ARTERIAL SHORT-TERM TRAVEL TIME PREDICTION USING BLUETOOTH DATA

Amir Mohammad Khoei
N 7623232

Supervisory Team
Prof. Edward Chung
Dr. Ashish Bhaskar

A thesis submitted in fulfilment of the requirements for the degree of Master of Engineering- BN72

School of Civil Engineering and Build Environment (CEBE)
Science and Engineering Faculty (SEF)
Queensland University of Technology
2014
Keywords

ARIMA, Bluetooth data, Bluetooth detection, travel time estimation, travel time prediction, SARIMA, urban arterial
Abstract

One of the most important elements of transportation studies is travel time analysis. These studies enable researchers to find solutions for traffic issues such as decreasing the congestion and increasing free flow on arterials and motorways. The ability to effectively monitor traffic performance also has proven to be a valuable asset for transport planning for several years. Particularly for increasing the efficiency of travel time analysis, it is important to gather reliable data for starting the study. While many data collection techniques exist, this research focuses on using data gathered by Bluetooth technology. Bluetooth data application has recently become more popular in transport studies as a cost efficient method for collecting and making traffic databases in transportation studies (D. and Vickich 2010).

As an important part of travel time analysis it is vital to have an accurate estimation of travel time on study routes. A variety of models for estimating travel time on motorways and arterials have been proposed by researchers. Most of the existing models on motorways are developed based on loop data, which are the oldest traffic data sources. Most of these researches have also been focused on estimating and predicting travel time on motorways. Travel time estimation on arterial links is challenging because of various reasons such as stop-and-go running conditions due to signals and non-conservation of flow due to the presence of mid-link sources and sinks (e.g. parking, side street). Based on comparing advantages and disadvantages of using Bluetooth technology in terms of traffic data collection, this research uses Bluetooth data for travel time estimation on arterial corridors. The estimated travel times by Bluetooth data has later been used to make historical databases for urban arterials, to be applied in prediction models.

Meanwhile, based on the reviewed literature in terms of travel time prediction models on motorways and arterials, this research has found the Seasonal Auto Regressive Integrated Moving Average (SARIMA) model as a good choice to be applied for the prediction of short term travel time on arterials. The model has been used previously on motorways. This research, after fitting a suitable SARIMA model on historical databases, evaluates the model in terms of predicting future short term travel time on arterials as well.
Consequently the research focuses on two main objectives: firstly testing Bluetooth data as ground truth data for travel time estimation on arterials and secondly applying and evaluating the SARIMA model in terms of predicting short term future travel time. Bluetooth travel time calculation confidence for estimating each minute’s travel time for the studied routes in this research is obtained between 44 to 85 percent in average. The accuracy, robustness and transferability of the SARIMA model are tested by applying the proposed model on three sites on Brisbane network and evaluating the error percentage comparing with real values. The results contain the detailed validation for different prediction horizons (5 min to 90 minutes). Travel time prediction results by SARIMA model is found more accurate than historical mean method’s results in 90 percent of times. In this case the SARIMA predicts short term travel time with error less than 40 percent for up to 30 minutes after real data, whereas the error by historical mean method is calculated as 50 percent. Therefore, these results would be a good guidance for planners in terms of using Bluetooth data as raw data for transportation studies or SARIMA model for predicting future values depending on how much accuracy is required.
Table of Contents

Keywords .................................................................................................................................................... i
Abstract ..................................................................................................................................................... ii
List of Figures ........................................................................................................................................... v
List of Tables ............................................................................................................................................ vi
List of Abbreviations ............................................................................................................................... vii
Statement of Original Authorship .............................................................................................................. viii
Acknowledgements ..................................................................................................................................... ix

CHAPTER 1: INTRODUCTION .................................................................................................................. 1
  1.1 Background ......................................................................................................................................... 1
  1.2 Aims and objectives ............................................................................................................................. 3
  1.3 Scope and limitations .......................................................................................................................... 4
  1.4 significance ......................................................................................................................................... 4
  1.5 Thesis outline ...................................................................................................................................... 4

CHAPTER 2: LITERATURE REVIEW ........................................................................................................... 7
  2.1 Offline Travel time estimation ............................................................................................................. 7
    2.1.1 Different traffic data collection methods ..................................................................................... 7
    2.1.2 Bluetooth Data Collection Method ........................................................................................... 10
    2.1.3 Bluetooth travel time data filtering ............................................................................................ 16
  2.2 Travel time prediction ......................................................................................................................... 17
    2.2.1 SARIMA Modelling .................................................................................................................... 18

CHAPTER 3: RESEARCH DESIGN ............................................................................................................. 23
  3.1 Case study ......................................................................................................................................... 23
  3.2 methodology ..................................................................................................................................... 26
    3.2.1 Calculating travel time for each section along a corridor .............................................................. 26
    3.2.2 Data accuracy calculation for estimating each section travel time .......................................... 30
    3.2.3 Whole corridor travel time estimation ....................................................................................... 32
    3.2.4 Data clustering and historical database creation ......................................................................... 33
    3.2.5 Identifying Normal Patterns for Three Study Routes ................................................................. 34
    3.2.6 Model creation and fitting suitable SARIMA model on data .................................................... 37
    3.2.7 Predicting future travel time values ............................................................................................ 40
    3.2.8 Prediction results evaluation .................................................................................................... 41

CHAPTER 4: RESULTS AND DISCUSSION .............................................................................................. 44
  4.1 Bluetooth data accuracy results ......................................................................................................... 44
  4.2 Whole corridor travel time estimates ................................................................................................. 55
  4.3 Travel time prediction .......................................................................................................................... 57
    4.3.1 Applying SARIMA ....................................................................................................................... 57
    4.3.2 Applying ARIMA .......................................................................................................................... 64
    4.3.3 Checking SARIMA model reliability .......................................................................................... 69

CHAPTER 5: CONCLUSION ....................................................................................................................... 72

APPENDICES ............................................................................................................................................ 77

BIBLIOGRAPHY ........................................................................................................................................ 97
List of Figures

Figure 1: MAC Address Matching (Blogg et al. 2010-9) ................................................................. 11
Figure 2: Travel time calculation between a pair of detectors .......................................................... 12
Figure 3: Progress of time-series prediction by SARIMA modelling ................................................. 20
Figure 4: Coronation Drive route map .............................................................................................. 23
Figure 5: Wynnum Road route map .................................................................................................. 24
Figure 6: Old Cleveland route map ................................................................................................... 25
Figure 7: Multiple detections of a unique MAC address ................................................................. 27
Figure 8: Vehicle stops along the section caused multiple detections ............................................. 27
Figure 9: Multiple detections for vehicles go along the local access routes ................................... 28
Figure 10: Sample daily travel time patterns for a month showing the clusters ............................... 34
Figure 11: Travel Time Profile for 8th of May 2012 (Coronation Dr) .............................................. 35
Figure 12: Sample historical database for predicting a day ............................................................. 36
Figure 13: Sample Auto Correlation Function graph for investigating “q” ................................. 38
Figure 14: Sample Auto Correlation Function graph for investigating “Q” ...................................... 39
Figure 15: Sample Partial Auto Correlation Function graph for investigating “p” ......................... 39
Figure 16: Sample Partial Auto Correlation Function graph for investigating “P” ......................... 40
Figure 17: Forecasting time series process by SARIMA model ....................................................... 41
Figure 18: Section's travel time points for a sample day on Coronation Drive after removing multiple detections (12th of October 2011) ................................................................. 45
Figure 19: Section's travel time points for a sample day on Coronation Drive after applying MAD filtering (12th of October 2011) ................................................................. 46
Figure 20: First section daily travel time profiles for August 2011 in Coronation Drive ............. 46
Figure 21: Travel Time series for first section of Coronation Dr after removing all Outliers and Smoothing ........................................................................................................................................ 47
Figure 22: All Mondays’ travel time profiles for section (1) Coronation Dr ................................. 48
Figure 23: Travel time profiles for all working Mondays in Coronation Drive after removing days with strange patterns ........................................................................................................ 49
Figure 24: Coefficient of variation graph for all working Mondays in section (1) Coronation Dr .... 49
Figure 25: Required number of data points for calculating travel time value within each 5 minutes time window in section (1) Coronation Dr ................................................................. 50
Figure 26: Number of available data points for each 5 minute time window for a sample day in section (1) ................................................................. 51
Figure 27: Travel time calculation error for each 5 minute time window by using available data (10th October 2011) ........................................................................................................ 52
Figure 28: Data point numbers within each 5 minute moving window (1st section of Coronation drive 8th Nov 2011) ................................................................. 52
Figure 29: Data point numbers within each 5 minute moving window (2nd section of Coronation drive 8th Nov 2011) ................................................................. 53
Figure 30: Data points numbers within each 5 minutes moving window (Whole corridor of Coronation drive 8th Nov 2011) ................................................................. 53
Figure 31: Daily travel time series for Coronation Drive (May 2012) .............................................56
Figure 32: Daily travel time series for Wynnum Road (May 2012) ....................................................56
Figure 33: Daily travel time series for Old Cleveland Road (May 2012) .............................................57
Figure 34: Historical database for predicting June 19th 2012 (Coronation Drive) ..............................58
Figure 35: June 19th forecasts within different prediction horizons by SARIMA ..................................61
Figure 36: Travel time prediction error by SARIMA within each prediction horizon (Morning peak period) .........................................................63
Figure 37: Travel time prediction error by SARIMA within each prediction horizon (Afternoon peak period) ........................................................................63
Figure 38: 4 continuous days from historical database for 19th June 2012 ..........................................65
Figure 39: June 19th forecasts within different prediction horizons by ARIMA ..................................68

List of Tables

Table 1: Travel time collection methods advantages and disadvantages .................................................9
Table 2: Steps for defining required number of data points for each time window on a sample section ........................................................................................................................................31
Table 3: Data status information for estimating travel time on case study routes during peak periods .........................................................................................................................................54
Table 4: Sample model's coefficients and equation ..................................................................................58
Table 5: Forecasted travel time series for 19th June 2012 within different prediction horizons by SARIMA ..................................................................................................................................59
Table 6: Error rate for forecasting 19th of June 2012 by SARIMA within each prediction horizon .......................................................................................................................................62
Table 7: Forecast travel time series for 19th June 2012 within different prediction horizons by ARIMA ..................................................................................................................................65
Table 8: Error rate for forecasting 19th of June 2012 by ARIMA within each prediction horizon .........68
Table 9: Evaluating reliability of SARIMA model ....................................................................................69
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT</td>
<td>Bluetooth</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of Variation</td>
</tr>
<tr>
<td>MAC</td>
<td>Media Access Control</td>
</tr>
<tr>
<td>MAD</td>
<td>Median Absolute Deviation</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean Absolute Percentage of Error</td>
</tr>
<tr>
<td>SARIMA</td>
<td>Seasonal Auto Regressive Integrated Moving Average</td>
</tr>
<tr>
<td>TT</td>
<td>Travel Time</td>
</tr>
</tbody>
</table>

---

2014 Amir Mohammad Khoei
Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

QUT Verified Signature

Signature: [Signature]

Date: 3, 07, 2014
Acknowledgements

I would like to express my special appreciation and thanks to my supervisory team, Professor Edward Chung and Dr Ashish Bhaskar, for their help and valuable advice throughout all stages of this thesis. My greatest thanks and appreciation goes to my parents for their endless love and support. Without them, I would not have had this opportunity. I would like to thank my lovely sister, who was also a great support to me in dealing with all the struggles and frustrations of my new life and studies in this country. Words cannot express how grateful I am to have you here beside me. A special thanks goes to my friends and colleagues at Queensland University of Technology, who were always there to help me to overcome various difficulties I encountered during different stages of this thesis.

Amir M Khoei
February 2014
Chapter 1: Introduction

This chapter after discussing briefly about background of travel time analysis on arterials and also Bluetooth technology application in traffic studies tries to clarify objectives and also significance of this research. Finally it mentions an outline for explaining the achievements of this research within each chapter.

1.1. BACKGROUND

One of the most important parameters in traffic studies is travel time. Travel time estimation and prediction have always been main topics of research for a long time (Bhaskar, Chung and Dumont 2008). Travel time estimation is transforming the observed traffic variables such as flow and occupancy into experienced travel time. Forecasting the future travel time values is named as prediction and is needed for a variety of Intelligent Transport Systems (ITS) applications such as advanced traveller information systems (Casey et al. 2014; Casey, Bhaskar and Chung 2012).

