Performing Short-Term Travel Time Prediction on Arterials

by

Soroush Salek Moghaddam

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

As urban centers become larger and more densely developed, their roadway networks tend to experience more severe congestion for longer periods of the day and increasingly unreliable travel times. Proactive traffic management (PTM) strategies such as proactive traffic signal control systems and advanced traveler information systems provide the potential to cost effectively improve road network operations. However, these proactive management strategies require an ability to accurately predict near-future traffic conditions. Traffic conditions can be described using a variety of measures of performance and travel time is one of the most valued by both travelers and transportation system managers. Consequently, there exists a large body of literature dedicated to methods for performing travel time prediction.

The majority of the existing body of research on travel time prediction has focused on freeway travel time prediction using fixed point sensor data. Predicting travel times on signalized arterials is more challenging than on freeways mainly as a result of the higher variation of travel times in these environments. For both freeways and arterial environments, making predictions in real-time is more challenging than performing off-line predictions, mainly because of data availability issues that arise for real-time applications.

Recently, Bluetooth detectors have been utilized for collecting both spatial (i.e. travel time) and fixed point (e.g. number of detections) data. Bluetooth detectors have surpassed most of the conventional travel time measuring techniques in three main capacities: (i) direct measurement of travel time, (ii) continuous collection of travel times provides large samples, and (iii) anonymous detection. Beside these advantages, there are also caveats when using these detectors: (i) the Bluetooth obtained data include different sources of outliers and measurement errors that should be filtered out before the data are used in any travel time analysis and (ii) there is an inherent time lag in acquiring Bluetooth travel times (due to the matching of the detections at the upstream and downstream sensors) that should be carefully handled in real-time applications.

In this thesis, (1) the magnitude of Bluetooth travel time measurement error has been examined through a simulation framework; (2) a real-time proactive outlier detection

algorithm, which is suitable for filtering out data anomalies in Bluetooth obtained travel times, has been proposed; (3) the performance of the existing real-time outlier detection algorithms has been evaluated using both field data and simulation data; and (4) two different data-driven methodologies, that are appropriate for real-time applications, have been developed to predict near future travel times on arterials using data obtained from Bluetooth detectors.

The results of this research demonstrate that (1) although the mean Bluetooth travel time measurement error is sufficiently close to zero across all the examined traffic conditions, for some situations the 95% confidence interval of the mentioned error approaches 35% of the true mean travel time; (2) the proposed proactive filtering algorithm appropriately detects the Bluetooth travel time outliers in real time and outperforms the existing data-driven filtering techniques; (3) the performance of different outlier detection algorithms can be objectively quantified under different conditions using the developed simulation framework; (4) the proposed prediction approaches significantly improved the accuracy of travel time predictions for 5-minutre prediction horizon. The daily mean absolute relative errors are improved by 18% to 24% for the proposed k-NN model and 8% to 14% for the proposed Markov model; (5) prevailing arterial traffic state and its transition through the course of the day can be adequately modeled using data obtained from Bluetooth technology.

Acknowledgment

Completion of this Ph.D. research was possible with the support of several people; I would like to express my sincere gratitude to all of them.

I am extremely grateful to my supervisor, Professor Bruce Hellinga, for his valuable guidance, scholarly input, continuous encouragement, and financial support. Having the privilege of working with him, I not only learned how to be a better professional, but also learned how to be a better person.

I would also like to thank my committee members, Dr. Lina Kattan, Dr. Liping Fu, Dr. Frank Saccomanno, and Dr. John Zelek for their time and valuable comments towards the improvement of this research.

I am thankful to my friends and colleagues Pedram Izadpanah, Reza Noroozi, Hossein Zarei, Amir Ghods, Akram nour, Ehsan Bagheri, Babak Mehran, Shahram Heydary, Usama Shahda, Tae J. Kwon, You-Jin Jung, Amir Zarinbal, and Rita Hu for their kind support and great friendship.

I owe my deepest gratitude towards my parents, Nahid Rahbar Roshandel and Alireza Salek Moghaddam, for their unconditional love, support, and motivation, without whom I would never have enjoyed so many opportunities. I love you Mom and Dad.

I want to thank my brother, Sam Salek Moghaddam, for his inspiration, support and friendship.

Finally, I would like to thank my family, my wife Roshanak and my daughter Yaas, for their love, support, and understanding. My dear Roshank, without your patience and sacrifice I would not have been able to peruse my aspirations and complete this research; thank you for being by my side. My little Yaas, you gave me joy and happiness during the past couple of years, and I am grateful for that.

Dedication

This thesis is dedicated to ambitious researchers all around the world who devoted their lives for the promotion of science.

Table of Contents

List of Figures

List of Tables

Publications

Portions of this thesis have been previously published in a variety of journal papers and conference proceedings or presented at conferences. The following provides a listing of the sections of the thesis that have been previously published in whole or in part and the citation for the publication.

Journal Publications

- Content of Chapter 3 (Section 3.1) and Chapter 5 (Section 5.1)

  Salek Moghaddam, S. and B. Hellinga, "Real-Time Prediction of Arterial Roadway Travel Times Using Data Collected by Bluetooth Detectors", 2014 (In Press), Transportation Research Record.

- Content of Chapter 4 (Section 4.2)

  Salek Moghaddam, S. and B. Hellinga, "Quantifying Measurement Error in Arterial Travel Times Measured by Bluetooth Detectors", 2013, Transportation Research Record, Vol. 2395, pp. 111 - 122.

- Content of Chapter 4 (Section 4.3)

  Salek Moghaddam, S. and B. Hellinga, "An Algorithm for Detecting Outliers in Bluetooth Data in Real-Time", 2014 (In Press), Transportation Research Record.

- Content of Chapter 4 (Section 4.4)

  Salek Moghaddam, S. and B. Hellinga, "Evaluating the Performance of Travel Time Outlier Detection Algorithms", 2013, Transportation Research Record, Vol. 2338, pp. 67 - 77.

Conference Presentations (Excluding TRB Presentations)

- Salek Moghaddam, S., R. Noroozi, and B. Hellinga, "Real Time Prediction of Roadway Travel Times for Traveler Information and Proactive Control", 2014, Canadian Institute of Transportation Engineers, Conference and Annual Meeting, Waterloo Region.

- Salek Moghaddam, S. and B. Hellinga, "Real Time Prediction of Travel Time on Arterials", 2013, ITS-Canada: 16th Annual Meeting, Toronto, Ontario.

- Salek Moghaddam, S. and B. Hellinga, "Estimating Travel Times on Arterials using Data Acquired from Bluetooth Detectors", 2012, ITS-Canada: 15th Annual Meeting, Quebec City, Quebec.

Chapter 1
Introduction

Traffic congestion is one of the significant challenges that transportation jurisdictions face in urban areas. Delay, extra fuel consumption and air pollution are among the most important negative impacts of congestion. Based on an estimation that has been done by Transport Canada for nine Canadian major cities, the delay, wasted fuel, and extra Green House Gas (GHG) emissions, caused by both recurrent and non-recurrent congestion, on average cost the country $5.18 billion, $0.32 billion, and $0.07 billion respectively (in 2000 dollar value) (iTRANS 2006). In total, these costs represented 0.81% of Canada's GDP.

Due to the limited amount of public funding for roadway construction, it is impossible for transportation agencies to provide sufficient roadway capacity in response to continuous increases of traffic demand. In this way, most of the efforts for alleviating traffic congestion in recent decades have focused on traffic management strategies. Congestion management approaches such as zone pricing, demand management, parking management, coordinated signal control, ramp metering, High Occupancy Vehicle (HOV) lanes, temporary shoulder use, speed harmonization, traveler guidance, and bidirectional lanes are widely implemented by traffic authorities around the world.

However, these individual strategies are often implemented either as isolated systems (i.e. their operation is not integrated with other systems) and/or their control systems are reactive (i.e. strategies are implemented only after an event occurs and has been detected). Proactive (active) traffic management (PTM/ATM) strategies that are integrated and applied to transportation corridors (rather than just single facilities), hold the potential to cost effectively improve the utilization of existing transportation infrastructure and improve the quality of service provided by this infrastructure.

By implementation of PTM strategies, the traffic control and regulatory features of the road network are adjusted or altered in order to handle the changes in traffic conditions. Based on prevailing and future traffic conditions, PTM can dynamically control traffic demand and available capacity by using a single or combination of available strategies (Sisiopiku et al. 2009). In this way, PTM improves the efficiency of the network, which reduces the negative impacts of

congestion (both recurrent and non-recurrent) and provides reliable travel times for all users (Brinckerhoff 2010).

The number of different PTM strategies can be found in the literature, including: proactive traffic signal control (Roozemond 2001); speed harmonization (Alessandri et al. 1999); and dynamic route guidance (Wang et al. 2003). A key component of any proactive traffic management strategy is the ability to accurately predict near-future traffic conditions. The prediction of near-future traffic conditions (as defined by link travel times) for signalized roadways (i.e. arterials) is the subject of this research.

1.1 Background

1.1.1 Travel Time definition

The amount of time that it takes for a vehicle to traverse a defined segment of a roadway is considered as its travel time. Individual vehicle travel time is the stochastic outcome of interactions between travel demand and network supply. Figure 1-1 shows some of the significant factors that can affect travel time on a road segment (Tu 2008 and SHRP 2 2009).

On arterials, these variations are expected to be higher than on freeways in part due to the effects of traffic control devices and midblock maneuvers. For example, an individual vehicle travel time depends on its arrival time to the signalized intersections along the segment (whether the vehicle has arrived during the red phase or green phase, and in the former case, how long it has waited there before proceeding). In this research, we are focused on the prediction of the average travel time for some near-future period, not the travel time of individual vehicles. Consequently, individual vehicle travel time data (with the mentioned high variation) is not useful in the prediction process. Instead, aggregating the data over a short time interval (e.g. 5 minutes) and using the average travel time of the vehicles during that time interval can be more helpful. In fact, these average travel times also are random variables but with lower variation.

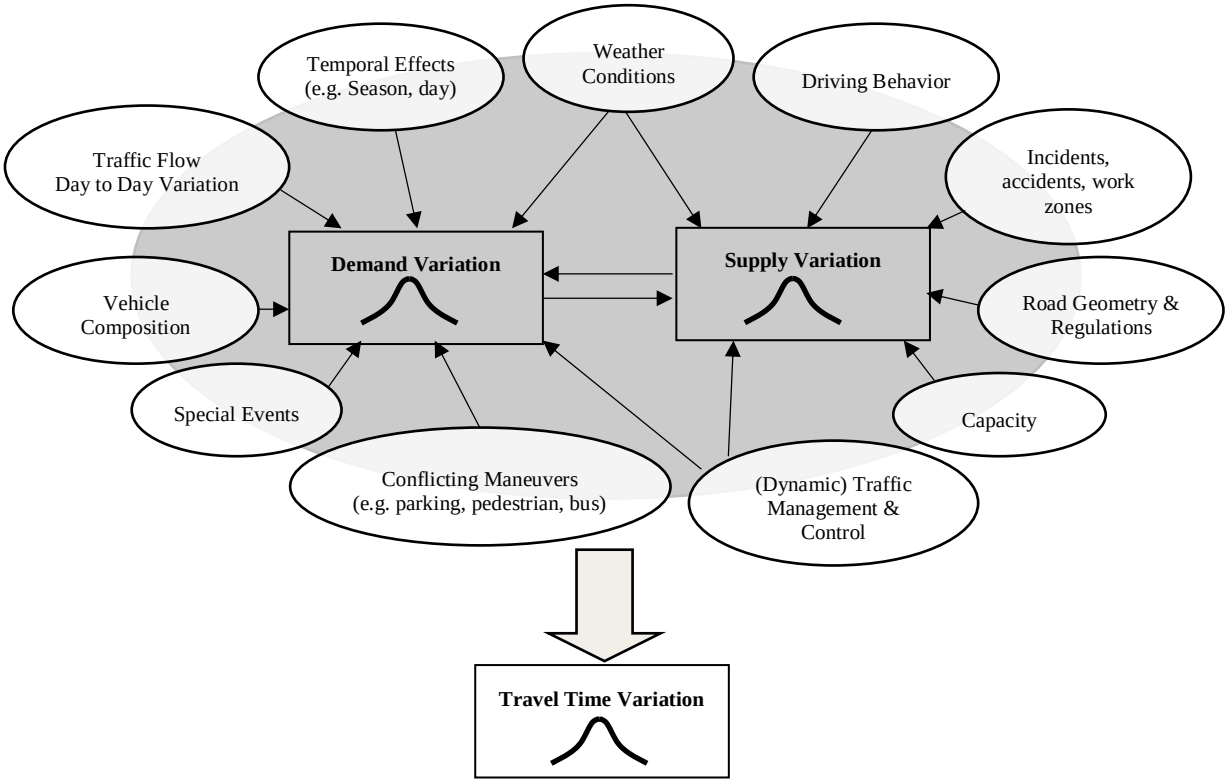


Figure 1-1: Overview of some important factors influencing travel time variability (Tu 2008 and SHRP 2 2009)

In this way, the travel time of vehicles over segment i during k^{th} time interval can be calculated based on the following equation:

$$t_{(i,k)} = \frac{1}{n} \cdot \sum_{j=1}^{n} t_{ij}^{k} \qquad \textbf{(1-1)}$$

Where, $t_{(i,k)}$ is the average travel time on segment i during polling interval k (in minutes) and t_{ij}^{k} is the travel time for the j^{th} vehicle entering segment i during time interval k. Based on this definition, the travel time for the schematic space-time diagram presented in Figure 1-2 can be calculated by averaging the travel times for vehicles *(j+1)* to *(j+4)*:

$$t_{(i,k)} = \frac{t_{j+1} + t_{j+2} + t_{j+3} + t_{j+4}}{4} \qquad \textbf{(1-2)}$$

By defining the travel time as the average of the travel times of vehicles entering the segment during any time interval, the results of travel time prediction can be used in a more meaningful

manner. For example, when providing traveler information, it is useful for drivers to know the expected time to travel a given segment of roadway before they enter the segment.

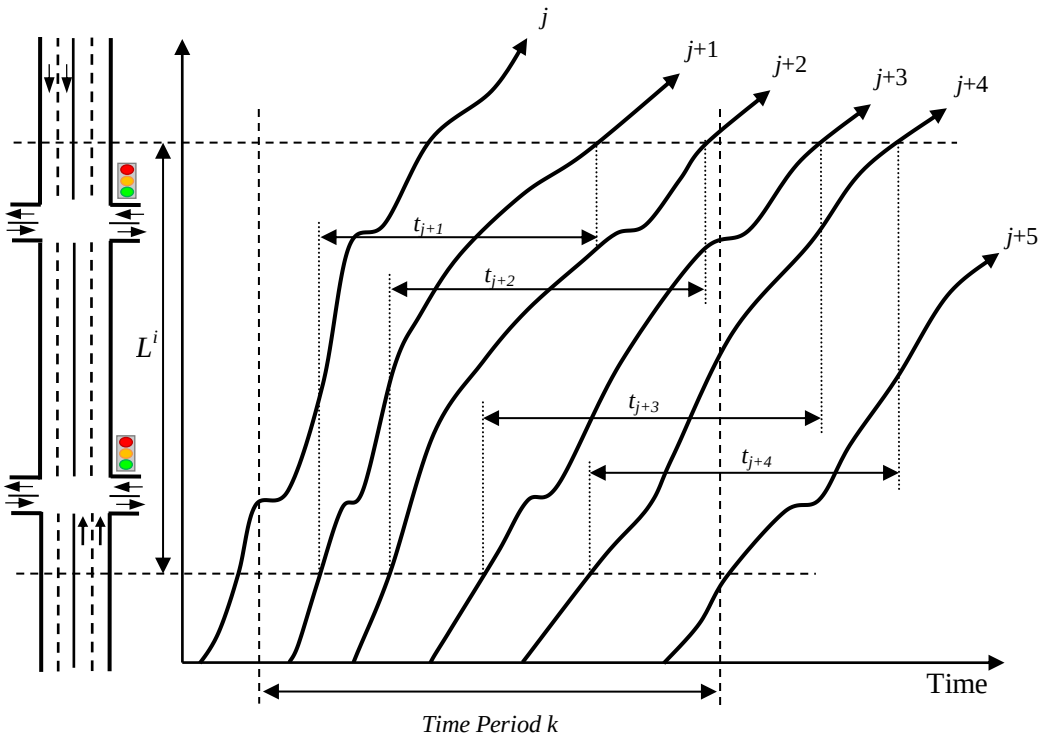


Figure 1-2: Space - time diagram for definition of travel time

1.1.2 Travel Time Estimation versus Prediction

In traffic engineering literature, "estimated" travel time over a specified segment of the roadway can be calculated using known traffic state information for a sample of completed trips over that segment. On the other hand, when the travel time calculations are done for unknown future traffic states, the result is referred to as "predicted" travel time (Van Lint 2004).

Figure 1-3 shows the trajectory of a sample of vehicles traversing road segment *i* during a specific period. Based on the definition, the travel time over time period *k* should be calculated by averaging the individual travel times of vehicles *(j+1)* to *(j+5)*. However, this is not possible as vehicles *(j+3)*, *(j+4)*, and *(j+5)* have not completed the segment yet and therefore their travel time information is not yet known. Therefore, the estimated travel time for time period *k* can be calculated by averaging only the travel times of vehicles *(j+1)* and *(j+2)* which have traversed the segment completely in this time period and their information is known:

$$t_{(i,k)} = \frac{t_{j+1} + t_{j+2}}{2} \qquad \textbf{(1-3)}$$

The travel time prediction can be done for different time horizons. Van Lint (2004) categorized these prediction horizons as: (1) real time; (2) short-term; and (3) long-term. In real time prediction, the travel time for incomplete trips is calculated using the traffic information for recently completed trips. For the case presented in

Figure 1-3, if the prior traffic state information along this segment (for example, traffic information for vehicles *j*, *(j+1)* and *(j+2)*) are utilized to predict the average time that is required by vehicles *(j+3)*, *(j+4)* and *(j+5)* to complete their trips over segment i, then the process could be categorized as "real time" prediction.

In short-term travel time prediction, the observed traffic states for recent time intervals along with historical data are used in order to predict the average travel time of vehicles for a time horizon less than one hour. For the example of

Figure 1-3, the average traffic state of vehicles during recent time intervals can be used to predict the travel time of a near future time horizon like $(k+1)^{th}$ time interval. The result of this prediction will be equivalent to the average travel time of vehicles *(j+6)*, *(j+7)*, and *(j+8)* which are a sample of vehicles that will traverse segment i in near future.

If the prediction time horizon is more than one hour, it is categorized as long-term prediction. Long-term travel time prediction models usually utilize historical traffic data as the input for forecasting future travel times.

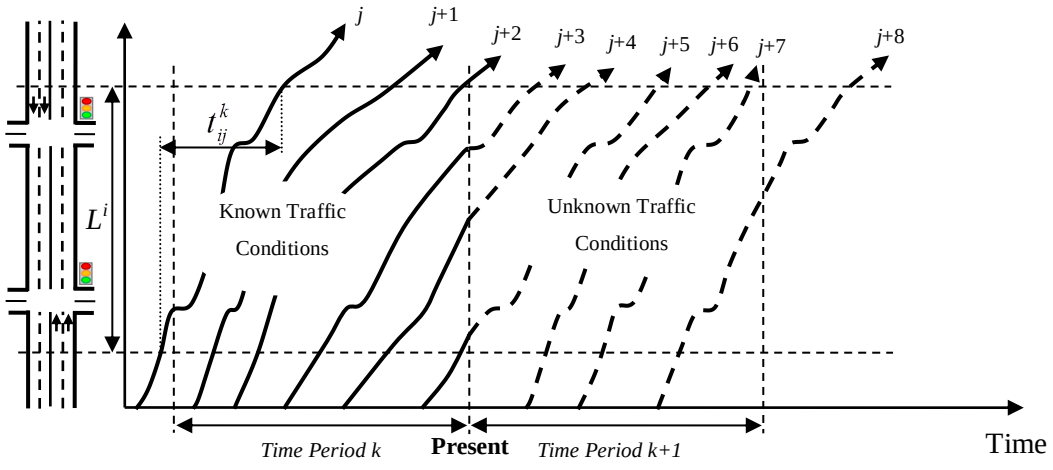


Figure 1-3: Estimation vs. prediction in space - time diagram

1.1.3 Potential challenges in travel time prediction

Performing short-term travel time prediction on arterials is complicated by a number of specific technical challenges. Some of these challenges exist for travel time prediction on both freeways and arterials while others are unique to the arterial road network in which traffic controls (such as traffic signals) are present. Some of the factors that complicate arterial travel time prediction are listed here and are discussed in more detail later in this document:

- Sampling error & Arrival time bias
- Free flow-congestion transition error
- Time lag problem
- Segment length effect
- Time interval length effect
- Signal control effect
- Coordination effect
- Signal timing variation over time

1.1.3.1 Sampling error & arrival time bias

Two types of error/bias may occur when using sample data: *sampling error* due to the size of the sample pool and *arrival time bias* due to the difference between the arrival distribution of the sample and the population. The sampling error is general and can happen on both freeways and arterials while the arrival time bias is just limited to arterials (Hellinga and Fu 1999). Here the term *arrival* refers to time of arrival of vehicles to the signalized intersection. Hellinga and Fu (2002) show that by using a stratified sampling technique the arrival time bias can be reduced significantly. It should be noted that the arrival time bias is reduced as the size of the sample increases.

1.1.3.2 Free flow-congestion transition error

Figure 1-4 illustrates the variation in segment travel time as the traffic conditions on the segment change. The transition from uncongested to congested conditions occurs at approximately time interval k. This transition is associated with a substantial increase in the segment travel time. If

only real time information (i.e. traffic states for recent time intervals) are considered for predicting the average travel time during the $(k+1)^{th}$ time interval, then there will be a significant amount of error in the predicted travel time as the model cannot predict this abrupt change due to the state transition. The same problem exists when the traffic state changes from congested to uncongested regimes. In Figure 1-4, this kind of transition occurs at time interval *(k+5)*. The non-linear relationship between travel time and traffic state suggests that it is necessary to be able to identify the traffic state (and the transition between states) in order to accurately prediction future travel times.

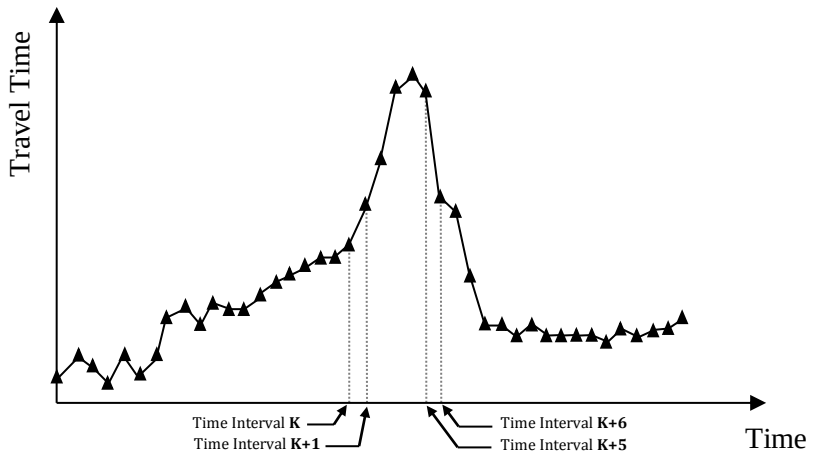


Figure 1-4: Variation in travel times associated with traffic state transitions

1.1.3.3 Time lag problem

The average travel time during each time interval can be calculated only when all vehicles that entered the segment in the time interval traverse the whole road segment. Based on this definition, it is probable that we cannot calculate the average travel times of recent time intervals as some of the vehicles that entered the segment have not yet passed through the entire segment at the current time (and as the result have not yet been detected at the downstream of the segment). To illustrate consider the sample trajectories of Figure 1-5.

At the present time, time interval *(k-2)* is the most recent time interval for which we can calculate the average travel time using the traversal times of all the detected vehicles (vehicles *(j+1)*, *(j+2)* and *(j+3)*). For the time interval *(k-1)* we cannot calculate the average travel time, but we are able to make an estimate of the average travel time using only a subset of the vehicles that entered the segment during period *(k-1)* (i.e. we can use the travel times of vehicles *(j+4)* to

(j+6) although a measured travel time for vehicle *(j+7)* is not yet available). Finally, for time interval k we cannot calculate or estimate the average travel time as none of the vehicles entering the segment in this period (vehicles *(j+8)*, *(j+9)* and *(j+10)*) have traversed the segment before time t.

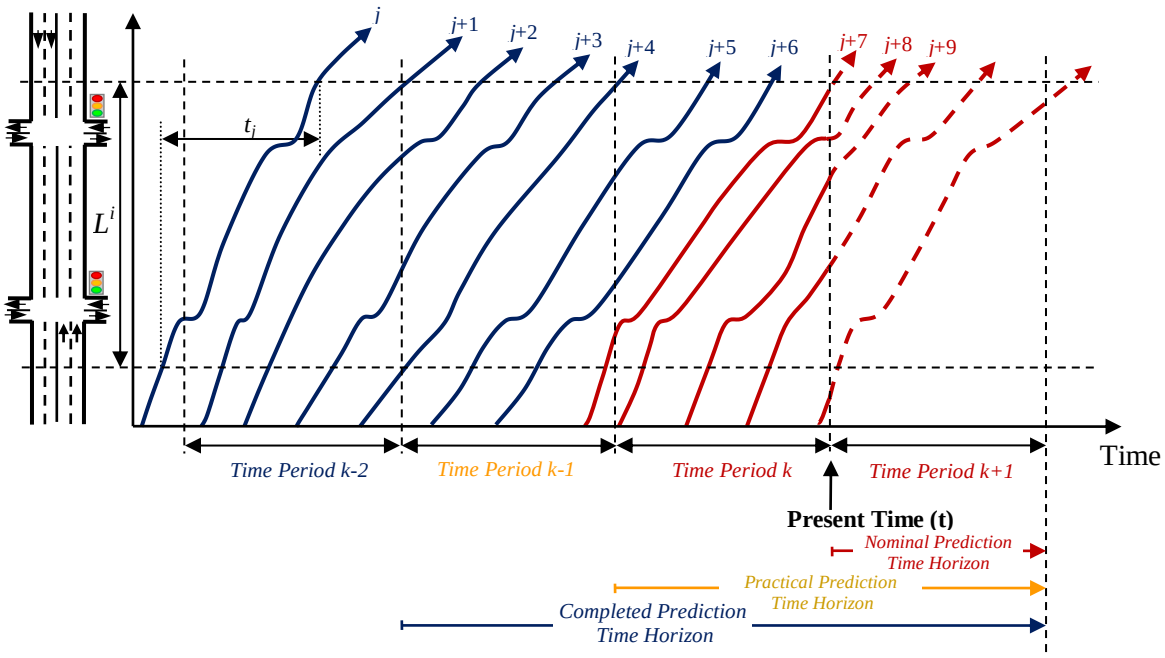


Figure 1-5: Time lag problem associated with the definition of the travel time

In the above example, as we are not able to calculate the average travel time for time interval k, we can only utilize the mean travel time estimations of prior time intervals. This issue can increase the nominal prediction time horizon by a time gap equal to at least one time interval (here time period k). In this way the prediction horizon will be equal to the summation of the mentioned time lag and the nominal prediction horizon.

The increase of prediction horizon can decrease the accuracy of the predictions, especially when there is a transition from free flow to congestion or vice versa. In other words, this source of error which can be referred as the time lag problem associated with the definition of travel time, has a compounding effect with the free flow-congestion transition error.

1.1.3.4 Segment length effect

The size of the above mentioned time lag on arterials is proportional to the length of the segment and the level of congestion. In order to reduce the time lag problem, the segment should

not be longer than approximately 2 km (in a 2 km segment with the average speed of 25 km/hr, vehicles can traverse the segment in 4.8 minutes). The shorter the segment (and therefore the shorter the traversal time), the smaller the time lag effect becomes. However, vehicle travel times are measured through the deployment of dedicated sensors and consequently, there is increasing infrastructure costs associated with decreasing the segment lengths. From a practical perspective, the minimum segment length should not be shorter than approximately 500 meters.

1.1.3.5 Time interval length effect

The length of the measurement time interval can affect the accuracy of travel time analysis. Choosing short time intervals (e.g. 1 minute) will result in biased samples. On the other hand, long time intervals (e.g. 1 hour) will cause heterogeneity in the collected samples. By averaging high and low travel times in non-homogenous samples, the stochasticity of the travel times may be ignored (Tu 2008 and Wang et al. 2003).

In addition, the selection of an inappropriate measurement time interval can lead to inaccuracies in identifying traffic states (Persaud et al. 1998; Lorenz and Elefteriadou 2000). For time intervals that are too short, short term instability in the traffic states may be considered as a state transition; however the instability is frequently short term and does not result in traffic breakdown (a traffic state transition). On the other hand, the change in traffic state cannot be detected when the measurement time period is too long. Considering these factors, the appropriate time interval duration has been reported to be between 5 and 15 minutes (Ulmer 2003).

1.1.4 Travel time characteristics on arterials

While extensive research has been done for travel time analysis on freeways, limited literature exists for arterials. The primary reason appears to be the greater complexity of travel time analysis on arterials compared to freeways. While traffic flows on freeways can be treated as uninterrupted flows, flows on arterials are impacted by conflicting maneuvers and more importantly by intersection delays. Due to these interruptions, the arterial travel times are more variable compared to freeways (Figure 1-6). On freeways, the main variations will be recorded during congestion while on arterials even in free flow conditions significant variability in travel times can be observed.

In general, travel time on arterials can be decomposed to two parts; (1) *mid-link travel time* which consists of free-flow travel time and mid-link delay, and (2) *intersection delay*. More details about these two elements of arterial travel time and their influencing factors are provided in Appendix A.

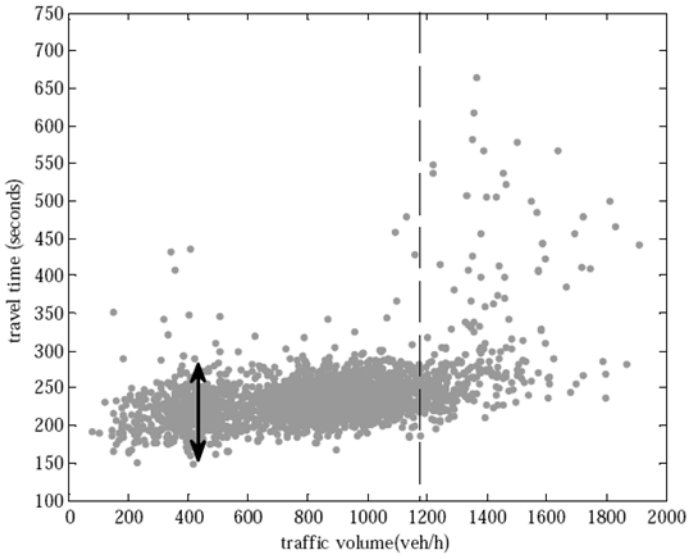


Figure 1-6: Travel time – outflow relationship on an urban arterial in the Netherlands during March 2010 - 10 minutes aggregation (Zheng 2011)

Furthermore, as discussed in Zheng (2011), travel time variability on arterials can be observed in three different time scales:

- Fast variation (time scale less than 2 minutes): Stochastic arrivals at the signalized intersections cause fast variation of travel times. In this way, even vehicles travelling a segment with very similar entry times may end up experiencing significantly different travel times (Figure 1-7).
- Medium fast variation (time scale 2-5 minutes): At signalized intersections when the flow rate is close to the capacity, even small variation of traffic flow can result in the formation of overflow queues during some cycles. In this way, in some cycles an initial queue may exist at the beginning of green interval while this may not be the case for some other cycles. The effect of this unstable overflow queues can cause medium fast variation in travel times when considering time intervals in the scale of 2 to 5 minutes.

– Slow variation (time scale 10–20 minutes): This type of travel time variation is mainly due to the stable changes in average traffic flow, and can be observed in 15 minutes time intervals. In fact, by increasing the length of time interval the fast and medium fast variation effect of other factors like arrival moments and unstable overflow queues will be eliminated and the de-noised travel times which has been referred as "underlying trend" can be achieved (van Hinsbergen et al. 2009).

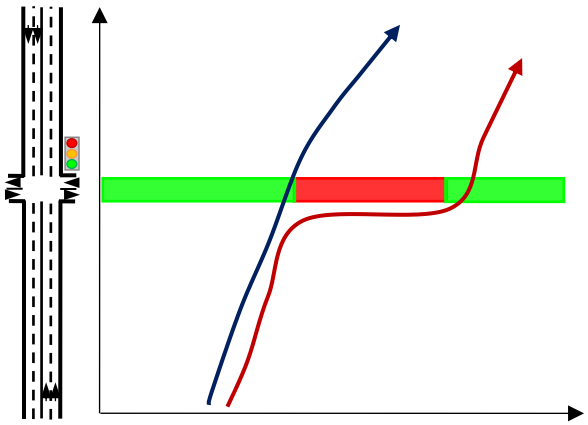


Figure 1-7: The effect of signal control and arrival time on intersection delay

As mentioned earlier, for the purpose of short-term travel time prediction, time intervals with the length between 5 and 15 minutes are recommended in the literature. In 5 minutes time intervals, both medium fast and slow travel time variations may exist.

1.1.5 Traffic surveillance devices

Travel time and speed data are valuable performance measures which can be utilized both in advanced traveler information systems (ATIS) and advanced traffic management systems (ATMS). In order to measure these data on roadways, different traffic sensing technologies have been developed. In general, these sensors can be classified in to three major categories:

– Fixed sensors: These sensors are usually installed at fixed locations along the roadway, and as the vehicles pass an individual or pair of sensors, one or more traffic characteristics (e.g. speed) can be measured. Loop detectors, Bluetooth scanners, automatic number plate recognition (ANPR) cameras and speed detectors can be included in this type of sensors.

- Mobile sensors: Trackable positioning equipment installed in probe vehicles or carried by a percentage of motor vehicles can provide travel time information of these vehicles along the traversed routes. GPS devices and cell phone sensors can be categorized as mobile sensors.
- Integrated sensors: Some systems use both fixed and mobile sensors in order to measure traffic states. Recently proposed vehicle infrastructure integration (VII) system which uses the data from both on-board equipment (OBE) and road side equipment (RSE) is a system that uses an integrated sensing system.

In recent years, Bluetooth detectors have been widely considered as an efficient and straightforward tool for measuring travel time and speed both on freeways and arterials. Each road side mounted detector continuously scans a specific radio frequency (the spectrum dedicated to the Bluetooth protocol). Bluetooth devices in detect mode (a mode in which they continuously transmit their unique machine access code (MAC)) and that pass within 50 to 60 meters of the detector are detected. The detector records the MAC address and the time of detection. The difference between the detection times of the same MAC address detected at two adjacent detector location can be used as the travel time along that segment. In this way, the travel times of vehicles having at least one Bluetooth device onboard can be measured (Figure 1-8).

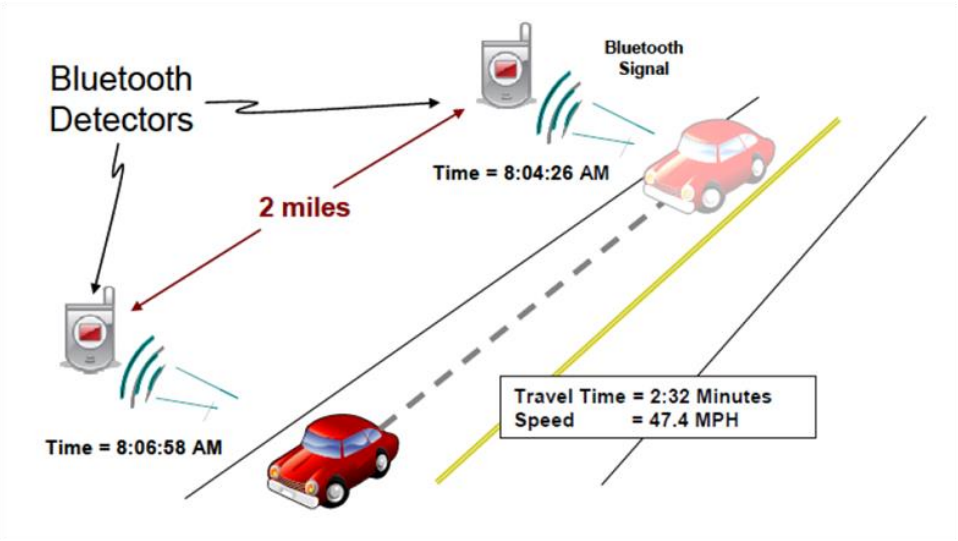


Figure 1-8: Travel time measurement using Bluetooth scanners (University of Maryland 2008)

The main advantages of using Bluetooth scanners are:

- It uses a straightforward technique for directly calculating the travel time.
- The detection process can be done anonymously.
- The technique does not violate privacy as there is no way to correlate a MAC address with a specific vehicle or with a specific person.
- One sensor can be used for both traveling directions.
- It reduces the data gathering costs. Compared to probe vehicle data collection, the cost per data point is reported to be 500 to 2500 times less expensive (University of Maryland 2008).

However, these detectors are impacted by the following sources of errors[1]:

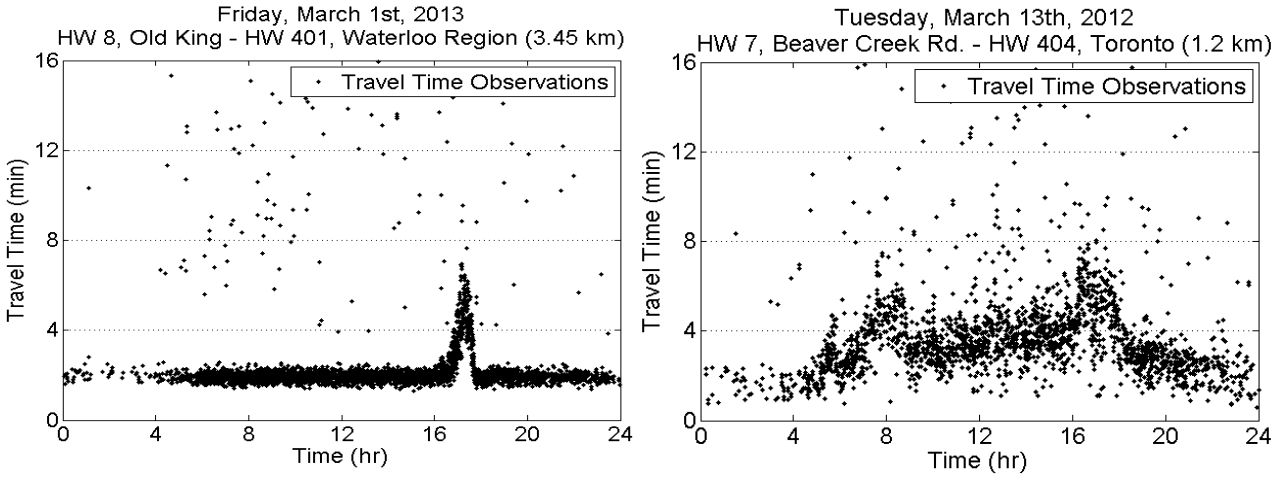
- The same Bluetooth device could be detected several times by a single scanner. The question is which moment should be considered as the true detection time. This problem becomes more crucial for short segments.
- A single vehicle may be carrying multiple Bluetooth devices which are detected. This may result in sampling bias.
- Detected Bluetooth devices may not be in an automobile. Instead, they may be contained within other motorized vehicles (taxies, buses, emergency vehicles) or be associated with non-motorized modes (cyclists, pedestrians). The travel times obtained from these other sources do not reflect the real traffic state of the segment and should be considered as outliers.
- Bluetooth detectors are not able to configure the route taken by the vehicle or if it made planned stops along the segment (mid-link sink) resulting in long travel times.

Given the characteristics of arterials versus freeways, it is expected that outliers will be more prevalent on arterials than on freeways. This, combined with the greater variation in individual vehicle travel times on signalized arterials (e.g. as a result of traffic control devices), means that the filtering of outliers is more difficult for arterials than for freeways. This concept is corroborated through the outlier pattern illustrated in Figure 1-9. In this figure the raw Bluetooth travel times, which include all the mentioned sources of errors, are presented for two different

[1] In this thesis we define "outliers" as those travel time measurements which are not associated with the traffic stream of interest. These include travel times associated with non-motorized vehicles, transit vehicles, vehicles making en-route stops, etc.

locations, one a freeway (Figure 1-9a) and the other a signalized arterial (Figure 1-9b). The arterial data exhibit much larger variability of travel times, greater number of anomalies, and more scattered outliers.

In order to detect the aforementioned outliers different filtering algorithms have been suggested in the literature (Robinson and Polak 2006; Dion and Rakha 2006; Van Boxel et al. 2011). More explanation on these filtering methods along with their associated advantages and shortcomings are discussed in Chapter 2.



(a) (b)

Figure 1-9: Travel time pattern on arterials vs. freeways

The other commonly utilized technology for travel time measurement is the use of probe vehicles equipped with GPS devices. The most useful output of this method of measurement is the trajectory information of probe vehicles along their traversed routes. However, there are some issues related to the accuracy of travel time data collected by GPS probe vehicles in urban environments:

- Positioning: At least 4 satellites are required for positioning a GPS device within 5-10 meters of its actual location. On urban arterials due to the presence of tall buildings and also other constructions like bridges and tunnels this requirement may not be satisfied (urban canyon effect). Moreover, the reflection of satellite signals by tall buildings can substantially degrade the accuracy of GPS positioning (Modsching et al. 2006).

- Travel times measured by probe vehicles cannot be directly used for a specific link or route as they are usually measured from a particular point on a link to another point on another link. In the literature, different algorithms have been proposed for decomposing the travel times measured by probe vehicles to a certain road segment (Hellinga et al. 2008; Zheng 2011).
- Nevertheless, the most important limitation is the need for vehicles to be equipped with a GPS logger and to provide this information in real time. This infrastructure is not readily available for a large proportion of the in use vehicle fleet and access to the data from those vehicles that are equipped is typically restricted, as these systems are proprietary.

1.2 Problem statement

In order to alleviate the negative impacts of traffic congestion (e.g. delay and environmental effects) and improve network performance (e.g. reliability and mobility), more efficient utilization of the existing arterial networks is required. As mentioned earlier, this can be accomplished through implementation of modern strategies like proactive signal control system and advanced user information system where the prediction of near future traffic condition (e.g. travel time) is the key component.

Travel time prediction on freeways has been extensively addressed through the traffic engineering literature (Izadpanah 2010; van Hinsbergen et al. 2009; Van Lint 2004; Van Lint et al. 2005), but very limited research has conducted to address the problem of travel time prediction on signalized arterials. This is mainly due to the complex nature of traffic flow on urban streets and high variability of travel times in this environment.

However, due to the recent developments in traffic surveillance technologies along with ever increasing traffic related problems on urban areas, during the last two decades more research has been conducted for modeling travel time on arterials.

In this regard, various traffic data sources and different modeling approaches have been employed. According to Zheng 2011), the available modeling approaches can be categorized into two main groups: Model-based methods and Data-driven methods.

Queuing theory based models (Takaba et al. 1991; Liu and Ma 2007), *traffic flow theory based models* (Oh et al. 2003; Skabardonis and Geroliminis 2005), and *cell transmission based models* (Lo 1999; Lo 2001) are among model-based methods. Although these approaches are really descriptive and are not sensitive to specific locations and times, they are very difficult to implement in real world problems. This is mainly due to the complexity of their structures and the variety of their parameters.

On the other hand, data-driven approaches include: *Pure statistical approaches* (Sisiopiku et al. 1994; Nakata and Takeuchi 2004; Davis et al. 1990; Billings and Yang 2006) and *Pattern recognition based approaches* (You and Kim 2000; Robinson and Polak 2005; Li and McDonald 2002; Liu 2008). The main disadvantages of these approaches includes, requiring extensive historical and real time data, lack of generalizability and transferability as well as not being sensitive to traffic processes.

Furthermore, the advances in real time travel time measurement technologies like Bluetooth detectors have provided the opportunity to access real-time travel time data with lower cost and higher accuracy. Continuous archiving of the collected travel time data provides invaluable historical information which can be used for prediction purposes.

In this research we propose a hybrid method in which we develop a data-driven algorithm sensitive to dynamic traffic features of the roadway while real time and historical Bluetooth obtained travel time measurements are utilized as the main source of data.

1.3 Research Goals and objectives

In order to perform short-term travel time prediction on arterials, the following research questions and objectives are defined (Figure 1-10):

Like any other travel time analysis, the development of an algorithm for short term prediction of travel time on arterials requires observed travel times. In recent years Bluetooth detectors have been deployed in many travel time analysis studies both in freeway and arterial environments (Barcelo et al. 2010; Haghani et al. 2010; Quayle et al. 2010). However, little research has been conducted to investigate the adequacy of its provided information for short term travel time prediction on arterials. Thus, the first research question is as follows:

Research question 1: Are the travel time data obtained from Bluetooth detectors adequate for short term travel time prediction on arterials or is it necessary to combine these data with data from other sources (e.g. signal timing data, spot detectors such as inductive loops, connected vehicle data, etc)?

Research Objective 1: Determine the adequacy of Bluetooth travel time data for near future travel time prediction on arterials (Box A in Figure 1-10).

Despite many promising features, there are some inherent sources of errors within this technology which can affect the accuracy of its collected data (especially on arterials). Multiple detection moments, confusion between cyclists and vehicles during congestion and en-route stops are some of these errors which should be addressed appropriately in real time. This leads to the second research question:

Research question 2: What are the sources of outliers in Bluetooth sensing technology and how can these outliers be detected on urban roads in real time?

Research Objective 2: Develop an algorithm to detect the probable travel time outliers collected by Bluetooth sensors on arterials in real time (Box A1 in Figure 1-10).

Traffic states on freeways can be dichotomized into two different regimes namely uncongested and congested. The congested state occurs whenever demand exceeds the capacity. However, on arterials the traffic states cannot be defined using the same logic (Wu et al. 2011; Yoon et al. 2007), as the traffic flow rarely exceeds the mid-link capacity (due to the presence of intersections). In fact, the traffic state on arterials can be distinguished based on the level of saturation for signalized intersections, as the delay for oversaturated and under-saturated conditions varies significantly.

Besides, one of the most critical challenges in short-term travel time prediction is the detection of traffic state transition for near-future time intervals. The anticipation of abrupt changes in travel time during transition periods not only will reduce the prediction error but also can be useful in more accurate detection of outliers. This leads to the third research question as follows:

Research question 3: Considering the available data, which criterion or set of criteria could best characterize the transition between traffic states on signalized arterials? Furthermore, how can we predict these transitions in near future time intervals?

Research Objective 3a: Define a set of criteria for describing the transition between "free-flow" condition and "congestion" (Box C2 in Figure 1-10).

Research Objective 3b: Develop and evaluate an algorithm for predicting the probable occurrence of these transitions in the coming time intervals (Box C3 in Figure 1-10).

Based on the findings through the former research objectives and by focusing on Bluetooth detectors as the main available source of data, the final research question is:

Research question 4: How can we predict near future travel times on a segment of an arterial using as an input travel time data acquired in real-time from Bluetooth detectors?

Research Objective 4: Develop and evaluate an algorithm to predict short-term travel time on a segment of an arterial using real time and historical travel time data provided by Bluetooth detectors (Box C1 and Box C4 in Figure 1-10).

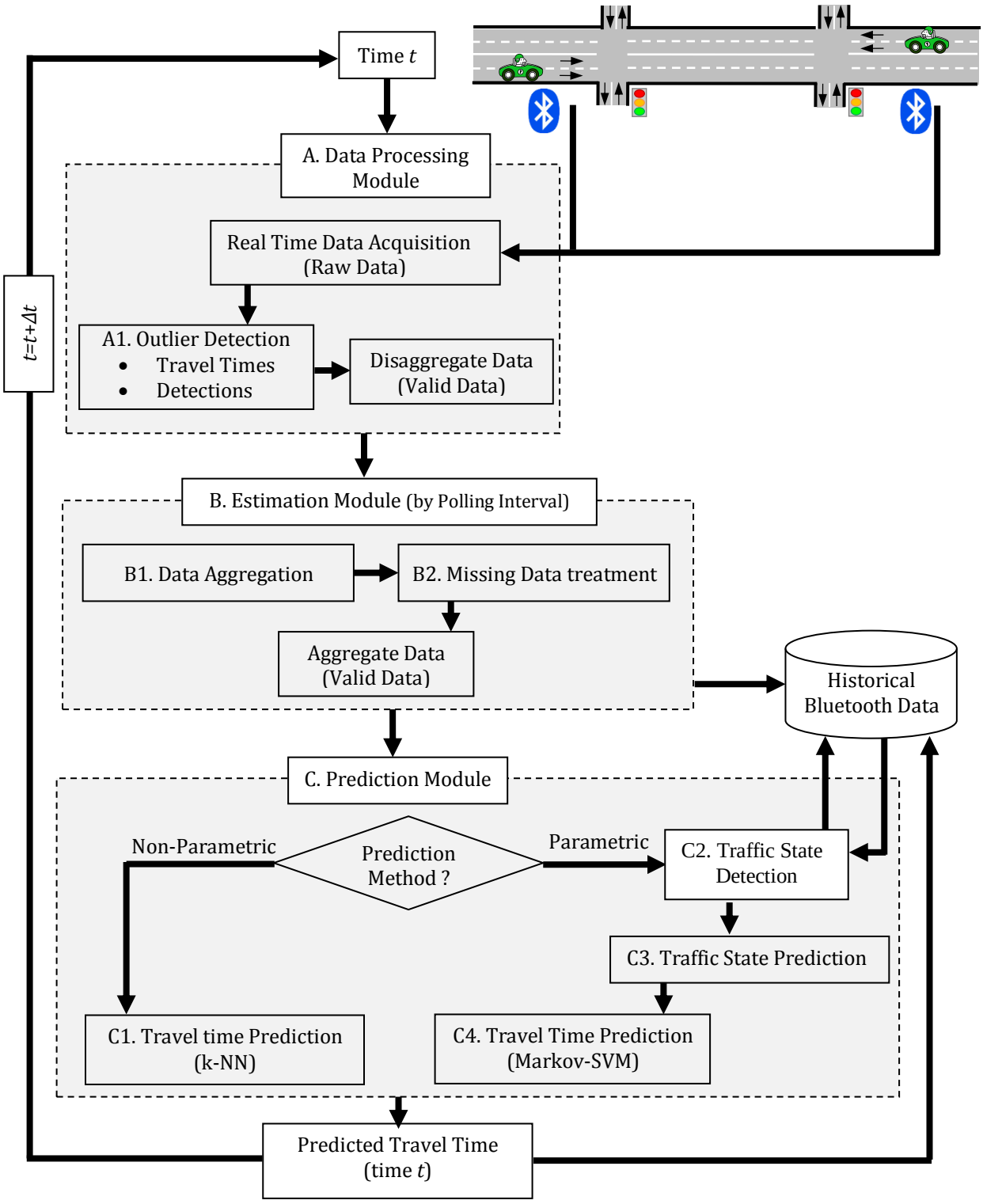


Figure 1-10: Proposed framework for travel time prediction on arterials

1.4 Thesis Outline

The reminder of this dissertation is organized as follows:

Chapter 2 identifies and describes the state of the art in the development of travel time prediction models on arterials.

Chapter 3 discusses the challenges in real time travel time analysis using Bluetooth technology and further presents the characteristics of the Bluetooth field data available in this research.

Chapter 4 identifies the Bluetooth travel time errors. Later, using a simulation approach, the magnitude of the measurement error in Bluetooth obtained travel times is quantified for arterial environments. Moreover, a proactive outlier detection algorithm is developed identifying the valid Bluetooth travel time observation in real time. Finally, the potential use of field data and simulation data in evaluation of travel time filtering techniques is investigated,

Chapter 5 describes the real time prediction approaches proposed to predict near-future travel times on arterials. The details on the formulation, calibration and evaluation of the developed Non-Parametric and Parametric travel time prediction approaches are also provided in this chapter.

Chapter 6 summarizes the research conclusions and contributions of this research and provides the recommendations for further studies.

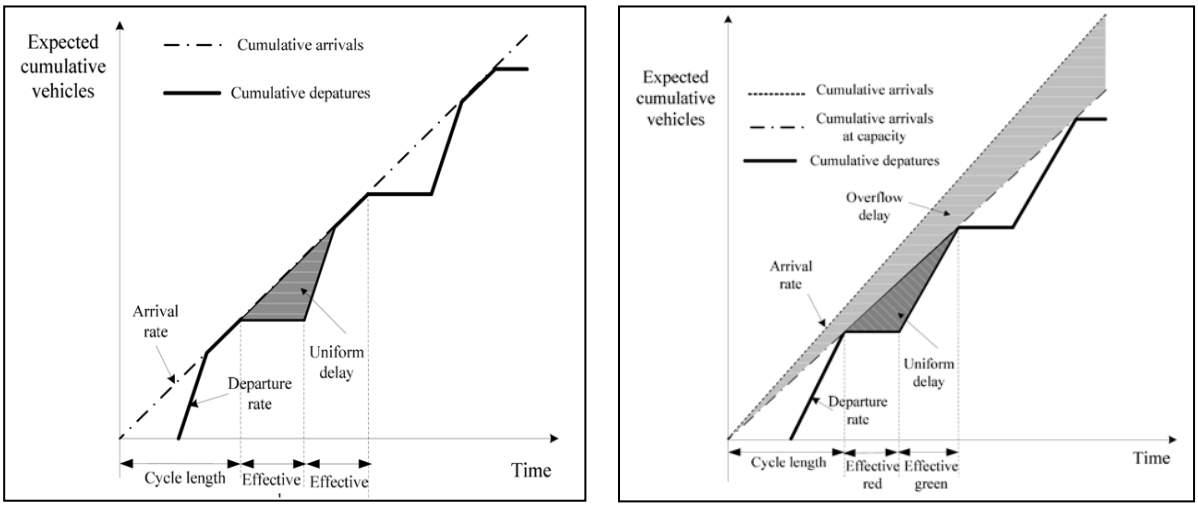
Chapter 2
Literature Review

This chapter identifies and describes previous studies relevant to the defined objectives of this research. The reminder of the chapter is organized as follows. First the basic traffic flow theory for signalized intersections is presented. Then, the concept of travel time measurement using Bluetooth technology including the available outlier detection algorithms is discussed. Further, the characteristics of traffic states on arterials are reviewed. Finally, the available research in short term travel time prediction on arterials is summarized.

2.1 Basic traffic flow theory for signalized intersections

As discussed in Chapter 1, the delay at signalized intersections explains a large part of the total delay experienced on arterials. Study of the relationship between the affecting factors and the delay at signalized intersections was the main focus of many research works during the last 60 years. Here, in order to gain more insight into this issue, an overview of the endeavored research has been provided.

Queue theory based models are among the best recognized models which have been utilized to simulate the operation of signalized intersections. Figure 2-1 shows the components of the delay at signalized intersections assuming deterministic uniform arrivals and departures. Based on the common terminology, the under-saturated delay is usually referred to as the uniform delay while the over capacity delay that vehicles experience during over-saturated conditions is referred to as overflow delay.



(a) Under-saturated conditions (b) Over-saturated conditions

Figure 2-1: Signalized intersection delay components with uniform arrivals and departures (deterministic queue theory) (Zheng 2011)

In real world the arrival of vehicles at signalized intersections is random. In order to capture the randomness of arrivals, Webster (1958) proposed the following delay function under steady state conditions (i.e. average arrival rate is less than the average service rate):

$$d = \frac{r^2}{2c(1-\rho)} + \frac{x^2}{2\lambda(1-x)} - .0.65\left(\frac{c}{\lambda^2}\right)^{\frac{1}{3}} x^{2+5\left(\frac{g}{c}\right)} \quad \textbf{(2-1)}$$

Where, r is the duration of red interval (sec), c is the cycle length (sec), ρ is the ratio of average arrival rate (veh/sec) to average service rate (veh/sec) ($\rho = \lambda/\mu$), and x is the ratio of approach arrivals to approach capacity ($x = \lambda c/\mu g$).

The first term in Webster's expression is the deterministic average delay, the second term is the random effect of arrivals (assuming Poisson distribution) and the last term is an empirical correction factor calculated through simulation.

When the traffic demand approaches the capacity, the steady state assumption in Webster's model no longer is valid. In order to properly estimate the delay during the over-saturated conditions, Kimber and Hollis (1979) used the coordinate transformation technique. In their proposed method, the Webster delay equation has been manipulated in such a way to be asymptotic to the over-saturation deterministic delay line (for $x >> 1$) (Figure 2-2).

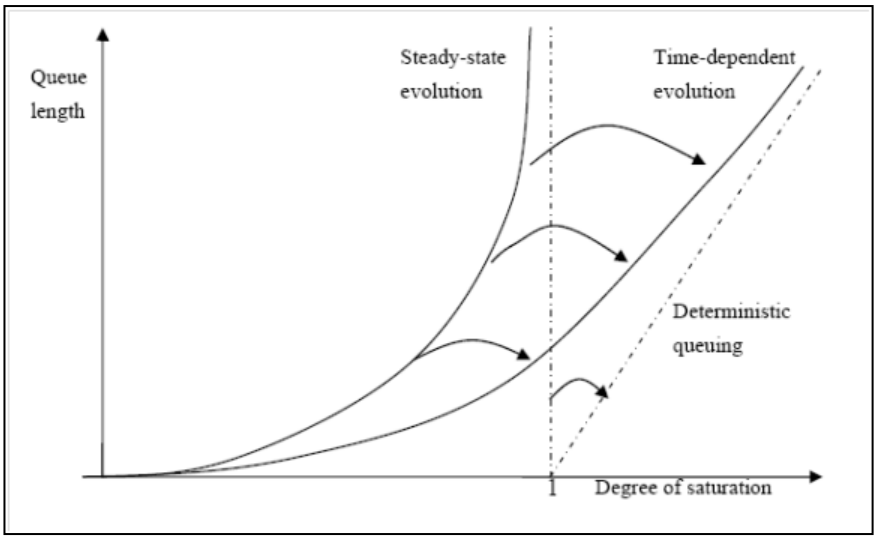


Figure 2-2: Illustration of coordinate transformation process (Viti 2006)

A general form of models developed using the co-ordinate transformation process has been given by Dion et al. (2004):

$$D = f_{PF} \cdot d_1 + d_2 + f_r \cdot d_3 \quad \textbf{(2-2)}$$

Where, d_1 is the uniform delay, d_2 accounts for the extra delay caused by the randomness of arrivals, d_3 explains the residual delay due to the existence of oversaturated initial queues at the beginning of analysis period, f_{PF} is the adjustment factor which accounts for progression in coordinated systems, and f_r is the adjustment factor for residual delay component.

