DEVELOPING SAMPLING STRATEGIES AND PREDICTING FREEWAY TRAVEL TIME USING BLUETOOTH DATA

by

Hasan M Moonam

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Engineering

at

The University of Wisconsin-Milwaukee

December 2016

ABSTRACT

DEVELOPING SAMPLING STRATEGIES AND PREDICTING FREEWAY TRAVEL TIME USING BLUETOOTH DATA

by

Hasan M Moonam

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professor Xiao Qin

Accurate, reliable, and timely travel time is critical to monitor transportation system performance and assist motorists with trip-making decisions. Travel time is estimated using the data from various sources like cellular technology, automatic vehicle identification (AVI) systems. Irrespective of sources, data have characteristics in terms of accuracy and reliability shaped by the sampling rate along with other factors. As a probe based AVI technology, Bluetooth data is not immune to the sampling issue that directly affects the accuracy and reliability of the information it provides. The sampling rate can be affected by the stochastic nature of traffic state varying by time of day. A single outlier may sharply affect the travel time. This study brings attention to several crucial issues - intervals with no sample, minimum sample size and stochastic property of travel time, that play pivotal role on the accuracy and reliability of information along with its time coverage. It also demonstrates noble approaches and thus, represents a guideline for researchers and practitioner to select an appropriate interval for sample accumulation flexibly by set up the threshold guided by the nature of individual researches' problems and preferences.

After selection of an appropriate interval for sample accumulation, the next step is to estimate travel time. Travel time can be estimated either based on arrival time or based on

departure time of corresponding vehicle. Considering the estimation procedure, these two are defined as arrival time based travel time (ATT) and departure time based travel time (DTT) respectively. A simple data processing algorithm, which processed more than a hundred million records reliably and efficiently, was introduced to ensure accurate estimation of travel time. Since outlier filtering plays a pivotal role in estimation accuracy, a simplified technique has proposed to filter outliers after examining several well-established outlier-filtering algorithms.

In general, time of arrival is utilized to estimate overall travel time; however, travel time based on departure time (DTT) is more accurate and thus, DTT should be treated as true travel time. Accurate prediction is an integral component of calculating DTT, as real-time DTT is not available. The performances of Kalman filter (KF) were compared to corresponding modeling techniques; both link and corridor based, and concluded that the KF method offers superior prediction accuracy in link-based model. This research also examined the effect of different noise assumptions and found that the steady noise computed from full-dataset leads to the most accurate prediction. Travel time prediction had a 4.53% mean absolute percentage of error due to the effective application of KF.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Alhamdulillah! Praise be to Allah for the successful completion of my thesis.

I would like to render my profound gratitude to Professor Xiao Qin for his invaluable guidance, extreme support, and inspiring words throughout the research work. He is one of the sincerest people I have ever seen in my life and I feel proud to be his student. I also want to thank Professor Yue Liu and Professor Jun Zhang for their time and guidance as the examining committee.

I am thankful to Elizabeth Schneider from the Wisconsin Department of Transportation as well, she kind heartedly agreed to hold a couple of meetings and provided me with the data for conducting this research.

I am truly grateful to my colleague Zhi Chen for his extraordinary suggestions. I would also like to acknowledge Mohammad Razaur Rahman Shaon and Zhaoxiang He for their valuable comments. Considering this is an opportunity; I would like to thank my friends and well-wishers who has been supporting and encouraging me throughout my journey in the United States.

Finally, I would like to express my sincerest thanks to my parents and family members, including a very special friend, for their unconditional love and support.

CHAPTER I. INTRODUCTION

Rapid expansion of cities followed by urban agglomeration has been a very common phenomenon witnessed in the U.S. in recent years. Agglomeration economies have created mega-regions, produced new employment opportunities, and stimulated urban economic development, while attracted more people to cities. As the majority of population of a country is now living in cities, providing sufficient infrastructures for people to live and travel is a daunting task (Edwards and Smith, 2008). The shortage of needed residential, commercial, and transportation infrastructure is further aggravated by improper planning and difficulty of determining employment, population, transportation and infrastructure growth (Duranton and Turner, 2012). Economic outlook and population growth of a city has a profound impact on the development and use of its transportation systems and transportation infrastructure. According to the U.S. Department of Transportation, in 2014, total highway travel was 3,025,656 million vehicle miles or 4,371,706 million passenger miles, which was slightly higher than the previous years in spite of the nationwide economic recession (National Transportation Statistics, 2016). According to the Urban Mobility Report 2015, total national congestion delay has been increased every year (except 2008 and 2009) since 1982 (Schrank et al., 2015). FIGURE 1 shows the trend in national congestion from 1982 to 2014.

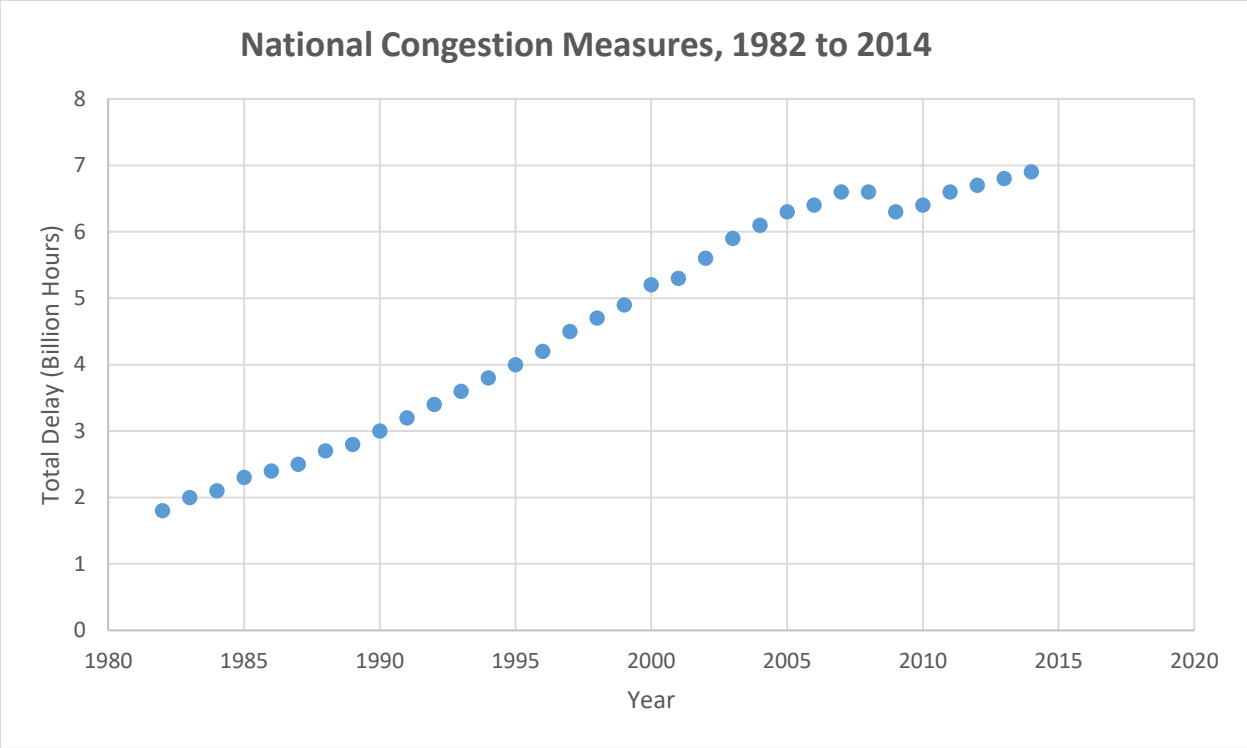


FIGURE 1 Trend of national congestion from 1982 to 2014.

(Data source: Urban Mobility Scoreboard 2015)

The urban areas can be divided into four groups based on population size, - very large with 3 million and over population, large with 1 million to less than 3 million populations, medium with 500,000 to less than 1 million populations and small with less than 500,000 populations. According to Bureau of Transportation Statistics (BTS), travel time indices were reported to be 1.32, 1.23, 1.18 and 1.14 for these 4 types of urban areas with a congestion cost of 5259, 1281, 474 and 191 million dollars, respectively in year 2014. The Travel Time Index is defined as the ratio of travel time in the peak period to the travel time at free-flow conditions. Therefore, a value of 1.32 indicates that a trip time in the peak hour would be 1.32 times of the free-flow travel time (National Transportation Statistics, 2016). Obviously, the larger the urban areas, greater is the peak hour travel time, as well as the congestion cost. However, congestion is not just a big city problem; it is

worse in areas of every size (Schrank et al., 2015). For different population sized cities, the following figure shows the congestion growth trend in hours of delay per auto commuter:

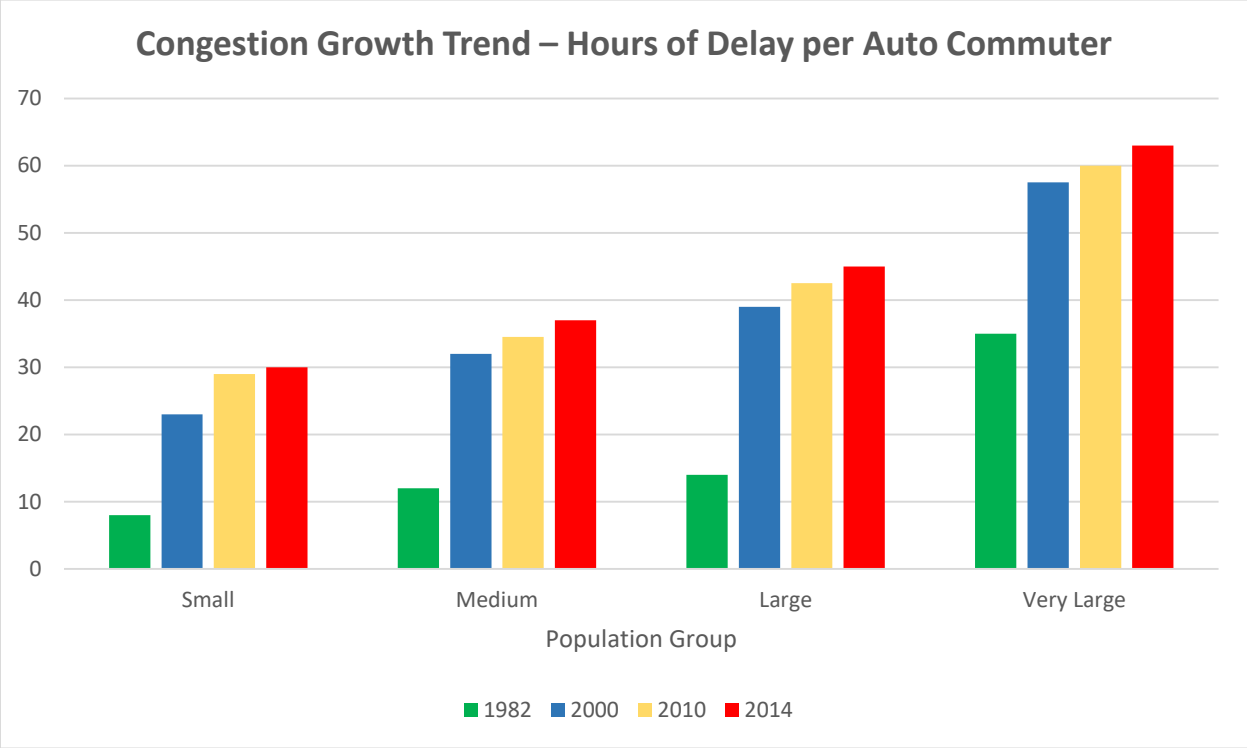


FIGURE 2 Congestion growth trend in different population sized cities.

(Source: Urban Mobility Scoreboard 2015)

Congestion means extra travel time and the uncertainty of travel, which adversely affects the daily routine of travelers. It is important to notify travelers of the probable travel time so that they can choose alternative routes and/or re-schedule their departure times and activities accordingly. Travel time, an important component of Advanced Traveler Information System (ATIS), is a key factor to determine the travel decisions in response to unexpected delays (Khattak et al., 1996). As a part of Intelligent Transport System (ITS), ATIS performs two basic functions, routing and navigation through continuous monitoring of transportation system and evaluation of its performance. Integrated with a traffic data collector- Advanced Transportation Management System (ATMS), ATIS provides information about recurrent and non-recurrent congestions (Qiu

and Cheng, 2007) which benefits individual travelers and a road network as a whole, especially at the time of traffic congestions (Levinson, 2003). Besides the measurement of transportation system performance, travel time has been used to assist to predict travel time and traffic state, which helps the traffic operations room in versatile ways.

In general, ITS integrates advanced communications, control, electronics, and computer hardware and software technologies into the transportation infrastructure and in vehicles (Sussman et al., 2000) and improves surface transportation system performance by using its electronic surveillance, communications, and traffic analysis and control technologies with various sensor technologies- inductive-loop detector, magnetic detector, video image processor, microwave sensor, and infrared sensor (Handbook, 2006). ATIS services include dynamic message signs, dynamic information (e.g. real-time travel times and congestion information), in-vehicle navigation systems and dynamic route guidance to reduce trip uncertainty (Sussman et al., 2000). In ATIS, fixed point traffic sensors, cellular geo-location technologies (Sussman et al., 2000), automatic vehicle identification (AVI) systems (e.g. Bluetooth readers, electronic toll collection tags, license plate readers, and signature re-identification based on detector or magnetometer measurement) are being used to measure and/or estimate the travel time (Xiao et al., 2014). *"Automatic vehicle identification (AVI) data allows individual travelers to be time stamped at different parts of the transportation system, enabling the quality of service to be measured by calculating travel times between pairs of points in the system"* (Day et al., 2012). Amongst all the available techniques, Bluetooth has emerged as one of the fastest growing data collection technologies whose market share is continuing to rise, mainly due to its cost effectiveness. Bluetooth is a probe based (Puckett and Vickich, 2010) AVI technique for collecting travel time data. Each Bluetooth device contains a unique electronic identifier known as a Media Access

Control (MAC) address. By mounting a simple antenna adjacent to the roadway, MAC addresses of other devices in range can be logged. When a logged MAC address matches at two consecutive stations, the difference in logging time stamps is used to estimate travel times and the corresponding traffic speed (Bachmann et al., 2013). The process of Bluetooth detection is illustrated in **FIGURE *3***:

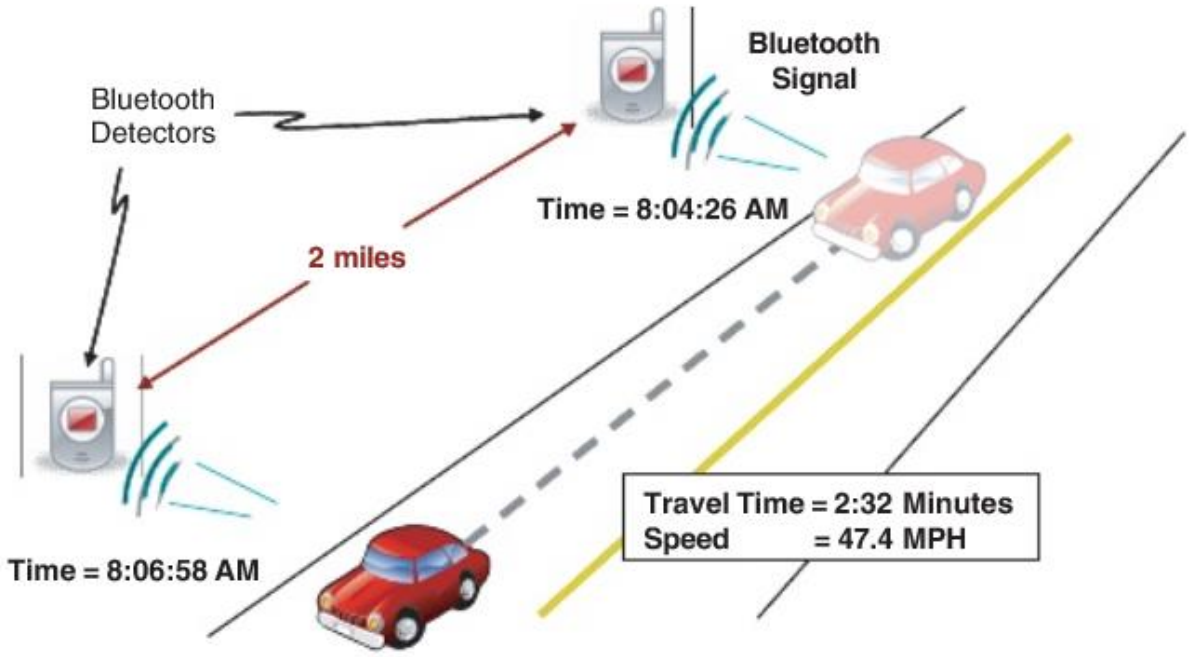


FIGURE 3 Process of Bluetooth traffic monitoring (Haghani et al., 2010).

Travel Time and its Estimation: Arrival vs. Departure time based travel time

Travel time is measured by the time elapsed when a traveler moves between two distinct spatial positions (Carrion and Levinson, 2012). Over a decade, various studies have attempted to define travel time estimation (e.g. instantaneous travel time, experienced travel time (Xiao et al., 2014), and predicted travel time (Bhaskar et al., 2011)), but the definitions lack clarity and have created confusion and inconsistency in data collection and analysis (Bhaskar et al., 2011, Toppen and

Wunderlich, 2003). Recently, however, one study made clear the distinction between arrival time-based travel time (ATT) and departure time-based travel time (DTT) (Kim et al., 2009). ATT and DTT have two different estimation algorithms for real time computation. ATT refers to the travel time associated with arrival at the destination, while DTT refers to the travel time associated with departure from the origin. Hence, a vehicle starting at 8:30am from place A and reaching place B at 9:00am will yield the ATT ($ATT_{AB@9:00AM}$) and DTT ($DTT_{AB@8:30AM}$) of link AB at 9:00am and 8:30am, respectively. In practice, ATT and DTT are link-dependent, not vehicle-dependent, as they refer to the link-based travel time. ATT and DTT may available simultaneously if many vehicles travel a link continuously. Practitioners usually treat ATT as the travel time in an ATIS due to the lack of available DTT; however, this reported ATT is one-step (step interval = travel time) earlier than the actual travel time to be experienced (DTT) by drivers. Although ATT and DTT differ slightly in a free-flow condition, the difference can sharply escalate at the onset and end of traffic congestion. FIGURE 4 shows the difference between ATT and DTT from free-flow to onset of congestion for a link of 3.1 mile with a free-flow travel time of 202 seconds.

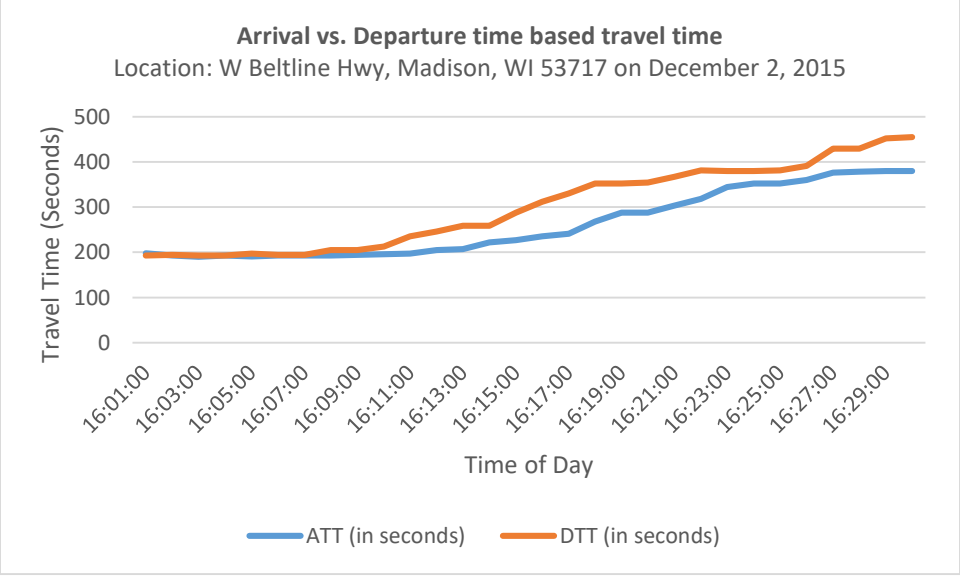


FIGURE 4 Arrival vs. Departure time based travel time.

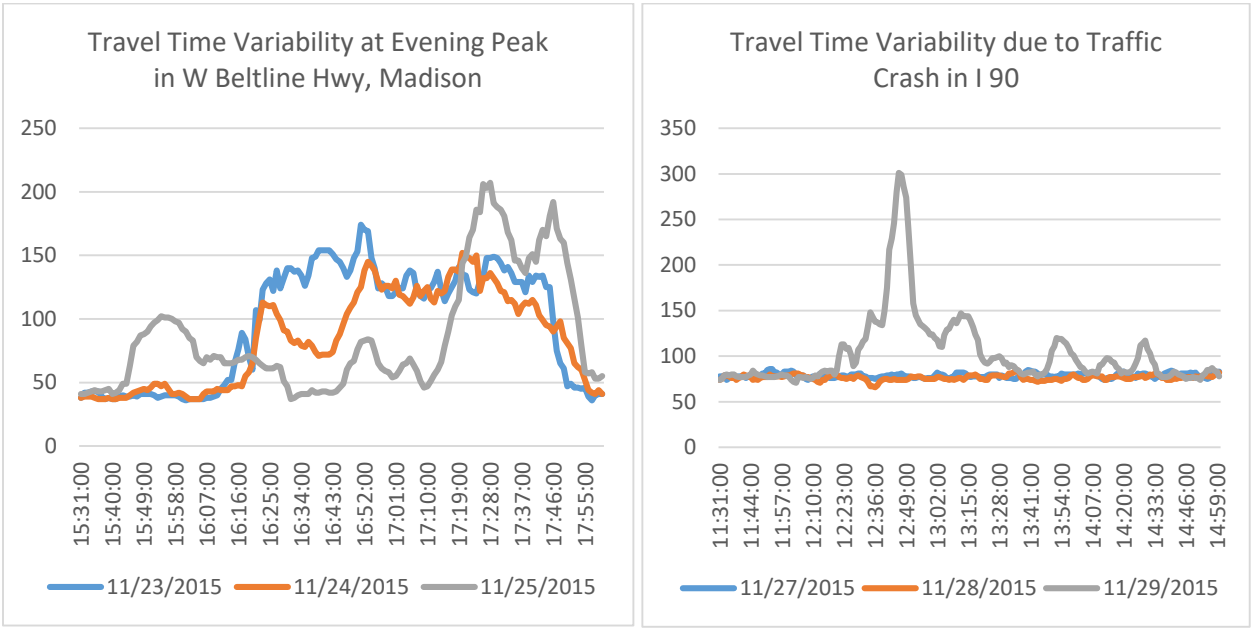
The graph presented above clearly states that the difference between ATT and DTT becomes higher at the onset of congestion. Data show that ATT usually lags behind DTT during a transition of traffic state, and the difference starts to decrease when the traffic state becomes stable. Information on travel time during a transitional state, as opposed to a stable state, is more important to the travelers. Ideally, the travel time in ATIS should be the prediction of travel time experienced by a traveler, or DTT. This predicted travel time also helps ensure proper and proactive operations and management of traffic in a network.

Travel Time Variability and Reliability

The variation of travel time on a route with the same origin and destination can be defined as the travel time variability. According to Arup et al., there are two distinguished components of travel time variability- incident related variability and day-to-day variability (Arup et al., 2004). The former one is random, whereas the latter one is predictable as it is demand and capacity related variability. Travel time variability is reciprocal to the reliability of travel time. According to Carrion and Levinson, higher the variability, lower the reliability, and hence, the unreliability can be defined as the measure of spread of the travel time probability distribution (Carrion and Levinson, 2012).

Outcomes of recurrent congestions are the day-to-day variability of travel time, which is somewhat anticipated by the travelers, specifically, the travelers who travel a specific route regularly. To adjust this anticipatable variation, travelers offset the added cost. The added cost consists of the value of travel time whose variability and reliability has been measured in several studies (Hensher and Truong, 1985, Eliasson, 2004) and was considered in cost-benefit analysis (CBA) of transportation projects (Kouwenhoven et al., 2014). A few studies have been conducted

to explore the quantitative methods of forecasting travel time variability (Sohn and Kim, 2009) and the effect of variability reduction (Eliasson, 2006). Travel time variability also includes the unpredictable variations that lead to the uncertainty of travel time. This randomness is the effect of non-recurrent congestions. FIGURE 5 (a) and (b) illustrates the effects of both recurrent and non-recurrent congestions respectively:



(a) (b)

FIGURE 5 a) Predictable (peak hour) b) unpredictable variability in travel time

Unreliable i.e. highly oscillating travel time is undesirable to travelers due to its cost in daily activities. Since it is impossible to remove the variations in travel time completely, informing the travelers in advance can be thought of as a better option. Therefore, at the beginning, the accurate prediction of travel time (DTT) using available real time ATT and then, short term forecasting (if necessary) might be an effective endeavor to ensure better assistance for travelers. However, the accuracy of predicted or forecasted travel time may vary with the variability of travel time. Higher accuracy can be achieved for travel time with lower variability and vice versa. As a

result, acceptable accuracy range for forecasted travel time should be narrower for the travel time with low variability than the travel time with high variability and vice versa (Toppen and Wunderlich, 2003).

