented Regression Model
# create a figure to get an idea of the data
### have to provide estimates for breakpoints.
my.seg <- segmented(Mod_lin,
seg.Z = ~ ATC,
npsi=1)
new_data<-as.data.frame(cbind(log(abs(residuals(my.seg))),fitted(my.seg)))
colnames(new_data)<-c("residual","fit")
# loess smoothing
sp<-0.9
lowess_values <- loess(residual ~ fit, data=new_data, span=sp)
phat <- predict(lowess_values)
new_data <- as.data.frame(cbind(new_data,phat))
std_est<-exp(new_data$phat)
```

```
var_est3<-exp(2*new_data$phat)
Normal.W<-var_est3/mean(var_est3)
W3<-1/Normal.W
t<-t+1
}
#************ Segmented  Model with two knots ***********************
# Weight vector
W4<-rep(1,length = 105120)
while(t<=stop){
Mod_lin.1 <- lm(BT~ATC+Bus.count, data =test,weights = W4)
my.seg.1 <- segmented(Mod_lin.1,
seg.Z = ~ ATC,
npsi=2)
new_data<-as.data.frame(cbind(log(abs(residuals(my.seg.1))),fitted(my.seg.1)))
colnames(new_data)<-c("residual","fit")
# loess smoothing
sp<-0.9
lowess_values <- loess(residual ~ fit, data=new_data, span=sp)
phat <- predict(lowess_values)
new_data <- as.data.frame(cbind(new_data,phat))
std_est<-exp(new_data$phat)
var_est4<-exp(2*new_data$phat)
Normal.W<-var_est4/mean(var_est4)
W4<-1/Normal.W
t<-t+1
}
#************ Segmented  Model with three knots ***********************
# Weight vector
W5<-rep(1,length = 105120)
while(t<=stop){
Mod_lin.2 <- lm(BT~ATC+Bus.count, data =test,weights = W4)
my.seg.2 <- segmented(Mod_lin.2,
seg.Z = ~ ATC,
npsi=3)
new_data<-as.data.frame(cbind(log(abs(residuals(my.seg.2))),fitted(my.seg.2)))
colnames(new_data)<-c("residual","fit")
# loess smoothing
sp<-0.9
```

```
lowess_values <- loess(residual ~ fit, data=new_data, span=sp)
phat <- predict(lowess_values)
new_data <- as.data.frame(cbind(new_data,phat))
std_est<-exp(new_data$phat)
var_est5<-exp(2*new_data$phat)
Normal.W<-var_est5/mean(var_est5)
W5<-1/Normal.W
t<-t+1
}

#*****************************************************
ggplot(new_data,aes(fit, residual)) +
xlab("Fitted") +
ylab("log(abs(residuals))")+
#   geom_point() +
geom_line(aes(y = phat),color = "red")+ theme_bw() +
theme( panel.grid.major = element_blank(),
panel.grid.minor = element_blank(),
axis.line = element_line(colour = "black"))
+ylim(c(0, 3))
#*****************************************************
# Comparision between models
AIC(Mod_lin,Mod_Qua,Mod_cub,my.seg,my.seg.1,my.seg.2)
BIC(Mod_lin,Mod_Qua,Mod_cub,my.seg,my.seg.1,my.seg.2)
```

B.3 Poisson regression model with Fourier basis

```
library(tidyverse)
library(anytime)
library(lubridate)
library(TSA)
library(janitor)
library(broom)

All_Data_BUS<-read.csv(file="C:\\Users\\BT_BUS.csv")
colnames(All_Data_BUS)<-c("BT", "Bus.count","ATC","day","time")
# Adding another variable as time of day and Type of day ( weekdays vs weekend)
```

