ARTERIAL TRAVEL TIME DISTRIBUTION ESTIMATION AND APPLICATIONS

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

YIHENG FENG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

ADVISORS: GARY A. DAVIS, JOHN HOURDOS

AUGUST 2011

Acknowledgements

I would like to express my deepest appreciation to my advisors, Prof. Gary A. Davis and Dr. John Hourdos. This thesis would not have been possible without their continuous instruction and support. Prof. Davis' truly scientist intuition inspires and enriches my domain knowledge in many areas. Dr. Hourdos' excellent guidance, caring and patience provide me with an excellent atmosphere for research.

I am greatly thankful to Dr. Henry X. Liu and Dr. David M. Levinson who teach excellent courses and provide great suggestions on my research. I also want to thank Dr. Nikolas Geroliminis who offered me the opportunity to join the transportation program at the University of Minnesota. Finally, special thanks go to Dr. William L. Cooper who served on my committee. His insightful opinions have improved this thesis.

I would also like to thank my office mates Hui Xiong and Indrajit Chatterjee for their great help during my two years study. Many thanks go to Mike Collins who worked on the same project with me. He helped me a lot in data analysis and algorithm programming.

Last but not least, my greatest gratitude goes to my family for all their love and care. For my parents who supported me in all my pursuits and encouraged me with their best wishes. And most of all for my dear wife, Xueyan Du who is always stand by and believes in me.

Abstract

Travel time estimation on signalized urban arterials has been one of the biggest challenges in transportation engineering. This thesis focuses on the characterization of arterial travel times by estimating the travel time distributions and collecting GPS data from probe vehicles to predict travel times.

The main factors that affect travel time patterns on an arterial link roughly fall into four categories: geometric structure, driving behavior, signal control and traffic demand. Four states of travel time for through-through vehicles are defined. State 1 (non-stopped) and State 3 (stopped) can be approximated using mixture normal densities while State 2 (non-stopped with delay) and State 4 (stopped with delay) can be approximated with uniform distributions. When prior travel time data is available, travel time distributions could be estimated by EM algorithm. Otherwise, they can be estimated based on signal control and geometric structure of the arterial.

Link travel times are then extended to route travel times. A method based on Markov Chain is proposed to estimate mean route travel time. Results suggest that the proposed method can capture the relationship of link travel times well and provide an accurate estimation of mean route travel time.

Combined with travel time data collected from GPS probe vehicles, a real-time traffic condition identification approach based on Bayes theorem is proposed. Numerical examples show a single GPS probe is able to identify real-time traffic condition successfully in most cases. In addition, GPS travel times can also be used to refine the existing travel time distributions using Bayesian update.

Finally, a comprehensive case study based on the NGSIM Peachtree Street Dataset is demonstrated. Travel time distributions estimated from signal timing and the geometry are considered as prior distributions. Traffic condition identification process is performed and the probability one travel time sequence belongs to each traffic condition is calculated. Data are then classified according to the posterior probabilities. Finally, a Bayesian update is run to calculate posterior distributions under each traffic condition combining with classified data. This update process can be repeated iteratively when new GPS data are available. The results obtained from Bayesian update are also compared to those estimated from EM algorithm. Overall the EM algorithm fits the data better than Bayesian update. However, sometime Bayesian approach could reflect the real world situation when some data is missing while EM does not.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

Travel time is a crucial variable both in traffic demand modeling and network performance measurement. In traffic demand modeling, travel time determines the route choice and trip distribution. From the network performance measurement point of view, travel time implies the congestion level and is the most direct traffic information to general road users. Therefore, how to estimate or predict travel time accurately has been one of the most popular topics in transportation engineering.

Travel time estimation could be roughly divided into freeway travel time estimation and arterial travel time estimation. This thesis only focuses on arterial travel time estimation. Several analytical models (Spiess, 1990; Xie *et al.*, 2001; Skabardonis and Dowling, 1997; National Research Council, 2000; Bureau of Public Roads, 1964) were proposed which relate arterial link travel time to traffic volume. Even with signal timing and volume information, these models can only provide the average travel time for all vehicles and are generally used in planning applications. In reality, different cluster of vehicles' travel time behave very differently. The average value couldn't reflect different components of travel time. Therefore, a travel time distribution from which performance measures such as the mean travel time, or the standard deviation can be derived, is preferred.

In order to estimate the arterial travel time distribution, travel time information for individual vehicles is required. One feasible way to collect travel time data is from the monitoring systems. In the United States, most urban freeways are now equipped with sensor systems that allow real time monitoring of traffic conditions. These generally consist of a combination of pavement based sensors and video detectors and can be used to identify breakdowns in the traffic flow as well as estimate travel times on the freeway system. For urban arterial systems however the ability to monitor traffic conditions and estimate travel times has lagged behind what is done on freeways, in large part due to the size of urban traffic environment. For example, up to the year 2009 the Twin Cities metro area of Minnesota had approximately 18,638 lane-miles of urban arterial (Minnesota DOT, 2009, 2010). Simply providing loop detectors at a density of 1 detector every 0.5 lane-mile (similar to the density on freeways) would require more than 37,000 detectors. Moreover, a great number of arterial links are less than 0.5 mile in length, this in turn increases the number of detectors needed for monitoring the arterial system.

These difficulties have led to an active interest in monitoring urban arterials using already-deployed sensors. One interesting possibility is to use probe vehicles. Certain types of vehicles, such as taxies or buses, which are easier to monitor have been suggested as probe vehicles (Dailey, 2002). However, using special purpose vehicles has disadvantages. Buses have to stop at each bus stop which overestimates travel times while taxis behave differently when they are empty or full. As a result, they are not able to represent the driving behavior of general road users. Recently, the increasing availability of vehicles which can access the Global Positioning System (GPS) and the development of wireless telecommunication technologies, have permitted general road users to serve as probes. In principle, a steadily increasing number of GPS-equipped vehicles could lead to reliable, accurate travel time. Information transmitted from those GPS probe vehicles not only can help calculate travel time distributions on arterials, but also provides real-time traffic condition information.