Travel Time Data Collection Technologies

The application of travel time information is mainly two folds - congestion measurement and real-time travel information. According to Turner, electronic distance-measuring instruments (DMIs), computerized and video license plate matching, cellular phone tracking, automatic vehicle identification (AVI), automatic vehicle location (AVL), and video imaging are some advanced techniques of collecting travel time data. Among these, non-expensive electronic DMIs and expensive computerized and video license plate matching are most applicable for congestion measurement and monitoring. Cellular phone tracking, AVI, and AVL systems may require a significant investment in communications infrastructure, but appropriate for real-time information. When an observer records his travel time at predefined checkpoints using a data collection vehicle, it is known as the probe vehicle technique (Turner, 1996).

As a probe based AVI technology, Bluetooth data is not immune to the sampling issue that directly affects the accuracy and reliability of the information it provides. Its sampling rate is very low and depends on many issues including configuration, installation, location etc. Due to the low sampling rate and sampling error, data may not be available for every minutes and a single outlier may affect the travel time sharply. Despite of having some sampling issues, Bluetooth data has several advantages over other data sources including anonymous and continuous data collection. Moreover, Bluetooth data can be used as a Ground Truth data after applying proper data processing algorithm (Haghani et al., 2010). Ideally, a smoothed average travel time of all vehicles traversing

a target link or route is termed as “ground truth” for that link or route due to the variation in traveling speed among different vehicles (Toppen and Wunderlich, 2003).

Research Gaps

For Bluetooth data, sample size depends on the penetration rate of detectable Bluetooth signals in the traffic stream and the total number of vehicles per unit time. Generally speaking, higher sample size usually represents the population better than a lower sample size. The dilemma is that real-time information requires the travel time to be updated on a frequent basis, which may contradict with the desire for a large sample size collected over a long period given a low penetration rate. Another problem regarding sample size involves the penetration rate that varies over time. Provided that a valid sample size is determined, the time-varying penetration rate results in a changing time interval for updating the travel time: higher penetration rate requires a shorter time interval for travel time than lower penetration rate and vice versa. Dynamic time interval for travel time update creates confusion to travelers as information should be updated neither too frequent nor too slow. Another undesirable feature is the computation complexity added by constantly finding the proper time period that contains target samples. A few studies have been conducted to address the low sampling challenges of travel time estimation using Bluetooth data, hardly any of these has even mentioned sampling interval.

Bluetooth data is sensitive to the sampling issue that directly influences the accuracy and reliability of the information it provides. As mentioned earlier, its sampling rate is very low and depends on many aspects including configuration, installation, location etc. Despite of having some limitations, Bluetooth Technology (BT) has become popular due to various reasons including low cost. Due to the low sampling rate and sampling error, data may not be available in every

minute and a single outlier may affect the travel time sharply. Accumulation of data in several minutes would help to overcome these limitations. However, accumulation is a trade-off between the real time sensitivity and accuracy of the travel time. Since the longer the aggregation time period is, the more real time essence of travel time is compromised, it is imperative to know about the loss of stochasticity, a predominant property of travel time. FIGURE 6 delineates the loss of this predominant property among different aggregation intervals of travel time:

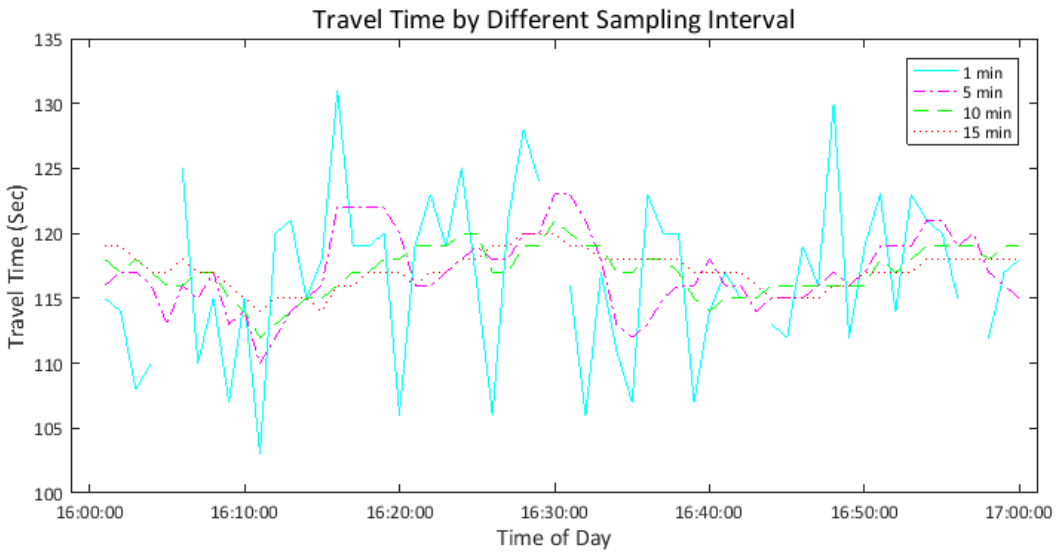


FIGURE 6 Variation in travel time stochasticity.

According to FIGURE 6, travel time variability decreases and coverage of intervals with no-sample increases in higher aggregation intervals e.g. 1-min, 5-min, 10-min and 15-min. The 1-min sampling interval yields excessive instability in travel time. Therefore, it is imperative to identify the appropriate time interval for sampling Bluetooth data that balance the need for accurate, reliable, and timely update of the travel time on freeways.

Travel time estimation is not a new topic. In general, the travel time in an ATIS is ATT due to the lack of available DTT; however, this reported ATT is one-step (step interval = travel time) earlier than the actual travel time to be experienced (DTT) by drivers, as mentioned earlier.

ATT usually lags behind DTT during a transition of traffic state, and the information on travel time during a transitional state, as opposed to a stable state, is more important to the travelers. Therefore, DTT should be manifested as real time travel time from the viewpoint of travelers. The travel time in ATIS should be the travel time experienced by a traveler, or DTT derived by the prediction from ATT. This predicted travel time also helps ensure proper and proactive operations and management of traffic in a network. Unfortunately, no prediction algorithm has been applied in any study to predict DTT from the available ATT. Moreover, very few studies have distinguished between DTT and ATT or attempted to estimate DTT (Kim et al., 2009).

Research Goal and Objectives

The overarching goal is to develop a comprehensive model for Bluetooth data preparation, accurate travel time estimation considering low sampling issues and thus, short-term travel time prediction on freeways based on the estimated travel time. To attain this research goal, the following objectives are proposed:

Firstly, identifying and recommending an appropriate time interval for sampling Bluetooth data that balance the need for accurate, reliable, and timely update of the travel time on freeways. Rather than determining a fixed sample size that leads to varying time interval for travel time update, a simple method has been proposed considering a balance between real time sensitivity and reliability of the travel time estimation. A framework should be developed to quantify the effect of aggregation on intervals in terms of high confidence sample rate, sample penetration rate and measure of succession.

Secondly, developing a dynamic filtering algorithm to estimate ATT and DTT reliably. This objective requires the development of an efficient computer algorithm to process, refine and integrate massive Bluetooth dataset to calculate the travel time.

Finally, proposing a prediction algorithm that is capable of predicting DTT from ATT accurately. It requires the discerning selection of appropriate method and critical analysis of its prediction performance.

Thesis outline

The remaining part of this thesis is organized as follows:

Chapter 2 explores the literature review on Bluetooth technology, outlier detection procedure, sampling requirements of low sample data, and travel time estimation and prediction.

Chapter 3 discusses about the massive dataset and its preparation. The preparation includes introduction of an appropriate filtering model supported by the principles from the existing dynamic travel time outlier filters. It also demonstrates the performance of proposed filtering model in travel time estimation briefly.

Chapter 4 examines the existing strategies to set sampling unit (sample size or sampling interval) for travel time estimation using Bluetooth data with low sample rates and inaugurates a novel procedure to find the suitable sampling unit.

Chapter 5 introduces travel time prediction model to predict DTT from ATT.

Chapter 6 summarizes the entire research, includes the contributions of the study and concludes the thesis with explanation regarding future research scopes.

CHAPTER II. LITERATURE REVIEW

The probe vehicle techniques, an intelligent transportation system (ITS) application, are primarily designed to collect data in real-time for traffic operations monitoring, incident detection, and route guidance applications including travel time. When driving a probe vehicle, different driving approaches (average, chasing and maximum car) can be followed to record the travel time using manual, DMI (Distance Measuring Instrument) or GPS (Global Positioning System) recorder. A probing technique, has both advantages like wide and flexibility of data collection area and disadvantages like single vehicle representation of entire traffic stream. Alternatively, probe vehicle data can be passively collected through manual or automatic vehicle identification, such as license plates, toll tags, cellular phone or Bluetooth signals. License plate matching techniques include three approaches – a) manual, b) video with manual transcription and c) automatic character recognition. Each approach has its positive and negative aspects. Cellular probe or wireless network data (handoff and location update) also has several pros and cons, which emphasizes the need for careful analysis of the characteristics of a network. Validation and evaluation can be difficult for these types of data due to lack of ground truth. In addition, the use of absolute or relative error is not straightforward to know the performance, because use of absolute error is unjust to high speed cases and relative error is unreasonable for low speed cases (Qiu and Cheng, 2007). After all, automatic vehicle identification (AVI) is the most accurate technique (depending on sufficient market penetration) where vehicles are equipped with transponders or toll tags (Toppen and Wunderlich, 2003).

Review on Bluetooth Technology

Amongst the AVI techniques, market penetration rate of Bluetooth devices is comparatively low. After compared to Automatic License Plate Recognition (ALPR) technology, it was found Bluetooth represented actual conditions very well at a fraction of the price of an ALPR despite of having low sampling rate (Wang et al., 2011). Other studies claim that the technology is powered by the effective matching capability, low cost and above all, sufficient accuracy with a professional setup (Turner et al., 1998, Araghi et al., 2015, Wang et al., 2011).

More studies found that the information collected using Bluetooth technologies could be subject to errors due to low penetration rate, communication range, location placement and installation, and some offered solutions. A study based on a 24 hours empirical data set on I-65 in Indianapolis has found that the MAC address is discoverable for 7.4% of the vehicles within 30′ and 6.6% of the vehicles between 102′ and 114′ (Brennan Jr et al., 2010). The freeway market penetration rate usually varies within a certain range, for instance, 5-11% (Quayle and Koonce, 2010) or 6.25% (Click and Lloyd, 2012) of total volume based on 24 hours counts. Communication range of Bluetooth devices is up to 300 feet, which can be affected by power rating, antenna quality, and obstructions between units etc. (Click and Lloyd, 2012, Bachmann et al., 2013). For instance, vertically polarized antennas with gains between 9dBi and 12dBi are the best antennas for travel time data collection (Porter et al., 2013). Limitations associated to MAC address scanners such as scanning frequency and maximum number of ID capturing in a same time frame can play vital role during data collection (Abedi et al., 2013). Moreover, the optimal number and location of Bluetooth sensors in a network for the reliability of the collected data(Asudegi, 2009) were thoroughly investigated and recommendations were made.

Malinovskiy et al. considered several types of Bluetooth detector antenna, several detector placement locations and Bluetooth device configurations (e.g. lane-length covered, antenna direction, opposite tandem, strength etc.) to estimate Bluetooth based travel time error on a short corridor for a 15-mins window (Malinovskiy et al., 2011). Detection zone, device mounting location, antenna direction and even, combination of mounting locations and antennas have significant impact on accuracy of travel time estimation. Bluetooth data quality related error can be generally classified into - spatial, temporal and sampling error (Malinovskiy et al., 2011, Mei et al., 2012). Spatial error indicates the lack of information about exact position of the vehicle at the time of detection. Temporal error includes multiple detection or no detection at all within the time range of up to 10.24 seconds after it enters the detection zone. Spatial and temporal error jointly lead to the measurement error. Sampling error refers to the low sampling rate that is unable to represent the population. In addition, Malinovskiy et al. considered sampling bias as a type of sampling error. Sampling bias includes error due to fast moving cyclists and bus passengers' Bluetooth devices, multiple Bluetooth devices in a single vehicle, vehicles with planned en route stops. Algorithms are designed to detect and remove some of the biases since a single outlier may affect the travel time sharply due to low sample rates. Intuitively, accumulation of several minute data would help to minimize the measurement error and a simple and robust outlier-filtering algorithm to overcome the limitation of sampling bias.

Review on Outlier Detection

The prime concern of outlier detection algorithms is to detect extreme travel times that result from sampling bias. Introducing a novel method named overtaking rule, Robinson and Polak filtered out vehicles that took indirect route, stopped en-route, were not restricted to normal traffic

regulations (e.g. emergency vehicles traveled over speed limit), traveled in an unusual fashion (Robinson and Polak, 2006). The percentile tests and deviation tests are other recognized outlier detection algorithms (Liu, 2008). Influenced by these methods, Liu introduced a generic algorithm to work offline. This algorithm needs future data set as well as whole dataset to define its parameters. However, the percentile test is a way to filter out outliers based on predefined percentile range (lower and upper limit). Therefore, the application of this method is subject to having prior knowledge of travel time distribution. In a deviation test, the range is defined by a critical distance (CD) from the median of the travel times within each period. Clark et al. applied percentile, deviation and traditional (modified) z- or t-statistical test (Clark et al., 2002). Traditional z- or t-statistical test outperformed other two methods, specially, in the presence of incident condition.

Fixed range outlier filtering methods are not suitable for travel time filtering due to local travel time turbulences, specially, at onset and end of congestion. Instead of imposing arbitrary bound, data driven real time adaptive bound (Dion and Rakha, 2006), moving average speed based lower and upper bound (Haghani et al., 2010) have been introduced by researchers in different adaptive algorithms. Compared to conventional algorithms, Dion and Rakha incorporated few simple but significant alterations in their proposed adaptive method. The main alteration includes the expansion of data validity window when three consecutive observations fall either above or below (same side) the validity window. Although this key adjustment helps to capture sudden changes in travel time trend, it is prone to the inclusion of extreme outliers. As a result, accuracy of travel time estimation would be compromised. Based on the algorithm proposed by Dion and Rakha, Moghaddam and Hellinga proposed a proactive method using pattern recognition model which showed superior performance (Moghaddam and Hellinga, 2014a). All these adaptive

methods are associated with some degree of complexity due to real time (online) applicability. For offline processing of dataset, a simplified version of the algorithm proposed by Dion & Rakha or Lie should be adequate to serve the purpose if applied appropriately.

Review on Sampling Techniques (Rate or Interval)

Sample size not only varies with the techniques of data collection but also varies with the types of studies or application. In a typical travel time study, sample size could be fixed by the researcher prior to data collection (Turner et al., 1998). In contrast, continuous samples are necessary for a real time application like travel time prediction.

Bluetooth technology has the advantage of collecting data continuously and anonymously (Moghaddam and Hellinga, 2014b). Different studies reflect researchers' efforts to find sampling requirement, more specifically, sample size for probe vehicles (Turner and Holdener, 1995, Chen and Chien, 2000, Li et al., 2005). Chen and Chien estimated the minimum sample size using statistical method and applied heuristic approach using CORSIM simulation to find the minimum number of required probe vehicle with a desired statistical accuracy. Their study suggests that 3-12 probe vehicles are required for each 5-min interval depending on traffic flow rate from low or high to moderate. Similar method based approaches have also been applied to define the minimum sample size considering cost, measurement error, true error and confident interval (Toppen and Wunderlich, 2003). Li et al. utilized Chris's probe vehicle sampling size model combining capacity constraints of wireless communication system (Li et al., 2005). The Chris's model (Ygnace and Drane, 2001) utilizes the information of traffic density, average link length and fraction of vehicles sampled to get the coverage. Ygnace and Drane studied cellular phones as probe vehicles to find the probe vehicle size to estimate travel time with 5% accuracy (Ygnace and Drane, 2001). In

addition, Jiang et al. studied the impact of probe vehicle sample size and sampling interval and concluded that the time interval had little effect for same sample size. They also demonstrated that the estimation error of average link travel time varied steadily when the sample size reached at a certain threshold (Jiang et al., 2006). Therefore, the accumulation of several minute samples are capable to provide reasonably reliable travel time regardless of population size. In a study, Click and Lloyd concluded that the intervals with sample size 8 or more possess higher confidence in Bluetooth data on rural freeways (Click and Lloyd, 2012). More accuracy can be ensured by applying appropriate methods of estimation. Araghi et al. estimated travel time in four different approaches- min, max, median and average travel time within two different sample intervals (15 and 30 mins) and found the min and median travel time were more robust in the presence of outliers (Araghi et al., 2015).

Although, studies regarding the impact of sampling interval and sample size are unavailable about Bluetooth data, studies related to probe vehicle explicitly exhibit that the sampling interval and sample size are interrelated. Therefore, a generic interval would not be effective for any dataset since sample size within an interval is uncontrollable. Sometimes, low sampling rate affects the minute-by-minute data availability. Nevertheless, accumulation of several minute data would increase number of samples within a certain time interval without affecting the penetration rate. Since accumulation of data is a trade-off between the real time sensitivity and accuracy of the travel time, fixing the accumulation time window i.e. sampling interval is excessively challenging. A data driven approach would be appropriate to decide the interval, ensuring minimum error in estimation of travel time.

Review on Travel Time Estimation

In transportation science, travel time, the time to traverse a specific route, can be differentiated in accordance with its measurement procedure. When a travel time is deduced from a spot based measurement (e.g. spot speed), it is known as instantaneous travel time. Contrary, an experienced travel time is a travel time that a traveler actually experienced; which is measured based on the start and end time of a journey. This experienced travel time considers traffic states and hence, it is the measure of free flow travel time and additional travel time (at no free flow state of traffic) due to the variability in traffic state. Measuring the experienced travel time is excessively simple for individual travelers but estimation based on samples to reflect the expected value of the actual travel time can be complicated due to several contributing factors.

As already discussed, Bluetooth is a relatively new technology that has become popular due to several reasons including low cost and anonymous detection of vehicles as well as travelers who carry Bluetooth devices accessible through designated roadside Bluetooth stations. Due to the technical complexity and limitations, none of the data collection procedure for travel time estimation including Bluetooth is error free. Consequently, collected data contains error that emerges the need for exploration of different robust techniques. In response, researchers introduced several outlier-filtering techniques (as studied earlier in this literature) to remove various types of error from the data. Once data is cleaned, it can be used to estimate experienced travel time using simple average method. Travel time data is available from those vehicles or travelers who have already traversed the route. Hence, the estimated travel time provides the most recent historic or current travel time as a real time travel time.

Since the most recently experienced travel time is the key to predict the future travel time, many studies have been conducted to ensure accurate estimation of travel time in a real time

fashion (Dion and Rakha, 2006, Skabardonis and Geroliminis, 2005, Li and McDonald, 2002, Lu and Chang, 2012, Sumalee et al., 2013, Moghaddam and Hellinga, 2014a). Those studies are diversified in terms of data, method and applicability. For instance, (Lu and Chang, 2012) and (Skabardonis and Geroliminis, 2005) applied traffic flow theory based model (queue and delay) to estimate travel time distribution in a signalized arterial using license plate recognition system data, and loop detector data respectively. (Moghaddam and Hellinga, 2014a), and (Dion and Rakha, 2006) detected outliers, removed and took simple average to estimate travel time for arterial, and freeway using Bluetooth, and AVI respectively. At the same time, researchers have also been involved in prediction of travel time as a sequential research of travel time estimation. As discussed in earlier chapter, DTT (will be experienced travel time) is more appropriate to be manifested as real time travel time rather than ATT (the most recently experienced travel time) from the viewpoint of travelers. Unfortunately, no research except (Kim et al., 2009) has been conducted to predict DTT from ATT. Although there is a research gap from the viewpoint of research goal and objectives, this topic has been over saturated in terms of, different methodologies resulted from a plenty of studies conducted over a decade.

Broadly, travel time prediction methods can be classified into two categories: the classical approach (Oda, 1990) which includes statistical (Rice and Van Zwet, 2004) and time series models (Al-Deek et al., 1998, Hamed et al., 1995), and the data-driven approach (Vlahogianni et al., 2014, Zheng and Van Zuylen, 2013, Zhang et al., 2014, Wu et al., 2004, Myung et al., 2011). Integration of different models, for example, a time-series model - auto regressive integrated moving average (ARIMA) with an embedded adaptive Kalman filter, has also been applied to develop a multistep travel time predictor (Xia et al., 2011). Due to the instability of traffic states, most classical approaches have shown to be incapable of better prediction, especially with regard to structured

and unstructured data (Vlahogianni et al., 2014). Since data-driven approaches fit easily with massive datasets, researchers have applied neural and Bayesian networks (Zheng and Van Zuylen, 2013, Oh and Park, 2011, Li and Rose, 2011, van Hinsbergen et al., 2009, Fei et al., 2011, Van Lint et al., 2005), fuzzy and evolutionary techniques (Zhang et al., 2014), support vector regression (Wu et al., 2004, Vanajakshi and Rilett, 2007), and k-nearest neighbor (Myung et al., 2011, Bustillos and Chiu, 2011) model to directly or indirectly predict travel time. Rather than using as a black-box or location-specific model, Van Lint et al. proposed for the state space neural network based on the lay-out of the freeway stretch of interest (Van Lint et al., 2005). However, according to Myung et al., the use of non-representative samples to train artificial neural-network (ANN) model may lead to the non-negligible error in prediction since this data driven approach needs to be trained using historical dataset (Myung et al., 2011). In addition, ANN is long training dependent using large historical dataset and non-transferable. Similarly, Support vector regression (SVM), a time-series forecasting format of support vector machine, also needs to be trained by a massive representative dataset.

Researchers have also applied the combination of several data driven techniques to achieve better performance. Li and Chen applied K-mean clustering to partition the dataset, CART-based classification to identify important variables and finally, neural network based approach to predict travel time in freeway with non-recurrent congestion (Li and Chen, 2014). Based on the variable selection method, they selected six scenarios and compared them in terms of mean absolute percentage error (MAPE); and found error rate of 6%-9%. Zou et al. applied MAPE, mean absolute error (MAE) and root mean square error (RMSE) as performance indexes to make a comparison among the output from auto regressive model (AR), vector auto regressive model (VAR), space-time diurnal truncated normal model (ST-D TN) and space-time diurnal lognormal model (ST-D

LN) (Zou et al., 2014). This probabilistic prediction method was able to provide the prediction related confidence intervals that reflected the level of uncertainty. Results showed that ST-D LN model performed better in case of predicting multiple time-steps ahead into the future.

One notable observation is that the vector auto regressive model has performed slightly better than other methods in case of 5-min ahead prediction. Unlike the neural networks and some hybrid methods that lack a good interpretation of the model due to the ‘‘black box’’ approach, this model has used theoretically interpretable travel time features (temporal and spatial correlation, diurnal pattern, and the non-negativity of the travel time). Temporal and spatial correlation feature has incorporated information from upstream and downstream locations (similar to vector auto regressive model) into the prediction model unlike the on-site predictor type univariate auto-regressive model. Although the application of diurnal pattern made the model robust in terms of interpretability, it might deteriorate the model performance with the change in pattern due to non-recurrent congestion.

To predict dynamic travel times, Elhenawy et al. proposed a simple and computationally efficient genetic programming algorithm that did not suffer from non-interpretability (Elhenawy et al., 2014). Nevertheless, data requirements to train the model would be specific to the roadway and traffic conditions. Moreover, the minimum data requirement is yet to be investigated.

Usually, performance deteriorates in case of multi-step prediction using the ANN method (Boné and Crucianu, 2002, Parlos et al., 2000). Even after introducing a noble method named time-delayed state-space neural network, Zeng and Zhang has found more than doubled error in the fifth step compared to the first step (Zeng and Zhang, 2013). Hence, Chen and Rakha applied particle filter, a sequential Monte Carlo method, which outperformed other methods including K-NN and Kalman filter (KF) in case of multi-step prediction (Chen and Rakha, 2014b). Their results showed

consistency in systematic error propagation whereas performances of different methods were somewhat reasonably close at only the beginning. The authors also introduced another recursive probabilistic method namely agent based modeling, which is a weighted average of prediction from different agents i.e. predictors generated by the model (Chen and Rakha, 2014a). Each agent used pattern recognition technique to predict the travel time. The concept is somewhat similar to Boosting technique (Schapire, 2003) of machine learning, a concept of combining rough and moderately inaccurate rules so that the combination can produce an accurate prediction rule. A method comparison showed that this agent based model performed better than historic average, instantaneous and k-NN (Chen and Rakha, 2015).