```
f=as.POSIXct(All_Data_BUS$time,tz="Europe/London", origin="1970-01-01")
Hour<-hour(f)
mon<-month(f)
DAY<-c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday")
test <- All_Data_BUS %>%
mutate(Hour = factor(Hour))%>%
mutate(DayWeek = factor(weekdays(as.Date(day)),levels=DAY)) %>% tibble()
#************************************************
# The number of observations in one week: 2016=12*24*7
BT.ts <-ts(test$BT,frequency=2016)

# Creating harmonic basis functions
fourier<-harmonic(BT.ts,m=43) .

# We use clean_names() from the R package janitor to get
# a clean tibble of Fourier

fourierbase<-clean_names(as.data.frame(fourier))
data<-cbind(test,fourierbase) %>% tibble()

#The above codes simply generate the desired data set with
# only BT, ATC and Fourier bases.
# Linear model first to get the initial values for Poisson regression.

fourier.lm <- lm(BT~ATC+ATC:(.-ATC-BT),data=data)
fourier.glm<-glm(BT~ATC+ATC:(.-ATC-BT),data=data,
                 family=poisson(link="identity")
                 ,start=fourier.lm$coefficients)
```

B.4 Poisson regression model with the periodic B-spline basis

```
All_Data_BUS<-read.csv(file="C:\\Users\\BT_BUS.csv")
colnames(All_Data_BUS)<-c("BT", "Bus.count","ATC","day","time")
# Adding another variable as time of day and Type of day ( weekdays vs weekend)
f =as.POSIXct(All_Data_BUS$time,tz="Europe/London", origin="1970-01-01")
Hour<-hour(f)
```

```
mon<-month(f)
DAY<-c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday")
test <- All_Data_BUS %>%
       mutate(Hour = factor(Hour))%>%
       mutate(DayWeek = factor(weekdays(as.Date(day)),levels=DAY)) %>%
       tibble()
#************************************************
# This part generates the periodic B-spline basis for one week
spline.base<-pbs::pbs(1:2016,df=N,Boundary.knots=c(0,2016),intercept=FALSE)
spline.design.matrix<-rbind(matrix( rep( t( spline.base ) , 52 ) ,
                      ncol = ncol(spline.base) , byrow = TRUE ),
                      spline.base[1:288,])
spline.data <- cbind(test,spline.design.matrix)
spline.lm <- lm(BT~ATC+ATC:(.-ATC-BT),data=spline.data)
spline.glm <-glm(BT~ATC+ATC:(.-ATC-BT),data=spline.data,
                  family=poisson(link="identity"),
                  start=spline.lm$coefficients)
```

B.5 Calibration with the classic estimator and the profile log-likelihood methods

```
All_Data_BUS<-read.csv(file="C:\\Users\\BT_BUS.csv")
colnames(All_Data_BUS)<-c("BT", "Bus.count","ATC","day","time")

# Adding another variable as time of day and Type of day ( weekdays vs weekend)
f =as.POSIXct(All_Data_BUS$time,tz="Europe/London", origin="1970-01-01")
Hour<-hour(f)
mon<-month(f)

DAY<-c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday")

test <- All_Data_BUS %>%
       mutate(Hour = factor(Hour))%>%
       mutate(DayWeek = factor(weekdays(as.Date(day)),levels=DAY)) %>%
       tibble()

#************************************************
```

```
BT.ts<-ts(test$BT,frequency=2016)
fourier<-harmonic(BT.ts,m=43)

fourier.data <- cbind(test,fourier) %>% tibble()
fourier.lm <- lm(BT~ATC+ATC:(.-ATC-BT),data=fourier.data)
fourier.glm <- glm(BT~ATC+ATC:(.-ATC-BT),data=fourier.data,
                   family=poisson(link="identity"),
                   start=fourier.lm$coefficients)
### Store the glm coefficient as the initial values for the glm calibration.
fourier.glm.initial<-fourier.glm$coefficients