To date numerous models for estimating travel time on motorways (van Lint and van der Zijpp 2003; Sun, Yang and Mahmassani 2008; Sun et al. 1999; Bhaskar, Qu and Chung 2014b, 2014a) and arterials (Bhaskar, Chung and Dumont 2012, 2011, 2010, 2009; Tsubota, Bhaskar and Chung 2013; Bhaskar et al. 2007; Bhaskar 2009) have been proposed. Most of the existing models on motorways are developed based on loop data, which are the oldest and widely used traffic data sources. But loop data are not proven to be accurate enough in terms of travel time estimation. With advancement in technology, other data sources such as Bluetooth (Bhaskar et al. 2013; Nantes et al. 2014; Bhaskar and Chung 2013b; Bhaskar et al. 2014; Tsubota, Bhaskar and Chung 2014; Bhaskar, Qu and Chung 2014a; Kieu, Bhaskar and Chung 2013) and WiFi scanners (Abedi, Bhaskar and Chung 2013; Abbott-Jard, Shah and Bhaskar 2013; Abedi, Bhaskar and Chung 2014) are being explored as complementary transport data. Travel time from Bluetooth data is compared with that from video cameras for motorways (Wang et al. 2011) and arterials (Mei, Wang and Chen 2012). This data provides significant benefit to the road operators for travel time estimation on road networks in a cost effective manner. In some studies (Haghani and Aliari 2011), travel time from traditional matching of Bluetooth scanners is considered as ground truth travel time. Bluetooth tracking is
not only being explored for car travel times estimation, but also for other applications such as bicycle travel time (Mei, Wang and Chen 2012), travel patterns of people movement in airports, shopping malls (MALNATI et al.; Malinovskiy and Wang 2011; O’Neill et al. 2006), work zone delays (Haseman, Wasson and Bullock 2010), Origin-Destination studies (J. Barceló et al. 2011; Blogg et al. 2010-9; Barceló et al. 2010), route choice analysis (Hainen et al. 2011; Carpenter, Fowler and Adler 2010), and freeway travel time variability (Marchouk, Mannering and Bullock 2011).

Travel time estimation on arterial links is also challenging because of various reasons such as stop-and-go running conditions due to signals; non-conservation of flow due to presence of mid-link sources and sinks (e.g. parking, side street) etc. Based on comparing the advantages and disadvantages of using Bluetooth technology in terms of traffic data collection, this research uses Bluetooth data for travel time estimation on arterial corridors. Consequently the estimated travel time is input for the prediction modelling for forecasting future travel time (Bajwa, Chung and Kuwahara 2003; Park, Rilett and Han 1999).

Besides collecting reliable data for travel time calculation, short term travel time prediction is a topic of interest in transport studies. This is important due to its applications in real time traffic monitoring and management, including traveller information systems. This research, after estimating travel time based on Bluetooth data, has applied a prediction model to forecast short term future travel time values in arterial corridors. The estimated travel time from case study routes are shown and stored as stationary recurrent time-series with a specific seasonality. Then the Seasonal Auto Regressive Integrated Moving Average (SARIMA) modelling is chosen to forecast future behaviour of travel time-series. SARIMA is one of the commonly used linear models for forecasting the short term future and working on univariate data as its input data. It is also proved that the SARIMA model can deal well with seasonality and trends in terms of monitoring historical data for predicting future (Maier and Dandy 1996; Tong and Liang 2005). In this research, the SARIMA model is applied on data gathered for up to eight months, considering seasonality and without this factor. This model is also applied on the Bluetooth data within different prediction horizon times. The prediction results by the model are evaluated by comparing them with the real values within each prediction horizon. Finally, the
capability of the model for predicting travel time on arterials is tested by applying it on various urban corridors.

1.2 AIMS AND OBJECTIVES

This research aims to use Bluetooth data from Brisbane city as ground truth data for traffic analysis along Brisbane’s signalised arterials. The main parts of this analysis are estimating the arterials travel time and producing an accurate prediction model for predicting short term (up to 30 or 90 minutes after real data) travel time values on them.

The research direction is towards the development of a reliable method which is capable of considering the past traffic condition on any signalised arterial corridor and accordingly predicts the future values for the same route. For this purpose, the following objectives are defined to be achieved in this thesis:

- Analyse Bluetooth data on arterial corridors and evaluate the accuracy of estimated travel time for the signalised urban corridors using this data
- Develop historical time series databases for the urban arterials and check whether daily travel time data shows any seasonality
- Develop a time series model for predicting the future based on created historical database
- Measure the accuracy of created model for forecasting future travel time value within different periods of time
- Check if the prediction results on arterial networks are more accurate when considering seasonality in daily patterns or without it
- Test the transferability of robustness of model by applying it on different corridors

Therefore, the key aim is to examine the potential of using the chosen model for travel time prediction on Bluetooth traffic data gathered from arterials. This is needed to be evaluated for all types of signalised arterials in terms of having different traffic conditions during a day.

Generally, this research, after evaluating the accuracy of travel time estimation by Bluetooth technology, tries to apply and examine the performance of the selected prediction model for forecasting future short term travel time on arterial corridors.
1.3 SCOPE AND LIMITATIONS

This research tries to predict future travel time values for urban routes based on analysing past traffic data at the same route. The primary step for achieving this goal is estimating the most accurate travel time value for required arterial routes. Level of accuracy for travel time estimation on signalised arterials using Bluetooth data is evaluated in the research as well. However, irrelevant of how much accuracy is obtained for the estimated travel time series by Bluetooth data, it is assumed that the created historical daily travel time series using Bluetooth data has sufficient accuracy. Then the next prediction steps are done, considering this assumption.

1.4 SIGNIFICANCE

Recently considerable rate of global population growth and migration of people from rural parts to the cities caused many transportation issues in big cities. As a valuable asset for decreasing the traffic issues such as rush hours’ congestions, it is vital to monitor performance of urban arterials efficiently. Particularly, travel time estimation and prediction are known as two primary steps in transportation planning which can lead to improvement of transportation efficiency. It is also important to prepare the most reliable raw data for increasing the accuracy of the travel time prediction. The outcomes of this research can provide guidance for planners in terms of using Bluetooth data as raw data for transportation studies. It will conclude whether it is practical and accurate enough to use Bluetooth data for estimating urban arterials travel time. The research also evaluates applying a linear prediction model on travel time data for predicting short term future on urban arterial routes. The results of this prediction can be applied in variety of ITS (Intelligent Transport System) applications such as advanced traveller information system, dynamic route guidance and advanced control measures.

1.5 THESIS OUTLINE

The thesis is structured in five chapters as described below;

Chapter 1 is an introduction to the thesis. Research aims and objectives are presented based on justification from the background information. The scope of the study and its limitation is also included in this chapter.
Chapter 2 presents a review of the available literature related to the subject of the thesis containing:

- Different methods of data collection for traffic studies, focusing on Bluetooth data collection method
- Choice of a suitable prediction model for forecasting future travel time based on characteristics of input data
- Explanation of the characteristics of the chosen model and its application on available traffic data

Chapter 3 discusses the research methodology followed in this thesis, for achieving the defined scopes in this research. The chapter aims to clarify the approach adopted in analysing the arterials’ traffic data and predicting future short term travel time based on historical data.

Chapter 4 discusses the results of each step explained in the methodology section for achieving the final goal of the research. Finally, based on the performed comparison among results gathered within each step of this thesis, accuracy of travel time estimation via Bluetooth data will be defined. The performance of the selected model for forecasting travel time in short term future on signalized arterials is also evaluated in this chapter.

Chapter 5 reports the final conclusions of this research based on the results discussed in Chapter 4 and also gives a recommendation for upgrading the final results for future studies.

**Summary:**

This chapter summarised the background of research in same field of travel time estimation on motorways and arterials for choosing the best method of prediction. Therefore by defining the main objectives of this study, it clarified the major issues upcoming in this research.
Chapter 2: Literature review

This chapter summarises the reviewed literatures regarding travel time data collection, estimation and prediction methods. These reviews include the chosen methods for filtering traffic data and also finding the best model for predicting future travel time series based on historical databases.

2.1 OFFLINE TRAVEL TIME ESTIMATION

Travel time estimation is an important study area in real time and predictive traveller information systems that can lead to the improvement of transportation efficiency. Travel time is an important parameter for evaluating the operating efficiency of the traffic networks, assessing the performance of traffic management strategies and developing the real time vehicle route guidance system (Kristin Tufte 2008). Travel time estimation is used as the main data resource for predicting travel time algorithms and analysing the historical travel time data. This enables researchers to estimate short term or long term travel time in arterials or motorway corridors. Therefore it determines offline performance measures for different policy applications. For example, this information is used by decision makers and transport planners in many project assessments. Travel time estimation is also used to form an important part of Advance Traveler Information and Transportation Management System (Barceló et al. 2010; Porter, Kim and Magaña 2011).

However, for increasing the efficiency of travel time estimation, it is important to prepare the most reliable raw data for starting the study.

2.1.1 Different traffic data collection methods

Loop Detectors

One of the most popular methods for estimating travel time is applying loop detectors. Loop detectors are always installed under the pavement of roads and consist of a large coil of electrically charged wire. Whenever a large metal object passes over the coil, it affects the coil’s electrical inductance, causing a change in the flow of electricity running through it and triggering the detector. Dual-loop detectors formed are by two consecutive single-loop detectors placed a set distance apart, so
they are capable to measure vehicle speed (Porter, Kim and Magaña 2011; Gibson 2009). Single loop detectors are usually used to measure the volume and occupancy on lanes but dual loop detectors are capable of measuring speed, length and consequently travel of vehicles (Gibson 2009). Most transport studies have used loop detectors for estimating travel time on motorways. Also, travel time estimated by loop detectors is not based on an individual vehicle’s travel time and can be faulty, as the estimated speed is based on the assumption of a common vehicle length. Although double loop detector estimates travel time value better than individual ones, but still in terms of congestion periods, which are the most important for estimating travel time, they cannot perform well (Marchouk and Mannering 2009).

**License Plate Matching**

This technique is based on capturing and recording license plates of any individual passing vehicles at specific places on roads equipped by camera. These cameras take license plate pictures and record them with corresponding time for calculating travel time later. Despite the high accuracy of this method in terms of estimating individual vehicle’s travel time, it is a quite expensive technique and produces a large amount of data. There would also be some external factors affecting the capture of license plates by the cameras, such as lane changing by cars at the time of capturing and environmental factors such as precipitation or angle and intensity of sun (Asudegi 2009).

**Floating Car and Probe Vehicles**

In this technique someone is hired to drive a vehicle along a specific defined route and record the time and distance for travel time calculation. This is a pretty simple method for estimating travel time on corridors but is has some disadvantages such as: inserting human error in data recording, limited numbers of possible measurement within each day and also not working well in testing a variety of traffic conditions within a day (Asudegi 2009).

**Electronic Distance Measuring Instruments (DMIs)**

This technique is quite similar to the probe vehicle method. The difference is that a sensor is attached to the probe vehicle, which moves along the route. The DMI receives pulses from a vehicle during its movement along the corridor. This recorded
time is maintained on a computer to use as travel time estimation (Asudegi 2009). This technique is easier and safer to use compared to a probe vehicle. It is also easier to process the data, as data is automatically received and saved electronically in databases and decreases human errors in estimation, but it still has the same disadvantages of probe vehicles.

**Automatic Vehicle Identification (AVI)**

This technique works with an in-vehicle transponder, some reader units installed on the road side, and a central processing unit. When a vehicle with attached transponder passes a road side reader unit, the reader captures the transponder information and transfers this data to the main processor. The main computer processes the data and finalizes the travel time. In this method, human error has been eliminated and the computation of travel time is easier and faster and can be installed in toll collection places. It still has some disadvantages such as high initial costs, vehicles needing to be provided by transponder tags and being limited by only few equipped corridors (Asudegi 2009; Tam and Lam 2011).

**Global Positioning System (GPS) Equipped Probe Vehicles**

This technique uses probe vehicles that are equipped by GPS for increasing the accuracy of data in the floating cars method (Blogg et al. 2010-9). Although the accuracy of finding travel trajectory by this method is pretty high and there is no need for equipment installation along roads, data collection and management cost by this method is still high (Asudegi 2009; Blogg et al. 2010-9).