KF has been widely used with different modification (e.g. adaptive KF (Guo et al., 2014) and extended KF (Liu et al., 2006)) in different studies including travel time prediction (Chien and Kuchipudi, 2003, Nanthawichit et al., 2003, Yang, 2005, Chen and Chien, 2001). KF, an optimal recursive data processing algorithm, incorporates all information that can be provided to it and process all available measurements to estimate current value of the variables of interest (Maybeck, 1990). By definition, it can be inferred that KF has two part- process (also known as system or state space) and measurement (also known as observation). Hence, the algorithm includes process update (also known as time update) and measurement update steps. Both the noises related to process and observation have to be white Gaussian.

Nanthawichit et al. formed their state space equation by declaring traffic density and space mean speed as state variables and observation equation by declaring traffic volumes and spot speeds as observation variables (Nanthawichit et al., 2003). Chen and Chien directly used travel time as their input variable in both state and measurement equations. Previous step travel time was multiplied by a transition matrix to obtain the state update equation (Chen and Chien, 2001).

similar studies were conducted using field data (Chien and Kuchipudi, 2003) rather than simulated data (Chen and Chien, 2001). Despite of showing promising results, the studies deficit in some degree of details about the case study including sources of process and measurement (variables') values. Although only Yang discussed about the effect of noise variance on prediction error, he did not mention about its variability with time-step (equations in article showed that the noise variances were time varying). In addition, no one discussed about the assumed or derived (including the rationale) values of error covariance that have greater effect on prediction performance in case of single step prediction. Regardless of methodologies, single step prediction provides somewhat similar output (as already mentioned) in most comparisons. Therefore, it is expected that the KF, applied with accurate assumptions, would provide reasonable performance in case of predicting travel time.

Moreover, KF can be indexed as a data induced statistical time series based noise-filtering technique. It requires a measurement update, to perform time update i.e. to predict the future step. Hence, the enough knowledge about the process, availability of measurement and reasonable assumptions regarding noise covariance are highly desirable to experience better performance. Moreover, KF provides the way to update the state space when new observations become available; and state space form is the key to handle structural time series models that are nothing more than regression models (Harvey, 1990). Therefore, KF has the robust power to apply with a regression driven time series model. Favorably, KF algorithm exactly matches with the concept of predicting departure time based travel time (DTT) from arrival time based travel time (ATT).

The knowledge regarding the KF noise is intuitively related to the instability in traffic conditions. As mentioned in previous chapter, different types of congestions: recurrent and non-recurrent are responsible for the instability in traffic conditions. The most unpredictable

congestions are the incident related congestions which are known as non-recurrent congestion (NRC).The congestions that exhibit a daily pattern during peak and off-peak periods are known as recurrent congestion (RC) (Anbaroglu et al., 2014). RC is somewhat predictable by travelers. Although NRC combined with RC makes the situation worsen, in some cases, researchers have showed great deal of success in detection, sometimes, distinction of RC and NRC (Hawas, 2007, Dowling et al., 2004, Skabardonis et al., 2003, Garib et al., 1997, Lin and Daganzo, 1997, Ritchie and Cheu, 1993). Accurate distinction is imperative to classify the KF noise related to different traffic state which may lead to the precise prediction of travel time.

CHAPTER III. DATA PREPARATION AND REDUCTION

Bluetooth data contains three variables: the MAC ID of the detector, MAC IDs of the detected devices, and the detection timestamp. In spite of a simple data format, a complex processing algorithm is required to produce the final dataset from the source, which stores the entire network data in a single table. For a logged MAC ID, recorded timestamps at two consecutive stations are processed to estimate travel time and the corresponding traffic speed. The complete processing algorithm can be subdivided into three major parts: a) data pre-processing, b) outlier filtering, and c) output generation. The detailed description of the processing algorithm is included below with a brief description of the data characteristics.

Data Description

The selected study area consists of a 62.8-mile long route, or approximately 47.5 miles on I-90 and the remaining on the Beltline Highway in Madison, Wisconsin. The route is equipped with 41 unequally spaced Bluetooth stations, resulting in 40 links. The first 21 links are on I-90, the 22nd link is on both corridors, and the remaining links are on the Beltline. The spacing varies from 1.3-3.4 miles on I-90 and 0.4-1.3 miles on the Beltline Highway. FIGURE 7 shows the study area and the stations' locations.

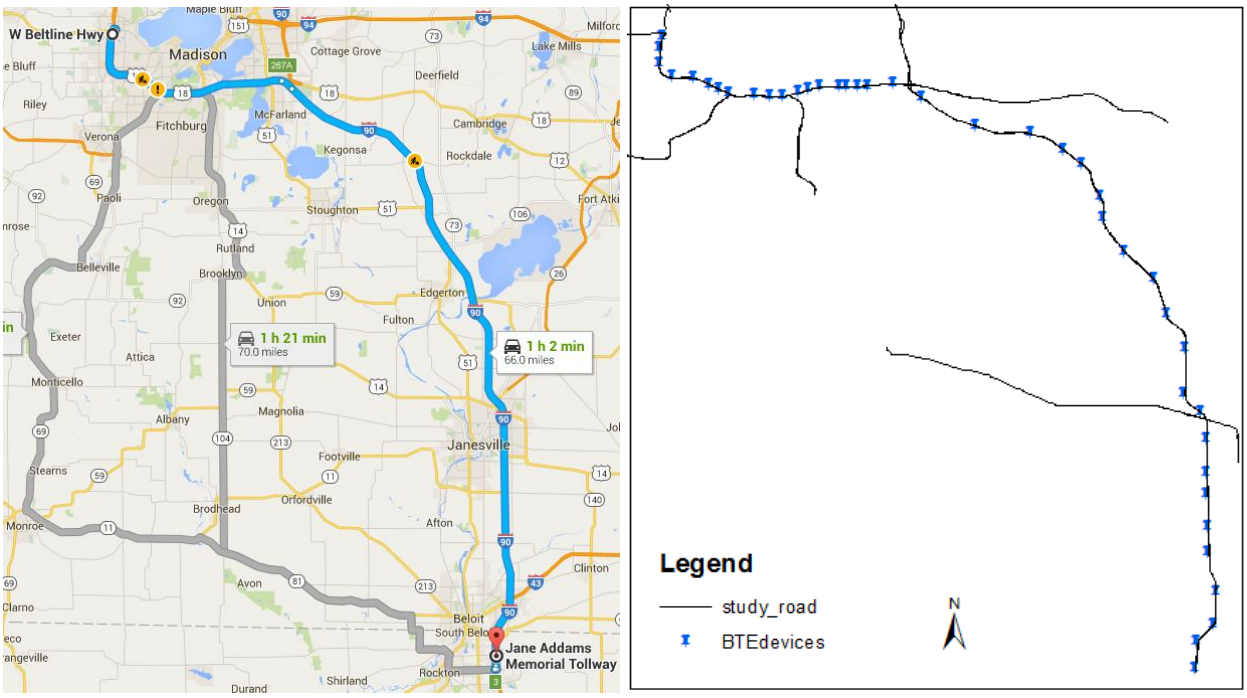


FIGURE 7 Study Area (Wisconsin, US) with the location of Bluetooth Devices/Stations.

Forty-seven days' worth of data (11/16/2015-01/01/2016) containing more than 100 million records was collected from traffic in both directions. Half of the records were from outside the study-area. Each station of the one hundred stations selected captured around one million records for 47 days, or 67,680 minutes. However, a large portion (approx. three-fourths) of the data are either corrupted or contaminated due to multiple detections and unsuccessful detections (i.e. not detected in two consecutive stations). Therefore, the average penetration rate would be roughly 3-4 samples per minute. Extensive efforts have been assumed to ensure the data quality.

Data Processing

Efficient and reasonable data processing was made possible by automation. Initially, data was queried from an Oracle database to remove unnecessary records like multiple detections, reducing the records to a reasonable number. Then, a Java application has been developed to further process

the previously filtered dataset. FIGURE 8 shows the complete procedure of data processing followed by the detail descriptions in following sections.

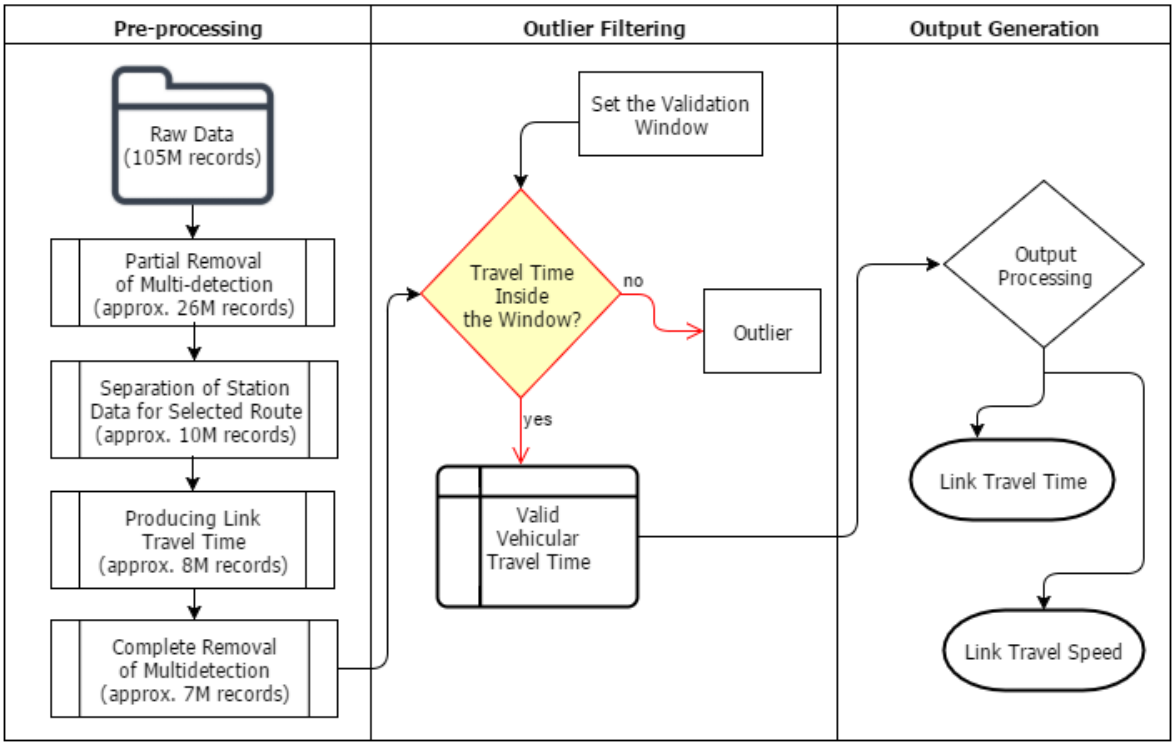


FIGURE 8 Data processing procedures

Pre-Processing

The primary goal of data pre-processing is to clean unnecessary records and prepare a smoothly workable dataset containing calculated travel time of each vehicle. A Bluetooth station usually detects a Bluetooth device in its range more than once. The number of such detections can increase significantly due to planned or unplanned stopping of vehicles. A general inspection of the dataset revealed that such detections usually vary two to four times. Detecting a vehicle at the end of a station and at the beginning of its upstream station would yield maximum spatial error. Since the study data lack of signal strength information, the record of first detection has been preserved carefully to minimize such error. If signal strength information were available, the strongest signal

(of each multi-detection) data would have been preserved to estimate the travel time. Since the detection is automatic, it is impossible to ensure a vehicle's detection at the points having same signal strength in different detection zones. Consequently, it is impossible to avoid spatial error of detection, even by using the strongest signal information. For instance, the strongest (among all detections) signals at S11 and S22 points (assumed) could provide the longer length (2.1 miles) than the actual link length (2 miles) and introduce spatial error (FIGURE 9).

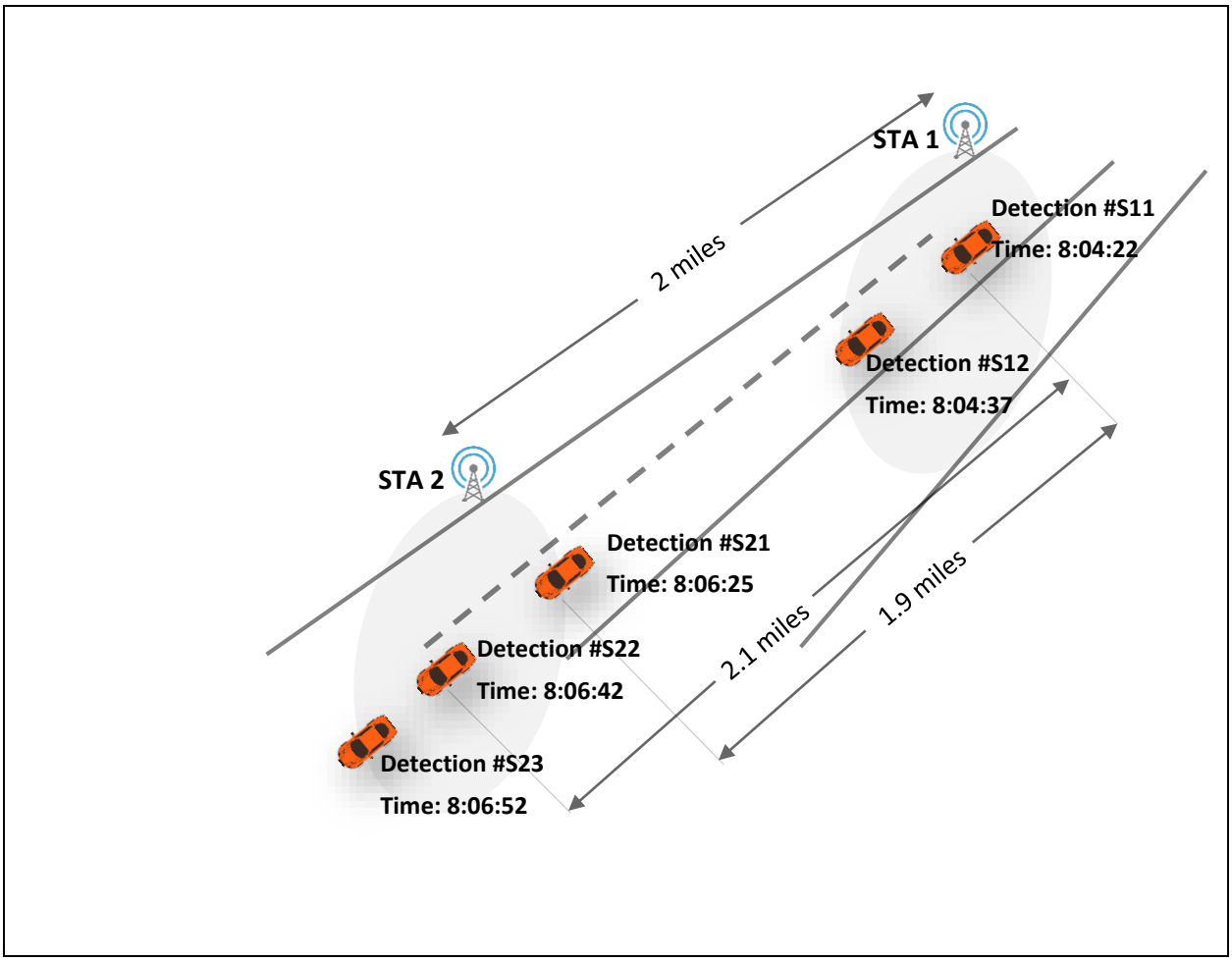


FIGURE 9 Detections of a vehicle at two consecutive Bluetooth stations.

In addition, first detections (at S11 and S21 points) may provide the shorter length (1.9 miles) than the actual link length (2 miles) and also introduce spatial error. Either way it may contain spatial error. Oracle queries helped clean up the multi-detection (records except the first one were

deleted), resulting in the total number of records decreasing from 105 million to 26 million. The data was then separated by each station for the selected routes, further reducing the records to 10 million. When travel time of each vehicle was estimated from the separated station data, unsuccessful detections were automatically ignored due to the vehicle's detection timestamps from two adjacent stations. Next, the reduced dataset of 8 million samples was processed through a robust Java-based pre-processing module that investigated each record individually and cleaned all unnecessary records based on the following principle:

> *A vehicle cannot be detected twice in a station, STA_1 without being detected at least once in its upstream station, STA_2 within the time gap of two detections in STA_1. If so, these are redundant detections for any single trip.*

For example, a vehicle detected on 08:59am, 09:01am and 09:08am at STA_1, and on 09:04am at STA_2. Detection on 09:01am is redundant since there is no detection at *upstream station* STA_2 in between 08:59am and 09:01am. In addition, none of the detections on 09:01am or 09:08am is a redundant detection since there is a detection at *upstream station* STA_2 on 09:04am. The pre-processed dataset of 7 million records containing the journey start time, end time and travel time of each vehicle was further processed to filter outliers.

Outlier Filtering

Outlier filtering is a challenging task due to the possibility of treating a good sample as an outlier. Various studies have been conducted previously to define and filter outliers, and a wide range of methods, simple to complex, have been introduced. In this study, a simple yet robust statistical approach has been applied in which the key of filtering outlier or selecting good samples is to define the validation window i.e. the upper and lower boundary (of travel times). The lower

boundary can be defined based on the assumption that a vehicle's speed cannot exceed more than the double of a posted speed limit, or it is an outlier

$$tt_{lowr} = {}^{tt_{ff}}/_{2} \quad (3.1)$$

Where, tt_{lowr} and tt_{ff} are lower bound and free flow travel time respectively.

Once the lower boundary is fixed, upper boundary should be dynamically defined by the samples (travel time) because of the inherent stochastic nature of travel time. A dynamic validation window works best in case of outlier filtering (Dion and Rakha, 2006) but the implementation is very challenging due to the unprecedented variability of travel time. Thus, the following equation is proposed:

$$tt_{uppr} = tt_e + n.\sigma_e \quad (3.2)$$

Where tt_{uppr}, tt_e and σ_e are upper bound, expected travel time and expected standard deviation of travel time (samples) respectively. $n = 1, 2, 3,$ ....= nth standard deviation.

The expected travel time tt_e is the predicted DTT and σ_e is the predicted standard deviation of DTT since the actual DTT is unavailable in real time. However, DTT and its standard deviation have been used as $tt_e \; and \; \sigma_e$ respectively, since the outlier filtering has been accomplished in offline. The unfiltered dataset was processed dynamically using time varying validation window. The time varying standard deviation σ_e was estimated using the samples from the latest 30mins data and n was set to 2. The decision for sampling previous 30 minutes with n equals to two was made from trial and error for better performance. It is assumed that the standard deviation of the upcoming samples remains similar to the estimated standard deviation of the recently observed samples for a shorter time period. Following figure shows the performance of this simple method:

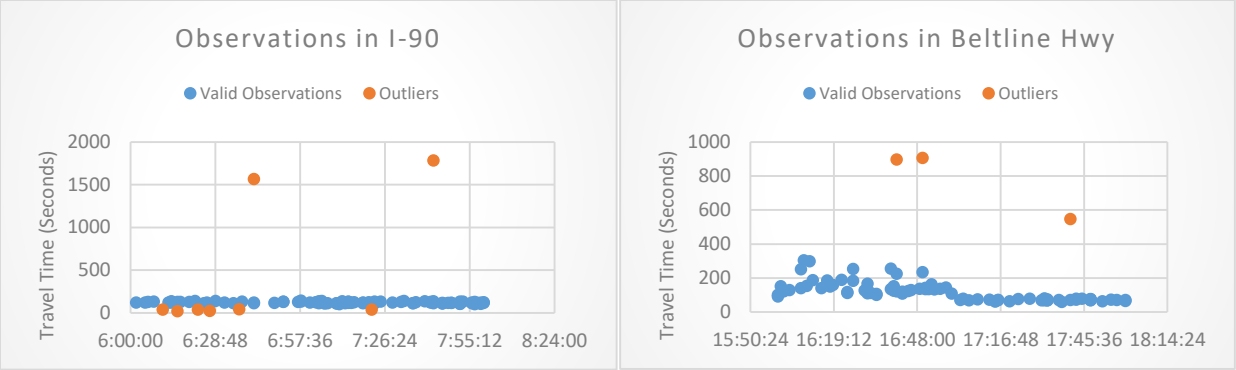


FIGURE 10 Performance of filtering algorithm during morning and evening peak hours.

Output Generation

Finally, a Java-based programming module produced the travel time and speed data using the outlier-filtered data. Since travel direction is pertinent to travel time, this study used northbound data.

CHAPTER IV. IDENTIFYING SAMPLING INTERVAL

Methodology

The methodology section details the process of estimating travel time by aggregating samples from several minutes and the method for selecting sampling interval.

Travel Time Aggregation

Estimating travel time in one-minute interval is straightforward, like taking the average of available samples. For an interval of more than one minute, two types of aggregation can be considered: a) simple average and b) moving average. The basic difference between these two estimation procedure is that the former one gives a single travel time for the aggregation interval which means same travel time for every minute within the interval while the latter one updates travel time at each minute regardless of interval size. The estimation process is shown in TABLE 1.

TABLE 1 Travel Time Aggregation Process

Time of Day	09:01	09:02	09:03	09:04	09:05	09:06	09:07	09:08	09:09	09:10
Travel Times of Samples	t_{11}, t_{12}	t_{21}	t_{31}, t_{32}	t_{41}	t_{51}	t_{61}, t_{62}, t_{63}		t_{81}	t_{91}	t_{101}
Simple Average	N/A					$\frac{t_{11}+t_{12}+t_{21}+t_{31}+t_{32}+t_{41}+t_{51}}{7}$				
Moving Average	N/A	N/A	N/A	N/A	N/A	$T_{09:06}$	$T_{09:07}$	$T_{09:08}$	$T_{09:09}$	$T_{09:10}$

$T_{09:06} = \frac{t_{11}+t_{12}+t_{21}+t_{31}+t_{32}+t_{41}+t_{51}}{7}$, $T_{09:07} = \frac{t_{21}+t_{31}+t_{32}+t_{41}+t_{51}+t_{61}+t_{62}+t_{63}}{8}$,
$T_{09:08} = \frac{t_{31}+t_{32}+\ldots+t_{63}}{7}$, $T_{09:09} = \frac{t_{41}+t_{51}+\ldots+t_{81}}{6}$, $T_{09:10} = \frac{t_{51}+t_{61}+\ldots+t_{91}}{6}$

Sampling Interval Selection

The accurate prediction of travel time is mostly constrained by the turbulence in traffic state. No algorithm is able to completely handle the randomness of traffic state i.e. travel time. As a result, researchers use average travel time over the longer time period (usually 15 minutes) to limit the erratic nature of travel time within a tangible proximity. The longer the time period is; the more real time essence of travel time is compromised. Compared to non-aggregated real time data, aggregated data may show better prediction accuracy that is not necessarily the best approximation of the real situation. Therefore, it is imperative to know the measure of succession (of stochasticity, a predominant property of travel time).

Selection of sampling interval is a two-step process: In the first step, the high confidence sample rate for all the sampling intervals will be determined. The high confidence sample rate refers to the percentage of intervals that contain sample sizes equal or more than a predefined required sample size. Then, the sample penetration rate will also be estimated for all the sampling intervals. The sample penetration rate refers to the percentage of intervals that contain at least one sample. Based on the high confidence sample rate and the sample penetration rate, a minimum sampling interval selection would be determined. In the second step, a sampling interval from a bunch of candidate intervals (e.g. 5-min, 10-min etc.) would be selected based on the measure of succession. The candidate intervals must be equal or higher than the selected minimum interval. To measure the succession, a simple but effective method based on travel time reliability measure has been applied.

High Confidence Sample Rate

The high confidence sample rate (R_{HC}) is the percentage of intervals that contain samples more than a predefined threshold. R_{HC} for a route can be expressed by,

$$R_{HC} = \frac{\sum_n N_{HC.i}}{nN} \tag{4.1}$$

where $N_{HC.i}$= total number of high confidence sample interval in a link i, n = number of links and N = total number of intervals within the analysis period.

Total number of high confidence intervals (N_{HC}) for a link can be estimated by,

$$N_{HC} = \sum_N I_{j(m/r)} \tag{4.2}$$

where $I_{j(m/r)}$ is a Boolean function to determine whether the j^{th} interval with m samples has high or low confidence given required minimum sample size, r.