# Computing Over-dispersion
DP<-c_hat(fourier.glm , method = "pearson")
#************************************************
# theta : parameter vector
theta<-fourier.glm$coefficients
# Sigma: Variance-covariance matrix
Sigma<-vcov(fourier.glm)
temp<-as.data.frame(fourier.data)
# vec is a corresponding to that particular row of the Fourier basis of time,
#for example, if I am assuming the first 5 min slot,
#it chooses the first row of  the Fourier basis, for the second 5 min slot,
#it is going to be 2, and etc.

t<-206  # particular t : for example 17:00:00 on Monday
vec<-temp[t,-c(1:2)]
names(vec)<- NULL
# Known Bluetooth
yzero=19

#********************# Delta method **************
#First part for computing SE for x0 using Delta method
# V1=(1,0,...,0)
V1<-c(1,rep(0, 87))
# V2=(0,1,the row of the Fourier basis for the particular time t)
V2<-c(0,1,vec)
V<-rbind(unname(V1),unlist(V2))
```

```
Cov_Mat_1<-(V) %*% Sigma %*% t(V)
Cov_Mat_1
# Extended cov matrix
# yzero considers as variance of e

Ext_Cov_Mat<-cbind(c(yzero,0,0),rbind(0,Cov_Mat_1))
Ext_Cov_Mat

#*******************************************************
# slope for that particular time slot
b1<-t(unlist(V2))%*% theta
b1
#*******************************************************
# intercept for that particular time slot
b0<-fourier.glm$coefficients[1]
#*******************************************************
#*******************************************************
# x0 = (y0-b0-e)/b1
# Estimation unknown ATC
x0<-(yzero-b0)/b1

# Partial derivate : dx0/db0= -1/b1
dx_b0<- -1/b1
# Partial derivate : dx0/db1=  b0-y0/(b1^2)
dx_b1<- (b0-yzero)/(b1^2)
# Partial derivate : dx0/de=  -1/b1
dx_e<- -1/b1
# derivate matrix
Derv_mat<-c(dx_b0,dx_b1,dx_e)

Var_x0<-t(Derv_mat )%*% Ext_Cov_Mat %*% Derv_mat
Var_x0
SE<-sqrt(Var_x0)
# Considering SE with over-dispersion
SE.DP<-SE*sqrt(DP[1])

# we use x0hat +/- 3 SE, this should give us
# a range of trail xo values for the profile log-likelihood
```

```
l<-max(round(x0-3*SE.DP),0)
u<-round(x0+3*SE.DP)
#************************************************
#//////// Profile log-likelihood  ////////////
vec_x0=seq.int(l,u, 1)
res=rep(0,length(vec_x0))

for(i in 1:length(vec_x0)){
### generate the specific data set by add one row
fourier.data.addone <-fourier.data%>%
                  add_row(BT = yzero, ATC = vec_x0[i],temp[t,-c(1:2)])
fourier.glm <- glm(BT~ATC+ATC:(.-ATC-BT),data=fourier.data.addone,
              family=poisson(link="identity"),start=fourier.glm.initial)
res[i] <- (logLik(fourier.glm))

}

result<-data.frame(Estimate.ATC=numeric(),LogLike=numeric())
result<-as.data.frame(cbind(vec_x0,res))

max_point<-result[which(result$res==max(result$res)),]