All these techniques have their own advantages and disadvantages. Some of the key points about each travel time collection technique are summarized in the following in Table 1.

<table>
<thead>
<tr>
<th>Data collection Method</th>
<th>Benefits</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop Detectors</td>
<td>Good for counting traffic volume and vehicles speed</td>
<td>Low accuracy of estimated travel time in congestions, Very difficult to estimate travel time on signalized arterials</td>
</tr>
<tr>
<td>License Plate Matching</td>
<td>Forming large sample size, useful for OD studies, giving speed profiles for study section</td>
<td>Mismatching license plates, Not good for high speed corridors, High initial costs</td>
</tr>
</tbody>
</table>

Table 1: Travel time collection methods advantages and disadvantages
through the peak period

<table>
<thead>
<tr>
<th>Floating cars</th>
<th>Easy performance</th>
<th>Human errors and limited samples, not measuring delay of cross street traffic turning onto the route</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Electronic Distance Measuring Instruments</strong></td>
<td>Cost effective compared with floating cars, access to detailed travel time in routes, easier data processing</td>
<td>Similar to floating cars (only omitting human errors)</td>
</tr>
<tr>
<td><strong>Automatic Vehicle Identification</strong></td>
<td>High accuracy of data gathering and processing, real travel time information, removing human errors</td>
<td>High initial costs, privacy concerns, limited amount of data</td>
</tr>
<tr>
<td><strong>GPS Equipped Probe Vehicles</strong></td>
<td>High resolution vehicle trajectory data, no indoor installation needs, precise scheduling, no limitation in routes, no human errors</td>
<td>Fail in tunnels and tree canopies, difficulty of data management, high initial costs and privacy concerns</td>
</tr>
</tbody>
</table>

**2.1.2 Bluetooth Data Collection Method**

Bluetooth, invented in 1994 by engineers from Ericsson, a Swedish company, enables sharing of music, images, and other data wirelessly over a Personal Area Network (PAN) defined by the device’s antenna (Bluetooth: About the technology 2011). A radio frequency refers to rate at which radio signals are transmitted. The effective signal range of a Bluetooth device defined by its antenna class is the range at which other Bluetooth devices may be discovered and connected. Bluetooth operates using low power and is intended to replace the cables which are commonly required to connect devices such as headsets and phones, keyboards and computers, and cameras and printers (Bluetooth: About the technology 2011). In contrast to more commonly used radio signals (TV, radio, etc.) which are broadcast over large areas, Bluetooth sends radio signals over short distances ranging from a minimum of one metre to more than 100 metres (Bluetooth: About the technology 2011). In early 2000, researchers investigated and tested Bluetooth for discovering movement of vehicles and applying it to intelligent transport services (Sawant et al. 2004; Nusser and Pelz 2000; Murphy, Welsh and Frantz 2002).

Bluetooth data application has recently become more popular in transport studies as a cost efficient method for collecting and making traffic databases in
transportation studies (D. and Vickich 2010). Interested readers should refer to Bhaskar and Chung (2013b) for detailed discussion on the use of Bluetooth as a complementary transport data. Each electronic device is enhanced by Bluetooth technology and used by automobile passengers, and has its own unique identifier number. This is a 48-bit, 12 alpha-numeric character, which is specific for the device manufacture and unique for each device (Porter, Kim and Mangaña 2011; Barceló et al. 2010). This number can be captured and registered by compatible Bluetooth scanners. These unique MAC (Media Access Control) addresses are captured by scanners to calculate travel time along an urban route when a vehicle containing a discoverable Bluetooth device passes through a scanner detection zone. Such a zone is the communication area defined by characteristics of the scanner antenna (Abedi, Bhaskar and Chung 2013; Khoei, Bhaskar and Chung 2013). Figure 1 is a schematic view of MAC address matching in central station via the Bluetooth data collection method. It shows how travel time value is calculated for each individual vehicle (MAC address) for a section between a pair of Bluetooth detectors.

![Figure 1: MAC Address Matching](image)

As the detection zone of each antenna is a broad area and can be a circle of even 150 m, the exact location of detection is unknown in this method. There is also
probability of a vehicle being detected in a zone several times because of having a long stop there. Therefore, each captured MAC address is registered with the corresponding detection time and the time of device presence in the detection area as duration. For calculating travel time between each couple of detection zones, the exit-to-exit method is used. Based on this method, the exit time of a vehicle within each detection zone is calculated by adding duration time to the time that vehicle is detected for the first time. Then for estimating travel time between upstream and downstream these two exit values are subtracted from each other (Bhaskar and Chung 2013a). Figure 2 shows the exit to exit scenario clearly.

![Figure 2: Travel time calculation between a pair of detectors](image)

**Bluetooth data collection method benefits**

During recent years using the Bluetooth technology has become more popular in traffic studies due to its advantages. Some of the advantages which make this method particular are:

- Data gathered by this method is based on estimating travel information from each vehicle that is tracked individually. So taking into account its easy implementation and relatively few costs, it is a good method for travel time data collection.

- Applying this new method does not need any extra construction work and equipment can be installed on current traffic devices in roads with minimum cost.
In this method, by applying some strategies (changing MAC address name) the probability of privacy conflict has decreased considerably.

In this method, sensors (readers) are capable of being moved and installed in different points of a route to increase the volume of collected data.

This method has the minimum rate of human error because of being fully automatic.

Finally using this method gives planners this opportunity to calculate average speed for each section of a road by constant monitoring during different periods of times.

---

**Bluetooth Data Collection Method Challenges**

Based on the literature published about traffic data collection by Bluetooth technology, it has some technical and methodological challenges which should be considered during the research. Some of these challenges are described as follows:

**Technical challenges**

Different methods of detector installation and antenna array can lead to considerable differences in rates of gathered data. Some of important points which have major effects on antenna detection are mentioned in the three following categories.

1. **Scanning Area (detection zone) or Antenna Configurations**

   - Based on concluded results from a test conducted by Michael O(Rodgers and Hunter 2011), Omni directional antenna has a better detection zone compared with directional antenna (Rodgers and Hunter 2011; Porter,Kim and Magaña 2011).

   - Vertically polarized antennas with gains between 9 and 12 dB are suitable choices to apply for a Bluetooth-based travel time data collection system (Porter,Kim and Magaña 2011).

   - According to an implemented experiment in Indianapolis for evaluating the influence of vertical sensor placement on data collection efficiency from Bluetooth, antenna with height of 2.3 m can lead to a threefold increase of capture rate. The research result suggests the minimum
height of 2.5 m for antenna installation (Brennan et al. 2010; Martchouk and Mannering 2009).

- For most geographic situations, when a standard 1 dB Omni-directional antenna couples with a class 1 adapter, the detection zone is between 100 to 120m (D. and Vickich 2010).

- During test correlation, radio range and vehicle speed should be noted. Bakula says, “When the effective range of detection in two locations grows to 100 or 200m on each side of radio detection, probability of detection grows to more than 94% (independent of vehicle speed). But when this range falls into less than 20m, probability of detection in two locations decreases to 6%, due to variety of vehicle speeds” (Bakula et al. 2012).

- Depending on different routes for testing, there would be different ranges of average speed. For roads with higher average speed and free flow such as freeways, it would be better to divide roads into larger segments. But for the arterial routes that have more congestion and less average speed, it would be better to define smaller segments. This will help to collect more accurate traffic data for each kind of network (Click and Lloyd 2012).

2. Location of scanners on the road

- Click and Lloyd (2012) implemented a test in a rural route with a wide median for finding the best placement (in median part or roadsides) of detectors on roads. They concluded the following result; for the roadways with larger median, installing two roadside antennas increases the capture rate of MAC addresses compared with installing one antenna in the median part of a similar road. (Click and Lloyd 2012)

- Based on a performed test by Michael (Rodgers and Hunter 2011) for comparing different types of antenna combinations to increase the captured MACs’ rate, the following result is derived:
1) Combination of two horizontal antennas that are installed apart from each other works better than a combination of vertical antennas with similar configuration. It means that horizontally separated antennas have bigger rates of detection compared with vertically separated ones.

2) Two antennas with no separation between them (attached together) would not have better performance than a single antenna.

3) Finally, arrays with an even number of antennas would have more capture rate than a combination of an odd number of antennas.

3. **Station Configuration**

Clock drift: Computers cannot keep perfect time, when they are isolated from a central time source. There are always drifts happening at times during a week period. So, when calculating a short distance travel time is required, these few seconds can impact on the results (Blogg et al. 2010-9).

Ping Cycle: Because of the short ping cycle of Bluetooth radio devices (about 0.1 second), they can be captured several times in a detection area. This extra detection rate increases when vehicle speed is less or the signal strength is high. Consequently, the final gathered data in this case needs more effort applied to be filtered efficiently (Blogg et al. 2010-9).

**Calculation Challenges**

Data gathered by Bluetooth technology have some outliers that should be cleaned before being used for travel time studies. These outliers are usually inserted in traffic data from different ways such as:

- **Unknown mode of transporting a device between two detection zones.** Bluetooth enabled devices can be transferred by pedestrian, cyclist, public transport users and vehicle users. When the main focus of study is on traffic data analysis for motorways or arterials, travel times inserted in data by pedestrian and cyclists would be considered as outliers in data.

- **Vehicle which has been detected at both upstream and downstream locations but did not pass through the required trajectory between two**
detectors. Or vehicle that has a stop along the section. These types of registered travel time values are counted as outliers.

- Multiple detection of one vehicle at a detection area, which can be caused by turning from detection zone to a bystreet and coming back to detection zone after a while.

- Multiple detection of one vehicle, which is caused by vehicle U turn between two detection areas (Wang et al. 2011; Haghani et al. 2010; Bhaskar and Chung 2013a).

2.1.3 Bluetooth travel time data filtering

Data used in this research is collected by Brisbane City Council from Bluetooth detectors installed along the urban arterials. By having access to this real data, the technical challenges of Bluetooth data collection are over come so far. Therefore as a primary step, this research focuses on filtering data from outliers as explained in the previous section. For removing multiple detections a minimum travel time value for passing a section is found. Therefore by finding the least value registered for a specific vehicle among multiple detections, which is more than defined minimum, the other data points for the same vehicle are removed (Khoei, Bhaskar and Chung 2013).

In terms of removing outliers from Bluetooth travel time data, Le Minh Kieu (2012) has compared two “Median Absolute Deviation” (MAD) and “Box & Whisker” methods for removing outliers from traffic data gathered by Bluetooth technology (Kieu, Bhaskar and Chung 2012). John Tukey proposed the Box and Whisker method in 1977 (Tukey 1977). This method is based on defining an upper bound and lower bound window for filtering data. Upper and lower bound values are found by investigating the 25th and 75th percentile of data points in this method. Accordingly values beyond these two boundaries are counted as outliers and are removed from data (Tsubota et al. 2011).

MAD (Median Absolute Deviation) also is the median of the absolute deviations from the data's median (Smith and Demetsky 1994). In this method upper bound value is calculated by adding $\sigma f$ to the median of data and likewise the lower bound is derived from subtraction $\sigma f$ from Median of data points, where $\sigma$ is the standard deviation from the MAD. A normally distributed data is approximated as
1.4826*MAD and MAD is the Median of the absolute deviations from the data median (Smith and Demetsky 1994).

Based on a comparison performed by Kieu et al. between two methods on Bluetooth data, the MAD method performs better than the Box and Whisker method in terms of removing more outliers from data points (Kieu, Bhaskar and Chung 2012). After removing all these outliers from Bluetooth data, it is ready to be used for further studies.

After travel time points calculated for each section are filtered, the average travel time for the sections can be calculated for each minute of day. Therefore by having all sections’ average travel time values, whole corridor (route containing several sections) travel time can be calculated and shown as a daily time series.

2.2 TRAVEL TIME PREDICTION

Travel time prediction is needed for a variety of intelligent transport systems (ITS) applications such as advanced traveler information systems, dynamic route guidance for assisting travelers in terms of choosing the best mode of travelling and decreasing the travel time to its minimum level. It can also be used as an advanced control measure for urban planners and traffic researchers to decrease congestion in roads and conduct vehicle movement to free flow.