The Boolean function,

$$I_{j(m/r)} = \begin{cases} 1, & m \geq r \\ 0, & otherwise \end{cases} \tag{4.3}$$

where m represents number of samples in j^{th} interval and r is the required minimum samples.

Sample Penetration Rate

Since the sample penetration rate refers to the percentage of intervals that contain at least one sample, it can be estimated by following the same method of estimating high confidence sample rate using r=0.

Measure of Succession

In general, succession refers to the action or process of inheriting a property. Therefore, the measure of succession indicates the measure of inheriting a foremost property. Within the travel time dataset, succession should be the measure of travel time variability or stochasticity inheritance since variability is a foremost property of travel time. Therefore, the inheritance i.e. the conveyance of variability with aggregation should be evaluated before selecting an appropriate time interval. The conveyance of travel time variability refers to the degree of variability conveyed to the candidate intervals after aggregation from the benchmark interval. Travel time variability, also referred as travel time reliability, is a measure of spread to the travel time distribution (Carrion and Levinson, 2012) which can be quantified by mean and standard deviation of travel times (Martchouk et al., 2010). Therefore, conveyance of travel time variability into the aggregated travel time can be estimated by travel time reliability test based on these statistical properties (mean and standard deviation).

Travel time reliability, a key performance indicator, is related to the properties of the day-to-day travel time distribution. The reliability measures include 90th or 95th percentile travel time, buffer index, planning time index, frequency that congestion exceeds some expected threshold, and several statistical measures of variability such as standard deviation and coefficient of variation (Reliability, 2006). It also includes some probabilistic approaches, tardy trip measures or misery index, and some other modern approaches (Van Lint et al., 2008, Clark and Watling, 2005, Guo et al., 2010). Objective and quantitative criteria are keys to selecting the appropriate measure for travel time reliability. In this study, coefficient of variation (CV) was applied to estimate the conveyance of travel time uncertainty. For each link i, CV is estimated by following equation,

$$cv_i = {}^{\sigma_i}/_{\mu_i} \tag{4.4}$$

The difference between the measure of uncertainty between an aggregated interval (k-min) and the benchmark interval is the measure of its conveyance. Therefore, travel time uncertainty/unreliability conveyance of link i is measured by,

$$d_{ik} = |cv_{ib} - cv_{ik}| \quad (4.5)$$

where cv_{ib} and cv_{ik} are the measure of reliability for the benchmark and an aggregated interval (k-min) respectively in link i. Assuming $d_{1k}, d_{2k}, \ldots d_{nk}$ are the change in travel time uncertainty/unreliability of the links $1, 2, \ldots i \ldots n$ for any aggregated interval of k-min. Let, the mean and standard deviation of these changes in uncertainties are μ_k and σ_k (where suffix k denotes the aggregation time interval).

Smaller μ_k corresponds to smaller average difference in CV for the entire route, which means higher conveyance of travel time reliability/variability. Smaller σ_k corresponds to smaller variations among the differences in travel time reliability/variability over the entire route, which means conveyance of travel time reliability/variability is consistent or somewhat similar in the network i.e. among all the links. However, larger σ_k refers to the situation where some links have higher loss in travel time reliability and some have lower. Therefore, conveyance of travel time reliability/variability is inconsistent within the network.

Results and Discussion

A study (Click and Lloyd, 2012) shows that 8 samples per 15-min are sufficient to provide a reliable and accurate travel time or speed estimation using Bluetooth data collected on rural freeways. The penetration rate of that study data was around 5-6%, which is a general penetration rate of Bluetooth data. Since 8 samples are sufficient for a 15-min interval, it will also be sufficient

for the intervals lower than 15-min. Therefore, to estimate the high confidence sample rate, the value of the required minimum sample size (r) parameter was set to 8 samples per interval. The high confidence sample rate and the sample penetration rate were estimated for 1-min to 15-min intervals.

FIGURE 11 represents that both the high confidence sample rate and the sample penetration rate increase with the increment of sampling interval while the increasing rate gradually decreases. It shows that the increasing rate of sample penetration rate becomes significantly low after 5-min aggregation. However, increasing rate of the high confidence sample rate decreases after 8min aggregation that transforms 54% of total intervals to high confidence intervals. Generally, the traffic flow rate is significantly low resulting free-flow before 6:00AM in the morning and after 9:00PM at night. Within this period, it would be extremely difficult to get 8 samples for a 15-min interval. Hence, a general expectation is that the 15 hours or 62.5% time of a day should be under high confidence surveillance to provide reasonably reliable and accurate information to the control rooms as well as travelers. Therefore, 10-min aggregation providing around 62% intervals of high confidence sample would be a great choice of minimum aggregation interval. This interval will also ensure 93% sample penetration rate which is higher than 88% (for 5-min interval).

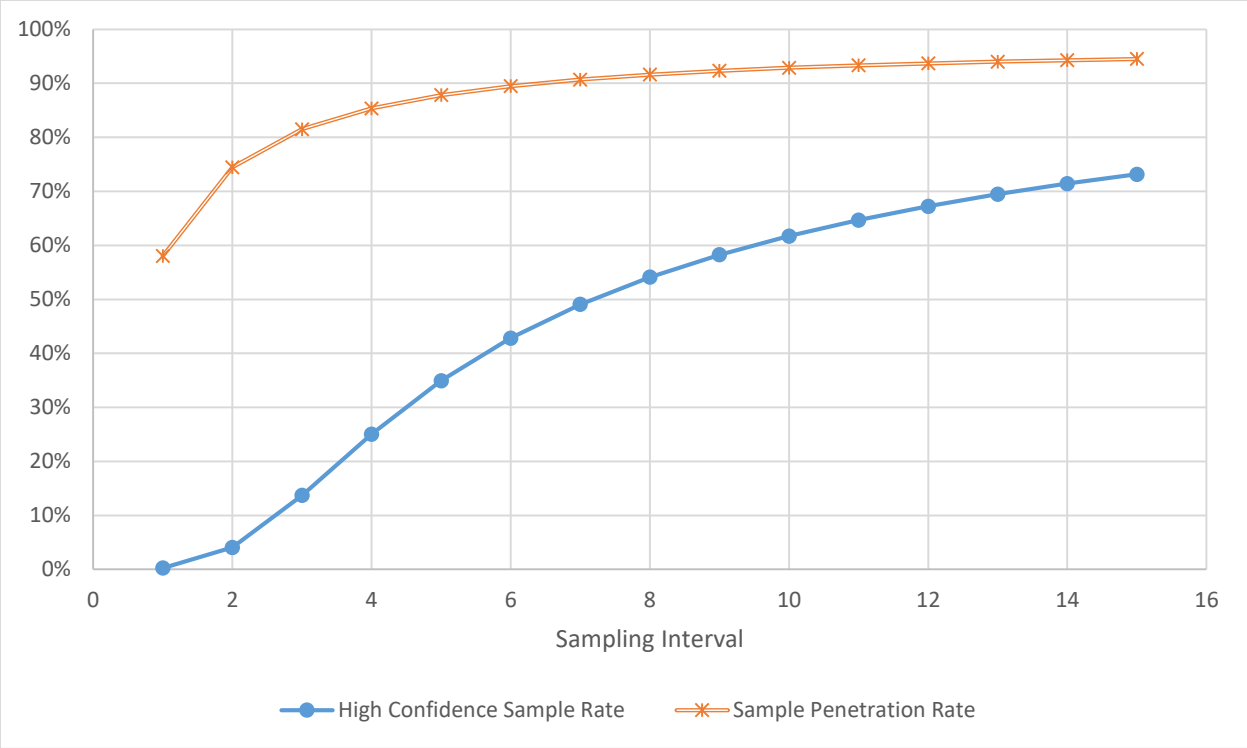


FIGURE 11 Change is sampling character for different intervals.

Benchmarked (1-min travel time) against 10-min and 15-min (multiplier of five minutes, as a general trend of travel time estimation interval) aggregations were examined for the travel time variability conveyance. Since two types of aggregation were considered, four sets of travel time data (10-min and 15-min simple and moving average) with the 1-min data as benchmark were evaluated by the reliability test. Since the study network is comprised of parts of two different corridors (I-90 and Beltline Highway, Madison, WI), the reliability test results of one corridor's links are significantly different from the scores of another corridor's, as shown in FIGURE 12. The results of 10-min and 15-min moving average are almost identical to the results of 10-min and 15-min simple average, respectively. Therefore, 10-min and 15-min simple average has not been included to ensure better presentation/visualization of the graph.

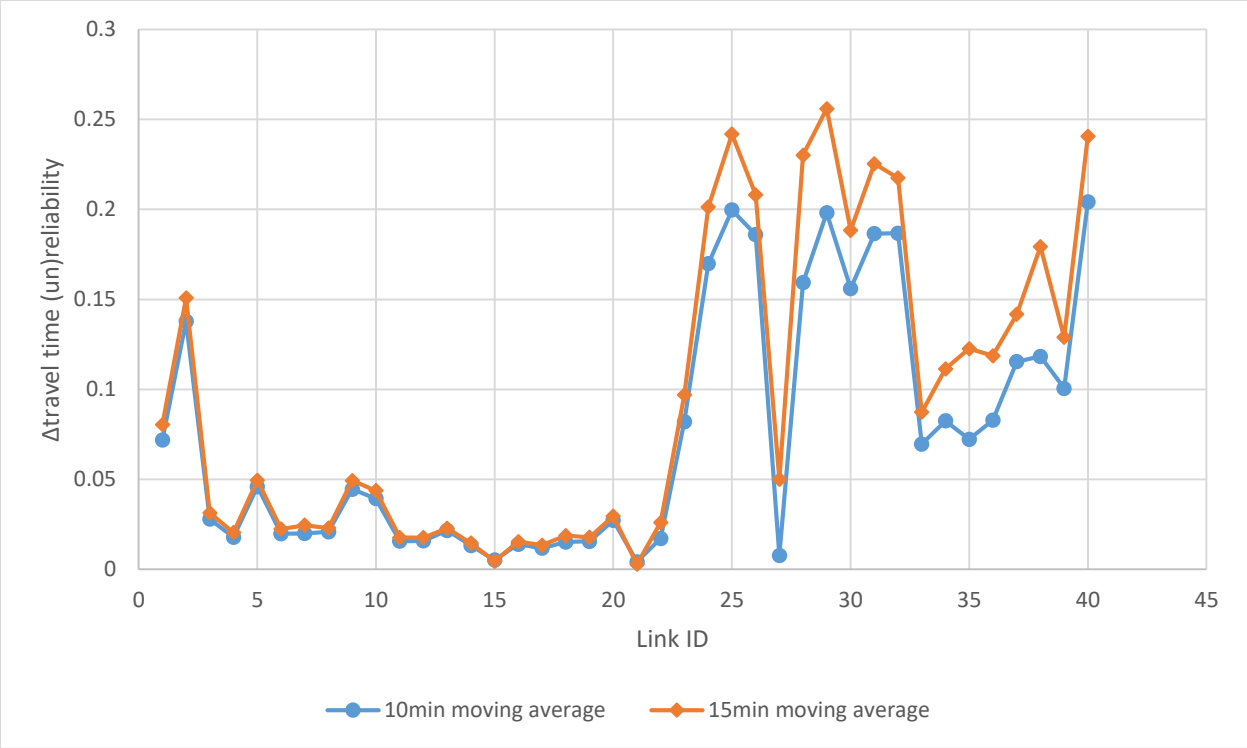


FIGURE 12 Results of reliability test in each link.

The first 22 links of the selected route are from I-90, which shows lower sensitivity towards aggregation than the rest 18 links from the Beltline Highway. Beltline Highway suffers from recurrent congestion during peak period while I-90 highway does not. Therefore, one may argue that the speeds of vehicles in a congested situation can show significantly lower variation than speeds in free flow condition. The relatively high variation of travel time in a free flow condition is potentially affected by the drivers' flexibility to drive at different speeds. Since the variation is higher at free flow condition, data aggregation in a longer time period reduces the variation in travel time significantly for the links that mostly experience the free-flow condition. However, that may not be the case due to two potential reasons: a) total congested (or peak) period is much shorter than the total free flow (or off-peak) period and b) since the posted speed is lower and the link lengths are much shorter in the Beltline Highway, the relative variation in driving speed as well as travel time might be higher at free flow condition. Moreover, the much shorter (0.4-1.3 miles)

links in Beltline Highway might contribute to the inclusion of spatial error in Bluetooth data. Since Bluetooth device has a detection zone covering a significant length, for instance, 300 feet (Click and Lloyd, 2012). From the local perspective, 10-min interval is preferred than 15-min interval.

The performances of moving and simple average approaches are also compared. The two different corridors have significantly different sensitivity towards the conveyance of reliability, it is better to examine the global/overall conveyance. Table 2 compares the conveyance of travel time variability/(un)reliability of the corridors and the entire route.

TABLE 2 Conveyance of (un)Reliability Property (Global Perspective)

Sampling Interval		10-min		15-min	
Averaging Method		Moving	Simple	Moving	Simple
I-90 corridor	MEAN	0.0207	0.0203	0.0228	0.0188
	STD. DEV.	0.0282	0.0285	0.0309	0.0320
Beltline Highway Corridor	MEAN	0.1110	0.1205	0.1559	0.1328
	STD. DEV.	0.0565	0.0527	0.0609	0.0696
Overall Route	MEAN	0.0750	0.0760	0.0936	0.0966
	STD. DEV.	0.0682	0.0674	0.0840	0.0905

TABLE 2 shows that the simple average of 15-min interval has the least distinction with the benchmark in I-90 while with the maximum variation (the highest standard deviation) over the different links in the corridor. The moving average aggregation of 10-min interval shows the least distinction with the benchmark in the Beltline Highway and exhibits the moderate variation (second lowest standard deviation) over the different links in the corridor. However, the moving average aggregation of 10-min interval also represents minimum deviation from the benchmark over the entire study route (two corridors). In addition, it shows the second lowest variation (slightly higher than the lowest) over the route. The global indicators (mean and standard

deviation) of moving and simple aggregations in 10-min interval are off by a negligible value (0.001) and one shows the most conveyance and another shows the least variation in conveyance over different links on the entire route. Therefore, 10-min moving or simple aggregation can be selected considering the patterns of update: former one will provide new/updated travel time at each minute and latter one will provide updated travel time at the end of each 10^{th} minute.

CHAPTER V. PREDICTING SHORT-TERM FREEWAY TRAVEL TIME

Methodology

The methodology section details the algorithms for ATT, DTT, travel speed estimation, and travel time prediction.

ATT, DTT and Speed Estimation

ATT and DTT refer to travel time and are associated with either the time of arrival at the destination or the time of departure from the origin, respectively. To get ATT and DTT at the same time, assume that two vehicles (Vehicle 1 and Vehicle 2) start at 8:30am and 9:00am from point A and end at 9:00am and 9:25am at point B, respectively. $ATT_{AB@9:00AM} = 30mins$ and $DTT_{AB@9:00AM} = 25mins$. Vehicle 1 that traveled link AB at 9:00am has already experienced a travel time of 30mins (ATT), and Vehicle 2 that traveled link AB at 09:00am will experience a travel time of 25mins (DTT). Since the DTT at 9:00am is unavailable until 9:25am, it is understandable that the DTT at 9:00am requires a prediction of travel time. FIGURE 13 illustrates the concept of estimating ATT and DTT for a link:

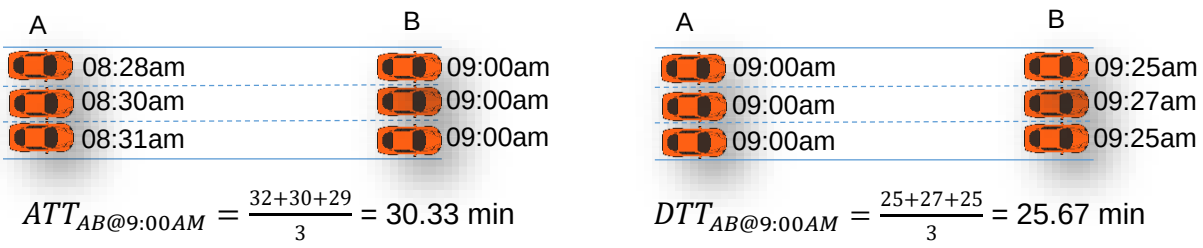


FIGURE 13 Estimating ATT and DTT of link AB at 9am.

The following equation estimates the space mean speed of a link AB with length L:

$$v = \frac{L}{\frac{1}{n}\sum_i tti} \tag{5.1}$$

where, n is the number of observations within any particular interval and tt_i is the travel time of i^{th} observation.

Travel Time Prediction

According to the literature, different methods have different advantages and disadvantages. Instantaneous, historic average and clustering over specific days with similar traffic patterns are widely used concept for travel time prediction due to their simplicity. However, these models possess low accuracy (Van Hinsbergen et al., 2007). In case of linear regression, parameters of prediction models are calibrated by historical data. The predictive performance of the K-nearest neighbor (K-NN) method largely depends on the similarities of the past traffic conditions (Smith et al., 2002). Modified time series such as adjustments based on wavelet transform ARIMA model (Huajun et al., 2010) can outperform traditional time series models (e.g. AR, ARMA, ARIMA etc.) owing to their capabilities to account for the non-stationarity nature in the traffic conditions (Vlahogianni et al., 2006).

Travel time recorded on a regular time interval can be treated as time series data. A time series model usually captures three components: trend, seasonality and irregularity (Harvey, 1990). "*The principle structural time series models are nothing more than regression models in which the explanatory variables are functions of time and the parameters are time varying"* (Harvey, 1990; p-10). Harvey also claims that the time series models need to be handled by the state space form that may include trend, seasonality and disturbance. When a new observation becomes

available, state can be updated by applying the Kalman filtering technique. The state space form empowers the model by making it extendable to handle data irregularities. In this study, three different methods: Kalman Filter, K-NN and Boosting have been applied to predict DTT from ATT.

Kalman Filter (KF)

KF combines all available observations and prior knowledge of a system in such a manner so that the errors are minimized statistically (Maybeck, 1990). KF is directly applicable in signal processing and control system because the prime goal is to filter out the noise form observations when they are available. This filtering technique has also been widely applied to time series forecasting with state space model (Durbin and Koopman, 2012, Hamilton, 1994, Harvey, 1990). In a time series model, the state space model could be formed following any procedure (e.g. linear regression, ARIMA) which predicts next step. The prediction will then be optimized based on KF when an observation becomes available. Since problems related to time series data deal with the prediction of k-step at its previous (k-1) step, the latest available observation z_{k-1} instead of z_k is applied in the filtering equation. In a time series model, the update equation (eq. 5.2) (Commandeur and Koopman, 2007, Durbin and Koopman, 2012) is expressed by,

$$x_t = x_{t-1} + K_{t-1}(z_{t-1} - x_{t-1}) \tag{5.2}$$

This equation clearly states a time latency: the next step is predicted using the current step. Therefore, sufficient knowledge about state space and noises is important for better prediction performance.

Although predicting DTT from ATT has two advantages over a conventional time series data: a) two data sources (ATT & DTT) and b) negligible time latency, sufficient knowledge about state space and noises are still key issues. Researchers (Chen and Chien, 2001, Chien and Kuchipudi, 2003, Nanthawichit et al., 2003, Yang, 2005, Chen and Rakha, 2014b) modeled the state-space/process of travel time prediction as a linear system to which KF was applied. Similar approach has been followed in this study. In following section, the proposed formulation of KF model has been discussed. Discussion also includes relevant assumptions and techniques to overcome the challenges related to applying KF in travel time prediction.

Proposed Formulation of KF Model: ATT is selected for observation, as it is the observation nearest the DTT to be predicted. ATT is assumed to be related to DTT linearly:

$$ATT_t = DTT_t + v_t \tag{5.3}$$

where, v denotes the observation noise.

The current traffic condition is more highly correlated with close (current) conditions than it is with distant conditions. In other words, the traffic conditions of any specific time will always contain more information about the conditions nearest that time. The proposed state-space equation is:

$$DTT_t = \Phi_{t-1} DTT_{t-1} + w_{t-1} \tag{5.4}$$

And the transition function, Φ_{t-1}:

$$\Phi_{t-1} = {DTT_{t-1}}/{DTT_{t-2}} \tag{5.5}$$

where w denotes the state-space noise.

The proposed KF model based on (Welch and Bishop, 2006) is described below:

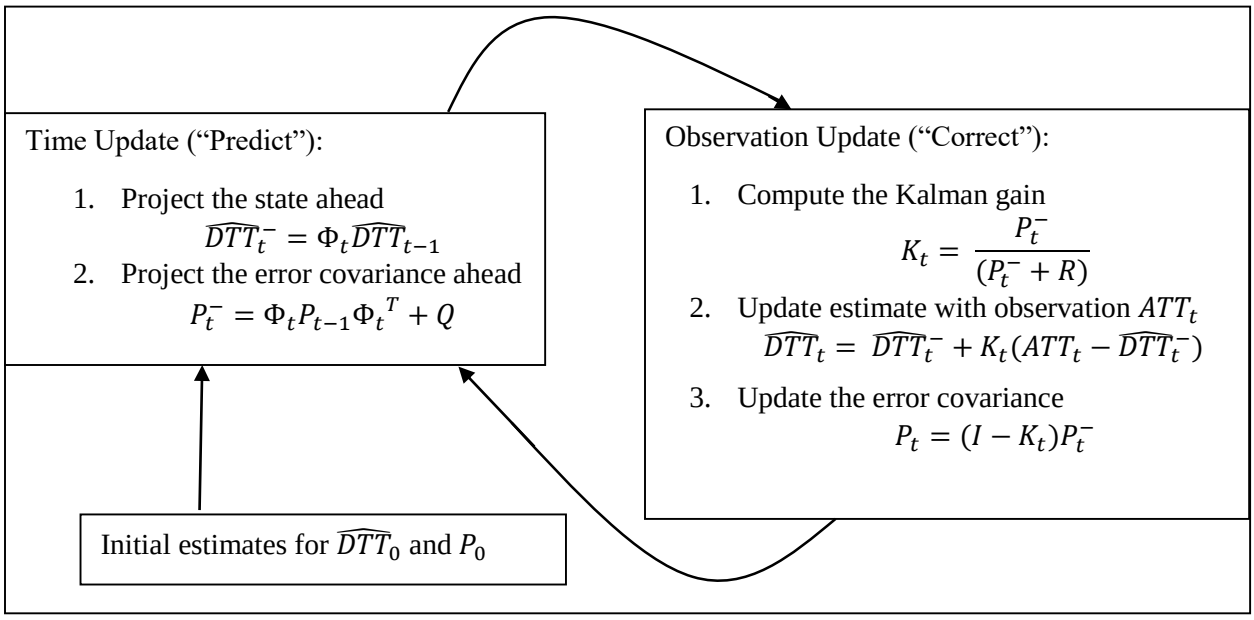


FIGURE 14 KF model.

The most recent DTT ($t-1\ and\ t-2\ steps$) is unavailable until the vehicles have finished travelling the route; therefore, the DTT from the latest and same historical day of week at $t-1\ and\ t-2\ steps$ are used to estimate Φ_t. w and v are assumed to be independent of each other and follow the normal probability distributions: $p(w) \sim N(0, Q)$ and $p(v) \sim N(0, R)$.

A priori ($\widehat{DTT}_t^-$), according to its definition, should be equal to a corresponding DTT. Therefore, the differences between a priori (i.e. state-space projection) and DTT are considered the state-space noise. the state-space noise (w_t) is measured by:

$$w_t = DTT_t - \widehat{DTT}_t^- \tag{5.6}$$

where $\widehat{DTT}_t^-$ = Corresponding a priori of DTT at time t.

The observation - after the noise is removed - should be equal to the predicted DTT or a posterior, according to eq. (4). Therefore, the difference between observation and predicted time

(i.e. DTT-ATT) is considered as the observation noise. The observation noise (v_t) at time t can be expressed as:

$$v_t = DTT_t - ATT_t \tag{5.7}$$

Considering computational simplicity and the availability of sufficient data, noise from historic all-days (instead of same-days) was used to estimate the noise covariance. Assumptions regarding the temporal characteristics of noise can be categorized into three types:

a) Steady Noise (SN): Regardless of time of day, noise is assumed to be the same for a day and estimated from the complete training dataset.
b) Contextual Noise (CN): Noise is assumed to vary by traffic state (free-flow, delay, recurrent, and non-recurrent congestions). Four covariance matrices (Q_{ff}, Q_{dl}, Q_{rc} andQ_{nrc}) are estimated depending on traffic conditions.
c) Time-varying Noise (TVN): Noise is assumed to vary with every time step of prediction; hence, the covariance matrix (Q_t) is estimated by the noise of a training dataset at time t.

K-Nearest Neighbor Method

Travel time at any timestamp has close relationship with the travel time of its close proximity in time. Therefore, DTT is modeled by the nearest ATT.

$$DTT_t = ATT_t + \Delta tt_t \tag{5.8}$$

where Δtt_t = predicted difference of ATT and DTT at a time-step t.