#range() simply get the upper limit and lower limit of all valid ATC counts.
R1<-range(vec_x0[max(res)-(res)<qchisq(.99,1)*DP[1]/2])
```

Bibliography

Abbott-Jard, M., Shah, H., and Bhaskar, A. (2013). Empirical evaluation of Bluetooth and Wi-Fi scanning for road transport. In *Australasian Transport Research Forum (ATRF), 36th*.

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 94–105.

Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis*, 8(3):261–275.

Araghi, B. N., Christensen, L. T., Krishnan, R., and Lahrmann, H. (2012). Application of Bluetooth technology for mode-specific travel time estimation on arterial roads: Potentials and challenges. In *Proceedings the Annual Transport Conference at Aalborg University*, volume 19, pages 1–15, Denmark.

Araghi, B. N., Hammershøj Olesen, J., Krishnan, R., Tørholm Christensen, L., and Lahrmann, H. (2015). Reliability of Bluetooth technology for travel time estimation. *Journal of Intelligent Transportation Systems*, 19(3):240–255.

Bachmann, C., Roorda, M. J., Abdulhai, B., and Moshiri, B. (2013). Fusing a Bluetooth traffic monitoring system with loop detector data for improved freeway traffic speed estimation. *Journal of Intelligent Transportation Systems*, 17(2):152–164.

Barcelö, J., Montero, L., Marqués, L., and Carmona, C. (2010). Travel time forecasting and dynamic origin-destination estimation for freeways based on Bluetooth traffic monitoring. *Transportation Research Record*, 2175(1):19–27.

Barceló Bugeda, J., Montero Mercadé, L., Bullejos, M., Serch, O., and Carmona Bautista, C. (2012a). Dynamic OD matrix estimation exploiting Bluetooth data in urban networks. In *Proceedings of the 14th international conference on Automatic Control, Modelling & Simulation, and Proceedings of the 11th international conference on Microelectronics, Nanoelectronics, Optoelectronics*, pages 116–121.

Barceló Bugeda, J., Montero Mercadé, L., Bullejos, M., Serch, O., and Carmona Bautista, C. (2012b). A Kalman filter approach for the estimation of time dependent OD matrices exploiting Bluetooth traffic data collection. In *Transportation Research Board 91st Annual Meeting Compendium of Papers DVD*, page 16p, Washington DC, United States. Transportation Research Board.

Berger, V. W. and Zhou, Y. (2014). Kolmogorov–Smirnov test: Overview. *Wiley Statsref: Statistics Reference Online*.

Bhaskar, A. and Chung, E. (2013). Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. *Transportation Research Part C: Emerging Technologies*, 37:42–72.

Bhaskar, A., Kieu, L. M., Qu, M., Nantes, A., Miska, M., and Chung, E. (2013). On the use of Bluetooth MAC scanners for live reporting of the transport network. In *Proceedings of the Eastern Asia Society for Transportation Studies, Volume 9 (The 10th Conference in Taipei)*, pages 1–20. Eastern Asia Society for Transportation Studies (EASTS).

Bhaskar, A., Kieu, L. M., Qu, M., Nantes, A., Miska, M., and Chung, E. (2015). Is bus overrepresented in Bluetooth MAC scanner data? is MAC-ID really unique? *International Journal of Intelligent Transportation Systems Research*, 13(2):119–130.

Bouguila, N. and ElGuebaly, W. (2009). Discrete data clustering using finite mixture models. *Pattern Recognition*, 42(1):33–42.

Brennan Jr, T. M., Ernst, J. M., Day, C. M., Bullock, D. M., Krogmeier, J. V., and Martchouk, M. (2010). Influence of vertical sensor placement on data collection efficiency from Bluetooth MAC address collection devices. *Journal of Transportation Engineering*, 136(12):1104–1109.

Brown, P. (1982). Multivariate calibration. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):287–308.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference – a Practical Information-theoretic Approach*. Springer New York, second edition.

Carpenter, C., Fowler, M., and Adler, T. J. (2012). Generating route-specific Origin–Destination tables using Bluetooth technology. *Transportation Research Record*, 2308(1):96–102.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. CRC Press.

Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):157–176.