Based on the traffic data characteristics, different methods of historical database algorithm, time series models and simulations are done before for forecasting future travel time (Okutani and Stephanedes 1984). Through using different types of input data for prediction process, different models have been assessed up until now. The input data can be categorised into two main groups of univariate or multivariate data. The multivariate data group consists of independent variables which have direct effects on the predicted travel time by mathematical relation based on historical data (Araghi, Krishnan and Lahrmann). In terms of multivariate data analysis, methods of linear and non-linear regression, Kalman filtering and artificial neural network (ANN) are applied for predicting travel time and bus arrival time (Box 1976; Araghi, Krishnan and Lahrmann).

There are also univariate prediction models which are designed to predict a dependent variable just by making a relationship between historical data and future values. The commonly used univariate models include: probabilistic estimation, time
series models such as ARIMA (Auto Regressive Integrated Moving Average) and Neural Network (Araghi, Krishnan and Lahrmann; Box 1976; MALNATI et al.; Okutani and Stephanedes 1984; Stephanedes, Kwon and Michalopoulos 1990). ARIMA and SARIMA (seasonal ARIMA) have proved to be performed effectively in terms of forecasting time series among linear models. Neural Network also shows good results in terms of forecasting future on both univariate and multivariate data.

It has been proven over recent years that both Neural Network and SARIMA (Seasonal Auto Regressive Integrated Moving Average) models can deal better with trends and seasonality in data among other models (Tong and Liang 2005). But Maier and Dandy (1996) believe that in terms of forecasting univariate time series, the SARIMA model is more suitable for short term forecasting than long term, compared with neural network (Reisen 1994; Maier and Dandy 1996).

Most of these statements are concluded from working on data gathered from motorways, irrespective of data type, just on time series. So their performance is not evaluated on urban arterial traffic data. Therefore the SARIMA model is chosen as a capable model to apply on arterial traffic data and forecast the short term future.

2.2.1 SARIMA Modelling

Predicting future behaviour of a time series is an important area of forecasting in terms of observing past behaviour, and collecting variables in order to analyse and develop a model that considers underlying relationships. Then, based on the past behaviour of time series, it is possible to extrapolate future time series. One of the most popular models in terms of forecasting time series in the last decades is the ARIMA model (Zhang 2003). The ARIMA model is popular not only because of the well-known Box–Jenkins methodology in processing the model, but for its specific statistical properties (Box 1976). The ARIMA model is also capable of implementing various types of exponential smoothing in order to simplify the computational procedure and makes using the model easier (McKenzie 1984). Besides this, the ARIMA model is very flexible in terms of representing a variety of time series in different modes of pure auto regressive (AR), pure moving average (MA) or a mixture of them (ARMA). By upgrading a model to SARIMA, it can be capable of considering the seasonality of historical time series for predicting future behaviour of such time series. The only weak point about ARIMA, which is highlighted in literature, is that it cannot recognize nonlinear patterns. Therefore in the case of some
complex problems, for which linear models do not have satisfactory approximation, ARIMA cannot perform well (McKenzie 1984; Zhang 2003).

The Auto Regressive Integrated Moving Average (ARIMA) method developed by Box and Jenkins in 1976 analyses stationary univariate time series (Box 1976). This statistical model is fitted on the data to make a better understanding of the trends and variations from the previous data points observed for predicting future values in time series. ARIMA is a generalized model of the auto regressive moving average (ARMA) method enhanced by a degree of differencing for changing non-stationary time series to stationary ones (Ong, Huang and Tzeng 2005). This method also becomes more useful for forecasting future values when it is fitted on seasonal time series. In this case the method is changing to SARIMA with consideration of a seasonality value for the historical data (Williams and Hoel 1999; Bowerman, O’Connell and Richard 1993).

The shorthand notation SARIMA (p, d, q) (P, D, Q)s is often used when an ARIMA model is factored into "local" and "seasonal" multiplicative terms where the seasonal period “s” is known. The indicators p, d and q are non-negative integers which refer to the order of auto regressive, differencing and moving average non-seasonal part of the model (Tong and Liang 2005). Respectively the same capital letters (P, D, and Q) are referring to the same action but in a seasonal part. The application of this model contains three main steps; (1) model identification: where p, d, and q are estimated from the autocorrelation function and partial autocorrelation function of the time series (2) parameter estimation: where the coefficients are chosen accurately or need to be changed and (3) model checking and validation through prediction results (Tong and Liang 2005).

The steps for predicting future values by SARIMA are shown in Figure 3.
Figure 3: Progress of time-series prediction by SARIMA modelling

Figure 3 represents the procedure in which to see firstly whether the historical time series are stationary or not, and then turn them into stationary if they are not. When the data is in equilibrium around the mean and the variance changes around the mean remains constant over time, the historical time series are stationary. Otherwise it is non-stationary and one or two differentiations would be needed to transform the data into stationary (Van den Bossche, Wets and Brijs 2004; Makridakis, Wheelwright and Hyndman 2008). Secondly, the first values for the $AR$ and $MA$ are identified by drawing ACF (Auto Correlation Factor) and PACF (Partial Auto Correlation Factor) graphs. The ACF plots the correlations between $X_t$ and $X_{t-k}$ against the lags ($k = 1, 2, 3 \ldots$) and then identifies possible MA terms (Makridakis, Wheelwright and Hyndman 2008). The PACF plots the coefficients in a regression of $X_t$ on $X_{t-1}$, $X_{t-2} \ldots X_{t-k}$ and identifies possible AR terms (Reisen 1994; Ong, Huang and Tzeng 2005). Likewise, ACF identifies possible MA terms for the first created model. Details of finding first values based on observing ACF and PACF are explained in detail in methodology part (Section 3.2.6).
After this stage, by fitting the model on half of the data and calculating the residuals for the second half, a diagnostics test is implemented to check whether the model is fitted well on the data. Accordingly, the final coefficients are defined and the identified model is ready for forecasting future lags.

The following equation shows the SARIMA linear mathematical formula in a general form:

Equation 1:

$$\varphi_p(B^s)\Phi(B)\nabla_s^p\nabla_dX_t = a + \Theta_q(B^s) \theta(B)a^t$$

In this equation, $B$ and $B^s$ denote ordinary and seasonal backward shift operators. Respectively, $\Phi(B)$, $\theta(B)$, $\varphi(B^s)$ and $\Theta(B^s)$, are polynomials (of $x$ & $a$ variables) in $B$ and $B^s$ of finite order $p$ and $q$, $P$ and $Q$, respectively. The “$a$” also is an estimated white noise by the model (AIDO0 2011). This equation shows how much the future values are dependent on their previous lags and the lags from past seasons. For example the equation for SARIMA (1,0,1)(1,0,1)_{12} would be:

Equation 2:

$$X_t = a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-12} + \Phi_1 X_{t-1} + \varphi_1 X_{t-12}$$

This sample equation shows when a SARIMA model with seasonality of 12 is fitted on historical data; the $X_t$ value is predicted by making the mentioned linear formula among the previous lags in the same season and previous season ($X_{t-12} & X_{t-12}$), while considering a white noise and constant from current and past season (Anderson 1976).

Application of SARIMA, considering seasonality and without, is explained in detail in section 3.2.6 of this thesis.

**Summary:**

This chapter after discussing about different applied techniques for traffic data collection, and assessing them in terms of their advantages and disadvantages, looked for a new method for collecting accurate traffic with considering the cost and
implementation process. This would be a massive help to find a solution for data lack issue in traffic studies. Finally the Bluetooth technology is chosen as an accurate cost effective method to be used in traffic studies for collecting data. Then after explaining the chosen model for filtering Bluetooth travel time data and making historical databases, this chapter focused on finding a potential accurate model for predicting future travel time values. As a big gap in traffic studies on arterials, it is vital to have an accurate forecast of travel time on arterials to apply it for real time applications in future. For this purpose this study investigated the SARIMA model for predicting future short term travel time based on historical databases. Finally it explained the logic behind the SARIMA model for predicting future values in time series and also how it is applied in this research.
Chapter 3: Research design

This chapter shows how the methods explained in the last chapter are applied in this research for analysing Bluetooth traffic data in details. Based on procedure used in this research, this chapter firstly introduced three chosen routes as case studies in this research. Then it is explained how traffic data is processed on each of these routes in terms of data analysing and predicting future values and also evaluating the prediction model results.

3.1 CASE STUDY

All the procedure explained in the following sections of this chapter try to analyse and predict travel time data for three different arterial corridors in Brisbane city. Primary analysis and predictions are performed on the first case study route and then the procedure is expanded on two other arterial corridors with different types of traffic conditions. These three case study arterials consist of:

**First Sample Route:**

The first chosen sample route is Coronation Drive.

![Figure 4: Coronation Drive route map](image)

2014 Amir Mohammad Khoei
Coronation Drive is a signalized arterial connecting the south western suburbs to Brisbane CBD and it always experiences heavy traffic conditions during morning and afternoon peak periods. For calculating travel time along this route, two sections are defined by choosing three detectors. As shown in Figure 4, the first section (A-B) starts from Toowong suburb and continues to Milton suburb with length of 2 km. The second section (B-C) with about 800 metres length is located between Milton and the “Go Between Bridge” entrance near Brisbane CBD. This route has three vehicle lanes for each direction (inbound and outbound) and contains four traffic lights in the whole corridor length (3 on A-B section and 1 on B-C section).

**Second Sample Route:**

The second chosen route to be studied in this research is Wynnum Road.

![Figure 5: Wynnum Road route map](image)

Wynnum Road is a North-East to West route which connects eastern suburbs to Brisbane CBD via the Story Bridge. It contains two sections. The first section (A-B) is defined between Morningside suburb and the Hawthorne area with about 2.5 kilometres length. The second section (B-C) is located between Hawthorne and the Story Bridge entrance, and is of about 3.2 kilometres length. This route has mostly
two vehicle lanes for each direction of inbound and outbound traffic. It also contains 10 traffic lights along the whole corridor (five on section A-B and five on section B-C).

**Third Sample Route:**

Old Cleveland Road is the third case study route in this research.

![Figure 6: Old Cleveland route map](image)

This corridor is the East West arterial, in which its first section (A-B) starts from Camp Hill and continues for about 3.3 kilometres to the busy suburb of Stones Corner. The second section (B-C) of this route with 1.2 kilometres length starts from Stones Corner and ends at Ipswich Road in Woolloongabba suburb. This corridor has two vehicle lanes in each direction. There are also 12 traffic lights along this section (eight in section A-B and four in B-C).
3.2 METHODOLOGY

3.2.1 Calculating travel time for each section along a corridor

In this study, data was collected from Bluetooth detectors installed along the urban corridors by the Brisbane City Council. This data represents unique MAC (Media Access Control) addresses captured by the Bluetooth detectors and then registered with their corresponding complete details of date and time of first detection, and duration of presence in the detection zone. Access to this data for each detector, means the same MACs are matched from upstream and downstream and the time spent by each vehicle to pass between the two detectors is calculated. A section is the distance between two continuous detectors installed on the route. These calculated time values between two detectors are counted as travel time for that section. This method enables several travel time values to be calculated for a unique vehicle (MAC address) in a short period of time (Smith and Demetsky 1994)

Calculated travel time values for each section contain many outliers at this stage. Most of these outliers are inserted to the section travel time values by multiple detections. As explained in section 2.1.2, the Bluetooth detectors capture any MAC address passing within their detection zone. Therefore, the probability of detecting a unique vehicle ID (MAC address) several times within a short period of time is high in different possible situations.

First situation: One device that was detected at a detection zone, exited the detection zone and returned to it after a short time. In this case there are several registered time-stamped values for one vehicle and this produces several unrealistic travel time values just for one vehicle. This problem usually happens in shorter sections located in areas with several local access routes. Figure 7 draws a schematic view for this situation. This shows how a vehicle (green dot) exits detection area at upstream and downstream intersections and return to them after a while. For example the vehicle shown in this figure is detected at least two times within each detection area.
Second situation: In this situation (showed in Figure 8) some vehicles start their trip between two detection zone by passing the first detection zone and entering the route, stopping somewhere along the section for a short period of time and then continuing the trip to the second detection zone. These types of unrealistic travel time values are usually seen in data in longer sections locating near shops, gas station or other services. Figure 8 also shows the vehicle (green dot) movement between two detection areas “A” and “B”. The green dot at the middle part of this section shows that the vehicle has a stop at the middle of road and then approached to the second detection area after a while.