Travel time difference, Δtt_t is predicted by the distance weighted k-NN method using historic all day data:

$$\Delta tt_t = \left(\frac{\sum w_d \Delta tt_d}{\sum w_d}\right)_t \tag{5.9}$$

where $\Delta tt_{d.t}$ is the difference between ATT and DTT on a historic day, d at a time-step, t and $w_{d.t}$ is the corresponding weight which is the measure of the similarity between two traffic patterns: the traffic pattern of the present day and of the historic day, d. This similarity is the reciprocal of the measure of the variation between those two traffic patterns. This variation is measured by the Euclidian squared distance of two n-dimensional vectors representing the latest travel times (ATT of n-steps) from the present day, p and historic day, d: $[ATT_t \ ATT_{t-1} \ ATT_{t-2} \dots \dots \dots \ ATT_{t-n+1}]_p$ and $[ATT_t \ ATT_{t-1} \ ATT_{t-2} \dots \dots \dots \ ATT_{t-n+1}]_d$. Since, the n-steps should be the steps that have the most significant impact on the current step t to reflect the traffic pattern, the determination of n is a heuristic approach. In this study, travel time was predicted for n = 1, 2, 3, …, 10.

Boosting: LSBoost

A weak classifier may perform well for a particular group of data; but may not be able to classify another group of data. If a set of weak classifiers perform better in individual subsets of a whole dataset, then they may collectively produce a strong learner. This is the idea behind the boosting influenced by the investigation of the limitations in learning Boolean functions (Kearns and Valiant, 1994). The boosting algorithm combines weak learners by a multiplicative weight-update technique to create a strong learner that does not require any prior knowledge about the performance of the weak learning algorithm (Freund and Schapire, 1997).

Based on above foundation, boosting algorithm has several forms including AdaBoost.M1 (for binary classification), AdaBoost.M2 (for more than two classifications), RobustBoost RUSBoost, and LSBoost or Bag (for regression). RobustBoost was designed to handle the challenges induced by outliers (Freund, 2009)Dealing with the data having class imbalance is another challenge in data science. Seiffert et al. introduced a novel hybrid algorithm to alleviate

the problem of class imbalance and named the method random undersampling boosting (RUSBoost) (Seiffert et al., 2008). LSBoost is a least square regression boost approach that fits regression ensembles to minimize mean-squared error. At every step, a new learner is fitted to the difference between the observed response and the aggregated prediction of all learners grown previously. The main difference between adaBoost and LSBoost is the replacement of the exponential loss function with the least square loss function (Friedman, 2001). The following figure represents the algorithm:

$Initialize, F_0(x) = \bar{y}$

$for\ m = 1\ to\ M\ do$:

$\tilde{y}_i = y_i - F_{m-1}(x_i),\ i = 1, N$

$(\rho_m, \alpha_m) = \arg\min_{\rho,\alpha} \sum_{i=1}^{N} [\tilde{y}_i - \rho h(x_i; \alpha)]^2$

$F_m(x) = F_{m-1}(x) + \rho_m h(x; \alpha_m)$

$endfor$

Output the final regression function $F_m(x)$.

FIGURE 15 The lest square regression boost algorithm (Friedman, 2001).

Measuring Prediction Performance

Commonly used methods to measure the performance indices are mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE). These indices can be defined by the following equations:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|\widehat{T}_t - T_t\right| \tag{5.10}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\widehat{T}_t - T_t}{T_t}\right| \times 100\% \tag{5.11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\widehat{T}_t - T_t)^2} \tag{5.12}$$

where n is the number of observations, and T_t and $\widehat{T}_t$ are the actual and predicted travel times at time t on a particular link.

Results and Discussion

The complete dataset in this study was divided into two sets: Training and Validation. Twenty-eight of forty-seven days' worth of data was used for the training dataset, and the rest of the data was used as the validation dataset. It was more appropriate to use speed data as opposed to travel time due to the variability in link lengths. The prediction performance index was utilized similar to the mean absolute error (MAE) and root-mean-square error (RMSE) to perceive global performance (i.e. the performance of the entire network). The mean absolute percentage error (MAPE) was calculated based on the travel time dataset in order to discern the local performance at each link. Quantifying improvement is impossible without knowing the actual mean absolute percentage error (AMAPE) or actual lag/gap, considering that using ATT as the prediction of DTT

is a naïve method. AMAPE (i.e. MAPE of ATT) was used as the benchmark for MAPEs generated by other methods.

Kalman Filter

Spatial noise characteristics for KF models should be assumed by considering that a) noise of different links in a particular corridor can have similar characteristics, and b) noise of different links, regardless of corridor, can have different characteristics. Previously, three categorical assumptions regarding temporal characteristics of noise have been discussed; therefore, the output of the KF model would be affected by six different estimation procedures of noise covariance regarding the spatial-temporal characteristics of noise. The six methods are: corridor-based steady noise (CB-SN), contextual noise (CB-CN), time-varying noise (CB-TVN), link-based steady noise (LB-SN), contextual noise (LB-CN), and time-varying noise (LB-TVN).

Appropriate noise characteristics of KF should be determined through the evaluation of local performance. FIGURE 16, FIGURE 17, and FIGURE 18 represent the local (i.e. each link) performance of the KF model for validation dataset with different noise assumptions. The MAPE of the prediction for each link, when evaluating criteria of local performance, should be smaller than that of the actual lag/gap (AMAPE or the benchmark).

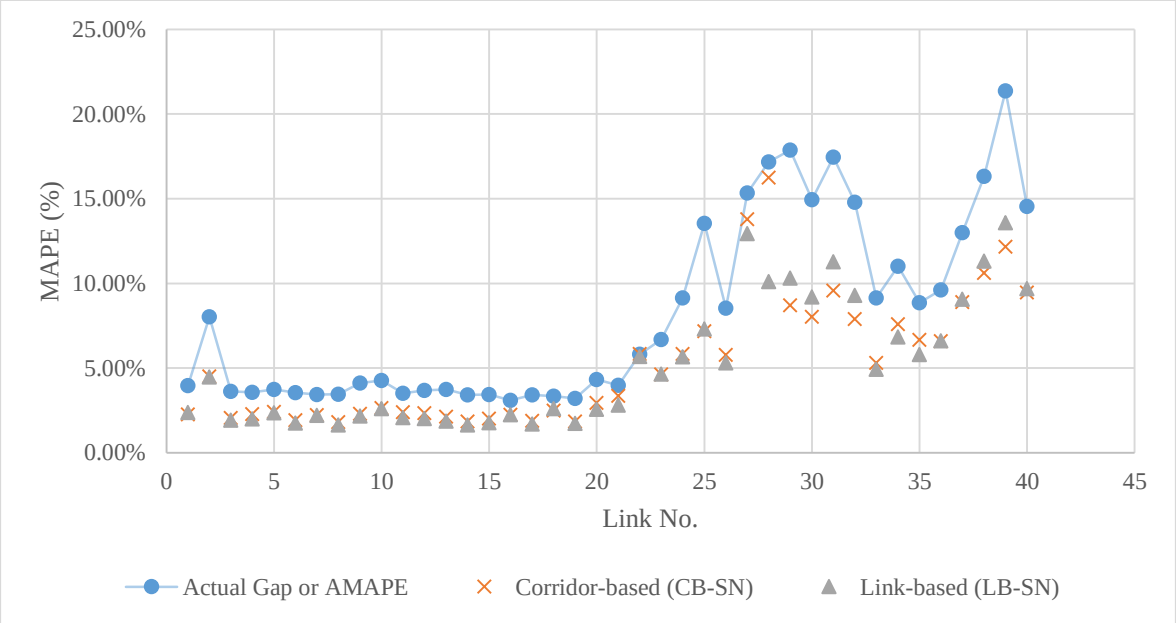


FIGURE 16 MAPE of KF at each link for steady noise assumption vs. actual gap.

In FIGURE 16, MAPEs of prediction by the KF model using CB-SN and LB-SN are compared to the actual lag/gap. The prediction performances of the KF model using both CB-SN and LB-SN for each individual link are acceptable, as no link shows a MAPE greater than the actual lag/gap.

In FIGURE 17, MAPEs of travel times predicted by the KF model using CB-CN and LB-CN are compared to the actual lag/gap. It is clear that the KF model with both CB-CN and LB-CN has a few links' MAPE greater than AMAPE. In general, the context-based noise assumption is supposed to perform better for the corridor that experiences congestion. The poor performance could be due to the stability of travel time resulting from the saturated traffic flow rate under congestions. At a saturated flow, the variations between ATT and DTT become similar to the variations in free flow or delay conditions.

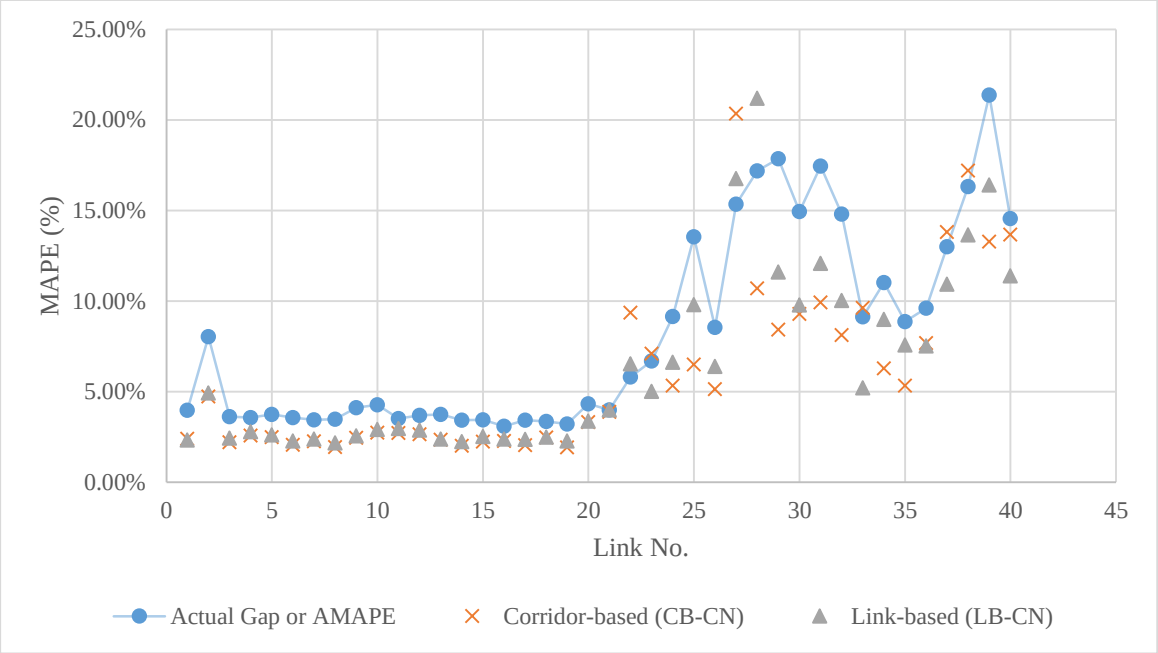


FIGURE 17 MAPE of KF at each link for contextual noise assumption vs. actual gap.

In FIGURE 18, MAPEs of travel times predicted by the KF model using CB-TVN and LB-TVN are compared to the actual lag/gap. The assumption of CB-TVN is invalid for a corridor with some links that experience traffic congestion. Despite the improvement from contextual noise assumption, FIGURE 18 shows that the noise homogeneity assumption of CB-TVN is violated at Link 28, as its performance exceeds the actual lag/gap (AMAPE).

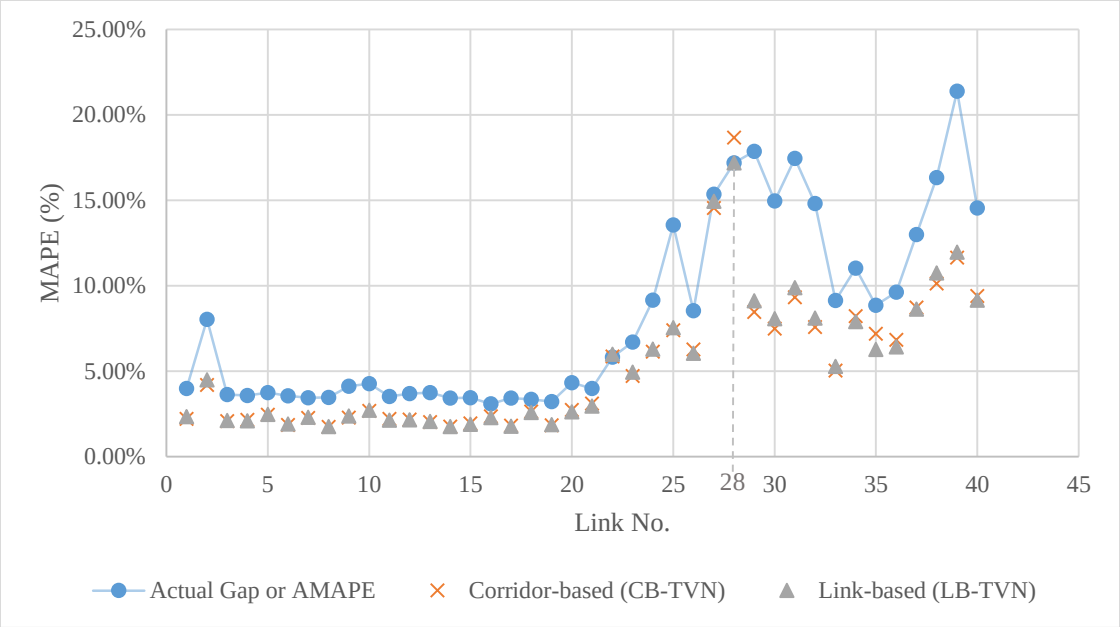


FIGURE 18 MAPE of KF at each link for time varying noise assumption vs. actual gap.

The above discussion provides a comprehensive description of local/link performance of different noise assumptions in a KF model. The unambiguous analyses reflect the suitability of CB-SN, LB-SN and LB-TVN. TABLE 3 shows the global performance of prediction expressed by the MAE and RMSE of speed data calculated from travel time, which reaffirms the most appropriate noise assumption for the dataset is LB-SN.

TABLE 3 Overall (Global) Performance of KF Model for Selected Noise Assumptions

Noise Assumption	Training Dataset		Validation Dataset	
	MAE	RMSE	MAE	RMSE
CB-SN	2.25	6.31	2.41	6.75
LB-SN	2.17	5.96	2.31	6.33
LB-TVN	2.15	6.78	2.45	7.47

The KF model with LB-SN assumption has a better prediction performance and more computational simplicity, which leads to operational efficiency in terms of run time. For instance,

the run time of KF with corridor-based noise assumptions is approximately 15mins, whereas link-based assumptions take only a minute or two. Corridor-based noise homogeneity assumption calls for extra processing of data since Bluetooth data is collected over each link; this way, noise covariance can be estimated over the entire corridor. This extra processing increases the run time significantly. The link-based model is the most suitable for an on-line application. KF with LB-SN was selected to predict DTT from ATT.

The MAPEs of the training and validation datasets using LB-SN KF are 4.27% and 4.53%, respectively, whereas, the actual lags/gaps (i.e. AMAPEs) are 6.43% and 6.70%, respectively, based on travel time dataset. Improvements are significant when the smaller actual lag/gap is considered, including a 50% reduction in AMAPE in some links. When the KF model with LB-SN assumption is applied, the validation dataset shows that a 40-50% gap between ATT and DTT is minimized in different links of the I-90 corridor, and a nearly 30-40% gap is minimized in different links of the Beltline Highway corridor.

The KF model with the LB-SN assumption has a superior prediction performance in cases of traffic state transition (e.g. onset and end of congestion). FIGURE 19(a) represents the DTT, ATT, and predicted travel time, and FIGURE 19(b) demonstrates the actual lag/gap and prediction error corresponding to FIGURE 19(a).

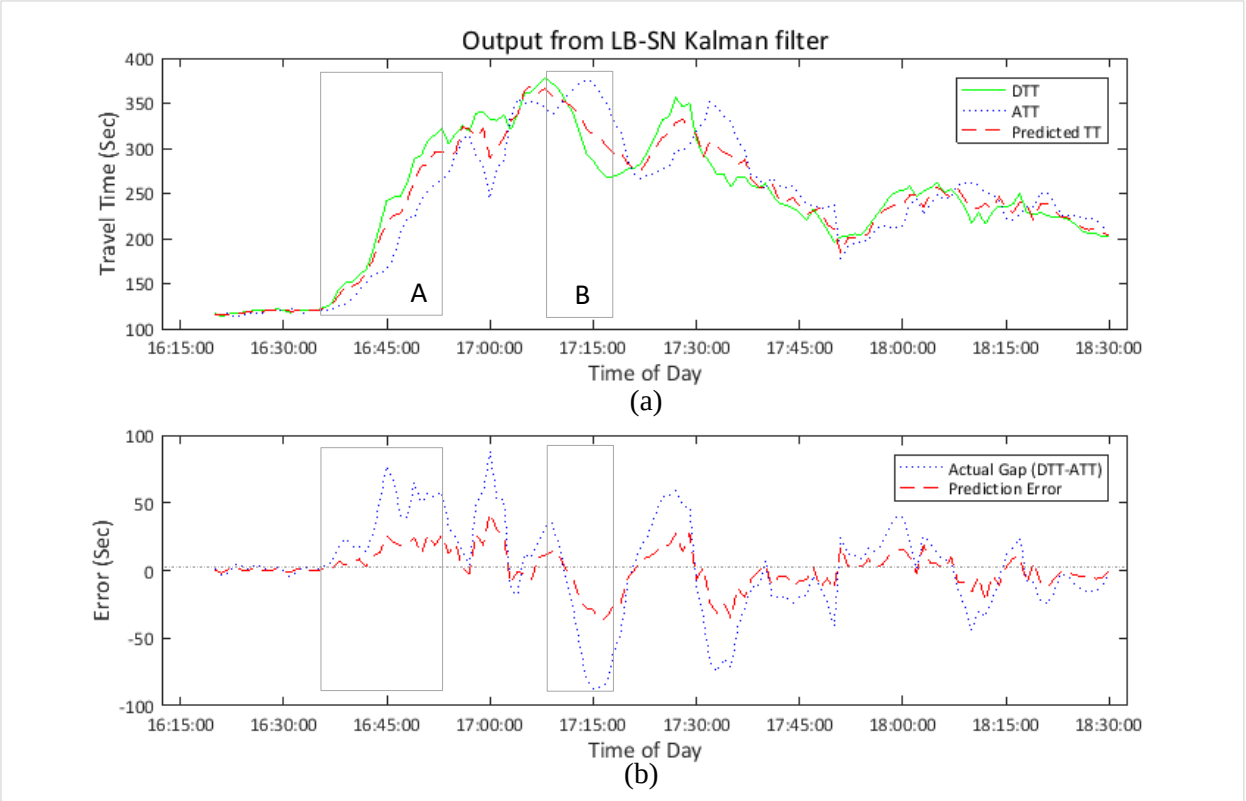


FIGURE 19 Prediction performance between free flow and congested conditions.

FIGURE 19 clearly depicts that the ATT, DTT, and predicted travel time are almost equal at free flow conditions (before 16:30:00), and are thus error free. Actual and prediction errors are negligible after 18:15:00 when congestion is stable. FIGURE 19(a) Box A shows the onset of congestion where the actual lag/gap is more than 50s, and FIGURE 19(b) shows prediction lag is less than 20s. Box B in FIGURE 19(a) shows the end of the congestion situation where the actual and prediction lags show little difference (less than 20s). However, at 17:15:00, a mid-point of the end of congestion, the ATT is off by 100s from DTT while the prediction error is around 30s. Despite of having a moderate improvement (around 40%) according to MAPE (global performance), the prediction shows excellent improvement (around 70%) in cases such as that in box B where there is a state transition. Since such cases cover shorter time periods compared to the complete study period, the overall performance (local or global) indices are unable to represent

the robustness of the prediction algorithm. Therefore, the selected noise assumption based on the KF model is capable of predicting travel time from the travelers' point of interest (the onset and end of congestion rather than free-flow or stable congested conditions).

K-Nearest Neighbor (k-NN) Method

A heuristic method has been applied for selecting distance vector in k-NN. Distance has been estimated for different n values (n = 1, 3 and 5). Prediction results (for validation dataset) of k-NN for different distance vectors has been presented in the figure below:

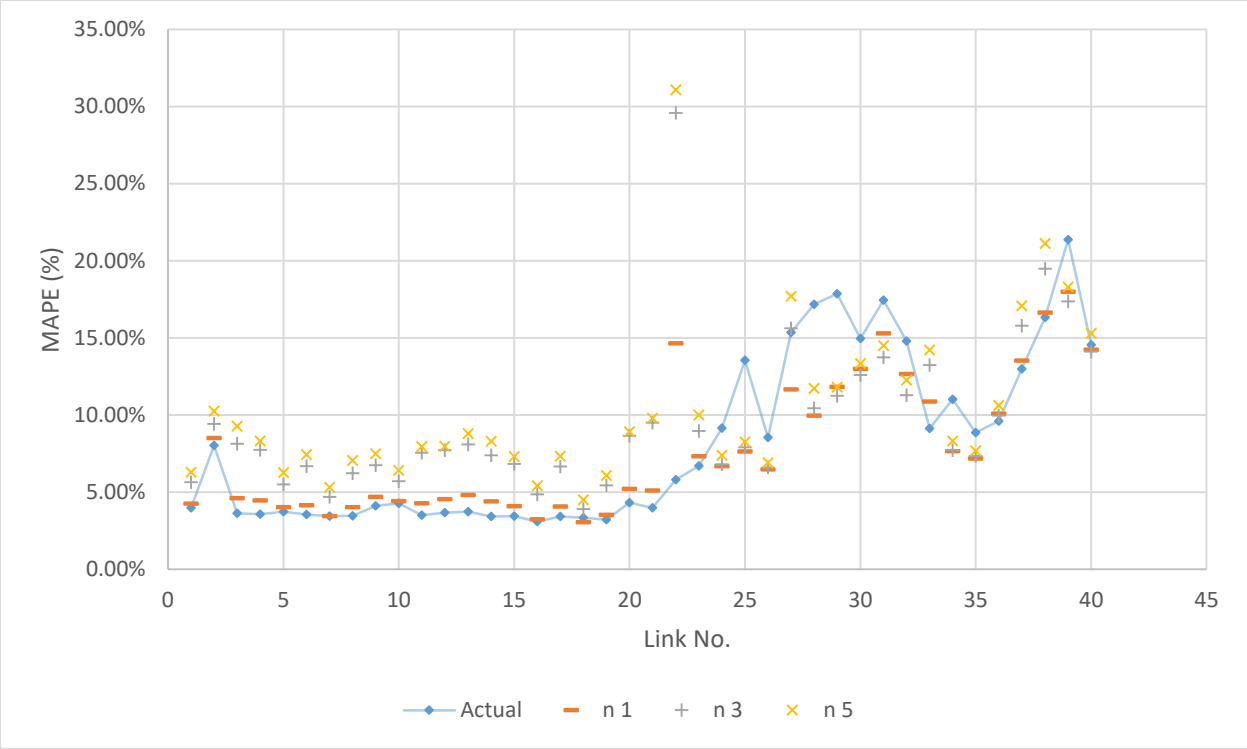


FIGURE 20 MAPE of k-NN model at each link for prediction vs. actual gap.

The graph shows that the k-NN performs poorly in I-90, a corridor that generally experiences free flow and comparatively better in Beltline Hwy, a corridor that generally faces recurrent congestion. However, overall average of performance decreases i.e. MAPE increases with the increase in n

value. The minimum MAPE (7.29%) is higher than the actual gap or AMAPE (6.70%) which actually refers to the lack of repeating traffic condition over the entire period. Therefore, k-NN is not an appropriate method to predict DTT from ATT when the traffic conditions in training dataset are not similar to the conditions in validation dataset.