The above mentioned approach has been included in most well-known capacity guides with some minor distinctions. Specifically, the Canadian Capacity Guide has only contemplated uniform and incremental stochastic delay components and ignored the initial delay part:

$$D = k_f \cdot d_1 + d_2 \quad \textbf{(2-3)}$$

$$d_1 = \frac{c\left(1-\frac{g}{c}\right)^2}{2\left[1-\frac{g}{c}\cdot\min(x,1.0)\right]} \quad \textbf{(2-4)}$$

$$d_2 = 15 t_e \left[(x-1)+\sqrt{(x-1)^2+\frac{240\,x}{C\,t_e}}\right] \quad \textbf{(2-5)}$$

Where, k_f is the progression factor, t_e is the evaluation time (min), and C is the Capacity ($S(g/c)$).

Although the discussed method has been widely implemented in the real world practices, the following issues still should be taken into account:

- The coordinate transformation technique is a heuristic approach and does not have a valid theoretical basis.
- The assumed Poisson arrival distribution is mainly applicable for isolated intersections with low traffic volumes (Viti 2006).
- Queue theory based models underestimate the delay due to their vertical queue formation assumption. Although developed shockwave theory based models (Michalopoulos et al. 1981) can address this issue, they fail to capture the stochasticity of the arrivals.

Regardless of the above mentioned shortcomings, the following points can still be observed after a thorough investigation of the discussed models:

- Signal timing specifications (g/c) affect the uniform delay more than they affect the overload delay.
- The level of coordination is mainly significant (in terms of the impact on delay) during under-saturated periods.
- The degree of saturation is one of the most important factors during the oversaturated periods.
- The evaluation period (which implies the duration of the congestion) can dramatically affect the overload delay.

2.2 Travel time measurement using Bluetooth scanners

The concept of travel time measurement using Bluetooth technology has been introduced in Chapter 1 of this document. In this section, first some real world implementation of the technology will be presented and then outlier detection methods proposed in the literature are discussed.

2.2.1 Case Studies

Haghani et al. (2010) investigated the accuracy of travel time data obtained from Bluetooth detectors as an efficient technology of freeway ground truth travel time data collection. In this regard, for different segment lengths and speed bins they statistically compared the mean travel times obtained from Bluetooth scanners and probe vehicles (collected on a segment of Interstate I-95 in Maryland). The results confirmed that the Bluetooth travel time data are not significantly different from those of probe vehicle.

In another study, Quayle et al. (2010) used this technology to perform a before-and-after analysis in order to quantify the impact of signal timing changes (along with correcting coordination and preemption settings) on a 2.5 mile signalized arterial (including 10 signalized intersections while on average carrying 15,000 to 20,000 vehicles per day) in Portland, Oregon over 27 days. In a separate analysis, they also compared the performance of Bluetooth scanners and probe vehicles in measuring travel time data, and concluded that not only does the former have the ability to collect travel times over long time periods but also it can gather considerably larger data sets at much lower costs.

Using Bluetooth detector technology, recent studies are focusing on active (real time data) and proactive (predicted data) signal adjustment systems in response to travel time observations (Day et al. 2010).

2.2.2 Outlier detection algorithms

Like any other measurement tool, Bluetooth detectors are also subject to different sources of outliers. Some of these sources have been presented in the first chapter of this document. The available literature on the outlier detection problem is mainly focused on capturing the outliers associated with *en-route stops* and *particular vehicles*.

The *multiple detection moments* problem cannot cause significant errors, as in this research instead of individual travel times; average travel times (during 5-15 minutes time intervals) will be utilized. It should be noted that the errors associated with the multiple detection moments can be assumed to be independently and identically distributed. Thus, the resulting individual errors will be cancelled out through averaging. In this way, for the purpose of consistency, the first detection moment at each station will be utilized for travel time calculation.

Due to the similarities of data obtained from automatic number plate recognition technology and those acquired through Bluetooth detectors, most of the algorithms available for the former are also applicable for the latter. This is a positive point as ANPR has been in use for more than two decades, and numerous outlier detection algorithms have been proposed for it during that time.

2.2.2.1 Overtaking rule approach

In order to identify those individual vehicles with en-route stops, Robinson and Polak (2006) proposed an overtaking rule approach which compares the travel time of each vehicle against that of some immediately following ones. The investigated travel time is valid if the following condition holds:

$$\tau_i \leq \tau_k + tt_{ik} + \Delta_C \qquad \forall k \in FV(i) \qquad \textbf{(2-6)}$$

Where, τ_i is the travel time of the subject vehicle, τ_k is the travel time of the following vehicle, Δ_C is the time threshold to determine the maximum allowable time that the subject vehicle can be overtaken by the following one (in order to being identified as a valid observation), $FV(i)$ is a set of following vehicles considered for target vehicle i, and tt_{ik} is the difference between arrival times of vehicles i and k at the upstream of the segment (considering the possibility of target vehicle being caught up by the following one because of for example a traffic light).

The performance of this algorithm depends on two parameters, Δ_C and $FV(i)$. Zheng (2011) examined the method accuracy by comparing the distribution of Bluetooth travel times identified as valid by Equation 2-6 with a distribution of travel times measured by vehicles equipped with GPS loggers (taxis). However, she found that the two distributions were not the same raising questions about the validity of the outlier identification scheme.

2.2.2.2 Statistical approaches

The available statistical based models can be classified in the following categories:

Percentile test: According to this test, an observation is considered as an outlier whenever it falls outside a predefined percentile range (e.g. lower than 10^{th} percentile and more than 90^{th}

percentile). In order to capture the time of day variability of travel time, this percentile range is defined for different time windows throughout the day. The investigated travel times are valid if one of the following conditions holds:

$$TT_i \le PT^l \quad or \quad PT^u \le TT_i \qquad \textbf{(2-7)}$$

Where, TT_i is the travel time of vehicle i, PL^l is the lower allowable percentile and PL^u is the upper allowable percentile. Using this test on a real world implementation, Clark et al. (2002) concluded that the percentile test fails to detect the outliers efficiently, as some specific percent of the data will be labeled as outliers while they may be actually valid data.

Deviation test: Based on this test an individual travel time will be detected as an outlier if its distance from the median (the median of the travel time observations in the investigated time window) exceeds a critical value. It is recommended to use median rather than mean as its value is less affected by the presence of outliers (Clark et al. 2002).

Different deviation based approaches have been proposed in the literature (Fowkes 1983; Clark et al. 2002; Haghani et al. 2010; Quayle et al. 2010). The main difference of the available methods in this category lies in their definition of deviation distance. In the approach presented by Clark et al. (2002), a range of valid travel time observations are determined around the median value. Borrowing the idea of standard confidence interval, the authors used median as the measure of location and quartile deviation (QD) as the measure of variation in order to find the range of valid observations:

$$QD = \frac{Q_3 - Q_1}{1.35} \qquad \textbf{(2-8)}$$

Where, Q_1 and Q_3 are the lower and upper quartile of the sample respectively and 1.35 is a conversion factor. Using the calculated quartile deviation, the valid travel time range can be determined as follows:

$$M_e \pm F_1(n)\, t_{\alpha/2,n^*}\, QD \qquad \textbf{(2-9)}$$

Where, M_e is the median of the observations in the investigated time window, $F_1(n)$ is the correction factor for the interval around the median (not the mean), $t_{\alpha/2,n*}$ is the approximate *t*-

statistics with $n*$ degree of freedom (which accounts for the use of QD instead of standard deviation). More details about the mentioned statistical theory along with the expressions for $F_1(n)$ and the tabulated values of n^* can be found in Freund, Miller, and Miller (1999).

The main issue with the deviation based approaches is their vulnerability to the existence of high number of outliers.

Adaptive filtering: Adaptive filtering techniques are enhanced versions of deviation tests where the validity window of the investigated time interval is established based on the variability and trend of travel times in the recent time periods. Dion and Rakha (2006) proposed two adaptive filtering algorithms which can be adopted for real-time applications.

Using a low-pass exponential smoothing algorithm, both methods estimate an expected smoothed mean travel time (*tts*) and smoothed travel time variance (σ^2_{stt}) for the current sampling interval (*k*), assuming a lognormal distribution for the travel time data:

$$tts_k = \begin{cases} e^{[(a).\ln(tt_{k-1})+(1-\alpha).\ln(tts_{k-1})]} & ; n_{vk-1} > 0 \\ tts_{k-1} & ; n_{vk-1} = 0 \end{cases} \quad \textbf{(2-10)}$$

$$\sigma^2_{stt_k} = \begin{cases} \alpha.(\sigma^2_{tt_{k-1}}) + (1-\alpha).(\sigma^2_{stt_{k-1}}) & ; n_{vk-1} > 1 \\ \sigma^2_{stt_{k-1}} & ; n_{vk-1} = \{0,1\} \end{cases} \quad \textbf{(2-11)}$$

In these equations, the expected mean travel time and its variance for the current time interval are calculated based on their associated observed and smoothed statistics, available from the previous time interval. The adaptive smoothing factor ($\alpha=f(n_{v,k-1},\beta)$), which is sensitive to the number of valid observations in the prior polling interval ($n_{v,k-1}$), determines the reliability of the observed travel times (*tt*) and furthermore should be calibrated based on local conditions (through calibration of β).

Using the smoothed average and variance of travel time, Dion and Rakha (2006) construct a validity window for the current time interval. Thus, those individual travel times (t_i) that lie within this validity interval are labeled as valid observations (Stt_k) and later can be considered for the next sampling interval calculation:

$$Stt_k = \{t_i \mid t_i \in k \ and \ tt_k^{\min} \le t_i \le tt_k^{\max}\} \quad \textbf{(2-12)}$$

$$tt_k^{\min} = e^{[\ln(tts_k) - n_\sigma \cdot (\sigma_{stt_k})]} \quad \textbf{(2-13)}$$

$$tt_k^{\max} = e^{[\ln(tts_k) + n_\sigma \cdot (\sigma_{stt_k})]} \quad \textbf{(2-14)}$$

The algorithm described thus far was considered as the "original" outlier detection algorithm developed by Dion and Rakha (2006).

These authors further argued that their original method is not appropriate for transient conditions (i.e., the formation or dissipation of the traffic congestion), when abrupt changes in the travel time are expected. This is mainly due to the fact that the mean and variance of travel time cannot be predicted only using real time data. In order to increase the capability of the developed algorithm to capture the sudden changes in the travel-time trend, Dion and Rakha (2006) proposed some adjustments to their original method which "allow the algorithm to consider as valid the third of three consecutive points outside the validity window, provided that all three observations are either above or below the validity window" (Dion and Rakha 2006). This adjusted approach is considered as the "modified" outlier detection algorithm developed by these authors.

Other statistical approaches: In a recently published paper, Van Boxel et al. (2011) proposed a model based approach for real-time detection of travel time outliers using the data obtained from Bluetooth scanners.

By assuming a constant level of market penetration of Bluetooth equipped vehicles in the traffic stream, a proxy density value has been calculated for a segment of the roadway:

$$\tilde{D}_{(i,k)} = \frac{\tilde{\tilde{v}}_{(i,k)}}{d_i} \quad \textbf{(2-15)}$$

Where, $\tilde{D}_{(i,k)}$ is the proxy density of segment i during time period k, $\tilde{\tilde{v}}_{(i,k)}$ is the number of vehicles detected by the upstream Bluetooth detector but not detected by the downstream detector at the end of time interval k, and d_i is the length of segment i.

Then, the average travel speed of vehicle j ($S_{(i,k)}^{j}$) can be calculated as follows:

$$S_{(i,k)}^{j} = \frac{d_i}{t_{(i,k)}^{j}} \quad \textbf{(2-16)}$$

Where, $t_{(i,k)}^{j}$ is the travel time of vehicle j.

As the proxy density and speed values have been calculated independently, it is possible to calibrate the following equation (a and b are model parameters) while assuming Greenshield's linear speed – density model holds (Greenshields 1935):

$$S_{(i,k)}^{j} = a \cdot \tilde{D}_{(i,k)} + b \quad \textbf{(2-17)}$$

Given that the presence of some outliers is expected, least mean sum of square error method cannot be applied for model calibration. Alternatively, least median square error is less sensitive to outliers (Rousseeuw, Leroy, and Wiley 1987) and can be used to find values for a and b ,and then the standardized residual values for all the data points can be calculated. By setting a lower and upper threshold for the standardize residuals, the points with residuals outside of these thresholds are identified as outliers.

This method is not applicable for road segments with midblock sinks and sources (which is the case for arterial segments with intermediate intersections and even for freeway segments including on or off ramps) because for these cases it is no longer possible to calculate the proxy density without having information about the midblock inflow or outflows.

2.3 Traffic State on Arterials

Unlike freeways, the issue of traffic states on arterials has not been addressed properly in the literature. Among the limited available studies, the work of Wu et al. (2011) and Yoon et al. (2007) are discussed in this section. These pairs of research have been chosen, as the former investigates the fundamental flow-occupancy relationship on arterials using loop detector and signal data, while the latter tries to establish a definition for traffic state on arterials using just GPS data.

Using two weeks of cycle-based flow-occupancy data obtained from advanced loop detectors located 400 ft upstream of the stop bar, Wu et al. (2011) investigates the impacts of signal operations on arterial fundamental diagram (AFD). The major findings of this research can be summarized as follow:

- Two capacity values have been observed in flow – occupancy diagram. This dual capacity issue can be clearly observed through Figure 2-3 for AM and PM peak hours. The authors then explain that this phenomenon is mainly due to the difference of g/c ratio, signal coordination level (bandwidth) and turning flow rates of the upstream intersection.
- Another interesting observation through the discussed AFD (Figure 2-3) is frequent data points with high occupancies. The authors showed that this issue is mainly due to the queue over detector (QOD) concept which is common on arterials. In order to remove the excessive occupancies (due to QOD), the available event based data have been used to reconstruct the trajectories of vehicles through a microscopic simulation. Having the vehicle trajectories the authors were able to exactly determine the duration of the times that the detectors were occupied just because of under-saturation queue.
- In order to attain a stable form of AFD, the effects of "two capacity values" and QOD have been filtered from the raw AFDs (Figure 2-4). The resulting stable AFDs can be used to correctly explain the state of the traffic on arterials.
- It has been reported that all the remaining high occupancy observations (infrequent observations) are related to downstream intersection queue spill back instances. This interesting result confirms that on arterials capacity can rarely be achieved as the flow rates are metered by signalized and unsignalized intersections along the roadway.

Based on these explanations, it is obvious that the conventional definition of traffic state which is applicable on freeways is not practical for the case of arterials.

Using GPS data, Yoon et al. (2007) tried to categorize arterial traffic state to "good" and "bad" classes. After investigating different traffic state indicators, they proposed spatio-temporal traffic status plot (Figure 2-5) as an appropriate state indicator. Each point in this plot represents an individual vehicle. The X coordinate is the overall mean speed (segment length divided by traversal speed) and the Y coordinate is the spatial mean speed (the arithmetic mean of instantaneous vehicle speeds along the segment - e.g. every 50 feet).

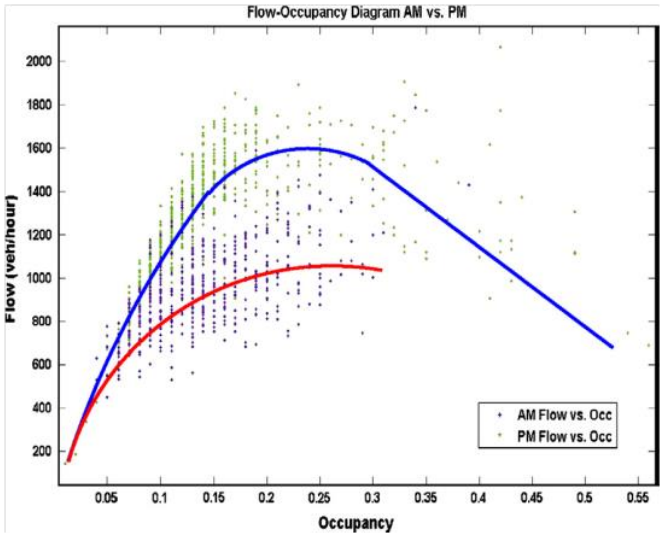


Figure 2-3: Cycle-based AFDs for AM and PM peak hours (2 weeks of data) (Wu et al. 2011)

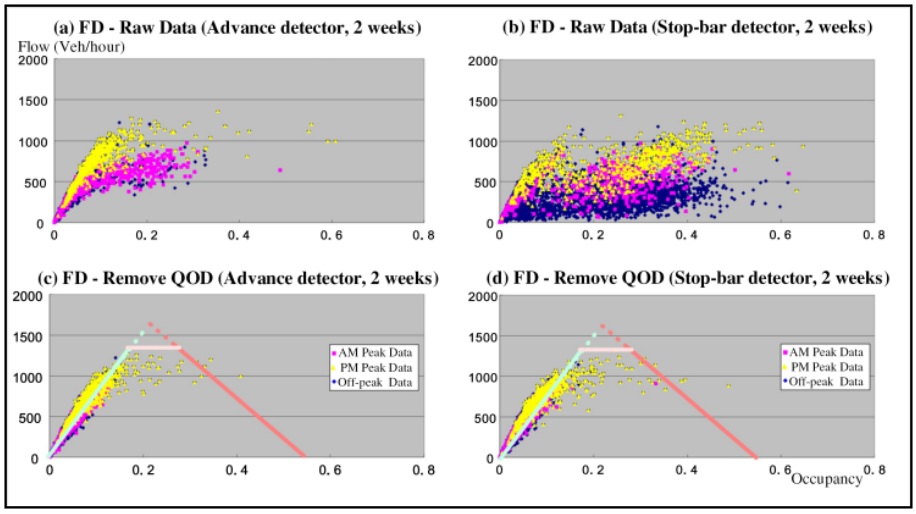


Figure 2-4: The stable AFD based on 2-weeks of data (Wu et al. 2011)

Further, the authors argued that the basic clustering algorithms cannot be utilized to classify the resulted data (spatio-temporal data) as the basic required assumptions of these algorithms do not hold for the available data (e.g. the number of clusters are not known, the data have significant overlap). Alternatively, the so called "threshold-based quadrant clustering" algorithm has been proposed. Based on this method, the spatio-temporal traffic status plot was divided into four spatially and temporally subspaces based on two thresholds (temporal speed and spatial speed thresholds).

The proposed temporal mean speed threshold can be calculated based on the following expression:

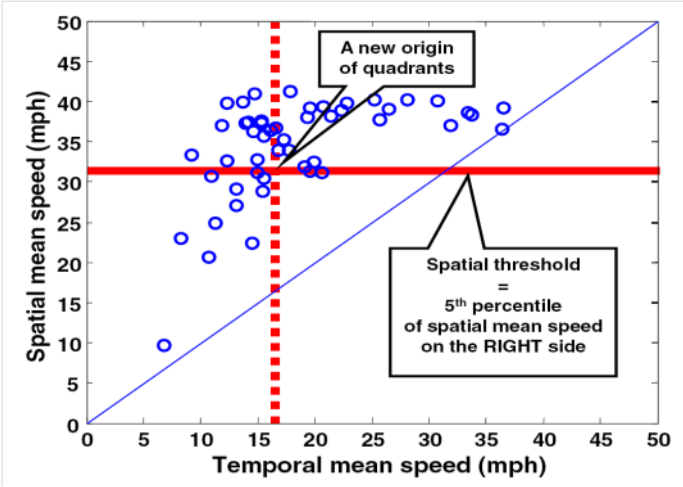


Figure 2-5: Spatio-temporal traffic status plot and four spatially and temporally subspaces (same road segment) (Yoon et al. 2007)

$$S_{(i)}^{cr} = \frac{L_i}{t_{(i)}^{5^{th}} + R_{(i)}} \qquad \textbf{(2-18)}$$

Where, $S_{(i)}^{cr}$ is the temporal speed threshold, L_i is the length of segment i, $t_{(i)}^{5^{th}}$ is the 5th percentile traversal time for segment i and $R_{(i)}$ is the red phase duration for the signalized intersection located at downstream of segment i. This red phase duration has been extracted from GPS data using 95th percentile value of recorded stopping durations.

Then, the spatial mean speed threshold has been set as the 5th percentile value of the spatial mean speed for those observations located at the right side of the temporal threshold.

Each of the resulting subspaces has its own interpretation. For example, the upper-right quadrant represents a "good" traffic state both temporally and spatially, and the lower right quadrant indicates good average temporal speed but with some "slow and go" periods. The authors considered the lower-left quadrant as the only "bad" traffic state.

However, this state classification algorithm has the following issues:

- In the upper left quadrant vehicles have low temporal but high spatial speeds. This can only happen when there is a significant delay in a small part of the segment (e.g. accident delays). In the context of traffic engineering these situations should be categorized as congestion state.

- This method cannot exactly be implemented using Bluetooth detector obtained data, as the mid-link instantaneous speeds cannot be measured through these devices.
- It can be observed (based on the spatio-temporal traffic status plot) that the traffic state is more explained through temporal mean speed threshold rather than spatial mean threshold.
- The state classification method is only applicable for those segments having just one signalized intersection.

2.4 Short term travel time prediction on arterials

As discussed earlier, the available travel time prediction models can be categorized into Model-based and Data-driven methods. The main difference between the two classes is that the model based approaches consider the traffic features of the roadway explicitly while the data driven ones do it implicitly. In this section a concise review of the most recognized approaches in each category is presented.

2.4.1 Model based approaches

In order to simulate the movement of vehicles along the roadway, model based methods aim to describe the governing dynamics of traffic flow. In this way, all the models in this category can be also referred to as simulation models.

Existing simulation models can be classified into three main groups, namely macroscopic, microscopic and mesoscopic simulation models. Using car following, lane changing and gap acceptance theories, microscopic traffic models re-construct the movement of each vehicle in the network. On the other hand, the macroscopic traffic models simulate the movement of the traffic stream using the relationship between traffic flow, density and speed at the aggregate level. Finally, the mesoscopic models set to reproduce the trajectory of individual vehicles through the speed-density relationship.

Although, the simulation models are able to explicitly explain the dynamic traffic features of the roadway, they require intensive input information and parameter calibration (Van Lint 2004). Moreover, real time application of simulation models adds extra complexities as it is required to estimate the OD and turning fractions dynamically using real time sensor data (Bierlaire and Crittin 2004; Antoniou 2004).

In order to reduce the analysis intricacy, most of the available simulation approaches on urban signalized arterials are bounded to macroscopic level. Shockwave based models and cell transmission models are among the most recognized models in this category.

2.4.1.1 Shockwave based model

Using shock (kinematic) wave theory, Skabardonis and Geroliminis (2005) developed a model to estimate travel time on urban signalized roads. Considering the spatial and temporal queuing, these researchers modeled the signal discharging process for different coordination levels. Nonetheless, the developed model requires extensive parameter calibration (e.g. free flow speed, capacity, jam density, congested wave speed) which makes it difficult to implement in real world applications.

2.4.1.2 Cell transmission model

Cell transmission is a macro simulation model which has been proposed by Daganzo (1995). The cell transmission model (CTM) is based on hydrodynamic theory which considers the traffic stream as a fluid. In this model the network is broken down to links and nodes (Figure 2-6). The traffic state (i.e. flow, speed and density) in each link is determined based on its available capacity, proportional priority rule for input links and first in / first out rule for output links (Daganzo 1995). The state of the link at time t is determined by its associated density which is estimated from sensor data (typically embedded loop detectors). The system progresses in time according to:

$$\rho_l(t+\Delta t) = \rho_l(t) + \frac{\Delta t}{\Delta x_l}\left[f_l^u(t) - f_l^d(t)\right] \qquad \textbf{(2-19)}$$

Where, Δx_l is the length of link l, Δt is the time interval, $f_l^u(t)$ is the flow rate at time t entering node from the upstream link u, $f_l^d(t)$ is the flow rate at time t exiting node from the downstream link d, and $\rho_l(t+\Delta t)$ is the density of traffic on link l at time $t+\Delta t$.

The application of CTM has been extended for signalized intersections by Lo (1999) and Lo (2001).

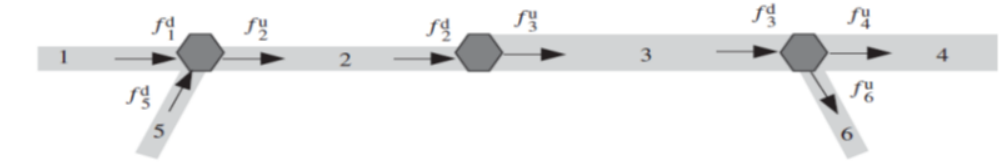


Figure 2-6: Road network representing a section of a freeway (Kurzhanskiy and Varaiya 2010)

The main challenge of applying CTM is the assumption of deterministic queue formation as well as extensive parameter calculations (e.g. real time OD matrix reconstruction). These challenges are particularly difficult for arterials.

2.4.2 Data driven based approaches

Data driven models consider the traffic process as a system and then try to relate the system responses (outputs or dependent variables, e.g. travel time) to system controls (inputs or independent variables, e.g. flow rate). Data driven approaches can be classified into two categories namely, pure statistical methods and pattern recognition approaches which will be discussed in the remainder of this chapter.

2.4.2.1 Pure statistical methods

2.4.2.1.1 Regression Methods

Regression analysis has been widely adopted for the purpose of travel time prediction on arterials (Gault and Taylor 1981; Sisiopiku et al. 1994; Nakata and Takeuchi 2004). In most of these studies, the travel time has been related to the traffic characteristic of the segment (occupancy, volume per capacity, free flow speed and upstream and downstream signal timings). The model which has been developed by Sisiopiku et al. (1994) also considered the influence of detector location as one of the significant factors:

$$t_i = \frac{L_i}{S_f} + a_0 + a_1\varphi_i + a_2 P_i^d + a_3 P_i^u + a_4 P_i^s \qquad \textbf{(2-20)}$$

Where, t_i is the predicted travel time on segment i, L_i is the length of the segment, S_f is the free flow speed, φ_i is the measured occupancy on the segment, P_i^d and P_i^u are the percentages of green time at the downstream and upstream signals, respectively, P_i^s is the ratio of the

detectors setback distance to the segment length, a_0, a_1, a_2, a_3, and a_4 are the model parameters.

The major positive point about the regression models is their simple structure which makes them attractive to be implemented in real world practice. However, the static nature of the regression models makes them unsuitable to predict dynamic traffic processes accurately.

2.4.2.1.2 Time series methods

Case studies

In order to capture the temporal correlation of travel time data, time series models have been developed by some researchers for short-term prediction purposes (Billings and Yang 2006; Yang 2007).

Billings and Yang (2006) developed a data derivative model for predicting travel time on arterials based on time series analysis. In this study, real time GPS probe data were utilized for model calibration. Due to non-stationary feature of the observed data, an ARIMA model structure was selected and calibrated for different segments of the roadway. The adequacy of the model has been tested based on residual analysis and Portmanteau lack of fit test. Based on the predicted and observed travel time trends (Figure 2-7), the authors claimed that the developed ARIMA model performs reasonably well, but did not report any value for the prediction error rate. Using the prediction results of different road segments, they further concluded that the performance of the model is better for segments with higher speed limit, longer length and lower cross-street traffic flow rates.

In another study (Yang 2007), a recursive Kalman filter technique was used to predict the arterial travel time based on travel times from previous time intervals. The required data were collected using GPS test vehicles on a 3.5 km arterial segment for seven months. The focus of this study was the analysis of the segment performance following special events (frequently held in the nearby convention center). The reported prediction mean absolute relative error (MARE) is 17.61 percent. However, these test results may be misleading as the road traffic conditions were quite similar for all events (e.g. congested). Consequently, it is expected that the model performance degrades when applied to all traffic conditions.

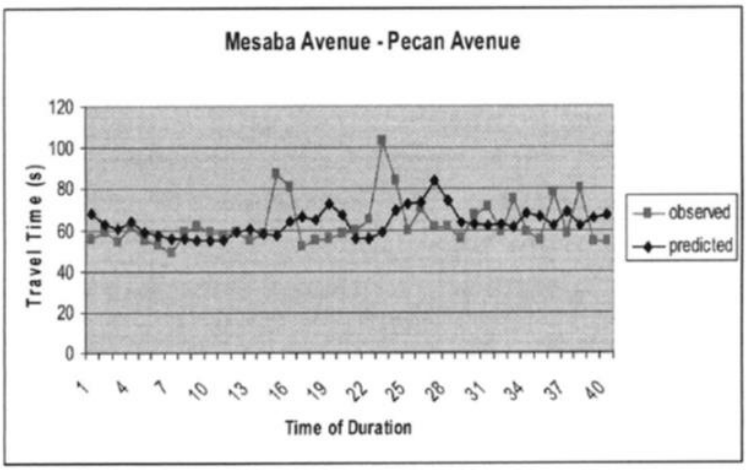


Figure 2-7: Comparison of ARIMA predictions and observed data (Billings and Yang 2006)

The most important advantage of time series models is that they can consider the time dependency of travel time data by including observations from previous intervals. However, just focusing on the trend of the data makes these models vulnerable to non-recurrent traffic behaviour or transient conditions (Figure 2-7).

The discrete Kalman filter

Kalman (1960) proposed his famous recursive algorithm for discrete linear filtering problems. Using a form of feedback control algorithm, the Kalman filter estimates the state of a process recursively. There are two equations for the Kalman filter: time update and measurement update equations. These linear stochastic equations are as follows:

$$\begin{cases} x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1} \\ w \sim N(0, Q_k) \end{cases} \quad \textbf{(2-21)}$$

$$\begin{cases} z_k = Hx_k + v_k \\ v \sim N(0, R_k) \end{cases} \quad \textbf{(2-22)}$$

Where, $x \in \eta^n$ is the actual state, $u \in \eta^l$ is the optional control input which determines the process dynamic, $z \in \eta^m$ is the measured state, w and v are the process and measurement noises, respectively, Q and R are the process noise and measurement noise variances respectively, and finally the $n \times n$ matrix A, the $n \times 1$ matrix B and the $m \times n$ matrix H are the coefficients of the mentioned equations.

In practice the operation of the Kalman filter requires the accomplishment of the following steps:

i. Using the time update equation (predictor equation), the current state and its related error covariance will be temporally projected forward and the *a priori* estimate both for process state ($\hat{x}_k$) and process error covariance (P_k) will be calculated:

$$\hat{x}_k = A\hat{x}_{k-1}^{+} + Bu_{k-1} \quad \textbf{(2-23)}$$

$$P_k = AP_{k-1}^{+}A^T + Q \quad \textbf{(2-24)}$$

ii. The Kalman gain factor (K_k) which has been derived through minimization of the *a posteriori* error covariance, will be computed:

$$K_k = P_k H^T (HP_k H^T + R)^{-1} \quad \textbf{(2-25)}$$

iii. Measurement update equation (corrector equation) applies the feedback which has been provided by a new noisy measurement (Z_k) into the *a priori* to estimate an improved *a posteriori* process state ($\hat{x}_k^{+}$) and process error covariance (P_k^{+}):

$$\hat{x}_k^{+} = \hat{x}_k + K_k(z_k - H\hat{x}_k) \quad \textbf{(2-26)}$$

$$P_k^{+} = (I - K_k H)P_k \quad \textbf{(2-27)}$$

In real world application, the measurement noise variance (R) is usually calculated off-line (prior to the filter operation) by taking some sample measurements. However, the process noise variance (Q) is more difficult to determine, as the true process cannot be observed directly. In practice these two filter parameters (R and Q) can be tuned in order to achieve the optimum performance of the filter. This tuning is usually conducted off-line in a process referred as system identification. Alternatively, the proposed covariance matching method developed by Myers and Tapley (1976) can be used instead of off-line tuning.

Furthermore, the Kalman filter can be generalized for those systems with nonlinear process and/or measurement relationship. This has been accomplished by using a method similar to Taylor series which linearizes the process and/or measurement equations around the current state

estimates. More details about this method, which is referred as extended Kalman filter, can be found in Welch and Bishop (2006).

The Kalman filter is an effective technique for estimating (also predicting) the state of a process especially when the precise nature of the modeled process is not known and also when the measurement noise is significant.

2.4.2.2 Pattern recognition based methods

2.4.2.2.1 k-Nearest Neighborhood

The theory

The *k* nearest neighborhood (k-NN) method was first referenced by Fix and Hodges (1951), but mostly due to the computation limitations, it just became popular by the early 1970s. This method is based on the simple idea of matching the investigated observation with those historical ones which have similar attributes (nearest neighbors). Thereafter, an appropriate estimator (usually the mean) of the nearest neighbors' outputs (which are known) can be considered as the current observation output. A simple clarification of k-NN having only one input variable and 5 nearest neighbors is depicted in Figure 2-8.

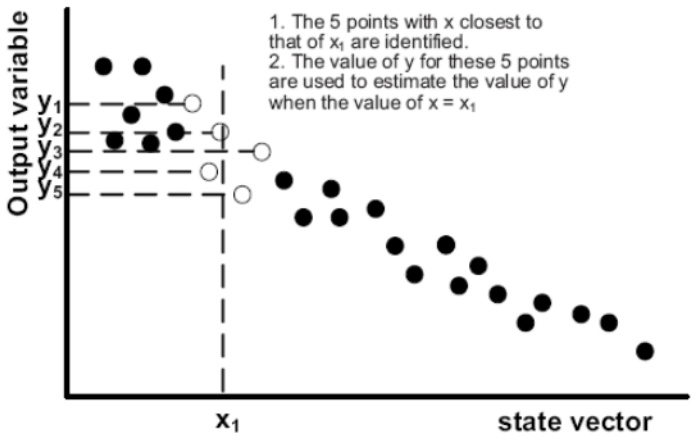


Figure 2-8: Example of the use of the k-NN method when k=5 (cited from Steve, 2005)

The following four key parameters should always be determined when using k-NN method:

- Feature Vector: The set of attributes which will be considered when comparing a current observation with the historical ones is the feature vector. Appropriate selection of attributes

in the feature vector is an important issue in the context of k-NN methodology. Selection of too few attributes may result in choosing historical observations with different underlying distributions. On the other hand, selection of too many attributes may cause the domination of irrelevant attribute(s) which will cause reduction in the effect of the relevant ones. The latter case is usually refereed as the "curse of dimensionality". More details about feature selection methods can be found in Kudo and Sklansky (2000).

- Distance Metric: The "closeness" of the current observation with any historical record should be determined using a distance metric. Euclidean, unit map, standardized Euclidean, Mahalanobis, city block, and Minkowski are among the most common distance metrics which can be used with multivariate data. However, Fukunaga and Hostetler (1973) suggest the Mahalanobis distance as the optimal one. In fact, this distance metric weights each variable using both its variance and its covariance with other variables.
- Number of nearest neighbors (k): It is identified in the literature that the optimum number of nearest neighbors is a function of sample size, dimensionality of the features space, and the underlying distribution of dependent variables against the independent ones (Fukunaga and Hostetler 1973). The same researchers also discussed about an existing trade-off between variance error (caused by the randomness of the output variable) and biased error (caused by the nonlinearity of the underlying process) in k-NN method. Further, they concluded that by increasing the value of k the variance error decreases, but the bias error increases. In this way, finding the optimum value of k is a nontrivial task which should be accomplished through optimization. More details about this issue can be found in Dasarathy (1991).
- Local estimation method: The estimated output for the current observation can be determined by applying an estimator function to the output of the k nearest neighbors. This estimator function is usually referred to as the "local estimation method". Arithmetic mean and median are among the most common local estimators in the literature, but both of them are prone to the effect of boundary bias. A robust alternative for the mentioned estimators is the use of locally weighted scatter plot smoothing (LOWESS), which fits a curve to the target values (outputs) of the nearest neighbors (Simonoff 1996; Härdle 1992).

Finally, it is to be noted that the k-NN relies on the following three main assumptions:

i. The observations that have been classified as the nearest neighbors have "similar a posteriori distributions" (Devijver and Kittler 1982). Thus, it is important to consider adequate attributes in the feature vector to assure the correct classification of observations.

ii. When using the k-NN method, a large historical database should be available. It has been shown whenever both k and N (number of historical data) tend to infinity, as long as k/N tends to zero, the chance of misclassification is minimal (Cover 1968). This is usually referred as the "Bayes risk", which occurs when the underlying process is completely known. However, the increase in the database size results in longer analysis times. Different techniques are available in the literature to reduce this problem (Dasarathy 1991).

iii. The distribution of the selected nearest neighbors around the input feature vector is random. This assumption does not hold for input feature vectors located at the boundary of existing feature space. We mentioned some techniques (e.g. LOWESS) for solving the k-NN boundary problem earlier (Simonoff 1996).

Case studies

Using the k -NN method, Robinson and Polak (2005) tried to predict the average travel time of vehicles for the next 15 minutes on a 1.06 km arterial segment. The segment is located at London's congestion charging scheme and includes 5 signalized intersections. The occupancy and flow rate data (for 37 separate days), which were measured through 3 single loop detectors along the segment, have been considered in the feature vector. Furthermore, the ground truth travel time values for individual vehicles have been obtained from automatic number plate recognition technique (ANPR) using two cameras mounted at either end of the segment. The reported mean absolute relative error (MARE) is 17.83 percent.

In another study, You and Kim (2000) applied k-NN method to predict the travel time on a 5.3 km arterial network (including 7 road links) located in Seoul, while using travel time data obtained from dedicated probe vehicles. The extracted data from the probe vehicles included distance, travel times and stopped times for each link.

The model prediction accuracy has been reported in terms MARE of 8 to 10 percent. However, these results are based on a set of 200 randomly selected prediction instances making these results difficult to interpret.

2.4.2.2.2 Neural network based approaches

Neural network based models like spectral basis neural network (SNN) (Park et al. 1999) and state-space neural network (SSNN) (van Hinsbergen and van Lint 2008; van Hinsbergen et al. 2009; Van Lint et al. 2005) have been utilized to predict short term travel time on freeways. All these models tried to learn the complex nonlinear relationship between the traffic state variables (along the route) and then use it for predicting travel time. On arterials, due to extra complexities mainly caused by the existence of intersections (e.g. turning fractions, signal timings), the application of neural networks was not very successful (Liu 2008).

Although neural network models are fast and easy to implement in real world problems, they suffer from three major problems, namely: over-fitting, lack of generalizability, and transferability.

2.5 Summary

In this chapter, four aspects of arterial travel time modeling are discussed, namely, modeling signalized intersections delays, applicable outlier detection algorithms, traffic state on arterials, and state of the art algorithms on short term arterial travel time prediction. The following list provides a synopsis of the main findings in this chapter:

- Signal timing specifications and level of coordination mainly affect intersection delay during under-saturated conditions while the ratio of volume to capacity is the most influential factor during oversaturation conditions.
- The performance of the existing outlier detection algorithms is not adequate for real-time applications. The modified adaptive filtering algorithm developed by Dion and Rakha (2006) is identified as the most efficient outlier detection algorithm from the literature. However, due to its heuristic nature, this algorithm frequently results in some unrealistic travel time estimations which restrict it from being efficiently implemented in automatic real-time applications.
- The conventional definition of traffic state on freeways is not applicable on arterials as traffic flow rarely exceeds the roadway capacity. In fact, other than traffic flow fluctuations, the formation of different traffic states on arterials highly depends on the performance of the signalized intersections.

- Travel time as the stochastic outcome of travel demand (i.e. traffic flow) and network supply (e.g. provided capacity at signalized intersections) interactions can be considered a good candidate factor for traffic state classification on arterials.
- Due to the complexity of their structures and the variety of their parameters, it is very difficult to implement model based approaches for real time arterial travel time prediction.
- Although the dependency of the travel time data can properly be captured by time series models, these models are vulnerable to abrupt changes in traffic conditions.
- The availability of a rich historical travel time database (which is the case for this research) suggests that pattern recognition based models may be most suitable for near future arterial travel time prediction purposes. However, working with a large database can result in long analysis times which may be troublesome in real time prediction processes.

Chapter 3

Real Time Acquisition of Bluetooth Data

This chapter describes in more detail the real time acquisition of Bluetooth data. The chapter is divided into two main sections. In the first section we discuss the differences between real-time and offline travel time analysis and specifically the additional challenges that real-time applications face. In the second section we identify the types of measurements that can be obtained from Bluetooth detectors and the characteristics of the available Bluetooth data.

3.1 Real time travel time analysis

3.1.1 Time lag problem in Bluetooth travel times

In practice, travel time predictions are made for the mean or expected travel time over some short aggregation interval (e.g., 5 minutes) rather than individual vehicle travel times. Furthermore, the mean travel time is defined as the average of the travel times of vehicles entering rather than exiting the segment during a given time interval. This definition of travel time is more useful in practice because we wish to provide travelers with the travel time that they will experience before they enter the segment.

Though this definition of travel time results in predictions that are more valuable for practical applications, it also introduces a computational challenge. The average travel time during each time interval can be calculated only when all vehicles that entered the segment in the time interval traverse the whole road segment. For most applications, the segment travel times exceed the analysis time interval duration, and consequently the travel times of vehicle entering the segment in interval k are not available until after interval k and we refer to this as a time lag.

It is necessary to differentiate between travel time predictions that are made in real-time and those that occur off-line (e.g. using historical data). When performing off-line predictions, all measurements made by the Bluetooth detectors up to the time interval for which the prediction is being made are assumed to be available. However, for real-time applications, the presence of the time lag problem described in the previous paragraph means that travel time measurements are not yet available for the current interval and possibly recent previous intervals. This time lag problem essentially increases the effective prediction time horizon by the length of the time lag

which in turn can decrease the accuracy of the predictions, especially when there is a transition from free flow to congestion or vice versa. More details on the mentioned time lag problem and its effect on the prediction horizon have been provided earlier in Chapter 1.

3.1.2 Real time processing of the Bluetooth detector measurements

The following primary technical challenges must be addressed when predicting arterial travel times in real-time:

- The presence of outliers
- Inconsistency in the order in which travel time data become available, as:
    - The vehicles may belong to different traffic movements at the upstream intersection and consequently may experience different levels of delay
    - The probabilistic nature of Bluetooth detection
    - Vehicles may overtake each other along the roadway

In order to address these challenges, we have developed a robust analysis framework for evaluating travel time prediction algorithms. The framework makes use of a database of historical Bluetooth detector measurements. However, the framework is constructed to provide the measurements to the prediction algorithm as if the data were being obtained in real-time. This framework is explained through the schematic diagrams provided in Figure 3-1 which illustrate the application of the process to a set of hypothetical data.

Consider Figure 3-1a. At the end of time interval *(k-1)*, there are two new observations (individual vehicle travel times) available. These observations need to be screened to determine if they are valid or invalid (e.g. outliers) observations. This screening can be completed by applying an appropriate outlier detection algorithm such as the one proposed in section 4.3. This outlier detection algorithm uses both the recent travel time estimates (here time interval *(k-2)*) and historical data to identify outliers.

When all the observed travel times in interval *(k-1)* are validated, the prediction process can be initiated. For example, the prediction algorithm proposed in section 5.1 uses the observed trend in the travel time estimations from recent time intervals (here time interval *(k-3)*, *(k-2)*, and *(k-1)*)

to find similar trends in the historical data and these historical data are used for predicting the travel time of the time interval *k* (Figure 3-1b).

As time elapses, the present time becomes the end of interval *k* and a set of new travel time measurements becomes available (Figure 3-1c). These new observations may belong to different time intervals (here they belong to time intervals *(k-2)*, *(k-1))*, and *k*) and therefore, their presence impacts outlier detection results and travel time estimates from previous time intervals (Figure 3-1d). The number of previous time intervals impacted is a function of the new data received, and therefore the outlier detection algorithm and travel time estimation processes must be reapplied starting from the earliest period for which new observations have been obtained (here time interval *(k-2)*).

Figure 3-1e to Figure 3-1g illustrate the results that are obtained from the application of the outlier detection algorithm and travel time estimation method for intervals *(k-2)*, (*k-1)*, and finally *k*, respectively. The effect of this updating can be observed by comparing the results in Figure 3-1c with those in Figure 3-1e. The upper bound of the validity window in interval *(k-2)* increases as a result of the new data obtained when the present time is at the end of interval *k*, and this causes a reclassification of one of the travel time observations (near the end of interval *(k-2*) from an outlier to a valid observation.

Finally, the travel time is predicted for interval *(k+1)* and the entire process repeats for the next time interval.

Considering the presented framework, the following factors can affect the accuracy of the travel time predictions in real time applications:

- Longer effective prediction horizons resulting from the time lag problem.
- Biased travel time estimates resulting from small sample sizes of the recent time intervals (e.g. in Figure 3-1g for time interval *k* there are only 2 observations available for calculating the estimated travel time).
- Unstable travel time estimates resulting from the recursive nature of the outlier detection.

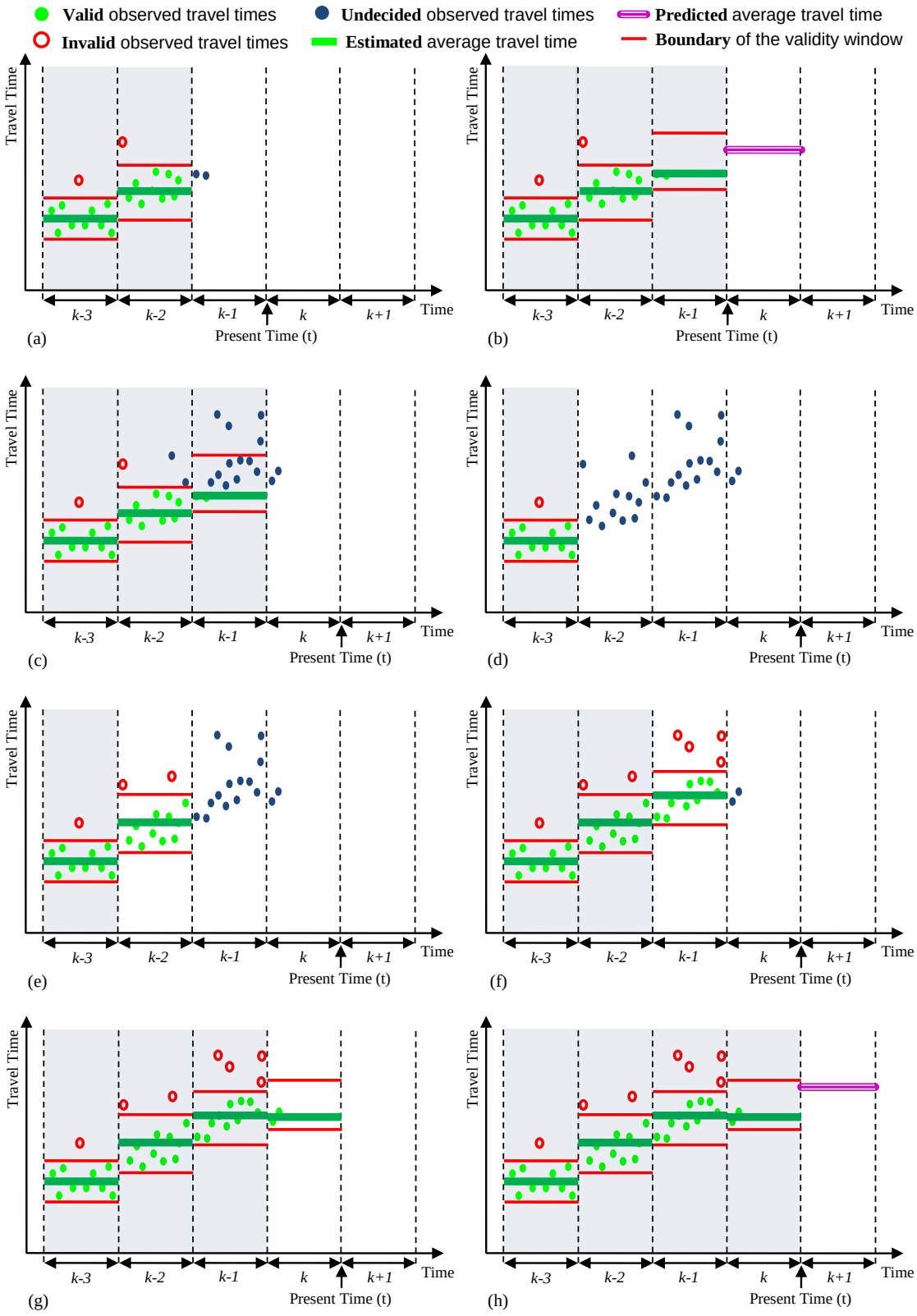


Figure 3-1: Real-time travel time prediction process using data obtained from Bluetooth detectors

3.2 Field data

3.2.1 Data types

Bluetooth detectors are most frequently deployed for the purpose of measuring individual vehicle travel times. However, these detectors collect several other types of information that may be valuable for travel time prediction and potentially other applications. Each time a Bluetooth equipped device is detected by a Bluetooth detector, the detector records the unique MAC address, the time of detection, and the Bluetooth detector ID. A device may be detected multiple times if it remains within the detection zone and this is reflected within the data as multiple records of the same MAC address at the same Bluetooth detector within the same consecutive period of time. To help clarify, we refer to each of these multiple detections as a *hit*. There is a strong correlation between the length of time that a device remains within the detection zone and the number of hits. Consequently, the sum of the number of hits across all devices can be thought of as a proxy for the total vehicle travel time over the time interval and over the length of roadway covered by the detection zone.

Though there may be multiple hits for a given device as it passes a Bluetooth detector, these hits are all associated with a single device (we assume each device is associated with a unique vehicle). When this same device is observed at a downstream detector, then we are able to compute the travel time for the vehicle by comparing the time stamp at the downstream detector with the time stamp at the upstream detector. There may be multiple hits at the upstream and at the downstream detectors. Typically simple heuristic approaches are used (e.g. first; last; median; etc.) to determine which one of the multiple hits at the upstream detector should be matched with one of the multiple hits at the downstream detector in order to compute a single travel time. These travel times can only be computed once the vehicle has been detected at the downstream detector.

Each series of hits at a detector represents the detection of a unique device (vehicle) and therefore is referred to as a *detection*. Though only a portion of the vehicles in the traffic stream contain Bluetooth devices that can be detected, the number of detections is expected to be correlated with traffic volume passing through the detection zone.

We refer to the number of travel time measurements between two consecutive Bluetooth detectors in a given time interval as *counts*.

On arterials, where at-intersection installation of Bluetooth detectors is the common practice (because of power source availability), data from single Bluetooth detectors (i.e. hits and detections) may provide an approximation of the congestion level of the intersection. Furthermore, unlike travel time and count data (i.e. data that can only be obtained from two consecutive Bluetooth detectors), these data do not suffer from any time lag problem. This property can potentially improve the accuracy of the predictions in transient conditions where sudden changes in travel times occur. The travel time prediction methods proposed in this research investigate the use of these various types of data that can be obtained from Bluetooth detectors.

3.2.2 Available data

As presented in Figure 3-2, field data were obtained from three Bluetooth detectors installed at signalized intersections along a 3.1 km section of an arterial roadway (Highway 7 in York Region, Toronto). Bluetooth data (i.e., travel times, counts, detections, and hits) were collected for both directions of the two road segments (resulting in 4 directional segments as defined in Table 3-1) for a period of 17 months (May 2011 through September 2012).

To construct the historical dataset, the travel time observations and the counted trips were filtered for data anomalies using an offline outlier detection algorithm proposed by Dion and Rakha (Dion and Rakha 2006).

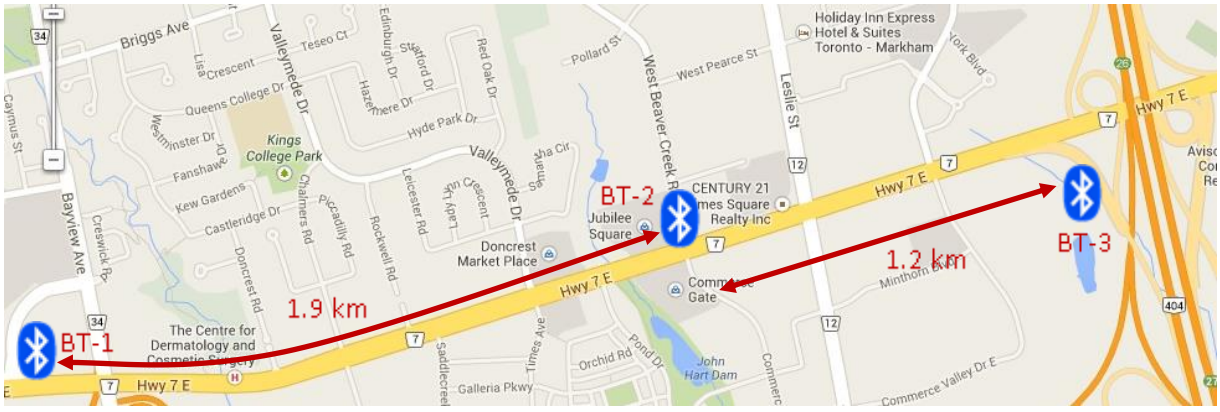


Figure 3-2: The location of Bluetooth scanners along the test network (Highway 7 - Toronto) (Source: Google Maps)

Table 3-1: The characteristics of the investigated sections

Segment	Stations		Length (km)	# of Signalized Intersections		# of Unsignalized Intersections
	From	To		Major	Minor	
1	Bayview Ave. (BT-1)	Beaver Creek Rd. (BT-2)	1.9	1	4	1
2	Beaver Creek Rd. (BT-2)	Bayview Ave. (BT-1)	1.9	1	4	1
3	Beaver Creek Rd. (BT-2)	HWY 404 (BT-3)	1.2	1	2	---
4	HWY 404 (BT-3)	Beaver Creek Rd. (BT-2)	1.2	1	2	---

Missing data (due to detector or communication failures) were imputed using the moving median technique (Montgomery and Runger 2010).

As we did not have access to the Bluetooth data in real time, we emulated the real time acquisition of Bluetooth data using the unfiltered field data set and the proposed framework described in the previous section (Figure 3-1).

At the end of each time period, the validity of the new observed travel times was examined using the real-time proactive adaptive filtering algorithm described in section 4.3. In this algorithm, unlike the conventional reactive techniques, the validity window of the time intervals with new observations is constructed based on their expected travel time trend, rather than that of the previous time intervals. This enhancement resolved the vulnerability of the reactive adaptive filtering techniques to sudden changes in the travel time trend. Finally, the outliers in the detections and hits were identified in real time using a threshold based method. Based on this approach, observations for which the number of detections or the number of hits exceeded a predefined threshold were removed from the set of valid observations.

In order to gain a better understanding of the travel time pattern on the investigated segments, the variation of 5-minute average travel time as a function of time of day is provided for all the available segments for a specific day (Wednesday, April 18th 2012) (Figure 3-3).

A review of the data depicted in this figure reveals the inconsistency of travel time patterns across the investigated road segments:

- The formation/recovery of the traffic congestion is usually more rapid on Segment 1 compared to the other investigated segments.
- The occurrence of extreme traffic congestion on Segment 2 is infrequent.

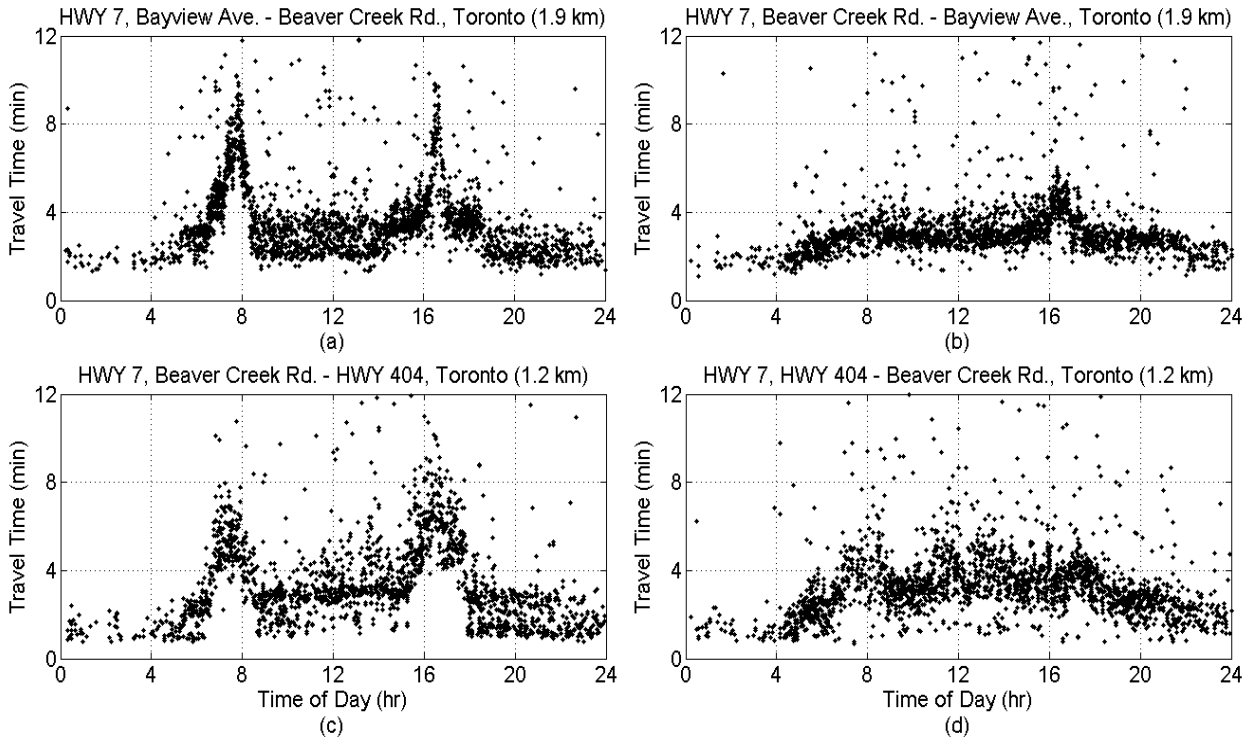


Figure 3-3: Travel time patterns on different segments for a sample day (Wednesday, April 18th, 2012)

- The variation within the valid travel time observations is relatively higher on Segments 3 and 4 compared to that of the first two segments.
- The traffic congestion on Segment 3 usually forms/recovers more gradually compared to the case of segment 1.
- The mild traffic congestion observed on Segment 4 usually holds from early in the morning to the mid-evening hours.