Third Situation: In this case some vehicles might pass the distance between two detection zones via another local access route. Therefore, vehicle IDs are detected at both detection zones but the calculated travel time value for the section is higher than the maximum required time for passing the section (Figure 9). This problem mostly happens in the sections located among several local access routes.
Fourth situation: There would also be a considerable rate of outliers inserted to calculated travel times for an urban arterial’s sections, caused by pedestrians and public transport users. This type of unrealistic data is unavoidable in urban traffic analysis by using Bluetooth technique.

Filtering phase:

In order to remove the explained outliers from calculated travel time values of a section’s data, two different scenarios are followed in this research. As a first step, a minimum and maximum value is defined for each section’s travel time based on normal expected travel time for the section. Based on this, the minimum calculated value stated in the mentioned interval for each vehicle is kept as real travel time value for that vehicle.

The second method used in this research for removing outliers is applying the MAD (Median Absolute Deviation) filtering method (Smith and Demetsky 1994). Having travel time values as a univariate data, MAD is the median of the absolute deviations from the data's median (Smith and Demetsky 1994). For this outlier detection method, the median of the residuals is calculated. Then, the difference is calculated between each historical value and this median. These differences are expressed as their absolute values, and a new median is calculated and multiplied by an empirically derived constant to yield the median absolute deviation (MAD). If a value is a certain number of MAD away from the median of the residuals, that value is classified as an outlier.
Equation 3:

\[ \text{MAD} = \text{median} | X_i - \text{median}(X) | \]
\[ \sigma = k \times \text{MAD} \]

In this formula \( X_i \) represents each data point and “median\((X)\)” is the median value for whole data points. The “k” is also a constant scale factor, which depends on the data distribution. This research created a moving time window (10 minutes) for applying MAD on data. This 10 minute time window considers five minutes before and five minutes after each minute of the day and calculates the MAD value for each minute. By adding \( \sigma \) to the median and subtracting it from the median, maximum and minimum margins are defined for each minute time window and only travel time values within this interval are kept as final data. The values beyond this interval are found and removed as outliers (Smith and Demetsky 1994).

This moving window is shifted from the first minute of each day to the last one and filters whole day travel time data.

As distribution of data within each time window in most cases is investigated as normal, the “k” factor is considered as 1.4826. For finding the distribution of data within each window, the normal probability plot of the data is generated for sample windows and it is found that it is normal distribution.

After removing outliers from each section’s travel time values by this method, the section’s travel time series is created. In this phase again, by using a moving average time window (containing three minutes before and three minutes after each minute of the day), the average of travel time values within each time window is calculated and represents that minute’s travel time value. For example, for the time 8:00 a window is defined from three minutes before (from 7:57) to three minutes after (to 8:03) and average calculated for all travel time values within this window, putting this average value as travel time for the nominated time. Finally, there would be just a data point for each minute of the day, representing that minute’s travel time and as such, the daily time series for each section of corridors are created. This sliding time window is created and moved from the first minute of day to the last minute (1440 minutes) for every day. This method of travel time calculation for each
minute is called Moving Average method. After calculating each minute’s travel
time value, daily travel time series are drawn for every day. Later for using the
created time series by this method in prediction model, the average value for five
minute intervals are replaced with each five lags in time series. This means that for
24 hours of a day, 288 (1440/5) lags are shown in time series instead of 1440 lags.
This is done because the minimum defined length for prediction horizon in the
prediction process would be 5 minutes.

3.2.2 Data accuracy calculation for estimating each section travel time

Before estimating whole corridor travel time values for any routes, it should be
checked whether the calculated travel time series for each section is reliable or not.
During the process of travel time estimation for a day, the parameter that defines the
reliability and accuracy of calculated travel time is the number of travel time data
points which are used to estimate each minute travel time. This means that it should
be some minimum number of data points inside each moving average time window
so that their average can be represented as the travel time value for that minute of the
day. For defining this minimum required number of data points for each time
window, the following statistical formula is used:

\[
 n = \left[ \frac{Z_{\alpha/2} \cdot CV}{E} \right]^2
\]

This formula can be used when the CV (Coefficient of Variation) is defined for
the population and a necessary sample size is needed to be established, with a
confidence of \((1 - \alpha)\), the mean value \(\mu\) within \(\pm E\) (Error Percentage). It means that
“\(n\)” gives the number of samples needed, when the probability of absolute relative
deviation of error being less than maximum error \(E\) is more than alpha (level of
significance). In this research, by defining maximum error as 10\% and having 90\%
confidence corresponding to \(\alpha = 0.05\), the \(Z\) coefficient is defined as 1.65
(Quiroga and Bullock 1998; Israel 1992).

For calculating the coefficient of variation (CV), all days from the same
clusters (explained in section 3.2.4) are put together and the CV is calculated for
each 15 minutes time interval. Based on these calculated CVs for the whole data, the
number of required data points for each five minutes sliding time window (2 minutes before and two minutes after each lag) is defined by using the formula in Equation 4. Here it is assumed that the calculated CV via selected days in same cluster is pretty equal to CV values for all population. Three defined steps are shown in Table 2.

Table 2: Steps for defining required number of data points for each time window on a sample section

<table>
<thead>
<tr>
<th>Step (1): All similar days together</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step (2): Calculating the CV for every 15 minutes</td>
</tr>
</tbody>
</table>
Step (3): Calculating the required numbers of data points for each 5 minutes time window

As shown in the last graph in Table 2, during off peak periods which have free flows in the section’s travel time profile, there are not many required data points for calculating travel time. However, during peak periods with higher rate of changes in travel time values, the number of required data points increases to around 60 points. In the next steps (section 4.1), whether there are enough data points for calculating travel time in each six minute time window or not is checked.

3.2.3 Whole corridor travel time estimation

After cleansing data and calculating travel time for each section separately, the next step is estimating the whole corridor travel time value by aggregating sections’ travel time. Estimating travel time with Bluetooth scanners along an arterial route of several links can be obtained by:

A) Direct matching the MAC-IDs only from upstream entrance to downstream exit of the route

B) Time-slice based travel time estimation from each link

As an example to illustrate the function of MAC address matching for travel time calculation using this technique, Bluetooth scanners are placed at intersections A, B and C where A, B and C are in order along a route. This provides travel time estimated between any two combinations of the scanner locations.

An estimation of travel time from A to C is obtained by direct matching MAC-ID’s at A and C. This straight-forward method usually has a low sample of travel time values due to loss/gain of vehicles at B. Another approach is to first estimate
time series of travel time from A to B and B to C and then apply the *time-slice* based travel time estimation model on the two series to estimate travel time from A to C (Khoei, Bhaskar and Chung 2013).

The time-slice method can be mathematically formulated as follows, considering the following variables:

- \( k \) is the time when a vehicle enters the first section;
- \( t_n \) is the entry time for the vehicle at \( n^{th} \) section;
- function \( t(i, t_i) \) is the travel time for the vehicle on section \( i \) when its entry time on section \( i \) is \( t_i \).

Time in which the vehicle completes its journey over the first section is given by \( t = k + t(1, k) \) and the entry time at the \( n^{th} \) section is given by:

\[
    t_n = k + t(1, k) + \sum_{i=2}^{n} t(i, t_i)
\]

The travel time estimated for the route (\( tt \)) is the sum of individual section travel times as below:

\[
    tt = \sum_{i=1}^{n} t(i, t_i)
\]

### 3.2.4 Data clustering and historical database creation

After calculating whole corridor travel time values for all days during the study period on three corridors, daily travel time profiles are drawn for each route. By observing the travel time patterns in daily graphs and also investigating variation of daily data, the similarity of daily time series are investigated. Based on this scenario, daily travel time profiles are classified into the following clusters:

- Day of week (Monday, Tuesday…)
- Working days
- Public holidays
- School holidays
As an example, Figure 10\(^1\) shows different cluster labels for daily travel time patterns.

![Sample daily travel time patterns for a month showing the clusters](image)

**Figure 10**: Sample daily travel time patterns for a month showing the clusters

Based on this method of clustering, similarity of travel time series is investigated within each cluster and historical databases for use in the prediction model is created.

### 3.2.5 Identifying Normal Patterns for Three Study Routes

For creating historical databases, firstly the normal patterns of time series changes are required to be found. Drawing daily travel time profiles for all routes during the study period, means these daily profiles are classified into the mentioned groups. By separating graphs into these groups, it is concluded that the travel time patterns on holidays and school holidays don’t follow the same pattern as working days. Usually on school holidays and specifically on holidays most of the corridors experience more free flow condition and do not show high peaks in their travel time profiles. This can be seen clearly in Figure 10 as well. As more congestion periods are observed on working day time series, the critical aspect of travel time prediction

---

\(^1\) The large size figure is available in Appendix (page 80)
is forecasting travel time for working days. Thus this research specifically focused on monitoring and identifying travel time behaviour on working days.

It is also found that most working days display similar recurrent travel time behaviour. Based on the identified pattern for working days, there is free flow between early morning hours and 7 am. Then congestion is experienced from 7:30 to 8:30; this congestion drops down and turns into a free flow just a bit higher than early morning free flow. This free flow usually continues till 4:30 pm when another congestion starts and does not finish till 6:30 pm. After that time, there is a free flow with pretty much the same value as before until midnight, as there are no more rush hours after the afternoon peak. Typical daily travel time series for each case study route are showed and discussed further in section 4.2.

Based on identified normal travel time behaviour for the three case study routes, some specific days are found in data that have different trends than the others. On these days sometimes a sudden congestion is seen during the free flow times, or sometimes morning peak or afternoon peak has some exaggerated high trend. As an example for this case, the daily travel time graph for 8th of May 2012 has an unexpected peak at midday around 12:30 pm. This problem can be seen clearly in Figure 11.

![Travel Time Profile for 8th of May 2012 (Coronation Dr)](image_url)

Figure 11: Travel Time Profile for 8th of May 2012 (Coronation Dr)
As shown in this figure, the unexpected congestion between 1:30 and 2:00 pm must be caused by an external reason. For investigating the cause of these abnormal trends in daily TT time series all information about the chosen arterials are studied. The information contains special weather condition reports like heavy rains or storm on specific days, and incident files. After studying all information about these specific days, it is found that most of these trends are caused by vehicle crashes, blocked lanes for construction works, traffic light faults or special weather conditions. Therefore, in the next steps of this study, these abnormal days are taken out of the historical data before using the database for further analysis.

After comparing daily TT time series within each cluster in terms of data variation at different times of day, the historical database is created. For this purpose, daily travel time series from the same days of the week (eg: Mondays) within the same group of working day, public holiday and school holiday are put continuously together. For example, for predicting the travel time for a sample Monday which is a working day, about 18 working Mondays, which had complete daily travel time series before that day, are collected and saved in a database. This is done by considering the abnormal day trends and removing them from the data. Now databases are ready to be used in the prediction model.

![Sample historical database for predicting a day](image)

Figure 12: Sample historical database for predicting a day
3.2.6 Model creation and fitting suitable SARIMA model on data

As discussed in section 2.2.1, the SARIMA \((p,d,q)x(P,D,Q)s\) model has different indicators which should be defined. In this research, for calibrating model on data and predicting the future values, the SARIMA model is created by writing codes in Matlab software. Some parts of the model are written as an ARIMA function in Matlab and some extra parts are added to the function for completing the model.

Before creation and applying suitable model on historical data, the time series estimated for each route are changed to average of travel time values for each five minute intervals. This means that each day time series has 288 \((1440/5)\) data points representing that day’s travel time values instead of 1440 \((\text{minutes of a day})\) data points.

**Stationary or non stationary (defining \(d & D\))**

At the first stage it should be defined if the historical time series are stationary or non-stationary. For this purpose, the travel time mean value for each day is calculated separately. These values are put continuously together to see if any increasing or decreasing trend is observed in the daily mean values. Performing the explained steps shows if the historical database pattern is stationary or non-stationary. If there is any increasing or decreasing trend in the data pattern, it is concluded to be non-stationary. However, if the calculated daily mean values fluctuate around a constant value and no trend is observed in the whole data profile, the historical data pattern is stationary.