Boosting: LSBoost

LSBoost has been applied to travel time dataset with an expectation that the Boosting could possibly be an alternative to KF for predicting DTT from ATT. In case of boosting, few variables (day of the week, time of day and previous n-step ATT) have been added to assist prediction instead of using ATT only. Different values of n (n = 2, 3, 5, 10) has been considered to select previous ATT. Thus, LSBoost has been applied to the dataset having variables: day of the week, time of day, DTT, several (n-steps) ATT. Unfortunately, no LSBoost model outperformed the selected KF model (LB-SN). However, the performance on different link of Beltline Highway corridor is comparatively much better than the corridor of I-90. Moreover, the results indicate that the n-value (i.e. parameter to combine several minute data) has negligible effects on prediction performance. FIGURE 21 represents the outcome of LSBoost in predicting travel time for validation dataset. To ensure visual clarity, only two selected results with the actual gap has been shown in the graph:

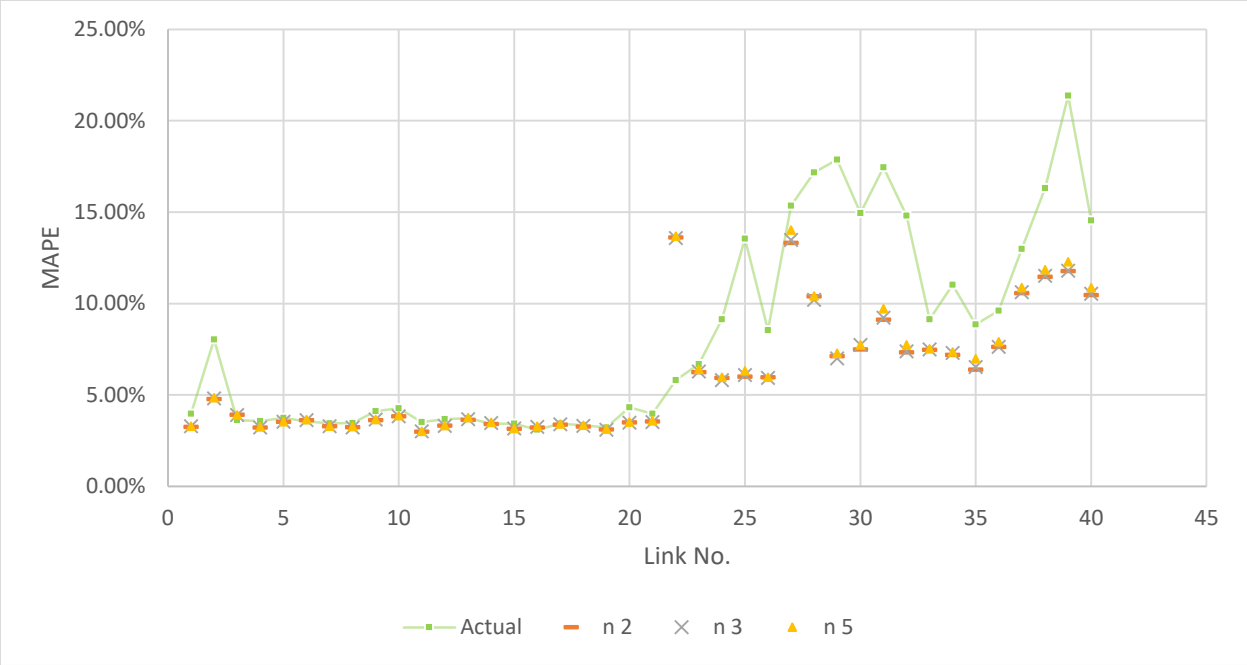


FIGURE 21 MAPE of LSBoost at each link for prediction vs. actual gap.

In the graph above, the line of Actual Gap is the threshold for MAPE after applying LSBoost algorithm. A link (22nd), a part of both I-90 and Beltline corridor, performed very poorly showing MAPE higher than that of AMAPE. Prediction has negligible effect on the I-90 since MAPE and AMAPE are mostly same. However, the performance of LSBoost prediction is notably well in Beltline Hwy. To decide whether LSBoost algorithm could be a better alternative of other prediction methods due to its prediction superiority in a congested corridor, extensive experiment using Bluetooth travel time data from the corridors experiencing different level of congestion is required. Therefore, LSBoost is not recommended at this point to predict DTT from ATT for freeways.

CHAPTER VI. CONCLUSION

In general, sample size is considered as the accurate sampling criteria which is extremely easy to apply in a controlled system or study. In an uncontrolled system (e.g. Bluetooth or Cellular probe), sample size based sampling unit would yield a variable sampling interval which will create complexity in advanced tasks (e.g. prediction of future travel time). This thesis introduced an empirical method to avoid the complexity by following the aforementioned quantitative approach. It demonstrated the varying nature of properties over the aggregation by different time intervals. The properties include the reliability of travel time which is affected by the sample size, the availability of travel time which is affected by the availability of sample, and the stochasticity of travel time which is an inherit property. The major advantage of this research is the scope of applying engineering judgement at some points.

This thesis proposed a framework for an excessively challenging task of selecting a sampling interval ensuring accuracy, reliability and preserving the primary property within a tangible proximity. This framework is applicable in any process, especially, uncontrolled process that requires sampling interval instead of sample size. Computation process of the high confidence sample rate and sample penetration rate would be directly applicable to other studies while in some cases a little alteration would be inevitable in computing the succession i.e. the inheritance of a foremost property.

For selected sampling interval, ATT needs to be estimated. ATT is the most available form of travel time, but DTT is the most desirable. The KF algorithm was used to predict DTT and thus assist motorists by providing a more accurate and reliable travel time. Although ATT and DTT

differ slightly depending on the flow of traffic, the variation becomes significant when the traffic state is in transition (e.g. moving from unstable to stable). The KF algorithm with steady noise assumption captured a state of transition property accurately and provided an excellent prediction. KF is exceptionally fast for link-based applications, making it extremely desirable for data sources that contain route travel time split into shorter links (e.g. loop detectors/Bluetooth data). Although the KF is applied (by default) to each link that is isolated as a different model, it demonstrates a higher level of accuracy and faster speed due to its flexibility, simplicity and compatibility with data characteristics. The application was demonstrated during peak periods on freeways covering two corridors – one with fewer transitions in traffic state and another with frequent transitions. Diversity in noise assumptions showed negligible impact on the former, while steady noise assumption showed better performance on the latter. Steady noise (SN) refers to a fixed covariance estimated from the complete training dataset, which indicates that the KF model performs better with the generalized noise assumption. Hence, the KF with LB-SN assumption was preferred over other assumptions. Exclusive performance was observed during a time when transitions in traffic state occurred more frequently. Since arterial highways are supposed to possess more state transitions, future research should examine the performance of predicting arterial highway travel time to test the robustness of this method.

Major Contributions

The major contributions of this research include the practical solutions to the critical problems relating to appropriate sampling interval selection and reliable travel time prediction using Bluetooth data. These contributions are discussed as follows:

Processing Big Data

This research has introduced a general guideline as well as a framework to process Big Data which is extremely simple in shape but exceptionally complex in processing due to containing overly redundant records. Presented processing principles are reusable in any continuous data collection system that produce Big Data like Bluetooth technology.

Sampling Interval Selection Framework

This research has demonstrated a framework to address the most challenging issue regarding the sampling requirements of the travel time estimation. Use of minimum sample size is the traditional method of ensuring data quality, which is inapplicable in travel time estimation due to the complexity associated with its further processing. This demanding task which was untouched for long time, has been confined in a simple frame to make the application easy for practitioners.

Outlier Filtering Algorithm

The study represents a simple, effective and fast filtering algorithm to remove the outliers. The proposed technique is based on well-known outlier filtering application proposed by (Dion and Rakha, 2006) and hence, reasonably accurate with extremely suitability in online application.

ATT to DTT Prediction

The study portrays the DTT as the true travel time that needs to be predicted. Different prediction methods including machine learning techniques was examined to guarantee the selection of the best method and eventually, a simple and better performing (in terms of, accuracy and speed)

method was nominated. Noise variation in KF method was analyzed critically and thus, the power of the noise assumptions has been revealed.

Online Applicability of Prediction Method

The entire process of predicting the true travel time (DTT) consists of several modules: raw data processing, outlier filtering, ATT estimation and finally, DTT prediction. This research recommended simple and online application friendly procedure for each module individually. Therefore, the complete process of producing DTT from raw Bluetooth data would be easily applicable online.

Future Research

Pertaining to this thesis, the potential research areas that should be the part of future research endeavors are presented below:

1. Due to the unavailability of signal strength in Bluetooth data, first detection (or first signal) was used to estimate travel time. A comparative analysis among travel times estimated from highest signal, first signal and a more reliable (ground-truth) source is imperative to compute the error rate.

2. The research was conducted on the freeways data. Arterial highways usually experience considerable variation in traffic state which is absolutely different than a traffic state on freeways. Therefore, the repetition of this study for an arterial highway is proposed.

3. Factoring the delay causality into travel time prediction would be another area of exploration and integration to prediction model. Crash-induced congestion has greatest

impact on travel time including a sudden sharp increase. If the delay causality can be predicted in a real time fashion with an acceptable accuracy, then the incident information inclusive prediction method would expectedly provide more accurate prediction.

REFERENCES

ABEDI, N., BHASKAR, A. & CHUNG, E. 2013. Bluetooth and Wi-Fi MAC address based crowd data collection and monitoring: benefits, challenges and enhancement.

AL-DEEK, H., D'ANGELO, M. P. & WANG, M. Travel time prediction with non-linear time series. Fifth International Conference on Applications of Advanced Technologies in Transportation Engineering, 1998.

ANBAROGLU, B., HEYDECKER, B. & CHENG, T. 2014. Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies,* 48**,** 47-65.

ARAGHI, B. N., PEDERSEN, K. S., CHRISTENSEN, L. T., KRISHNAN, R. & LAHRMANN, H. 2015. Accuracy of travel time estimation using Bluetooth technology: Case study Limfjord tunnel Aalborg. *International Journal of Intelligent Transportation Systems Research,* 13**,** 166-191.

ARUP, O., BATES, J., FEARON, J. & BLACK, I. 2004. Frameworks for modelling the variability of journey times on the highway network. Report for Department of Transport, London, UK.

ASUDEGI, M. 2009. *Optimal number and location of Bluetooth sensors for travel time data collection in networks.*

BACHMANN, C., ABDULHAI, B., ROORDA, M. J. & MOSHIRI, B. 2013. A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. *Transportation research part C: emerging technologies,* 26**,** 33-48.

BHASKAR, A., CHUNG, E. & DUMONT, A. G. 2011. Fusing loop detector and probe vehicle data to estimate travel time statistics on signalized urban networks. *Computer-Aided Civil and Infrastructure Engineering,* 26**,** 433-450.

BONÉ, R. & CRUCIANU, M. 2002. Multi-step-ahead prediction with neural networks: a review. *9emes rencontres internationales: Approches Connexionnistes en Sciences,* 2**,** 97-106.

BRENNAN JR, T. M., ERNST, J. M., DAY, C. M., BULLOCK, D. M., KROGMEIER, J. V. & MARTCHOUK, M. 2010. Influence of vertical sensor placement on data collection efficiency from bluetooth MAC address collection devices. *Journal of Transportation Engineering,* 136**,** 1104-1109.

BUSTILLOS, B. & CHIU, Y.-C. 2011. Real-time freeway-experienced travel time prediction using N-curve and k nearest neighbor methods. *Transportation Research Record: Journal of the Transportation Research Board***,** 127-137.

CARRION, C. & LEVINSON, D. 2012. Value of travel time reliability: A review of current evidence. *Transportation research part A: policy and practice,* 46**,** 720-741.

CHEN, H. & RAKHA, H. Agent-based modeling approach to predict experienced travel times. Transportation Research Board 93rd Annual Meeting, 2014a.

CHEN, H. & RAKHA, H. 2015. Real-Time Freeway Travel-Time Prediction. *Engineering & Technology Reference,* 1.

CHEN, H. & RAKHA, H. A. 2014b. Real-time travel time prediction using particle filtering with a non-explicit state-transition model. *Transportation Research Part C: Emerging Technologies,* 43**,** 112-126.

CHEN, M. & CHIEN, S. 2000. Determining the number of probe vehicles for freeway travel time estimation by microscopic simulation. *Transportation Research Record: Journal of the Transportation Research Board,* 61-68.

CHEN, M. & CHIEN, S. 2001. Dynamic freeway travel-time prediction with probe vehicle data: Link based versus path based. *Transportation Research Record: Journal of the Transportation Research Board,* 157-161.

CHIEN, S. I.-J. & KUCHIPUDI, C. M. 2003. Dynamic travel time prediction with real-time and historic data. *Journal of transportation engineering,* 129**,** 608-616.

CLARK, S., GRANT-MULLER, S. & CHEN, H. 2002. Cleaning of matched license plate data. *Transportation Research Record: Journal of the Transportation Research Board,* 1-7.

CLARK, S. & WATLING, D. 2005. Modelling network travel time reliability under stochastic demand. *Transportation Research Part B: Methodological,* 39**,** 119-140.

CLICK, S. M. & LLOYD, T. Applicability of bluetooth data collection methods for collecting traffic operations data on rural freeways. Transportation Research Board 91st Annual Meeting, 2012.

COMMANDEUR, J. J. & KOOPMAN, S. J. 2007. *An introduction to state space time series analysis*, OUP Oxford.

DAY, C. M., BRENNAN, T. M., HAINEN, A. M., REMIAS, S. M. & BULLOCK, D. M. 2012. Roadway system assessment using Bluetooth-based automatic vehicle identification travel time data.

DION, F. & RAKHA, H. 2006. Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. *Transportation Research Part B: Methodological,* 40**,** 745-766.

DOWLING, R., SKABARDONIS, A., CARROLL, M. & WANG, Z. 2004. Methodology for measuring recurrent and nonrecurrent traffic congestion. *Transportation Research Record: Journal of the Transportation Research Board,* 60-68.

DURANTON, G. & TURNER, M. A. 2012. Urban growth and transportation. *The Review of Economic Studies,* 79**,** 1407-1440.

DURBIN, J. & KOOPMAN, S. J. 2012. *Time series analysis by state space methods*, Oxford University Press.

EDWARDS, T. & SMITH, S. 2008. *Transport problems facing large cities*.

ELHENAWY, M., CHEN, H. & RAKHA, H. A. 2014. Dynamic travel time prediction using data clustering and genetic programming. *Transportation Research Part C: Emerging Technologies,* 42**,** 82-98.

ELIASSON, J. Car drivers' valuations of travel time variability, unexpected delays and queue driving. Proceeding of European Transport Conference, 2004, 2004.

ELIASSON, J. Forecasting travel time variability. European Transport Conference, 2006.

FEI, X., LU, C.-C. & LIU, K. 2011. A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. *Transportation Research Part C: Emerging Technologies,* 19**,** 1306-1318.

FREUND, Y. 2009. A more robust boosting algorithm. *arXiv preprint arXiv:0905.2138*.

FREUND, Y. & SCHAPIRE, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences,* 55**,** 119-139.

FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics***,** 1189-1232.

GARIB, A., RADWAN, A. & AL-DEEK, H. 1997. Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering,* 123**,** 459-466.

GUO, F., RAKHA, H. & PARK, S. 2010. Multistate model for travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board***,** 46-54.

GUO, J., HUANG, W. & WILLIAMS, B. M. 2014. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C: Emerging Technologies,* 43**,** 50-64.

HAGHANI, A., HAMEDI, M., SADABADI, K., YOUNG, S. & TARNOFF, P. 2010. Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record: Journal of the Transportation Research Board***,** 60-68.

HAMED, M. M., AL-MASAEID, H. R. & SAID, Z. M. B. 1995. Short-term prediction of traffic volume in urban arterials. *Journal of Transportation Engineering,* 121**,** 249-254.

HAMILTON, J. D. 1994. *Time series analysis,* Princeton university press Princeton.

HANDBOOK, T. D. 2006. -Volume-I, US Dept. *Transp., Fed. Highway Admin., Washington, DC*.

HARVEY, A. C. 1990. *Forecasting, structural time series models and the Kalman filter,* Cambridge university press.

HAWAS, Y. E. 2007. A fuzzy-based system for incident detection in urban street networks. *Transportation Research Part C: Emerging Technologies,* 15**,** 69-95.

HENSHER, D. A. & TRUONG, T. P. 1985. Valuation of travel time savings: a direct experimental approach. *Journal of Transport Economics and Policy***,** 237-261.

HUAJUN, W., LEI, S. & HONGYING, L. Adjustments based on wavelet transform ARIMA model for network traffic prediction. Computer Engineering and Technology (ICCET), 2010 2nd International Conference on, 2010. IEEE, V4-520-V4-523.

JIANG, G., GANG, L. & CAI, Z. Impact of probe vehicles sample size on link travel time estimation. Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE, 2006. IEEE, 505-509.

KEARNS, M. & VALIANT, L. 1994. Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM (JACM),* 41**,** 67-95.

KHATTAK, A., POLYDOROPOULOU, A. & BEN-AKIVA, M. 1996. Modeling revealed and stated pretrip travel response to advanced traveler information systems. *Transportation Research Record: Journal of the Transportation Research Board***,** 46-54.

KIM, J., RHO, J. & PARK, D. 2009. On-line estimation of departure time-based link travel times from spatial detection system. *International Journal of Urban Sciences,* 13**,** 63-80.

KOUWENHOVEN, M., DE JONG, G. C., KOSTER, P., VAN DEN BERG, V. A., VERHOEF, E. T., BATES, J. & WARFFEMIUS, P. M. 2014. New values of time and reliability in passenger transport in The Netherlands. *Research in Transportation Economics,* 47**,** 37-49.

LEVINSON, D. 2003. The value of advanced traveler information systems for route choice. *Transportation Research Part C: Emerging Technologies,* 11**,** 75-87.

LI, C.-S. & CHEN, M.-C. 2014. A data mining based approach for travel time prediction in freeway with non-recurrent congestion. *Neurocomputing,* 133**,** 74-83.

LI, R. & ROSE, G. 2011. Incorporating uncertainty into short-term travel time predictions. *Transportation Research Part C: Emerging Technologies,* 19**,** 1006-1018.

LI, W., CHUANJIU, W., XIAORONG, S. & YUEZU, F. Probe vehicle sampling for real-time traffic data collection. Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE, 2005. IEEE, 222-224.

LI, Y. & MCDONALD, M. Link travel time estimation using single GPS equipped probe vehicle. Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on, 2002. IEEE, 932-937.

LIN, W.-H. & DAGANZO, C. F. 1997. A simple detection scheme for delay-inducing freeway incidents. *Transportation Research Part A: Policy and Practice,* 31**,** 141-155.

LIU, H. 2008. *Travel time prediction for urban networks,* TU Delft, Delft University of Technology.

LIU, H., VAN ZUYLEN, H., VAN LINT, H. & SALOMONS, M. 2006. Predicting urban arterial travel time with state-space neural networks and Kalman filters. *Transportation Research Record: Journal of the Transportation Research Board***,** 99-108.

LU, Y. & CHANG, G.-L. 2012. Stochastic Model for Estimation of Time-Varying Arterial Travel Time and Its Variability with Only Link Detector Data. *Transportation Research Record: Journal of the Transportation Research Board***,** 44-56.

MALINOVSKIY, Y., LEE, U.-K., WU, Y.-J. & WANG, Y. Investigation of Bluetooth-based travel time estimation error on a short corridor. Transportation Research Board 90th Annual Meeting, 2011.

MARTCHOUK, M., MANNERING, F. L. & SINGH, L. 2010. Travel time reliability in Indiana.

MAYBECK, P. S. 1990. The Kalman filter: An introduction to concepts. *Autonomous robot vehicles.* Springer.

MEI, Z., WANG, D. & CHEN, J. 2012. Investigation with Bluetooth sensors of bicycle travel time estimation on a short corridor. *International Journal of Distributed Sensor Networks,* 2012.

MOGHADDAM, S. & HELLINGA, B. 2014a. Algorithm for detecting outliers in Bluetooth data in real time. *Transportation Research Record: Journal of the Transportation Research Board,* 129-139.

MOGHADDAM, S. & HELLINGA, B. 2014b. Real-time prediction of arterial roadway travel times using data collected by Bluetooth detectors. *Transportation Research Record: Journal of the Transportation Research Board,* 117-128.

MYUNG, J., KIM, D.-K., KHO, S.-Y. & PARK, C.-H. 2011. Travel time prediction using k nearest neighbor method with combined data from vehicle detector system and automatic toll collection system. *Transportation Research Record: Journal of the Transportation Research Board,* 51-59.

NANTHAWICHIT, C., NAKATSUJI, T. & SUZUKI, H. 2003. Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway. *Transportation Research Record: Journal of the Transportation Research Board,* 49-59.

NATIONAL TRANSPORTATION STATISTICS. 2016. *U.S. Department of Transportation, Bureau of Transportation Statistics, National Transportation Statistics* [Online]. Available: http://www.bts.gov/publications/national_transportation_statistics/ [Accessed May 21 2016].

ODA, T. An algorithm for prediction of travel time using vehicle sensor data. Road Traffic Control, 1990., Third International Conference on, 1990. IET, 40-44.

OH, C. & PARK, S. 2011. Investigating the effects of daily travel time patterns on short-term prediction. *KSCE Journal of Civil Engineering,* 15**,** 1263-1272.

PARLOS, A. G., RAIS, O. T. & ATIYA, A. F. 2000. Multi-step-ahead prediction using dynamic recurrent neural networks. *Neural networks,* 13**,** 765-786.

PORTER, J. D., KIM, D. S., MAGAÑA, M. E., POOCHAROEN, P. & ARRIAGA, C. A. G. 2013. Antenna characterization for Bluetooth-based travel time data collection. *Journal of Intelligent Transportation Systems,* 17**,** 142-151.

PUCKETT, D. D. & VICKICH, M. J. 2010. Bluetooth®-based travel time/speed measuring systems development.

QIU, Z. & CHENG, P. State of the art and practice: cellular probe technology applied in advanced traveler information system. 86th Annual Meeting of the Transportation Research Board, Washington, DC, 2007.

QUAYLE, S. & KOONCE, P. 2010. Arterial Performance Measures Using MAC Readers—Portland's Experience. *North American Travel Monitoring r_report. htm*.

RELIABILITY, T. T. 2006. Making It There on Time, All the Time. *Federal Highway Administration, USA*.

RICE, J. & VAN ZWET, E. 2004. A simple and effective method for predicting travel times on freeways. *Intelligent Transportation Systems, IEEE Transactions on,* 5**,** 200-207.

RITCHIE, S. G. & CHEU, R. L. 1993. Simulation of freeway incident detection using artificial neural networks. *Transportation Research Part C: Emerging Technologies,* 1**,** 203-217.

ROBINSON, S. & POLAK, J. 2006. Overtaking rule method for the cleaning of matched license-plate data. *Journal of transportation engineering,* 132**,** 609-617.

SCHAPIRE, R. E. 2003. The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*. Springer.

SCHRANK, D., EISELE, B., LOMAX, T. & BAK, J. 2015. Urban mobility scorecard. *College Station: Texas A&M Transportation Institute and INRIX*.

SEIFFERT, C., KHOSHGOFTAAR, T. M., VAN HULSE, J. & NAPOLITANO, A. RUSBoost: improving classification performance when training data is skewed. Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008. IEEE, 1-4.

SKABARDONIS, A. & GEROLIMINIS, N. 2005. Real-time estimation of travel times on signalized arterials.

SKABARDONIS, A., VARAIYA, P. & PETTY, K. 2003. Measuring recurrent and nonrecurrent traffic congestion. *Transportation Research Record: Journal of the Transportation Research Board***,** 118-124.

SMITH, B. L., WILLIAMS, B. M. & OSWALD, R. K. 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies,* 10**,** 303-321.

SOHN, K. & KIM, D. 2009. Statistical model for forecasting link travel time variability. *Journal of Transportation Engineering,* 135**,** 440-453.

SUMALEE, A., PAN, T., ZHONG, R., UNO, N. & INDRA-PAYOONG, N. 2013. Dynamic stochastic journey time estimation and reliability analysis using stochastic cell transmission model: Algorithm and case studies. *Transportation Research Part C: Emerging Technologies,* 35**,** 263-285.

SUSSMAN, J. M., PEARCE, V., HICKS, B., CARTER, M., LAPPIN, J. E., CASEY, R. F., ORBAN, J. E., MCGURRIN, M. & DEBLASIO, A. J. 2000. What Have We Learned About Intelligent Transportation Systems?

TOPPEN, A. & WUNDERLICH, K. 2003. *Travel time data collection for measurement of advanced traveler information systems accuracy*, Mitretek Systems.

TURNER, S. 1996. Advanced techniques for travel time data collection. *Transportation Research Record: Journal of the Transportation Research Board***,** 51-58.

TURNER, S. M., EISELE, W. L., BENZ, R. J. & HOLDENER, D. J. 1998. Travel time data collection handbook.

TURNER, S. M. & HOLDENER, D. J. Probe vehicle sample sizes for real-time information: The Houston experience. Vehicle Navigation and Information Systems Conference, 1995. Proceedings. In conjunction with the Pacific Rim TransTech Conference. 6th International VNIS.'A Ride into the Future', 1995. IEEE, 3-10.

VAN HINSBERGEN, C., VAN LINT, J. & SANDERS, F. Short term traffic prediction models. PROCEEDINGS OF THE 14TH WORLD CONGRESS ON INTELLIGENT TRANSPORT SYSTEMS (ITS), HELD BEIJING, OCTOBER 2007, 2007.