The day to day variation of travel time can be investigated through Figure 3-4. In this figure, the travel time vs. time of day diagrams have been produced for segment 1 during six consecutive days. It can be observed that the temporal variations of travel times as a function of time of day are not completely consistent from day to day, even when comparing working days.

The available historical dataset includes different travel time patterns. For the sample case of Segment 3, the existing diversity in the trend of travel time observations can be examined through Figure 3-5. Four different travel time patterns are recognized in this figure, a weekend traffic congestion (Figure 3-5a), a regular recurrent traffic congestion (Figure 3-5b), an extended recurrent traffic congestion (Figure 3-5c), and a non-recurrent traffic congestion (Figure 3-5d).

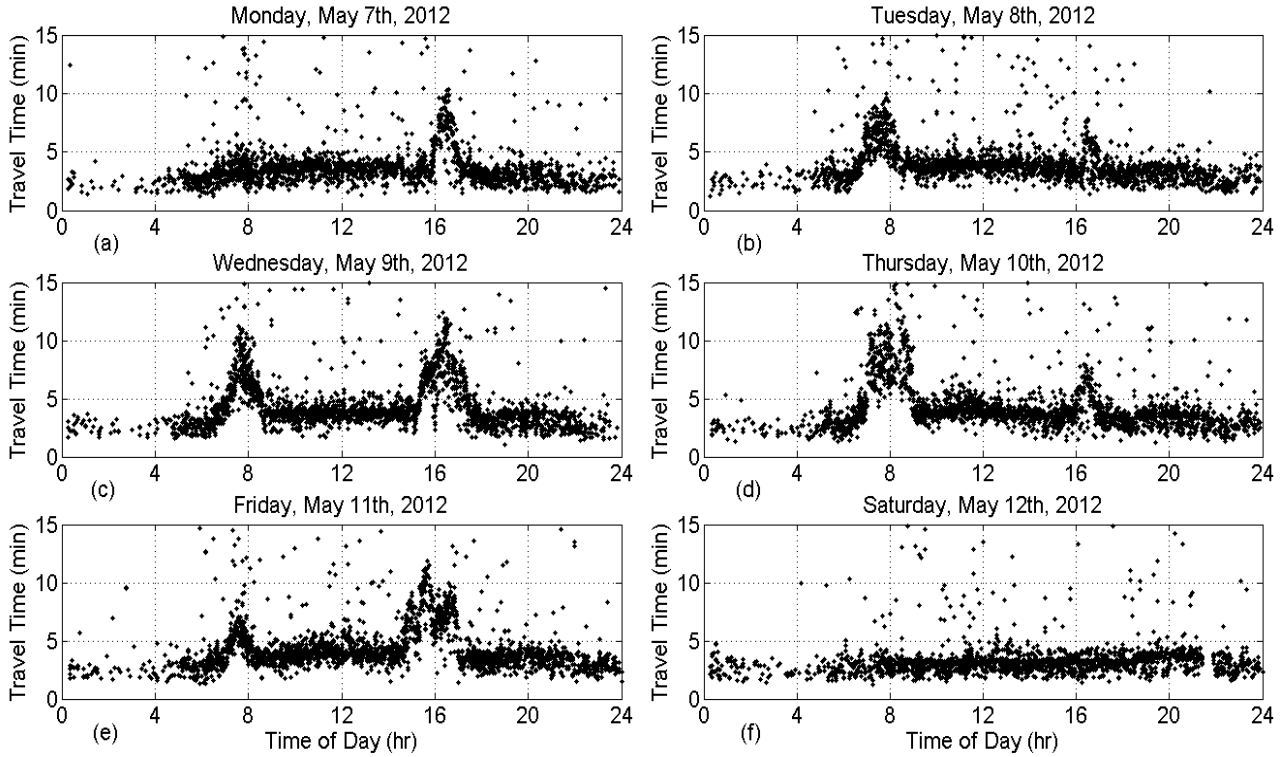


Figure 3-4: Day to day variation of travel time on a sample segment

HWY 7, Bayview Ave. – Beaver Creek Rd., Toronto (1.9 km)

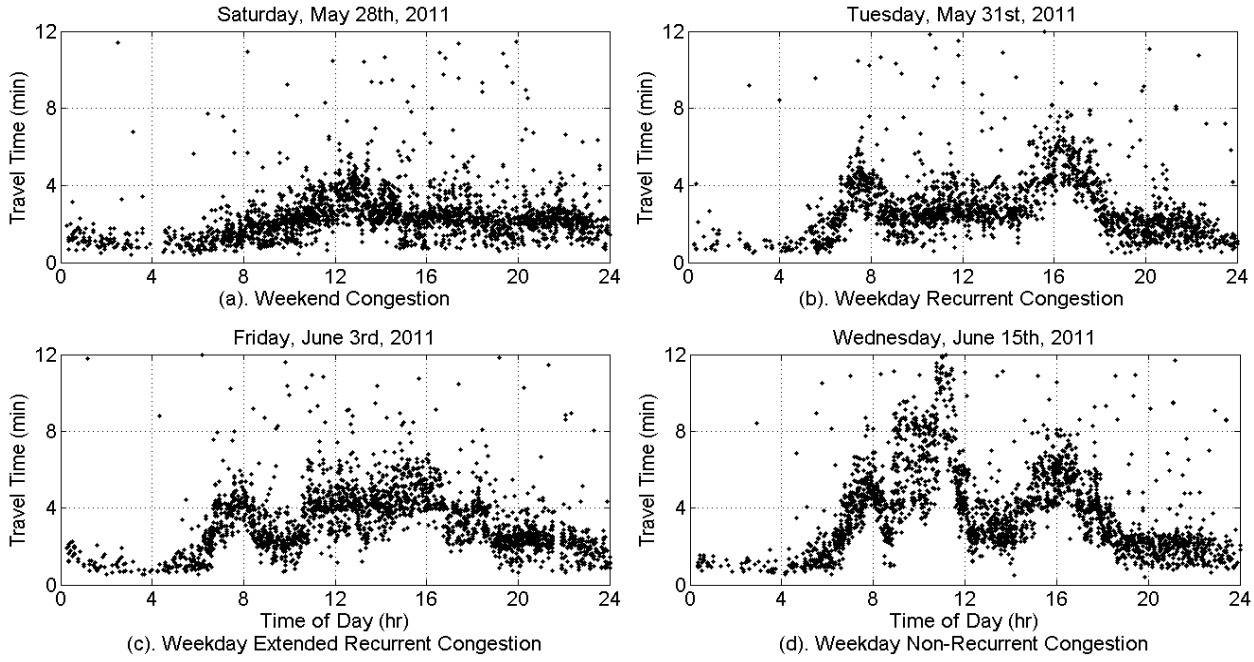


Figure 3-5: Diversity of travel time patterns on a sample segment

HWY 7, Beaver Creek Rd. – HWY 404, Toronto (1.2 km)

Chapter 4

Travel Time Estimation

The main focus of this chapter is to improve the accuracy of travel time estimations. The chapter is divided into four main sections. Section 1 discusses errors inherent to the Bluetooth travel time measurement approach. Section 2 quantifies the Bluetooth measurement error in arterial travel times is quantified. Section 3 presents the proposed real-time filtering algorithm for detecting Bluetooth travel time outliers. Finally, Section 4 presents the use of field data and simulation data for evaluating travel time outlier detection algorithms.

4.1 Bluetooth travel time errors

Bluetooth detectors provide several advantages over other sensor technologies, including direct measurement of travel time, anonymous detection, and cost effectiveness; however, like any other measurement tool the Bluetooth detectors are also subject to different sources of outliers and measurement errors. Specifically, Bluetooth obtained travel times are prone to the following sources of error:

- *Sampling error:* Using Bluetooth sensing technology, we are only able to capture the travel times of a sample of vehicles in the traffic stream. The accuracy of the sample average travel time (during a specific time interval) as an estimate of the population mean travel time is a function of the number of observations in the sample. The sample size required to achieve a specific desired level of accuracy is highly dependent on the variation of observed travel times during the investigated time interval (Montgomery and Runger 2010) and is not always achievable on arterial segments.
- *Sampling bias:* The detection time at each Bluetooth scanner is the only piece of information that can be acquired for a detected Bluetooth device. Thus, Bluetooth detectors are not able to distinguish between a single Bluetooth device in a personal automobile and (i) multiple devices in a single vehicle, (ii) the devices that are being transported using a mode other than an automobile (e.g., bus or bike), and (iii) the vehicles with planned en-route stops. Moreover, due to the difference between the arrival distribution of the sample and population (the term arrival refers to the arrival time of vehicles at signalized

intersections), another type of bias, which is referred as arrival time bias, may occur on arterials (Hellinga and Fu 1999). Thus, considering all these biases, the observed Bluetooth travel times may not necessarily reflect the real traffic state of the roadway.

- *Measurement error:* A single road side Bluetooth detector continuously searches for Bluetooth devices passing through its detection zone. Therefore, depending on the duration of the time that a device resides in the detection area, it may be detected once, several times or even not be captured at all. As these detections may occur anywhere along the scanning area, it is quite possible that the detection locations do not match that of the scanner (temporal effect). The time difference associated with these imprecise detections is referred to as detection time measurement error. This measurement error occurs at both the upstream and downstream Bluetooth detectors and therefore gives rise to an error in the travel time measurement (Figure 4-1).

In the case of multiple detection instances, the issue of measurement error becomes more complex, as it is not obvious which moment should be considered as the more accurate detection time. In this research, following the common practice, the first detection moments at consecutive stations are utilized for travel time calculation. In this way, the discussed detection time error specifically corresponds to the error associated with the first detection moment.

Finally, it should be noted that the radius of the detection zone varies based on the antennae specifications of the scanner. In this way, different antennae configurations at the upstream and downstream scanners may result in more complications in the discussed measurement errors (spatial effect).

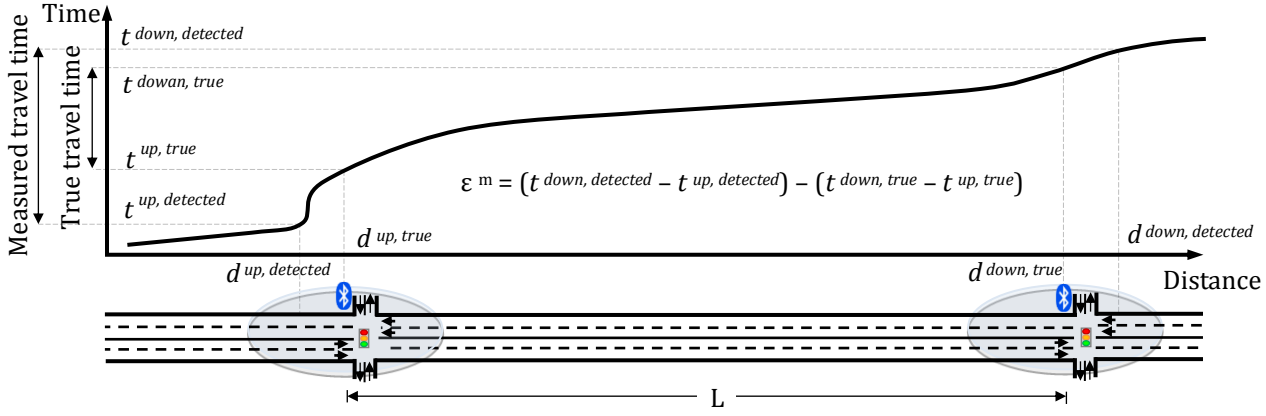


Figure 4-1: Measurement error in Bluetooth obtained travel times

4.2 Quantifying measurement error in arterial travel times measured by Bluetooth detectors[1]

4.2.1 Measurement error diagnosis

As mentioned earlier, the magnitude of the measurement error for an individual travel time observation depends on the magnitude of the detection time errors at the upstream and downstream scanners. The detection time error itself can be determined based on the location based probability of the first detection. This probability can be calculated based on the probabilistic characteristics of channel hopping behaviour and those factors affecting the speed profile of the device within the detection zone.

Based on the Bluetooth protocol, Bluetooth devices perform a single scan of 18 frequencies every 1.28 seconds. Given that there are 32 allocated frequencies in the Bluetooth spectrum, during each polling interval the probability of detection at the vicinity of the detector is 0.5625 (18/32) (Bakula et al. 2012). As long as the moving Bluetooth device can communicate with the roadside mounted Bluetooth detector effectively, it can be detected according to the mentioned probability. However, the signal strength decays over distance. Based on the literature, the received power transmitted from the detector (a class 1 Bluetooth device), starts to fall below the minimum appropriate receiver strength sensitivity (-70 to -85 dBM) at a distance of approximately 50 meters from the detector. Based on this discussion, in this research we reasonably assumed that the probability of detection around the detector follows an isosceles trapezoid pattern (Figure 4-2a).

The speed profile of the vehicles within the detection area highly depends on the location at which the Bluetooth detector is installed (i.e., mid-block vs. at-intersection). The following sections investigate these two installation locations separately.

[1] The content of this section is published in the following journal paper:

- Salek Moghaddam, S. and B. Hellinga, "Quantifying Measurement Error in Arterial Travel Times Measured by Bluetooth Detectors", 2013, Transportation Research Record, Vol. 2395, pp. 111 - 122.

4.2.1.1 Mid-block location

On arterials, mid-block traffic flow rate rarely exceeds the capacity of the link (being controlled by the upstream traffic signal). Therefore, it is reasonable to assume that vehicles maintain a constant speed (posted speed) when traveling along these segments.

When the Bluetooth detector is located mid-block, the probability of the first detection during the K^{th} polling interval for a sample device travelling with the constant posted speed can be calculated based on its location at different polling intervals. In this way, as depicted in Equation 4-1, the probability of the first detection during the K^{th} polling interval can be calculated by multiplying the probability of being detected at the K^{th} interval and the probabilities of not being detected at the prior intervals (Figure 4-2a):

$$P_{(K)}^{F.D} = \begin{cases} \left[\prod_{\kappa=1}^{K-1}\left(1-p\cdot(\frac{\kappa\cdot\tau_k\cdot S}{3.6(L-B)})\right)\right]\cdot\left(p\cdot(\frac{K\cdot\tau_k\cdot S}{3.6(L-B)})\right) & ; K \le \frac{3.6(L-B)}{\tau_k\cdot S} \\ \left[\prod_{\kappa=1}^{(\frac{3.6(L-B)}{\tau_k\cdot S})}\left(1-p\cdot(\frac{\kappa\cdot\tau_k\cdot S}{3.6(L-B)})\right)\right]\cdot\left(\prod_{\kappa=(\frac{3.6(L-B)}{\tau_k\cdot S})+1}^{K-1}(1-p)\right)\cdot p & ; \frac{3.6(L-B)}{\tau_k\cdot S} < K < \frac{3.6(L+B)}{\tau_k\cdot S} \\ \left[\prod_{\kappa=1}^{(\frac{3.6(L-B)}{\tau_k\cdot S})}\left(1-p\cdot(\frac{\kappa\cdot\tau_k\cdot S}{3.6L})\right)\right]\cdot\left(\prod_{\kappa=(\frac{3.6(L-B)}{\tau_k\cdot S})+1}^{\kappa=(\frac{3.6(L+B)}{\tau_k\cdot S})}(1-p)\right)\cdot & \\ \left[\prod_{\kappa=(\frac{3.6(L+B)}{\tau_k\cdot S})+1}^{K-1}\left(1-p\cdot(\frac{2L-\kappa\cdot\tau_k\cdot S}{3.6(L-B)})\right)\right]\cdot\left(p\cdot(\frac{2L-K\cdot\tau_k\cdot S}{3.6(L-B)})\right) & ; K > \frac{3.6(L+B)}{\tau_k\cdot S} \end{cases}$$

(4-1)

Where:

$P_{(K)}^{F.D}$ = the probability of the first detection during the K^{th} polling interval

p = the probability of detection in the vicinity of the detector (i.e. p = 0.5625)

τ_k = the length of scanner polling interval (i.e., 1.28 sec)

S = the speed of the vehicle carrying the Bluetooth device in km/hr (e.g., 50 km/hr)

L = the detection radius of the scanner in meter (e.g., 100 m)

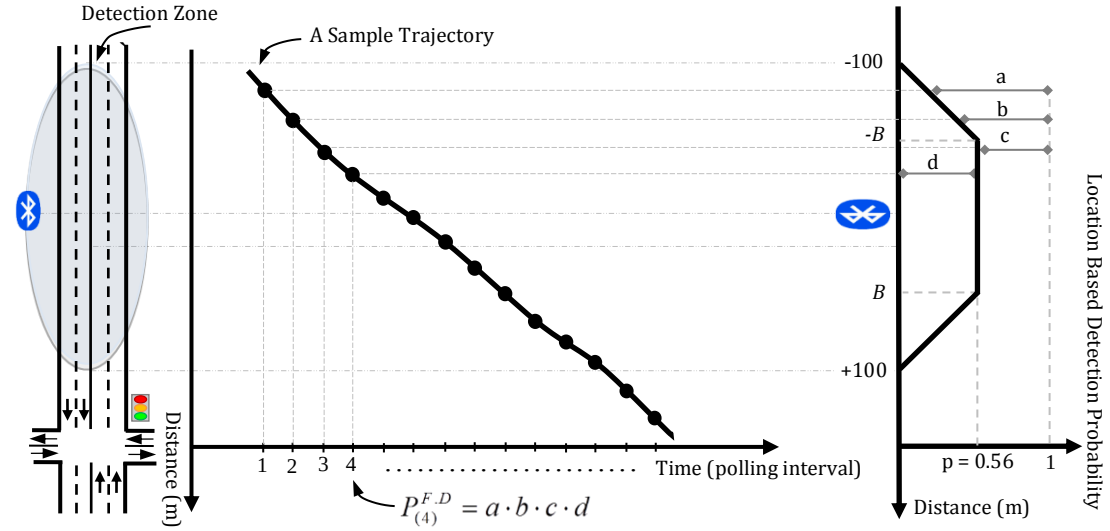
B = half the length of the small base in the trapezoid shape probability of detection (e.g., 50 m)

Following this procedure, the location based probability of the first detection and its associated cumulative probability are developed for two different sample trajectories ($S = 50 \& 80\ km/hr$) (Figure 4-2b-e). The results in this figure suggest that when vehicles travel at constant speed, the distribution of the location of the first detection is uni-modal and though somewhat skewed, the vast majority of first detections occur upstream of the detector. It can be observed that as the speed of the vehicle increases, the distribution becomes more highly skewed.

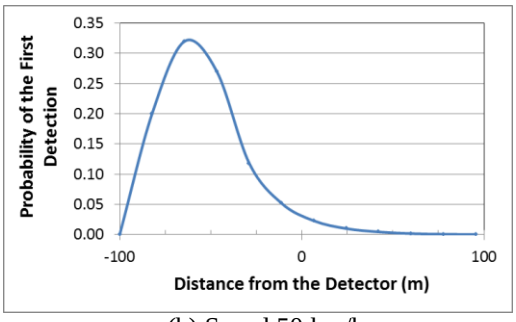
4.2.1.2 At-intersection location

Due to the power source restrictions, Bluetooth detectors are usually installed close to the signalized intersections where the movement of the vehicles may be interrupted by traffic signals. The changes in the speed profile of vehicles approaching a signalized intersection can be approximated using shockwave theory (Figure 4-3).

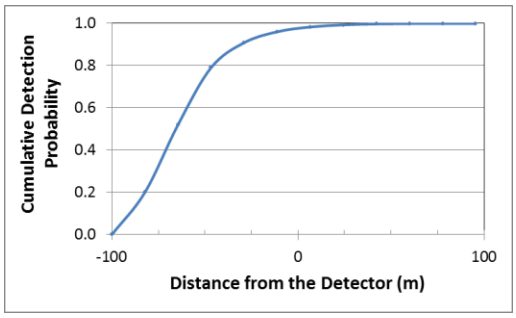
Figure 4-3a shows the shockwave diagram for an under-saturated signalized intersection approach and Figure 4-3b shows the diagram for an oversaturated approach. Generally, a sample vehicle approaching a signalized intersection travels at a constant speed (posted speed) before joining the tail of the queue at the intersection (pattern 1). At this stage, the vehicle is required to stop in the queue (pattern 2). For under-saturated conditions, when the signal turns green and the backward recovery shockwave encounters the vehicle, the vehicle is able to discharge from the intersection at the speed of capacity (pattern 3). For over-saturated conditions, the vehicle is able to move downstream but is unable to cross the stop-line before the signal turns red and therefore must stop again (pattern 2). Based on shockwave theory, the duration of these speed patterns highly depends on the arrival time pattern of vehicles (governed by the coordination level along the corridor), level of saturation, and signal timing specifications (Figure 4-3).



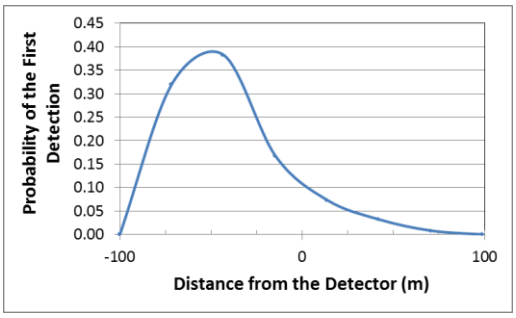
(a)



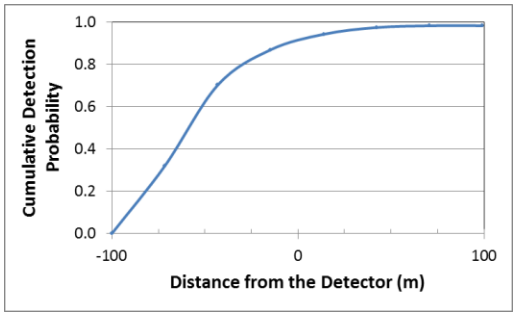
(b) Speed 50 km/hr



(c) Speed 50 km/hr



(d) Speed 80 km/hr



(e) Speed 80 km/hr

Figure 4-2: Location based probability of the first detection for mid-link installation scheme (sample trajectory)

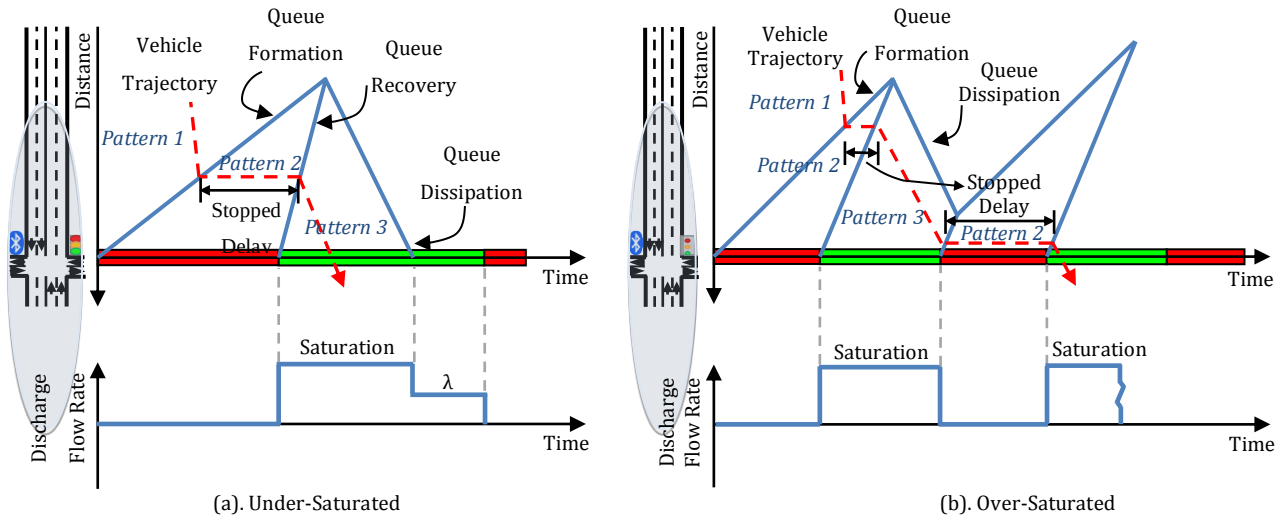


Figure 4-3: Investigating the speed profile of vehicles at signalized intersections

Based on these speed patterns, the probability of the first detection (for a sample device) can be formulated. During the first and third speed regimes, the calculations are quite similar to those of mid-block installation (constant speed). However, for the second speed pattern, when the device is stationary (waiting at the queue), the probability of the first detection can be formulated based on the Binomial probability function. Based on the Binomial function, the probability of having at least one detection (success) throughout the stationary mode (from the beginning of this speed pattern till the K^{th} interval) is equivalent to the complementary probability of having no detection during this period (Equation 4-2):

$$P_{(K)}^{F.D} = \left[\prod_{\kappa=1}^{\left(\frac{3.6(L-B)}{\tau_k \cdot S}\right)} \left(1 - p \cdot \left(\frac{\kappa \cdot \tau_k \cdot S}{3.6(L-B)}\right)\right)\right] \cdot \left(\prod_{\kappa=\left(\frac{3.6(L-B)}{\tau_k \cdot S}\right)+1}^{\left(\frac{3.6(L-d_S)}{\tau_k \cdot S}-1\right)} (1-p)\right) \cdot \left[1 - \left(\begin{pmatrix} \frac{t_S}{\tau_k} \\ 0 \end{pmatrix} \cdot (p)^0 \cdot (1-p)^{\left(t_S / \tau_k\right)}\right)\right]$$

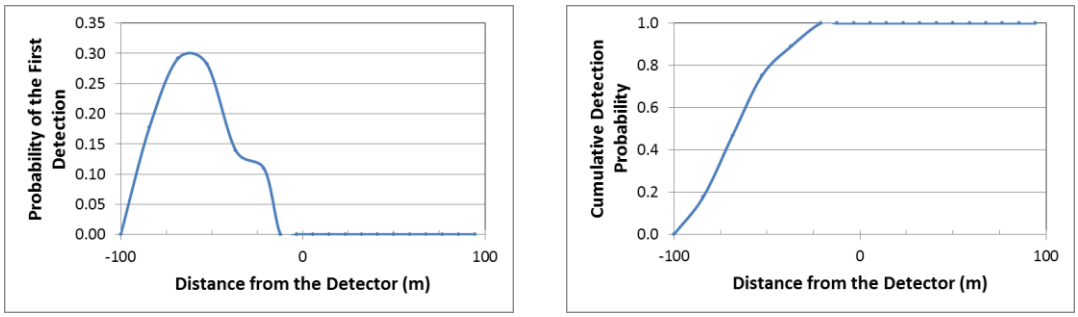
$$; \frac{3.6(L-d_S)}{\tau_k \cdot S} \le K < \frac{3.6(L-d_S)}{\tau_k \cdot S} + \frac{t_S}{\tau_k} \quad \& \quad B > ds \qquad \textbf{(4-2)}$$

Where:

t_S = duration of the stopped delay

d_S = distance from the detector to the location of stopped delay

Using this equation, the location based probability of the first detection and its associated cumulative probability are developed for a sample trajectory (under-saturated condition) (Figure 4-4). A comparison of the results in Figure 4-4a to those in Figure 4-2b or Figure 4-2d shows the impact that the signalized intersection has on the probability of first detection. The probability distribution is now no longer smooth and all devices are detected upstream of the detector.



(a) (b)

Figure 4-4: Location based probability of the first detection for at-intersection installation (sample trajectory; under-saturated conditions)

4.2.2 Modeling framework

The previous section describes the underlying mechanism by which Bluetooth detections occur and by which detection measurement errors arise. In this section, we describe a framework in which we numerically compute the distributions of arterial travel time measurement error as a function of the traffic conditions along the arterial.

The framework consists of three elements, namely: (1) simulation of vehicle trajectories; (2) simulation of Bluetooth detections; and (3) calculation of travel time measurement errors.

4.2.2.1 Simulation of vehicle trajectories

As demonstrated in the previous section, the location of the first detection of a Bluetooth device is highly dependent on the pattern of the vehicle trajectory which is a function of the traffic characteristics. Therefore, vehicle trajectories were generated through the use of the VISSIM micro-simulation model for a variety of signal control and traffic conditions. An arterial roadway containing two signalized intersections was modeled. Each signal operated under two-phase

fixed time control. Given that in most applications, Bluetooth travel times are obtained for major arterials, we assumed a constant g/c ratio for the through movement on the arterial of 2/3. Trajectories were extracted for all vehicles approaching the downstream intersection only. These trajectories consisted of the position of the vehicle relative to the downstream intersection stop line every 0.2 seconds.

A total of 45 simulation scenarios were defined. We control for three traffic factors at the downstream intersection approach, namely: (1) v/c ratio (0.5, 0.7, 0.85, 1.0, 1.1); (2) cycle length (60, 90, 120 seconds); and (3) quality of progression.

Three levels of progression were modeled, namely: well-coordinated (WC); moderately coordinated (MC); and poorly coordinated (PC). These three levels correspond to HCM (TRB 2010) arrival categories of AT6, AT3, and AT1, respectively. Coordination was controlled by varying the offset of the downstream signal relative to the upstream signal.

A single simulation run was conducted for each scenario. Each simulation was run for 60 minutes with the first 5 minutes for a warm-up period, and data were extracted from the remaining 55 minutes.

4.2.2.2 Simulation of Bluetooth detections

The next step in the framework is to simulate Bluetooth detections at a Bluetooth detector. As illustrated in Figure 4-2a, the location of the first detection of a Bluetooth device is a function of the location-based probability of detection and the vehicle trajectory. Consequently, the location of first detection of each individual vehicle was generated through a Monte Carlo simulation process. The Bluetooth scanning interval is 1.28 seconds. Therefore at each scanning interval, the location of each individual vehicle was extracted from the recorded trajectories. Then using the location-based probability of detection (as defined in Figure 4-2a), we determine for each vehicle if it is detected. The location and time of the first detection of each vehicle is recorded.

We also record the time at which each vehicle passes the location of the Bluetooth detector. The difference between the time of the first detection and the time at which the vehicle passed the location of the Bluetooth detector is the detection time measurement error.

This process is repeated for all 45 scenarios and empirical distributions of the detection time measurement error for each scenario were compiled.

4.2.2.3 Calculation of Travel Time Measurement Errors

Travel time measurement error is the summation of the detection time measurement errors at the upstream and downstream detectors. Consequently, we used a Monte Carlo simulation approach to generate detection time measurement errors at the upstream and downstream detector by sampling from the distributions developed in the previous step.

In most applications, arterial travel times are aggregated in order to generate a sufficiently large sample size and to reduce the variability that is caused, at least in part, by the cyclic nature of the traffic signal operations. A typical aggregation period is 5 minutes and therefore this period was chosen in this study.

The total hourly volume of through vehicles for one direction of a typical 4 lane arterial was calculated on the basis of the v/c and g/c ratios for each scenario. A travel time measurement error was generated for each vehicle for the investigated scenarios using the Monte Carlo simulation approach.

The total number of travel time measurement errors (N_i) for a single scenario (i) were distributed to each of the twelve 5-minute aggregation periods (k) using the following approach. The average number of observations in each 5-minute period was computed as $\bar{n}_{i,k} = N_i / 12$. However, the number of observations in a single period k ($n_{i,k}$) is subject to random variation and therefore we assume $n_{i,k}$ follows a Normal distribution with mean of $\bar{n}_{i,k}$ and standard deviation which is sensitive to the v/c ratio.

Lastly, we aggregate the travel time measurement errors to compute the mean travel time measurement error for each aggregation period for each scenario.

This process is repeated for 100 hours to generate 1200 aggregate travel time measurement errors for each of the investigated scenarios. Using these data, we generated distributions of travel time measurement errors that are expected to be contained within 5-minute aggregated Bluetooth measured arterial travel times.

4.2.3 Error Distributions

The resulted error distributions are presented in two different sections. The first section examines the detection time errors. The second examines the travel time errors.

4.2.3.1 Detection Time Errors

Figure 4-5 illustrates the distribution of detection time errors and the distribution of first detection location for three of the 45 scenarios. These three scenarios were selected as they depict the range of detection time errors that were observed. The conditions of each scenario are defined in each figure heading in term of v/c ratio; cycle length; and quality of progression. A number of observations can be made from these results:

- For low v/c, well coordinated operation, and low cycle length (Figure 4-5a & b), the detection error distribution exhibits a single mode and has relatively low mean and standard deviation. Approximately 4.1 percent of vehicles are detected for the first time downstream of the detector.
- For high v/c, medium coordination operation, and medium cycle length (Figure 4-5c & d), the distribution of error is bimodal and has a higher mean and standard deviation compared to the case of low v/c. Approximately 2.6 percent of vehicles are detected for the first time downstream of the detector.
- For high v/c, poor coordination operation, and high cycle length (Figure 4-5e & f), the distribution of error is again bimodal but with an even larger standard deviation than the other two cases. Approximately 3.1 percent of vehicles are detected for the first time downstream of the detector.

The previous observations suggest that the distribution of detection measurement error is dependent on several traffic factors. Consequently, an investigation of these dependencies was conducted using linear regression. Given that the spread of the distribution is of interest, but the distributions may not be unimodal, the dependent variable selected to capture the spread was the 90th percentile – 10th percentile detection time error (seconds). The independent variables were the different levels of v/c, cycle length and coordination level reflected by the 45 scenarios. A linear model structure with the first order interaction terms was calibrated to the data. All independent variables are binary and only those terms that were statistically significant were retained (Table 4-1). The model was used to investigate the influence of v/c, cycle length, and quality of progression on the spread of the distribution of detection time measurement errors. Several observations can be made on the basis of these results (depicted in Figure 4-6):

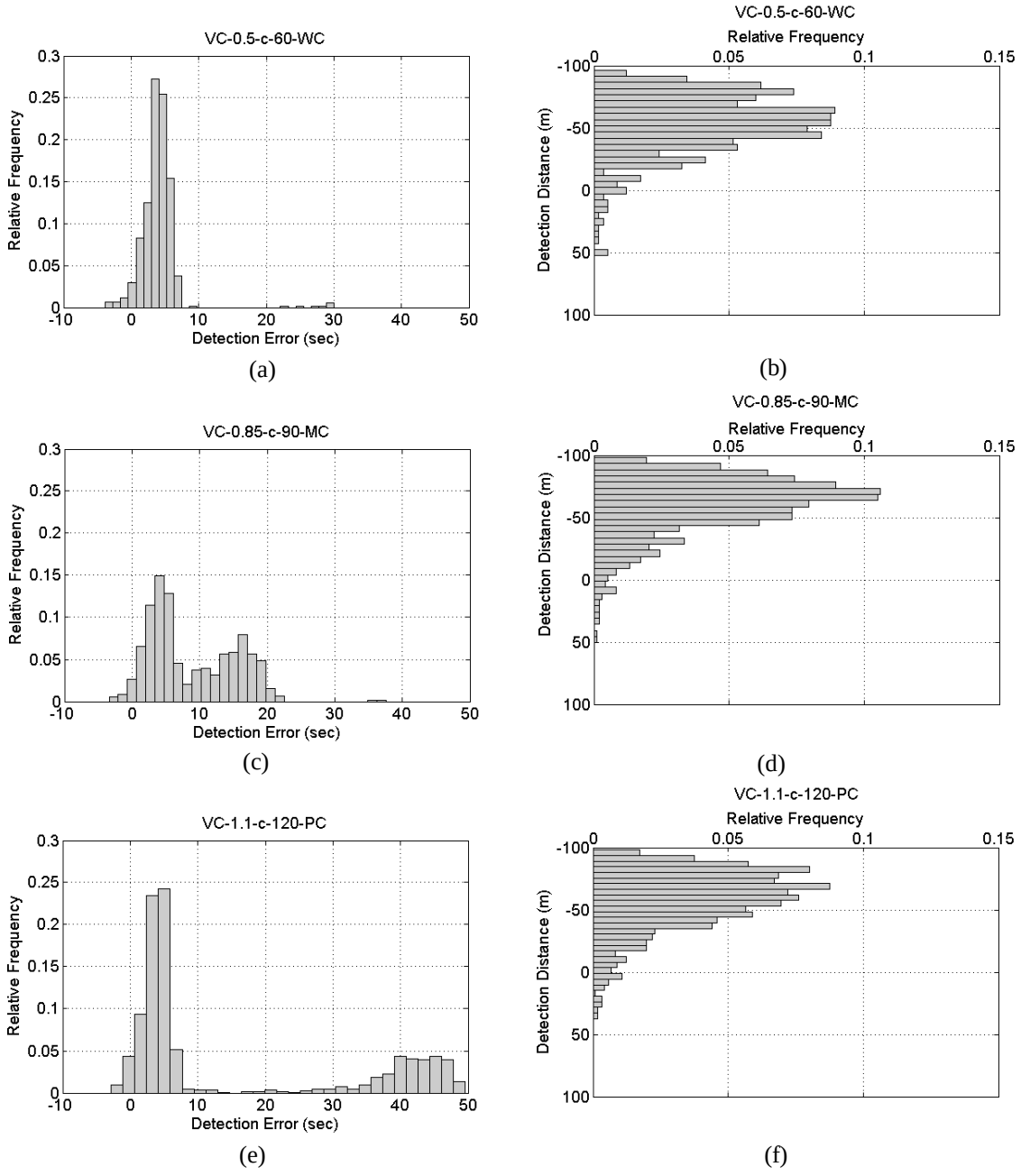


Figure 4-5: The distribution of detection time measurement error and detection location

- The overall trends resulting from the model appear to be logical. The spread of the distibution increases with increasing v/c, increasing cycle length, and decreasing quality of progression.
- At high v/c levels, the quality of progression does not have a significant impact. This is consistent with expectation as coordination has the greatest impact on delays when the approach is under-saturated.

- Some inconsistancies can be observed. For example, at v/c = 1 and the section is well coordinated, the distribution becomes wider compared to v/c = 0.85. This increase is the result of excluding the interaction terms with significance levels lower than 95% (some of these interaction terms would be included in the model if the acceptable significant level were reduced to 85%). Nevertheless, these inconsistancies appear to be minor.

Table 4-1. Linear regression model showing spread of detection time error distributions

Variable Name	Description	Coefficient (sec)	t*
Constant		43.24	14.70
β_{VC1}	v/c = 0.5	-9.57	-2.52
β_{VC2}	v/c = 0.7	-9.88	-2.60
β_{c1}	Cycle Length (c) = 60 sec	-21.62	-6.36
β_{c2}	Cycle Length (c) = 90 sec	-9.74	-2.87
$\beta_{VC2,c1}$	v/c = 0.7 & c = 60	8.37	2.20
$\beta_{VC1,WC}$	v/c= 0.5 & Well Coordinated (WC)	-22.43	-5.12
$\beta_{VC1,MC}$	v/c = 0.5 & Medium Coordinated (MC)	-10.73	-2.45
$\beta_{VC2,WC}$	v/c = 0.7 & WC	-22.15	-5.05
$\beta_{VC2,MC}$	v/c = 0.7 & MC	-11.26	-2.57
$\beta_{VC3,WC}$	v/c = 0.85 & WC	-24.21	-5.52
$\beta_{VC3,MC}$	v/c = 0.85 & MC	-13.55	-3.09
$\beta_{VC4,MC}$	v/c = 1.0 & MC	-14.61	-3.33
$\beta_{c1,WC}$	c = 60 & WC	8.75	2.58
No. of Observations = 45		Adjusted R^2 = 0.90	

* Significant at 95% confidence level

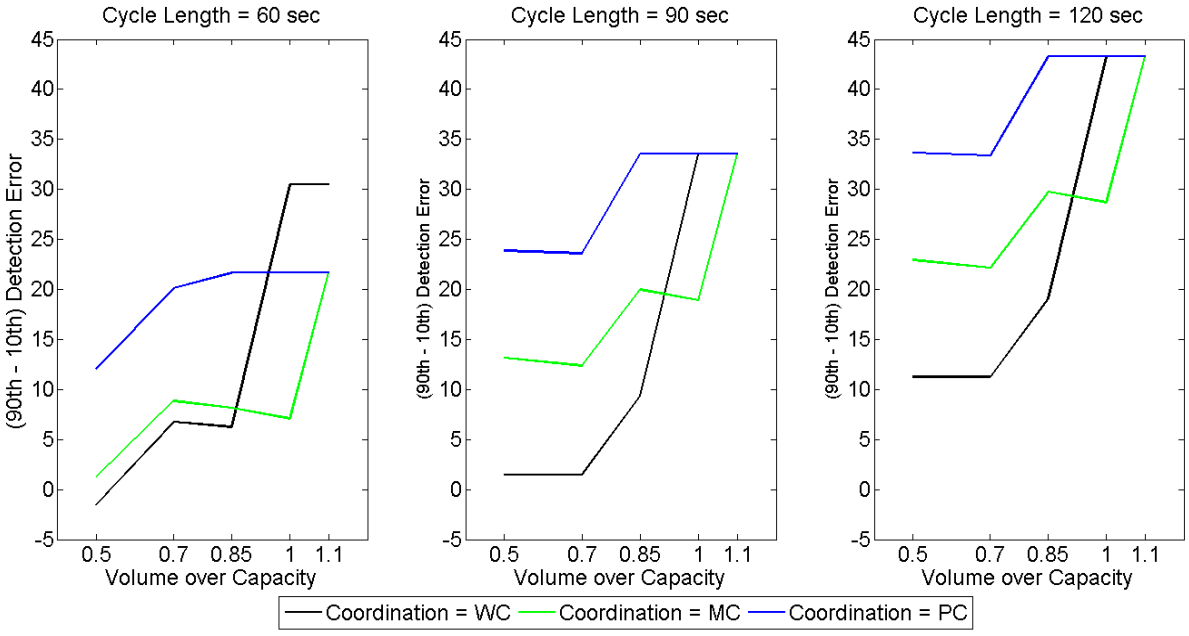


Figure 4-6: Impact of cycle length, v/c, and coordination on the spread of the detection time error distribution

4.2.3.2 Travel Time Measurement Error

The previous section examined the characteristics of the detection time error. However, in practice we use Bluetooth detector to obtain travel time and we are ultimately interested in the magnitude of the travel time measurement error. As discussed earlier, using the proposed framework, we generated distributions of travel time measurement errors (the distribution of average 5-minute travel time measurement error for 1200, 5-minute periods) for 225 different scenarios. Scenarios were defined in terms of the v/c ratio at the upstream signal (5 levels), v/c ratio at the downstream signal (5 levels), cycle length (3 levels) and coordination (3 levels - cycle length and coordination level were the same for both the upstream and downstream intersections). Analysis of the distributions showed that they are uni-modal and roughly Normal in shape. Consequently, we elected to characterize the distributions in terms of the mean and standard deviation.

Figure 4-7 illustrates the distribution of the average of travel time measurement errors from each of 225 distributions. It is notable that: (1) the mean is approximately zero; and (2) the variance is very small. Therefore, there is little value in developing a model for estimating the average of travel time measurement errors as a function of the scenario characteristics as this average is nearly constant.

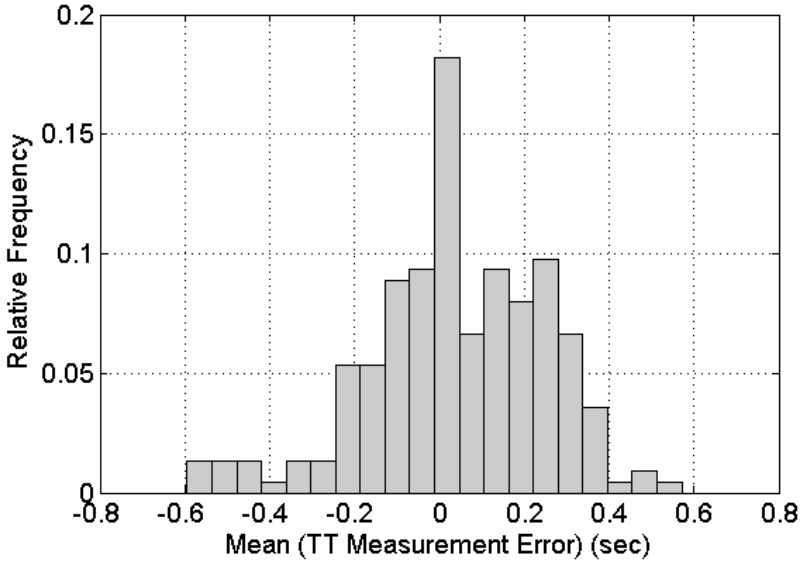


Figure 4-7: The distribution of average of travel time measurement errors (aggregated level)

However, examination of the standard deviations of the average 5-minute travel time measurement errors reveals that the standard deviation is not constant for all 225 scenarios. This can be illustrated by depicting the distribution of average 5-minute travel time measurement errors for three sample scenarios (Figure 4-8). Figure 4-8a shows a scenario for which the standard deviation is very small and travel time errors range from approximately -10 to +10 seconds. Figure 4-8c shows a scenario for which the standard deviation is much larger and travel time errors range from approximately -40 to +40 seconds. These results show that the standard deviation is sensitive to traffic characteristics of the roadway.

A multiple linear regression model was calibrated to estimate the standard deviation of 5-minute average travel time measurement error as a function of the traffic factors (Table 4-2). All independent variables are binary and only significant terms have been retained. Interpreting the impact of the different traffic factors is not straightforward because of the interaction terms. Consequently, we illustrate the use of the model to investigate the influence of travel time measurement error for a typical Bluetooth deployment on an arterial roadway.

4.2.4 Example Application

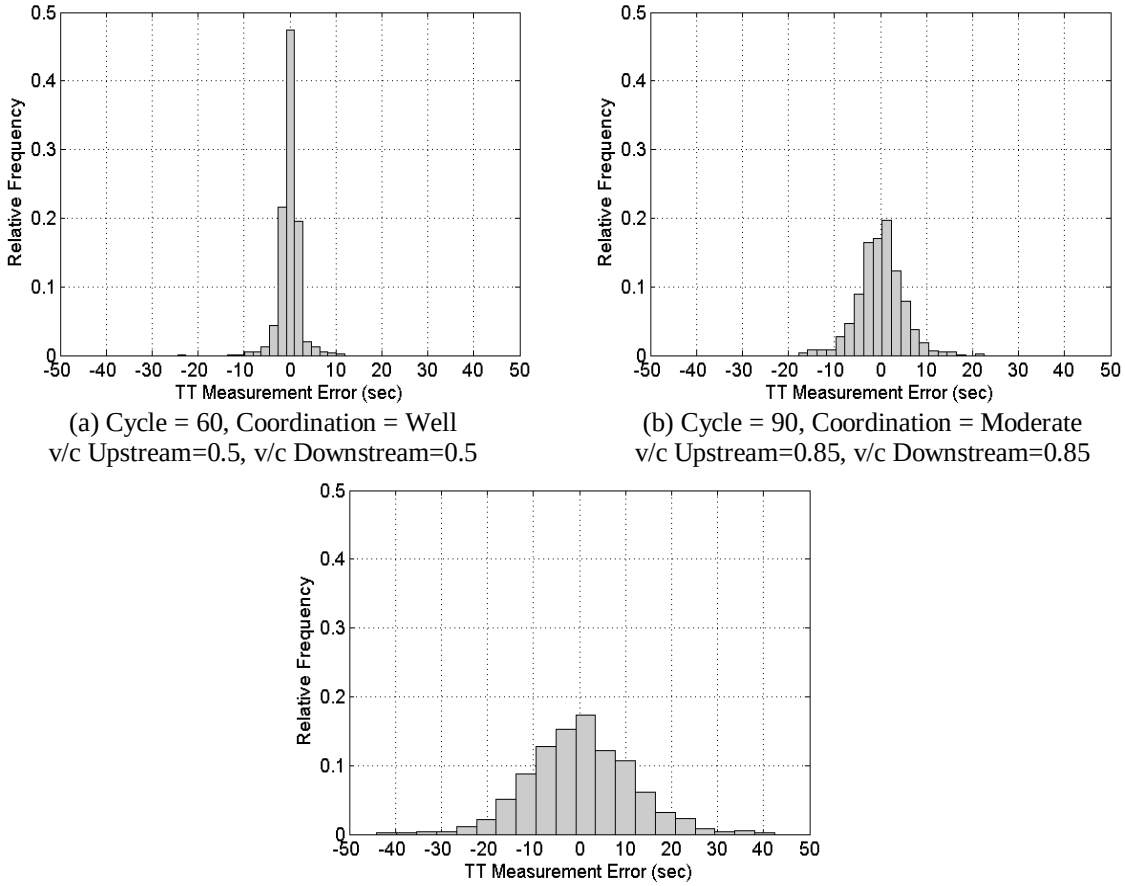
In practice, typically we are interested in understanding the magnitude of the measurement error for a given section length and traffic conditions relative to the true mean travel time. If the measurement error is small, then it can be ignored. However, if the error is expected to be large, then changes to the detector deployment could be made to reduce these errors. The results described in the previous sections can be used to carry out this type of analysis.

Consider a 1.5 km long arterial segment with signalized intersections at the upstream and downstream boundaries. No assumptions are made about the presence of other intermediate intersections. Bluetooth detectors are located at the upstream and downstream intersections. The speed limit is 50 km/h.

We do not know the true average travel time, but we can estimate it by assuming the travel time consists of two components, free speed travel time along the length of the arterial between the two intersections, and delay experienced at the downstream intersection. Based on the length and posted speed, the free speed travel time is expected to be 108 seconds. The delay at the

downstream signalized intersection can be estimated using the HCM delay equation and the signal timings, v/c, and level of coordination.

The multivariate regression model described in the previous section can be used to estimate the standard deviation of the 5-minute average travel time measurement error. Assuming the distribution can be adequately described by the Normal distribution, we compute the 95 percent confidence interval of the measurement error as ± 1.96 times the standard deviation estimated by the regression model. We represent this as the relative error by dividing by the estimate mean travel time.



(a) Cycle = 60, Coordination = Well
v/c Upstream=0.5, v/c Downstream=0.5

(b) Cycle = 90, Coordination = Moderate
v/c Upstream=0.85, v/c Downstream=0.85

(c) Cycle = 120, Coordination = Poor
v/c Upstream=1.1, v/c Downstream0.5

Figure 4-8: The distribution of measurement error for 3 sample traffic conditions

Table 4-2: Linear Regression Model Showing Standard Deviation of Travel Time Error Distributions

Variable Name	Description	Coefficient (sec)	t*
Constant		12.71	34.26
β_{VC_U1}	Upstream v/c = 0.5	-1.48	-3.38
β_{VC_U2}	Upstream v/c = 0.7	-1.38	-3.17
β_{VC_U3}	Upstream v/c = 0.85	-1.43	-3.27
β_{c1}	Cycle Length (c) = 60 sec	-7.97	-21.32
β_{c2}	Cycle Length (c) = 90 sec	-3.68	-9.84
β_{WC}	Well-Coordinated (WC)	-1.69	-4.53
β_{MC}	Medium-Coordinated (MC)	-1.68	-4.49
β_{VC_D1}	Downstream v/c = 0.5	3.82	8.74
β_{VC_D2}	Downstream v/c = 0.7	2.77	6.34
β_{VC_D3}	Downstream v/c = 0.85	1.89	4.32
β_{VC_D4}	Downstream v/c = 1.0	1.16	2.66
β_{VC_U1,VC_D1}	Upstream v/c = 0.5 & Downstream v/c = 0.5	-1.08	-2.35
β_{VC_U2,VC_D1}	Upstream v/c = 0.7 & Downstream v/c = 0.5	-1.08	-2.34
β_{VC_U3,VC_D1}	Upstream v/c = 0.85 & Downstream v/c = 0.5	-1.09	-2.37
$\beta_{VC\_U1,c1}$	Upstream v/c = 0.5 & c = 60	2.31	6.49
$\beta_{VC\_U1,c2}$	Upstream v/c = 0.5 & c = 90	0.85	2.39
$\beta_{VC\_U2,c1}$	Upstream v/c = 0.7 & c = 60	2.22	6.23
$\beta_{VC\_U2,c2}$	Upstream v/c = 0.7 & c = 90	1.10	3.08
$\beta_{VC\_U3,c1}$	Upstream v/c = 0.85 & c = 60	2.41	6.76
$\beta_{VC\_U3,c2}$	Upstream v/c = 0.85 & c = 90	1.24	3.49
$\beta_{VC\_U4,c1}$	Upstream v/c = 1.0 & c = 60	0.75	2.11
$\beta_{VC\_U1,WC}$	Upstream v/c = 0.5 & WC	-7.29	-20.46
$\beta_{VC\_U1,MC}$	Upstream v/c = 0.5 & MC	-4.02	-11.29
$\beta_{VC\_U2,WC}$	Upstream v/c = 0.7 & WC	-6.56	-18.42
$\beta_{VC\_U2,MC}$	Upstream v/c = 0.7 & MC	-4.35	-12.20
$\beta_{VC\_U3,WC}$	Upstream v/c = 0.85 & WC	-6.38	-17.90
$\beta_{VC\_U3,MC}$	Upstream v/c = 0.85 & MC	-4.50	-12.62
$\beta_{VC\_U4,WC}$	Upstream v/c = 1.0 & WC	-1.90	-5.32
$\beta_{VC\_U4,MC}$	Upstream v/c = 1.0 & MC	-3.49	-9.81
$\beta_{c1,WC}$	c = 60 & WC	4.69	16.99
$\beta_{c1,MC}$	c = 60 & MC	3.31	12.00
$\beta_{c2,WC}$	c = 90 & WC	1.88	6.82
$\beta_{c2,MC}$	c = 90 & MC	1.60	5.80
$\beta_{VC\_U1,c1}$	Downstream v/c = 0.5 & c = 60	-0.96	-2.71
$\beta_{VC\_U2,c1}$	Downstream v/c = 0.7 & c = 60	-0.78	-2.19
$\beta_{VC\_U1,WC}$	Downstream v/c = 0.5 & WC	-1.53	-4.30
$\beta_{VC\_U1,MC}$	Downstream v/c = 0.5 & MC	-1.12	-3.15
$\beta_{VC\_U2,WC}$	Downstream v/c = 0.7 & WC	-0.87	-2.45
$\beta_{VC\_U2,MC}$	Downstream v/c = 0.7 & MC	-0.72	-2.02
$\beta_{VC\_U3,WC}$	Downstream v/c = 0.85 & WC	-0.87	-2.45
No. of Observations = 225	Adjusted R^2 = 0.96		

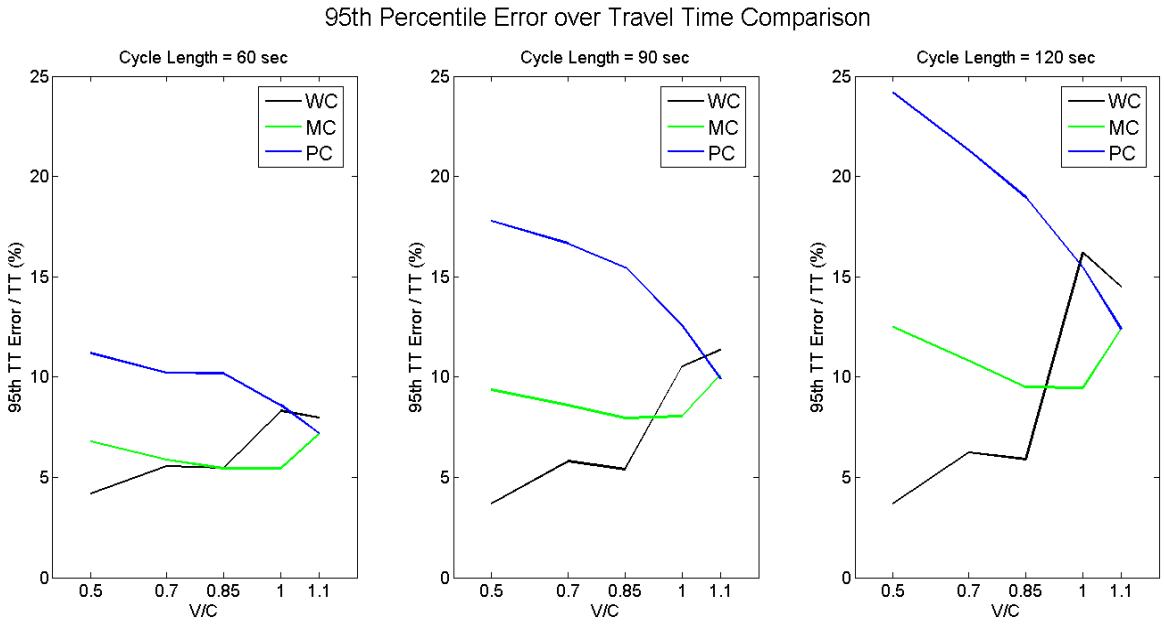
* Significant at 95% confidence level

The results of this approach are shown in Figure 4-9a. Each graph is for a different cycle length, the different series are for different levels of progression, and the magnitude of the v/c ratio at both the upstream and downstream detectors are indicated on the horizontal axis. Several observations can be made from these results:

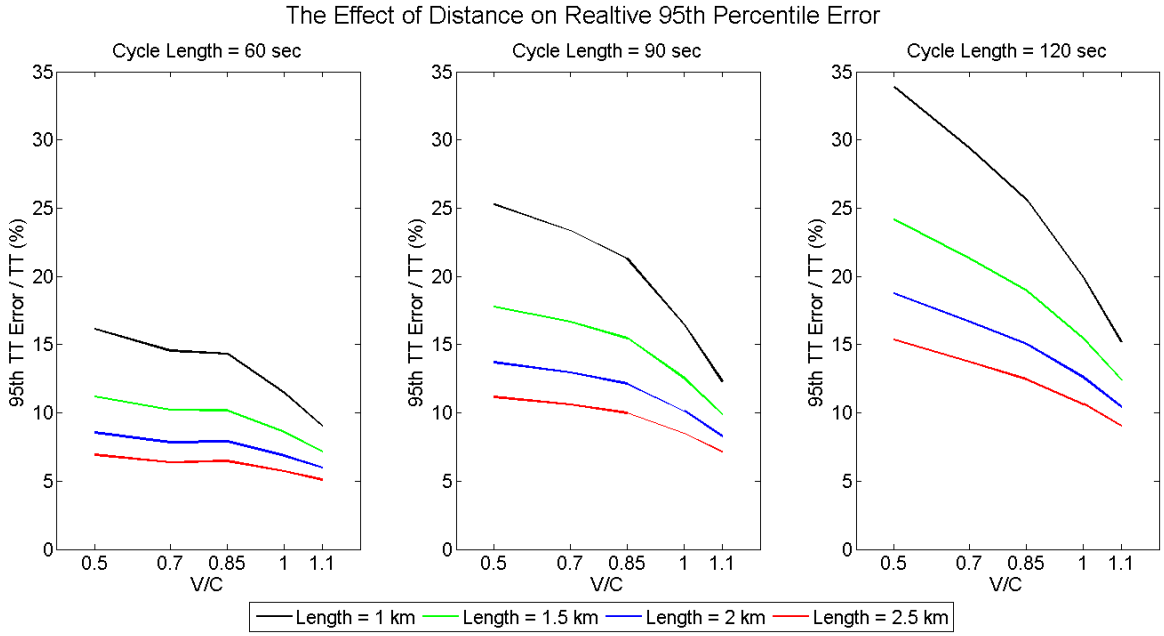
- For almost all scenarios the 95% confidence interval of the measurement error is greater than 5% of the true mean travel time.
- However, for some conditions (i.e. poor coordination, high cycle length and high v/c) the 95% confidence interval of the measurement error approaches 25% of true mean travel time.
- The magnitude of the ratio of the 95% confidence interval of the measurement error to the true mean travel time tends to increase as the cycle increases.
- The effect of coordination is more pronounced for low v/c.