This test is done based on two scenarios. The first one is checking each day’s travel time profile to see whether it is stationary or not. In this case if every single day pattern is non-stationary, one or two levels of differencing is needed to transform the data trend into a stationary one (defining ‘d’ indicator). Based on the second scenario, if the whole database pattern is non-stationary, again one or two level of seasonal differencing is needed to remove the seasonal trend from the whole historical database (defining ‘D’ indicator). Accordingly seasonality is defined as 24 hours. Therefore, by testing the historical data behavior based on these two scenarios, the “d & D” indicators are defined.

**Defining \(“p, P, q, Q”\) indicators**
The next step is defining suitable values for indicators “p, P, q, Q”. As explained before, these indicators define how much the previous continuous lags and seasonal lags contribute to forecasting the next lags. These indicators primarily are defined using Auto Correlation Factor (ACF) and Partial Auto Correlation Factor (PACF) graphs. After defining the “d & D” indicators of the model, firstly the historical database is differenced every time these indicators are defined (seasonal & no seasonal). Then, ACF and PACF graphs are drawn for the data. Basically, the “q” is defined as the point in the ACF graph where no more spike can be observed after that. Using the same method, the “Q” is defined from the seasonal ACF graph. For example, Figure 13 shows a sample ACF graph. In this graph there are two more considerable spikes at lags 1 and 6. But the difference between lag one and lag 6 is still too high. So we can conclude that there is no more huge lag after lag one. Therefore, based on this graph the initial value for “q” can be defined as 1.

![Sample Auto Correlation Function graph for investigating “q”](image)

**Figure 13: sample Auto Correlation Function graph for investigating “q”**

Based on the same scenario, the behaviour of the ACF graph is assessed from a seasonal aspect as well. For example Figure 14 again shows two huge spikes at lags one and lag 288 (first lag of next season). But again a huge spike is happened at lag one. In this case initial value for “Q” can be defined as 1 or 2.
Repeating the same procedure for PACF graphs, the related values for “p” and “P” are obtained for each database. For example based on the PACF graph shown in Figure 15, this gives the primary value of 2 for “p”. Likewise in a seasonal manner, the graph shown in Figure 16 also gives 1 as the first value for the “P”.

Figure 14: Sample Auto Correlation Function graph for investigating "Q"

Figure 15: Sample Partial Auto Correlation Function graph for investigating "p"
After defining the SARIMA model initial indicators, the model is created and calibrated on half of the created historical database. Thereafter for the next half of data the sum of the squared residuals ($SS_{res}$) from the fitted model is calculated. The sum of the squared differences from the mean of the real values ($SS_{tot}$) is also defined. Finally the coefficient of determination “$R^2$” is calculated for the model based on the following equation.

Equation 5:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

The calculated value for $R^2$ indicates how closely values are obtained from fitting a model match the dependent variable, the model is intended to predict.

Therefore, after comparing the $R^2$ for different models with different initial indicators, the final suitable indicators are found.

### 3.2.7 Predicting future travel time values

After finding the suitable indicators for each database, the created model is used for predicting travel time after the real data. In this case, the real data for each route is available until an exact time and future values after that time are predicted by the model.

Figure 16: Sample Partial Auto Correlation Function graph for investigating "P"
To predict short term travel time values, different prediction horizons are defined in this research. Prediction horizons are the number of lags (minutes), which will be predicted after real available travel time values. For example in the case of a 15 minute prediction horizon, real data is available until 7:45 and the value for 15 minutes later, which is 8:00, is predicted. Then for predicting 8:01, the real data till 7:46 is used and forecast 15 minutes later. Therefore, through defining different prediction horizons, accuracy of short term prediction is tested in variety of time periods. For this purpose the model predicts future values within prediction horizons of 5, 15, 30, 45, 60, 90 minutes.

Generally, the whole process of predicting future travel time series by SARIMA model is summarized in the following flowchart.

Figure 17: Forecasting time series process by SARIMA model

### 3.2.8 Prediction results evaluation

Several random days from each of case study corridors databases are chosen and the prediction process is performed on their data. Prediction is done within every single prediction horizon for each day. To evaluate the selected model results within
each prediction horizon, the percentage of error during congestion periods is calculated for each predicted day. For this purpose, Mean Absolute Percentage of Error (MAPE) is calculated for each prediction result. The following equation is used as the MAPE equation for comparing real travel time values and the predicted ones. In this formula $R_t$ represents the real travel time values and $F_t$ stands for the predicted ones.

Equation 6:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{R_t - F_t}{R_t} \right| \times 100$$

Finally, the prediction error rates from all three corridors are accumulated within each prediction horizon and the performance of model is evaluated based on the following conditions:

- Comparing results of prediction with considering seasonality and without it
- Comparing results with historical mean method results

Passing these steps for all three corridors means the accuracy and transferability of the model is tested for different types of arterials experiencing different traffic situations.

Summary:

This chapter showed how Bluetooth data are used in this research to calculate final travel time series for each day on selected arterials. Then it clarified how these time series are clustered and saved in databases for being used later in future predictions. Finally based on the characteristics of data showed in this chapter the final suitable SARIMA model is applied on time series and future short term travel time values are predicted. It is also explained that how the prediction results are assessed to check the reliability of prediction model on different routes.
Chapter 4: Results and discussion

This chapter illustrates how much accuracy is obtained by Bluetooth data for estimating travel time on arterial routes. Then with assuming to have the minimum required accuracy for the final travel time series for each route, it explains that how SARIMA model is used on data with and without considering seasonality to predict future values. Finally it shows how predicted values are compared to real data in terms of finding the error of prediction by the chosen model.

In this chapter firstly, the minimum required sample size for estimating each minute’s travel time value is defined based on the required confidence level (90%). For this purpose, the CV value is calculated for each 15 minute interval and it is assumed that this calculated CV represents the whole data’s CV for applying in Equation 4. Thereafter, by comparing the required number of data points with the available rate of travel time data points for each route, it is evaluated whether 90 percent confidence in travel time estimation is met.

Secondly, by putting the number of available data points for each minute’s sample size and also CV value in Equation 4, the error of travel time estimation for each minute is calculated. Accordingly the average error rate for estimating travel time for each route is calculated separately. It is also tested if there is at least one travel time data point registered for each minute of day or any gap is seen during the congestion periods.

Finally, in order to having minimum five minute horizon in prediction phase, the time series with a minute interval, is transformed to a time series with 5 minute interval, where each data point represents average travel time for 5 minutes and the new time series are used in prediction model. Therefore, prediction model have applied on the new historical time series and prediction results are evaluated by comparing them with the real data.

4.1 BLUETOOTH DATA ACCURACY RESULTS

As discussed in the previous methodology section (Section 3.2.1), after calculating travel time values for each section, the data is filtered by removing multiple detections and then the MAD method is applied for removing the rest of the
outliers. The following figures show the results of the data filtering process for a sample day (October 12th, 2011) at Coronation Drive. Figure 18 shows how data point distribution looks for a section after removing the multiple detections from the sample day. As shown in the two following figures, there are many outliers existing in arterial data due to different situations explained in section 3.2.1.

![Figure 18](image1.png)

Figure 18: Section's travel time points for a sample day on Coronation Drive after removing multiple detections (12th of October 2011)

Figure 19 also shows the same section data distribution after removing outliers by applying the MAD method (section 3.2.1).

![Figure 19](image2.png)
Figure 19: Section's travel time points for a sample day on Coronation Drive after applying MAD filtering (12th of October 2011)

As seen in this figure, for some minutes during a day, there are several travel time values registered, but for some other lags (minutes in daily graph), there are few registered data points or no data points at all. As long as these gaps (time periods with no registered point) occur during early morning or late nights, they can be ignored. However, missing data during congestion periods increase the calculation errors for estimating corridor travel time values in the next steps.

Figure 20: First section daily travel time profiles for August 2011 in Coronation Drive

Figure 20 also shows daily travel time profiles for the first section of Coronation Drive corridor after passing all mentioned filtering steps (section 3.2.1) during a study month. These types of profiles are created for all sections along the studied corridors during the eight months of study.

In the next step, based on the method explained in the methodology part, a sliding window with five minutes width (two minutes before and two minutes after each lag including the lag) is created to take average of all existing data points in the window and replace them as that minute’s travel time value. This sliding window is shifted from the first minute of the day to the last one and the specific value representing that lag’s travel time is calculated.
Figure 21 shows the time series created for the first section of Coronation Drive based on the explained sliding window scenario. In this figure, each minute of the day has only a data point representing travel time value for that section if a vehicle starts its trip within the section at that specific time.

According to section 3.2.2, the steps for defining the minimum required number of data points for each time window have done. For calculating the coefficient of variation (CV), all days within the same clusters explained in section 3.2.4, are put together and the CV is calculated for each 15 minute time interval. It is assumed here that the calculated CV based on collected study period data, shows the general CV for whole database of any specific day. For example, the following graph shows how the travel time profiles for all working Mondays during August 2011 to June 2012 from section one Coronation Drive are gathered for CV calculation.
As can be seen in this figure, there are some days which have an extra-ordinary trend, specifically during the congested hours. These days have a huge variance for the data during the peak periods. The normal peaks seen in daily travel time trends are showing congestions on arterials, but the high variance in data caused by exaggerated peaks might be considered as abnormal trend. Therefore, after looking through the recorded incidents and weather files for the same days, it is found that in all of these days there have been some problems that had caused unexpected congestion. These problems can be nominated as; vehicle crashes, traffic lights faults, blocked lanes for construction works and heavy rain and storm. After finding these days, they are removed from the databases and the CV value is calculated for the rest of working Mondays’ data.
As shown in Figure 23, after removing those days with strange patterns (according to section 3.2.4), the overall patterns for travel time during all working Mondays are fairly much following the same trend. Therefore, calculated CV values for each 15 minute time interval within this section are illustrated in the next figure.

Figure 24: Coefficient of variation graph for all working Mondays in section (1) Coronation Dr
Based on these calculated CVs for the whole data and defining the maximum acceptable error percentage, the number of required data points for each five minute sliding time window is defined by using the formula shown in Equation 4. In this research the required percentage of confidence is defined as 90%. Therefore, firstly the required number of data points for sampling is defined. Figure 25 shows the number of required data points for the previous sample section on Coronation Drive. Also by having the number of available travel time data points within each sample size, the real calculation error rate in terms of data sampling for estimating each minute’s travel time value is investigated in the next step.

As shown in this graph, during off peak periods which have free flow in the route, there are not many required data points for calculating travel time. During peak periods, in which a higher rate of change is seen in travel time values, the number of required data points increases to around 60 points per 5 minutes. Now, with having access to these values, it is important to see whether enough data points are available for calculating travel time in each five minute time window or not. Figure 26 shows the number of available data points for each five minute window of a sample day.
By comparing the trends shown in Figure 25 and Figure 26, it is investigated that just a few minutes of the day have enough available data points based on requiring 90% of confidence in travel time estimation. In other periods of a day, there are not enough data points or confidence for calculating travel time for each five minutes sliding window. Figure 27 also shows the error that is seen in travel time calculation for each minute, by using the available data points within each five minute time window. This error is calculated based on comparing the number of available data points and also required numbers of data via the method explained in section 3.2.2.
Figure 27: Travel time calculation error for each 5 minute time window by using available data (10th October 2011)

Figure 27 shows that the hypothesis of having 90 percent confidence in travel time estimation using Bluetooth data is not achieved in this sample route.

The next step after obtaining the explained results is to check if at least one data point is existing for each minute of the day for estimating the travel time (specifically during peak periods). To address this concern, the number of available data points for calculating travel time within each five minute sliding window is drawn in some graphs for some random days along the defined corridors. This helped to investigate in which sample routes there is a lack in travel time data for estimating section travel time. The following figures show the drawn profiles for each section and also whole corridor of Coronation drive for a sample day.

![Graph](image)

Figure 28: Data point numbers within each 5 minute moving window (1st section of Coronation drive 8th Nov 2011)
Figure 29: Data point numbers within each 5 minute moving window (2nd section of Coronation drive 8th Nov 2011)

Figure 30: Data points numbers within each 5 minutes moving window (Whole corridor of Coronation drive 8th Nov 2011)
Figure 28 and Figure 29 show the number of available data for each moving window and Figure 30 shows the available data points within the window for the whole corridor. In all these graphs the blue points represent travel time profile and the red dots show the number of data points within each sliding window. Comparing these graphs together proves that section-by-section travel time calculation has a higher chance of obtaining travel time points than has corridor-based. Section-by-section TT calculation has a higher sample size, but lower accuracy in terms of travel time estimation due to built in noise of the Bluetooth data (section 3.2.1), whereas corridor-based has a lower sample but higher individual vehicle travel time accuracy.