VAN HINSBERGEN, C., VAN LINT, J. & VAN ZUYLEN, H. 2009. Bayesian training and committees of state-space neural networks for online travel time prediction. *Transportation Research Record: Journal of the Transportation Research Board***,** 118-126.

VAN LINT, J., HOOGENDOORN, S. & VAN ZUYLEN, H. J. 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C: Emerging Technologies,* 13**,** 347-369.

VAN LINT, J., VAN ZUYLEN, H. J. & TU, H. 2008. Travel time unreliability on freeways: Why measures based on variance tell only half the story. *Transportation Research Part A: Policy and Practice,* 42**,** 258-277.

VANAJAKSHI, L. & RILETT, L. Support vector machine technique for the short term prediction of travel time. Intelligent Vehicles Symposium, 2007 IEEE, 2007. IEEE, 600-605.

VLAHOGIANNI, E. I., KARLAFTIS, M. G. & GOLIAS, J. C. 2006. Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume. *Transportation Research Part C: Emerging Technologies,* 14**,** 351-367.

VLAHOGIANNI, E. I., KARLAFTIS, M. G. & GOLIAS, J. C. 2014. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies,* 43**,** 3-19.

WANG, Y., MALINOVSKIY, Y., WU, Y.-J., LEE, U. K. & NEELEY, M. 2011. Error modeling and analysis for travel time data obtained from Bluetooth MAC address matching. *Department of Civil and Environmental Engineering, University of Washington*.

WELCH, G. & BISHOP, G. 2006. An introduction to the kalman filter. Department of Computer Science, University of North Carolina. Chapel Hill, NC, unpublished manuscript.

WU, C.-H., HO, J.-M. & LEE, D.-T. 2004. Travel-time prediction with support vector regression. *Intelligent Transportation Systems, IEEE Transactions on,* 5**,** 276-281.

XIA, J., CHEN, M. & HUANG, W. 2011. A multistep corridor travel-time prediction method using presence-type vehicle detector data. *Journal of Intelligent Transportation Systems,* 15**,** 104-113.

XIAO, Y., QOM, S., HADI, M. & AL-DEEK, H. 2014. Use of Data from Point Detectors and Automatic Vehicle Identification to Compare Instantaneous and Experienced Travel Times. *Transportation Research Record: Journal of the Transportation Research Board***,** 95-104.

YANG, J.-S. Travel time prediction using the GPS test vehicle and Kalman filtering techniques. American Control Conference, 2005. Proceedings of the 2005, 2005. IEEE, 2128-2133.

YGNACE, J.-L. & DRANE, C. Cellular telecommunication and transportation convergence: a case study of a research conducted in California and in France on cellular positioning techniques and transportation issues. Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE, 2001. IEEE, 16-22.

ZENG, X. & ZHANG, Y. 2013. Development of Recurrent Neural Network Considering Temporal-Spatial Input Dynamics for Freeway Travel Time Modeling. *Computer-Aided Civil and Infrastructure Engineering,* 28**,** 359-371.

ZHANG, X., ONIEVA, E., PERALLOS, A., OSABA, E. & LEE, V. C. 2014. Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction. *Transportation Research Part C: Emerging Technologies,* 43**,** 127-142.

ZHENG, F. & VAN ZUYLEN, H. 2013. Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research Part C: Emerging Technologies,* 31**,** 145-157.

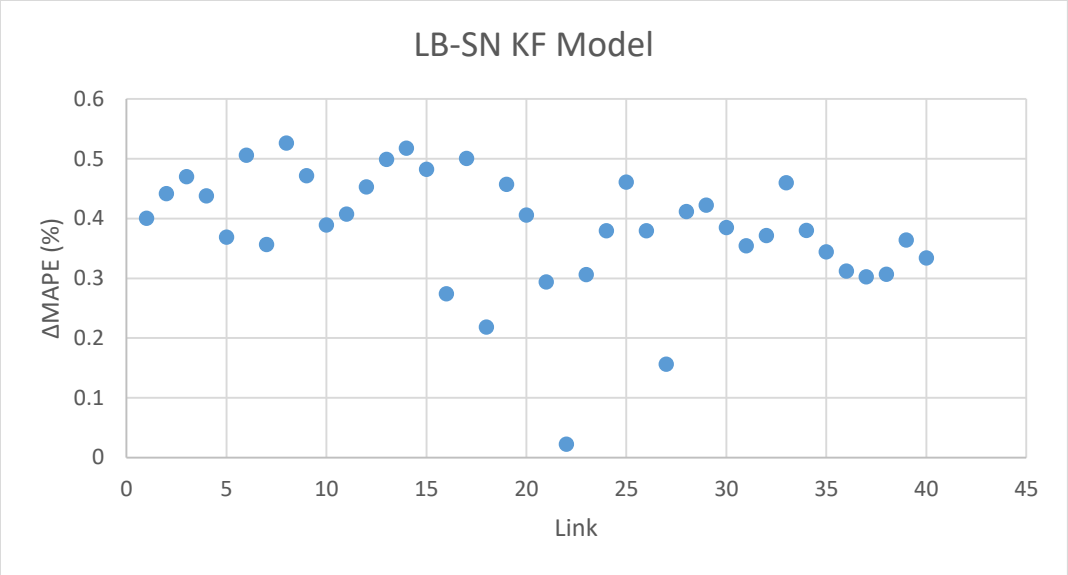
ZOU, Y., ZHU, X., ZHANG, Y. & ZENG, X. 2014. A space–time diurnal method for short-term freeway travel time prediction. *Transportation Research Part C: Emerging Technologies,* 43**,** 33-49.

APPENDIX A PERFORMANCE OF DIFFERENT PREDICTION METHODS

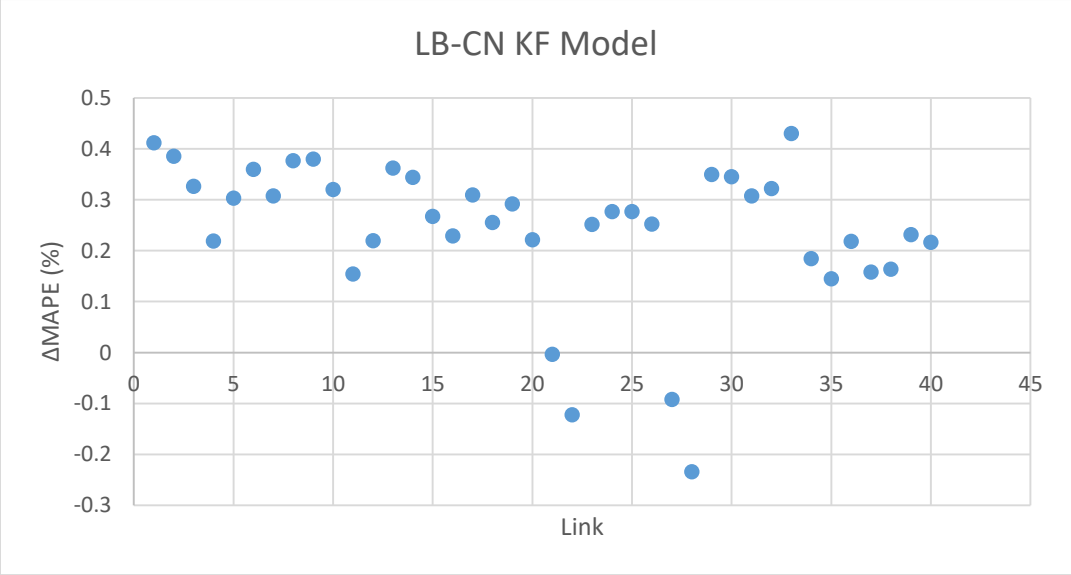
Performance in Prediction of Travel time									
			Training			Validation		
			MAE	MAPE	RMSE	MAE	MAPE	RMSE
Actual Error (or Actual Gap)			5.98	6.43%	17.52	6.29	6.70%	19.31
Corridor Based Model	Steady Noise		3.71	4.25%	13.40	4.08	4.57%	15.63
	Context Based Noise		4.04	4.23%	17.01	4.39	4.46%	19.51
	Time Varying Noise		3.51	4.20%	12.99	3.87	4.54%	14.92
Link Based Model	Steady Noise		3.78	4.27%	15.77	4.11	4.53%	18.51
	Context Based Noise		4.20	5.09%	14.29	4.55	5.47%	16.05
	Time Varying Noise		3.28	3.82%	13.42	3.96	4.57%	17.93
Weighted KNN	Value of distance parameter (n)	1	N/A	N/A	N/A	6.45	7.29%	18.31
		3	N/A	N/A	N/A	8.83	8.55%	28.70
		5	N/A	N/A	N/A	9.19	9.12%	29.32
		5	N/A	N/A	N/A	9.85	9.88%	30.42
		15	N/A	N/A	N/A	9.31	9.58%	29.59
LSBoost	Value of the parameter n (to combine previous n no of ATT)	2	4.58	5.17%	13.64	5.14	5.58%	17.55
		3	4.59	5.18%	13.61	5.15	5.60%	17.49
		5	4.63	5.26%	13.60	5.22	5.72%	15.25
		10	4.63	5.32%	13.48	5.32	5.87%	17.42

Performance in Prediction of Speed								
			Training			Validation		
			MAE	MAPE	RMSE	MAE	MAPE	RMSE
Actual Error (or Actual Gap)			2.40	4.29%	4.22	2.48	4.39%	4.31
Corridor Based Model	Steady Noise		2.00	3.46%	5.67	2.14	3.69%	6.08
	Context Based Noise		2.01	3.52%	4.56	2.13	3.69%	4.81
	Time Varying Noise		2.01	3.47%	5.94	2.16	3.70%	6.38
Link Based Model	Steady Noise		1.84	3.28%	4.88	1.95	3.43%	5.18
	Context Based Noise		1.90	3.38%	3.95	1.98	3.50%	4.09
	Time Varying Noise		1.88	3.22%	6.00	2.11	3.63%	6.44
Weighted KNN	Value of distance parameter (n)	1	N/A	N/A	N/A	3.57	6.12%	7.38
		3	N/A	N/A	N/A	4.94	8.39%	11.81
		5	N/A	N/A	N/A	5.25	8.91%	12.51

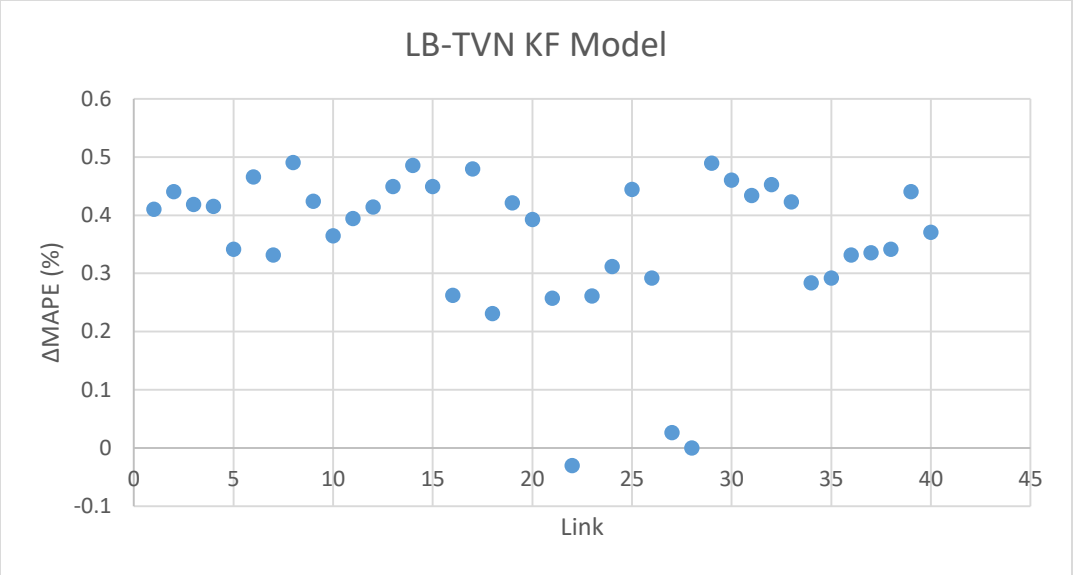
APPENDIX B PERFORMANCE OF KF MODELS



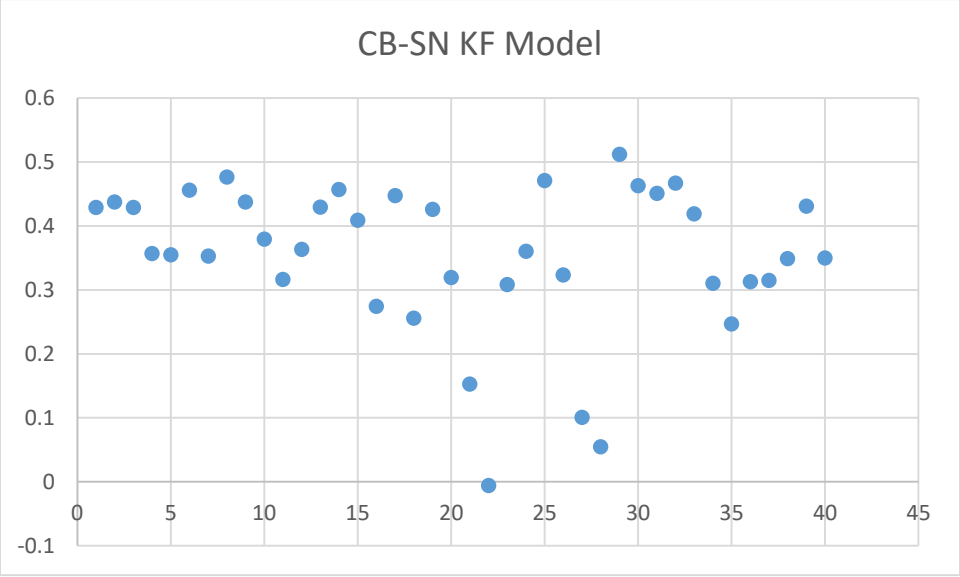
Prediction Improvement (% change in MAPE) Compared to the Actual Gap (or AMAPE) for LB-SN KF model



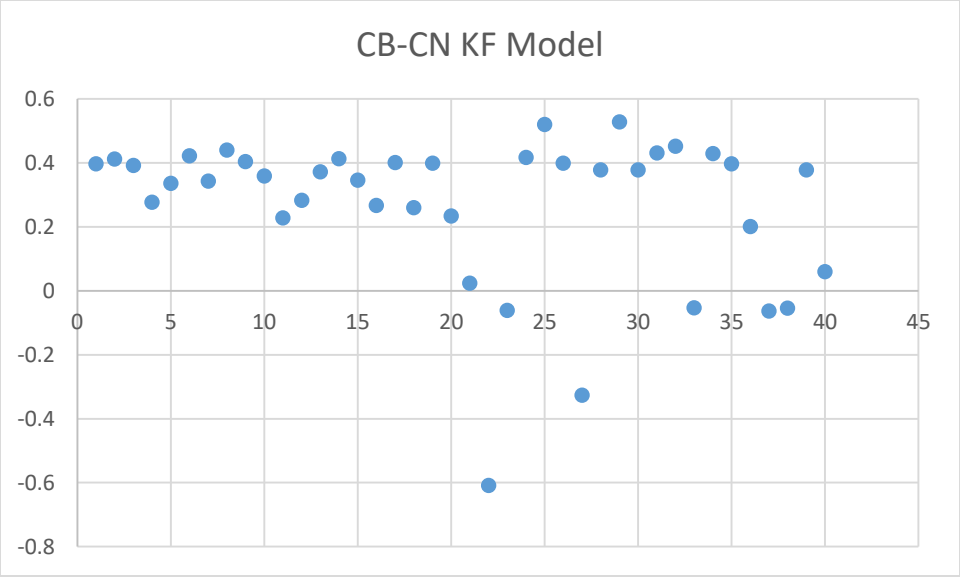
Prediction Improvement (% change in MAPE) Compared to the Actual Gap (or AMAPE) for LB-CN KF model



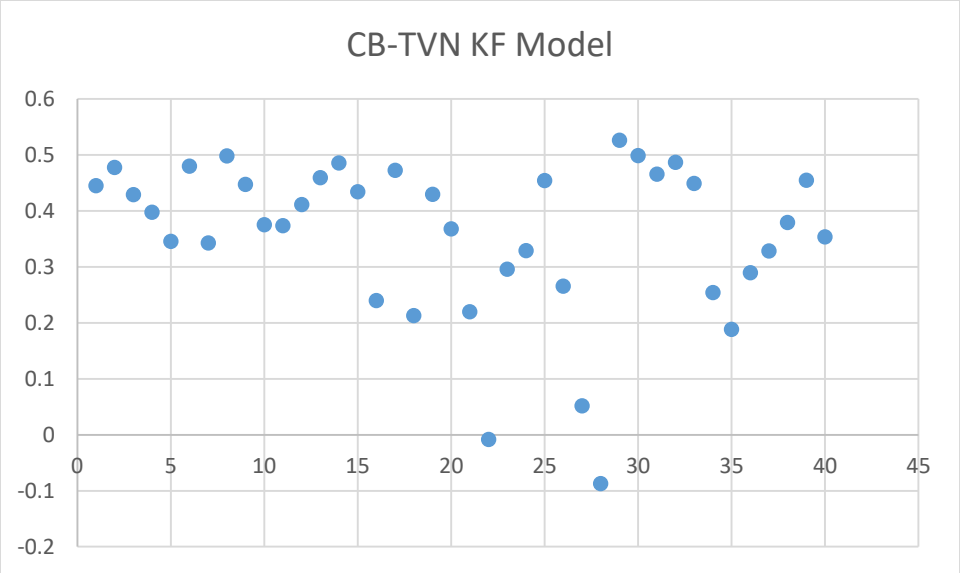
Prediction Improvement (% change in MAPE) Compared to the Actual Gap (or AMAPE) for LB-TVN KF model



Prediction Improvement (% change in MAPE) Compared to the Actual Gap (or AMAPE) for CB-SN KF model



Prediction Improvement (% change in MAPE) Compared to the Actual Gap (or AMAPE) for CB-CN KF model



Prediction Improvement (% change in MAPE) Compared to the Actual Gap (or AMAPE) for CB-TVN KF model

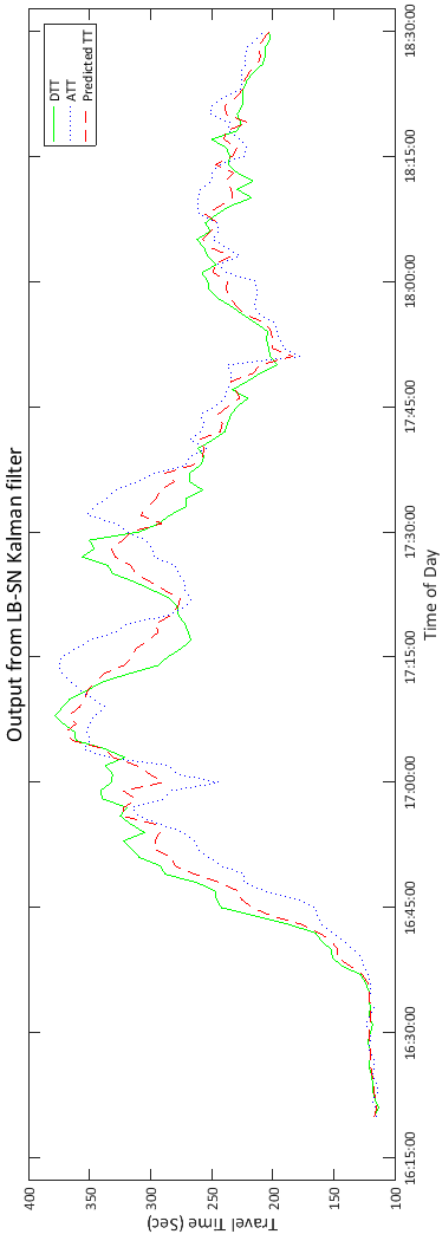
APPENDIX C Output of LSBoost

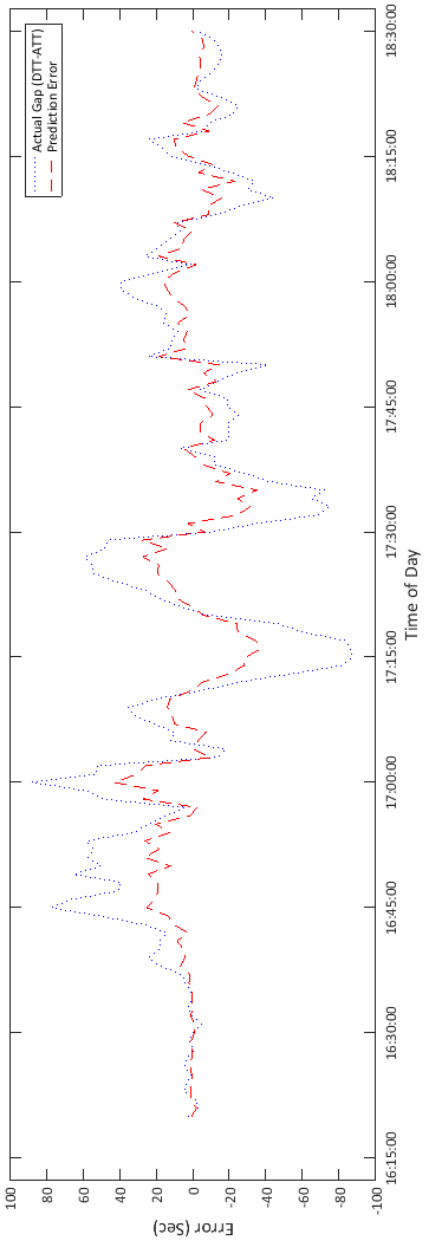
Link based output of LSBoost when trained and validated with two ATT (n=2), time of day, and day of week as four independent variables and DTT as the dependent variable:

Bluetooth Station ID		Training Prediction			Testset Prediction			Overall (trn+test) Prediction		
From	To	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
591	605	3.10	3.06%	8.59	3.31	3.25%	9.72	3.19	3.14%	9.06
605	601	4.61	4.58%	13.24	4.67	4.78%	13.89	4.63	4.66%	13.50
601	590	4.74	3.77%	11.80	4.83	3.93%	11.92	4.78	3.83%	11.85
590	597	4.53	2.96%	11.48	5.16	3.22%	13.54	4.78	3.07%	12.35
597	586	3.01	3.22%	7.15	3.49	3.54%	9.39	3.20	3.35%	8.12
586	614	3.73	3.08%	9.27	4.92	3.63%	14.67	4.20	3.30%	11.73
614	583	2.45	2.90%	6.04	3.31	3.29%	12.07	2.79	3.06%	8.97
583	635	3.63	2.99%	9.19	4.02	3.24%	9.83	3.79	3.09%	9.45
635	588	3.65	3.36%	9.12	4.04	3.62%	10.27	3.81	3.46%	9.60
588	600	3.17	3.61%	7.34	3.40	3.83%	7.74	3.26	3.70%	7.51
600	603	4.73	2.83%	12.37	5.02	3.00%	12.89	4.85	2.90%	12.58
603	592	4.86	3.15%	12.30	5.24	3.32%	13.06	5.01	3.22%	12.61
592	584	4.72	3.48%	12.14	5.02	3.64%	12.87	4.84	3.54%	12.43
584	634	4.33	3.29%	11.32	4.41	3.41%	11.27	4.36	3.34%	11.30
634	585	4.05	2.92%	10.69	4.44	3.16%	11.32	4.21	3.01%	10.95
585	628	2.57	3.01%	6.48	2.75	3.22%	6.94	2.64	3.09%	6.67
628	598	3.92	2.95%	9.83	4.84	3.37%	13.75	4.29	3.12%	11.56
598	595	2.09	3.03%	5.06	2.30	3.28%	5.67	2.18	3.13%	5.31
595	593	3.13	2.82%	7.93	3.52	3.11%	9.25	3.29	2.94%	8.49
593	589	5.10	3.16%	12.94	5.76	3.50%	14.65	5.36	3.30%	13.65
589	587	5.76	3.06%	15.60	7.12	3.55%	19.04	6.30	3.26%	17.06
587	602	16.78	13.11%	31.46	17.43	13.62%	32.98	17.04	13.31%	32.08
602	617	5.26	6.24%	14.77	5.05	6.26%	13.87	5.18	6.25%	14.41
617	1166	2.57	5.84%	8.39	2.47	5.93%	7.85	2.53	5.88%	8.17
1166	1162	1.96	5.91%	6.61	2.13	6.01%	9.86	2.01	5.94%	7.75
1162	751	2.28	6.23%	7.35	2.17	5.97%	8.63	2.25	6.16%	7.74
751	749	5.34	6.93%	19.12	13.11	13.31%	45.63	7.44	8.65%	28.79
749	644	3.19	6.69%	13.88	5.14	10.40%	22.37	3.72	7.69%	16.60
644	663	3.35	6.81%	12.34	3.53	7.13%	14.73	3.42	6.94%	13.36