The main difficulty with reliance on GPS probes is that the travel time provided by a single probe is essentially a sample, of size one, from the prevailing distribution of travel times. This leads to questions regarding the density of probes needed to produce useful sample sizes. Unfortunately, so far the GPS vehicle density is quite low and moreover one GPS vehicle could only provide information regarding one component of travel time at each link. Therefore, rather than attempt to estimate the travel time distribution from scratch, it may be possible to combine limited information from GPS probes with prior knowledge from signal timing and arterial geometry information to reconstruct the travel time distribution.

1.2 Literature Review

Existing literature on travel time estimation focus mainly on two aspects: freeway travel time estimation and arterial travel time estimation. Due to the interrupted natures of traffic flow on urban streets, arterial travel time estimation is more challenging.

Zhang (1997) summarized the early arterial travel time estimation models. He divided them into five different categories:

Regression-type link travel time models

Regression-type models consider the travel time is linearly or nonlinearly related to other traffic parameters such as traffic demand, link capacity, and signal timing. Detector occupancy, signal delay and volume are considered to be closely related to travel time in these models (Gault, 1981; Young, 1988).

Dynamic input-output link travel time models

Dynamic input-output models estimate travel time based on time series methods and traffic data, and it can be used to estimate both link and route travel times. Basically, it is based on the traffic flow measurements at two locations (Strobel, 1977). Compared to the regression-type models, dynamic input-output models require fewer site-specific parameters and therefore can be applied to broader circumstances.

Sandglass link travel time models

Usami *et al.* (1986) first proposed this model for an oversaturated links. These models divide travel time into two parts: free flow travel time and congested travel time. Free flow travel time is calculated using a constant speed; for congested section, travel times are the deterministic queuing delays which are related to link length, traffic volume and signal timing. Many later researchers have used similar ideas to develop their own travel time estimation approaches.

As an expansion of the Sandglass type of model, Skabardonis and Geroliminis (2006) further divided the total delay caused by signalized intersections into three categories: a) the delay of a single vehicle approaching a signalized intersection without any interaction with other vehicles, b) the delay because of the queues formed at the intersection, and c) the oversaturation delay. Their paper also discussed some extreme situations such as long queues which extended over the detector and spillovers when queue from the downstream signal blocked the outflow from the upstream intersection stop line.

Link travel time estimation based on pattern matching

This approach (Bohnke and Pfannerstill, 1986) requires that each vehicle generates a unique signature on an inductive loop detector when passing through its detection zone. Then the sequence of such signatures extracted from an upstream detector is compared with successive sequences of signatures extracted from a downstream detector to find two sequences with the most matches. The time difference between the two sequences is the average travel time.

BPR – Type model

The BPR function (Bureau of Public Roads, 1964) is widely used in estimating link travel time in planning models, especially in traffic assignment. It consists of free-flow travel time and demand-dependent delay:

$$t = t_0(1 + \alpha(\frac{q}{c})^{\beta}) \tag{1.1}$$

Where, *t* is link travel time,

t_0 is free flow travel time,

q is link flow,

c is link capacity, and

α and β are two parameters (in practice use, α=0.15 and β=4)

The Highway Capacity Manual (HCM, 2000) also provided a similar approach to calculate signal delays. However, these models tend to produce unreliable estimation under oversaturation conditions (TRB, 1994).

Zhang (1999) derived the arterial travel time from journey speed. He analyzed a real-world data and concluded that critical V/C ratio, loop detector occupancy, and green band width were the main factors that affect average-link journey speeds.

Probe vehicles are also used widely in travel time estimation. Dailey and Cathey (2002) used a mass transit system as a speed sensor. This approach requires a "transit database" which contains the schedule times and geographical layout of every route and time point. In addition, all the transit vehicles used as probes should be equipped with a transmitter such as GPS receiver, and report back to control center periodically. Liu and Ma (2009) proposed a virtual probe vehicle method for arterial travel time estimation. A virtual probe is a simulated vehicle that is released from the origin to the destination at certain time point. The behavior of the virtual vehicle depends on its position and speed, the current signal status, and queue length ahead. The vehicle could accelerate, decelerate or maintain its original speed from one time step to the next. This approach is used for congested arterials, and the authors also claimed that the travel time estimation errors can be self-corrected in their model because the trajectory differences between a virtual vehicle and real one can be reduced when both vehicles meet at a red signal phase and/or a vehicle queue. Recently Hainen et al. (2011) tried to use Bluetooth equipped vehicles as probes to acquire individual travel times. The probe vehicles were required to install Bluetooth devices. When they came through an upstream/downstream detection point, the MAC address of the Bluetooth device was recorded. The travel times were then acquired by matching the MAC addresses.

Herrera et al. (2010) collected data from GPS-enabled Nokia N95 phones to estimate path travel time. They claimed that a 2-3% penetration of cell phones in the driver population was enough to provide accurate measurement. In the following research, they extracted travel time distributions from raw GPS measurements and presented the arterial network by a probabilistic model with expectation maximization (EM) algorithm for learning the parameters. (Hunter et al., 2009) One limitation of their work was that they assumed the link travel times were independent. However, in reality, there are strong correlations between links. Later, Herring et al. (2010) presented two algorithms that based on logistic regression and Spatio-Temporal Auto Regressive Moving Average (STARMA) to describe the interactions between the states of arterial links in an urban network.