Following the same procedure, the effect of segment length (i.e. distance between the upstream and downstream detectors) on the Bluetooth measurement error can be investigated. The impact of segment length on the ratio of the 95% confidence interval of the measurement error to the true mean travel time is depicted in Figure 4-9b for the case of poor coordination. These results show that, as expected, the ratio increases as the section length decreases (because the true mean travel time becomes smaller). From these results we observe that the ratio approaches 35% for a 1km section.

We emphasize that the proposed model reflects measurement errors only and therefore the magnitude of the travel time estimation error in practice will be larger than estimated by the model due to the presence of various outliers. Thus, the measurement confident interval ratios provided in Figure 4-9b can be viewed as a lower bound on the confident interval for the travel time error.



(a)



(b)

Figure 4-9: The 95th percentile travel time measurement error over the segment travel time, (a) different cycle length and coordination level, (b) different cycle length and segment length

4.3 Detecting Outliers in Bluetooth Data in Real Time[1]

As mentioned in Chapter 2, a number of travel time outlier detection algorithms exist in the literature. Although most of these algorithms have been developed primarily to address the outliers associated with other detection technologies (e.g., Automatic number plate recognition, automatic vehicle identification via transponders or toll tags), they are still applicable for travel times obtained from Bluetooth detectors.

Among the existing filtering techniques, a number of statistical-based methods are capable to be implemented in real-time applications. In these methods, outliers are identified on the basis of variation in the travel times over the recent past and the extent to which the current observation deviates from the recent observations. The mentioned statistical approaches include percentile test, deviation test and, and adaptive filtering.

The performance of these statistical-based algorithms can be qualitatively investigated though the schematic diagrams presented in Figure 4-10. The percentile test labels a predetermined and fixed proportion of the observations as outliers. As illustrated in Figure 4-10a, the percentile test is inefficient, as the designation of observations as valid or outlier is not dependent on the value of the observation and often leads to mis-classification of observations. The deviation test declares observations as outliers if they fall outside of an established validity window. This validity window is usually defined on the basis of the mean/median and the standard deviation of the observations in the investigated time interval. Generally, deviation tests that use median instead of the mean perform better, as median is less vulnerable to the outliers (Figure 4-10b and Figure 4-10c). However, the existence of a relatively large proportion of outliers in a given interval can adversely impact the accuracy of the median based deviation tests (Figure 4-10d).

In the adaptive filtering algorithms, which are enhanced versions of deviation tests, the validity window of the investigated time interval is established on the basis of the variability and trend of travel times in the recent time periods. These algorithms usually outperform the deviation tests, as they are less affected by local travel time turbulences (Figure 4-10e). Nevertheless, the reactive nature of the existing adaptive filtering algorithms makes them inappropriate for

[1] The content of this section is accepted for publication in the following journal paper:

- Salek Moghaddam, S. and B. Hellinga, "An Algorithm for Detecting Outliers in Bluetooth Data in Real-Time", 2014 (In Press), Transportation Research Record.

transient conditions (i.e., the formation or dissipation of the traffic congestion), when abrupt changes in the travel time are expected (Figure 4-10f).

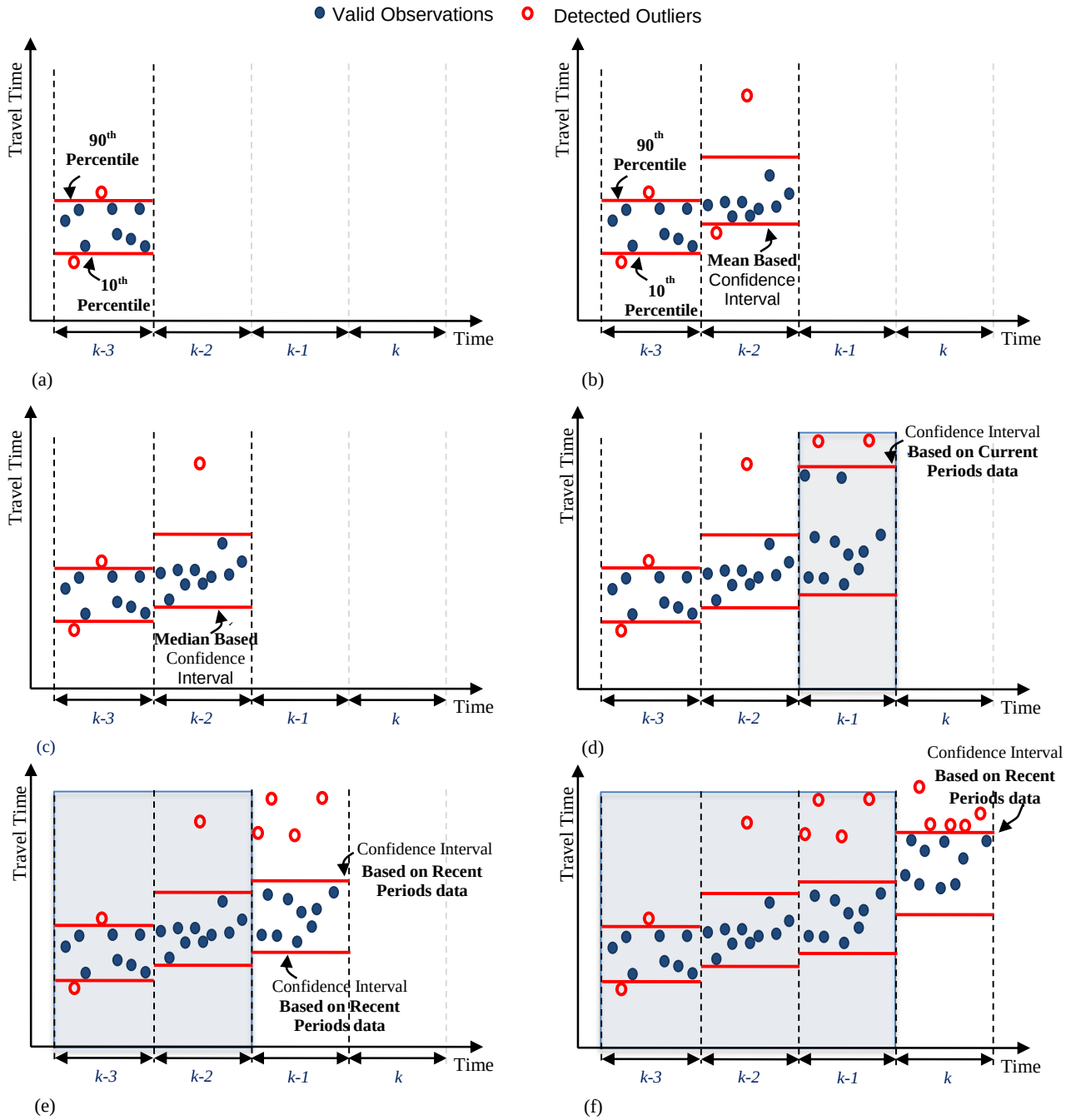


Figure 4-10: Illustration of operating characteristics of existing travel time outlier detection algorithms

In order to increase the capability of the adaptive filtering algorithms in capturing sudden changes in the travel-time trend, Dion and Rakha (2006) proposed some modifications to the existing reactive approaches. In their proposed algorithm, they introduce a subjective rule that if three consecutive outliers violate the validity window from the same side (i.e. all fall below or above the validity window), then the third outlier is labelled as a valid observation. Dion and Rakha (2006) showed that this modification significantly improves the performance of the filter and it appears to perform better than other filters described in the literature. Throughout the remainder of this section we use Dion and Rakha's algorithm as the benchmark algorithm against which we compare our proposed algorithm. However, we also note that the modification proposed by Dion and Rakha may result in some inconsistencies in the trend of estimated average travel times. In particular, the inclusion of an extreme outlier as a valid point can dramatically change the average travel time of an investigated time interval (especially when the number of observations is small) and therefore undermine the accuracy of the estimated travel times.

4.3.1 Proposed Model

As discussed in the previous section, reactive adaptive filtering techniques usually fail when the travel time of the traffic stream changes rapidly. This is illustrated in the schematic diagram of Figure 4-11a in which the reactive algorithm follows the recent trend of travel time and mislabels 4 observations in time interval *k*.

To enhance the performance of these algorithms we propose to predict abrupt changes in travel time by using both recent travel time trend and similar historical patterns. This concept is illustrated in Figure 4-11b. The travel time trend identified from historical data is depicted by the arrow at the end of period *k-1*. This historical trend is combined with the trend in the recent data (e.g. from *k-1*) in order to determine the validity window for period *k*. In this case, the historical data suggests an increase in travel times and consequently the validity widow boundaries increase and the 4 observations mislabeled by the reactive filter algorithm can now be correctly labeled as valid observations.

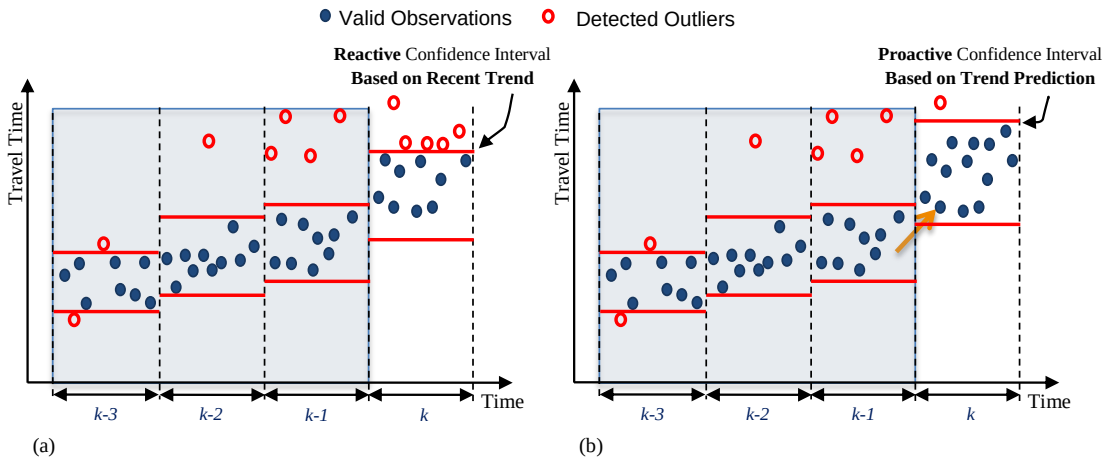


Figure 4-11: The schematic representation of the proposed outlier detection algorithm

4.3.1.1 Model Formulation

The proposed model is formulated to estimate the upper and lower boundaries of the validity window for each time interval and this requires the following steps:

- Calculating the expected average of the individual travel times for the investigated time interval.
- Calculating the expected variance of the individual travel times for the investigated time interval.
- Knowledge of the distribution of individual travel times.
- Constructing a confidence interval around the mean using the above mentioned information.

To calculate the expected mean and variance of individual travel times for the investigated time interval, we propose a hybrid method, which combines the predicted and estimated travel times to obtained a fused mean travel time ($\tau_{(i,k)}$) and a fused variance of travel time ($\sigma^2(\tau_{(i,k)})$). In the calculation of the fused variance of travel time, we conservatively ignored the covariance of the predicted and estimated travel times (due to the expected positive correlation of these two terms).

$$\tau_{(i,k)} = \lambda \cdot t_{(i,k-\Delta)} + (1-\lambda) \cdot \hat{t}_{(i,k)} \qquad \textbf{(4-3)}$$

$$\sigma^2(\tau_{(i,k)}) = \begin{cases} \lambda^2 \cdot \sigma^2(t_{(i,k-\Delta)}) + (1-\lambda)^2 \cdot \hat{\sigma}^2(t_{(i,k)}) & \text{if } n^v_{(i,k-\Delta)} > 1 \\ \hat{\sigma}^2(t_{(i,k)}) & \text{if } n^v_{(i,k-\Delta)} \in \{0,1\} \end{cases} \quad \textbf{(4-4)}$$

Where $\tau_{(i,k)}$ is the fused mean travel time for segment *i* (the investigated segment) during time interval *k* (the investigated time interval), $t_{(i,k-\Delta)}$ is the estimated mean travel time for segment *i* during the most recent time interval for which observations are available, $\hat{t}_{(i,k)}$ is the predicted mean travel time for segment *i* during time interval *k*, $\sigma^2(\tau_{(i,k)})$ is the fused variance of individual travel times for segment *i* during time interval *k*, $\sigma^2(t_{(i,k-\Delta)})$ is the variance of travel times for segment *i* during the most recent time interval for which observations are available, $\hat{\sigma}^2(t_{(i,k)})$ is the predicted variance of the individual travel times for segment *i* during time interval *k*, $n^v_{(i,k-\Delta)}$ is the number of travel time observations for segment *i* during the most recent time interval for which observations are available, Δ is the time lag, and λ is the model parameter. Following the approach proposed by Dion and Rakha (2006) this parameter is calculated on the basis of $n^v_{(i,k-\Delta)}$ and another parameter, β.

$$\lambda = 1 - (1-\beta)^{n^v_{(i,k-\Delta)}} \quad \textbf{(4-5)}$$

Individual travel times are not normally distributed. However, it has been suggested in the literature that travel time observations can be appropriately represented by the Log-normal distribution, as these data are usually skewed to the right (Li et al. 2006). Using one month Bluetooth travel time data collected on one of the investigated segments, we examined the distribution of individual travel times. Consistent with the literature, the resulted distribution is skewed to the right (Figure 4-12).

Adopting the aforementioned Log-normal assumption, we can establish a confidence interval for the individual travel times:

$$\Pr(LB < t^k_{ij} < UB) = 1 - \alpha \quad \textbf{(4-6)}$$

Where, *LB* and *UB* are the lower and upper bounds of the constructed confidence interval, respectively, t^k_{ij} is the travel time observation for vehicle *j* which enters the segment *i* during time interval *k*, and α is the significance level. These limits can be calculated through the use of

the inverse lognormal distribution function, assuming that the calculated fused mean and variance are good estimates of the population mean and variance:

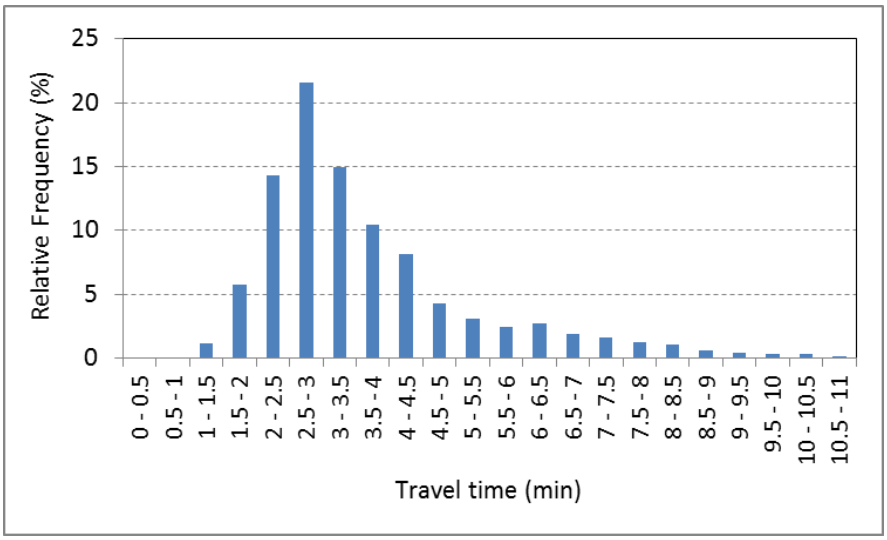


Figure 4-12: The distribution of individual travel times
(One month data for Bayview Ave. - Beaver Creek Rd. (1.9 km))

$$LB = LogNormal^{-1}\left[\alpha/2, \overset{N}{\tau}_{(i,k)}, \overset{N}{\sigma}(\tau_{(i,k)})\right] \quad \textbf{(4-7)}$$

$$UB = LogNormal^{-1}\left[(1-\alpha/2), \overset{N}{\tau}_{(i,k)}, \overset{N}{\sigma}(\tau_{(i,k)})\right] \quad \textbf{(4-8)}$$

$$\overset{N}{\tau}_{(i,k)} = \ln\left(\frac{\tau^2_{(i,k)}}{\sqrt{\sigma^2(\tau_{(i,k)}) + \tau^2_{(i,k)}}}\right) \quad \textbf{(4-9)}$$

$$\overset{N}{\sigma}(\tau_{(i,k)}) = \sqrt{\ln\left(\frac{\sigma^2(\tau_{(i,k)})}{\tau^2_{(i,k)}} + 1\right)} \quad \textbf{(4-10)}$$

Where, $\overset{N}{\tau}_{(i,k)}$ and $\overset{N}{\sigma}(\tau_{(i,k)})$ are the mean and standard deviation of the associated normal distribution, respectively. In each time interval (e.g., time interval *k*), individual travel time observations are labeled as valid if they lie within the constructed validity window.

The proposed model consists of two main tasks:

- Estimation of the mean ($t_{(i,k-\Delta)}$) and variance ($\sigma^2(t_{(i,k-\Delta)})$) of valid travel time observations for the most recent time interval for which observations are available. This is a straight forward task to accomplish.

- Prediction of the mean ($\hat{t}_{(i,k)}$) and variance ($\hat{\sigma}^2(t_{(i,k)})$) of the individual travel times for the investigated time interval. This is really a quasi-prediction process. The interval for which the predictions are made is actually in the past. However, prediction is still required as the validity of observations at the investigated time interval (interval *k*) has not been determined yet. The predictions can be made using any appropriate travel time prediction approach. The particular prediction model we have used in this research is an adaptation of a model developed for dynamically predicting near-future arterial times which is described later in this thesis (Section 5.1).

4.3.1.2 Prediction model

The quasi-prediction process is performed using a data driven pattern recognition model. The k-NN (k Nearest Neighbor) model is defined by four elements, namely: the feature vector; the distance metric; number of nearest neighbors to be selected; and Local Estimation method. As will be described later, a full factorial experimental design was used to find optimal values for model parameters. These same values are adopted for the quasi-prediction process here.

The feature vector consists of two types of attributes: (1) The trend in the travel time as observed during the most recent 15 minutes (i.e. three 5-minute intervals); and (2) The summation of hits during the most recent 15 minutes (i.e. three 5-minute intervals) at both the upstream and downstream detectors of the investigated segment.

We selected the commonly used Mahalanobis distance as the distance metric because it considers both the variance and covariance of the feature vector variables. The number of nearest neighbors was set equal to 50. The weighted arithmetic mean of the travel times from the 50 nearest neighbors is considered as the local estimation method for the mean and provides a value for $\hat{t}_{(i,k)}$. The 45 days prior to the date of the test day were considered as the historical dataset and the search space was restricted to 4 hour prior and 4 hour after the time of day of the interval being investigated.

This k-NN was originally developed for predicting the mean travel time; however, in this research we also need to predict variance of individual travel times (i.e. $\hat{\sigma}^2(t_{(i,k)})$). Consequently, the variance of the target individual travel times has been estimated by pooling the individual travel time observations from all of the 50 nearest neighbors and calculating their variance (estimation of the population variance based on the sample variances).

4.3.2 Model Calibration

The only parameter in the proposed model that required calibration is β from Equation 4-5. Calibration of β was done through a sensitivity analysis. This sensitivity analysis is performed by finding the value of beta that minimizes the estimation error. We don't know the true mean travel times so we treated the estimated travel times obtained from the Benchmark method (Dion and Rakha 2006) as the ground truth. Error is measured as the mean absolute relative error (MARE) between the mean travel times estimated from the proposed model ($t_{(i,k)}$) and the Benchmark ($t^*_{(i,k)}$):

$$MARE_{(i)} = 100 \times \frac{1}{N} \sum_k \frac{\left| t_{(i,k)} - t^*_{(i,k)} \right|}{t^*_{(i,k)}} \quad \textbf{(4-11)}$$

Where, N is the number of estimation instances and i represents the investigated segment.

We recognize that the Benchmark model fails to perform well under some conditions and therefore we selected data from 5 days for which the Benchmark model performed well.

Calibration was performed using the data from 4 segments from the 5 days. The daily MARE is computed across the 5 days for each segment separately for each value of beta. Figure 4-13 shows that for each of the 4 segments, a value of beta equal to 0.2 provides the minimum MARE.

The sensitivity of the results to the value of parameter beta is further illustrated in Figure 4-14a-f for data from two different directions of the same roadway segment (Highway 7 between Beaver Creek Rd. and Highway 404) on November 9th, 2011.

On the Eastbound direction travel times increase very rapidly during the PM peak. When $\beta = 0$, implying that the validity window is based only on the historical data (i.e., k-NN model

predictions), the outlier detection model performs poorly as it incorrectly labels the valid congested travel times in the PM peak as outliers. However, when $\beta = 1$, implying that the validity window is based only on the latest available time interval estimates, the model appears to perform correctly during the PM peak as well as during the rest of the day. The model also appears to perform well for $\beta = 0.2$. Under this condition, the validity window is based on the fusion of the historical data and the latest available data from the current day.

The graphs in the right hand column of Figure 4-14 show the model performance for data collected on the Westbound direction. On this day the mean travel time has greater variations with time of day, but travel times change less rapidly. In this case, the performance of the model is better when $\beta = 0$ than when $\beta = 1$. This is mainly due to the high fluctuation of the travel time trend throughout the course of the day (as discussed in Chapter 1, these fluctuations are relatively more common on arterials compare to freeways). Once more the use of beta equal to 0.2 to combine the two estimations (k-NN and latest observations) seems to cause the model to perform well throughout the day.

In general, the results support the conclusion that the fusion of the estimates from the historical and current data with $\beta = 0.2$ provides best model performance.

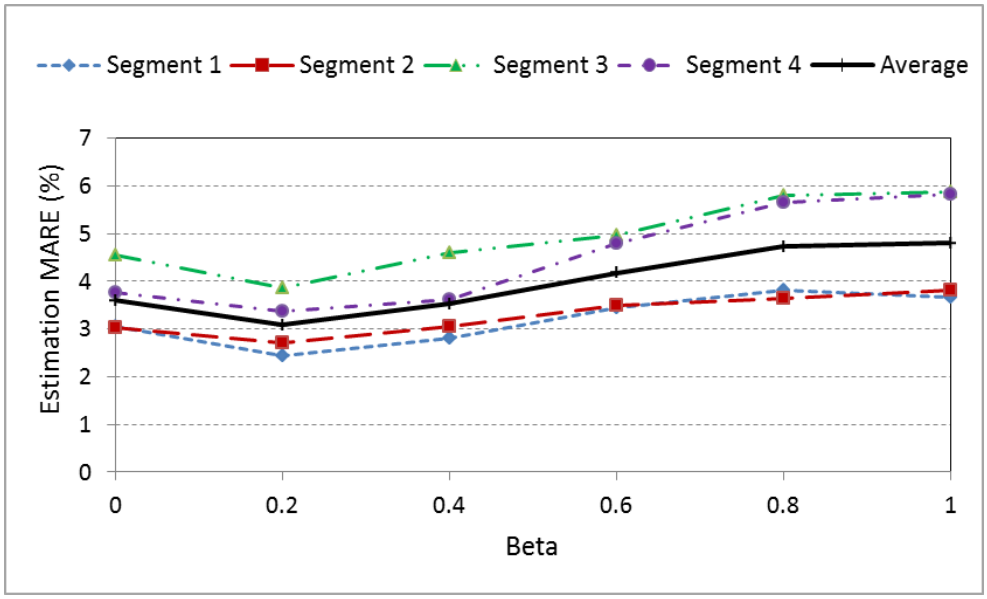


Figure 4-13: Impact of model parameter β, based on the daily estimation MARE (Average over five random days)

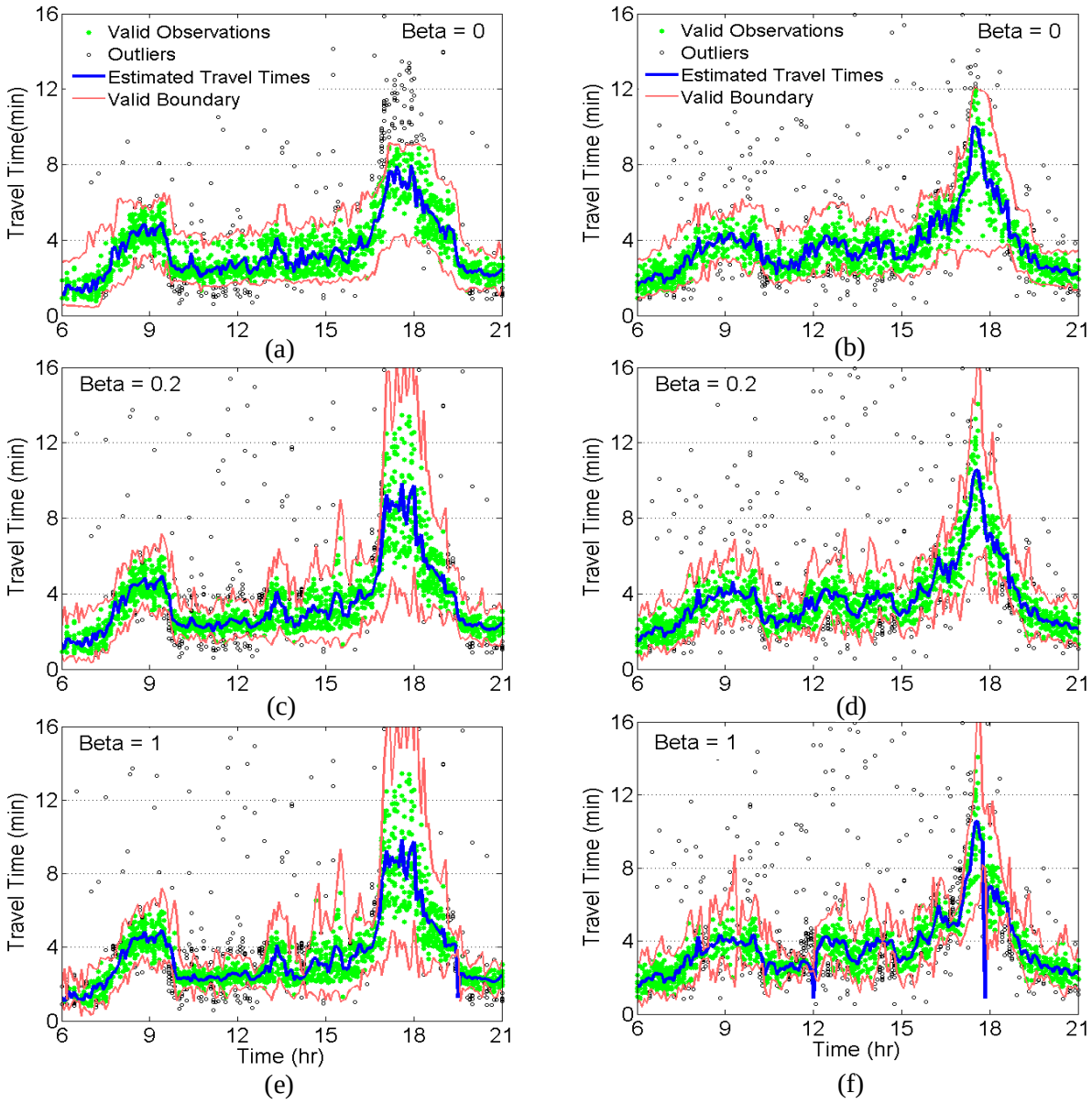


Figure 4-14: Impact of model parameter β, based on travel time trend (2 sample observations)

4.3.3 Model Evaluation

The evaluation of outlier detection models can be conducted using either field data or simulated data. The challenge when using field data is determining truth (i.e. whether or not a specific measured travel time is an outlier). It may be possible to collect travel time data via some other sensor technology, such as automatic license plate recognition (ANPR) system, as an

independent measure of the travel times. However, this approach suffers from the following complications:

- It is very difficult to associate a travel time measured by the alternative technology with a specific travel time obtained from the Bluetooth detector as there is typically no way to ensure the same vehicle was measured by both sensor systems.
- The alternative sensor technology will also be subject to measurement error and the magnitude and distribution of these errors may not be known.
- Outliers in the data occur from a variety of sources, not only measurement error. Sampling bias and vehicles making en-route detours also contribute to the presence of outliers. Data from the alternative sensor technology may also be subject to these sources of outliers.

The use of simulated data is appealing because it is possible to know what the true travel time was as this can be extracted from the simulation. However, the sources of errors that occur in the field data must be introduced into the simulated data and the magnitude and distribution of these errors usually are not well understood. The potential application of simulated data in evaluating the performance of the outlier detection algorithms has been investigated in Section 4.4.

To evaluate the performance of the proposed proactive outlier detection algorithm, we have elected to use field data, but rather than quantitatively evaluate the performance of the proposed algorithm, we have assessed the performance by observing the behaviour of the algorithm and comparing this behavior to the benchmark model on the same set of field data. This is the same approach adopted by some other researchers (e.g. Dion and Rakha (2006).

4.3.3.1 Evaluation Results

The calibrated outlier detection algorithm and three algorithms from the literature (median based deviation test (Clark et al. 2002); Dion and Rakha's original adaptive filter (Dion and Rakha 2006); and Dion and Rakha's modified adaptive filter, which we refer to as the Benchmark (Dion and Rakha 2006)) were applied to 10 days of data from the field data set. These days were different from those used in the calibration process. Figure 4-15 shows the outlier detection results of the four models for two sample days.

It can be observed that the methods from the literature fail to perform well under some conditions. For the median based deviation test the instability of the validity window results in

frequent high jumps in the estimated travel times (Figure 4-15a-b). The original Dion and Rakha's adaptive filter, as expected, fails to capture the sudden changes of travel time (Figure 4-15c-d). The Dion and Rakha's modified adaptive filter also suffers from some occasional high jumps in the estimated travel times (Figure 4-15e-f).

For example, in Figure 4-15e at approximately 3 pm, the upper bound of the validity window becomes extremely large such that an extreme outlier is incorrectly labeled as valid point and the inclusion of this outlier results in the estimated mean travel time becoming approximately 9 times the true value. This problem arises when a relatively large number of outliers exist and the variation of travel time is high – conditions that may occur frequently on arterials. For these conditions, the heuristic method employed by the modified Dion and Rakha's model to make it more responsive (i.e. labeling the third consecutive outlier as valid regardless of its value) can lead to highly unrealistic results. Figure 4-15f illustrates similar occurrences of this problem.

In their original paper, Dion and Rakha acknowledged the existence of some extreme outliers even after the application of their algorithm. To handle these cases they suggested a complementary overtaking rule approach, but this method limits the application of the algorithm for offline analysis only.

An examination of Figure 4-15g-h indicates that for these days, the proposed method is able to perform well without suffering from the issues displayed by the methods from the literature.

In addition to illustrating algorithm performance through graphs, we attempted to quantify performance of the proposed model relative to the Benchmark model (i.e. Dion and Rakha's modified filter). It is not possible to directly evaluate the model performance because the truth of which observations are outliers is not known. Instead, we selected 10 days for which the Benchmark was observed to suffer from obviously poor performance (i.e. large error in the mean travel time) during some time intervals. We manually fixed the observed errors by interpolating the travel time values of the neighbor time intervals. The resulting time series of mean travel times were considered as the ground truth and the proposed and Benchmark methods were compared to these data. We performed our comparison based on two different measures of effectiveness (MOE), namely mean absolute relative error (MARE) computed over the whole day, and the MARE computed over only those time intervals for which the Benchmark method was observed to perform poorly (i.e., periods for which the mean travel time was interpolated).

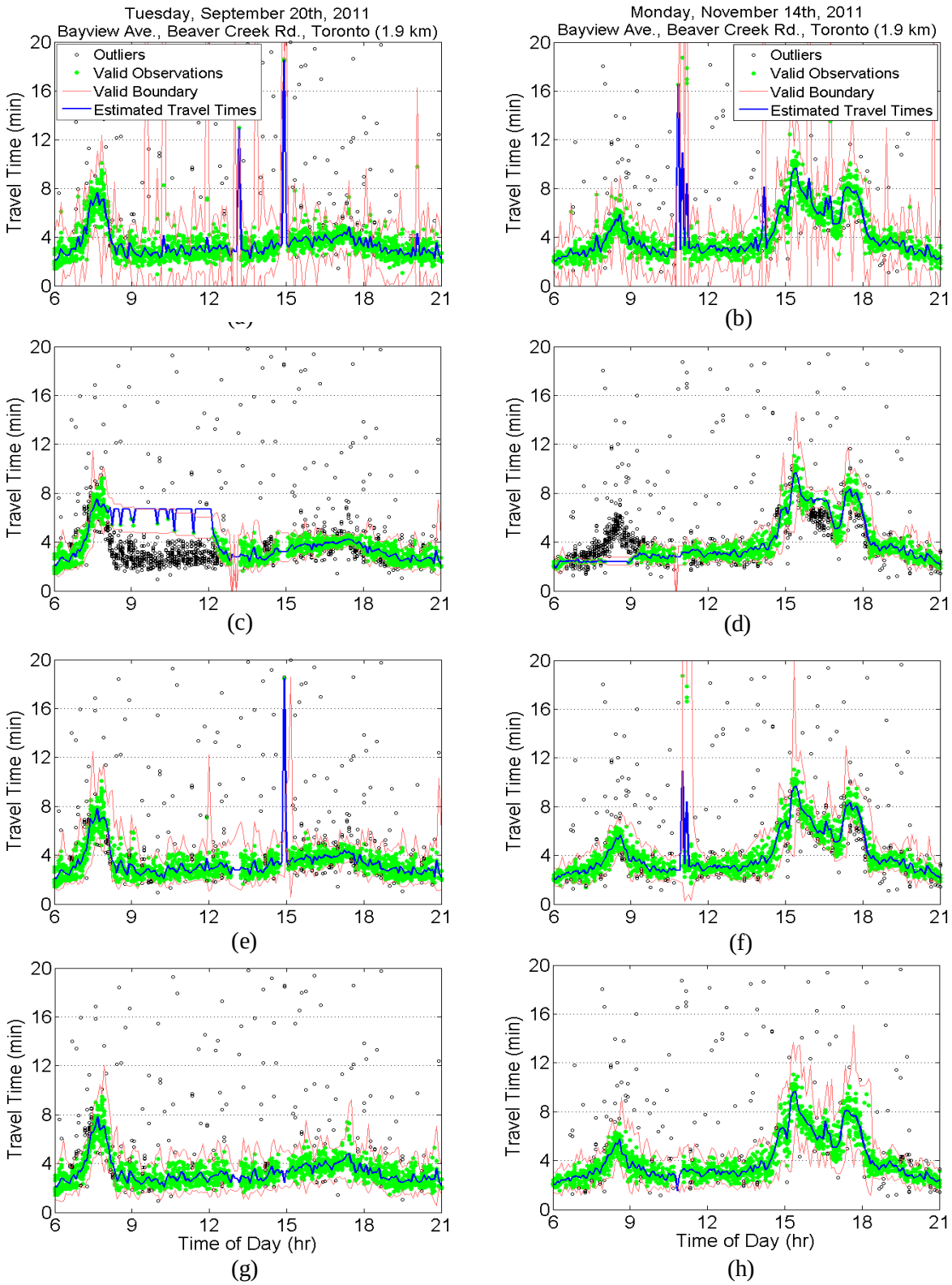


Figure 4-15: Illustration of the performance of the median based model (first row); Dion and Rakha original filter (second row); Dion and Rakha modified filter (third row); and proposed method (last row) (Highway 7, Toronto)

The results, provided in Table 4-3 show that the proposed method dramatically outperforms the Benchmark in terms of the second MOE (i.e. when the Benchmark method is observed to perform poorly, the proposed model performs well). Furthermore, the comparison of the daily MAREs shows that the proposed method performs as well as the Benchmark in all other time intervals during the day. It should be noted that the latter comparison is not a fair one as for most of the intervals the Benchmark absolute relative errors are basically zero.

It is worth noting that the mentioned sudden jumps in the trend of estimated travel times resulted from the Benchmark method are not rare occurrences; in fact, these poor performances happen almost frequently in the available dataset.

Table 4-3: Comparison of the performance of the proposed algorithm and the benchmark method

Test Day	MARE (%)			
	Benchmark Method		Proposed Method	
	Selected time periods	Entire Day	Selected time periods	Entire Day
June 6th 2011 (Bayview Ave. – Beaver Creek Rd.)	195.4	6.4	29.4	6.3
June 24th 2011 (Bayview Ave. – Beaver Creek Rd.)	177.3	1.2	21	2.1
September 20th 2011 (Bayview Ave. – Beaver Creek Rd.)	185.9	2.5	53.1	3.0
November 14th 2011 (Bayview Ave. – Beaver Creek Rd.)	180.6	2.4	10.4	2.7
June 6th 2011 (Beaver Creek Rd. – Bayview Ave.)	106.9	1.9	12.9	2.0
May 30th 2011 (Beaver Creek Rd. – Hwy 404)	92.7	1.7	29.3	5.8
June 2nd 2011 (Beaver Creek Rd. – Hwy 404)	145.8	2.4	27.6	5.0
June 28th 2011 (Beaver Creek Rd. – Hwy 404)	185.0	4.3	47.7	4.7
June 2nd 2011 (Hwy 404 – Beaver Creek Rd.)	94.7	3.8	27	5.3
June 6th 2011 (Hwy 404 – Beaver Creek Rd.)	113.6	5.1	33.2	5.6
Mean	147.8	3.2	29.2	4.3
Confidence Interval Lower Bound	119.7	2.0	20.1	3.2
Confidence Interval Upper Bound	175.9	4.3	38.2	5.3

Due to the iterative nature of travel time outlier detection in real time applications (discussed in Chapter 3), the validity of some individual travel time observations may change throughout the analysis period. These changes may introduce some level of instability into the recent trend of travel time estimates and further reduce the accuracy of the travel time predictions.

Following the above discussion, the stability of the results provided by the proposed filtering algorithm is investigated against that of the Benchmark for two sample days through Figure 4-16. In this figure, those outliers which were incorrectly included as valid points are indicated by Blue squares and the mislabeled valid points are denoted by green circles. The results associated with the Benchmark and Proposed methods are presented in the left and right hand columns respectively. From these results the followings two observations can be made:

- The number of mislabeled travel time observations is higher for the Benchmark method. This is mainly due to the heuristic nature of this method which has been discussed earlier in this document.
- In some cases the estimated travel time can be significantly affected by the incorrect inclusion of some extreme outliers (Figure 4-16a – right after 4:00 pm).

The evaluation results suggest that the proposed outlier detection algorithm is robust and it can be incorporated within an automatic real-time travel time estimation/prediction framework.

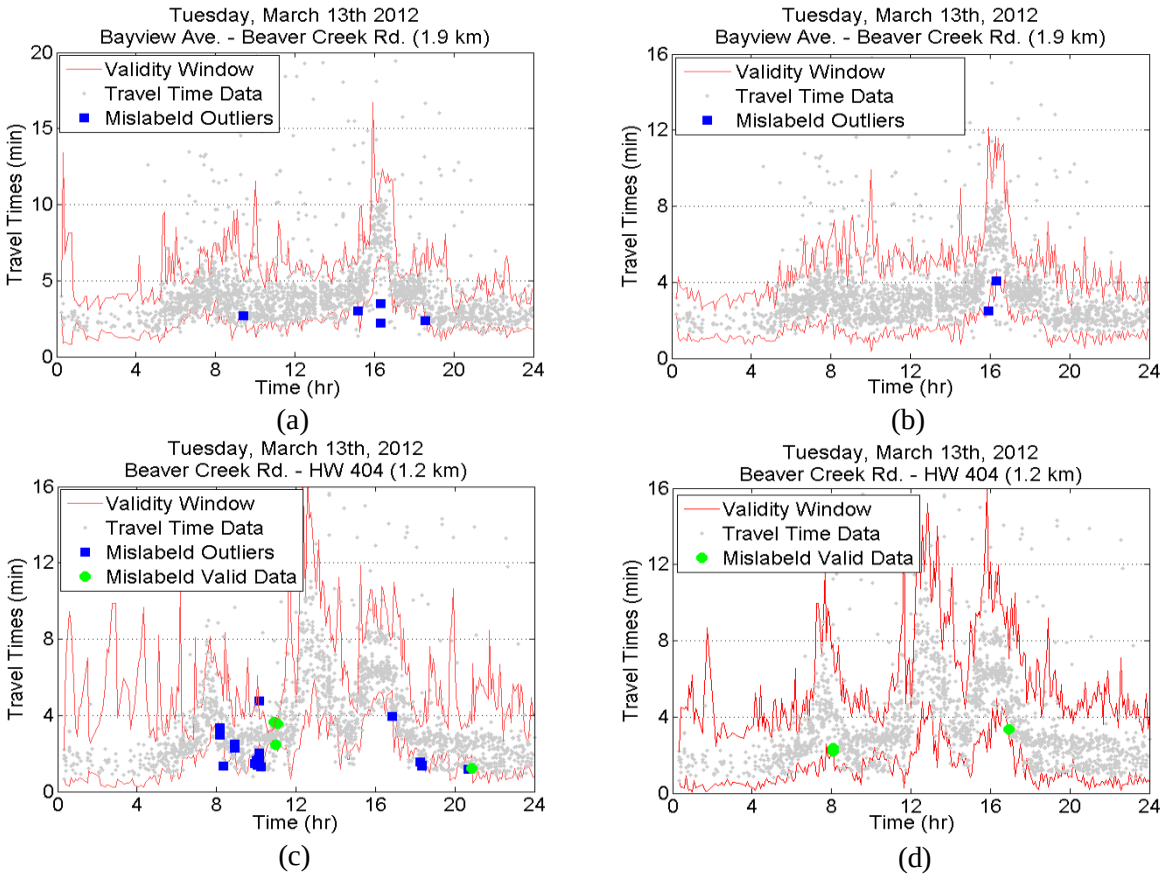


Figure 4-16: Stability of the resulted travel time validation for the Benchmark (left column) and the Proposed method (right column) - two sample days (HWY 7, Toronto)

4.4 Evaluating the Performance of Travel Time Outlier Detection Algorithms (Simulation Approach)[1]

As mentioned earlier, simulation data can be used to evaluate the performance of outlier detection algorithms. The knowledge of true travel times and the possibility of investigating the performance of each algorithm under different traffic conditions are the main advantages of using such data. The potential application of simulated data in evaluating the performance of the outlier detection algorithms has been investigated in this section.

The developed simulation framework in this section is only appropriate for evaluating algorithms relying on the current data. In order to investigate the performance of the proposed outlier detection algorithm (presented in Section 4.3), the current framework needs to be extended to permit the evaluation of algorithms that make use of historical travel times. However, as almost all other real-time outlier detection algorithms in the literature rely only on current data, the proposed algorithm can be effectively utilized for evaluating their performance.

The developed simulation framework consists of three components, namely: (1) Generation of true travel times; (2) Generation of outliers; and (3) Application and evaluation of the outlier algorithms. Each of these components is described in the following sections.

4.4.1 Generation of true travel times

The true travel times (i.e. travel times without error) are generated using the VISSIM micro-simulation model. For the example application in this research, a segment of a 4 lane arterial in the City of Waterloo, Ontario, Canada was modeled (Figure 4-17). The segment is 1.62 km in length and contains three signalized intersections, each with actuated left turns. Hourly turning movement counts and signal timing data were obtained for each intersection and used within the model to simulate typical weekday conditions from 6am to 8 pm (14 hours).

The time when a vehicle passed the intersection at Westmount Rd and the time taken to travel from Westmount Road to Philip Street were extracted from the VISSIM model for each vehicle. These data represent the true travel times.

[1] The content of this section is published in the following journal paper:

- Salek Moghaddam, S. and B. Hellinga, "Evaluating the Performance of Travel Time Outlier Detection Algorithms", 2013, Transportation Research Record, Vol. 2338, pp. 67 - 77.

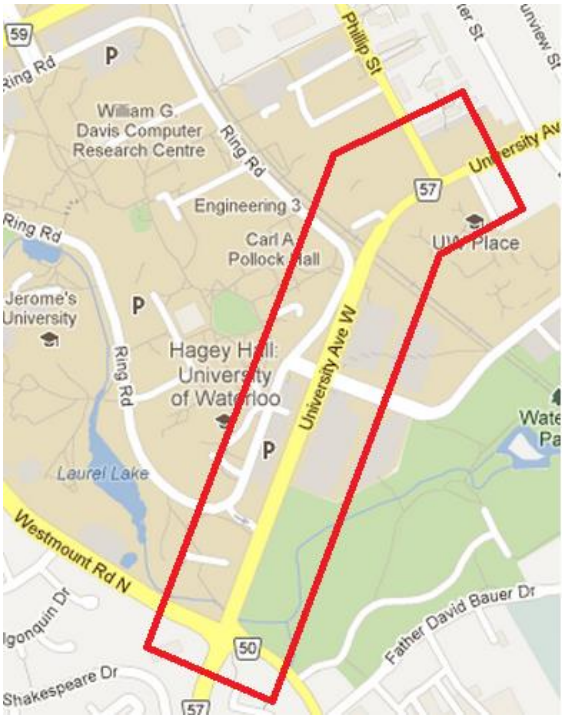


Figure 4-17: The simulated arterial segment - University Ave. (Westmount to Phillip), Waterloo (Source: Google Maps)

4.4.2 Generation of Outliers

The second component of the proposed framework exogenously generates outliers using a Monte Carlo simulation process. Four types of outliers are considered, namely: (1) En-route stops; (2) Non-auto observations; (3) Multiple devices in a single vehicle; and (4) measurement error.

4.4.2.1 En-route stops

En-route stops result in measured travel times that are longer than the true travel time by an amount equal to the duration of the en-route stop. We define (P_e) as the fraction of vehicles that make en-route stops. The duration of the en-route stop is assumed to follow a log-normal distribution with a mean of 15 minutes and a standard deviation of 10 minutes.

4.4.2.2 Non-auto Observations

Non-auto observations may consist of travel times obtained from Bluetooth devices carried via a mode other than private automobile (e.g. walk, bike, bus, etc.). In this study, we conservatively consider public transit buses as the only available non-auto mode. Including other non-auto

modes would introduce more and larger outliers; however, the appropriate distribution for the travel times of these other non-auto modes is not known.

We define P_T as the percent of buses in the traffic stream. A previous study (Salek Moghaddam et al. 2011) used Waterloo public transit AVL (automatic vehicle location) data to examine the public transit bus speeds and travel times. On the basis of those results, we model public transit bus speeds as a Normal distribution with a mean of 0.6 times the speed of personal automobiles and a standard deviation of 2 km/h. A transit vehicle is assumed to carry 25 passengers each of whom has the same probability of carrying a discoverable Bluetooth device.

4.4.2.3 Multiple devices in one vehicle

As described in Section 4.1, bias may occur when multiple devices residing in the same vehicle are detected. We define P_m as the fraction of the number of true travel times for which two measured travel times are obtained.

4.4.2.4 Measurement Error

Measurement error occurs as a result of the imprecise detection time of Bluetooth detectors for all of the individual travel time observations (including the generated outliers) at both upstream and downstream detectors. Empirical detection time error distributions were developed using a combination of VISSIM simulations and a Monte Carlo simulation technique. These distributions were developed for 5 different levels of v/c (0.5, 0.7, 0.85, 1.0, 1.1), 3 cycle lengths (60, 90, 120 seconds), and 3 levels of coordination (Well-coordinated; Moderately coordinated; and Poorly coordinated[1]) for a total of 45 different scenarios. A Monte Carlo simulation was performed using the 45 detection time error distributions to generate measurement errors for each 5-minute aggregation time period using the cycle length, v/c ratio, and level of coordination associated with the intersection for that time of day (the procedure explained in Section 4.2 of this thesis).

Finally, Bluetooth penetration rate (P_B) is considered as the fourth parameter in this framework.

Figure 4-18a-b illustrates the results of applying the proposed framework to the study arterial segment. Figure 4-18a illustrates the measured data assuming a Bluetooth penetration rate (P_B)

[1] The levels of coordination correspond to the HCM arrival types of AT6, AT3, and AT1 respectively.

of 10% and Figure 4-18b shows the same data but for P_B equal to 15%. All observations (i.e., Valid points and Outliers) include measurement error. In other words, the various outlier points contain measurement error plus the associated outlier error.

As expected, outliers labeled as "En-route stops" contain the largest errors and therefore may be the easiest (and most important) to identify as outliers. The errors associated with multiple devices and non-auto modes (i.e. transit) tend to be much smaller.

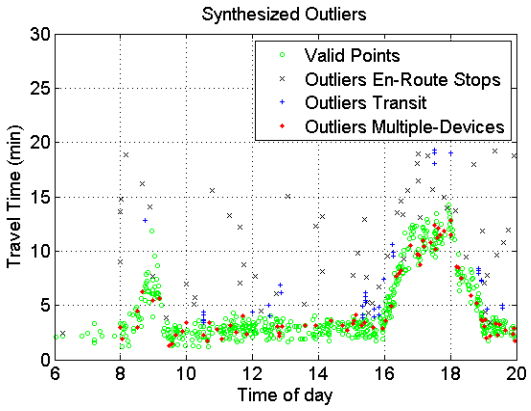
4.4.3 Application of the Outlier Algorithms

Two outlier detection algorithms proposed by Dion and Rakha (2006) have been coded and applied to the data illustrated in Figure 4-18a-b (the detailed formulation of these methods is provided in Chapter 2). These algorithms are labeled within this study as Method 1 (Dion and Rakha original filter) and Method 2 (Dion and Rakha modified filter).

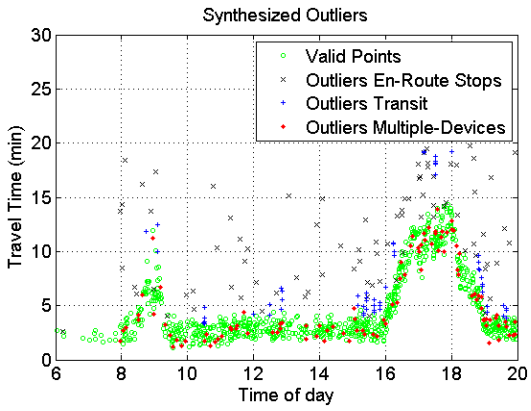
Figure 4-18c-d shows the results of applying Method 1 to the two sets of data illustrated in Figure 4-18a-b, respectively. Figure 4-18e-f shows the results of applying Method 2 to the same sets of data.

In each graph (Figure 4-18c-f), the output from the outlier detection algorithm is depicted in terms of valid points (those that have not been designated as outliers by the algorithm) and outliers (those that have been designated as outliers by the algorithm). The solid lines labeled "Valid boundary" are computed by the outlier detection algorithms and used to classify observations as "valid" or as "outlier". Several observations can be made from an examination of these results:

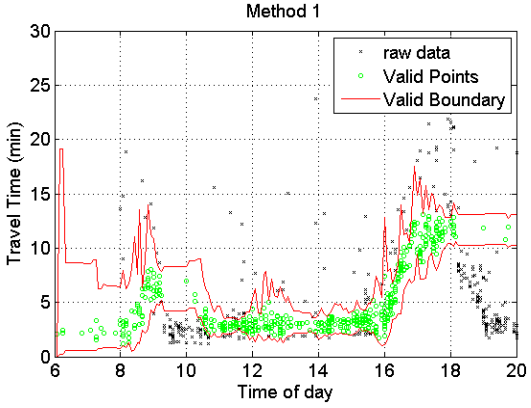
- The behaviour of the outlier detection algorithms appears to be quite sensitive to the sample size (as controlled by the Bluetooth penetration rate, P_B). This is evident by comparing the results in Figure 4-18c with those in Figure 4-18d.
- It is difficult to determine which outlier detection algorithm (Method 1 or Method 2) is better for the case of higher Bluetooth penetration rate.
- It is not clear whether or not the outlier detection algorithms respond differently to the different sources of errors.



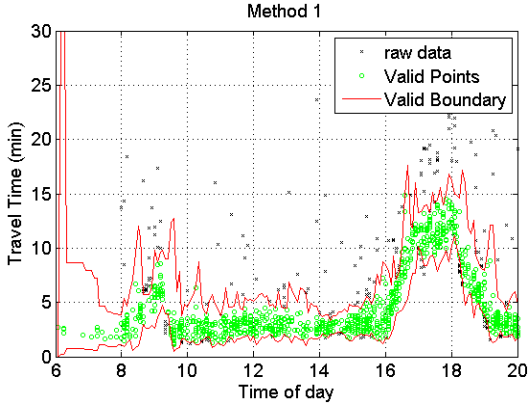
(a) Synthesized Data (P_B = 0.1)



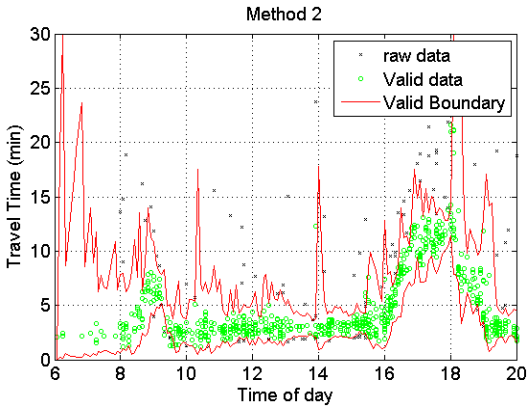
(b) Synthesized Data (P_B =0.15)



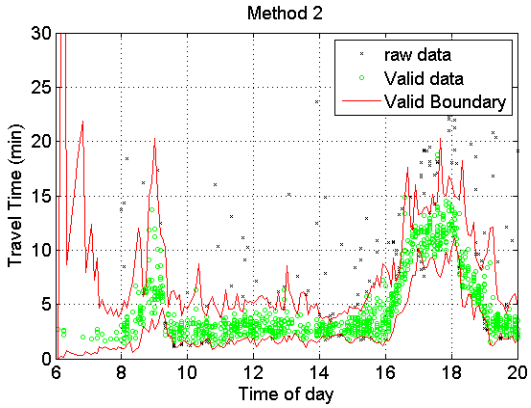
(c) Method 1: (P_B = 0.1)



(d) Method 1: (P_B = 0.15)



(e) Method 2: (P_B = 0.1)

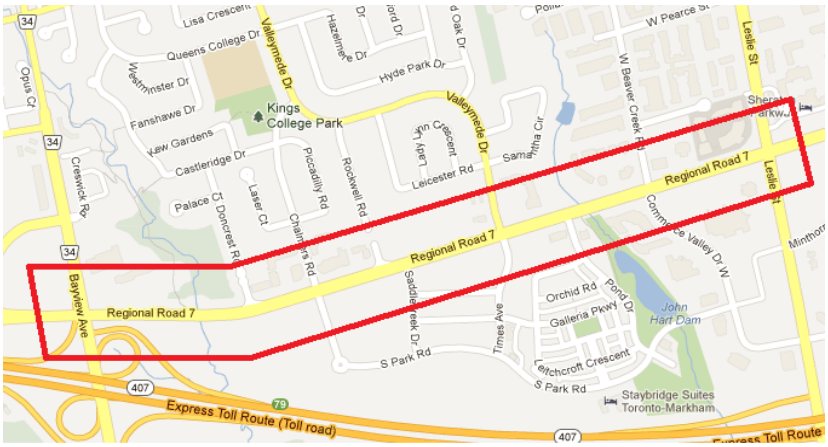


(f) Method 2: (P_B = 0.15)

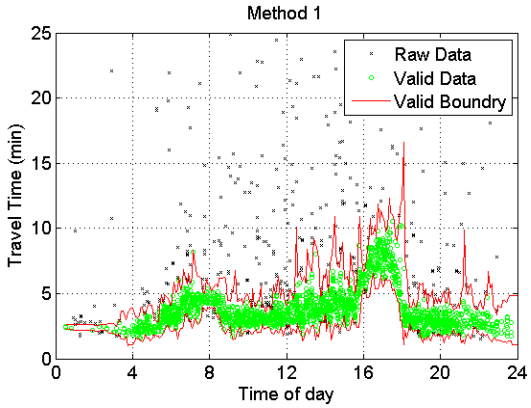
Figure 4-18: Generated outliers and investigated methods performance (Bluetooth penetration = 10% & 15%)

To provide a context for the interpretation of the results in Figure 4-18, we also applied Method 1 and Method 2 to a set of field data (Figure 4-19). These Bluetooth measurements were obtained from a 2.25 km long section of urban arterial in Toronto, Ontario, Canada. The result of

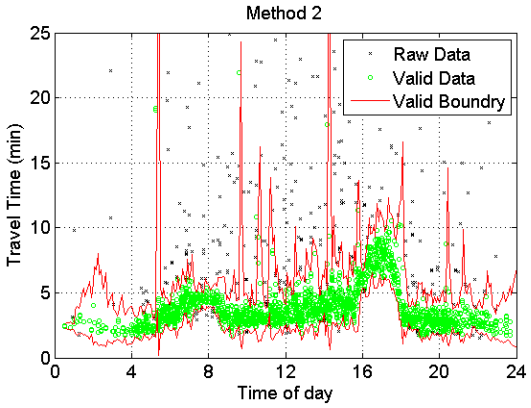
applying Methods 1 and 2 are shown in Figure 4-19b-c respectively. It can be argued that variation exhibited by the field data is larger than the variation contained within the synthesized data shown in Figure 4-18. However, the data in Figure 4-18 do not contain outliers associated with modes other than auto and bus. Furthermore, there were construction activities active in the study section shown in Figure 4-19a and this likely introduced larger variations in the valid data as well as additional outliers associated with detecting construction vehicles, etc. It should also be noted that it is not possible to conclude on the basis of the results in Figure 4-19-c which method is superior as the true travel times are unknown.