663	640	5.11	7.04%	16.11	5.23	7.49%	17.27	5.16	7.22%	16.59
640	645	5.71	9.36%	18.94	5.32	9.12%	17.48	5.55	9.27%	18.37
645	651	4.54	7.30%	15.70	4.30	7.34%	14.81	4.44	7.32%	15.35
651	648	8.09	7.54%	19.64	7.61	7.47%	16.80	7.90	7.51%	18.54
648	650	2.85	6.74%	10.49	3.12	7.20%	11.90	2.96	6.93%	11.08
650	604	2.98	6.18%	11.58	2.80	6.39%	9.92	2.87	6.30%	10.64
604	750	4.88	7.26%	16.26	5.14	7.63%	16.89	5.03	7.48%	16.63
750	658	7.78	9.74%	19.95	9.08	10.57%	27.97	8.30	10.08%	23.51
658	746	9.79	10.74%	24.63	12.48	11.47%	49.95	10.87	11.03%	36.89
746	747	7.19	10.73%	20.91	9.50	11.79%	40.76	8.12	11.16%	30.49
747	642	7.19	9.98%	20.48	7.45	10.46%	19.99	7.30	10.18%	20.28
Over the network		4.58	5.17%	13.64	5.14	5.58%	17.55	4.81	5.34%	15.31

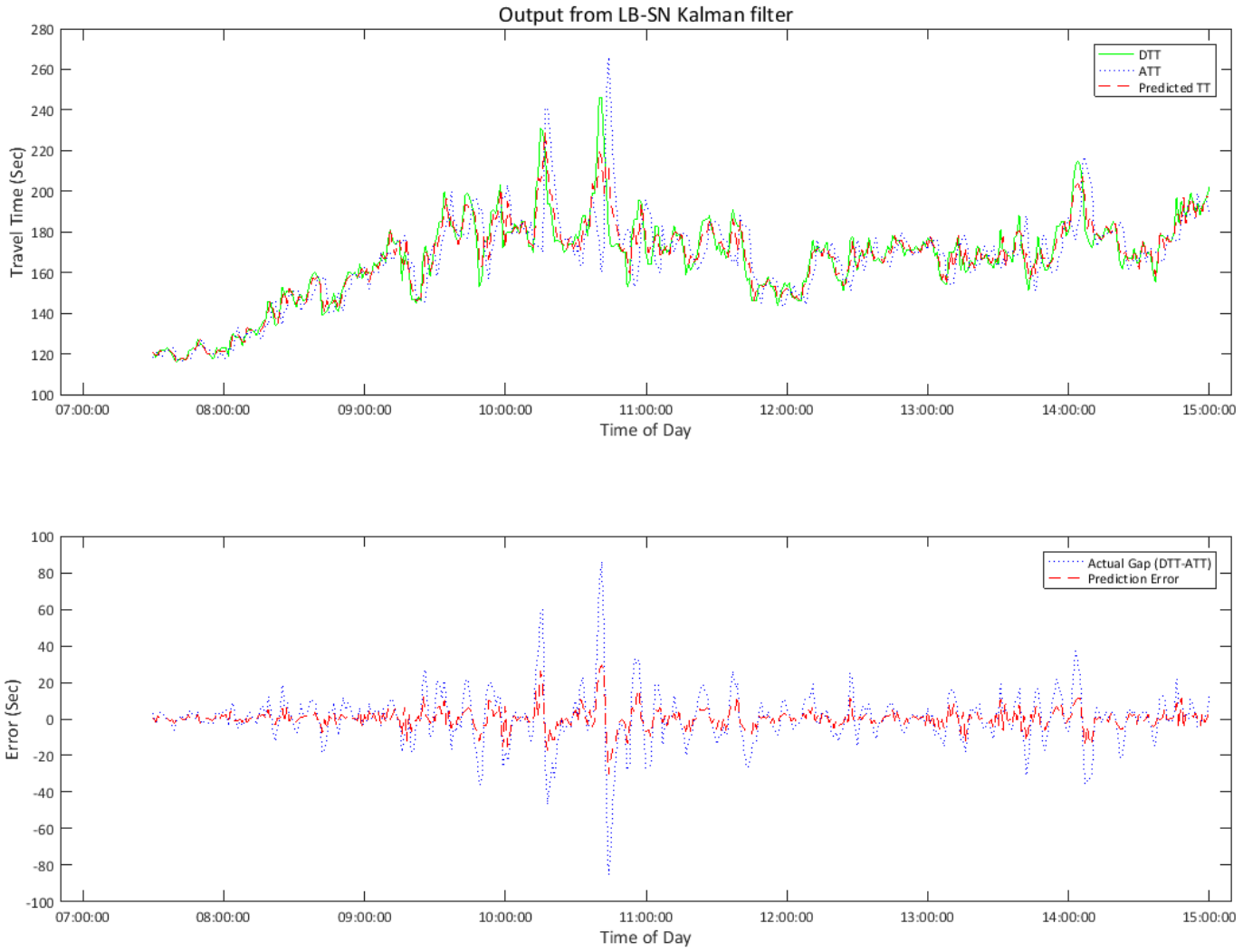
APPENDIX D PREDICTION PERFORMANCE OF LB-SN KF MODEL



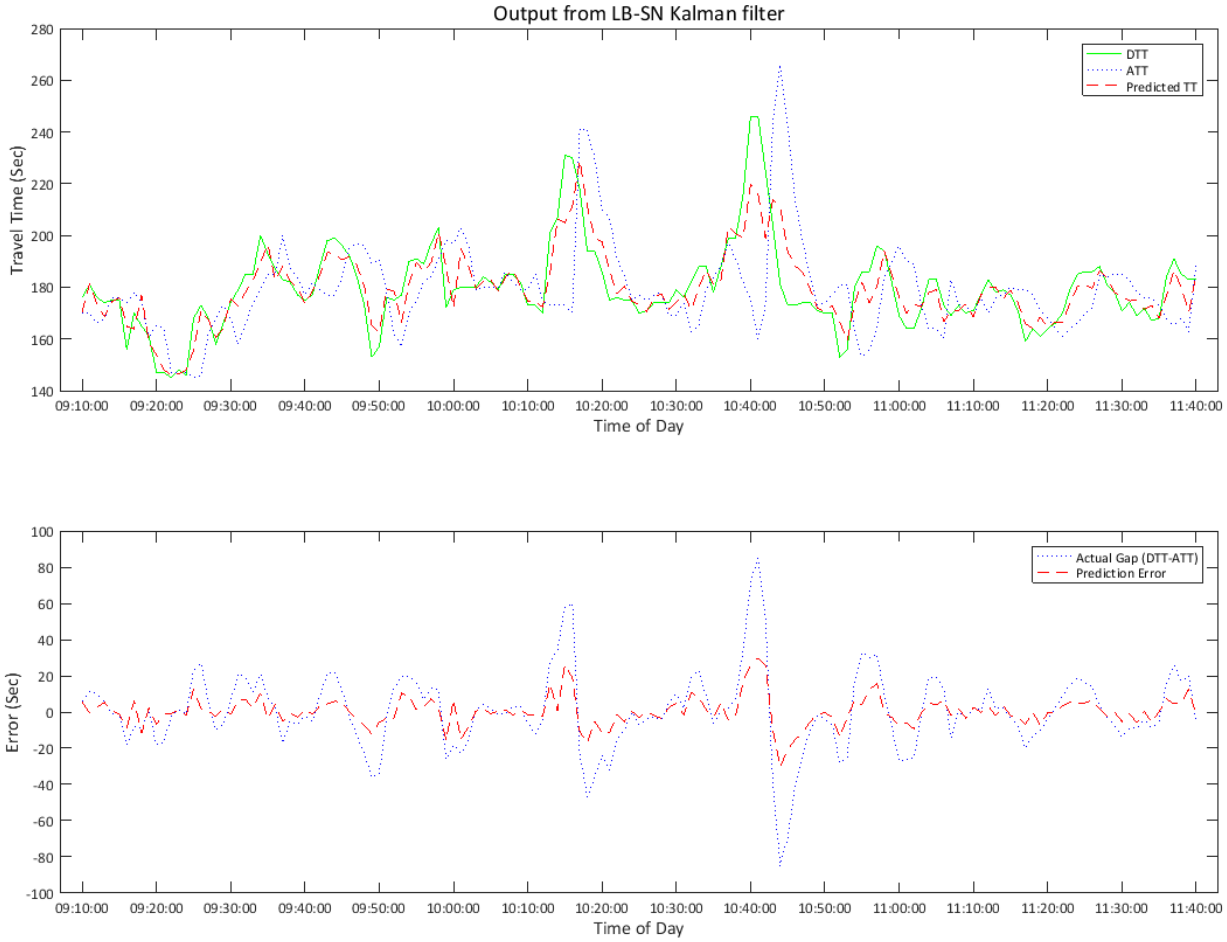


Prediction performance on I-90 corridor (8th link of study route MAC0583 to MAC0635).
Date: 18 December 2015

Note: There was an incident reported on 16:30:00

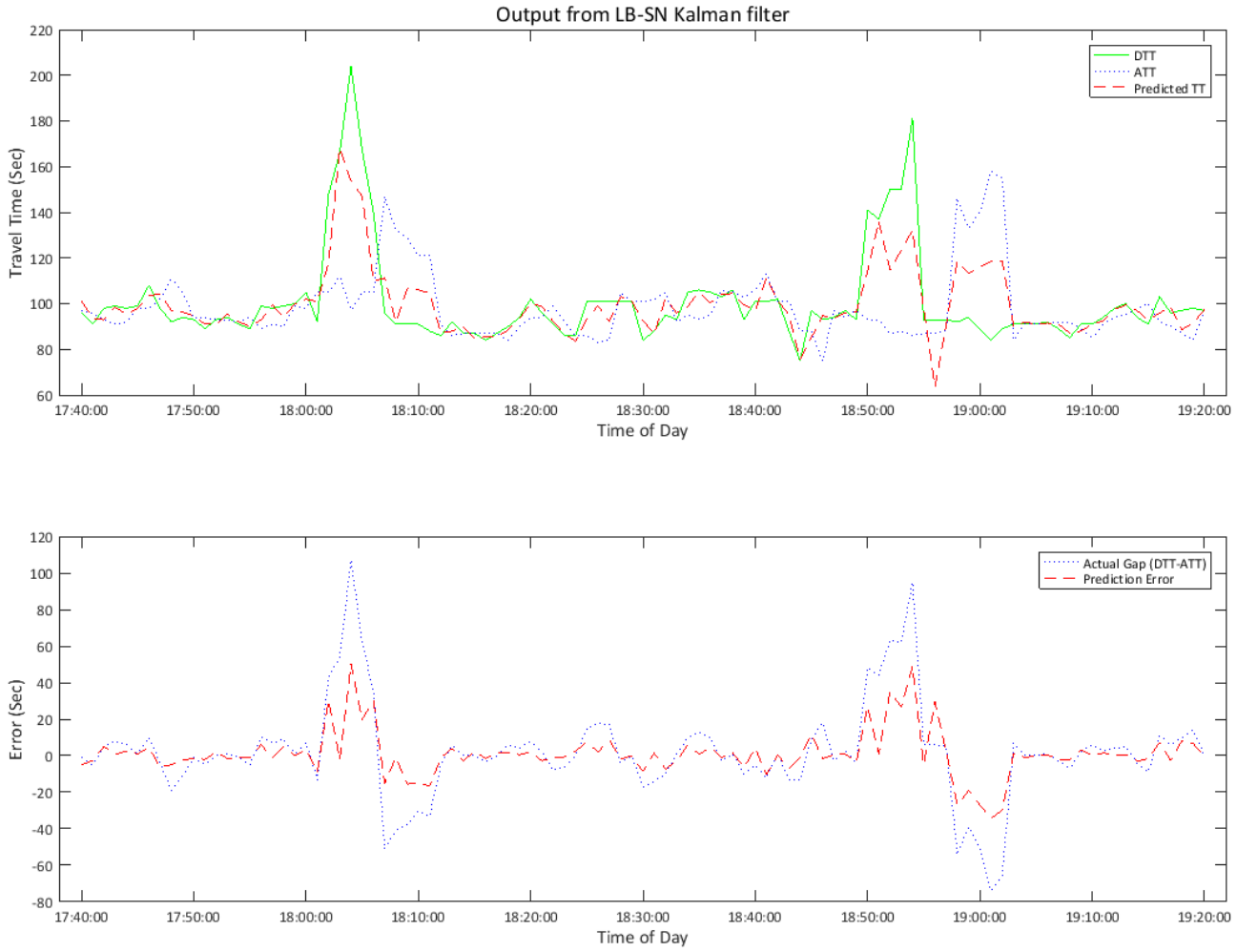


Prediction performance on I-90 corridor (8th link of study route MAC0583 to MAC0635).
Date: 28 December 2015

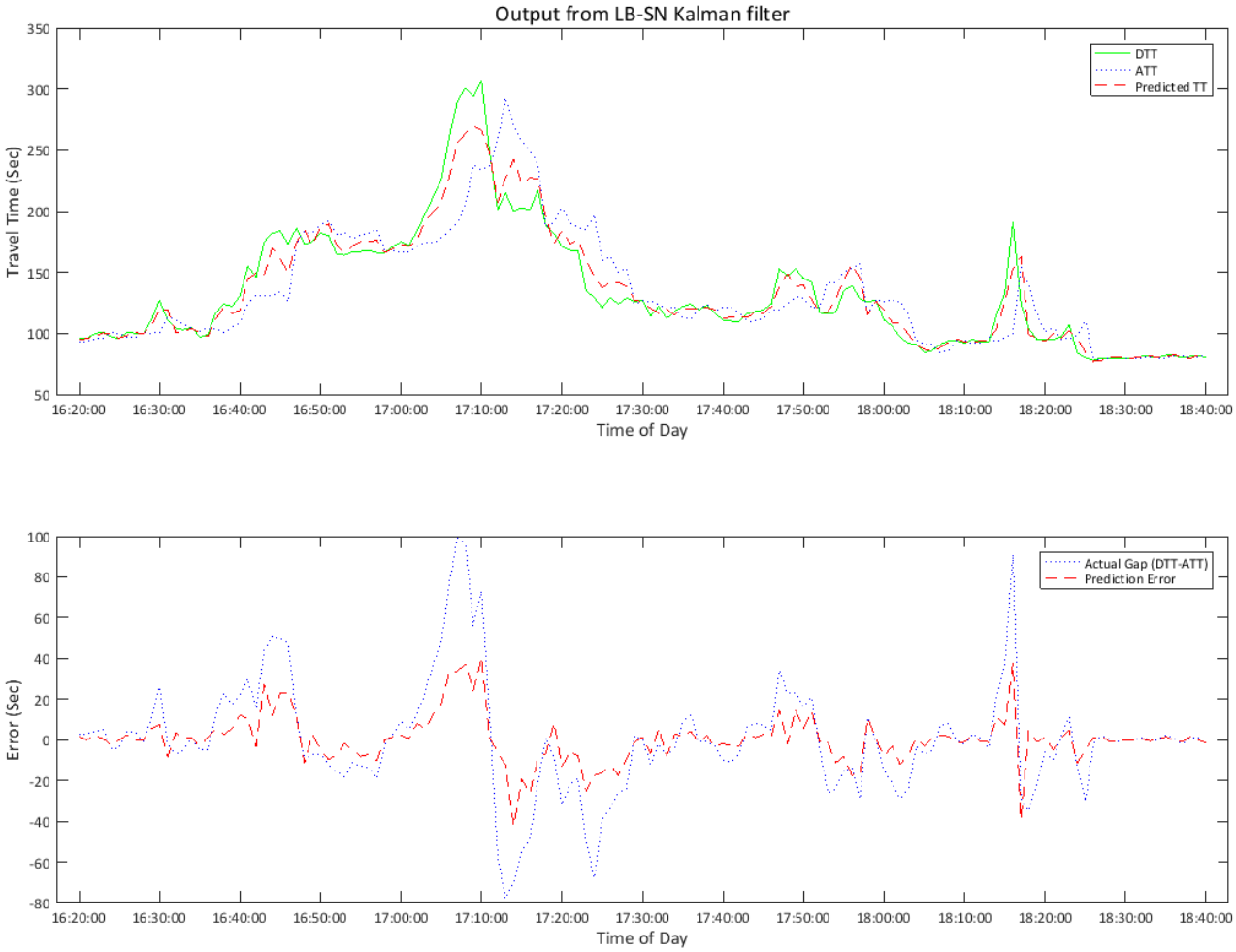


Prediction performance on I-90 corridor (8th link of study route MAC0583 to MAC0635).
Date: 18 December 2015

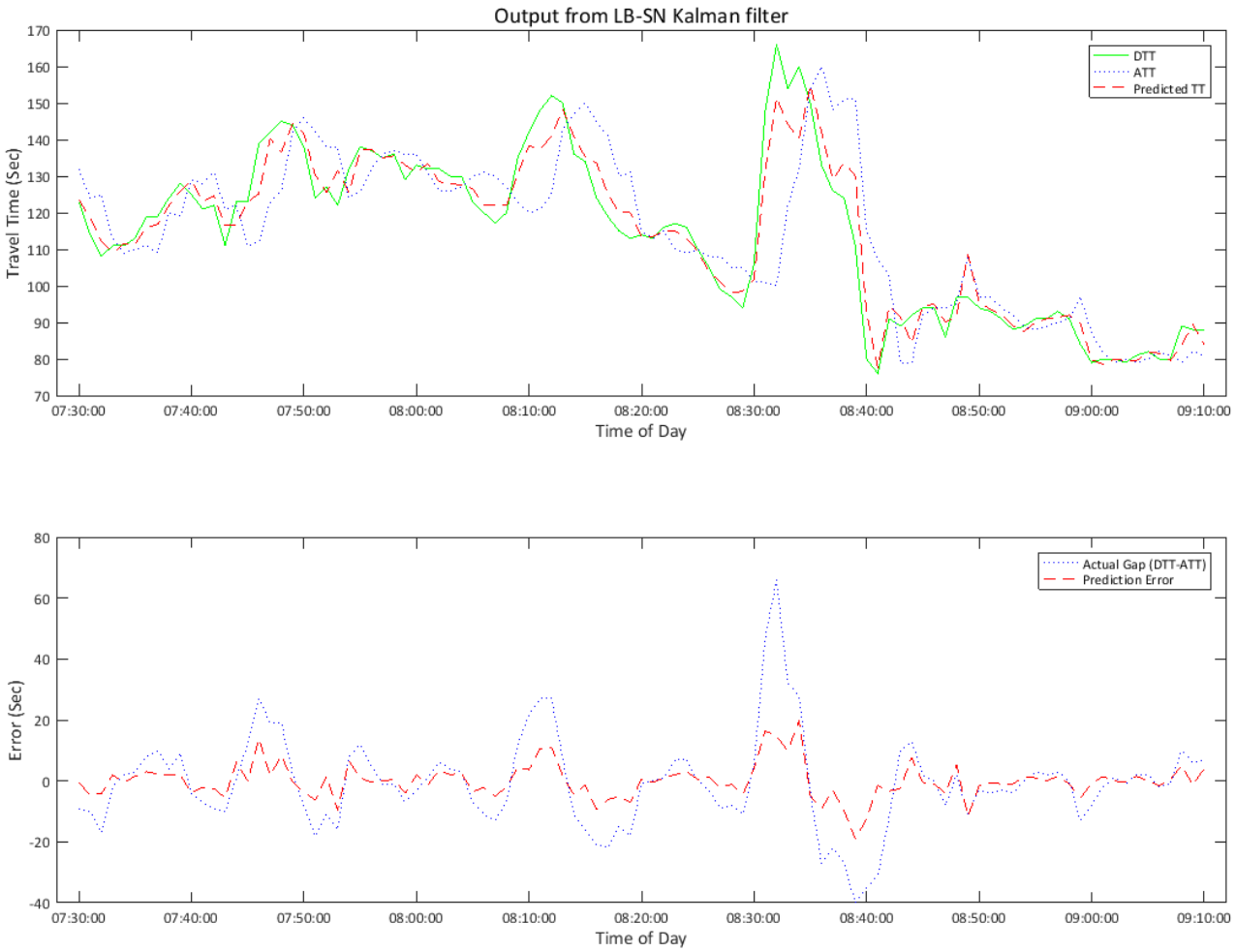
Note: Closer look of the previous picture



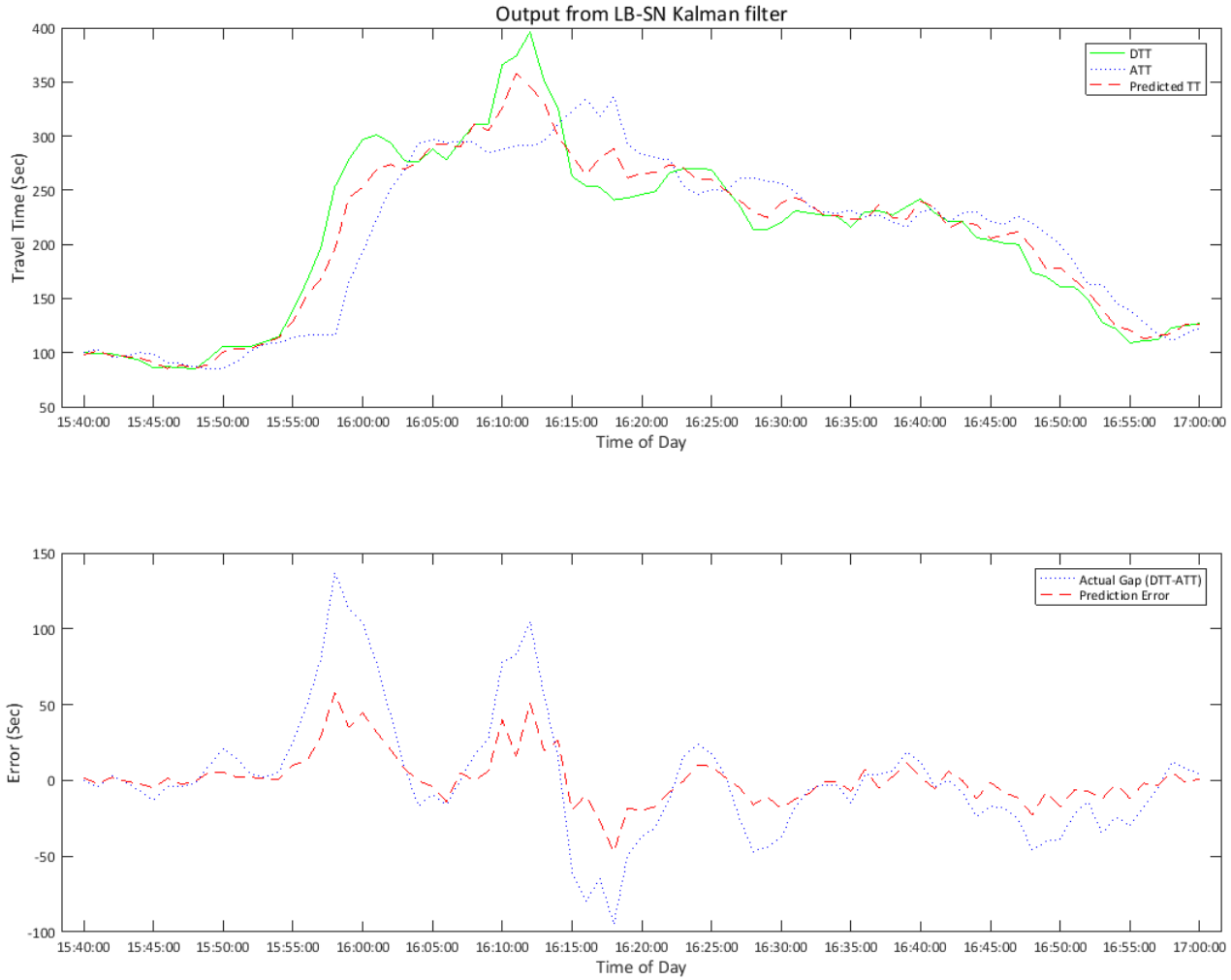
Prediction performance on Beltline Hwy (33rd link of study route MAC0651 to MAC0648).
Date: 27 November 2015



Prediction performance on Beltline Hwy (33rd link of study route MAC0651 to MAC0648).
Date: 01 December 2015



Prediction performance on Beltline Hwy (33rd link of study route MAC0651 to MAC0648).
Date: 15 December 2015



Prediction performance on Beltline Hwy (33rd link of study route MAC0651 to MAC0648)

on 18 December 2015