Statistical modeling, such as Markov Chain and Bayesian models, has also been used. There are some applications of Markov Chain in travel time estimation on freeways and travel time reliability. Yeon et.al (2008) proposed a model based on discrete time Markov chains to estimate travel time on a freeway. They simply divided the system states into 0 (uncongested) and 1 (congested). The travel time of one route was the summation of each link's travel time at different states. They believed the transition matrix would converge to a steady-state no matter what the initial state was. Consequently, a uniform transition matrix was used in their study. Dong and Mahmassani (2009) studied a similar case from a different perspective. They focused on the probabilities of flow breakdown on freeways at peak hour. Lin et.al (2003) proposed a method to estimate travel time on arterials, suing a Markov chain model. Similar to previous work, they reduced the continuous delay experienced by drivers at each intersection into two distinctive states: zero-delay and a state of nominal delay. The parameters of the transition matrix of their Markov process was the flow condition at the intersection, the proportion of net inflows into the arterial from the cross streets, and the signal coordination level. A cell transmission model (CTM) was used in the paper to validate the model.

Jintanakul et al. (2009) proposed using Bayesian mixture model to estimate freeway travel time. They considered that travel time came from either a faster component or a slower component and tried to use a mixture normal distribution to approximate the travel time probability density. They established a hierarchical Bayesian mixture model to estimate the parameters. They assumed the variances of the normal distributions to be constants to simplify the problem. However, in most real-world cases, variances change under different traffic conditions. This model does not apply to arterial travel time estimation.

1.3 Context of Study

The objective of this thesis is to develop and test methods utilizing data collected from GPS probe vehicles to characterize arterial travel time. There are two major research interests.

- Trip travel time estimation

Given the origin and destination of a trip, how to estimate the total travel time is very attractive to both transportation engineers and general road users. The trip travel time is related to many factors. For example, different routes result in different travel times. For a specific route, travel times under different traffic conditions are different as well. Note that one route usually consists of multiply arterial links. In order to estimate the route travel time, we need to find the patterns of travel times of each arterial link under different traffic conditions first. Furthermore, when signal coordination is in effect, the travel times between links are not independent.

- Real-time traffic condition identification

Traffic conditions change throughout a day because of the different signal controls and traffic volume. In order to estimate the real-time travel time, we need to know the operating traffic condition. The foundation of distinguishing different traffic conditions is again to characterize the travel times of individual arterial links from historical data. In addition, travel times collected from GPS probe vehicles are also needed to provide real-time information. Finally, the operating traffic condition could be identified by combining the historical travel time patterns and the real-time GPS travel time.

To find the answers to these problems, we need to characterize the travel times of single arterial links and collect real-time GPS data. One way to describe arterial travel time precisely and accurately is to use probability distributions. Travel time distributions could provide many measurements including mean travel time, and standard deviation. The travel time distributions should be estimated as a function of the geometry, the signal settings and the level of congestion to represent different traffic conditions. The real-time GPS data which reflect the current traffic condition are then used to refine the corresponding travel time distributions on a learning procedure. As more GPS data become available, the travel time distributions will provide updated and more accurate measurements. Note that GPS data are collected under different traffic conditions. Before utilizing these data to update the travel time distributions, they need to be assigned to the right cluster.

1.4 Thesis Organization

This thesis focuses on the characterization of arterial travel times by estimating the travel time distribution and GPS data from probe vehicles to predict travel times. Two major themes are addressed by this thesis:

1). Travel time characterization and travel time distribution estimation. Empirical travel time distributions tend to be multi-modal, and can be modeled using mixture distributions. Two approaches are proposed to estimate link travel time distribution with or without prior data.

2). After constructing a set of travel time distributions, GPS probe data can be used to identify the correct distribution, and estimate the mean route travel time, and real-time traffic condition. The parameters of travel time distribution can also be updated using a Bayesian model.

The rest of the thesis is organized as follows: Chapter 2 first introduces the data used throughout the thesis and then describes the travel time patterns on signalized arterial links. Chapter 3 introduces two approaches to estimate link travel time distributions. Chapter 4 extends individual links to routes which are defined as sequences of consecutive links, and proposes a model for estimating mean route travel time. In Chapter 5, two applications which combine travel time distributions with GPS data are illustrated. Chapter 6 provides a comprehensive case study using the NGSIM Peachtree Street dataset, while Chapter 7 concludes this thesis and lays out the directions for further research.

Chapter 2 Characterization of Travel Time Patterns on Signalized Arterial Links

2.1 Data Introduction

The NGSIM Peachtree Street dataset is used intensively throughout this work. The NGSIM program, led by Federal Highway Administration, is aimed at developing a core of useful and open behavioral algorithms which improve the quality and performance of microscopic simulation tools (NGSIM Community, 2011). As part of this program, detailed traffic data were collected on a segment of Peachtree Street, in Atlanta Georgia, on November 8th, 2006. (Cambridge Systematics, 2007) This arterial section is approximately 2,100 feet in length, with five intersections and two to three through lanes in each direction. The section is divided into six links, numbered from one to six running from south to north, divided by the neighboring intersections. Links 1 and 6 turned out to be too short for useful analysis, so our study area is only limited to links 2 through 5. Four of the five intersections are signalized, with intersection 4 being un-signalized. Figure 2.1 shows the geometric structure of the study area. The Peachtree Street data consisted of two 15-minute time periods, 12:45 p.m. to 1:00 p.m. and 4:00 p.m. to 4:15 p.m. These two periods were taken to represent two different traffic conditions, which we call Noon and PM. Note that Noon and PM are only names of traffic conditions, and the traffic conditions are not discriminated by time.

The data consisted of detailed individual vehicle trajectories with time and location stamps, from which the link travel times of individual vehicles could be calculated. In this study, link travel time refers to the time point a vehicle enters the arterial link to the time point this vehicle passes the stop-bar at the end of the link. Intersection travel time is excluded.