(a) Highway 7 (Leslie to Bayview), Toronto (2.25 km) – Source: Google Maps



(b) Outlier Detection (Method 1)



(c) Outlier Detection (Method 2)

Figure 4-19: Outlier detection using investigated algorithms - Highway 7, Leslie to Bayview, Toronto (Tuesday Sep 20th 2011)

4.4.4 Experiment Design

The discussion in the previous section suggests that a more rigorous outlier detection algorithm evaluation methodology must be applied. The first step in this process is to define a set of performance measures.

Three performance measures are considered:

- *Percent detected outliers:* This is computed as the number of true outliers correctly labeled as outliers by the algorithm, divided by the total number of true outliers. This measure was computed separately for each of the 3 categories of outliers.
- *Percent wrong detections:* This is computed as the number of valid observations that have been labeled as outliers by the algorithm, divided by the total number of valid observations.
- *Relative travel time improvement index:* The problem with the previous two performance measures is that they are not sensitive to the impact that the incorrect identification of outliers has on the estimate of the true travel time. Consequently, we propose a new performance measure which we term the relative travel time improvement index which is computed as follows:

$$RTTI = \sum_{k} \left(\frac{\left| t_{(i,k)}^{outlier} - t_{(i,k)}^{true} \right|}{t_{(i,k)}^{true}} - \frac{\left| t_{(i,k)}^{\text{detected}} - t_{(i,k)}^{true} \right|}{t_{(i,k)}^{true}} \right) \quad \textbf{(4-12)}$$

Where:

RTTI = daily relative travel time improvement index

$t_{(i,k)}^{outlier}$ = average of measured travel times including measurement error and outlier error on segment *i* and within interval *k* (minutes)

$t_{(i,k)}^{true}$ = average of the true travel times including no error on segment *i* and within interval *k* (minutes)

$t_{(i,k)}^{detected}$ = average of all valid measurements (i.e. excluding measurements labeled as outliers) on segment *i* and within interval *k* (minutes)

As discussed earlier, the ultimate goal for applying the outlier detection algorithms is to improve travel time estimations at the aggregate level (e.g., 5 minutes), as in traffic engineering studies we are more concerned with the aggregated travel times rather than individual observations (because of their high variation). *RTTI* is specifically defined to capture this aggregate level improvement in travel time estimations.

We illustrate the behavior of each of these three measures of performance by computing the measures for the data shown in Figure 4-18 (Table 4-4). Consider first the results for the Bluetooth penetration rate of 10% (corresponding to Figure 4-18c and Figure 4-18e). Method 1 provides a higher performance in terms of percent detected outliers for all three categories of outlier than does Method 2. However, Method 1 performs much more poorly than Method 2 in terms of incorrectly labeling valid points as outliers (i.e. high Percent Wrong Detections). These incorrectly labeled points are visible in Figure 4-18c as a cluster from 9 – 11 am and a cluster from 18 – 20 pm. Examining the *RRTI* results indicates that Method 1 actually performs worse than not using an outlier algorithm at all (as indicated by the negative value of *RTTI*). Method 2 provides a 9% improvement in the 5-minute average travel times.

When the Bluetooth penetration rate increases to 15%, both outlier detection algorithms show improved performance, implying that the performance of the algorithms, particularly Method 1, are quite sensitive to the Bluetooth penetration rate.

Table 4-4: Performance measures for data in Figure 4-18

Bluetooth Penetration Rate		10%		15%	
Outlier Algorithm		Method 1	Method 2	Method 1	Method 2
RTTI		-14.8%	9.3%	11.1%	12.7%
Percent Detected Outliers	En-route Stop	78.5%	69.2%	79.9%	72.2%
	Transit	59.5%	33.3%	50.0%	34.2%
	Multiple Device	28.0%	0.0%	12.3%	10.5%
Percent Wrong Detections		29.5%	6.9%	8.1%	5.8%

The results in Table 4-4 are associated with a single set of data and therefore cannot be used to make conclusions about the relative performance of Method 1 versus Method 2, Therefore, we use the proposed framework to conduct a multi factor comparison of the two outlier detection

algorithms in order to capture the effect of all the synthesized sources of errors. The proposed framework was used to generate 81 different scenarios using the following parameter values:

- % vehicles with en-route stops: P_e = {5%, 15%, 25%}
- % buses in traffic stream: P_T = {1%, 3%, 10%}
- % vehicles with multiple devices on board: P_m = {10%, 30%, 50%}
- Bluetooth penetration rate: P_B= {1%, 5%, 15%}[1]

Each scenario consisted of data generated for 14 hours (6am – 8 pm) for the westbound direction of the study arterial. Each scenario was simulated 12 times to generate a total of 972 data sets.

The two outlier detection algorithms were applied to all 972 sets of data. For each application *RTTI* was computed for each algorithm.

4.4.5 Comparison Results

The performance of the two outlier detection algorithms (Method 1 and Method 2) were evaluated by calibrating a multiple linear regression model in which the dependent variable was *RTTI* and the independent variables were binary variables associated with the different levels and first order interaction terms of the four scenario factors, P_e, P_T, P_m, and P_B. The regression results are provided in Table 4-5 for both Method 1 and Method 2. The Student-t statistic values marked with an asterisk indicate coefficients that are statistically significant at the 95 percent confident level.

The inclusion of the first order interaction terms makes it somewhat difficult to directly interpret the impact of the independent variables on the performance of the outlier algorithm. Consequently, the model (only significant terms) was used to estimate *RTTI* for all 81 scenarios and these results are illustrated in Figure 4-20.

A number of observations can be made on the basis of these results. For the results from Method 1 we observe:

[1] According to the recent literature, the Bluetooth penetration rate varies between 2 to10% depending on the roadway location and characteristics.

- Higher Bluetooth penetration rates result in better performance for all the cases. This is also evident from the negative sign of the coefficients for the 1% and 5% Bluetooth penetration variables.
- The method performs better when the proportion of en-route stops is higher. This is also demonstrated by the negative sign for 5% and 15% en-route stop variable. Travel times for vehicles that make an en-route stop tend to be significantly longer than the true travel time and therefore these outliers are easier to identify.
- The fraction of travel time observations from buses in the traffic stream does not appear to have a significant impact on the algorithm performance.
- The presence of multiple travel time measurements for the same vehicle does not have a significant impact on the algorithm performance.
- The amount of variation in the data explained by the regression model is rather low (R^2 = 0.47) but is not unreasonable for this kind of modeling.

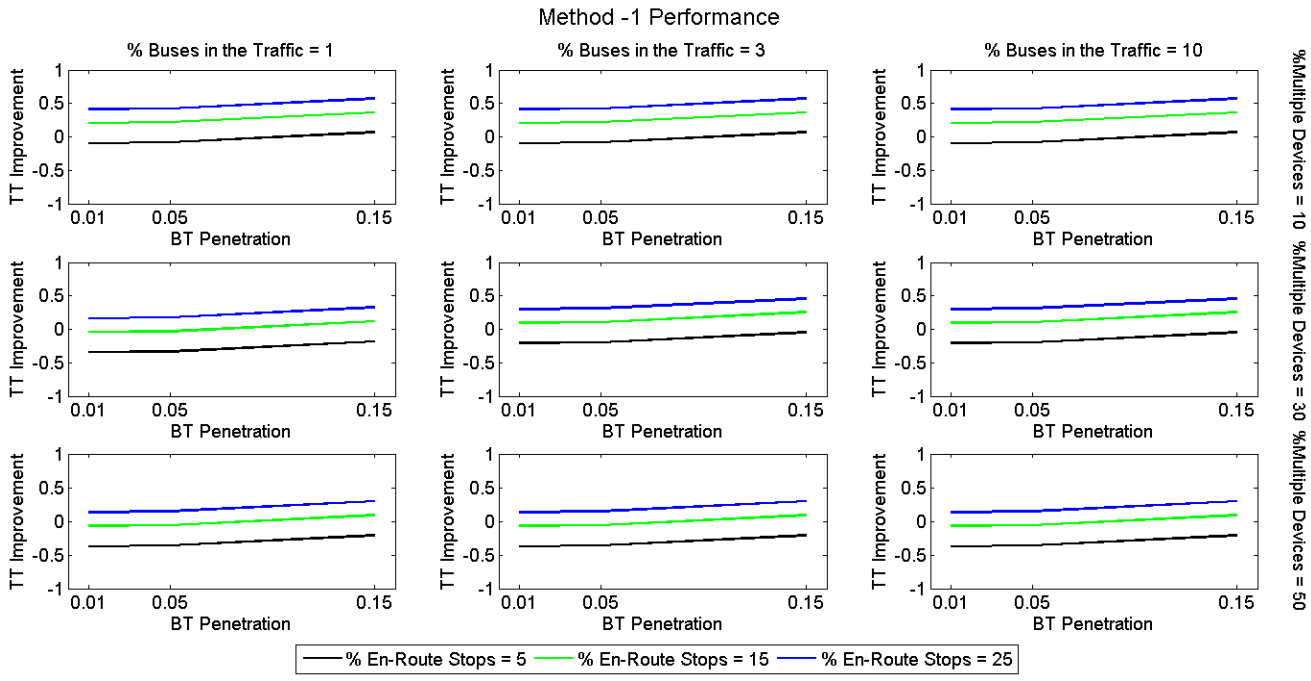
For the results from Method 2 we observe:

- Method 2 appears to achieve maximum performance at a lower level of Bluetooth penetration factor because increasing the penetration rate from 5% to 15% does not provide a significant improvement.
- Similar to Method 1, Method 2 performs better when the proportion of en-route stops is higher and the presence of transit vehicle travel times does not have a significant impact on the method performance.
- The presence of multiple travel time observations from the same vehicle has a statistically significant impact. The smaller proportion of multiple observations, the better the algorithm performance.
- The adjusted R^2 value for Method 2 is relatively high suggesting that the model explains a large proportion of the variation in the observed data.

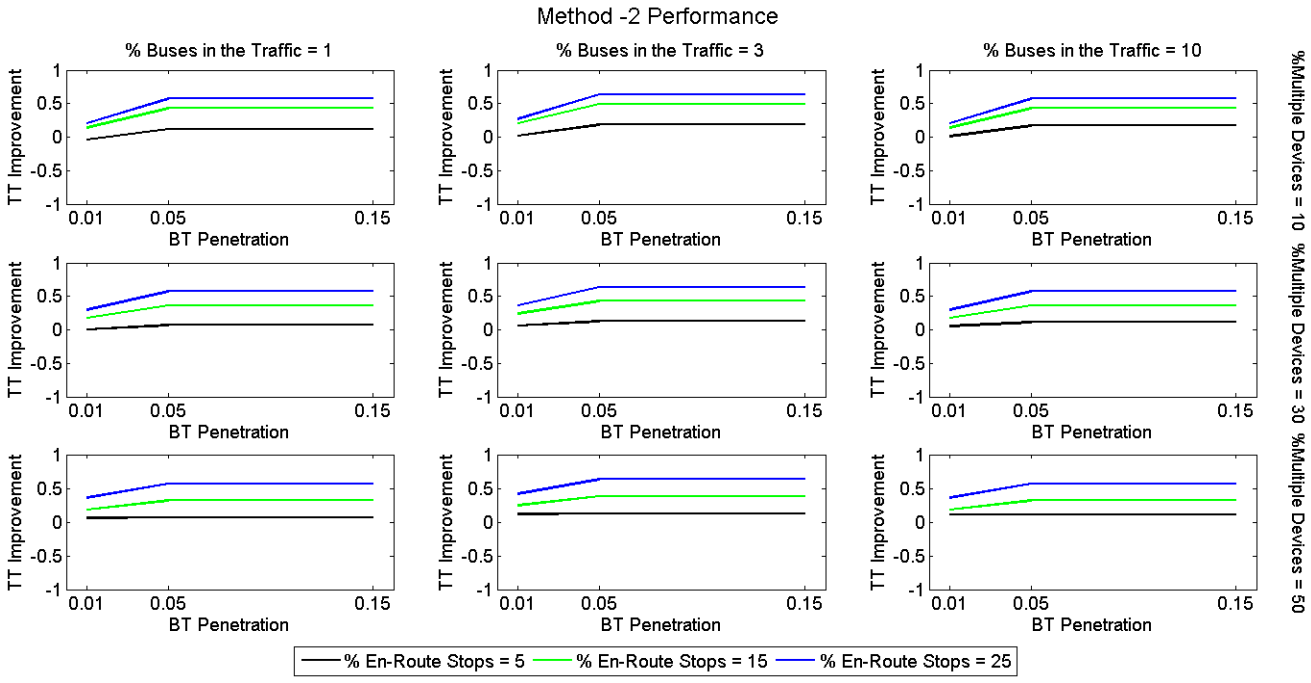
Table 4-5: Linear Regression Model Showing Significance of Factor Levels (Method 1 vs. Method 2)

Variable Name	Description	Method 1		Method 2	
		Coefficient	t-stat	Coefficient	t-stat
Constant		0.293	5.90*	0.570	25.39*
β_{BT1}	BT Penetration = 1%	-0.159	-2.83*	-0.213	-8.40*
β_{BT2}	BT Penetration = 5%	-0.148	-2.64*	-0.041	-1.62
β_{EnR1}	En-Route Stops = 5%	-0.507	-9.05*	-0.456	-18.02*
β_{EnR2}	En-Route Stops = 15%	-0.206	-3.67*	-0.254	-10.02*
β_{Bus1}	Buses in Traffic = 1%	-0.022	-0.39	0.038	1.48
β_{Bus2}	Buses in Traffic = 3%	0.018	0.31	0.064	2.52*
β_{MlD1}	Multiple Devices = 10%	0.279	4.98*	0.003	0.12
β_{MlD2}	Multiple Devices = 30%	0.167	2.97*	0.009	0.37
$\beta_{BT1,EnR1}$	BT = 1% & En-Route = 5%	-0.031	-0.60	0.206	8.79*
$\beta_{BT1,EnR2}$	BT = 1% & En-Route = 15%	-0.006	-0.11	0.079	3.38*
$\beta_{BT2,EnR1}$	BT = 5% & En-Route = 5%	0.009	0.17	0.041	1.75
$\beta_{BT2,EnR2}$	BT = 5% & En-Route = 15%	0.015	0.29	0.032	1.34
$\beta_{BT1,Bus1}$	BT = 1% & Bus = 1%	0.097	1.86	-0.026	-1.10
$\beta_{BT1,Bus2}$	BT = 1% & Bus = 3%	0.026	0.50	-0.025	-1.07
$\beta_{BT2,Bus1}$	BT = 5% & Bus = 1%	0.055	1.06	-0.035	-1.51
$\beta_{BT2,Bus2}$	BT = 5% & Bus = 3%	-0.024	-0.46	-0.010	-0.43
$\beta_{BT1,MlD1}$	BT = 1% & Mltp-Devices = 10%	-0.009	-0.18	-0.154	-6.55*
$\beta_{BT1,MlD2}$	BT = 1% & Mlp-Devices = 30%	0.003	0.06	-0.057	-2.41*
$\beta_{BT2,MlD1}$	BT = 5% & Mlp-Devices = 10%	-0.022	-0.42	0.001	0.06
$\beta_{BT2,MlD2}$	BT = 5% & Mlp-Devices = 30%	0.047	0.91	-0.017	-0.70
$\beta_{EnR1,Bus1}$	En-Route = 5% & Bus = 1%	-0.039	-0.75	-0.047	-2.00*
$\beta_{EnR1,Bus2}$	En-Route = 5% & Bus = 3%	-0.079	-1.53	-0.051	-2.17*
$\beta_{EnR2,Bus1}$	En-Route = 15% & Bus = 1%	-0.006	-0.12	0.014	0.60
$\beta_{EnR2,Bus2}$	En-Route = 15% & Bus = 3%	-0.037	-0.72	-0.014	-0.58
$\beta_{EnR1,MlD1}$	En-Route = 5% & Mltp-Devices = 10%	-0.014	-0.28	0.054	2.32*
$\beta_{EnR1,MlD2}$	En-Route = 5% & Mlp-Devices = 30%	-0.021	-0.40	-0.003	-0.12
$\beta_{EnR2,MlD1}$	En-Route = 15% & Mlp-Devices = 10%	-0.066	-1.26	0.113	4.81*
$\beta_{EnR2,MlD2}$	En-Route = 15% & Mlp-Devices = 30%	-0.072	-1.38	0.048	2.06*
$\beta_{Bus1,MlD1}$	Bus = 1% & Mltp-Devices = 10%	-0.047	-0.90	0.026	1.09
$\beta_{Bus1,MlD2}$	Bus = 1% & Mlp-Devices = 30%	-0.136	-2.62*	0.022	0.95
$\beta_{Bus2,MlD1}$	Bus = 3% & Mlp-Devices = 10%	-0.029	-0.56	0.013	0.57
$\beta_{Bus2,MlD2}$	Bus = 3% & Mlp-Devices = 30%	-0.097	-1.87	-0.002	-0.08
No. of Observations = 972		Adjusted R^2 = 0.47		Adjusted R^2 = 0.71	

* Significant at 95% confidence level

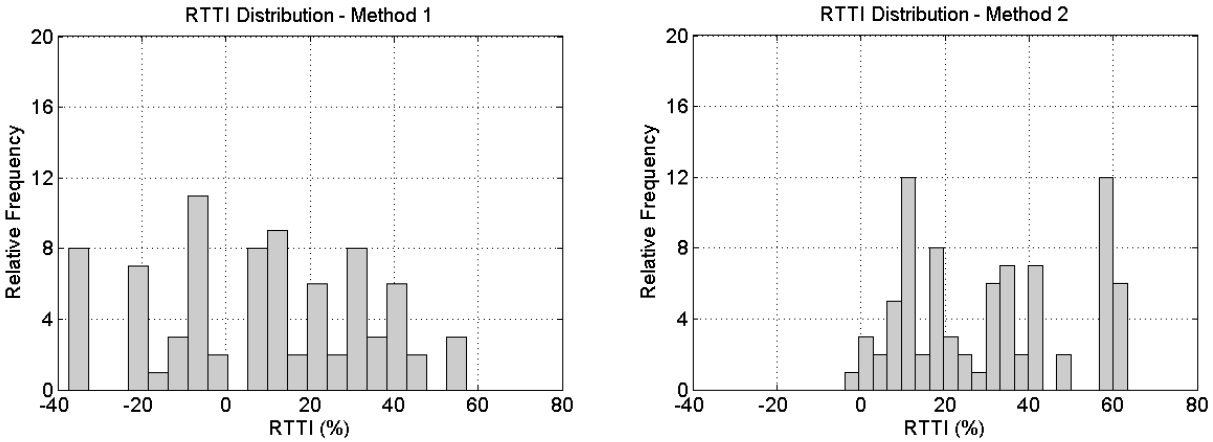


(a) Method 1



(b) Method 2

Figure 4-20: The performance of the investigated methods under different factor levels



(a) Method 1 (b) Method 2

Figure 4-21: The relative travel time improvement distributions

Figure 4-21 shows the distribution of *RTTI* for both of the outlier algorithms. Both distributions are based on the regression model estimates for the 81 factor combinations. *RTTI* index for Method 2 ranges from just less than zero to approximately 60% indicating that for almost all conditions examined, the application of Method 2 outlier detection algorithm provides benefits. Method 1 *RTTI* ranges from -40% to approximately 60% indicating that in some cases the use of outlier detection algorithm Method 1 provides results that are much worse than not using any detection algorithm at all. In fact, the likelihood that using Method 1 results in worse results than not applying any outlier detection algorithm at all is rather high – approximately 30%. We can also observe that for some conditions Method 1 provides benefits that are almost as large as those provided by Method 2. From these results we can conclude that Method 2 is clearly superior to Method 1 as the maximum improvement it provides is similar to that provided by Method 1 without a high likelihood of providing worse results than not applying an algorithm at all.

4.5 Chapter Summary

4.5.1 Measurement Error in Bluetooth Travel Times

In this research we have developed an approach for evaluating the magnitude of detection time and travel time measurement errors for Bluetooth acquired travel times. The approach is based on the characteristics of the Bluetooth communication protocol and vehicle trajectory

characteristics which are dependent on the location of the Bluetooth detectors, as well as the traffic volumes, signal operations, and quality of progression on the arterial.

The research described in this chapter has provided 3 important outcomes:

- We have proposed an objective method by which the measurement errors (detection time and travel time) can be modeled.
- We have demonstrated that the mean travel time error is sufficiently close to zero across all traffic conditions examined, that it can be ignored for practical purposes.
- We have developed a multivariate regression model that can be applied to assist in planning the deployment of Bluetooth detectors and in assessing the magnitude of the measurement error affecting Bluetooth acquired travel times.

4.5.2 Detecting Outliers in Bluetooth Data in Real Time

The literature review conducted as part of this study has revealed that very few algorithms have been proposed that: (i) can be applied in real time; (ii) are suitable for application to signalized arterials; and (iii) can be applied to Bluetooth data.

The most promising model was proposed by Dion and Rakha (2006). However, when applied in real-time, this algorithm is susceptible to large errors when the number of travel time observations in a time period is relatively small and the variation in the travel times is relatively large. These conditions tend to occur relatively frequently in Bluetooth travel time data acquired on signalized arterials.

The proposed outlier detection algorithm is formulated to address this problem. The algorithm is a pro-active adaptive filtering technique that fuses the quasi-predictions coming from a k-NN model with the latest available travel time estimates to construct an appropriate validity window.

The superiority of the proposed algorithm over the Benchmark algorithm has been demonstrated using field data in which a variety of travel time patterns are included.

4.5.3 Evaluating the Performance of Travel Time Filtering Algorithms using Simulation

In this study we have proposed a simulation framework to evaluate the performance of outlier detection algorithms under different conditions. The strength of this approach is that the

performance of the algorithm can be objectively quantified because the true travel times are known.

As part of this framework, we have proposed a new measure (*RTTI*) by which to quantify the performance of outlier detection algorithms. *RTTI* is a more robust measure of performance than percent outliers detected or percent wrong detections because it measures the impact of removing the outliers in terms of the relative improvement in the travel time estimate.

The proposed framework was used to carry out a structured evaluation of an existing pair of outlier detection algorithms. The results of this evaluation provide insights to the strengths and limitations of these models, and provided clear evidence of how the models perform over a wide range of conditions. On this basis it is possible to conclude which model is superior and to know the range of improvements that the algorithm is likely to provide. Lastly, the analysis of the performance of a given outlier detection algorithm can be used to improve the algorithm with the goal to achieve a distribution of *RTTI* that has low standard deviation and a large positive mean (implying the algorithm consistently provides large improvements for a wide range of conditions).

Chapter 5

Travel Time Prediction

In this chapter we focus on the real-time prediction of near future travel times on signalized arterials using the Bluetooth estimated travel times obtained from the outlier detection algorithm proposed in Chapter 4. In this chapter we examine two different prediction approaches, namely: (1) Non-Parametric method and (2) Parametric method.

The proposed Non-Parametric method uses the k-nearest neighborhood (k-NN) technique to identify historical data from which an understanding of the near-future traffic patterns can be extracted. On the other hand, the proposed Parametric method models the prevailing roadway traffic condition and its progression pattern using a combination of Support Vector Machine (SVM) classification technique and Markov Chain stochastic process.

More details on the justification, calibration, and performance evaluation of the Non-Parametric and Parametric methods are provided in the following sections.

5.1 Modeling Travel Time as a Pattern Recognition Problem (Non-Parametric Approach)[1]

The proposed non-parametric travel time prediction method is based on the k-nearest neighborhood (k-NN) technique. We selected this approach for the following reasons:

- Using only Bluetooth data for the purpose of prediction, we are restricted to use one of the data driven approaches.
- The high variability of travel time on arterials cannot be exclusively captured through time series models.
- The availability of a rich historical dataset provides us with the possibility of choosing one of the pattern recognition models.
- Unlike several other pattern recognition algorithms, the k-NN technique has a solid theoretical foundation, is relatively easy to implement, and can be readily explained/interpreted (unlike some “black box methods”).

[1] The content of this section is accepted for publication in the following journal paper:

- Salek Moghaddam, S. and B. Hellinga, "Real-Time Prediction of Arterial Roadway Travel Times Using Data Collected by Bluetooth Detectors", 2014 (In Press), Transportation Research Record.

- Due to its flexible non-parametric structure, k-NN method not only is able to dynamically capture the underlying trend of travel time but also is adequately transferable.

5.1.1 Model Structure

As discussed in Chapter 2, the goal of using the k-NN method is to find a set of historical traffic conditions similar to the current one and then derive the target variable (e.g. travel time for the next time interval) from the set of selected historical data. The k-NN method has four main elements which are discussed in the following paragraphs:

1. Feature vector
2. Distance metric
3. Number of nearest neighbors to select
4. Local estimation method.

The *feature vector* defines the set of characteristics (variables) that are used for selecting the k most similar historical instances. In this study, we considered the following four characteristics:

1. Estimated mean travel time during recent time intervals (temporal scale) for the investigated segment and its upstream segment (spatial scale) ($[t]^{(i,k)}_{1\times((n_k+1)(n_i+1))}$ - defined in Equation 5-1);
2. Counted trips during recent time intervals for the investigated segment and its upstream segment ($[cnt]^{(i,k)}_{1\times((n_k+1)\cdot(n_i+1))}$ - defined in Equation 5-1);
3. Number of detections during recent time intervals at the upstream and downstream detectors of the investigated segment ($[DT]^{(i,k)}_{1\times(2\times(n_k+1))}$ - defined in Equation 5-1);
4. The summation of hits during recent time intervals at the upstream and downstream detectors of the investigated segment ($[SH]^{(i,k)}_{1\times(2\times(n_k+1))}$ - defined in Equation 5-1)

Given that we are seeking to find historical data that exhibit the same travel time pattern (trend) as has been observed to the present time on the current data, we consider data from a number of recent time intervals rather than just the current interval. We also consider spatial interactions.

Based on the above explanations, the examined feature vector has been formulated through the following equation:

$$\left[FV\right]^{(i,k)} = \left[[t]^{(i,k)}_{1\times((n_k+1)\cdot(n_i+1))} \quad [count]^{(i,k)}_{1\times((n_k+1)\cdot(n_i+1))} \quad [DT]^{(i,k)}_{1\times(2\times(n_k+1))} \quad [SH]^{(i,k)}_{1\times(2\times(n_k+1))} \right]_{1\times((n_k+1)\cdot(2n_i+6))} \quad \textbf{(5-1)}$$

Where, $\left[FV\right]^{(i,k)}$ is the constructed feature vector for segment *i* at the end of time interval *k*, n_k is the number of prior time intervals, and n_i is the number of upstream segments.

The inclusion of variables with different dimensions may result in bias towards the variables with larger magnitudes (i.e. scale effect) and consequently, all the variables in the feature vector have been normalized based on the mean and standard deviation calculated using a random sample extracted from the historical dataset:

$$[\overset{SD}{FV}]^{(i,k)} = \frac{(\left[FV\right]^{(i,k)} - E_{(FV^H)})}{Var_{(FV^H)}} \quad \textbf{(5-2)}$$

Where, $[\overset{SD}{FV}]^{(i,k)}$ is the standardized feature vector for segment *i* at the end of time interval *k*, $E_{(FV^H)}$ is the expected value of the features over the historical dataset, and $Var_{(FV^H)}$ is the variance of the features over the historical dataset.

The k-NN requires the use of a *distance metric* to measure the degree of similarity between the feature vector of the current data and the historical data. We use the Mahalanobis distance (Equation 5-3) which considers both the variance and covariance of the feature vector variables. The consideration of the variance reduces the effect of those variables with high variations while the common effects of correlated variables can be excluded by considering the covariance of the feature vector variables:

$$MD_{(i,k)} = ((\overbrace{\left[\overset{SD}{FV}\right]^{(i,k)}_h - \left[\overset{SD}{FV}\right]^{(i,k)}_c}^{\Delta FV^h}) \cdot \Sigma^{-1}_{(FV^H)} \cdot (\left[\overset{SD}{FV}\right]^{(i,k)}_h - \left[\overset{SD}{FV}\right]^{(i,k)}_c)^T)^{\frac{1}{2}} \quad \textbf{(5-3)}$$

Where, $MD_{(i,k)}$ is the Mahalanobis distance between historical record *h* and current prediction instance *c*, $[\overset{SD}{FV}]^{(i,k)}$ is the feature matrix, and $\sum_{(FV^H)}$ is the variance – covariance matrix of historical feature variables.

After calculation of distance metric for each prediction instance in the selected historical database, the k instances with the lowest distance are chosen as the set of *nearest neighbors*.

The weighted arithmetic mean of the target travel times is considered as the *local estimation method*:

$$\hat{t}_{(i,k+1)} = \frac{\sum_{h \in NN} n^{h}_{(i,k+1)} t^{h}_{(i,k+1)}}{n_{NN}} \quad \textbf{(5-4)}$$

Where, $\hat{t}_{(i,k+1)}$ is the predicted travel time on segment *i* during time interval *k+1* (next future time interval), $t^{h}_{(i,k+1)}$ is the observed travel time on segment *i* during time interval *k+1* for historical record *h*, $n^{h}_{(i,k+1)}$ is the number of individual travel times which have been utilized to calculate $t^{h}_{(i,k+1)}$, *NN* is the set of nearest neighbors, and n_{NN} is the number of nearest neighbors.

The performance of the k-NN model can be significantly affected by both the variables included in the feature vector and the model parameters (e.g., n_{NN}). Therefore, an optimization process is required to first select the variables to include in the features vector and second to calibrate the best parameter values.

5.1.2 Model Calibration

5.1.2.1 Experimental Design

Based on the defined model structure, the following eight terms have been considered as the model parameters (note that travel time data were always included and therefore this parameter does not appear as one of the variables in the list). The optimum value of each parameter has been investigated through an optimization process.

1. Number of prior days (prior to the test date) that has been considered as the historical dataset (size of historical dataset) – 4 levels: 15, 30,45, and 60 days
2. Number of upstream segments (spatial scale) – 2 levels: 0 and 1
3. Number of prior time intervals (temporal scale) – 5 levels: 0, 1, 2, 3, and 4
4. Inclusion of the *counted trip* data – 2 levels: included and not-included

5. Inclusion of the *number of detections* data – 2 levels: included and not-included
6. Inclusion of the *summation of hits* data – 2 levels: included and not-included
7. Number of nearest neighbors – 4 levels: 1, 10, 20, and 50
8. The length of the search window around the current time - 9 levels: 0 min, 15 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr, 16 hr, and 24 hr

An experimental design method was adopted for determining the optimal value of each parameter. The experimental design method provides benefits over theoretical and heuristic optimization methods, namely: shorter analysis time; possibility of investigating both main effects and interactions; and possibility of investigating the statistical significance of the examined factors.

The length of the search window (factor #8) is assumed not to have significant interaction with other parameters, so the optimum value of this factor is investigated separately through a sensitivity analysis. The result of this analysis is presented in Figure 5-1, where the accuracy of the k-NN travel time prediction method (i.e., the daily MARE[1]) is examined against different values of the search window. These accuracies are calculated by averaging the daily MAREs obtained from the application of the model over 5 random days and 4 arterial segments. According to this diagram, the optimum value of this parameter happens at 4 hours search window around the prediction time.

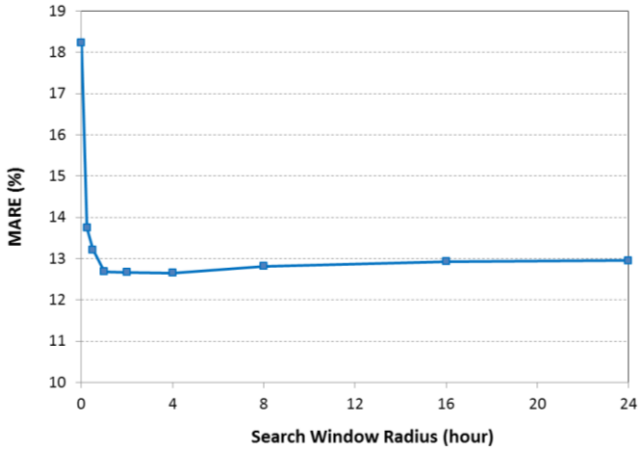


Figure 5-1: The effect of search window length on the accuracy of the kNN prediction method (Investigated over five random days)

[1] Mean absolute relative error

The other seven factors were investigated through a full factorial type of analysis. All the analysis is performed using the developed real-time travel time analysis framework. For the purpose of optimization the model parameters, the outlier detection algorithm proposed by Dion and Rakha (2006) is used because the proposed outlier detection algorithm is itself sensitive to the investigated parameters.

All combinations of the discussed parameters (7 parameters) have been examined except the simultaneous inclusion of the number of detections and sum-hits (due to their high correlations). In this way, 1120 scenarios have been tested on 4 arterials segments over 10 randomly selected test days (44,800 daily k-NN runs). The random selection of the test days was only intervened once to verify if the Dion and Rakha's outlier filtering algorithm performs well on the test days. The results of 9 days have been considered in the optimization process and the results from the one remaining day are utilized for cross-validation.

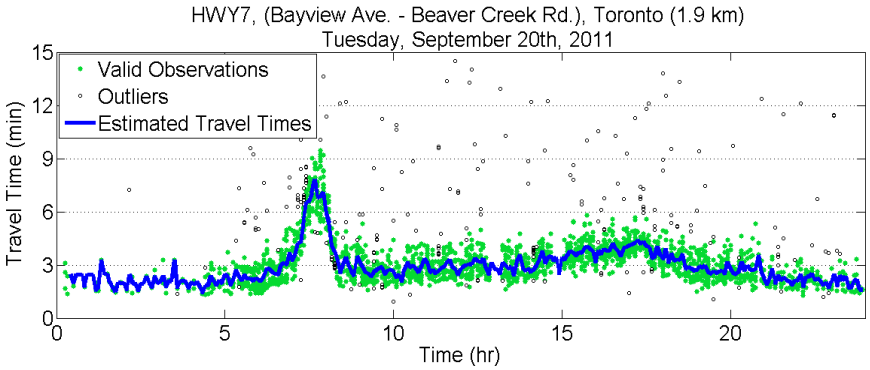
Figure 5-2 illustrates the data for four of the 10 days used to optimize the model parameters. These days contain a variety of different travel time patterns (distinct AM and PM peaks, recurrent and non-recurrent congestion, and extended PM peak congestion).

A linear regression model is calibrated to the results for investigating the main effects and interactions of the parameters. Factor levels are defined as different binary variables. To exclude the variation associated with different test days and different road segments, these two factors are considered as blocking variables.

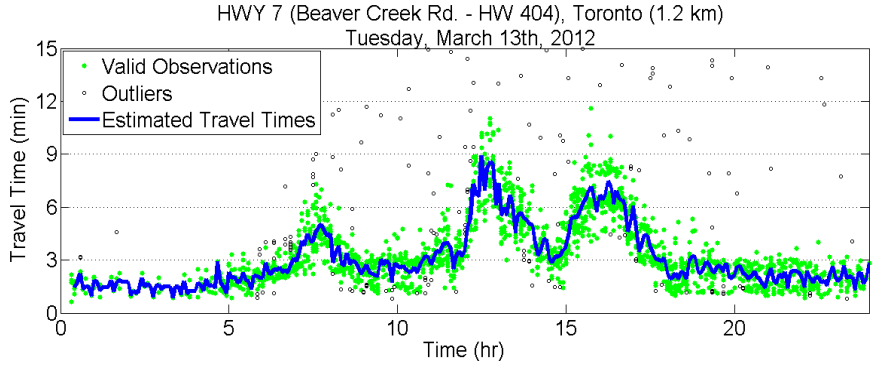
The daily mean absolute relative error (MARE) of the predictions is considered as the response variable (Equation 5-5). The calibrated coefficients and the confidence intervals of the statistically significant terms in the regression are provided in Table 5-1. Interpretation of these results is provided in the next section.

$$MARE_{(i)} = 100 \times \frac{1}{N} \sum_{k} \frac{\left| \hat{t}_{(i,k)} - t_{(i,k)} \right|}{t_{(i,k)}} \qquad \textbf{(5-5)}$$

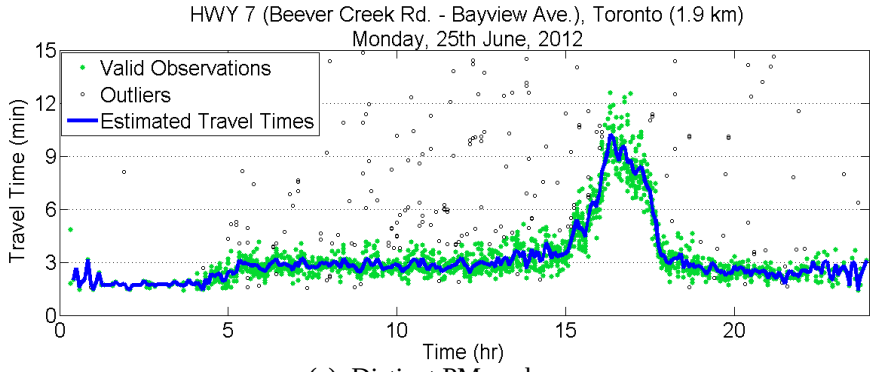
Where, $MARE_{(i)}$ is the mean absolute relative error for segment i, $\hat{t}_{(i,k)}$ is the average predicted travel time over time interval k on segment i, $t_{(i,k)}$ is the average measured travel time over time interval k on segment i, and N is the number of prediction instances.



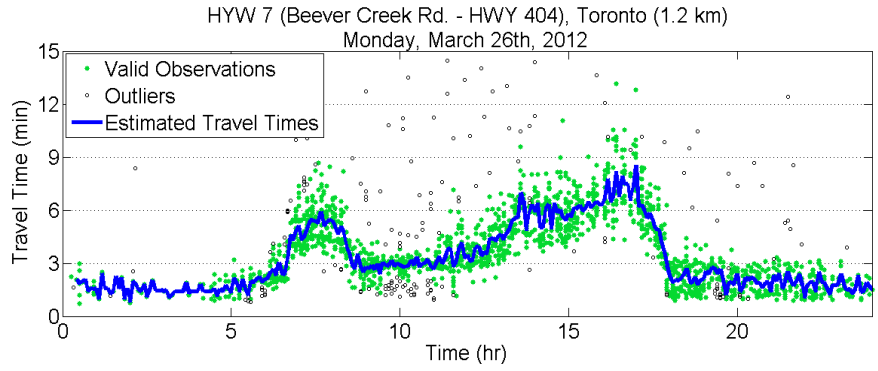
(a). Distinct AM peak



(b). Recurrent & non-recurrent congestion



(c). Distinct PM peak



(d). Distinct AM peak and extended PM peak

Figure 5-2: Sample of field data used for model calibration (HWY 7, Toronto)

Table 5-1: Linear Regression Model Showing the Calibration Results of the Proposed k-NN Prediction Model – Response variable: MARE (%)

Variable Name	Description	Coefficient	Confidence Interval*
Constant		19.18	(19.09,19.27)
β_{DAY1}	Day # 1	-0.31	(-0.37,-0.26)
β_{DAY2}	Day # 2	-0.13	(-0.18,-0.07)
β_{DAY3}	Day # 3	-1.42	(-1.48,-1.36)
β_{DAY4}	Day # 4	-1.08	(-1.14,-1.03)
β_{DAY5}	Day # 5	2.08	(2.02,2.13)
β_{DAY6}	Day # 6	0.87	(0.81,0.92)
β_{DAY7}	Day # 7	-0.65	(-0.71,-0.6)
β_{DAY8}	Day # 8	1.76	(1.71,1.82)
$\beta_{SEGMENT1}$	Segment # 1	-1.45	(-1.49,-1.41)
$\beta_{SEGMENT2}$	Segment # 2	-3.51	(-3.54,-3.47)
$\beta_{SEGMENT3}$	Segment # 3	1.93	(1.89,1.96)
β_{PRIOR_DAYS3}	# Prior days = 45	0.17	(0.1,0.24)
β_{PRIOR_DAYS4}	# Prior days = 60	0.22	(0.15,0.29)
$\beta_{PRIOR\_INTERVAL2}$	# Prior intervals = 1	-0.94	(-1.02,-0.86)
$\beta_{PRIOR\_INTERVAL3}$	# Prior intervals = 2	-1.10	(-1.17,-1.02)
$\beta_{PRIOR\_INTERVAL4}$	# Prior intervals = 3	-1.14	(-1.21,-1.06)
$\beta_{PRIOR\_INTERVAL5}$	# Prior intervals = 4	-1.03	(-1.12,-0.95)
β_{COUNT2}	Count data = Included	0.45	(0.36,0.54)
$\beta_{DETECTION2}$	Detection data = Included	-0.88	(-0.96,-0.8)
β_{SUM_HITS2}	Sum-Hits data = Included	-0.89	(-0.96,-0.81)
β_{NN2}	# Neighbors = 10	-4.11	(-4.18,-4.04)
β_{NN3}	# Neighbors = 20	-4.40	(-4.47,-4.33)
β_{NN4}	# Neighbors = 50	-4.54	(-4.62,-4.46)
$\beta_{PRIOR\_DAYS4,PRIOR\_INTERVAL5}$	# Prior days = 60 & # Prior intervals = 4	-0.09	(-0.16,-0.01)
$\beta_{UPSTREAM\_SEGMENT1,COUNT2}$	Upstream Segment = 1 & Count data = Included	0.12	(0.08,0.15)
$\beta_{PRIOR\_INTERVAL2,COUNT2}$	# Prior intervals = 1 & Count data = Included	-0.25	(-0.34,-0.16)
$\beta_{PRIOR\_INTERVAL3,COUNT2}$	# Prior intervals = 2 & Count data = Included	-0.25	(-0.34,-0.16)
$\beta_{PRIOR\_INTERVAL4,COUNT2}$	# Prior intervals = 3 & Count data = Included	-0.26	(-0.35,-0.17)
$\beta_{PRIOR\_INTERVAL5,COUNT2}$	# Prior intervals = 4 & Count data = Included	-0.25	(-0.34,-0.16)
$\beta_{PRIOR\_INTERVAL2,DETECTION2}$	# Prior intervals = 1 & Detection data = Included	0.33	(0.24,0.42)
$\beta_{PRIOR\_INTERVAL3,DETECTION2}$	# Prior intervals = 2 & Detection data = Included	0.42	(0.33,0.51)
$\beta_{PRIOR\_INTERVAL4,DETECTION2}$	# Prior intervals = 3 & Detection data = Included	0.44	(0.35,0.53)
$\beta_{PRIOR\_INTERVAL5,DETECTION2}$	# Prior intervals = 4 & Detection data = Included	0.47	(0.38,0.56)
$\beta_{PRIOR\_INTERVAL2,SUM\_HITS2}$	# Prior intervals = 1 & Sum-Hits data = Included	0.32	(0.24,0.41)
$\beta_{PRIOR\_INTERVAL3,SUM\_HITS2}$	# Prior intervals = 2 & Sum-Hits data = Included	0.34	(0.25,0.43)
$\beta_{PRIOR\_INTERVAL4,SUM\_HITS2}$	# Prior intervals = 3 & Sum-Hits data = Included	0.41	(0.33,0.5)

Table 5-1: Linear Regression Model Showing the Calibration Results of the Proposed k-NN Prediction Model – Response variable: MARE (Cont.)

Variable Name	Description	Coefficient	Confidence Interval*
$\beta_{PRIOR\_INTERVAL5,SUM\_HITS2}$	# Prior intervals = 4 & Sum-Hits data = Included	0.46	(0.37,0.55)
$\beta_{COUNT2,DETECTION2}$	Count & Detection data = Both Included	0.18	(0.1,0.25)
β_{COUNT2,SUM_HITS2}	Count & Sum-Hits data = Both Included	0.18	(0.11,0.25)
$\beta_{COUNT2,DETECTION2,SUM\_HITS2}$	Count & Detection & Sum-Hits data = All Included	0.48	(0.41,0.55)
$\beta_{PRIOR\_DAYS3,NN2}$	# Prior days = 45 & # Neighbors = 10	-0.13	(-0.22,-0.04)
$\beta_{PRIOR\_DAYS3,NN3}$	# Prior days = 45 & # Neighbors = 20	-0.16	(-0.25,-0.06)
$\beta_{PRIOR\_DAYS3,NN4}$	# Prior days = 45 & # Neighbors = 50	-0.21	(-0.3,-0.12)
$\beta_{PRIOR\_DAYS4,NN2}$	# Prior days = 60 & # Neighbors = 10	-0.15	(-0.24,-0.05)
$\beta_{PRIOR\_DAYS4,NN3}$	# Prior days = 60 & # Neighbors = 20	-0.17	(-0.27,-0.08)
$\beta_{PRIOR\_DAYS4,NN4}$	# Prior days = 60 & # Neighbors = 50	-0.23	(-0.33,-0.14)
$\beta_{COUNT2,NN2}$	Count data = Included & # Neighbors = 10	-0.25	(-0.33,-0.17)
$\beta_{COUNT2,NN3}$	Count data = Included & # Neighbors = 20	-0.26	(-0.34,-0.18)
$\beta_{COUNT2,NN4}$	Count data = Included & # Neighbors = 50	-0.28	(-0.35,-0.2)
$\beta_{DETECTION2,NN4}$	Detection data = Included & # Neighbors = 50	0.08	(0.02,0.15)
$\beta_{SUM\_HITS2,NN4}$	Sum-Hits data = Included & # Neighbors = 50	0.12	(0.06,0.19)
No. of Observations = 40,320	Adjusted R^2 = 0.83		

** 95% Confidence level*

5.1.2.2 Calibration Results

A number of observations can be made on the basis of the results provided in Table 5-1:

- As expected, the inclusion of detection data (either number of detections or summation of hits) has decreased the prediction error (their coefficients are statistically significant).
- Increasing the number of neighbors (n_{NN}) from 1 to 50 decreases the prediction error, although the rate of this improvement is the highest in moving from 1 to 10 neighbors. This finding is graphically shown in Figure 5-3a.
- Contrary to expectation, the coefficient associated with including the upstream segment (spatial scale) was not statistically significant (although its first order interaction with count data was significant) and therefore including this attribute in the feature vector does not improve the model predictions. This may be due to the fact that two out of the four segments do not have any upstream segment.

Figure 5-3b illustrates the effect that number of prior time intervals has on the prediction error for a number of feature vector configurations. From this figure, we can make the following observations:

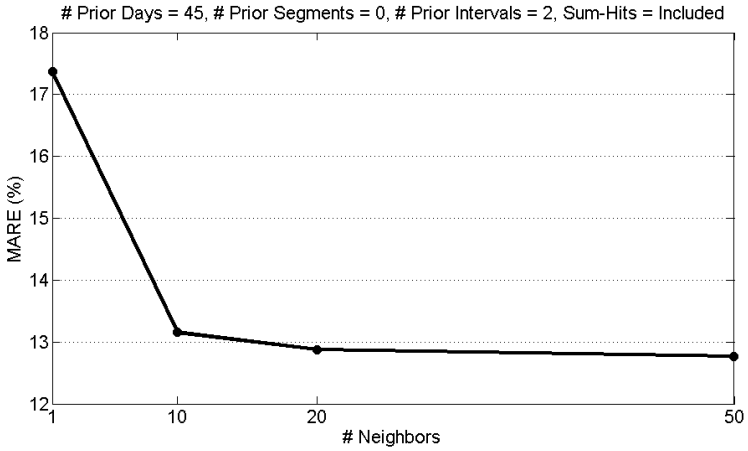
- Combining *detections*, or *sum hits* with travel time data improves the model performance. The best model performance occurs for the combination of travel time data with *sum hits*.
- For all of the feature vector variations, the inclusion of prior time intervals (temporal scale) improves model performance when considering one or two prior time intervals. Including more than two time intervals degrades model performance (possibly due to the curse of dimensionality).

On the basis of the results in Table 5-1 and Figure 5-3, the optimum values for the model parameters are selected as follows:

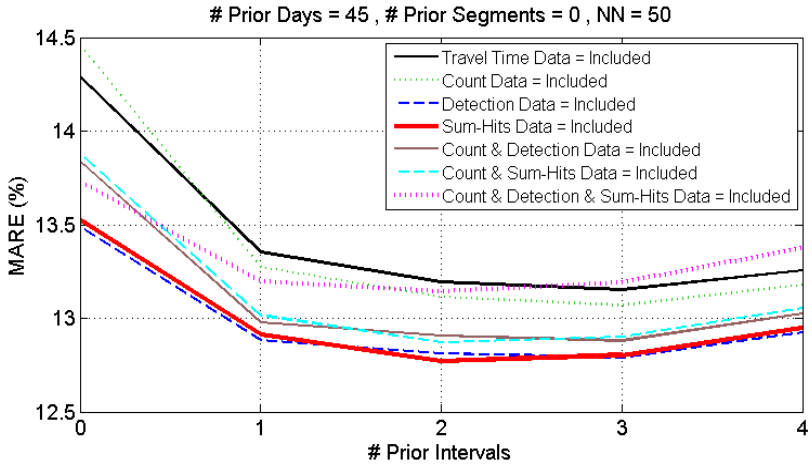
- Size of historical dataset: 45 prior days
- Number of upstream segments: zero
- Number of prior time intervals: 2 time intervals
- Number of nearest neighbors: 50 instances
- The length of the search window around the current time: 8 hours (4 hours before and 4 hours after than the current time)

The regression model coefficients provided in Table 5-1 provide an indication of the relative importance of the different parameters. However, assessing the relative importance using these coefficients is complicated by the presence of the interaction terms. To help illustrate the relative importance, the relative contribution of optimizing each model parameter is illustrated in Figure 5-4. These contributions are determined using the developed regression model and represent the condition of all days and segments; however to calculate the base prediction error, the segment blocking variable is set equal to the segment of Highway 7 between HWY 404 and Beaver Creek Rd. and the day blocking variable was set equal to Tuesday, September 20th, 2011. The column labeled “Initial Error” reflects the model performance using the set of initial parameter values. As indicated in the figure, the model provides prediction accuracy (MARE) of

approximately 19%. The column labeled "Optimized Error" shows that the model accuracy improved to a MARE of approximately 13% when the optimal parameter values are used.



(a) Prediction error vs. # of neighbors



(a) Prediction error vs. # of prior time intervals categorized by the inclusion of different attributes in the feature vector

Figure 5-3: Model calibration results

Figure 5-4 also illustrates the contribution that the individual parameters values have had on this improvement. Seventy-two percent of the improvement is attributed to optimizing the number of neighbors (n_{NN}) from 1 in the initial model to 50 in the optimized model. The smallest contribution to the improvement is associated with the number of prior historical days suggesting that having a large historical database (i.e. 45 days) versus a smaller database (i.e. 15 days) is not nearly as important as selecting 50 historical data series (neighbors) versus just one.

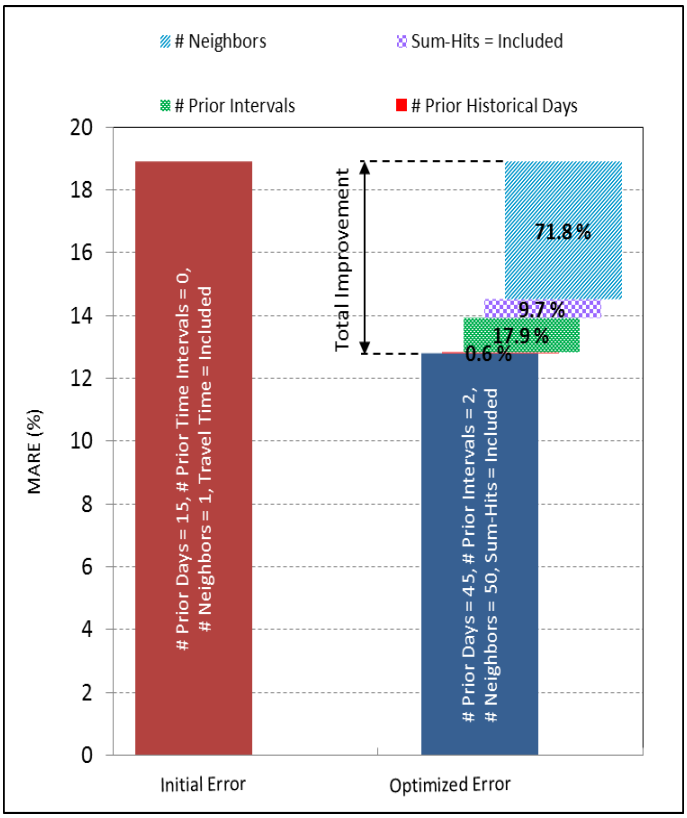


Figure 5-4: Relative contribution of optimized parameters on the performance of the k-NN model

The performance of the optimized model is also examined by analyzing model performance on the cross-validation day. Table 5-2 compares the performance of the optimized k-NN model with a benchmark method. The benchmark method simply considers the average travel time of the last available time interval as the prediction for the next time interval. The performance of both the optimized k-NN model and the Benchmark model is assessed with respect to the performance of all 1120 k-NN models considered within the parameter optimization process. For example, for the first segment, the Benchmark model provides a MARE of 13% and the optimized k-NN model provides a MARE of 10.6% representing an 18.5% reduction in the prediction error. The columns labeled "Relative Rank" indicate the proportion of the 1120 k-NN models that provided a smaller MARE for this particular segment and day. Almost 75% of the 1120 k-NN models examined in the parameter optimization process provided better performance than the Benchmark model. However, only 3% provided better performance than the optimized model.

Table 5-2: Cross validation results showing the relative performance of the optimized k-NN model (Cross-validation day: Monday, November 14th, 2011)

Segment	Benchmark Model		Optimized k-NN Model		Percent Improvement (MARE)**
	MARE (%)	Relative Rank* (%)	MARE (%)	Relative Rank* (%)	
Segment 1 HWY 7 (Bayview Ave. - Beaver Creek Rd)	13.0	74.7	10.6	3.2	18.5
Segment 2 HWY 7 (Beaver Creek Rd. - Bayview Ave.)	13.3	75.5	10.2	2.1	23.3
Segment 3 HWY 7 (Beaver Creek Rd. - HWY 404)	18.1	73.2	15.0	9.5	17.1
Segment 4 HWY 7 (HWY 404 - Beaver Creek Rd.)	15.4	75.4	12.8	10.4	16.9

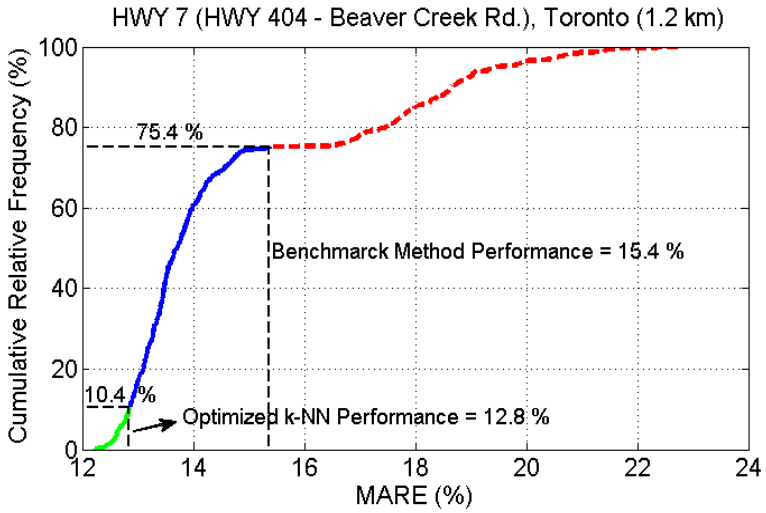
* *Relative performance of the model compared to the 1120 k-NN parameter combinations examined in the optimization process.*

** *Percent improvement of the optimized k-NN model over the benchmark model in terms of MARE.*

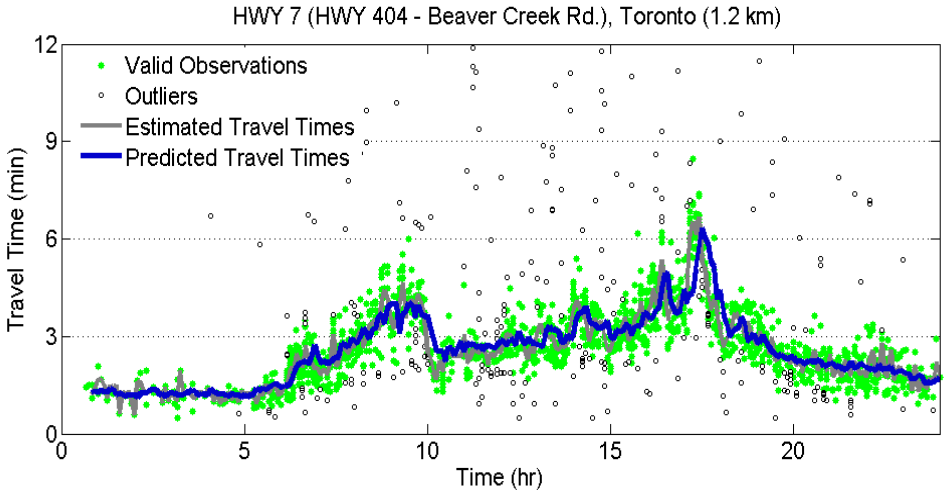
Similar results for segment 4 are provided in Figure 5-5a. Figure 5-5b further illustrates the model results for the validation day. The figure shows the predicted travel times as well as the individual observations and the true travel times. The outliers have been labeled by the proposed outlier detection algorithm and the predicted travel times have been obtained from the application of the optimized k-NN model.