Drawing the same graphs for all three case study routes led this research to conclude that in most cases, Coronation Drive has the least one observed travel time point for each minute during the peak periods. But for Wynnum Road and Old Cleveland Road, sometimes there is no registered travel time data point for some minutes of the day during peak periods. All these statements are concluded by summarized information about available data for the three case study routes in Table 3. This table shows how much Bluetooth travel time data is registered for estimating case study routes TT during the peak periods. The percentages shown in this table are gathered, based on testing 10 random days on each route.

Table 3: Data status information for estimating travel time on case study routes during peak periods

<table>
<thead>
<tr>
<th>Data status</th>
<th>90% confidence in TT estimation</th>
<th>Having minimum one data point per each minute</th>
<th>Average error of TT estimation by available data points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronation Dr</td>
<td>70%</td>
<td>88%</td>
<td>15%</td>
</tr>
<tr>
<td>Old Cleveland Rd</td>
<td>30%</td>
<td>54%</td>
<td>47%</td>
</tr>
<tr>
<td>Wynnum Rd</td>
<td>35%</td>
<td>47%</td>
<td>56%</td>
</tr>
</tbody>
</table>

Based on this, it is found that travel time estimated by Bluetooth data does not have the required confidence level (90%). It is also found that in some cases, for two case study routes, even one registered data point is not available for estimating travel
time during congested periods. This means that some gaps exist in data during the peak periods for these routes. In the last column of the table also, by putting the number of available data points and the CV value in Equation 4, the average error rate of travel time estimation by current data is calculated. Based on this information, the average error of travel time estimation with available data is calculated between 15 to 56 percent for three routes during peak periods. Consequently, confidence level of travel time estimation by available data is averagely between 44 to 85 percent. This means that the available Bluetooth data for these urban arterials in most cases are not reliable to estimate corridors travel time with considering 90% required confidence.

**Assumptions:**

In this study, before going through the next steps of historical database creation and predicting future travel time based on Bluetooth data, the following assumption is made:

- For calculating each minute’s travel time in cases that do not have enough data points within the moving window, the width of the window is increased from five minutes to up to 20 minutes during the free flow period. It is assumed that the final estimated value would be nearest in value to the real travel time value for that minute.

- For estimating whole corridor travel time values and use in the prediction model, it is assumed that the time series created for each section based on Bluetooth data are accurate enough.

### 4.2 WHOLE CORRIDOR TRAVEL TIME ESTIMATES

Based on the Time-slice method explained before (section 3.2.3), with having each section’s travel time series, time series for the whole corridor are created for all three case study urban routes. The following three figures show typical daily corridor time series for a sample month on different routes.
Figure 31: Daily travel time series for Coronation Drive (May 2012)

Figure 32: Daily travel time series for Wynnum Road (May 2012)
Figure 33: Daily travel time series for Old Cleveland Road (May 2012)

These graphs\(^2\) show how travel time series are classified and saved in the database for investigating each day’s TT pattern. The procedure is done for all three defined corridors from November 2011 to Jun 2012. Then the historical databases are created separately for each route/day (based on the method in section 3.2.4) to predict specific day travel time values.

### 4.3 TRAVEL TIME PREDICTION

With considering daily seasonality in historical databases, the seasonal ARIMA model is applied on databases first. Then the performance of ARIMA without seasonality is also checked on the same database later.

#### 4.3.1 Applying SARIMA

Figure 34 shows the historical database for forecasting travel time future values on 19\(^{th}\) of June 2012 on the Coronation Drive corridor. This specific day is a working Tuesday. As shown in this figure, 19 available working Mondays before this specific day within the same cluster and with identified normal pattern (section 3.2.4) are put continuously together. The first half of this data is later used to calibrate and fit the created prediction model on. The second half also is used to check whether the model is fitted well on the data.

\(^2\) The large size graphs are available in Appendix (page 81 to 83).
After passing the steps for finding SARIMA model indicators (section 3.2.6) for predicting this day, a SARIMA (1,0,1)(1,0,1)288 is calibrated on the first nine days of the historical data shown in Figure 34. Defined seasonality of 288 refers to 1440 minutes of a day divided by 5. Then, based on the method explained in section 3.2.5, the R2 is calculated for the calibrated model as 0.81. This means that the fitted model predicts 81% of the variance in the historical data. This model is approved for predicting the defined day as it has the highest R2 value comparing with tested models with different indicators on this specific historical database.

Table 4: Sample model's coefficients and equation

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimated Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.021</td>
</tr>
<tr>
<td>‘AR’</td>
<td>0.968</td>
</tr>
<tr>
<td>Seasonal ‘AR’</td>
<td>-0.245</td>
</tr>
<tr>
<td>‘MA’</td>
<td>0.642</td>
</tr>
<tr>
<td>Seasonal ‘MA’</td>
<td>-0.690</td>
</tr>
<tr>
<td>Variance</td>
<td>71.265</td>
</tr>
</tbody>
</table>

Figure 34: Historical database for predicting June 19th 2012 (Coronation Drive)
\[ X_t = 0.96X_{t-1} - 0.24X_{t-288} + 0.64a_{t-1} - 0.69a_{t-288} - 0.021 \]

Then, according to the section 3.2.6, the model is used to predict travel time values or 19th of June within a variety of prediction horizons.

Table 5 shows the figures containing real travel time trend and predicted ones within different horizons for the chosen day. Graphs 1 to 6 within this table show how predicted travel time series are changed by changing the prediction horizon from 5 to 90 minutes.

Table 5: Forecasted travel time series for 19th June 2012 within different prediction horizons by SARIMA

<table>
<thead>
<tr>
<th>Horizons (Minutes)</th>
<th>Forecast</th>
<th>Real</th>
<th>Travel Time [seconds]</th>
<th>Time [Minutes]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) 5 Minute forecast</td>
<td>Real: Blue, Forecasted: Green</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) 15 Minute forecast</td>
<td>Real: Blue, Forecasted: Red</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4: Results and discussion

(3) 30 Minute forecast
Real: Blue
Forecasted: Light Blue

(4) 45 Minute forecast
Real: Blue
Forecasted: Purple

(5) 60 Minute forecast
Real: Blue
Forecasted: Yellow
In each row of this table the actual travel time profile and the predicted one is shown for each prediction horizon separately. This gives a schematic idea about the difference between real travel time trend and predictions. Based on the profiles in figures shown in Table 5, prediction error by the created model is increased by increasing the prediction horizon length. This can be proved by seeing the increase in the gap between real travel time profile and the predicted ones in each figure of this table. Figure 35 also shows all predicted trends in one view.

Figure 35: June 19th forecasts within different prediction horizons by SARIMA

For evaluating the model performance for each day it is very important to calculate the error rate encountering prediction process during the peak periods. Therefore, error rate for travel time prediction on the sample day is calculated based
on the MAPE method (Section 3.2.7) and is illustrated in Table 6 for the peak periods.

Table 6: Error rate for forecasting 19th of June 2012 by SARIMA within each prediction horizon

<table>
<thead>
<tr>
<th>MAPE for peaks Prediction</th>
<th>First peak (6:30-9:00 am)</th>
<th>Second peak (3:30-6:00 pm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 mins Forecast</td>
<td>4.5%</td>
<td>4.2%</td>
</tr>
<tr>
<td>15 mins Forecast</td>
<td>18.5%</td>
<td>26.4%</td>
</tr>
<tr>
<td>30 mins Forecast</td>
<td>31.6%</td>
<td>26.4%</td>
</tr>
<tr>
<td>45 mins Forecast</td>
<td>29.1%</td>
<td>33.4%</td>
</tr>
<tr>
<td>60 mins Forecast</td>
<td>19.5%</td>
<td>32.1%</td>
</tr>
<tr>
<td>90 mins Forecast</td>
<td>19%</td>
<td>18%</td>
</tr>
</tbody>
</table>

Illustrated information in this table also proves that error of prediction by model increases when prediction horizon is bigger. For evaluating the performance of the SARIMA model, travel time on different days on the three defined routes are predicted by SARIMA modelling. Then the accumulated error rate for all these days (during peak periods) is gathered to check the reliability of the model within each prediction horizon. For this purpose, travel time series for seven days on Coronation Drive route, six days on Wynnum Road and six days on Old Cleveland Road are predicted and error rate of prediction for each day is calculated separately by the MAPE method. Then the quantity of error during the morning peak time and after noon peak time is accumulated together for each prediction horizon. The following figures show the rate of associated error with SARIMA modelling for forecasting travel time series on arterial corridors focused on congested periods.
Figure 36: Travel time prediction error by SARIMA within each prediction horizon (Morning peak period)

Figure 37: Travel time prediction error by SARIMA within each prediction horizon (Afternoon peak period)
The calculated errors for predictions by SARIMA are shown for five to 90 minute prediction horizons separately in Figure 36 and Figure 37. The last vertical bar on the right side of each graph is related to the calculated error for predicted travel time values using the historical mean method. The historical mean method is a simple naive method for forecasting an approximate future value in time series based on calculating the average values from historical data. According to this graph, distribution of prediction error rate by SARIMA is acceptable as long as it is not above the historical mean error. From the charts it can be concluded that the SARIMA prediction model works efficiently for predicting short term future up to 30 minutes. After that the prediction error by model will go beyond the historical mean method error and there would be no point in applying SARIMA for predicting the future values.

4.3.2 Applying ARIMA

The variety of travel time profiles in congestion buildup and congestion dissipation in the historical database necessitated a check on the prediction model performance without considering seasonality. Figure 38 shows this variety in congestion buildup and dissipation in four continuous days in the 19th June historical database. In this figure line “A” shows congestion start time and lines “B” and “C” define the approximate borders for peak and congestion ending. In all these graphs, time is divided by five. Therefore each horizontal point in these graphs represents five minutes. As shown in this figure, these four continuous days have different start and end times for their congestions. Based on these graphs’ difference in congestion start times or end times is varied from three to seven points. By multiplying these points to five minutes, it would be 15 to 35 minutes difference between the congestion start or end times for each two continuous days. Therefore, the ARIMA model without considering seasonality is applied on same databases.
Based on this idea, for predicting travel time values on the same sample day (19th June 2012), an ARIMA (1,0,1) is calibrated and applied on the same database. Then travel time profile for the day is predicted by the new model and figures of results are gathered in Table 7.

Table 7: Forecast travel time series for 19th June 2012 within different prediction horizons by ARIMA

In all the graphs shown in this table vertical axes show travel time in seconds and horizontal axes show time of day divided by 5 in minutes

(1) 
5 Minutes forecast
Real: Blue
Forecasted: Green
(2) 15 Minutes forecast
Real: Blue
Forecasted: Red

(3) 30 Minutes forecast
Real: Blue
Forecasted: Light Blue

(4) 45 Minutes forecast
Real: Blue
Forecasted: Purple
Based on the same scenario explained for SARIMA model in section 4.3.1, the figures shown in this table also illustrate the gap between the real travel time profile and predicted one. Figures shown in this table illustrate that the created ARIMA model on the same database has better results in terms of predicting time series shape. This means that the predicted travel time profiles within each prediction horizon have more similarity with the real one. However, in this case, by increasing the prediction horizon from five minutes to 90 minutes, an increasing lag is located between the real TT profile and predicted ones. Figure 39 shows these statements better.
Figure 39: June 19th forecasts within different prediction horizons by ARIMA

A simple comparison between the shown images in Figure 35 & Figure 39 clarifies that the predicted trends, without considering seasonality, are more matched with the real travel time profile in terms of their shapes. But the only issue seen in the predicted time series by ARIMA is the increasing lags between real data and predicted ones in this case. In this situation the created lag between the real travel time profile and the predicted one is measured as the length of prediction horizon. This means that for an example, predicted travel time series within 15 minutes, prediction horizon has a 15 minutes lag with the real time series.

Therefore, for calculating the error of prediction the MAPE method is applied as discussed for SARIMA prediction results. For having a quantifiable criterion for comparing the two models’ results on the same sample day, the encountering error for forecasting the day’s travel time by ARIMA (without seasonality) is gathered in Table 8.