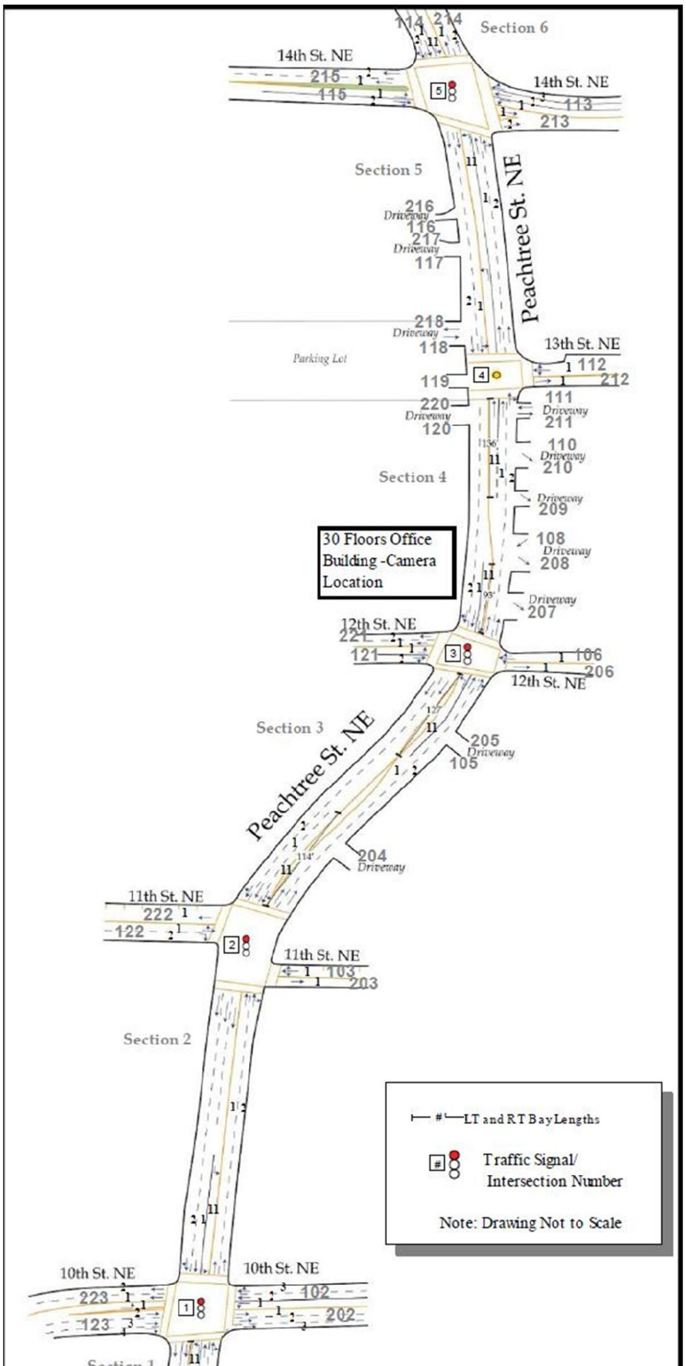


Figure 2.1 Study Area Schematic

Source: NGSIM Peachtree Street (Atlanta) Data Analysis Report (4:00 p.m. to 4:15 p.m.)

2.2 Characterization of travel time

Travel time patterns on signalized arterial links can vary in complexity under different situations. The main factors that affect travel time patterns on an arterial roughly fall into the following four categories:

- Geometric structure of the arterial including link length, number of lanes, speed limit, roadway alignment, and driveways.
- Driving behaviors including lane changing rate, car following behavior and driver aggressiveness.
- Signal control strategy including cycle length, phase splits, effective green time and offsets.
- Traffic demand such as traffic volume of the main arterial, the turning movement flows at the intersections, and the volume ratio between main arterial and the side street.

For a given location, over the course of the day it is reasonable to think that the geometric structure of the arterial is unchanged and the driving behavior remains similar, leaving the traffic demand and signal control as the determinants of travel time distributions. That is, under one specific traffic condition (combination of signal control strategy and given traffic demand level), it is believed that the travel time distributions are similar. As a result, a set of different travel time distributions are needed for different combinations of demand and signal timing.

Figure 2.2 (a) shows a typical travel time histogram for a signalized arterial link from the NGSIM Peachtree Street dataset (Link 2 Northbound at Noon). This histogram shows the travel time of vehicles which go through to the downstream link. Different colors in the figure represent the different origins of the vehicles. For example, green vehicles come from the through direction of the upstream intersection while brown and grey vehicles come from the side streets of the upstream intersection. Refer to Appendix A to see the list of colors and corresponding origins. It can be seen that travel times of green vehicles cluster around 10 seconds and 70 seconds, and that travel times of grey and brown vehicles tend to fall between those of green vehicles. Apparently, travel times from different origins are different due to different encounters with the different phases of signal control.

Restricting attention to through-through vehicles, defined as vehicles traveling through from the upstream boundary to the downstream, as shown in Figure 2.2 (b), gives a clear two-peak pattern as well as several additional travel time points located between the two peaks or beyond the second peak. The two peaks are the major components of travel time, representing non-stopped vehicles and stopped vehicles. Non-stopped vehicles refer to vehicles which pass the end of the link without stopping, while stopped vehicles refer to vehicles stopping at the end of the link for a red phase. The scattered points between the two peaks or beyond the second can be recognized as two additional components: non-stopped with delay and stopped with delay, where the additional delay is not caused by the signal control, but by some “incidents”. For example, some through vehicles can be delayed by a preceding vehicle making a permitted right turn or left turn. Other reasons such as slow moving vehicles, violating the traffic rule, or vehicles entering or exiting from driveways can also cause such delays. All link travel time diagrams for two traffic conditions, Noon and PM, of the NGSIM Peachtree dataset are listed in Appendix A.