The optimized model does not perform better than all the other k-NN models because (i) the optimization is based on minimizing error across 9 different days; and (ii) the cross-validation day was not part of the parameter optimization data set.

The results in Figure 5-5 and Table 5-2 suggest that (i) optimizing the model parameters has a large impact on performance; (ii) the proposed k-NN model is robust (performance remains relatively constant when applied to other days of data); and (iii) the proposed model performs significantly better than the Benchmark method.



(a). Relative performance ranking of the optimized k-NN model
(Applied on the validation day)



(b). Travel time prediction vs. time of day
(applying the optimized k-NN model on the validation day)

Figure 5-5: Sample cross-validation results – validation day: Monday, November 14th, 2011 (HWY 7 – HWY 404 - Beaver Creek Rd., Toronto (1.2km))

5.1.3 Model Evaluation

The results provided in the previous section are based on a single cross-validation day; consequently, following the procedure presented in Figure 5-6, the performance of the proposed non-parametric approach (k-NN model) has been evaluated further using 10 randomly selected validation days (data from the calibration were excluded from consideration in this validation). Table 5-3 presents the results of this validation in terms of MARE for the Benchmark method and the Proposed method.

Table 5-3 reports the mean of the MARE as well as the confidence interval. The mean and confidence interval are determined by treating the MARE from each of the 10 days as a single observation. When the upper confidence limit of the optimized model is smaller than the lower confidence limit of the benchmark method, we can confidently state that the Proposed model provides superior performance (in terms of a lower MARE) over the Benchmark model.

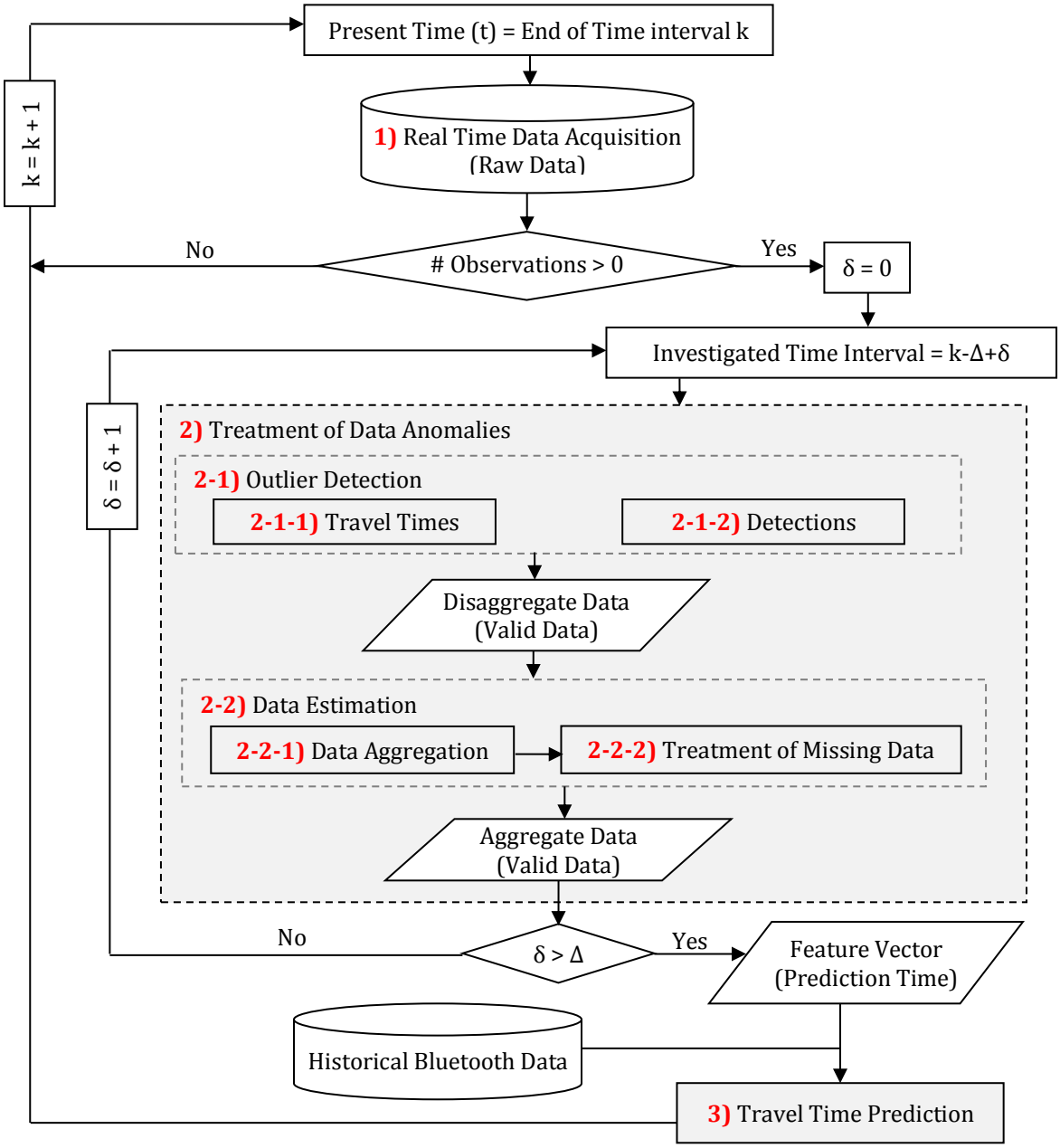


Figure 5-6: Proposed framework for travel time prediction on arterials (Non-Parametric Approach)

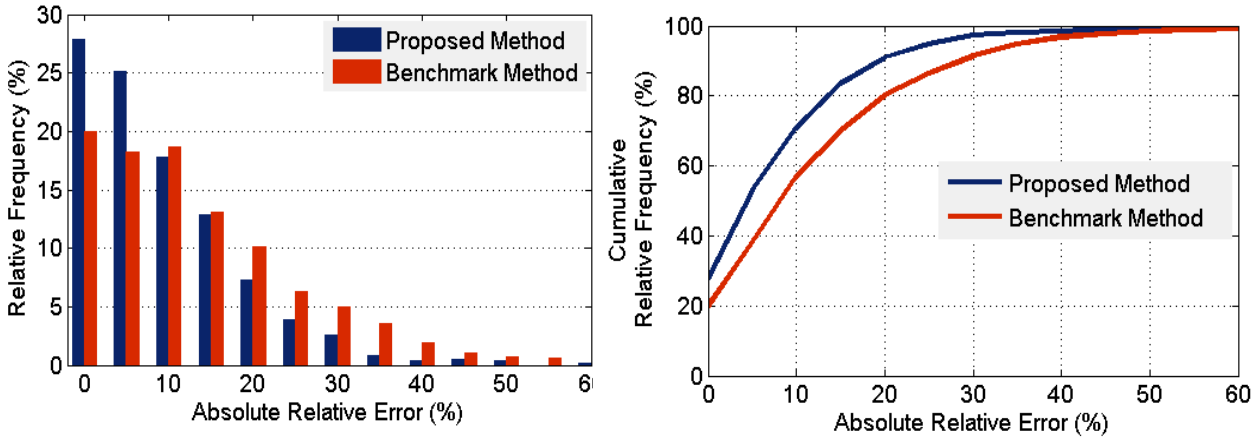
Table 5-3: Relative performance of the proposed method in terms of daily MARE (Investigated over 10 validation days)

Segment	Benchmark Method			Proposed Method			Absolute Improvement Over Benchmark (%)		Relative Improvement Over Benchmark (%)	
	Mean	90% Lower Bound	90% Upper Bound	Mean	90% Lower Bound	90% Upper Bound	Mean	90% Upper Bound	Mean	90% Upper Bound
Segment 1 HWY 7 (Bayview - Beaver Creek)	14.38	13.80	14.97	11.23	10.70	11.76	3.15	3.21	21.93	21.47
Segment 2 HWY 7 (Beaver Creek - Bayview)	11.82	10.82	12.82	8.90	7.99	9.80	2.92	3.02	24.72	23.56
Segment 3 HWY 7 (Beaver Creek - HWY 404)	16.98	16.27	17.68	14.03	13.30	14.76	2.95	2.92	17.35	16.52
Segment 4 HWY 7 (HWY 404 - Beaver Creek)	16.25	15.40	17.10	12.23	11.28	13.17	4.02	3.93	24.76	22.97

The performance of the Proposed and Benchmark methods has been further investigated through the empirical probability distribution and the empirical cumulative probability distribution of the 5 minute time interval prediction errors presented in Figure 5-7. The sample error distributions provided in this figure reflects the prediction accuracies for Segment 4 (Hwy 7 – HWY 404 - Beaver Creek Rd.).

For both of the distributions provided in this figure, the horizontal axis is the absolute relative error of the 5 minutes time interval predictions. The vertical axes in Figure 5-7a and Figure 5-7b are the relative frequency and the cumulative relative frequency, respectively. Based on this figure, the superior performance of the proposed k-NN method over the Benchmark can be verified. For example, according to Figure 5-7b, in 80% of the prediction instances (5 minute time intervals) the Proposed k-NN model resulted in accuracies (ARE[1]) less than 13% while this value for the Benchmark was 20%.

[1] Absolute relative error



(a). Relative Frequency vs. Absolute Relative Error

(b). Cumulative Relative Frequency vs. Absolute Relative Error

Figure 5-7: Prediction error distribution - Proposed k-NN model vs. Benchmark (Segment 4 (HWY 7 – HWY 404 - Beaver Creek Rd.) – Investigated over 10 validation days)

In a further investigation, the efficiency of the k-NN technique has been examined through Figure 5-8, which compares the prediction accuracies (MARE) against the k-NN similarity ranking (based on the distance metric). In this figure, the vertical axis is the prediction MARE and the horizontal axis is the order of 50 nearest neighbors that have been considered for the prediction of mean travel time. The red square indicates the prediction error for the case that the most similar 50 historical instances have been considered as the set of nearest neighbors (Proposed method). On the other hand, the blue triangles demonstrate the prediction errors associated with the higher orders of 50 neighbors being considered as the set of nearest neighbors.

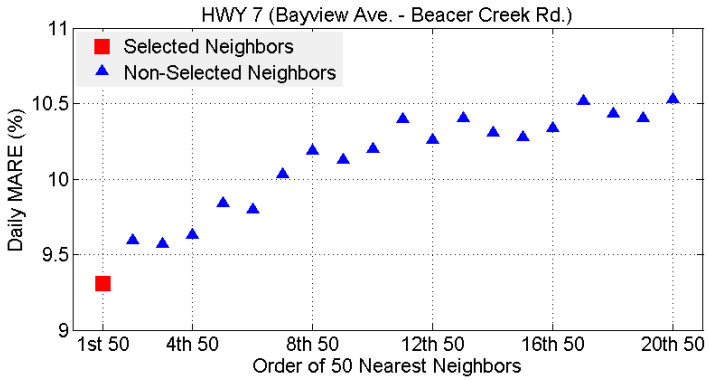
According to Figure 5-8, on all the investigated segments the prediction errors almost monotonically increased when using more distant (higher ordered) historical instances as the set of nearest neighbors. In other words, the prediction MARE and the calculated Mahalanobis distance metric were highly correlated (positive correlation). This finding confirms the overall efficiency of the implemented k-NN technique as an appropriate method for predicting near-future travel times on arterials.

Following the same kind of analysis, the efficiency of the k-NN method is further examined for different traffic conditions (i.e., free-flow, transient, and congestion). For the

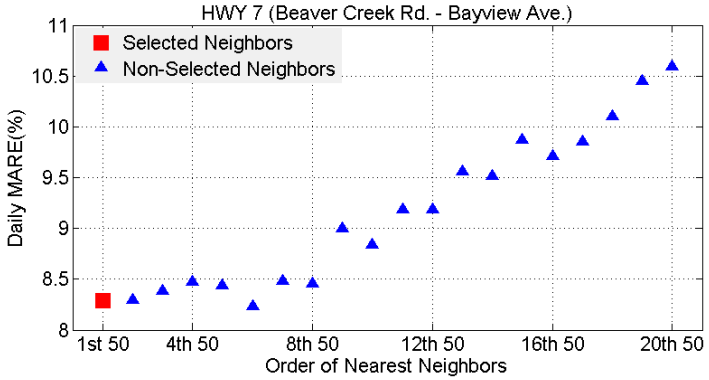
sample travel time trend presented in Figure 5-9a, three different time intervals belonging to different traffic conditions are identified. The k-NN similarity ranking diagrams associated with each of these traffic conditions are presented in Figure 5-9b to Figure 5-9d. These results confirm that the use of the k-NN model is appropriate during free flow conditions (Figure 5-9b) and congested conditions (Figure 5-9d). However, when traffic conditions are in transient state (Figure 5-9c) the correlation between the Mahalanobis distance ranking and the prediction accuracy is much less strong. There are groups of historical instances that do not have the minimal Mahalanobis distance but they do provide a lower prediction error.

Generally, the results provided in Figure 5-9 suggest that the traffic conditions (or traffic state) may significantly affect the performance of the k-NN method. This observation suggests that having (or estimating) the traffic state may improve the accuracy of the travel time predictions. The Parametric approach proposed in the next section explores the potential to improve the accuracy of travel time prediction by explicitly consider the state of traffic.

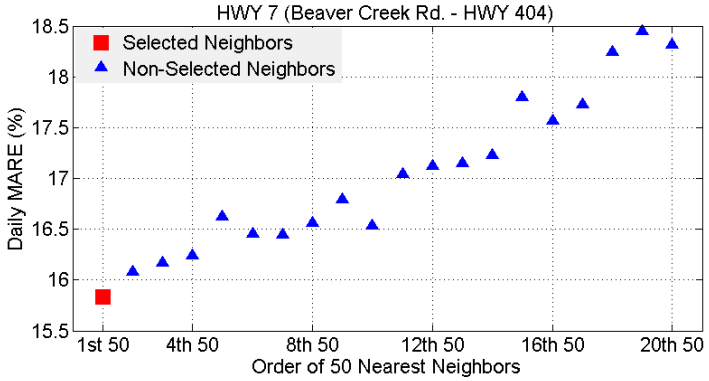
(a). Segment 1



(b).Segment 2



(c). Segment 3



(d).Segment 4

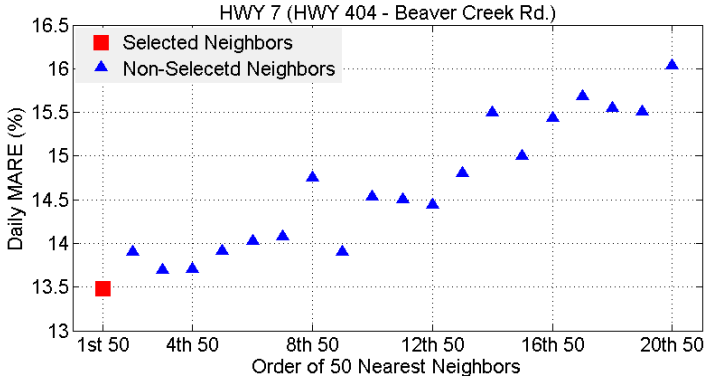


Figure 5-8: Daily prediction MARE Vs. k-NN similarity ranking (based on the distance metric) (Investigated segments – Sample day: Tuesday, March 13th, 2012)

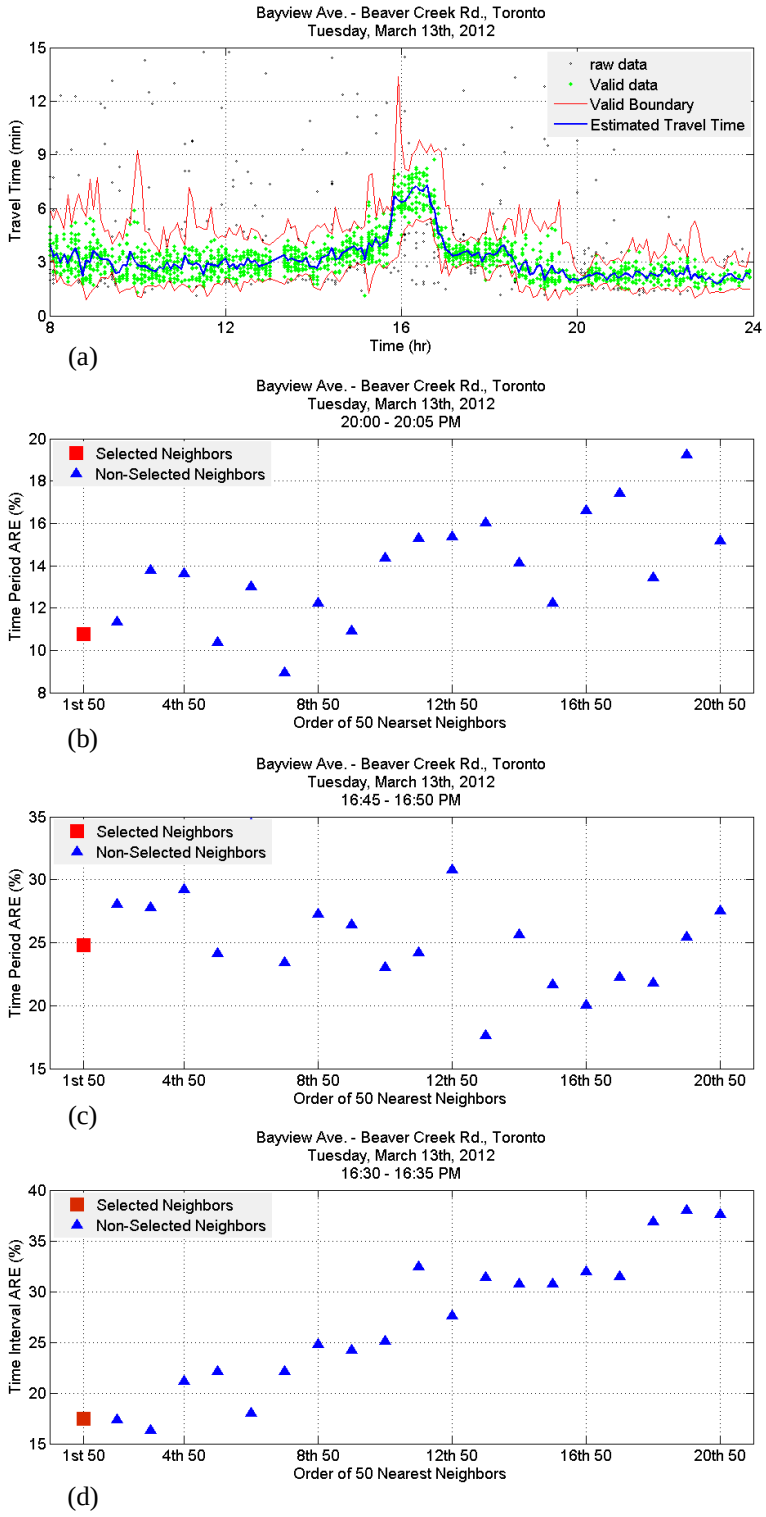


Figure 5-9: The performance of the proposed k-NN model for different time intervals, (a) sample daily travel time trend, (b) free-flow, (c) transient, (d) congestion

5.2 Modeling Travel Time as a stochastic process (Parametric Approach)[1]

Roadway travel time data are not independently nor identically distributed (iid). In other words, successive travel time observations are typically highly correlated, and their distribution varies depending on the roadway traffic states. Consequently, it is hypothesized that the accuracy of travel time prediction algorithms can be improved by explicitly considering the prevailing roadway traffic state and its progression pattern.

In the non-parametric k-NN model described in Section 5.1, changes in the roadway traffic states were endogenously addressed by including a set of explanatory variables in the feature vector. Conversely, in this section, we propose to model the traffic state transitions exogenously following a parametric approach. To reduce the level of modeling complexity, we propose to divide the travel time data into three different traffic states (e.g., free-flow, transient and congestion) and assume that the travel time data in each state are identically distributed, and model them separately. The travel time predictions from these models could then be combined using an appropriate fusing technique.

The real time applications of the proposed parametric approach involves, (i) determining the current traffic state (i.e., latest available time interval) and (ii) predicting the future traffic state (i.e., next time interval).

In this research, the detection of the current traffic state is modeled as a classification problem and the prediction of the future traffic state is modeled as a stochastic process.

The following section describe the development of the proposed traffic state detection and prediction models along with the results obtained from the application of the proposed Parametric model.

5.2.1 Real Time Traffic State Determination

5.2.1.1 Ground Truth Traffic States

Like any other modeling procedure, the first step in the detection of current traffic state is the knowledge of the ground truth state (for all the time intervals in the training dataset). As

[1] The content of this section is contained in a paper submitted for publication in the ASCE Journal of Transportation Engineering.

discussed in Chapter 2, the conventional definition of traffic state on freeways is not applicable on arterials. Moreover, the arterial traffic state determination procedure suggested by HCM requires both the roadway travel time and the level of services for all the intermediate intersections. Using Bluetooth technology it is not possible to follow this procedure, as we only have access to the travel time data collected along the investigated arterial segment.

In this research an off-line heuristic procedure is developed to label traffic states on the basis of the travel times measured by the Bluetooth detectors. This offline procedure can be automatically applied to all the observations in the historical dataset (i.e., 5 minutes estimated travel times).

The arterial travel time data obtained from the Bluetooth technology suffers from some levels of temporal instability for a number of time intervals which can be misleading in the state labeling procedure. These unstable fluctuations, which are more common during the transient traffic state (as opposed to the free flow or congested states), occur primarily as a result of sampling error and the high variability of travel times on arterial roadways. To handle this problem, a combination of smoothing and a rule based approach is considered.

Generally, three different traffic states are recognized, namely, "State 1: free-flow", "State 2: transient", and "State 3: congestion". As the discussed instabilities are more frequent at the verge of "transient", the proposed state labeling procedure is separated into two stages.

In the first stage, the traffic is classified into either "free-flow" (State 1) or "transient-congested" states (a combination of States 2 and 3). The original travel time observations are smoothed using a Kernel smoothing method (considering a number of prior and subsequent travel times), and then using a set of pre-defined heuristic rules, the traffic state is determined by comparing the smoothed travel time against a threshold value. This threshold value represents the HCM travel time value between LOS C and LOS D[1].

In the second stage, the observations labeled as "transient-congested" in Stage 1 are re-classified into "transient" (State 2) or "congested" (State 3) traffic states. This classification

[1] Based on the HCM procedure, the travel time threshold between LOS C and LOS D is 2.0 times the free-flow travel time.

uses the same procedure as in Stage 1 except that the threshold representing the boundary between the HCM LOS E and LOS F[1] is used. The complete details of this heuristic procedure are provided in the following paragraphs:

Step 1: Apply Kernel Smoothing (Nadaraya-Watson kernel-weighted average) on estimated travel times ($t_{(i,k)}$) to calculate the smoothed estimated travel times ($t_{(i,k)}^{Smoothed}$)

$$t_{(i,k)}^{Smoothed} = \frac{\sum_{p=1}^{N} \mathrm{K}(k,p) t_{(i,k)}}{\sum_{p=1}^{N} \mathrm{K}(k,p)} \quad \textbf{(5-6)}$$

$$\mathrm{K}(k,p) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(k-p)^2}{2b^2} \right) \quad \textbf{(5-7)}$$

Where, $t_{(i,k)}^{Smoothed}$ is the smoothed travel time on segment *i* during the investigated time interval (i.e., time interval *k*), $t_{(i,k)}$ is the travel time on segment *i* during the investigated time interval, $\mathrm{K}(.,.)$ is the radial basis function kernel, *p* is the p^{th} time interval, *k* is the investigated time interval, *b* is the band width around the investigated time interval (e.g., *b*=2 – two time intervals before and two time intervals after the investigated time interval), and *N* is the total number of time intervals (*N* = 288 – for the case of 5-minute time intervals).

Step 2: Determine if the traffic state at the investigated segment is "State 1: free-flow" or "State 2&3: Transient & Congested"

2-1) Determine the traffic state based on both the estimated and smoothed travel times

(Using the HCM travel time threshold – the boundary between LOS C and LOS D)

2-1-1) Label based on the estimated travel times

$$\begin{cases} t_{(i,k)} < 2 \times t_{(i)}^{f} & then \quad \therefore State_{(i,k)} = State\ 1 \\ t_{(i,k)} \geq 2 \times t_{(i)}^{f} & then \quad \therefore State_{(i,k)} = State\ 2\&3 \end{cases}$$

2-1-2) Label based on the smoothed travel times

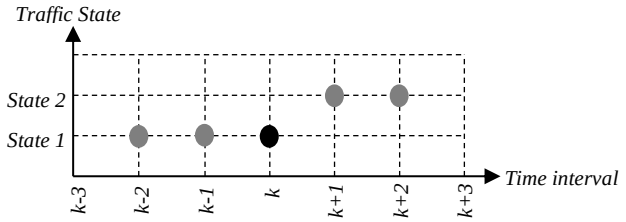
[1] Based on the HCM procedure, the travel time threshold between LOS E and LOS F is 3.3 times the free-flow travel time.

$$\begin{cases} t_{(i,k)}^{Smoothed} < 2 \times t_{(i)}^{f} & then \quad \therefore State_{(i,k)}^{Smoothed} = State\ 1 \\ t_{(i,k)}^{Smoothed} \geq 2 \times t_{(i)}^{f} & then \quad \therefore State_{(i,k)}^{Smoothed} = State\ 2\ \&\ 3 \end{cases}$$

2-2) The purpose for labelling the traffic states is to identify the persistent traffic condition during the time period of interest. As a result of small sample sizes or short term travel time fluctuations, it is possible for the smoothed or estimated travel time in a given time interval to reflect a traffic state that is inconsistent with the preceding or following traffic states. The following sets of heuristic rules are used to adjust the labels from steps 2-1-1 to avoid these inconsistences. In each of the following illustrations, the circles represent the state label from Step 2-1-1. The solid black circle is the label for the time period of interest. All these rules are mutually exclusive, meaning only one rule is applicable for any given situation. Each rule has two outcomes, namely (1) to keep the state label as determined from Step 2-1-1 or (2) to change the state label from the state determined from step 2-1-1 to the alternate state.

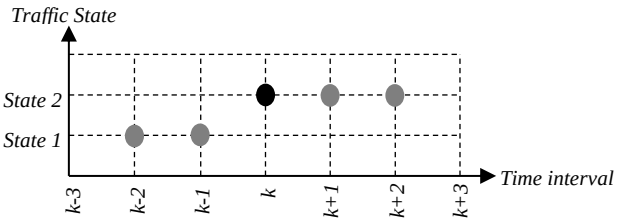
2-2-1) **State$_{(i,k)}$ =** State$_{(i,k-2)}$ = State$_{(i,k-1)}$

∴ Keep **State$_{(i,k)}$**



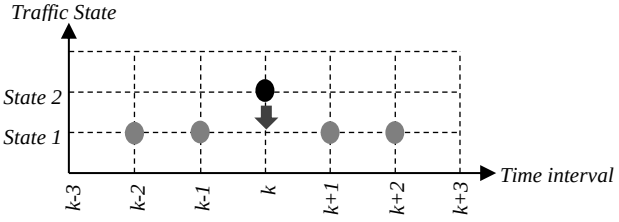
2-2-2) **State$_{(i,k)}$ =** State$_{(i,k+1)}$ = State$_{(i,k+2)}$

∴ Keep **State$_{(i,k)}$**



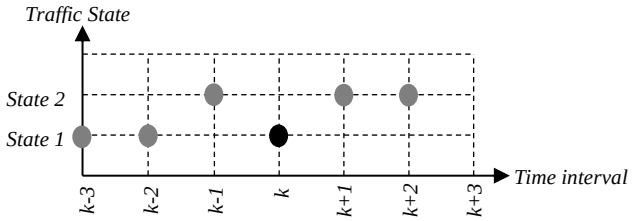
2-2-3) **State$_{(i,k)}$ ≠** State$_{(i,k-2)}$ = State$_{(i,k-1)}$ = State$_{(i,k+1)}$ = State$_{(i,k+2)}$

∴ Change **State$_{(i,k)}$**



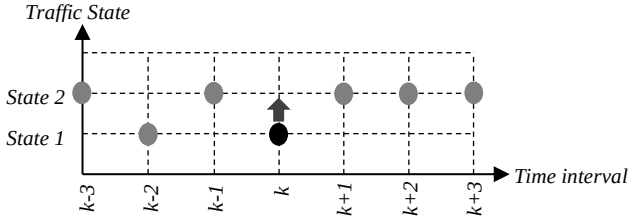
2-2-4) **State$_{(i,k)}$** = State$_{(i,k-3)}$ = State$_{(i,k-2)}$ ≠ State$_{(i,k-1)}$ = State$_{(i,k+1)}$ = State$_{(i,k+2)}$

∴ Keep **State$_{(i,k)}$**



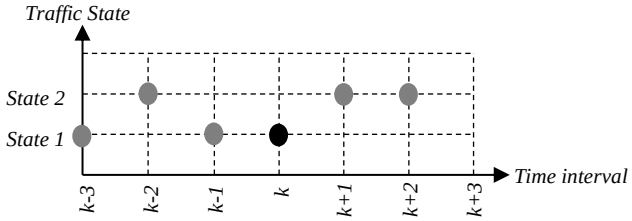
2-2-5) **State$_{(i,k)}$** = State$_{(i,k-2)}$ ≠ State$_{(i,k-3)}$ = State$_{(i,k-1)}$ = State$_{(i,k+1)}$ = State$_{(i,k+2)}$ = State$_{(i,k+3)}$

∴ Change **State$_{(i,k)}$**



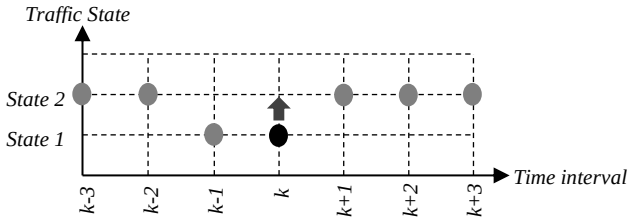
2-2-6) **State$_{(i,k)}$** = State$_{(i,k-3)}$ = State$_{(i,k-1)}$ ≠ State$_{(i,k-2)}$ = State$_{(i,k+1)}$ = State$_{(i,k+2)}$

∴ Keep **State$_{(i,k)}$**



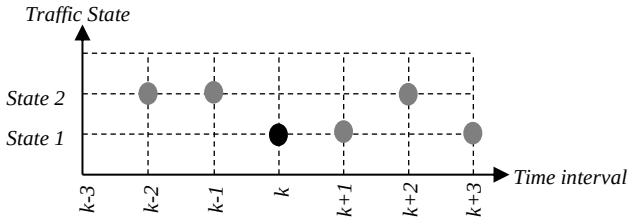
2-2-7) **State$_{(i,k)}$** = State$_{(i,k-1)}$ ≠ State$_{(i,k-3)}$ = State$_{(i,k-2)}$ = State$_{(i,k+1)}$ = State$_{(i,k+2)}$ = State$_{(i,k+3)}$

∴ Change **State$_{(i,k)}$**



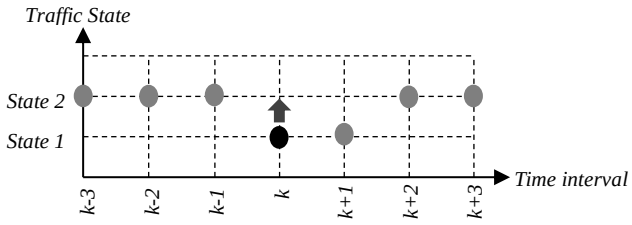
2-2-8) **State$_{(i,k)}$** = State$_{(i,k+1)}$ = State$_{(i,k+3)}$ ≠ State$_{(i,k-2)}$ = State$_{(i,k-1)}$ = State$_{(i,k+2)}$

∴ Keep **State$_{(i,k)}$**



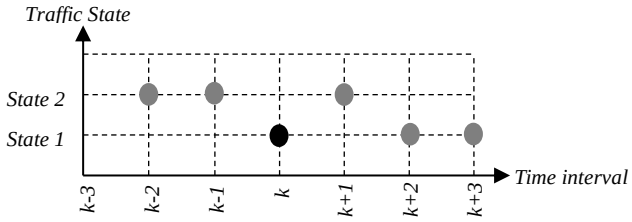
2-2-9) **State$_{(i,k)}$** = State$_{(i,k+1)}$ ≠ State$_{(i,k-3)}$ = State$_{(i,k-2)}$ = State$_{(i,k-1)}$ = State$_{(i,k+2)}$ = State$_{(i,k+3)}$

∴ Change **State$_{(i,k)}$**



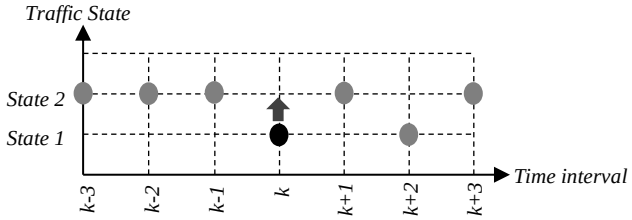
2-2-10) **State$_{(i,k)}$** = State$_{(i,k+2)}$ = State$_{(i,k+3)}$ ≠ State$_{(i,k-2)}$ = State$_{(i,k-1)}$ = State$_{(i,k+1)}$

∴ Keep **State$_{(i,k)}$**



2-2-11) **State$_{(i,k)}$** = State$_{(i,k+2)}$ ≠ State$_{(i,k-3)}$ = State$_{(i,k-2)}$ = State$_{(i,k-1)}$ = State$_{(i,k+1)}$ = State$_{(i,k+3)}$

∴ Change **State$_{(i,k)}$**



2-2-12) Other conditions

$\therefore State_{(i,k)} = State_{(i,k)}^{Smoothed}$

Step 3: Determine if the traffic state at the investigated segment is "State 2: Transient" or "State 3: Congested"

(Only applicable to the intervals labeled as "State 2&3: Transient & Congested" in step 2)

3-1) Determine the traffic state based on both the estimated and smoothed travel times

(Using the HCM travel time threshold – the boundary between LOS E and LOS F)

3-1-1) Label based on the estimated travel times

$$\begin{cases} t_{(i,k)} < 3.3 \times t_{(i)}^{f} \quad then \quad \therefore State_{(i,k)} = State\ 2 \\ t_{(i,k)} \geq 3.3 \times t_{(i)}^{f} \quad then \quad \therefore State_{(i,k)} = State\ 3 \end{cases}$$

3-1-2) Label based on the smoothed travel times

$$\begin{cases} t_{(i,k)}^{Smoothed} < 3.3 \times t_{(i)}^{f} & then \quad \therefore State_{(i,k)}^{Smoothed} = State\ 2 \\ t_{(i,k)}^{Smoothed} \geq 3.3 \times t_{(i)}^{f} & then \quad \therefore State_{(i,k)}^{Smoothed} = State\ 3 \end{cases}$$

3-2) Adjust the traffic states following a heuristic procedure similar to the one just presented in step 2-2.

Using the proposed approach the traffic state for all the observations in the historical dataset was determined. The performance of the applied state labeling procedure has been investigated manually for 16 random daily travel time trends. The satisfactory performance of the heuristic labeling procedure for one of these days is presented in Figure 5-10. In this evaluation the true traffic state is determined manually by looking at the trend of travel time along the whole day. As indicated in Figure 5-10a, by focusing on the stable trend of estimated travel time, the applied labeling method correctly disregarded two main instabilities (one at approximately 2:00 AM and the other at 5:00 PM) while capturing the stable changes occurring during the morning and evening peak periods. The proposed off-line state labelling process appears to perform quite well. Over the whole day (288 five-minute intervals) the algorithm appears to only misclassify the traffic state for 3 intervals (Figure 5-10b to Figure 5-10d) which is equivalent to a 1.1% daily misclassification rate. It should be noted that the mean and the 90^{th} percentile misclassification rates across all the investigated daily patterns (16 days) were 2.3 and 2.9%, respectively. We also note that because the true traffic state is not known, these statistics reflecting the accuracy of proposed off-line state classification process are somewhat subjective. However, our purpose at this stage is to demonstrate that the labelling is reasonably accurate and robust.

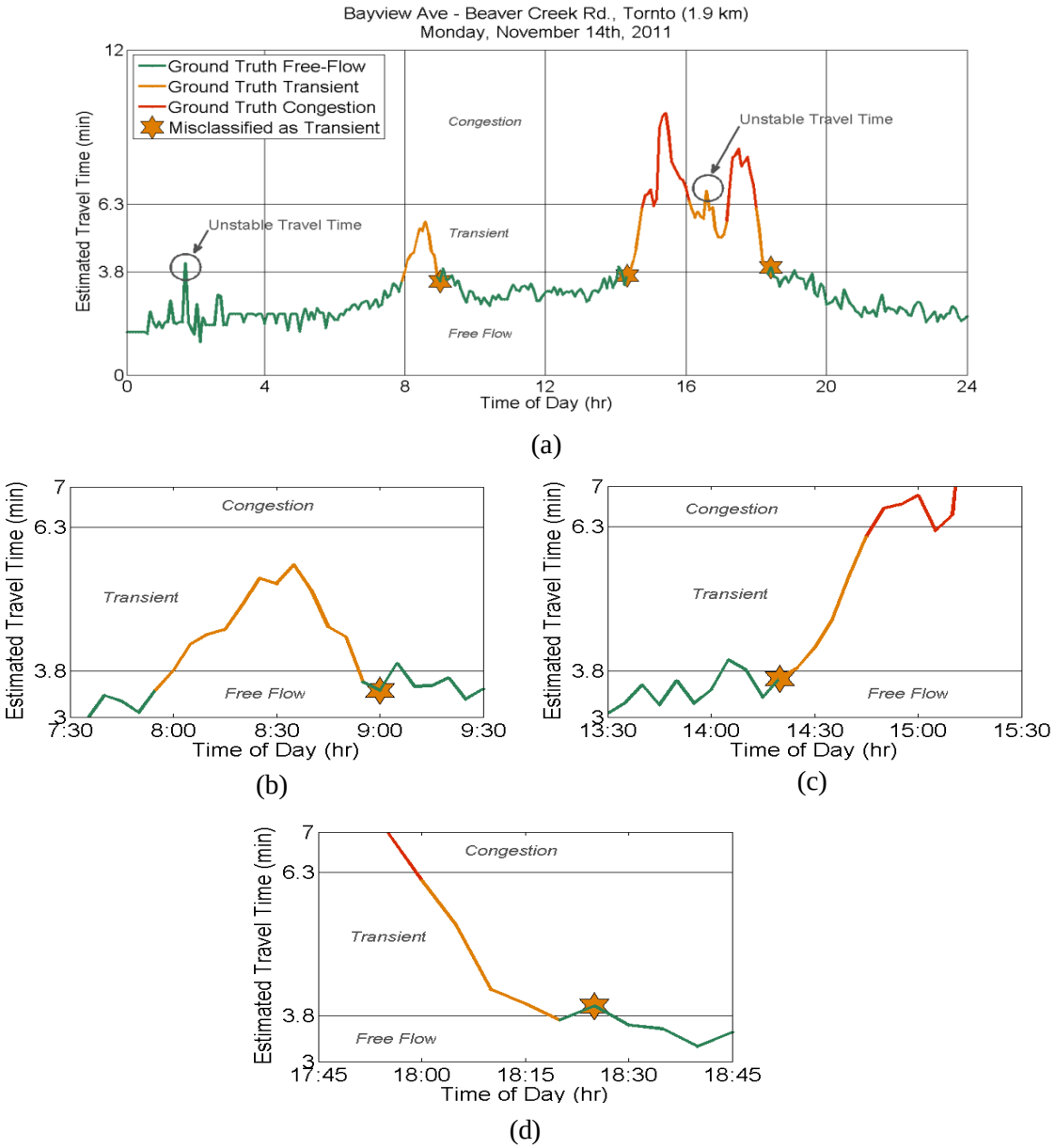


Figure 5-10: The performance of the off-line traffic state labeling algorithm for a random day (Monday, November 14, 2011 – Bayview Ave. – Beaver Creek Rd., Toronto (1.9 km))

5.2.1.2 Model Structure

In this research, the detection of the current traffic state is modeled as a classification problem. Similar to the off-line labeling procedure, the real-time traffic state detection also involves two steps. The first step consists of applying a support vector machine technique, in which the traffic state is categorized as either "state 1: free-flow" or "state 2-3: transient-

congestion". The second step, consisting of the application of a simple moving mean technique, is used to further classify the intervals labelled as "state 2-3: transient-congestion" in Step 1 into either "state 2: transient" or "state 3: congestion"[1].

Among the common classification methods (e.g., LDA[2], NB[3], LR[4], NN[5], SVM[6]), SVM is used in this work for the following reasons:

- SVM does not make any assumption about the distribution of the class features (unlike LDA, NB, and LR – where the first two assume the normality of the class features and the last one assumes Weibull distribution for the class utilities).
- The over fitting issue is less problematic for SVM (unlike NN).
- SVM is a mathematical optimization technique which converges to a unique solution for the classifier hyperplane (unlike NN).
- SVM requires the minimum number of parameters to be calibrated compare to other classification methods (e.g., LDA).

Developing a SVM model involves four different steps, namely: (1) feature selection; (2) determining the nonlinearity level of the classifier; (3) model calibration; and (4) model validation. To characterize the classifier between the free-flow state and the transient-congested state, the following features have been examined:

- Estimated mean travel time for the investigated segment (the last 3 time intervals with available travel time observations)
- Estimated mean travel time for the upstream segment (the last 3 time intervals with available travel time observations)

[1] Due to the relative stability of the travel time at the boundary of transient and congestion, the changes of traffic state between these two classes can be adequately characterized using moving mean technique.
[2] Linear Discriminant Analysis
[3] Naïve Bayes
[4] Logistic Regression
[5] Neural Networks
[6] Support Vector Machine

- The proportion of the summation of hits over the number of detections at the upstream and downstream detectors of the investigated segment[1] (the last 3 time intervals).
- The time lag between the last individual travel time observation and the current time.
- The number of counted trips for the investigated segment (the last 3 time intervals).
- The number of counted trips for the upstream segment (the last 3 time intervals).

To address the inherent Bluetooth time lag problem, a set of lag-free features (the last 4 items) are included in the above list. For the specific case of the number of counted trips, the last 3 time intervals[2] are considered regardless of the availability of travel time observations. In this way, for the time intervals with no available travel time observations (due to the long time lag), the number of counted trips is equal to zero.

The possible nonlinearity of the SVM classifier can be captured using appropriate Kernel functions. In this research, four common types of Kernels are examined, namely, linear (Kernel 1), quadratic (Kernel 2), 3rd order polynomial (Kernel 3), and RBF[3] (Kernel 4). The general format of the classifier hyperplane using a Kernel function can be written as follows:

$$g(x) = \beta^T \Phi(x) + \beta_0 \quad \textbf{(5-8)}$$

Where, x is the feature vector, β is the feature vector coefficients, β_0 is the hyperplane intercept, and $\Phi(x)$ is the mapping function projecting the feature vector to a higher dimension. Ultimately, the applied Kernel function can be represented as the inner product of the mapping functions:

$$\mathrm{K}(x, x^*) = \left\langle \Phi(x), \Phi(x^*) \right\rangle \quad \textbf{(5-9)}$$

The coefficients defining the optimum classifier hyperplane (i.e., β and β_0) are determined through a calibration process. In this way, a set of training data (including both the feature

[1] This proportion can approximately reflect the average delay at the location of each detector.
[2] By following the findings of Section 5.1, the temporal trend of each feature is captured along a 15 minutes time span (i.e., three 5-minute time intervals).
[3] Radial Basis Function

values and the ground truth labels) is considered, and the SVM optimization procedure is followed[1].

Finally, the performance of the developed model is examined over a new set of data (other than the training dataset) following a simple (one fold) cross-validation procedure.

5.2.1.3 Model Calibration and Validation

The model calibration has been completed examining six different sets of feature vectors (Table 5-4). The spatial characteristics of the traffic states have been captured by developing separate models for different segments. These considerations result in the calibration of 96 different SVM models (6 feature sets × 4 segments × 4 kernels = 96 models). To keep the required calibration time in a reasonable range, one month worth of data is considered for training the mentioned models (data for August 2012).

Table 5-4: The included features in the examined SVM models

Features	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Estimated Mean Travel Time (Investigated Segment – 3 Time Intervals)	✓	✓	✓	✓	✓	✓
Estimated Mean Travel Time (Upstream Segment – 3 Time Intervals)	---	---	---	---	✓	✓
Summation of Hits over Number of Detections (Upstream Detector – 3 Time Intervals)	---	✓	---	---	---	---
Summation of Hits over Number of Detections (Downstream Detector – 3 Time Intervals)	---	✓	---	---	---	---
Time Lag – Real-Time Analysis (Investigated Segment – Investigated Time Interval)	---	---	✓	---	---	---
Number of Counted Trips - Real-time Analysis (Investigated Segment – 3 Time Intervals)	---	---	---	✓	---	✓
Number of Counted Trips - Real-time Analysis (Upstream Segment – 3 Time Intervals)	---	---	---	---	---	✓

✓ Included --- Not Included

The performance of the developed models is evaluated both against each other and two benchmark methods (using 14 month validation data from June 2011 to July 2012). The first benchmark method, called the Naïve method, labels the traffic state of the investigated time

[1] Using a quadratic programming optimization, SVM finds a hyperplane for which its distance from the misclassified points is minimal. More details about the SVM calibration process can be found at Duda et al. (2012).

interval by simply comparing its estimated travel time with the threshold suggested by the HCM (i.e., the boundary between LOS C and LOS D). The second benchmark method, called the Moving Mean method, labels the traffic state by comparing the average estimated travel times from the last 3 time intervals to the HCM threshold.

The daily mean misclassification rates for the free-flow observations ($MMCR_F$), transient-congested observations ($MMCR_{TC}$), and all observations ($MMCR_{Overal}$) are considered as the measures of effectiveness reflecting the performance of the classification models:

$$MMCR_F = 100 \times \frac{1}{N_d} \sum_d (1 - \frac{F_d^{Correct}}{F_d^{All}}) \quad \textbf{(5-10)}$$

$$MMCR_{TC} = 100 \times \frac{1}{N_d} \sum_d (1 - \frac{TC_d^{Correct}}{TC_d^{All}}) \quad \textbf{(5-11)}$$

$$MMCR_{Overal} = 100 \times \frac{1}{N_d} \sum_d (1 - \frac{F_d^{Correct} + TC_d^{Correct}}{F_d^{All} + TC_d^{All}}) \quad \textbf{(5-12)}$$

Where, $F_d^{Correct}$ is the number of free-flow observations associated with day d which are classified correctly, $TC_d^{Correct}$ is the number of transient-congested observations which are classified correctly, F_d^{All} is the number of all free-flow observations, TC_d^{All} is the number of all transient-congested observations, and N_d is the number of evaluated days.

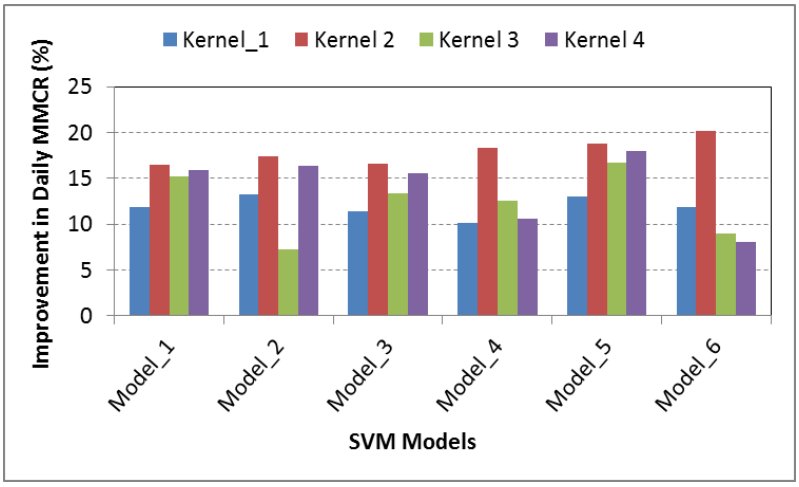
To complete this modeling procedure, the optimum kernel transformation and feature vector is determined using a sensitivity analysis.

The effect of different kernel transformations on the accuracy of the developed SVM models is investigated through Figure 5-11. In this figure, the performance of the examined models, while using different kernel functions (Kernel 1: linear, Kernel 2: quadratic, Kernel 3: 3rd order polynomial, Kernel 4: RBF), has been compared with the benchmark methods. The following main observations can be made on the basis of these results:

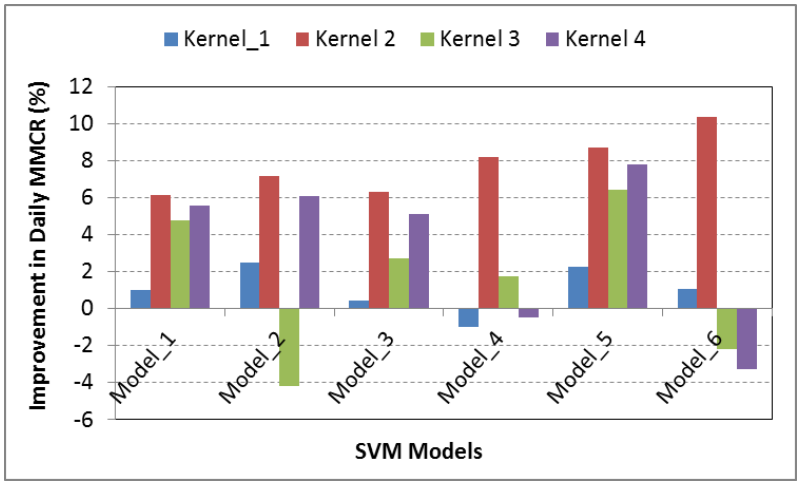
- The performance of the examined models under all kernel transformations was better than the Naïve method. The superiority of the investigated models against the moving mean method was limited to the linear and quadratic kernels, implying that

the possible nonlinearity of the feature space can be handled using a low order kernel function, as adding more complexities to the model degraded its performance.

- The best performance of the models was always achieved when using the quadratic kernel (Kernel 2). Thus, we continued this traffic state detection modeling using the quadratic kernel.



(a). Overall improvement over the Naïve Method



(b). Overall improvement over the Moving Mean Method

Figure 5-11: Relative performance of the SVM models compared to the Benchmark methods using different Kernels (Overall performance - Investigated over 14 month validation days).

To determine the optimum set of features, the performance of six different feature vectors was investigated. Table 5-5 reports the mean and the upper bound confidence interval of the daily misclassification rates obtained from applying the discussed models over 14 months of

validation days using the quadratic kernel. Based on the provided results, the following observations can be made:

- For overall and free flow conditions, the performance of all SVM models was superior to that of the benchmark methods.
- For transient-congested conditions, the performance of only two SVM models (Model 4 and Model 6) was superior to that of the benchmark methods.
- The best performances during the free flow, transient-congested and overall conditions were obtained from Model 5[1], Model 4[2], and Model 6[3], respectively.

Table 5-5: The Performance of the developed classification models in terms of daily misclassification rate (Quadratic Kernel - Investigated over 14 month validation days)

Classification Models	Misclassification Rate (%)					
	Free Flow		Transient - Congested		Overall	
	Mean	95th Percentile	Mean	95th Percentile	Mean	95th Percentile
Model 1	2.6	2.7	6.5	7.2	3.1	3.2
Model 2	2.6	2.7	6.1	6.8	3.0	3.1
Model 3	2.5	2.6	6.5	7.2	3.1	3.2
Model 4	2.8	2.9	5.1	5.7	3.0	3.1
Model 5	2.4	2.5	6.8	7.5	3.0	3.1
Model 6	2.6	2.7	5.4	6.0	2.9	3.0
Naïve	3.2	3.3	6.2	6.8	3.7	3.8
Moving Mean	2.8	2.9	6.0	6.5	3.3	3.4

The SVM calibration procedure does not provide the ability to make conclusions regarding the statistical significance of the calibrated parameters. To address this issue, the statistical significance of the included features has been exogenously investigated through the relative improvement matrix presented in Table 5-6.

[1] Model 5 includes the travel time data of the investigated and the upstream segments.
[2] Model 4 includes the travel time and count data of the investigated segment.
[3] Model 6 includes the travel time and count data of the investigated and upstream segments.

Table 5-6: Relative performance improvement matrix for the investigated classification models

(Quadratic Kernel - Investigated over 14 month validation days)

(a) Free-Flow State (Relative performance improvement (%))

Classification Models	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Naive	Moving Mean
Model 1	---	1.9	-1.4	7.5	-8.8	-0.4	20.8	8.7
Model 2	-2.0	---	-3.4	5.6	-10.9	-2.3	19.3	6.9
Model 3	1.4	3.3	---	8.8	-7.2	1.1	22.0	10.0
Model 4	-8.1	-6.0	-9.6	---	-17.5	-8.5	14.4	1.3
Model 5	8.1	9.8	6.7	14.9	---	7.7	27.2	16.0
Model 6	0.4	2.3	-1.1	7.8	-8.4	---	21.1	9.0
Naïve	-26.3	-23.9	-28.1	-16.9	-37.4	-26.8	---	-15.4
Moving Mean	-9.5	-7.4	-11.1	-1.3	-19.1	-9.9	13.3	---

(b) Transient-Congested State (Relative performance improvement (%))

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Naive	Moving Mean
Model 1	---	-7.0	-0.1	-26.6	4.1	-20.4	-4.7	-8.9
Model 2	6.6	---	6.5	-18.3	10.4	-12.5	2.1	-1.8
Model 3	0.1	-6.9	---	-26.5	4.2	-20.3	-4.7	-8.8
Model 4	21.0	15.5	20.9	---	24.3	4.9	17.3	14.0
Model 5	-4.3	-11.6	-4.4	-32.0	---	-25.5	-9.2	-13.6
Model 6	16.9	11.1	16.9	-5.2	20.3	---	13.0	9.5
Naïve	4.5	-2.2	4.5	-20.9	8.5	-14.9	---	-4.0
Moving Mean	8.2	1.7	8.1	-16.2	12.0	-10.5	3.8	---

(c) Overall (Relative performance improvement (%))

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Naive	Moving Mean
Model 1	---	-1.1	-0.2	-2.3	-2.8	-4.7	16.4	6.1
Model 2	1.1	---	0.9	-1.1	-1.7	-3.5	17.4	7.2
Model 3	0.2	-0.9	---	-2.0	-2.6	-4.5	16.6	6.3
Model 4	2.2	1.1	2.0	---	-0.6	-2.4	18.3	8.2
Model 5	2.8	1.7	2.5	0.6	---	-1.8	18.7	8.7
Model 6	4.5	3.4	4.3	2.4	1.8	---	20.2	10.3
Naïve	-19.7	-21.0	-19.9	-22.4	-23.1	-25.3	---	-12.3
Moving Mean	-6.5	-7.7	-6.8	-8.9	-9.5	-11.5	11.0	---

Statistically significant **Superior Performance** – 95 percent confidence level

In this table, the relative performance of each model has been compared against that of other models and the benchmark methods. The statistical significance of the relative performance was determined using the right tailed t-test. Based on the results in Table 5-6, the feature vector including the travel time and count data of the investigated and its upstream segment (Model 6) is considered as the proposed feature set.

The performance stability of the optimized model (Model 6) over the validation months is further investigated in Figure 5-12. According to this figure, for all the validation months, the proposed model provided better results than the benchmark methods. However, the month to monthly variation in the relative performance was relatively high (14% to 26% for the case of comparison with the naïve method and 4% to 16% for the case of comparison with the moving mean method).

In a further investigation, the daily MMCR cumulative distributions of the proposed SVM model and the benchmark methods are compared over all the 14 month validation days.

Figure 5-13 shows the performance comparison separately for the free flow state, transient – congested state, and overall conditions. This figure shows that the proposed model consistently provides better performance (i.e., cumulative curve is above and to the left) than the two benchmark methods for all traffic conditions.