Chen, J.-H. and Hung, W.-L. (2015). An automatic clustering algorithm for probability density functions. *Journal of Statistical Computation and Simulation*, 85(15):3047–3063.

Chen, K. and Miles, J. C. (2004). *ITS Handbook 2004: Recommendations from the World Road Association (PIARC)*. Artech House.

Cherchali, A., Gudelis Jr, M. J., Lester, W. G., and McLaughlin, R. J. (2010). Technique for automated MAC address cloning. US Patent 7,787,455.

Chiou, J.-M. and Müller (1999). Nonparametric quasi-likelihood. *The Annals of Statistics*, 27(1):36–64.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Colberg, K., Suh, W., Anderson, J., Zinner, S., Guin, A., Hunter, M., and Guensler, R. (2014). Lane bias issues in work zone travel time measurement and reporting. *Transportation Research Record*, 2458(1):78–87.

Cotten, D., Codjoe, J., and Loker, M. (2020). Evaluating advancements in Bluetooth technology for travel time and segment speed studies. *Transportation Research Record*, 2674(4):193–204.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika*, 70(1):269–274.

Cragg, S. (2013). Bluetooth detection – cheap but challenging. In *Scottish Transport Applications and Research Conference (STAR), Apr. 2013*.

Crawford, F., Watling, D., and Connors, R. (2017). Assessing the feasibility of using Bluetooth data to examine the repeated travel behaviour of road users. *Submitted for Publication*.

Crawford, F., Watling, D. P., and Connors, R. D. (2018). Identifying road user classes based on repeated trip behaviour using Bluetooth data. *Transportation Research Part A: Policy and Practice*, 113:55–74.

Davidson, R. and MacKinnon, J. G. (1993). Estimation and Inference in Econometrics. *Oxford University Press*.

Davis, J. C. and Sampson, R. J. (1986). *Statistics and Data Analysis in Geology*. Wiley New York.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag.

Dean, C. B. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457.

Díaz, J. J. V., González, A. B. R., and Wilby, M. R. (2015). Bluetooth traffic monitoring systems for travel time estimation on freeways. *IEEE Transactions on Intelligent Transportation Systems*, 17(1):123–132.

Diebold, J. (1995). *Transportation Infostructures: the Development of Intelligent Transportation Systems*. Greenwood Publishing Group.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press.

Erkan, I. and Hastemoglu, H. (2016). Bluetooth as a traffic sensor for stream travel time estimation under Bogazici Bosporus conditions in Turkey. *Journal of Modern Transportation*, 24(3):207–214.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. CRC Press.

Fisher, R. A. (1950). The significance of deviations from expectation in a Poisson series. *Biometrics*, 6(1):17–24.

Franssens, A. (2010). Impact of multiple inquires on the Bluetooth discovery process: and its application to localization. Master's thesis, University of Twente.

Friesen, M. R. and McLeod, R. D. (2015). Bluetooth in intelligent transportation systems: a survey. *International Journal of Intelligent Transportation Systems Research*, 13(3):143–153.

Frodigh, M., Johansson, P., and Larsson, P. (2000). Wireless ad hoc networking: the art of networking without a network. *Ericsson Review*, 4(4):249.

Gelman, A. and Hill, J. (2006). *Data Analysis using Regression and Multilevel/hierarchical Models*. Cambridge University Press.

Gideon, S. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Good, I. J. (1992). Rational decisions. In *Breakthroughs in Statistics*, pages 365–377. Springer.

Haartsen, J. (1998). Bluetooth – the universal radio interface for ad hoc, wireless connectivity. *Ericsson Review*, 3(1):110–117.

Hazelton, M., Mcveagh, M., and Van Brunt, B. (2021). Geometrically aware dynamic Markov bases for statistical linear inverse problems. *Biometrika*, 108(3):609–626.

Hounsell, N., Shrestha, B., Piao, J., and McDonald, M. (2009). Review of urban traffic management and the impacts of new vehicle technologies. *IET Intelligent Transport Systems*, 3(4):419–428.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323.