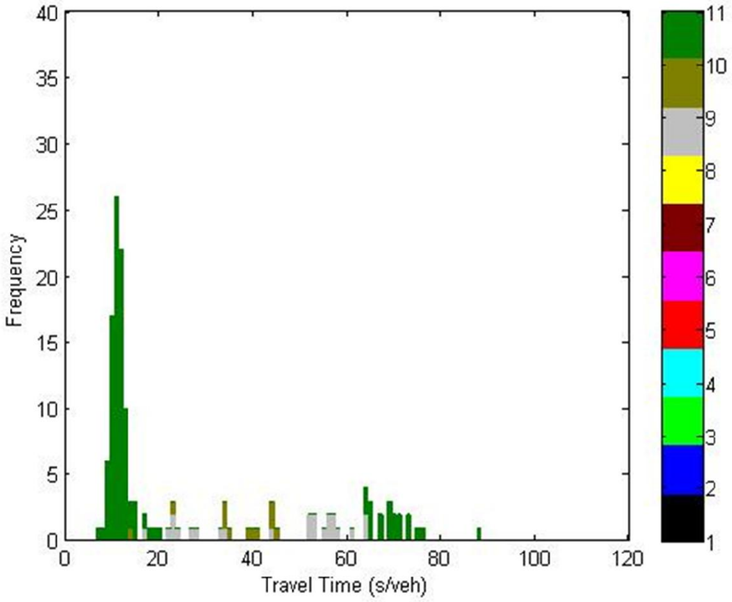
In this thesis, we will restrict attention to through-through vehicles and use the phrase travel time state to represent different travel time components. In short, four states of travel time are defined:

State 1: non-stopped,

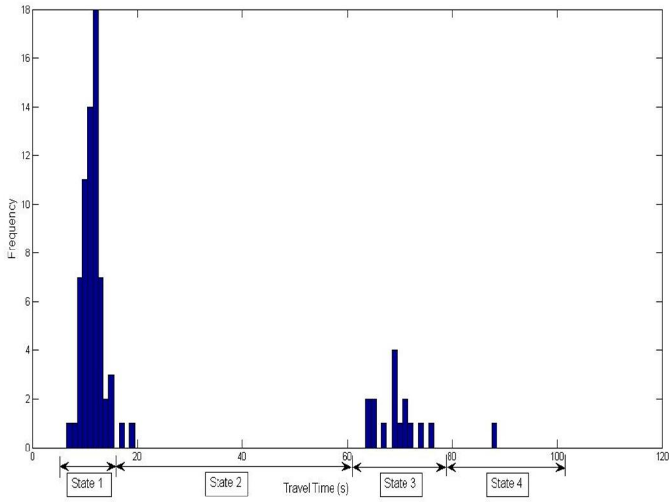
State 2: non-stopped with delay,

State 3: stopped,

State 4: stopped with delay.



(a)



(b)

Figure 2.2 Travel Time Histograms (a) All – Through (b) Through – Through

Vehicles which make turns at the end of the intersection or vehicles from minor streets may have different states. The travel time patterns involving turning vehicles are therefore much more complicated than through – through vehicles. For right turning vehicles, usually right turn can be made when the signal is red and sometimes there is no separate right turn bay. For left turning vehicles, similar geometry issues may apply and sometimes the left turn signal is divided into protected and permitted intervals. Both left turning and right turning vehicles are required to yield to through traffic. Those complexities make it much more difficult to analyze the travel time patterns of turning vehicles. Moreover, through-through vehicles are usually the main component of the total traffic. Therefore, for simplicity, in this thesis, only through-through vehicles are taken into consideration.

Chapter 3 Link Travel Time Distribution Estimation

3.1 Estimation with EM algorithm

Chapter 2 stated that the travel time histogram for through-through vehicle follows a two peak pattern. Therefore, a bimodal distribution could be used to fit the travel time histogram. Empirical characterization of travel time distributions on signalized arterials shows that this bimodal distribution can be approximated using a mixture of normal densities (Davis and Xiong, 2007; Xiong and Davis, 2009). Maximum likelihood estimates of mixture model parameters can be accomplished using the estimation-maximization (EM) algorithm (McLachlan & Krishnan, 1997). The EM algorithm is briefly introduced here.

Let $f(TT)$ be the probability density function (*pdf*) of a mixture normal distribution:

$$\mathrm{f(TT) = p \times f_n(TT) + (1\text{-}p) \times f_s(TT)} \qquad (3.1)$$

Where, p is the portion of non-stopped vehicles,

$f_n(TT)$ is *pdf* of non-stopped vehicles (first peak), $f_n(TT) \sim N(\mu_1, \sigma_1^2)$ which follows a normal distribution with mean μ_1 and variance σ_1^2, and

$f_s(TT)$ is *pdf* of stopped vehicles (second peak), $f_s(TT) \sim N(\mu_2, \sigma_2^2)$ which follows a normal distribution with mean μ_2 and variance σ_2^2.

Given the travel time of a specific vehicle, it is unknown whether it belongs to non-stopped vehicles or stopped vehicles. The travel times are then grouped into m intervals, where $m = round(TT_{max} - TT_{min}) + 1$. Consequently, the width of each interval is 1 second. Assume $n_1 \ldots n_m$ are the number of travel times that falls into intervals $[a_0, a_1], \ldots, [a_{m-1}, a_m]$.