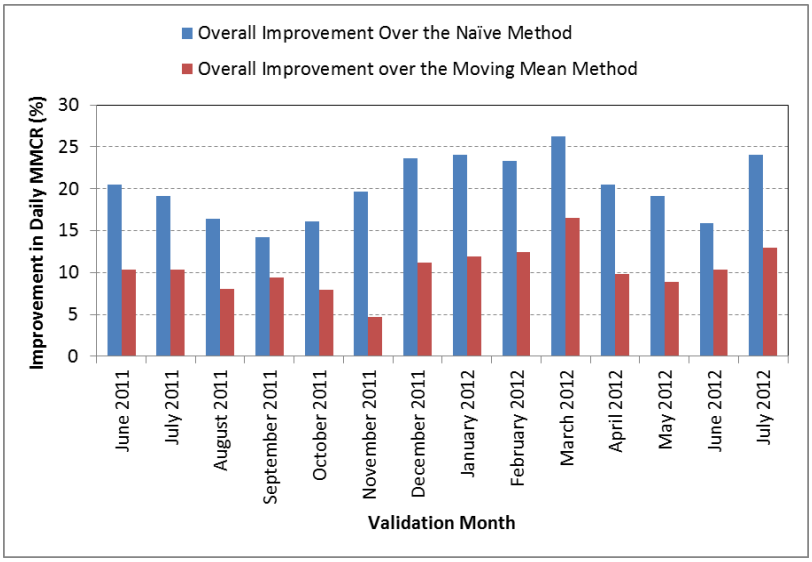
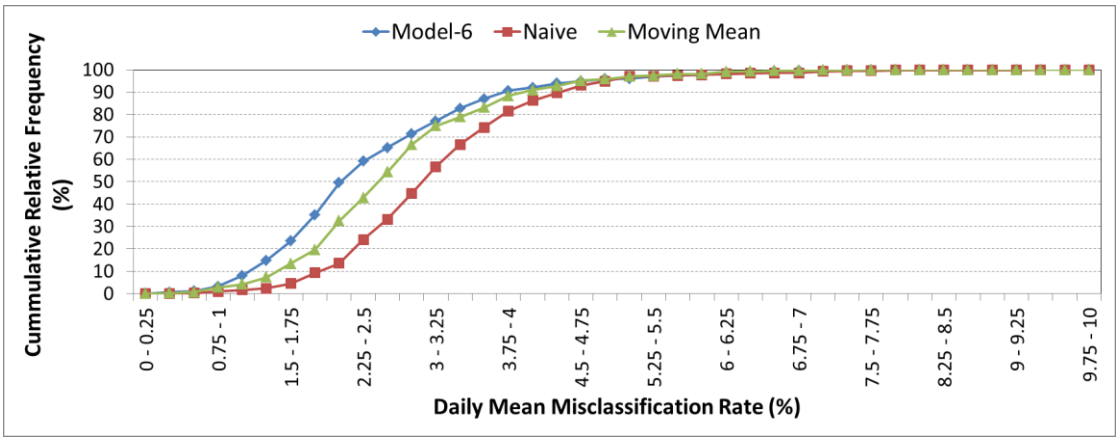
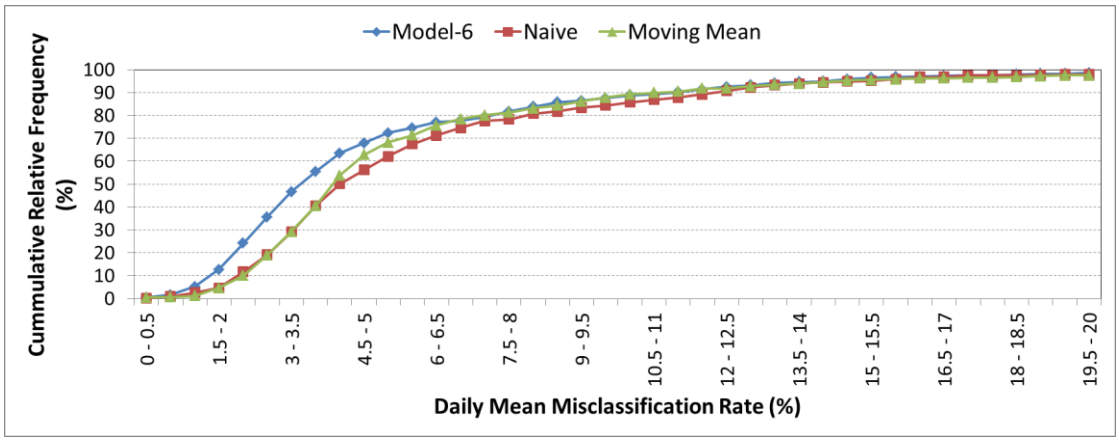


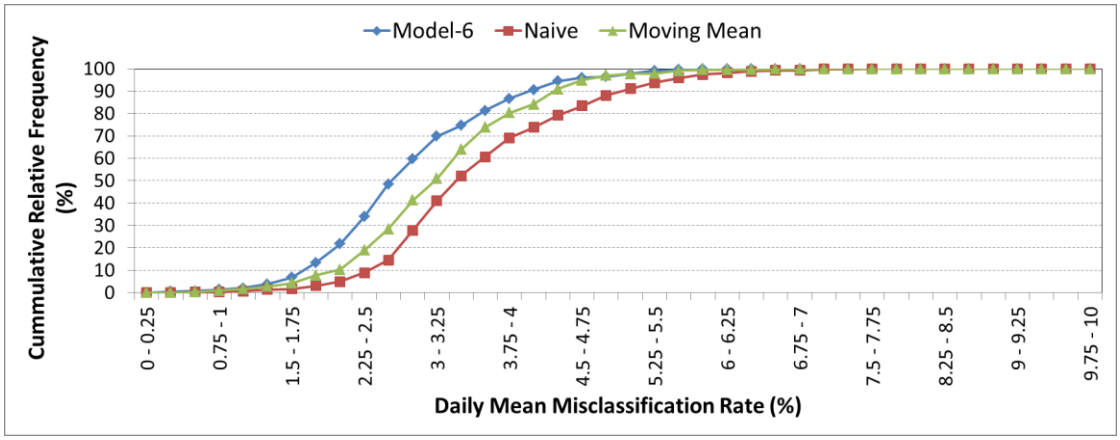
Figure 5-12: Monthly variation in the relative performance of the proposed SVM model compared to the Benchmark methods (Overall performance - Model 6 - Quadratic Kernel)



(a). Free-Flow State



(b). Transient – Congested State



(c). Overall

Figure 5-13: Daily MMCR cumulative distribution – Proposed model vs. Benchmark methods (Quadratic Kernel – Different traffic states - Investigated over 14 month validation days)

5.2.2 Real Time Traffic State Prediction

5.2.2.1 Model Structure

The time at which the traffic state on a roadway starts to change is a probabilistic event. Therefore, the transition of traffic states on roadways should be treated as a stochastic process. In this research, the stochastic nature of the traffic state transition is modelled using a multi-state Markov Chain. According to the Markov property, the conditional distribution of the future state of a process is independent of the past states. This property contradicts the nature of traffic state transition on arterials, where the stable changes in the traffic states are the result of recent traffic conditions on the roadway. This contradiction is avoided by including appropriate covariates in the proposed model.

In this study, the traffic state is defined as a discrete variable, but the transition between different states may occur at any time. To capture these properties, a continuous time Markov Chain with discrete state space is considered.

In continuous Markov Chain models, the time spent in each traffic state is a non-negative real value and follows an exponential distribution. Based on this property, the transition probability of the next time period can be calculated as follows:

$$\Pr(\Delta t) = e^{Q \times \Delta t} \quad \textbf{(5-13)}$$

Where, the operator $e^{()}$ represents matrix exponential, Q is the transition intensity matrix, and Δt is the length of the time interval. The transition intensity matrix reflects how quickly the transitions between traffic states happen and can be written as follows:

$$Q = \left[\lambda_{ij}\right] = \begin{bmatrix} -\sum_{k \neq 1} \lambda_{1k} & \cdots & \lambda_{1j} & \cdots & \lambda_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_{i1} & \cdots & -\sum_{k \neq i} \lambda_{ik} & \cdots & \lambda_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_{n1} & \cdots & \lambda_{nj} & \cdots & -\sum_{k \neq n} \lambda_{nk} \end{bmatrix}_{n \times n} \quad \textbf{(5-14)}$$

Where, n is the number of traffic states and λ_{ij} is the transition intensity between state i and state j which can be defined through the following equation:

$$\lambda_{ij} = \begin{cases} \lambda_{ij}^{base} \times e^{\sum_{s \in S} \beta_{s,ij} \cdot x_s} & \forall\ i \neq j \\ -\sum_{k \neq i} \lambda_{ik} & \forall\ i = j \end{cases} \quad \textbf{(5-15)}$$

Where, λ_{ij}^{base} is the base transition intensity between state i and state j, x_s is the s^{th} independent variable, $\beta_{s,ij}$ is the model coefficient for the s^{th} independent variable when traffic transitions from state i to state j, and S is the number of included independent variables.

According to the above equation, the effect of the system dynamics can be captured through the inclusion of explanatory covariates. In this module, the same Bluetooth obtained information included in the classification model is examined as the model covariates. The magnitude and the statistical significance of the investigated variables are determined through a calibration process. The maximum likelihood technique has been considered for this calibration practice, while the variances of the estimated coefficients are utilized to confirm the statistical significance of the investigated covariates. More details about the calibration of Markov Chain models can be found at Duda et al. (2012).

5.2.2.2 Model Calibration

The Markov model was calibrated assuming the three traffic states (i.e., free-flow, transient, and congested) as defined earlier in Section 5.2.1.1. In order to converge to a set of optimum parameter values, it is necessary to include adequate transition observations between different traffic states. In this way, all the available 15 month Bluetooth data are considered in the model training. Moreover, the spatial characteristics of the traffic state transitions have been captured by calibrating separate models for different segments.

The observed relative transition frequencies between different traffic states are presented in Table 5-7. The values provided in this table can be simply interpreted as the conditional transition probabilities between different traffic states given the current traffic state. For instance, when the current traffic state is labeled as “transient” on segment 1, the observed probability for transition to “free-flow” is 10.1%, remaining in “transient” is 84.5%, and transition to “congested” is 5.4%. It should be noted that in the calculation of the presented

relative frequencies, the rare transition observations between states 1 and 3 have been excluded from the modeling process.

Table 5-7: Observed relative traffic state transition frequencies – 5-minute time interval

Segment 1				Segment 2			
From	To			To			From
	State 1	State 2	State 3	State 1	State 2	State 3	
State 1	0.973	0.027	0[1]	0.988	0.012	0	State 1
State 2	0.101	0.845	0.054	0.084	0.878	0.037	State 2
State 3	0[1]	0.126	0.874	0	0.117	0.883	State 3
Segment 3				Segment 4			
From	To			To			From
	State 1	State 2	State 3	State 1	State 2	State 3	
State 1	0.943	0.057	0	0.933	0.067	0	State 1
State 2	0.041	0.888	0.071	0.027	0.877	0.096	State 2
State 3	0	0.112	0.888	0	0.142	0.858	State 3

[1] The rare transition observations between states 1 and 3 have been excluded from the modeling process.

As the simplest version, the observed transition probabilities can be modelled as a Markov process, considering constant state transition intensities. According to this model structure, which is referred as the "base model", the effect of recent traffic conditions on the transition intensities is ignored. The estimated intensities for this model are presented in Table 5-8. The 95 percent confidence intervals of each estimate and the value of the log-likelihood function of the developed models are also included in this table. The reported log-likelihood values for the base model can be later utilized to evaluate the performance of the alternative models.

Using the base model, the transition probabilities between different traffic states are regenerated for each segment and the results are provided in Table 5-9. A comparison of these estimated probabilities with the observed probabilities presented in Table 5-7 indicates that the base model performs reasonably well in terms of replicating the observed transition probabilities.

Table 5-8: Estimated transition intensities – 5-minute time interval (Base model)

Segment 1			
	State 1	State 2	State 3
State 1	-0.027 (-0.028,-0.025)[1]	0.027 (0.025,0.028)	0.000
State 2	0.101 (0.096,0.107)	-0.155 (-0.163,-0.149)	0.054 (0.050,0.059)
State 3	0.000	0.126 (0.116,0.136)	-0.126 (-0.136,-0.116)
-2 × log-likelihood	29,773.88		
Segment 2			
	State 1	State 2	State 3
State 1	-0.012 (-0.013,-0.011)	0.012 (0.011,0.013)	0.000
State 2	0.084 (0.078,0.091)	-0.122 (-0.130,-0.114)	0.037 (0.033,0.042)
State 3	0.000	0.117 (0.104,0.132)	-0.11 (-0.132,-0.104)
-2 × log-likelihood	15,603.17		
Segment 3			
	State 1	State 2	State 3
State 1	-0.057 (-0.061,-0.054)	0.057 (0.054,0.061)	0.000
State 2	0.041 (0.038,0.043)	-0.112 (-0.116,-0.108)	0.071 (0.068,0.074)
State 3	0.000	0.112 (0.107,0.117)	-0.112 (-0.117,-0.107)
-2 × log-likelihood	45,929.91		
Segment 4			
	State 1	State 2	State 3
State 1	-0.067 (-0.071,-0.063)	0.067(0.063,0.071)	0.000
State 2	0.027(0.025,0.029)	-0.123 (-0.127,-0.119)	0.096 (0.093,0.100)
State 3	0.000	0.142 (0.137,0.147)	-0.142 (-0.147,-0.137)
-2 × log-likelihood	52,231.13		

[1] 95 percent confidence interval for the estimated base intensities

Table 5-9: Estimated relative traffic state transition frequencies – 5-minute time interval

Segment 1				Segment 2			
From	To			To			From
	State 1	State 2	State 3	State 1	State 2	State 3	
State 1	0.975	0.024	0.001	0.989	0.011	0.000	State 1
State 2	0.093	0.860	0.047	0.079	0.888	0.033	State 2
State 3	0.006	0.109	0.885	0.005	0.104	0.891	State 3
Segment 3				Segment 4			
From	To			To			From
	State 1	State 2	State 3	State 1	State 2	State 3	
State 1	0.945	0.053	0.002	0.936	0.061	0.003	State 1
State 2	0.038	0.899	0.064	0.024	0.891	0.084	State 2
State 3	0.002	0.100	0.898	0.002	0.125	0.874	State 3

The transition intensities of the base model are calibrated to best represent the transition between different traffic states for an average case including different traffic conditions. In this way, the base model is not sensitive to the recent changes of traffic condition on the investigated roadway. As mentioned earlier, the effect of the recent traffic conditions on the transition intensities can be captured by including appropriate covariates in the model. In this research the contribution of different variables has been systematically investigated and the positive effect of the following covariates has been confirmed (Table 5-10):

- Estimated travel time for latest available time interval ($t_{(i,k-\Delta)}$)[1]:

  The magnitude of the estimated travel time (for the latest available time interval of the investigated segment) relative to the defined traffic state thresholds can be quite informative in predicting the future state of traffic on arterials. However, the existence of the time lag problem in Bluetooth travel times may negatively affect the efficiency of this variable.

- Estimated travel time for prior time intervals ($t_{(i,k-\Delta-1)}$ or $t_{(i,k-\Delta-2)}$):

  The rate of change in the magnitude of the estimated travel times from recent time intervals (for the segment under investigation) may be helpful in determining the future traffic state. This effect has been captured by the simultaneous inclusion of the estimated travel time during the latest available time interval and one of its prior time intervals. In fact, this consideration adds a short memory to the proposed Markov model.

- Upstream Segment estimated travel time ($t_{(i-1,k-\Delta')}$)[2]:

  To partially address the Bluetooth time lag issue, the upstream segment estimated travel time (during the latest available time interval) has been included as one of the models covariates. This consideration can be helpful, as the travel time variation on arterials is mainly affected by the upstream traffic flow fluctuation.

[1] *(k-Δ)* is the time period corresponding to the latest available travel time estimation on the investigated segment.
[2] *(k-Δ')* is the time period corresponding to the latest available travel time estimation on the upstream segment.

- Number of counted trips for the last time interval ($count_{(i,k)}$):

  One of the explanatory variables that can be utilized for addressing the Bluetooth time lag issue is the number of counted trips during the last time interval (Investigated segment). By including this variable, even for the time intervals with no travel time observations (due to the long time lag), some valuable information can be still provided to the model, as the number of counted trips in this case is equal to zero.

The calibration results of the "proposed model" are provided in Table 5-10. In this table, the 95 percent confidence interval of each estimate is provided in parenthesis and the significant covariates are provided in bold format.

Table 5-10: Calibration results for the proposed Markov model including covariates

Covariate			State 1	State 2	State 3
Segment 1	Base Intensities	State 1	-0.0017 (-0.0020,-0.0014)[1]	0.0017 (0.0014,0.0020)	0
		State 2	23.42 (17.69,31.00)	-23.42 (-31.01,-17.69)	0.0005 (0.0003,0.0009)
		State 3	0	511.3 (215.1,1216.0)	-511.3 (-1216.0,-215.1)
	Travel Time (Investigated Segment)	State 1	0	**0.471**[2] (0.448,0.495)	0
		State 2	**-1.609** (-1.689,-1.529)	0	**1.174** (1.064,1.285)
		State 3	0	**-1.289** (-1.425,-1.154)	0
	2nd Prior Interval Travel Time (Investigated Segment)	State 1	0	**0.408** (0.381,0.435)	0
		State 2	**0.289** (0.243,0.335)	0	**-0.186** (-0.258,-0.115)
		State 3	0	**0.173** (0.119,0.228)	0
	# of Counted Trips (Investigated Segment)	State 1	0	**0.050** (0.032,0.068)	0
		State 2	**0.052** (0.035,0.069)	0	**-0.086** (-0.115,-0.057)
		State 3	0	-0.004 (-0.030,0.022)	0
	-2 × log-likelihood		25,616.62		
Segment 2	Base Intensities	State 1	-0.0004 (-0.0005,-0.0003)	0.0004 (0.0003,0.0005)	0
		State 2	108.4 (61.9,189.7)	-108.4 (-189.7,-61.9)	2×10^{-6} (6×10^{-7},6×10^{-6})
		State 3	0	5394 (1389,20950)	-5394 (-20950,-1389)
	Travel Time (Investigated Segment)	State 1	0	**0.495** (0.464,0.526)	0
		State 2	**-1.516** (-1.632,-1.399)	0	**2.535** (2.306,2.765)
		State 3	0	**-1.070** (-1.288,-0.851)	0
	Prior Interval Travel Time (Investigated Segment)	State 1	0	**0.466** (0.425,0.506)	0
		State 2	0.0339 (-0.015,0.083)	0	**-0.701** (-0.847,-0.555)
		State 3	0	**-0.375** (-0.486,-0.264)	0
	Travel Time (Upstream Segment)	State 1	0	**0.182** (0.146,0.218)	0
		State 2	**-0.141** (-0.194,-0.087)	0	**0.052** (0.003,0.101)
		State 3	0	-0.011 (-0.052,0.030)	0
	-2 × log-likelihood		13,322.38		

[1] 95 percent confidence interval for the estimated base intensities
[2] Significant covariate coefficients are determined in **bold** format

Table 5-10: Calibration results for the proposed Markov model including covariates (Cont.)

Covariate			State 1	State 2	State 3
Segment 3	Base Intensities	State 1	-0.0071 (-0.0085,-0.0059)	0.0071 (0.0059,0.0085)	0
		State 2	76.51 (63.43,92.28)	-76.51 (-92.28,-63.43)	0.0007 (0.0005,0.0010)
		State 3	0	6305 (3659,10860)	-6305 (-10860,-3659)
	Travel Time (Investigated Segment)	State 1	0	**0.367** (0.337,0.397)	0
		State 2	**-1.492** (-1.618,-1.365)	0	**1.148** (1.039,1.258)
		State 3	0	**-1.965** (-2.098,-1.832)	0
	2nd Prior Interval Travel Time (Investigated Segment)	State 1	0	**0.457** (0.391,0.523)	0
		State 2	**-0.533** (-0.650,-0.416)	0	**0.251** (0.200,0.302)
		State 3	0	**-0.226** (-0.283,-0.170)	0
	Travel Time (Upstream Segment)	State 1	0	**0.054** (0.021,0.087)	0
		State 2	**-0.599** (-0.680,-0.518)	0	**0.045** (0.017,0.072)
		State 3	0	**-0.157** (-0.195,-0.118)	0
	# of Counted Trips (Investigated Segment)	State 1	0	**0.134** (0.116,0.152)	0
		State 2	0.011 (-0.012,0.035)	0	**-0.015** (-0.025,-0.005)
		State 3	0	**0.0314** (0.014,0.049)	0
	-2 × log-likelihood		38,626.35		
Segment 4	Base Intensities	State 1	-0.0100 (-0.0115,-0.0087)	0.0100 (0.0087,0.0115)	0
		State 2	52.41 (45.29,60.65)	-52.41 (-60.65,-45.29)	0.0001 (8×10^{-5},0.0001)
		State 3	0	20.46 (14.72,28.44)	-20.46 (-28.44,-14.72)
	Travel Time (Investigated Segment)	State 1	0	**0.383** (0.362,0.403)	0
		State 2	**-1.383** (-1.556,-1.209)	0	**1.668** (1.572,1.765)
		State 3	0	**-0.714** (-0.801,-0.628)	0
	Prior Interval Travel Time (Investigated Segment)	State 1	0	**0.379** (0.357,0.400)	0
		State 2	**-1.287** (-1.460,-1.114)	0	**0.430** (0.397,0.462)
		State 3	0	**-0.348** (-0.409,-0.287)	0
	# of Counted Trips (Investigated Segment)	State 1	0	**0.133** (0.116,0.149)	0
		State 2	**0.035** (0.010,0.060)	0	**-0.071** (-0.088,-0.054)
		State 3	0	**0.025** (0.012,0.038)	0
	-2 × log-likelihood		43,968.74		

[1] 95 percent confidence interval for the estimated base intensities

[2] Significant covariate coefficients are determined in **bold** format

Based on the results provided in Table 5-10, it is clear that the effect of estimated travel time ($t_{(i,k-\Delta)}$) is statistically significant for all the transition intensities estimated for the investigated segments (each segment includes 4 non-zero transition intensities for a specific covariate). The log linear effect of this covariate is positive for the transitions from lower states to upper states (i.e., from traffic state 1 to 2 and 2 to 3) implying that the higher the travel time is, the more likely the traffic changes to a more congested state. On the other hand, the effect is negative for the transitions from upper states to lower states (i.e., from traffic state 2 to 1 and 3 to 2) meaning that the lower the travel time is, the more likely the traffic changes to a less congested state.

The effect of the estimated travel time over the prior time intervals ($t_{(i,k-\Delta-1)}$ or $t_{(i,k-\Delta-2)}$) is also presented in Table 5-10 (prior time interval for segments 2 & 4 and 2nd prior time interval for segments 1 & 3). All the estimated transition intensities are statistically significant except one transition for segment 2 (transition from state 2 to 1).

The sign of the estimated effects for this covariate is not consistent over the investigated segments. On segments 3 & 4, the log linear effect of the prior intervals covariate is positive for the upper diagonal (i.e., from traffic state 1 to 2 and 2 to 3) indicating that the increase in the travel time in the prior interval will result in a higher chance of transition from less congested states to more congested states. The effect is negative for the lower diagonal (i.e., from traffic state 2 to 1 and 3 to 2) indicating that a decrease in the prior interval travel time will result in higher chance of transition from more congested states to less congested states.

This pattern does not hold for all the transitions on segments 1 & 2. For example, the transition intensity from state 2 to 3 is negative for both of these segments, implying that an increase in the magnitude of the recently estimated travel times results in an increased likelihood of transitioning from state 2 to 3. The difference in the pattern of congestion formation and recovery on the investigated segments resulted in the observed inconsistencies. Generally, the formation and recovery of the congestion on segments 1 and 2 happens more rapidly compared to the case of segments 3 and 4 where these changes occur more gradually.

Among the investigated segments, two of them have an upstream segment. According to the results provided in Table 5-10, the effect of the upstream estimated travel time ($t_{(i-1,k-\Delta')}$) for almost all the transition levels of these segments was statistically significant (7 out of 8 transition intensities). The signs of the estimated transition intensities for this covariate are exactly similar to those estimated for the investigated segment travel time covariate and are consistent with expectation.

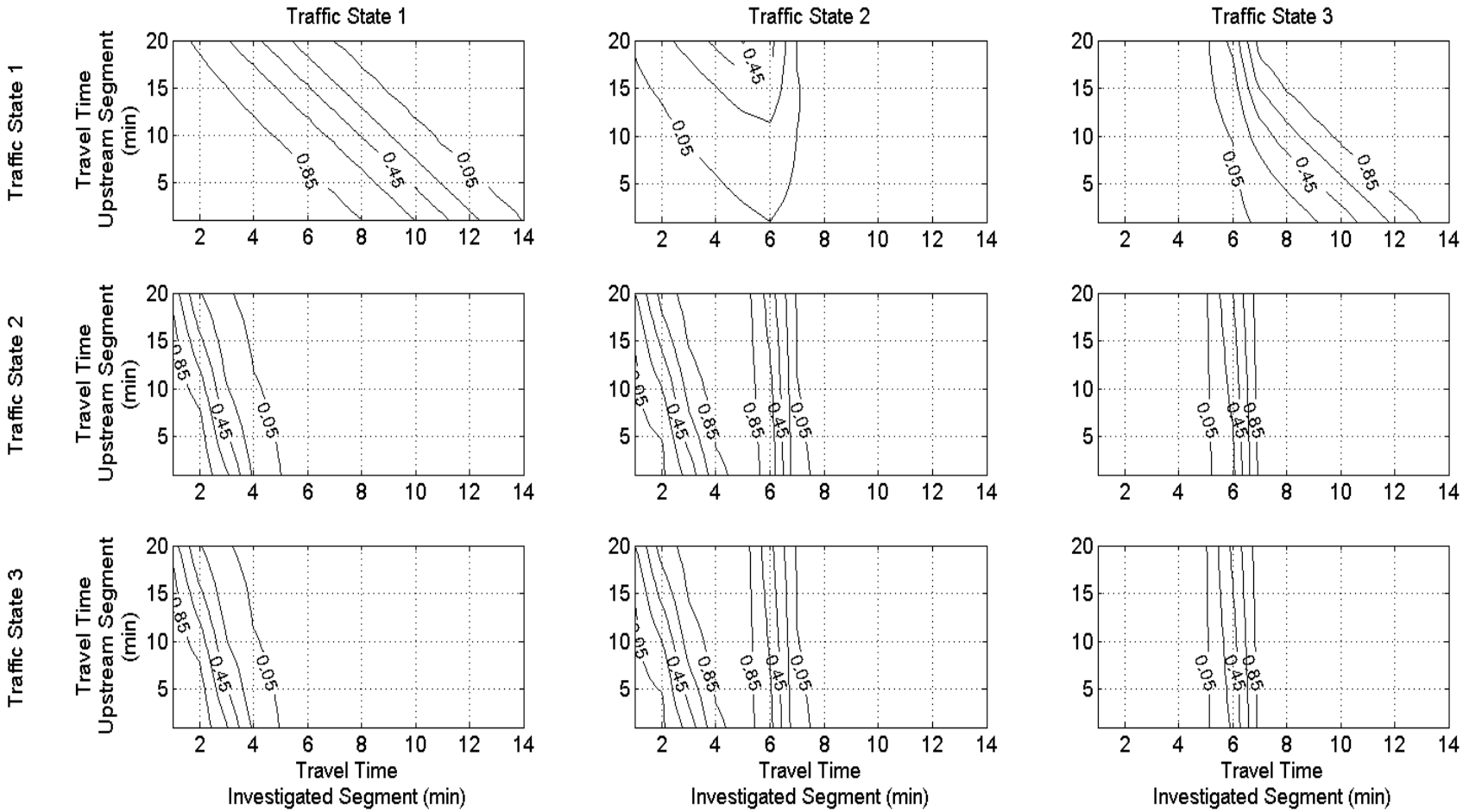
The significant contribution of the upstream segment travel time in the presented Markov Chain model appears to contradict the calibration results of the k-NN model, for which the inclusion of this variable did not statistically improve the prediction accuracies. This contradiction can be explained as follows:

- The response variable in the k-NN model is the near-future travel time; however, in the Markov Chain model the response variable is the near-future traffic state. These two variables are correlated, but still are not the same and may be characterized differently.
- For the k-NN approach, a single model is developed and the effect of different segments has been modeled through the blocking variables. On the other hand, for the Markov Chain model, separate models have been developed for each segment. This difference in the modeling approach, along with the fact that only two of the investigated segments have upstream segments, might contribute to the observed inconsistency.

As the last covariate, the effect of the number of counted trips ($count_{(i,k)}$) is presented in Table 5-10. Excluding segment 2, this covariate is included in the models developed for the remaining 3 segments. The high correlation between the upstream travel time covariate and the investigated number of counted trips covariate prevented us from including the latter in the model developed for segment 2.

For almost all the intensity levels the effect of the number of counted trips covariate was statistically significant (10 out of 12 transition intensities). The log linear effect of this covariate is positive for the transition from state 1 to 2, indicating that the higher number of counted trips will increase the chance of transition from "free flow" to "transient" states. For transitions from state 2 and 3 where relatively longer time lag periods are expected, the pattern of this effect changes. The positive sign of the transition intensities from state 2 to 1 and 3 to 2 implies that the lower number of counted trips (longer time lags) will increase the chance of transition from more congested states to less congested states. On the other hand, the negative sign of the transition from state 2 to 3 indicates that the higher number of counted trips will increase the chance of transition from "transient" to "congested".

Finally, using the developed Markov model, the estimated transition probabilities for a 5 minute time interval has been generated for segment 2 and the results are graphically presented in Figure 5-14. In order to calculate these probabilities, the value of the prior interval travel time has been fixed to be equal to 4 minutes.



Traffic State 1 = Free Flow; Traffic State 2 = Transition; Traffic State 3 = Congestion

Figure 5-14: Estimated transition probabilities for 5-minute time intervals (prior interval travel time = 4 minutes)

5.2.3 Application in Real Time Travel Time Prediction

5.2.3.1.1 Model Framework

Roadway travel time is a random variable with a non-symmetrical distribution. Some previous research suggests that the right skewed distribution of travel times can best be characterized as a log-Normal distribution (Li et al. 2006). However, the variability of the travel time observations changes under different traffic conditions.

Using the available Bluetooth data, the empirical distribution of the arterial travel times (5 minutes average) has been produced for each of the three traffic states (Figure 5-15). According to this figure, the travel time distribution patterns over the investigated traffic states are not similar. The distribution for "State 1" is almost symmetric and varies over a short range (1.5 – 4.0 minutes). In State 2 travel time observations spread over a relatively longer range (3.5 – 7.5 minutes) and follow a bimodal distribution. Finally, State 3 exhibits a right skewed distribution with a wide variation range (6.0 – 18.0 minutes) and is clearly different from the distributions associated with States 1 and 2.

The knowledge of traffic state can be helpful in travel time prediction applications. However, some of the traffic states (notably State 3) still exhibit considerable travel time variation, and consequently stable changes in the trend of travel time may not be adequately characterized using only near future traffic state information.

To reduce the variability of the travel time distribution, a set of restraining conditions has been applied. As the first condition, the traffic states during both the recent and future time intervals are considered. Following a similar approach developed by Noroozi and Hellinga (2014), the empirical probability density function of travel time can be written as the mixed distribution of different traffic states:

$$f(t_{(k+1)} \mid State_{(k-\Delta)} = m) = \sum_{n} \left((pr_{(k+1)}^{mn})^{(\Delta+1)} \times f(t \mid State_{(k-\Delta)} = m \ \& \ State_{(k+1)} = n) \right) \quad \textbf{(5-16)}$$

Where, $f(t \mid State_{(k-\Delta)} = m)$ is the empirical probability density function of travel time for the next time interval (i.e., time interval *(k+1)*) given the traffic during the last available time interval (i.e., time interval *(k-Δ)*) is in state *m*, $pr_{(k+1)}^{mn}$ is transition probability from traffic

state m to traffic state n for the next time interval (i.e., time interval $(k+1)$), and $f(t \mid State_{(k-\Delta)} = m \ \& \ State_{(k+1)} = n)$ is the empirical probability density function of travel time given the traffic during the last available time interval is in state m and the traffic during the next time interval is in state n (historical distributions).

In this research we are specifically interested in predicting the near future average travel times, which is equivalent to formulating the expected value of the above empirical travel time distribution through the following equation:

$$E(t_{(k+1)} \mid State_{(k-\Delta)} = m) = \sum_{n} \left((pr_{(k+1)}^{mn})^{(\Delta+1)} \times E(t \mid State_{(k-\Delta)} = m \ \& \ State_{(k+1)} = n) \right) \quad \textbf{(5-17)}$$

Where, the operator $E()$ represents the mathematical expectation.

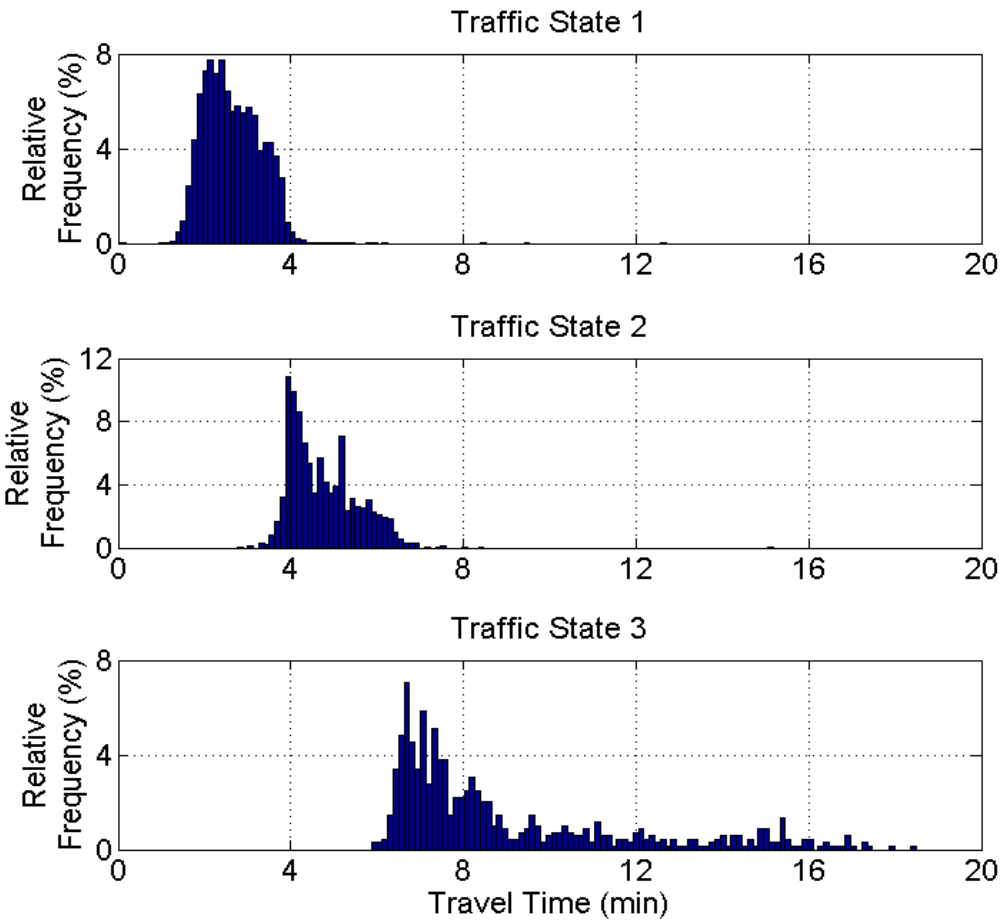


Figure 5-15: Empirical distribution of travel time for the defined traffic states (Segment 1, HWY 7 – Bayview Ave. - Beaver Creek Rd.)

Finally, the discussed high travel time variability has been further reduced by constraining the historical travel times (the conditional probability density function provided on the right hand side of Equation 5-16) to two weeks prior to the prediction day and a two hour window around the prediction time.

The above procedure has been incorporated in our real-time prediction framework, using the developed SVM and Markov models (Figure 5-16). Based on this approach, the most recent traffic state ($State_{(k-\Delta)}$) is determined by applying the trained SVM model and the transition probabilities ($pr^{mn}_{(k+1)}$) are estimated through the developed Markov model.

5.2.3.1.2 Model Evaluation

Following the procedure presented in Figure 5-16, the performance of the proposed parametric travel time prediction approach has been evaluated using 10 new validation days[1]. Table 5-11 presents the results of this validation in terms of MARE for four different methods, namely, the proposed parametric model (i.e., Markov model), the non-parametric model proposed in Section 5.1 (i.e., the optimized k-NN model), 2hr Window method, and the Benchmark method.

Using the offline labeling procedure explained in section 5.2.1.1, the traffic state of each time period can be determined. Based on this new information, the performance of the examined prediction models has been investigated during different traffic states (Table 5-11).

The Benchmark method is the same method introduced earlier in section 5.1.2.2, and the 2hr Window method simply considers the average travel time of the historical records belonging to two weeks prior to the prediction day and a two hour window around the prediction time. This is the same dataset considered for constructing the probability density function provided on the right hand side of Equation 5-16. In fact, by comparing the performance of the proposed Markov model and the 2hr Window method the explanatory capabilities of the proposed Markov model (i.e., from the inclusion of the traffic state information) can be investigated.

[1] These new validation days are different than those considered earlier for the evaluation of the proposed non-parametric travel time prediction approach.

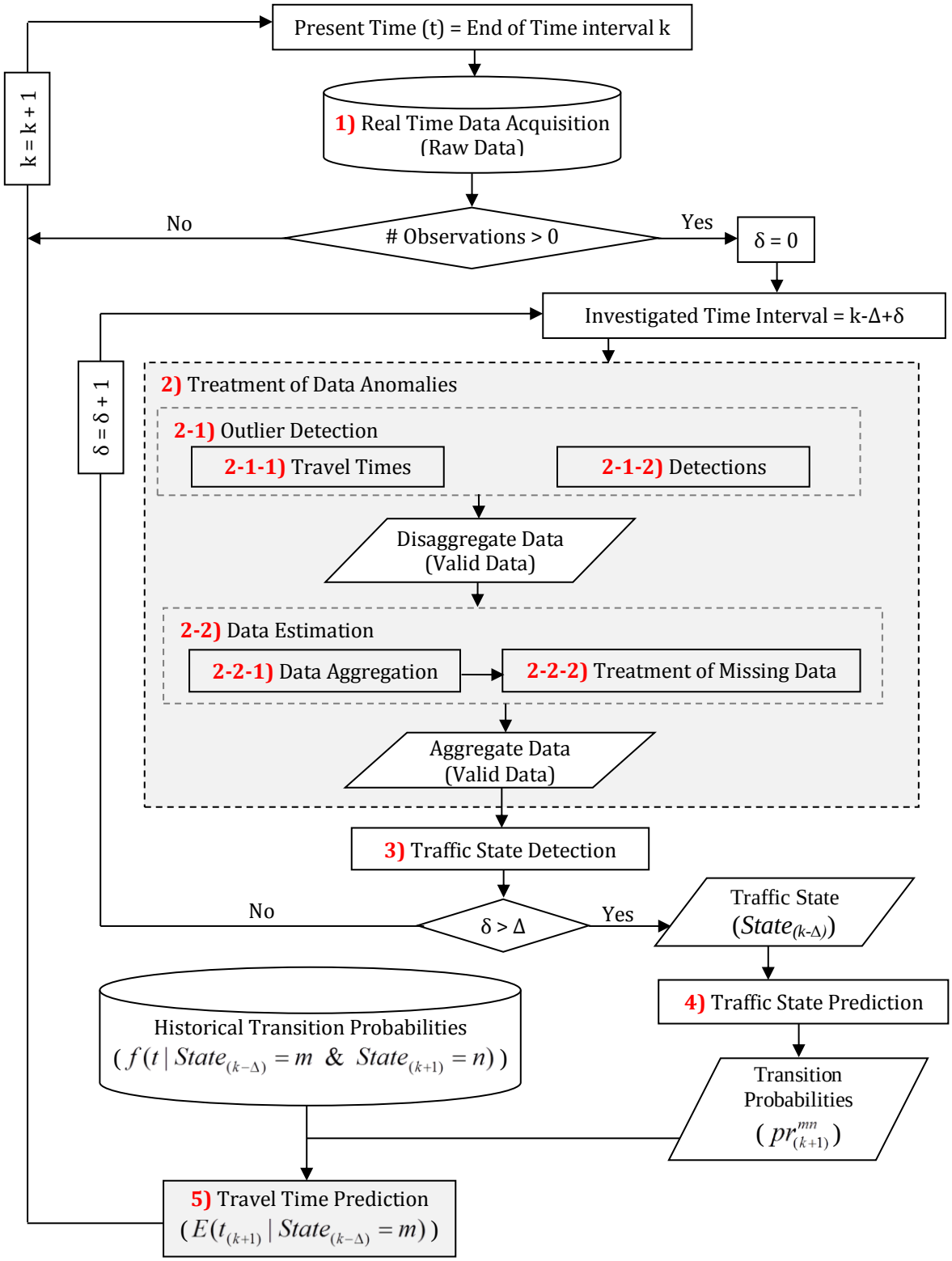


Figure 5-16: Proposed framework for travel time prediction on arterials (Parametric Approach)

Table 5-11: Performance of the Proposed models relative to the Benchmark models in terms of MARE (Different traffic states – Investigated over 10 validation days)

State	MOE		Segment 1 HWY 7 (Bayview - Beaver Creek)				Segment 2 HWY 7 (Beaver Creek - Bayview)			
			k-NN	Markov	2hr Window	Benchmark	k-NN	Markov	2hr Window	Benchmark
State 1	average		10.59	12.00	14.33	15.15	9.63	10.81	12.30	13.01
	Improvement over Benchmark (%)	Absolute	4.64	3.16	0.82	---	3.39	2.20	0.72	---
		Relative	30.59	20.83	5.44	---	26.02	16.94	5.50	---
State 2	Average		13.85	17.42	14.41	15.68	15.50	16.05	17.62	14.79
	Improvement over Benchmark (%)	Absolute	1.84	-1.73	1.27	---	-0.71	-1.26	-2.83	---
		Relative	11.71	-11.05	8.10	---	-4.78	-8.53	-19.12	---
State 3	Average		17.70	23.15	34.50	14.11	18.83	21.05	48.53	16.00
	Improvement over Benchmark (%)	Absolute	-3.58	-9.03	-20.39	---	-2.83	-5.06	-32.53	---
		Relative	-25.39	-64.00	-144.45	---	-17.69	-31.60	-203.38	---
Overall	Average		11.19	13.21	15.41	14.68	10.20	11.37	13.58	13.21
	90% Lower Bound		10.54	12.30	13.69	13.36	9.09	10.07	10.94	12.28
	90% Upper Bound		11.83	14.13	17.13	16.00	11.31	12.66	16.22	14.15
	Improvement over Benchmark (%)	Absolute	3.49	1.47	-0.73	---	3.02	1.85	-0.37	---
		Relative	23.80	10.01	-4.96	---	22.82	13.98	-2.80	---

State	MOE		Segment 3 HWY 7 (Beaver Creek - HWY 404)				Segment 4 HWY 7 (HWY 404 - Beaver Creek)			
			k-NN	Markov	2hr Window	Benchmark	k-NN	Markov	2hr Window	Benchmark
State 1	Average		19.78	21.17	25.74	23.94	18.78	20.61	23.43	25.02
	Improvement over Benchmark (%)	Absolute	4.15	2.77	-1.80	---	6.24	4.41	1.59	---
		Relative	17.35	11.56	-7.53	---	24.93	17.61	6.35	---
State 2	Average		11.84	14.25	20.75	14.71	10.73	12.93	16.15	13.35
	Improvement over Benchmark (%)	Absolute	2.87	0.46	-6.04	---	2.62	0.42	-2.79	---
		Relative	19.52	3.12	-41.04	---	19.64	3.16	-20.92	---
State 3	Average		12.17	14.25	15.27	14.83	11.96	15.22	16.28	16.13
	Improvement over Benchmark (%)	Absolute	2.66	0.58	-0.44	---	4.17	0.91	-0.15	---
		Relative	17.95	3.88	-2.96	---	25.86	5.62	-0.95	---
Overall	Average		15.49	17.35	22.05	18.85	13.81	16.28	19.43	18.09
	90% Lower Bound		14.68	16.48	20.28	18.16	12.90	15.14	17.68	17.13
	90% Upper Bound		16.31	18.22	23.81	19.54	14.72	17.41	21.19	19.05
	Improvement over Benchmark (%)	Absolute	3.36	1.51	-3.19	---	4.28	1.81	-1.34	---
		Relative	17.83	7.98	-16.93	---	23.67	10.01	-7.42	---

Statistically significant **Superior Performance** over the Benchmark method – 90 percent confidence level

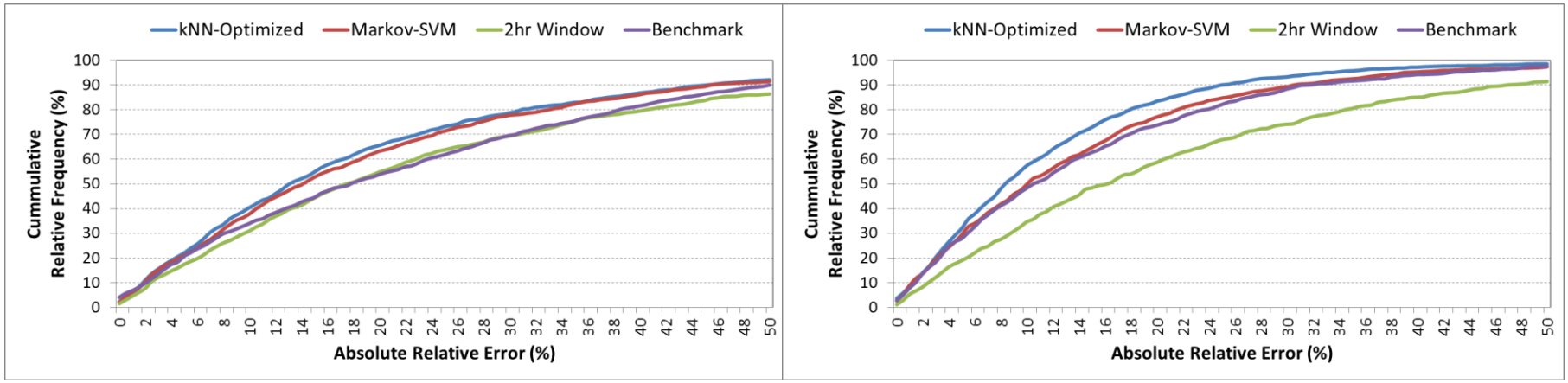
A number of observations can be made on the basis of the results provided in Table 5-11:

- On all the investigated segments, the performance of the proposed Markov model for both "State 1" and "Overall" traffic conditions was statistically superior to that of the Benchmark.
- The performance of the proposed Markov model relative to the Benchmark during "State 2" and "State 3" is not consistent over all the investigated segments. On the first two segments and during the mentioned two traffic states, the Benchmark model provides better prediction accuracies compared to the proposed Markov model. The more rapid formation and recovery of traffic congestion on these segments compared to the last two segments compounded with the existence of the time lag issue, resulted in the inferior performance of the proposed Markov model.
- Among the examined models, the performance of the 2hr Window method was the worst. This condition holds on all the investigated segments and during most of the traffic states. The observed superior performance of the Markov model over the 2hr Window method confirms the explanatory capabilities of the former through the inclusion of information regarding the traffic state transition.
- The number of prior weeks and the length of the time window which has been considered in the Markov model did not result from an optimization method. Rather these values were selected on the basis of engineering judgment and are intended to demonstrate the capability of the developed parametric algorithm in providing reasonable prediction accuracies using a small historical dataset (approximately 10.5% improvement over the Benchmark in daily MARE).
- Among the examined models, the performance of the Optimized k-NN model was the best. This condition holds on all the investigated segments and during most of the traffic states. The superiority of the Optimized k-NN model over the Markov model is mainly the result of its flexible non-parametric nature. Due to this flexible characteristic of the Optimized k-NN model, it was possible to include more explanatory variables in the feature vector.

- The relative performance of the Optimized k-NN model against the Benchmark is consistent over the two validation datasets. The improvement in daily MARE for the validation days examined here and those considered in section 5.1.2.2 is approximately 22%.

The performance of the examined models has been further investigated through the empirical cumulative distribution of the prediction errors presented in Figure 5-17. In this figure, the vertical axis is the relative frequency and the horizontal axis is the absolute relative error of the 5 minutes time interval predictions. The sample cumulative error distributions provided in this figure reflect the prediction accuracies for different traffic states on Segment 3. Based on this figure, the superior performance of both proposed methods over the Benchmark and 2hr Window methods can be verified.

Finally, the proposed Markov and k-NN models have been applied to predict near future travel times for a sample day on all the investigated segments (Figure 5-18). In this figure, the dark grey points represent the estimated travel times, the red line shows the Markov model prediction trend, and the blue line presents the k-NN model prediction trend. More sample daily prediction trends showing the predicted values resulted from the proposed k-NN and Markov models are provided in Appendix B.



(a) Traffic State 1

(b) Traffic State 2

(c) Traffic State 3

(d) All Traffic States

Figure 5-17: Prediction error distribution - Proposed models vs. Benchmark methods

(Different traffic states – Segment 3 (HWY 7 – Beaver Creek Rd. - HWY 404) – Investigated over 10 validation days)

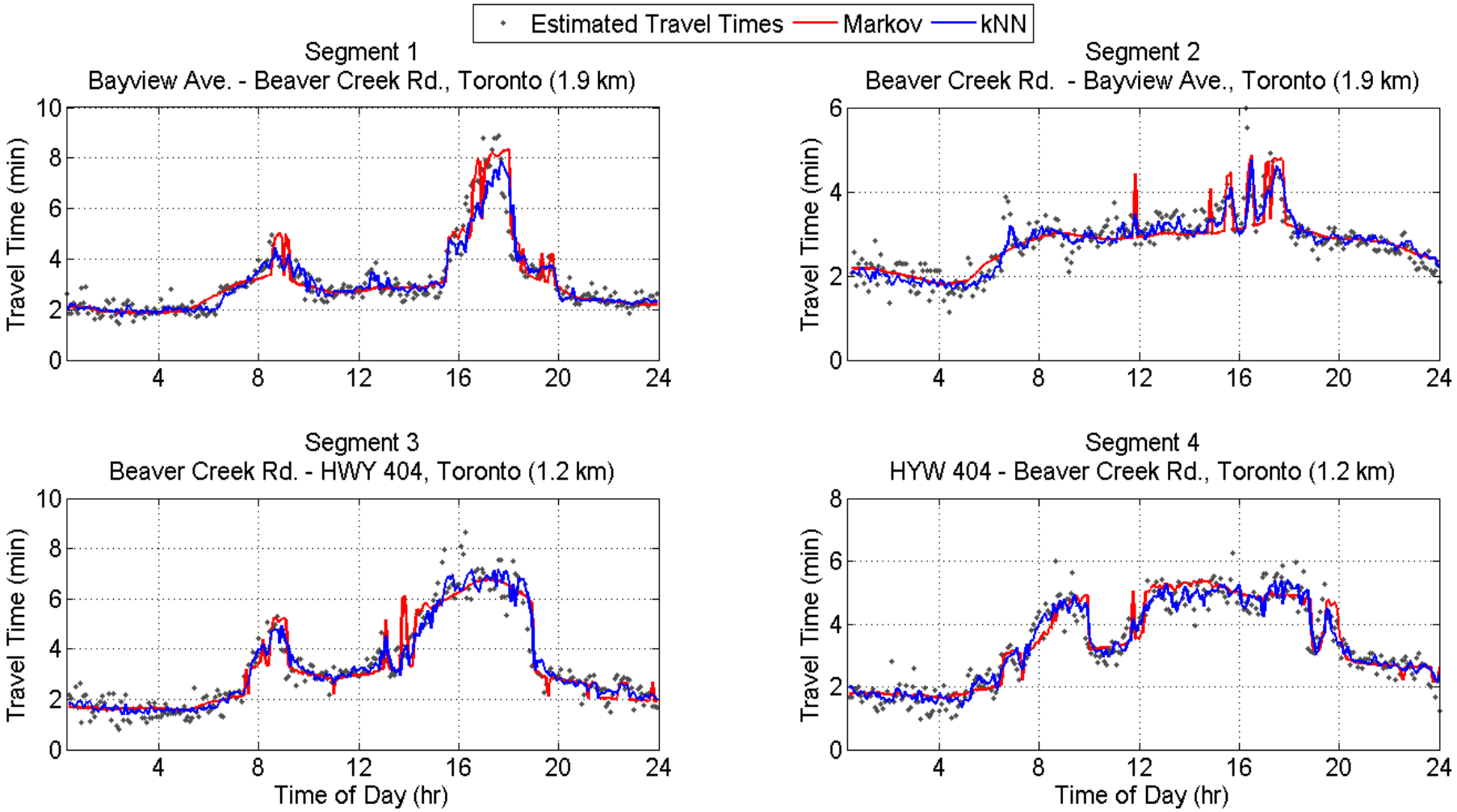


Figure 5-18: Travel time prediction - Markov model vs. k-NN model

(Sample day: Friday, February 17th, 2012)

5.3 Chapter Summary

This research has described two new methods for predicting near-future travel times on arterial roadways using data obtained from Bluetooth detectors. The performance of both methods is evaluated for real time applications using a robust framework which is devised to mimic the real-time acquisition of the Bluetooth data.

5.3.1 Modeling Travel Time as a Pattern Recognition Problem

The following observations and conclusions are made on the basis of the analysis described in this section:

- A Non-Parametric k-NN model is proposed to predict the arterial travel times. The model parameters have been optimized using an objective process (i.e., experimental design). This approach permits the investigation of the relative importance of different data types within the feature vector.
- The calibration results demonstrate the importance of using the lag-free detection data collected by single Bluetooth detectors in conjunction with measured travel times.
- The stability of the calibration results is confirmed through a cross validation process where the performance of the optimized k-NN model is assessed against the performance of all 1120 k-NN models considered within the parameter optimization process.
- The superior performance of the optimized k-NN model over a Benchmark method is verified using two different validation datasets each including 10 days' worth of data for 4 arterial segments.
- The superior performance of the optimized k-NN model over the developed Markov model is confirmed using 10 days validation data along 4 arterial segments.
- The overall efficiency of the implemented k-NN technique is demonstrated for all the investigated arterial segments.

5.3.2 Modeling Travel Time as a Stochastic Process

The following observations and conclusions can be made on the basis of the analysis described in this section:

- An off-line heuristic procedure is developed to automatically label traffic states on the basis of the travel times measured by the Bluetooth detectors.
- An optimum SVM classification model is developed to detect the most current roadway traffic state in real time. The optimum performance of the SVM model resulted when (1) using the quadratic kernel transformation and (2) selecting recent travel time and count data of the investigated and its upstream segments as the feature vector.
- The performance of the optimized SVM model is examined against two Benchmark methods (using 14 month validation days), and its consistent superior performance is confirmed.
- A continuous time Markov Chain model with three states is developed to model the transition between different traffic states. Using a 15 month training dataset, the calibration process is completed for the investigated segments and the significant covariates are identified.
- Using the optimized SVM and Markov models the conditional empirical probability density function of travel time on the investigated segment is formulated and its expected value is considered as the target predicted travel time.
- The superior performance of the optimized parametric model (Markov-SVM) over two Benchmark methods is confirmed using 10 days of validation data along 4 arterial segments.
- Despite the superior performance of the proposed Markov-SVM model, the results indicate that the proposed non-parametric model (k-NN) model was superior to the proposed Markov-SVM model.

Chapter 6
Conclusions and Recommendations

6.1 Conclusions

Near future travel time is one of the key traffic state indicators which can be utilized in proactive traffic management (PTM) strategies. However, short term prediction of travel times requires real time collection of these data. Among the limited available techniques, Bluetooth detectors are considered as a viable technology for acquiring vehicle travel times in real time. However, like any other measurement tool, these detectors are also subject to different sources of outliers and measurement errors. To improve the accuracy of travel time predictions, it is required to discern these anomalies from the valid travel time observations in real time.

Performing short-term travel time prediction on arterials per se is a non-trivial task, mainly due to the complex nature of traffic flow on urban streets and high variability of travel times in this environment. The mentioned prediction process, using noisy Bluetooth obtained travel time measurements, becomes more complicated for real time applications.

In this thesis we addressed these challenges and proposed an integrated framework for real time prediction of travel time on arterials using Bluetooth data. The main contributions of this research and the direction for further studies are described in this chapter.

6.2 Major Contributions

The work described in this thesis has resulted in the following five major contributions: (1) Quantification of the measurement error in Bluetooth travel times under various arterial traffic conditions; (2) Development of an algorithm to detect the Bluetooth travel time outliers in real-time; (3) Development of a simulation framework to objectively evaluate the performance of travel time outlier detection algorithms; (4) Development of two different algorithms for real time prediction of travel time on arterials, and (5) Development of an algorithm to detect the prevailing arterial traffic state and characterize its transition pattern

for the near-future. The following sections describe each of these contributions in more detail.

1. *Quantifying the measurement error in Bluetooth travel times under various arterial traffic conditions:* This research introduced a state of the art simulation approach which is utilized to characterize the measurement error in Bluetooth travel times collected on signalized urban roadways. Using the developed simulation framework, we identified three important control parameters (v//c, cycle length, and coordination level) influencing the magnitude of the investigated measurement errors and examined the resulting error distributions. We further developed a multiple regression model that is applied to assess the magnitude of the measurement error affecting Bluetooth acquired travel times.

2. *Development of an algorithm to detect the Bluetooth travel time outliers in real-time:* This research demonstrated the limitations of the existing travel time filtering algorithms (those that are capable of being implemented in real-time applications) using field data. To appropriately detect travel time outliers, a proactive adaptive filtering technique is proposed. The novel idea behind the proposed algorithm is the construction of a proactive validity window using quasi-predictions in conjunction with recent travel time estimates. We demonstrated that the addition of the novel quasi-prediction element improved the ability of the proposed algorithm to capture sudden changes in the trend in travel times. The performance of the proposed outlier detection algorithm was further demonstrated using field data.

3. *Development of a simulation framework to objectively evaluate the performance of travel time outlier detection algorithms:* In this study an innovative simulation framework was developed to objectively evaluate the performance of outlier detection algorithms under different conditions. As part of this framework, we proposed a new robust measure (*RTTI*) to quantify the performance of outlier detection algorithms. Following an experimental design procedure, the proposed evaluation framework provides clear evidence of how the examined outlier detection algorithms perform over a wide range of conditions. On this basis it is possible to conclude which filtering

technique is superior and to know the range of improvements that the algorithm is likely to provide.

4. *Development of two different algorithms for real time prediction of travel time on arterials:* In this research two new travel time prediction methods are developed. Using Bluetooth obtained data, the proposed models are utilized to perform near-future travel time prediction on arterials. A robust approach is followed to mimic the real-time acquisition of Bluetooth data through which the performance of the developed prediction algorithms is examined. A Non-Parametric k-NN model is considered as the first prediction approach. The objective optimization of the proposed k-NN model has been completed following an efficient full factorial analysis which allowed us to investigate the significance of the model parameters (both main effects and interactions) with a reasonable number of prediction runs.

   The second prediction approach is a Parametric model. By merging the results obtained from a classification model and a stochastic process, the conditional empirical probability density function of travel times is formulated. The expected value of the generated probability density function has been considered as the arterial travel time predictions. While requiring a small size historical dataset, the proposed Parametric prediction model outperformed two Benchmark methods.

   Our results show that of the two proposed model (Non-Parametric and Parametric), the Non-Parametric model outperforms the Parametric model.

5. *Development of an algorithm to detect the prevailing arterial traffic state and characterize its transition pattern for the near-future:* In this study, the progression of the traffic state is modelled. Using Bluetooth obtained data, the detection of the prevailing traffic state is modeled as a classification problem (using SVM technique) and the transition of traffic state is modeled as a stochastic process (using Markov Chain). The performance of the calibrated SVM and Markov models has been demonstrated through separate evaluation processes.