Jones, G. (2008). Inverse regression from longitudinal data. In *JSM Proceedings*, pages 1533–1538, Alexandria, VA. American Statistical Association.

Kasten, O. and Langheinrich, M. (2001). First experiences with Bluetooth in the smart-its distributed sensor network. In *Workshop on Ubiquitous Computing and Communications, PACT*, volume 1.

Knief, U. and Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 53:2576–2590.

Laharotte, P.-A., Billot, R., Come, E., Oukhellou, L., Nantes, A., and El Faouzi, N.-E. (2014). Spatiotemporal analysis of Bluetooth data: Application to a large urban network. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1439–1448.

Lebreton, J.-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, 62(1):67–118.

Leduc, G. (2008). Road traffic data: Collection methods and applications. Technical Report 1, Institute for Prospective Technological Studies. Working Papers on Energy, Transport and Climate Change.

Liu, Y., Xia, J. C., and Phatak, A. (2020). Evaluating the accuracy of Bluetooth-based travel time on arterial roads: A case study of Perth, Western Australia. *Journal of Advanced Transportation*, vol. 2020, Article ID 9541234, 19 pages. https://doi.org/10.1155/2020/9541234.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1):151–169.

Lusa, L. and Ahlin, Č. (2020). Restricted cubic splines for modelling periodic data. *PLoS One*, 15(10):e0241364.

Malinovskiy, Y., Lee, U.-K., Wu, Y.-J., and Wang, Y. (2011). Investigation of Bluetooth-based travel time estimation error on a short corridor. In *Transportation Research Board 90th Annual Meeting Compendium of Papers DVD*, page 19p, Washington DC, United States. Transportation Research Board.

Malinovskiy, Y., Wu, Y.-J., Wang, Y., and Lee, U. K. (2010). Field experiments on Bluetooth-based travel time data collection. In *Transportation Research Board 89th Annual Meeting Compendium of Papers DVD*, page 17p, Washington DC, United States. Transportation Research Board.

Michau, G. (2016). *Link dependent origin-destination matrix estimation: nonsmooth convex optimisation with Bluetooth-inferred trajectories*. PhD thesis, Université de Lyon.

Michau, G. E., Nantes, A., Chung, E., Abry, P., and Borgnat, P. (2014). Retrieving dynamic origin-destination matrices from Bluetooth data. In *Transportation Research Board 93rd Annual Meeting Compendium of Papers DVD*, page 11p, Washington DC, United States. Transportation Research Board.

Minnen, J., Glorieux, I., and van Tienoven, T. P. (2015). Transportation habits: evidence from time diary data. *Transportation Research Part A: Policy and Practice*, 76:25–37.

Moghaddam, S. S. and Hellinga, B. (2014). Real-time prediction of arterial roadway travel times using data collected by Bluetooth detectors. *Transportation Research Record*, 2442(1):117–128.

Murphy, P., Welsh, E., and Frantz, J. P. (2002). Using Bluetooth for short-term ad hoc connections between moving vehicles: a feasibility study. In *Vehicular Technology Conference. IEEE 55th Vehicular Technology Conference. VTC Spring 2002 (Cat. No. 02CH37367)*, volume 1, pages 414–418. IEEE.

Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465.

Nguyentrang, T. and Vovan, T. (2017). Fuzzy clustering of probability density functions. *Journal of Applied Statistics*, 44(4):583–601.

Nicolai, T. and Kenn, H. (2007). About the relationship between people and discoverable Bluetooth devices in urban environments. In *Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st*

International Symposium on Computer Human Interaction in Mobile Technology, pages 72–78.

Nielsen, F. (2016). Hierarchical clustering. In *Introduction to HPC with MPI for Data Science*, pages 195–211. Springer.

Nusser, R. and Pelz, R. M. (2000). Bluetooth-based wireless connectivity in an automotive environment. In *Vehicular Technology Conference Fall 2000. IEEE VTS Fall VTC2000. 52nd Vehicular Technology Conference (Cat. No. 00CH37152)*, volume 4, pages 1935–1942. IEEE.

Osborne, C. (1991). Statistical calibration: a review. *International Statistical Review/Revue Internationale de Statistique*, pages 309–336.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, 19(1):1–16.