Let ψ denotes the parameter set $(\mathrm{p}, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. Then the probability that an individual vehicle travel time falls in the j^{th} interval is given by

$$P_j(\psi) = \int_{a_{j-1}}^{a_j} f(TT|\psi)\, dTT, j = 1, \ldots, m \qquad (3.2)$$

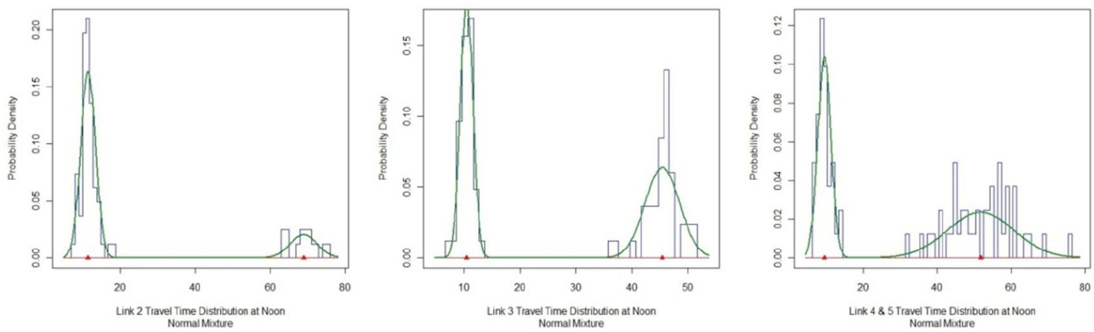
The grouped data then follow a multinomial distribution with likelihood function

$$\mathrm{L}(\psi) = \frac{\mathrm{n}!}{\mathrm{n}_1!\ldots\mathrm{n}_\mathrm{m}!}\{\mathrm{P}_1(\psi)\}^{\mathrm{n}_1} \ldots \{\mathrm{P}_\mathrm{m}(\psi)\}^{\mathrm{n}_\mathrm{m}} \qquad (3.3)$$

and the log-likelihood function is

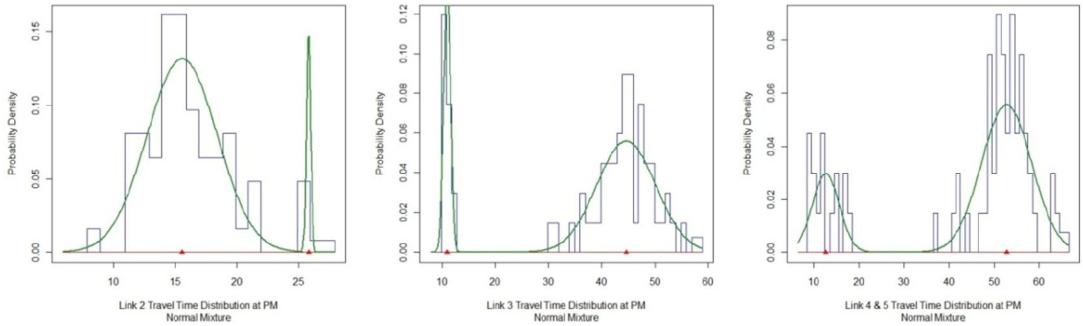
$$\log \mathrm{L}(\psi) = \sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{n}_\mathrm{j} \log(\mathrm{P}_\mathrm{j}(\psi)) + \log(\frac{\mathrm{n}!}{\mathrm{n}_1!\ldots\mathrm{n}_\mathrm{m}!}) \qquad (3.4)$$

Maximum likelihood estimation can now be used to compute the estimates. This method was accomplished using the R software package mixdist (Du, 2002).This routine uses the standard maximum likelihood estimation method and combines the estimation maximization (EM) algorithm with a Newton-type method (Venables, 2005). Figure 3.1 and Figure 3.2 show the estimated travel time distributions (normal mixture) of state 1 and state 3 of links 2 to 5 under two traffic conditions Noon and PM of the NGSIM Peachtree Street dataset. As mentioned above, intersection 4 is an un-signalized intersection, and left turns at intersection 4 were forbidden, so link 4 and link 5 were treated as one single link, giving us three consecutive links. Note that all travel time distributions in Figure 3.1 and Figure 3.2 show that the traffic conditions are not oversaturated. Under oversaturated traffic conditions, the distribution may have a third peak representing vehicles in the residual queue (stopped twice).



(a) (b) (c)

Figure 3.1 Normal Mixture Approximation of Travel Time State 1 & 3 at Noon



(a) (b) (c)

Figure 3.2 Normal Mixture Approximation of Travel Time State 1 and 3 at PM

The mixture distributions in Figure 3.1 and Figure 3.2 only represent stopped and non-stopped components, which are vehicles in state 1 and state 3. Since the number of observations in state 2 and state 4 is very small, it is difficult to identify a unimodal distribution for these. So it is assumed that the probabilities of delay times caused by different 'incidents' in state 2 and state 4 are equally likely, which indicates travel times in state 2 or state 4 follow uniform distributions. The upper boundary and lower boundary of these two uniform distributions can be defined as follows. Suppose the non-stopped component is normally distributed with mean and variance μ_1 and σ_1^2 while the stopped component is normally distributed with mean and variance μ_2 and σ_2^2. If an observed travel time is located beyond $\mu_1+3\sigma_1$ which is approximately the 99th percentile for the state 1, then it is reasonable to think that this travel time does not belong to state 1, and $\mu_1+3\sigma_1$ is considered as the lower boundary of the uniform distribution for state 2 (non-stopped with delay). Based on this rule, the upper boundary for non-stopped with delay, and the lower boundary for stopped with delay can be defined. The upper boundary for state 4 (stopped with delay) is assumed not to exceed a certain value TT_{max} which is the longest possible travel time in one arterial link.

3.2 Prior Estimation based on signal control and geometry

The previous section described how to estimate a travel time distribution based on individual vehicle travel times. However, under many circumstances, travel times of individual vehicles are not readily available. As an alternative, this section proposes an approach to compute prior of travel time distributions for state 1 and state 3 based on the signal control and geometric structure of the arterial link.

Some assumptions are made to simplify the problem. First, the signals of upstream and downstream intersections are coordinated, which means that the signals should have the same cycle length. Second, queues can be fully discharged during one cycle so that non-oversaturated traffic condition applies. Third, the turning rate from side streets is negligible. This assumption assures that vehicles from side streets should not affect the travel time on the main arterial too much.

Given the assumption that the travel time distribution for state 1 and state 3 can be approximated using mixture of normal densities as shown in Equation 3.1, five parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p)$ need to be estimated.