Table 8: Error rate for forecasting 19th of June 2012 by ARIMA within each prediction horizon

<table>
<thead>
<tr>
<th>MAPE for peaks Prediction</th>
<th>First peak (6:30-9:00 am)</th>
<th>Second peak (3:30-6:00 pm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 mins Forecast</td>
<td>7.9%</td>
<td>13.3%</td>
</tr>
<tr>
<td>15 mins Forecast</td>
<td>29%</td>
<td>24.2%</td>
</tr>
<tr>
<td>30 mins Forecast</td>
<td>36.5%</td>
<td>29%</td>
</tr>
<tr>
<td>45 mins Forecast</td>
<td>39%</td>
<td>36%</td>
</tr>
<tr>
<td>60 mins Forecast</td>
<td>48.8%</td>
<td>44.6%</td>
</tr>
<tr>
<td>90 mins Forecast</td>
<td>50.6%</td>
<td>47.7%</td>
</tr>
</tbody>
</table>
As can be seen in Table 8, the error rate for predicting the same day by ARIMA has increased a lot due to the lag caused between real and predicted travel time profiles. The same thing has happened to other sample days used in the prediction process by SARIMA. Therefore, it is concluded that the ARIMA model would not be applicable to use for prediction future travel time values on urban arterials.

### 4.3.3 Checking SARIMA model reliability

In the case of considering seasonality for forecasting travel time on arterials, depending on how much the day travel time profile is similar to the days before, the accuracy of prediction is different. Therefore, it is required to check the reliability of prediction by SARIMA using $E_{90}$ indicator. $E_{90}$ is the 90th percentile of absolute percentage error of predicted values. $E_{90}$ value shows the value that 90% of predictions have less error than that value. The represented result figures in Table 9 approve the previous results assessment for all three routes with different traffic conditions.

**Table 9: Evaluating reliability of SARIMA model**

<table>
<thead>
<tr>
<th>Prediction Horizon (Minute)</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>45</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{90}$ value</td>
<td>10</td>
<td>15</td>
<td>30</td>
<td>45</td>
<td>60</td>
<td>90</td>
</tr>
</tbody>
</table>

*90th percentile of absolute percentage error of predicted values by SARIMA for morning peak period*
Chapter 4: Results and discussion

Represented figures in Table 9 are an approval of result reliability for all three routes with different traffic conditions. According to these charts over the morning and afternoon peaks, in 90 percent of cases the predicted values until 30 minutes have less error than the historical mean method error. This is concluded by comparing the bars from 5, 15 and 30 minutes prediction horizons with the historical mean bar. Therefore based on the discussed results from all three case study corridors, the SARIMA model shows acceptable prediction results up to 30 minutes ahead during peak hours. It also shows that the travel time predictions by SARIMA method in at least 90% of the cases have better results than historical mean, with a considerable rate of difference in calculated errors.

Summary:

In this chapter firstly accuracy of travel time estimation by using Bluetooth data is evaluated. It is concluded that in some routes the required 90 percent accuracy is not met by Bluetooth data. Then by using the calculated time series for each route and creating historical database for the route, they are clustered and SARIMA model with considering seasonality and without seasonality is applied on database for predicting future. Based on this comparison the SARIMA model works much better than ARIMA in terms of predicting future travel time values on arterials until 30 minutes later.
Chapter 5: Conclusion

This research firstly analysed real traffic data gathered by Brisbane City Council exploiting Bluetooth technology. This means that by dividing urban arterials into small sections between each continuous couple of installed Bluetooth detectors and matching the same captured devices from downstream and upstream, travel time for each individual vehicle on that section is calculated. Then, after identifying and removing outliers from the data, travel time series for whole corridor is estimated and saved in databases for further calculations. This data is clustered into different groups of days of week, working day, holiday and school holiday. Then daily travel time series from the same clusters are taken out from databases later to use in prediction models. Therefore, based on the reviewed literature and characteristics of data, the Seasonal Auto Regressive Integrated Moving Average model is created, calibrated on data and applied to forecast future travel time values for some random days. This model is used to forecast future travel time series, both with consideration of the seasonality factor in historical data and also without consideration. The seasonal model and non-seasonal one are tested on the same databases and travel time on some random days from three different arterial corridors with a variety of traffic conditions is predicted. Based on calculating error rate associated with each prediction, two models are compared and both are evaluated.

5.1 Bluetooth Data Reliability

In terms of Bluetooth data reliability for estimating travel time on urban arterials, first of all a target of required 90 percent accuracy was outlined for sampling data. For this purpose a minimum required number of data points are defined for each sample window for estimating each minute’s travel time value. However, after creating sample size for estimating travel time on each minute of the day, it is found that the 90 percent confidence target is not met by available Bluetooth data in majority of cases. Based on comparing available data points with the required number of data points within each sample window, for some corridors error of estimation even reached to 70% during the peak periods. This means that some gaps are observed during the peak times in daily travel time data. Therefore,
based on the results of this study, the current rate of observed Bluetooth devices is not accurate enough for estimating travel time on Brisbane urban arterial corridors. Average confidence level for travel time estimation by available data for three case study routes is defined between 44 to 85 percent in this research. Depending on the required estimation accuracy in traffic studies it should be defined whether Bluetooth data are accurate enough to be used as the main data source or not.

5.2 SHORT TERM TRAVEL TIME PREDICTION

After Bluetooth data reliability investigation in this study, it is assumed that the travel time series estimated by using Bluetooth data are accurate and represent the real travel time values on the selected study routes. Accordingly, the next steps of travel time forecasting are done on the created database from Bluetooth data.

For predicting short term travel time on urban arterials, separate databases are created for each route. According to the explained defined clusters in analyses, historical data for predicting each day is chosen and the SARIMA model considering seasonality and without considering it is applied on the data. Finally, travel time series for several random days on each of the three case study corridors are predicted within different prediction horizons.

Based on the results taken from all three case study routes with different traffic conditions, the SARIMA model considering seasonality in travel time series’ historical data performs well up to 30 minutes after real data. This statement is driven by comparing the error rate of prediction by SARIMA to the error rate of the historical mean method. This comparison showed that the error of predicting travel time series by SARIMA increases by increasing the prediction horizon. But as the error of prediction gets more than the historical mean method, it is not acceptable to use SARIMA for predicting travel time more than 30 minutes after real data. All these conclusions are focused on congestion hours, which are more critical to be predicted in travel time series.

Testing the model without considering the seasonality factor in predictions turned the model to an ARIMA model. So by running the suitable ARIMA model on the same databases, the associated error for prediction by ARIMA is calculated as well. In this case prediction error by ARIMA has increased dramatically due to the
lag between real and predicted travel time values. This means that application of the ARIMA model for predicting travel time on urban arterials is not approved.

5.3 SARIMA MODEL RELIABILITY

As the main objective of this research for predicting future travel time values on arterials, reliability of the SARIMA model is also evaluated under different conditions. As the model is applied on different routes with a variety of travel time conditions, the 90th percentile of error for each prediction horizon’s predictions is calculated. Therefore, based on calculated 90th percentile of error for each prediction horizon, travel time prediction by the SARIMA model in at least 90% of the cases concludes more accurate results compared to the historical mean method with a considerable rate of difference in errors. The 90th percentile value for peak periods is calculated between 10 to 40 percent for 5 to 30 minutes prediction horizon, which is still less than 50 percent error of historical average method’s results. This means that in the worst case for 30 minutes prediction horizon, 90 percent of results by SARIMA have less error than 40 percent. The associated error within historical mean method is calculated as 50 percent. Accordingly SARIMA predict short term travel time better than historical mean method for up to 30 minutes after real data.

Finally, depending on the required application of the SARIMA model on travel time series - prediction or regression -considering seasonality is necessary for online travel time prediction (having no data after the current time) on routes. In terms of using the model for regression purposes such as finding missing data in travel time series in offline mode, there is no need to consider seasonality in calculations. Generally, according to the discussed results in this thesis, both these two models can show good performance in terms of time series prediction or regression methods, based on recurrent behaviour in historical databases.

5.4 FURTHER FUTURE RESEARCH

Consequently, it is concluded that the Seasonal Auto Regressive Integrated Moving Average model could potentially be applied as a linear model on historical databases for predicting short-term travel time on arterial networks. However, future studies would be needed for improving the accuracy of prediction results via consideration of incidental conditions on arterials. There would also be good potential for further study in terms of increasing the confidence of corridor travel
time estimation by improving the rates of travel time data points. In this case, fusing data from a multisource such as Bluetooth and loops is recommended.
Appendices

Sample SARIMA code for forecasting short term travel time within different prediction horizons:

%function SARIMA_Model(HDN,MonthDay)
% This function fit a SARIMA(1,0,1) model on the data and the forecast
% future TT values for each 5 minutes in different prediction horizons)
% HDN= Historical Data points Number = the number of data points save in
% the historical data base file. for example the file "Whole_Data" in
% "Jun7_Data_TW" file contains 7592 data point.
% Default calling syntax:SARIMA_Model(7592,'Jun7')

HDN =1566;

model = arima('D',0,'Seasonality',287,'ARLags',1,'SARLags',287,'MALags',1,'S
MALags',287)
    model = arima('D',0,'ARLags',1,'MALags',1)

% infilename = [MonthDay '_Data_TW.mat'];
% load(infilename);

fit=estimate(model, Jun4_peaks_data_TW);

Outfilename = [MonthDay '_forecasted101.mat'];

%%%% %%%% forecasting by fttied model:

%%%% %%%%% for 5 mins forecast:
Forecasted5=[];
for i=0:1:87
    [Yf YMSE] = forecast(fit,1,'Y0',final_peaks_jun4(1:HDN+i,1));
    Forecasted5=[Forecasted5;Yf];
end
for 15 mins forecast:

```matlab
Forecasted15=[];
for i=0:1:87
    [Yf YMSE] = forecast(fit,3,'Y0',final_peaks_jun4(1:HDN+i,1));
    Forecasted15=[Forecasted15;Yf(3,1)];
end
```

for 30 mins forecast

```matlab
Forecasted30=[];
for i=0:1:87
    [Yf YMSE] = forecast(fit,6,'Y0',final_peaks_jun4(1:HDN+i,1));
    Forecasted30=[Forecasted30;Yf(6,1)];
end
```

for 75 mins forecast

```matlab
Forecasted75=[];
for i=0:1:87
    [Yf YMSE] = forecast(fit,9,'Y0',final_peaks_jun4(1:HDN+i,1));
    Forecasted75=[Forecasted75;Yf(9,1)];
end
```

for 70 mins forecast

```matlab
Forecasted60=[];
for i=0:1:87
    [Yf YMSE] = forecast(fit,12,'Y0',final_peaks_jun4(1:HDN+i,1));
    Forecasted60=[Forecasted60;Yf(12,1)];
end
```

Forecasted90=[];
for i=0:1:87
    [Yf YMSE] = forecast(fit,18,'Y0',final_peaks_jun4(1:HDN+i,1));

    Forecasted90=[Forecasted90;Yf(18,1)];
end

save(Outfilename);
% end
Figure 20: First section daily travel time profiles for August 2011 in Coronation Drive
Figure 31: Daily travel time series for Coronation Drive (May 2012)
Figure 32: Daily travel time series for Wynnum Road (May 2012)
Figure 33: Daily travel time series for Old Cleveland Road (May 2012)
Daily travel time graphs for three case study routes:

Coronation Dr (Aug2011-Jun2012)

Coronation Dr (Aug2011-Jun2012)
Coronation Dr (Aug 2011-Jun 2012)

May 10152-10087-10221

Coronation Dr (Aug 2011-Jun 2012)

Jun 10152-10087-10221
Old Cleveland Rd (Nov2011-Jun2012)

Jun10680-10653-10463
Bibliography


Bajwa, S. I., E. Chung and M. Kuwahara. eds. 2003. A travel time prediction method based on pattern matching technique. *21st ARRB and 11th REAAA Conference, Cairns, Australia,* Cairns, Australia.,


Barceló, J., L. Montero, L. Marqués and C. Carmona. 2010. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth

2014 Amir Mohammad Khoei
http://dx.doi.org/10.3141/2175-03.

http://library.epfl.ch/theses/?nr=4416.


http://dx.doi.org/10.3141/2121-05.

http://dx.doi.org/10.1007/s13177-010-0012-y.

http://dx.doi.org/10.1111/j.1467-8667.2010.00697.x.

http://dx.doi.org/10.3141/2308-06.