Peterson, B. S., Baldwin, R. O., and Kharoufeh, J. P. (2006). Bluetooth inquiry time characterization and selection. *IEEE Transactions on Mobile Computing*, 5(9):1173–1187.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects Models in S and S-Plus*, pages 3–56. Springer, New York.

Porter, J. D., Kim, D. S., Magaña, M. E., Poocharoen, P., and Arriaga, C. A. G. (2013). Antenna characterization for Bluetooth-based travel time data collection. *Journal of Intelligent Transportation Systems*, 17(2):142–151.

Puckett, D. D. and Vickich, M. J. (2010). Bluetooth-based travel time/speed measuring systems development. Technical report, Texas Transportation Institute. University Transportation Center for Mobility.

Purser, K. (2016). Exploring travel time reliability using Bluetooth data collection: A case study in San Luis Obispo, California. Master's thesis, California Polytechnic State University.

Quayle, S., Koonce, P., DePencier, D., and Bullock, D. (2010). Arterial performance measures with media access control readers: Portland, Oregon, pilot study. *Transportation Research Record: Journal of the Transportation Research Board*, 2192(1):185–193.

Ramsay, J. O. (2004). Functional data analysis. *Encyclopedia of Statistical Sciences*, 4.

Remias, S. M., Hainen, A. M., Mathew, J. K., Vanajakshi, L., Sharma, A., and Bullock, D. M. (2017). Travel time observations using Bluetooth MAC address matching: A case study on the Rajiv Gandhi roadway: Chennai, India. Technical report, Purdue University, West Lafayette, Indiana.

Rokach, L. and Maimon, O. (2005). Clustering Methods. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 321–352. Springer.

Rosenberg, P. S., Katki, H., Swanson, C. A., Brown, L. M., Wacholder, S., and Hoover, R. N. (2003). Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. *Statistics in Medicine*, 22(21):3369–3381.

Sawant, H., Tan, J., Yang, Q., and Wang, Q. (2004). Using Bluetooth and sensor networks for intelligent transportation systems. In *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749)*, pages 767–772. IEEE.

Stevanovic, A., Olarte, C. L., Galletebeitia, Á., Galletebeitia, B., and Kaisar, E. I. (2015). Testing accuracy and reliability of MAC readers to measure arterial travel times. *International Journal of Intelligent Transportation Systems Research*, 13(1):50–62.

Student (1919). An explanation of deviations from Poisson's law in practice. *Biometrika*, 12(3–4):211–215.

Tahmasseby, S. (2015). Traffic data: Bluetooth sensors vs. crowdsourcing—a comparative study to calculate travel time reliability in Calgary, Alberta, Canada. *Journal of Traffic and Transportation Engineering*, 3(2):63–79.

Tai, V., Thao, N., and Ha, C. (2016). Clustering for probability density functions based on genetic algorithm. In *Applied Mathematics in Engineering and Reliability, Proceedings of the 1st International Conference on Applied Mathematics in Engineering and Reliability (Ho Chi Minh City, Vietnam, May 2016)*, pages 51–57.

Van Boxel, D., Schneider IV, W. H., and Bakula, C. (2011). Innovative real-time methodology for detecting travel time outliers on interstate highways and urban arterials. *Transportation Research Record*, 2256(1):60–67.

Vo, T. (2011). *An investigation of Bluetooth technology for measuring travel times on arterial roads: a case study on Spring street.* PhD thesis, Georgia Institute of Technology.

Wang, S. (2013). pbs: Periodic b splines. *R Package version :https://CRAN.R-project.org/package=pbs*, 1.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica*, 9(2):475–499.

Young, S. E. (2012). Bluetooth traffic detectors for use as permanently installed travel time instruments. Technical report, Maryland. State Highway Administration.

Zhou, H., Benz, R. J., Voigt, A., and Mao, A. C. (2016). Bluetooth travel time/speed data analysis for Houston-Galveston regional transportation study. In *Transportation Research Board 95th Annual Meeting Compendium of Papers*, Washington DC, United States. Transportation Research Board.