6.3 Future Research

The following topics improve and complement this research and therefore are recommended for future research:

1. Knowledge of the distribution of measurement error for Bluetooth obtained travel times can be useful for removing the measurement error through the filtering techniques such as the Kalman filter. The potential improvement that can be achieved through these techniques should be investigated.
2. The performance of outlier detection algorithms cannot be objectively quantified when applied to field data. Consequently, it is recommended to extend the simulation framework developed in Section 4.4 so that it can be used to evaluate outlier detection algorithms that make use of historical travel time data (such as the algorithm proposed in Section 4.3). Such a framework would allow for the direct quantitative comparison of different algorithms.
3. It is recommended that the results of the proposed k-NN pattern recognition model be fused with those of a time series model (e.g., Kalman Filter model). It is speculated that the time-series model can react more quickly to the variations exhibited in the data from the current day, particularly when the traffic is not changing states (e.g. moving from free flow to congested or vice versa). The fusion of the predictions provided by the k-NN and Time series model can be done using the proposed Markov Chain model.
4. The proposed prediction models are configured to provide predictions of travel time for each arterial segment independently. Following an approach similar to the trajectory method, the prediction models proposed in this thesis can be extended to predict travel times along a route consisting of multiple segments.
5. Bluetooth obtained data are considered as the only input for the developed SVM and Markov Chain models. The performance of these models can be enhanced using other informative lag-free features reflecting the prevailing roadway traffic conditions (e.g., estimate of delay for intermediate intersections – obtained from technologies like probe vehicles or connected vehicles).

6. The performance of the developed models is not empirically validated for another roadway section. However, we believe that at least some of the most important findings are transferable. For example, for the case of proposed k-NN model, it was hypothesized that the use of detection data, which do not suffer from a time lag as do measured travel times, would improve the accuracy of predictions. The results from the study section confirmed this hypothesis by demonstrating that statistically significant improvements in prediction accuracy are achieved from the inclusion of detection data. We have no reason to assume that this finding cannot be generalized to other arterial roadways.

   Regardless of the discussed capacity of the developed models, some of the elements of these models still need to be calibrated for other arterials. To address this issue it is recommended that the proposed models be applied to a different roadway (ideally, an arterial roadway with different traffic pattern and road classification) to investigate the transferability of the model to other locations.

References

Alessandri, A., A. di Febbraro, A. Ferrara, and E. Punta. 1999. Nonlinear optimization for freeway control using variable-speed signaling. Vehicular Technology, IEEE Transactions on 48 (6): 2042-52.

Antoniou, C. 2004. On–line calibration for dynamic traffic assignment. PhD Thesis, Faculty of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Bakula, C., W. H. Schneider IV, and J. Roth. 2012. Probabilistic model based on the effective range and vehicle speed to determine bluetooth MAC address matches from roadside traffic monitoring. Journal of Transportation Engineering 138 : 43.

Banks, J. H. 2002. Review of empirical research on congested freeway flow. Transportation Research Record: Journal of the Transportation Research Board 1802 : 225-32.

Barcelo, J., L. Montero, L. Marqués, P. Marinelli, and C. Carmona. 2010. Travel time forecasting and dynamic OD estimation in freeways based on bluetooth traffic monitoring. Paper presented at Transportation Research Board 89th Annual Meeting, .

Bierlaire, M., and F. Crittin. 2004. An efficient algorithm for real-time estimation and prediction of dynamic OD tables. Operations Research 52 (1): 116-27.

Billings, D., and J. S. Yang. 2006. Application of the ARIMA models to urban roadway travel time prediction-a case study.

Brinckerhoff, P. 2010. Synthesis of active traffic management experiences in europe and the united states.

Clark, S. D., S. Grant-Muller, and H. Chen. 2002. Cleaning of matched license plate data. Transportation Research Record: Journal of the Transportation Research Board 1804 : 1-7.

Cover, T. 1968. Estimation by the nearest neighbor rule. IEEE Transactions on Information Theory 14 (1): 50-5.

Daganzo, C. F. 1995. The cell transmission model, part II: Network traffic. Transportation Research Part B: Methodological 29 (2): 79-93.

Dasarathy, B. V. 1991. Nearest neighbor ({NN}) norms:{NN} pattern classification techniques.

Davis, G. A., N. L. Nihan, M. M. Hamed, and L. N. Jacobson. 1990. Adaptive forecasting of freeway traffic congestion. Transportation Research Record: Journal of the Transportation Research(1287): 29-33.

Day, C. M., R. Haseman, H. Premachandra, TM Brennan, J. Wasson, J. R. Sturdevant, and D. M. Bullock. 2010. Visualization and assessment of arterial progression quality using high resolution signal event data and measured travel time. Transportation Research Board Paper ID: 10-39.

Devijver, P. A., and J. Kittler. 1982. Pattern recognition: A statistical approach. London: Prentice/Hall International.

Dion, F., and H. Rakha. 2006. Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. Transportation Research Part B: Methodological 40 (9): 745-66.

Dion, F., H. Rakha, and Y. S. Kang. 2004. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. Transportation Research Part B: Methodological 38 (2): 99-122.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2012. Pattern classification. 2nd ed.John Wiley & Sons.

Fix, E., and JL Hodges. 1951. Discriminatory analysis: Nonparametric discrimination: Consistency properties. 4.

Fowkes, AS. 1983. The use of number plate matching for vehicle travel time estimation. Paper presented at PTRC Proceedings of the 11th Annual Conference, University of Sussex.

Freund, J. E., I. Miller, and M. Miller. 1999. John E. freund's mathematical statistics. 6th ed.Prentice Hall.

Fukunaga, K., and L. Hostetler. 1973. Optimization of k nearest neighbor density estimates. Information Theory, IEEE Transactions on 19 (3): 320-6.

Gault, H. E., and IG Taylor. 1981. The use of the output from vehicle detectors to assess delay in computer-controlled area traffic control systemsTransport Operations Research Group, University of Newcastle upon Tyne, Department of Civil Engineering.

Greenshields, BD. 1935. A study in highway capacity, highway research board. Paper presented at Proceedings of the Highway Research Board.

Haghani, A., M. Hamedi, K. F. Sadabadi, S. Young, and P. Tarnoff. 2010. Data collection of freeway travel time ground truth with bluetooth sensors. Transportation Research Record: Journal of the Transportation Research Board 2160 : 60-8.

Härdle, W. 1992. Applied nonparametric regression. Vol. 19Cambridge Univ Press.

Hellinga, B., and L. Fu. 1999. Assessing expected accuracy of probe vehicle travel time reports. Journal of Transportation Engineering 125 : 524-30.

Hellinga, B., P. Izadpanah, H. Takada, and L. Fu. 2008. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. Transportation Research Part C: Emerging Technologies 16 (6): 768-82.

Hellinga, B. R., and L. Fu. 2002. Reducing bias in probe-based arterial link travel time estimates. Transportation Research Part C: Emerging Technologies 10 (4): 257-73.

iTRANS. 2006. Cost of non-recurrent congestion in canada: Final report. Transport Canada, TP 14664E.

Izadpanah, P. 2010. Freeway travel time prediction using data from mobile probes. A Thesis Presented to the University of Waterloo in Fulfillment of the Thesis Requirement for the Degree of Doctor of Philosophy in Civil Engineering Department of Civil and Environmental Engineering Waterloo, Ontario, Canada.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82 (1): 35-45.

Kimber, RM, and E. M. Hollis. 1979. Traffic queues and delays at road junctionsTraffic Systems Division, Traffic Engineering Department, Transport and Road Research Laboratory.

Kudo, M., and J. Sklansky. 2000. Comparison of algorithms that select features for pattern classifiers. Pattern Recognition 33 (1): 25-41.

Kurzhanskiy, A. A., and P. Varaiya. 2010. Active traffic management on road networks: A macroscopic approach. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368 (1928): 4607.

Li, Ruimin, Geoffrey Rose, and Majid Sarvi. 2006. Using automatic vehicle identification data to gain insight into travel time variability and its causes. Transportation Research Record: Journal of the Transportation Research Board 1945 (1): 24-32.

Li, Y., and M. McDonald. 2002. Link travel time estimation using single GPS equipped probe vehicle. Paper presented at Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on.

Lin, W. H., A. Kulkarni, and P. Mirchandani. 2004. Short-term arterial travel time prediction for advanced traveler information systems. Journal of Intelligent Transportation Systems 8 (3): 143-54.

Liu, H. 2008. Travel time prediction for urban networks. PhD Thesis, Faculty of Civil Engineering and Geosciences, Transportation and Planning Section, Delft University of Technology, Delft, the Netherlands.

Liu, H. X., and W. Ma. 2007. Time-dependent travel time estimation model for signalized arterial network. Paper presented at Transportation Research Board 86th Annual Meeting.

Lo, H. K. 2001. A cell-based traffic control formulation: Strategies and benefits of dynamic timing plans. Transportation Science 35 (2): 148-64.

Lo, H. K. 1999. A novel traffic signal control formulation. Transportation Research Part A: Policy and Practice 33 (6): 433-48.

Lorenz, M., and L. Elefteriadou. 2000. A probabilistic approach to defining freeway capacity and breakdown. Paper presented at Fourth International Symposium on Highway Capacity.

Michalopoulos, P. G., G. Stephanopoulos, and G. Stephanopoulos. 1981. An application of shock wave theory to traffic signal control. Transportation Research Part B: Methodological 15 (1): 35-51.

Modsching, M., R. Kramer, and K. ten Hagen. 2006. Field trial on GPS accuracy in a medium size city: The influence of built-up. Paper presented at 3rd Workshop on Positioning, Navigation and Communication.

Montgomery, D. C., and G. C. Runger. 2010. Applied statistics and probability for engineers. 5th ed.Wiley.

Myers, K., and B. Tapley. 1976. Adaptive sequential estimation with unknown noise statistics. Automatic Control, IEEE Transactions on 21 (4): 520-3.

Nakata, T., and J. Takeuchi. 2004. Mining traffic data from probe-car system for travel time prediction. Paper presented at Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

Noroozi, Reza, and Bruce Hellinga. 2014. Real-time prediction of near-future traffic states on freeways using a markov model. Paper presented at Transportation Research Board 93rd Annual Meeting.

Oh, J. S., R. Jayakrishnan, and W. Recker. 2003. Section travel time estimation from point detection data. 82nd TRB Annual Meeting, Washington D.C., USA.

Park, D., L. R. Rilett, and G. Han. 1999. Spectral basis neural networks for real-time travel time forecasting. Journal of Transportation Engineering 125 (6): 515-23.

Persaud, B., S. Yagar, and R. Brownlee. 1998. Exploration of the breakdown phenomenon in freeway traffic. Transportation Research Record: Journal of the Transportation Research Board 1634 : 64-9.

Quayle, S. M., P. Koonce, D. DePencier, and D. M. Bullock. 2010. Arterial performance measures with media access control readers: Portland, oregon, pilot study. Transportation Research Board 89th Annual Meeting (CD-ROM), Transportation Research Board, Washington, D.C.

Robinson, S., and J. Polak. 2006. Overtaking rule method for the cleaning of matched license-plate data. Journal of Transportation Engineering 132 : 609.

Robinson, S., and J. W. Polak. 2005. Modeling urban link travel time with inductive loop detector data by using the k-NN method. Transportation Research Record: Journal of the Transportation Research Board 1935 : 47-56.

Roozemond, D. A. 2001. Using intelligent agents for pro-active, real-time urban intersection control. European Journal of Operational Research 131 (2): 293-301.

Rousseeuw, P. J., A. M. Leroy, and J. Wiley. 1987. Robust regression and outlier detection. Vol. 3Wiley Online Library.

Salek Moghaddam, S., R. Noroozi, J. M. Casello, and B. Hellinga. 2011. Predicting the mean and variance of transit segment and route travel times. Transportation Research Record: Journal of the Transportation Research Board 2217 : 30-37.

SHRP 2. 2009. L04 project: Incorporating reliability performance measures in operations and planning modeling tools - task 1: Fundamental issues of incorporation of travel time reliability in modeling tools. Prepared for the Strategic Highway Research Program 2, Transportation Research Board of the National Academies, (Preliminary Draft Technical Memorandum).

Simonoff, J. S. 1996. Smoothing methods in statistics. New York: Springer Verlag.

Sisiopiku, V. P., N. M. Rouphail, and A. Santiago. 1994. Analysis of correlation between arterial travel time and detector data from simulation and field studies. Transportation Research Record: Journal of the Transportation Research Board(1457): 166-73.

Sisiopiku, V. P., A. J. Sullivan, G. Fadel, University of Alabama at Birmingham. Dept. of Civil, Construction, and Environmental Engineering, and University Transportation Center for Alabama. 2009. Implementing active traffic management strategies in the USUniversity Transportation Center for Alabama.

Skabardonis, A., and N. Geroliminis. 2005. Real-time estimation of travel times on signalized arterials. Paper presented at Proceedings of the 16th International Symposium on Transportation and Traffic Theory.

Takaba, S., T. Morita, T. Hada, T. Usami, and M. Yamaguchi. 1991. Estimation and measurement of travel time by vehicle detectors and license plate readers. Paper presented at Vehicle Navigation and Information Systems Conference, 1991, .

TRB. 2010. Highway capacity manual. 5th ed. Washington D.C.: Transportation Research Board (TRB), National Research Council.

Tu, H. 2008. Monitoring travel time reliability on freeways. PhD Thesis, Faculty of Civil Engineering and Geosciences, Transportation and Planning Section, Delft University of Technology, Delft, the Netherlands.

Ulmer, J. M. 2003. Freeway traffic flow rate measurement: Investigation into impact of measurement time interval. Journal of Transportation Engineering 129 : 223-9.

University of Maryland. 2008. Bluetooth traffic monitoring technology:Concept of operation & deployement guidlines. Center for Advanced Transportation Technology.

Van Boxel, D., IV Schneider, and C. Bakula. 2011. Innovative real-time methodology for detecting travel time outliers on freeways and urban arterials. Paper presented at Transportation Research Board 90th Annual Meeting.

van Hinsbergen, C. P. I., A. Hegyi, JWC van Lint, and HJ van Zuylen. 2009. Application of bayesian trained neural networks to predict stochastic travel times in urban networks. Paper presented at 16th ITS World Congress and Exhibition on Intelligent Transport Systems and Services.

van Hinsbergen, C. P. I. J., and J. W. C. van Lint. 2008. Bayesian combination of travel time prediction models. Transportation Research Record: Journal of the Transportation Research Board 2064 : 73-80.

van Hinsbergen, C. P. I. J., J. W. C. van Lint, and HJ van Zuylen. 2009. Bayesian training and committees of state-space neural networks for online travel time prediction. Transportation Research Record: Journal of the Transportation Research Board 2105 : 118-26.

Van Lint, JWC. 2004. Reliable travel time prediction for freeways. PhD Thesis, Faculty of Civil Engineering and Geosciences, Transportation and Planning Section, Delft University of Technology, Delft, the Netherlands.

Van Lint, JWC, SP Hoogendoorn, and HJ Van Zuylen. 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transportation Research Part C: Emerging Technologies 13 (5-6): 347-69.

Viti, F. 2006. The dynamics and the uncertainty of delays at signals. PhD Thesis, Faculty of Civil Engineering and Geosciences, Transportation and Planning Section, Delft University of Technology, Delft, the Netherlands.

Wang, Y., M. Papageorgiou, and A. Messmer. 2003. Predictive feedback routing control strategy for freeway network traffic. Transportation Research Record: Journal of the Transportation Research Board 1856 : 62-73.

Webster, FV. 1958. Traffic signal settings. road research technical paper no. 39. Road Research Laboratory, Her Majesty's Stationery Office, Berkshire, England.

Welch, G., and G. Bishop. 2006. An introduction to the kalman filter. University of North Carolina at Chapel Hill, Chapel Hill, NC.

Wu, X., H. X. Liu, and N. Geroliminis. 2011. An empirical analysis on the arterial fundamental diagram. Transportation Research Part B: Methodological 45 : 255-66.

Yang, J. S. 2007. Application of the kalman filter to the arterial travel time prediction: A special event case study. Control and Intelligent Systems 2007 35 (1): 79-85.

Yoon, J., B. Noble, and M. Liu. 2007. Surface street traffic estimation. Paper presented at Proceedings of the 5th international conference on Mobile systems, applications and services.

You, J., and T. J. Kim. 2000. Development and evaluation of a hybrid travel time forecasting model. Transportation Research Part C: Emerging Technologies 8 (1-6): 231-56.

Zheng. 2011. Modelling urban travel times. PhD Thesis, Faculty of Civil Engineering and Geosciences, Transportation and Planning Section, Delft University of Technology, Delft, the Netherlands.

Appendix A
Components of Arterial Travel Time

Mid-link travel time

Free flow travel time is a function of distance and posted speed limit but also can be affected by other factors such as traffic mix, lane geometry (e.g. lane width), horizontal curvature, vertical grade, intersection spacing, weather, road surface condition, etc.

On arterials, mid-link delay seldom occurs as a result of traffic flow. This is mainly due to the fact that generally on arterials midblock flow is lower than the capacity of the link, which is controlled by the upstream signal. Near-linear relationship between flow and density on arterial links, suggesting the near-constant speed of vehicles, has been supported by different researchers (Wu et al. 2011; Banks 2002; Lin et al. 2004). In fact, other factors like bus maneuvers at bus stops, parking vehicles, crossing pedestrians and cyclists, and turning vehicles from cross streets have much greater contribution to mid-link delay on arterials than the volume.

Intersection delay

The most important component of arterial travel time is the intersection delay (signal delay). Stochastic arrivals and departures along with stochastic flow rate can cause large variations in delays that vehicles may experience at signalized intersections during oversaturated and even under-saturated conditions.

Signal control effect

The effect of signal control on the intersection delay is depicted in Figure A-1 (replication of Figure 1-7). In this figure vehicle A which has arrived at the intersection just seconds before the end of the green time, was able to proceed through the intersection without delay as there was no unmet traffic demand at the signal. On the other hand vehicle B which has arrived just a few seconds after vehicle A, arrived at the beginning of the red interval and has to wait intersection at almost the same time may experience very different levels of delays.

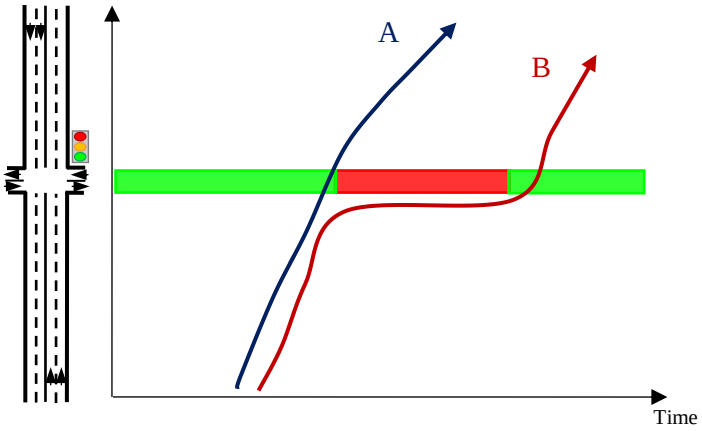


Figure A-1: The effect of signal control and arrival time on intersection delay

For under-saturated conditions the experienced delay msay vary between a maximum of red phase duration and a minimum of zero. For the case of deterministic arrival times, this variation is linear (Figure A-2), and the delay is a function of just arrival time and signal control characteristics. However, for oversaturated conditions the experienced delay will also be affected by the existing queue at the beginning of the red interval (initial queue). The delay variation for oversaturated conditions (again assuming deterministic arrival and service times) is depicted in Figure A-3.

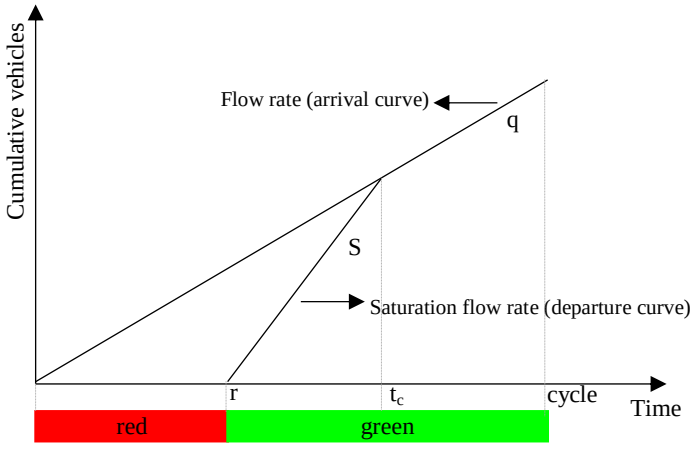


Figure A-2: Variation in intersection delay due to arrival time and signal control characteristics (Under-saturated conditions – Deterministic arrival)

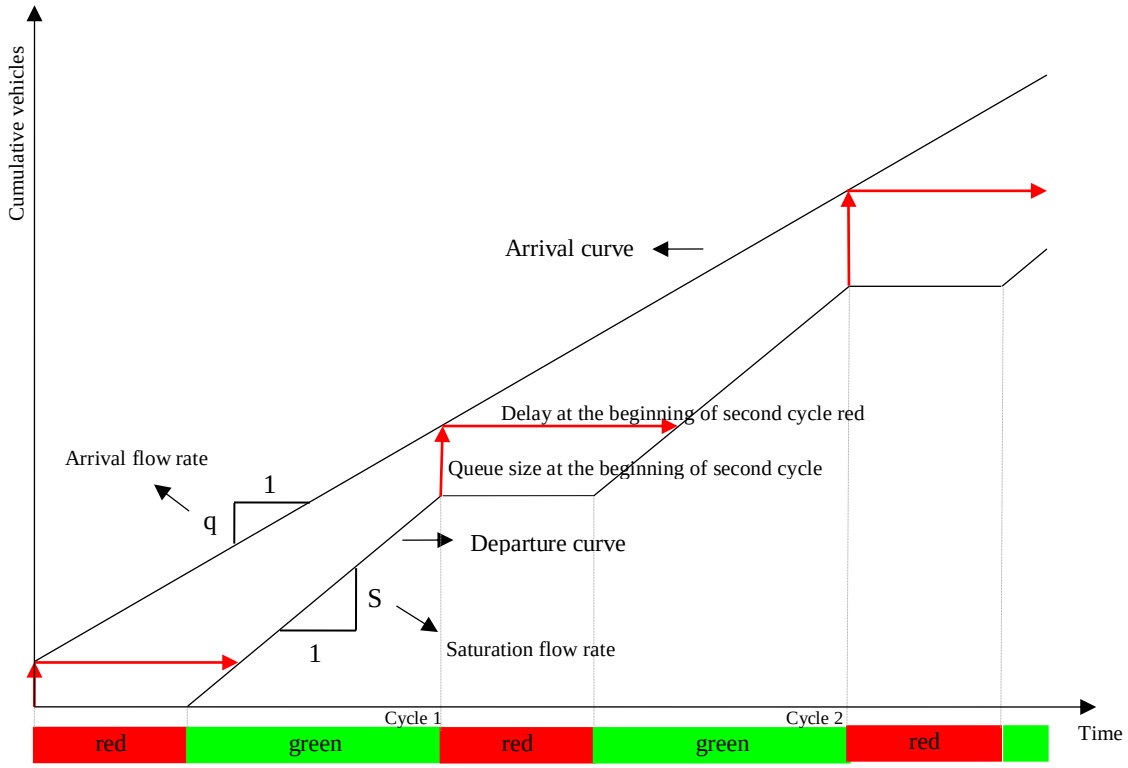


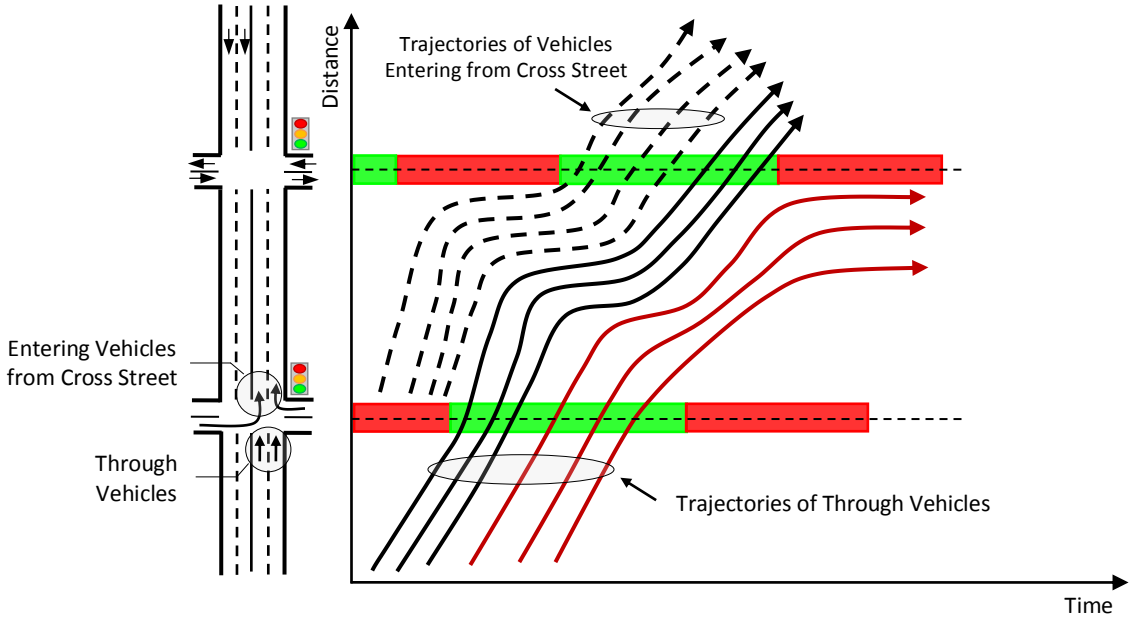
Figure A-3: Variation in intersection delay due to arrival time and signal control characteristics and initial queue (Oversaturated conditions – Deterministic arrival)

Coordination effect

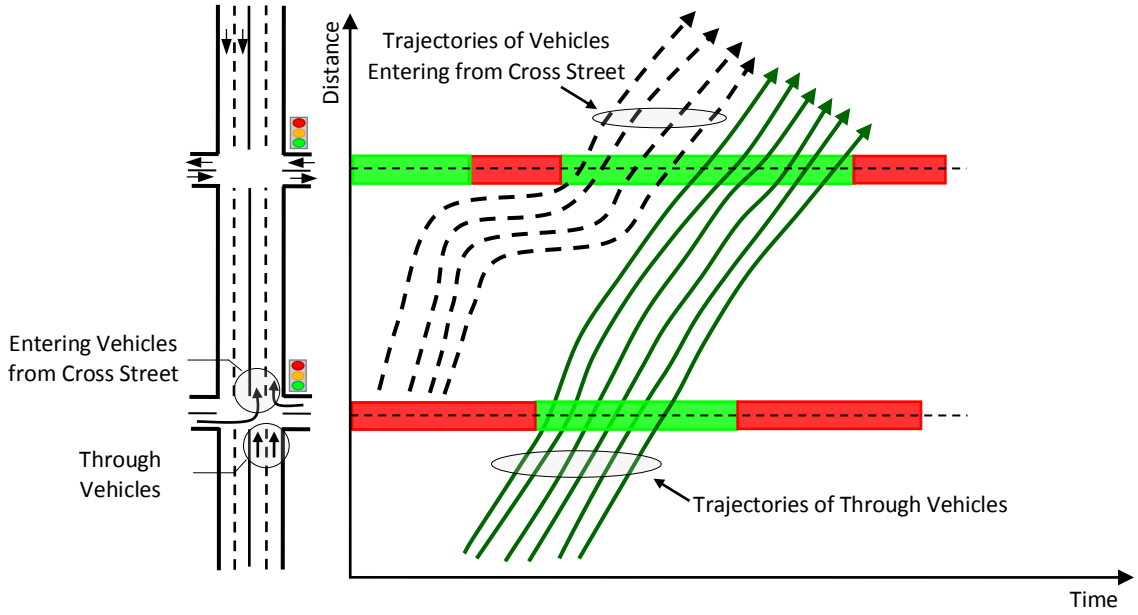
The level of signal coordination along an arterial is another important factor which can affect the travel time of individual vehicles traveling along that corridor. On well-coordinated corridors the timing schemes of signalized intersections have been set to ensure that through vehicles will arrive when the traffic signal is green and the initial queue has been dissipated. The distance between adjacent intersections and the flow rate of upstream intersection turning movements are the most important factors that should be considered in signal coordination calculations. Consequently, more coordination in traffic signal system leads to lower intersection delay along the corridor (coordinated arrival times).

The effect of signal coordination on intersection delay can be more clarified through an example. Figure A-4a shows a poorly coordinated traffic signal, as through vehicles are required to join the queue of existing upstream intersection turning vehicles and wait for a

while before being able to proceed through the intersection. On the other hand, in the well-coordinated traffic signal depicted in Figure A-4b through vehicles will reach the downstream signal after the queue has dissipated.



(a): Poorly Coordinated Corridor



(b): Well-Coordinated Corridor

Figure A-4: The effect of corridor coordination on intersection delay

Signal timing variation over time

Other than the described factors affecting travel time on arterials (i.e. mid-link travel time & intersection delay), the changes in signal splits (i.e. green & red phases) over time can also affect the calculated average travel time of different time intervals. In fact, due to signal split changes, different time intervals may not contain the same numbers of signal cycles. This is illustrated through Figure A-5. In this figure, the cycle length has remained the same between different time intervals and just the durations of red and green intervals have been changed (as a result of left turn signal actuation). Based on this figure, it can be observed that changes in signal splits per se can cause variability in average delay in different intervals.

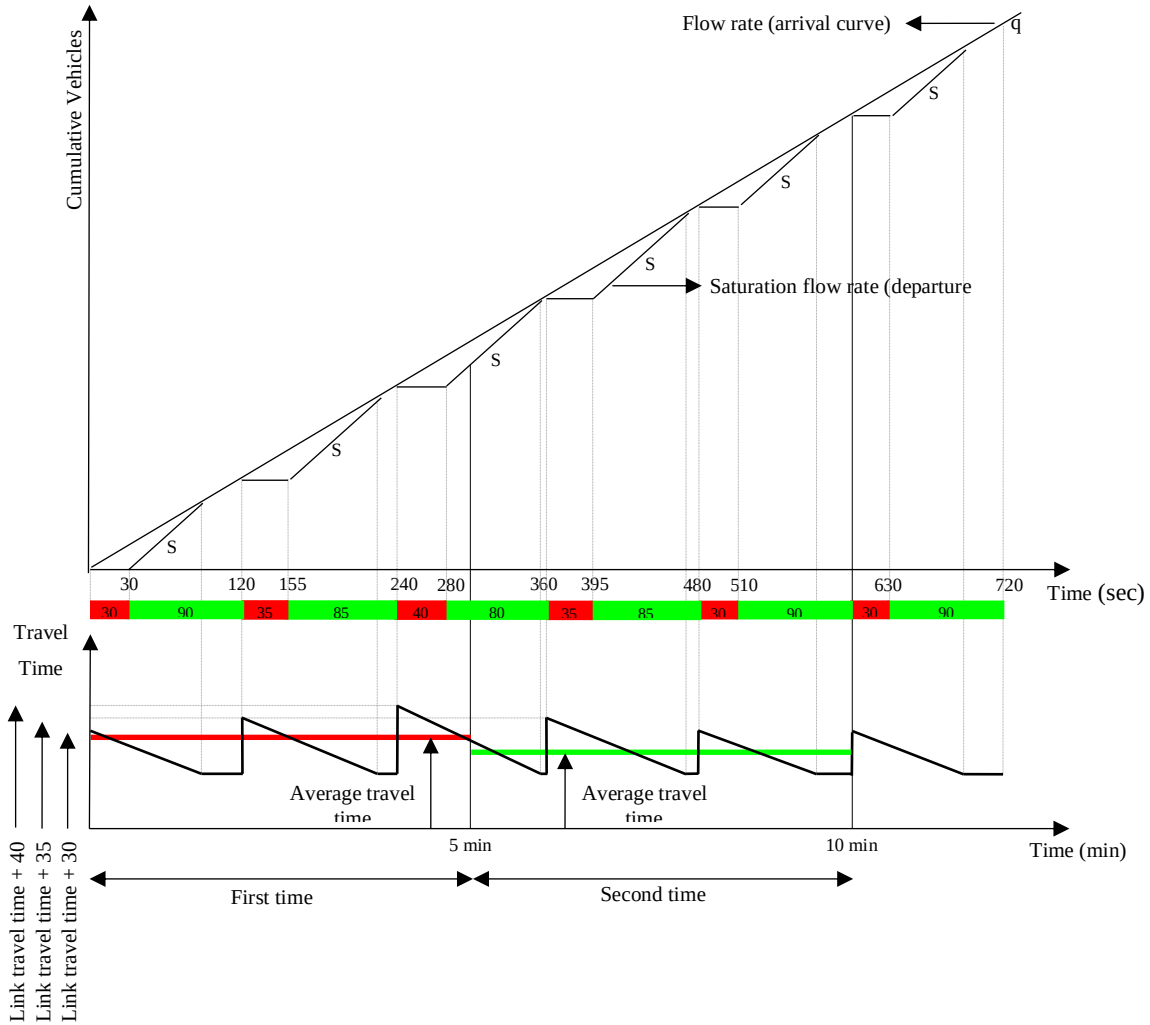


Figure A-5: Variability in average travel time for through movements due to changes in red and green time

Appendix B

Sample Travel Time Prediction Results

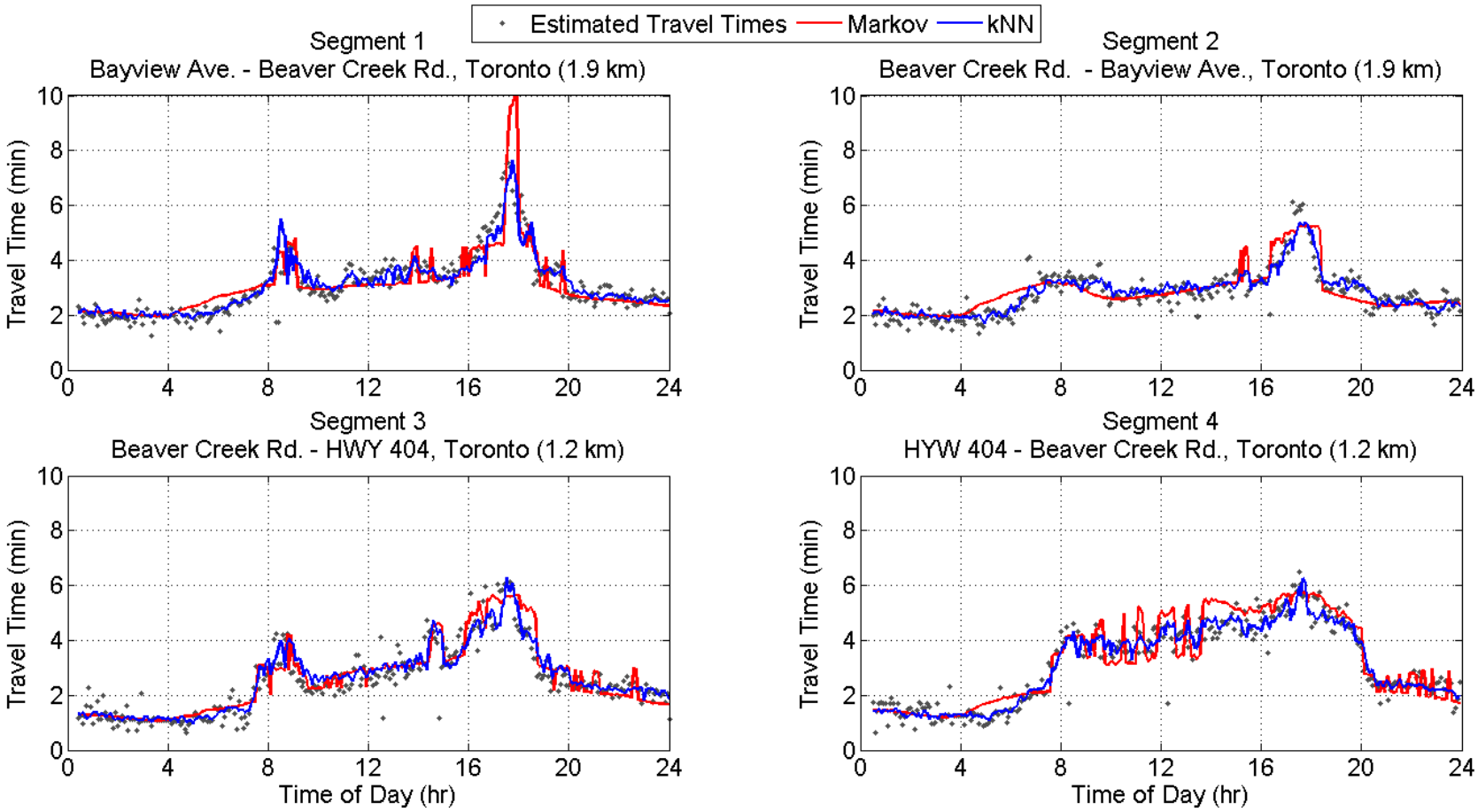


Figure B-1a: Travel time prediction - Markov model vs. k-NN model

(Sample day: Friday, November 11th, 2011)

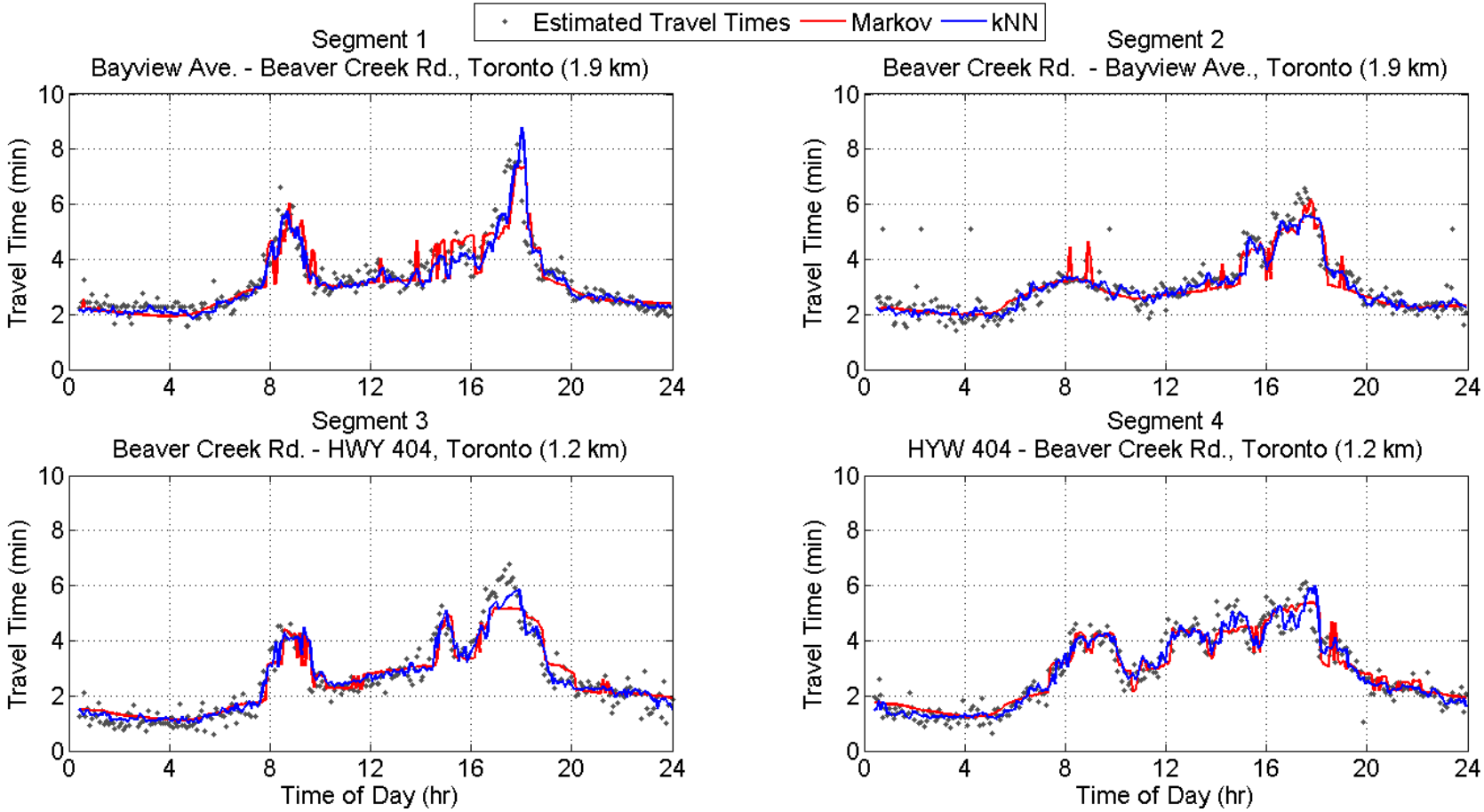


Figure B-1b: Travel time prediction - Markov model vs. k-NN model

(Sample day: Wednesday, December 14th, 2011)

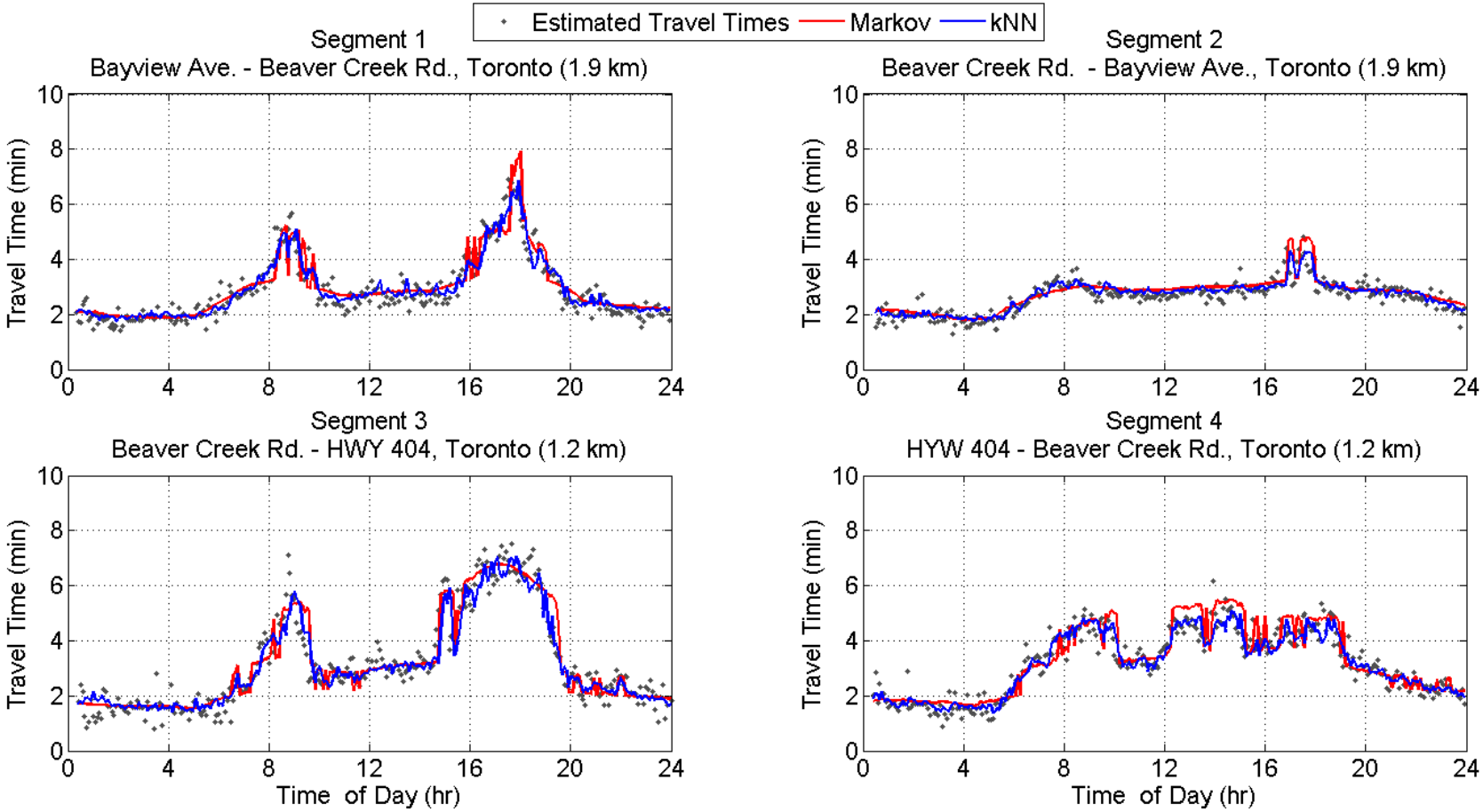


Figure B-1c: Travel time prediction - Markov model vs. k-NN model

(Sample day: Thursday, February 16th, 2012)

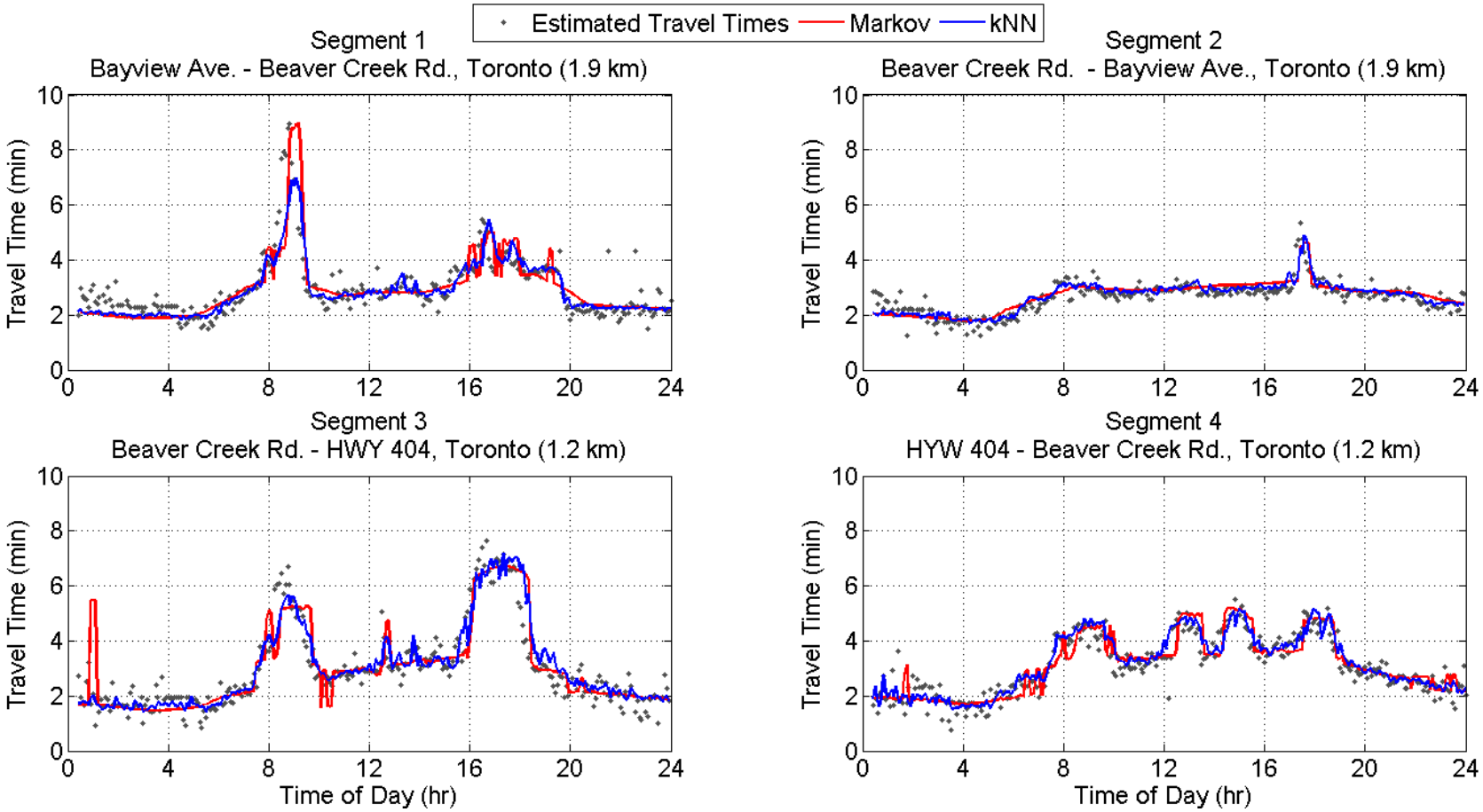


Figure B-1d: Travel time prediction - Markov model vs. k-NN model

(Sample day: Thursday, March 8th, 2012)

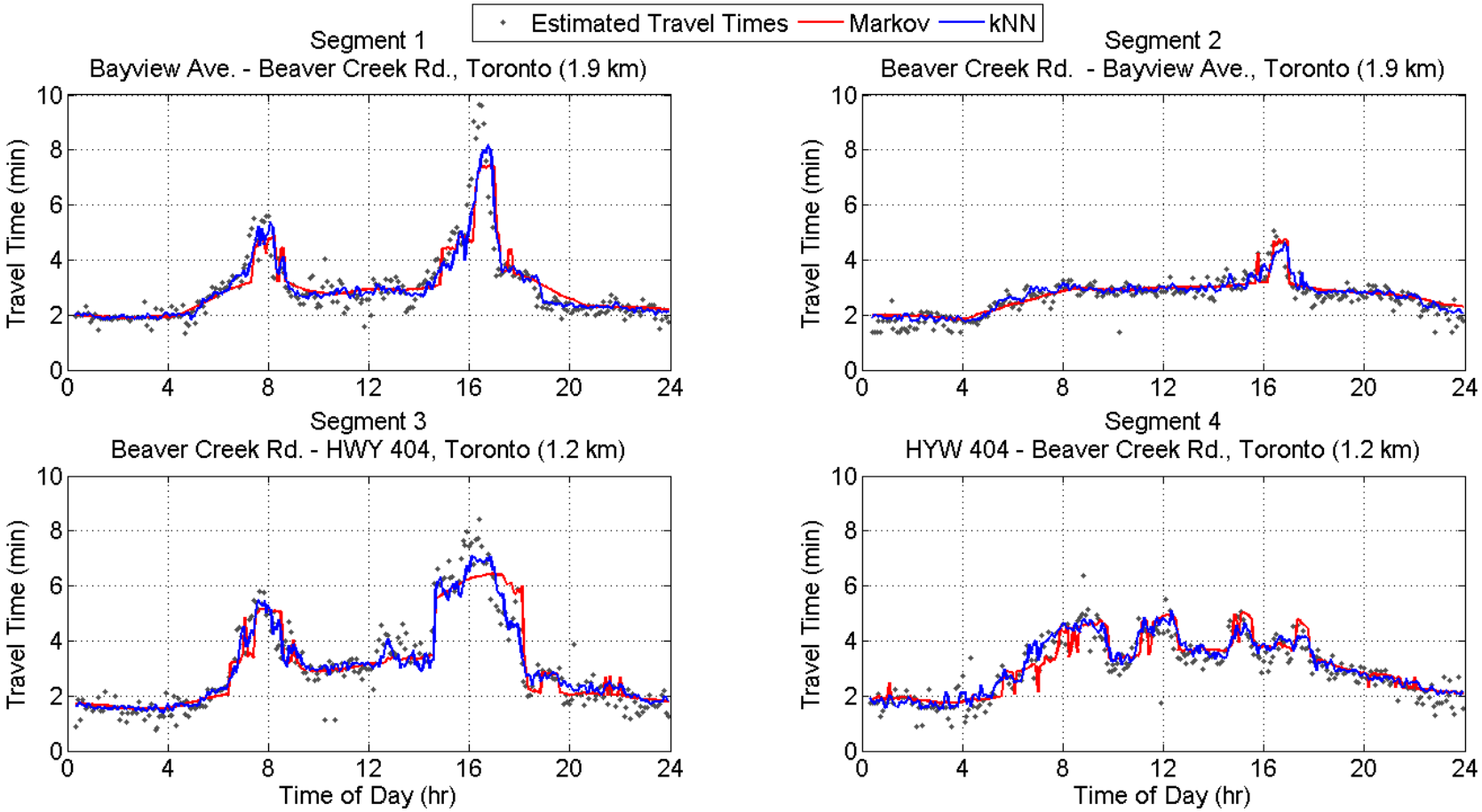


Figure B-1e: Travel time prediction - Markov model vs. k-NN model

(Sample day: Thursday, March 22nd, 2012)

General Symbol List

$[cnt]^{(i,k)}_{1\times((n_k+1)\cdot(n_i+1))}$:Counted trips vector during recent time intervals for segment *i*

$[DT]^{(i,k)}_{1\times(2\times(n_k+1))}$: Number of detections vector during recent time intervals at the upstream and downstream detectors of segment *i*

$e^{(\,)}$: Matrix exponential operator

E(): Mathematical expectation operator

$E_{(FV^H)}$: Expected value of the feature variables over the historical dataset

$f(t \mid State_{(k-\Delta)} = m)$:

Empirical probability density function of travel time for next time interval *(k+1)* given the traffic during the last available time interval *(k-Δ)* is in state *m*

$f(t \mid State_{(k-\Delta)} = m \;\&\; State_{(k+1)} = n)$:

Empirical probability density function of travel time for next time interval *(k+1)* given the traffic during the last available time interval *(k-Δ)* is in state *m* and the traffic during the next time interval is in state *n*.

$F_d^{Correct}$: Number of free-flow observations which are classified correctly

F_d^{All} : Number of free-flow observations

$[FV]^{(i,k)}$: Feature vector for segment i at the end of time interval k

$[\overset{SD}{FV}]^{(i,k)}$: Standardized feature vector for segment *i* at the end of time interval *k*

Φ(*x*): Mapping function projecting the feature vector *x* to a higher dimension

i: Investigated segment

$i-n_i$: Most upstream considered segment

k: Last time interval before present time

$k+1$: Next time interval after present time

$k-\Delta$: Time interval corresponding to the latest available observed travel times

$\mathrm{K}(.,.)$: Radial basis function kernel

L_i: Length of segment i

LOS: Prevailing level of service on the roadway

$MD_{(i,k)}$: Mahalanobis distance between a historical record and current prediction instance for segment *i* at the end of time interval *k*

$MMCR_F$: Daily mean misclassification rates for free-flow observations

$MMCR_{TC}$: Daily mean misclassification rates for transient-congested observations

$MMCR_{Overal}$: Daily mean misclassification rates for all observations

n_i: Number of considered upstream segments

n_k+1: Numbers of considered prior time intervals

$n^h_{(i,k+1)}$: Number of valid individual travel times which have been utilized to calculate $t^h_{(i,k+1)}$

$n^v_{(i,k)}$: Number of valid travel time observations for segment *i* during time interval *k*

n_{NN}: Number of nearest neighbors in the k-NN model

NN: Set of nearest neighbors

$pr^{mn}_{(k+1)}$: Transition probability from traffic state *m* to traffic state *n* for next time interval *(k+1)*

P_B: Bluetooth penetration rate

P_e: Fraction of vehicles that make en-route stops

P_m: Fraction of the number of true travel times for which two measured travel times are obtained.

P_T: Percent of buses in the traffic stream

$P_{(K)}^{F.D}$: Probability of the first detection during the K^{th} polling interval

Q: Markov model transition intensity matrix

RTTI: Daily relative travel time improvement index

$[SH]_{1\times(2\times(n_K+1))}^{(i,k)}$: Summation of hits vector during recent time intervals at the upstream and downstream detectors of segment *i*

$State_{(i,k)}$: Traffic state for segment *i* during time interval *k*

$\sigma^2(\tau_{(i,k)})$: Fused variance of individual travel times for segment *i* during time interval *k* (for the proposed outlier detection algorithm)

$\sigma^2(t_{(i,k)})$: Variance of travel times for segment *i* during time interval *k*

$\overset{N}{\sigma}(\tau_{(i,k)})$: Standard deviation for the corresponding normal distribution when the fused travel time is distributed log-normally (segment *i* during time interval *k* - for the proposed outlier detection algorithm)

$\hat{\sigma}^2(t_{(i,k+1)})$: Predicted variance of the individual travel times for segment *i* during next time interval *(k+1)*

$\sum_{(FV^H)}$: Variance – covariance matrix of historical feature variables

$t_{(i,k)}$: Average observed travel time on segment i during time interval k using incomplete travel time set

$t_{(i,k)}^{detected}$: Average of all valid measurements (i.e. excluding measurements labeled as outliers) on segment *i* and within interval *k*

$t_{(i)}^{f}$: Free-flow travel time for segment *i*

$t_{(i,k+1)}^{h}$: Observed average travel time on segment *i* during next time interval *(k+1)* for historical record *h*

t_{ij}^{k}: Travel time for j^{th} vehicle entering segment i during time interval k

$t_{(i,k)}^{outlier}$: Average of measured travel times including measurement error and outlier error on segment *i* and within interval *k*

$t_{(i,k)}^{Smoothed}$: Smoothed travel time on segment *i* during time interval *k*

$t_{(i,k)}^{true}$: Average of the true travel times including no error on segment *i* and within interval *k*

$\hat{t}_{(i,k+1)}$: Predicted average travel time on segment *i* during next time interval *(k+1)*

$[t]_{1\times((n_k+1)\cdot(n_i+1))}^{(i,k)}$: Average travel time vector during recent time intervals for segment *i*

$TC_d^{Correct}$: Number of transient-congested observations which are classified correctly

TC_d^{All}: Number of transient-congested observations

$\tau_{(i,k)}$: Fused mean travel time for segment *i* during time interval *k* (for the proposed outlier detection algorithm)

$\tau_{(i,k)}^{N}$: Mean value for the corresponding normal distribution when the fused travel time is distributed log-normally (segment *i* during time interval *k* - for the proposed outlier detection algorithm)

v/c: Volume over capacity ratio

$Var_{(FV^H)}$: Variance of the feature variables over the historical dataset

λ_{ij}: Transition intensity between state *i* and state *j*

λ_{ij}^{base}: Base transition intensity between state *i* and state *j*