- Estimation of μ_1

μ_1 is the mean travel time of non-stopped vehicles which can be considered as free flow travel time. It could be estimated by section length and speed limit:

$$\mu_1 = L/V_L \tag{3.5}$$

Where, L is the section length, and

V_L is the speed limit of the arterial link

- Estimation of σ_1

σ_1 is the standard deviation of the free flow travel time. Unfortunately, there is no direct way to estimate the standard deviation of travel time without data. However, since travel time is the reciprocal of speed given a section length, σ_1 could be converted from the standard deviation of speed. Empirical studies suggest that standard deviation of speed varies a lot under different demands. For example, it shows a U-shape

(Nezamuddin *et al.* ,2009) relation with respect to traffic volume. That is, when either the traffic demand is low or high, the standard deviation increases. When the traffic demand is within intermediate ranges, the standard deviation remains low and stable. Existing literatures (Box and Oppenlander, 1976; Oppenlander, 1963) suggest a standard deviation of 5-7 mph for speed distribution under intermediate range of demand.

- Estimation of μ_2 σ_2

μ_2 and σ_2 are mean travel time and standard deviation of stopped vehicles, which are closely related to the signal settings of the upstream and downstream intersections.

By assumption, the travel time of stopped vehicles also follows a normal distribution. According to the properties of normal distribution, if one wants to estimate the mean and standard deviation, instead of estimating directly, they could estimate the lower and upper boundary of 99th percentile first and calculate the mean and standard deviation accordingly. Figure 3.3 shows the probability density function (*pdf*) of a normal distribution. Suppose the lower and upper boundary of the 99th percentile are a and b respectively. Then by the symmetry property, the mean could be calculated as $\mu = (a+b)/2$. Moreover, the distance from the mean to the 99th percentile is 3σ. Therefore, standard deviation could be calculated as $\sigma = (b-\mu)/3$ or $\sigma = (\mu-a)/3$.

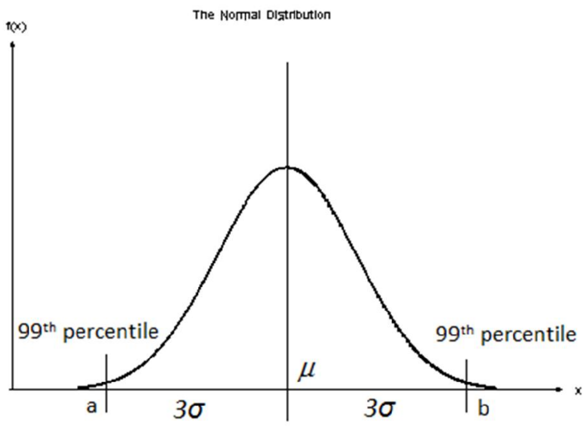


Figure 3.3 Normal Distribution *pdf*

Now the key question is how to estimate the upper and lower boundary of the 99th percentile of link travel time. It can be estimated by adding the upper and lower boundary of signal delay to free flow travel time.

To estimate upper and lower boundary of signal delay, we need to know the signal control plan such as cycle length, green and red split and the offset between upstream and downstream intersections. An example time-space diagram of two consecutive intersections is shown in Figure 3.4. The arrow on the right is the movement direction.

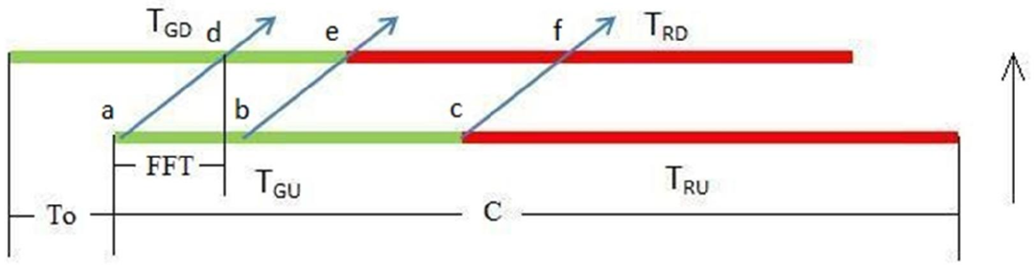


Figure 3.4 Time-space Diagram for Two Consecutive Intersections

Where, $\mathbf{T_{GD}}$ and $\mathbf{T_{RD}}$ are the green/red time of downstream intersection

$\mathbf{T_{GU}}$ and $\mathbf{T_{RU}}$ are the green/red time of upstream intersection

C is the cycle length

To is the offset between two signals

FFT is the non-stopped vehicle mean travel time (equals to μ_l)

By looking at the time-space diagram, one can easily see that vehicles pass during T_{GD} are non-stopped vehicles while vehicles reach the downstream intersection during T_{GD} are stopped vehicles. The estimation of the upper and lower boundaries of signal delay can be categorized into the follow three cases:

Case I $\mathbf{To + \mu_l < T_{GD} < To + T_{GU} + \mu_l}$

$D_{upper} = T_{RD}$ $\qquad\qquad$ $D_{lower} = C - (To + T_{GU} + \mu_l)$

Case II $T_{GD} < To + \mu_I$

$D_{upper} = C - (To + \mu_I)$

if $To + T_{GU} + \mu_I < C$ -> $D_{lower} = C - (To + T_{GU} + \mu_I)$

if $To + T_{GU} + \mu_I > C$ -> $D_{lower} = 0$

Case III $T_{GD} > To + T_{GU} + \mu_I$

$D_{upper} = 0$ $D_{lower} = 0$

Where, D_{upper} and D_{lower} are upper and lower boundaries of signal delay

In case I, where $\mathbf{To} + \mu_I < \mathbf{T_{GD}} < \mathbf{To} + \mathbf{T_{GU}} + \mu_I$, the green red split at the downstream intersection is between the green time start of the upstream intersection plus the free flow travel time, and the green time end of upstream intersection plus free flow travel time. That is, point e is between point a and point c in Figure 3.4. Suppose vehicle 1 passes upstream intersection at point b and arrives downstream intersection at the beginning of the red signal (point e). Since there is no residual queue from the previous cycle, vehicle 1 must be the first in the queue and thus would not experience extra queue discharging time. Then the longest signal delay time is considered as the duration of red signal at downstream intersection. Suppose vehicle 2 passes the upstream intersection at time point c and arrives downstream intersection at time point f. Its signal delay time is from time point f until the end of red signal at downstream intersection. Only if no vehicle passes the upstream intersection between time point b and c, would vehicle 2 be the only vehicle in the queue. Otherwise, vehicle 2 would experience extra queue discharging delay. So the shortest delay time is considered as the length of the red time from the time point when vehicles which passed the upstream intersection at the end of the green time arrive the downstream intersection to the end of red signal (from point f to the end of G_{RD}). Case II and Case III are analyzed in a similar way.

Therefore the lower and upper boundary of 99th percentile of travel time could be approximated as:

$$TT_{upper} = D_{upper} + \mu_I + 3\sigma_I + T_R \quad (3.6)$$

$$TT_{lower} = D_{lower} + \mu_1 - 3\sigma_1 + T_R \tag{3.7}$$

Where, TT_{upper} and TT_{lower} are the upper and lower boundary of travel time respectively, and

T_R is the start delay.

From Figure 3.3, $\mu_1 + 3\sigma_1$ is the upper boundary of 99th percentile of non-stopped vehicle (free flow) travel times and travel time of stopped vehicles could be treated as free flow travel time plus signal delay. Therefore, TT_{upper} could also be considered as the upper boundary of 99th percentile of travel time for the stopped vehicles. For the same reason, TT_{lower} is the lower boundary of 99th percentile of travel time.

Then, μ_2 and σ_2 can be calculated as:

$$\mu_2 = (TT_{upper} + TT_{lower})/2 \tag{3.8}$$

$$\sigma_2 = (\mu_2 - TT_{lower})/3 \quad \text{or} \quad \sigma_2 = (TT_{upper} - \mu_2)/3 \tag{3.9}$$

- Estimation of *p*

Since the parameter *p* is the proportion of non-stopped vehicles and *1-p* is the proportion of stopped vehicles, it can be estimated by combining the signal timing information and arrival rate at the entrance of the corridor, as shown in Figure 3.5.

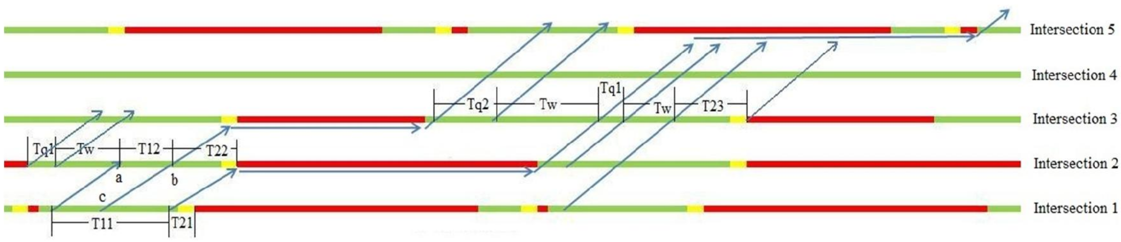


Figure 3.5 Example on Estimation of *p*

Suppose Intersection 1 is the entrance of the corridor. Based on the time-space diagram and the free flow speed, (shown as the slope of the arrows in the figure) vehicles passing Intersection 1 during T_{11} are non-stopped vehicles while vehicles pass during T_{21} are stopped vehicles. Let the average arrival rates (veh/s) during T_{11} and T_{21} be q_1 and q_2

respectively. The average arrival rate could be obtained for example from loop detectors at the stop bar. Therefore the number of non-stopped vehicles on the link between Intersection 1 and Intersection 2, $N_{11}=T_{11}\times q_1$, and the number of stopped vehicles $N_{21}=T_{21}\times q_2$. Then ratio of the first link $p_1=N_{11}/(N_{11}+N_{21})$.

The situation gets more complicated when calculating the ratio for the second link. The vehicles pass during T_{21} become a queue at Intersection 2. When the signal turns green at Intersection 2, it takes time T_{q1} to discharge the queue which is propagated during T_{21}. As a result, stopped vehicles in link 1 become non-stopped vehicles in link 2. After the queue is fully discharged, there is a period of time T_w when no vehicles pass Intersection 2 because the next platoon from intersection 1 will not arrive at Intersection 2 until time point a. So vehicles which pass Intersection 2 between point a and point b (T_{12}) are non-stopped vehicles while vehicles passing Intersection 2 between point b and the green time end (T_{22}) are stopped vehicles. The total number of non-stopped vehicles in link 2, say, N_{12} is equal to number of vehicles passing during T_{q1} and T_{12}. So N_{12} = $N_{21}+T_{12}\times q_1$. The number of stopped vehicles of link 2, $N_{22} = T_{22}\times q_1$. Then ratio of non-stopped vehicles at link 2, $p_2=N_{12}/(N_{12}+N_{22})$. The subsequent links are analyzed in a similar way.

Different signal settings and different offsets will result in different scenarios. For example, there are two periods of unused green time T_w at intersection 3. However, the analysis technique is the same and can be applied to most real world situations.

Figure 3.6 shows a comparison of travel time distributions estimated from signal timing and geometry (red curve) and from data using EM algorithm (blue curve). The NGSIM Peachtree Street Dataset Noon traffic condition was used for the test. Although the estimation of some parameters, such as the mean of the second peak in Section 3, is not very accurate, they are good enough for a prior distribution for further updating. A detailed description of how to update travel time distributions will be given in Chapter 5.

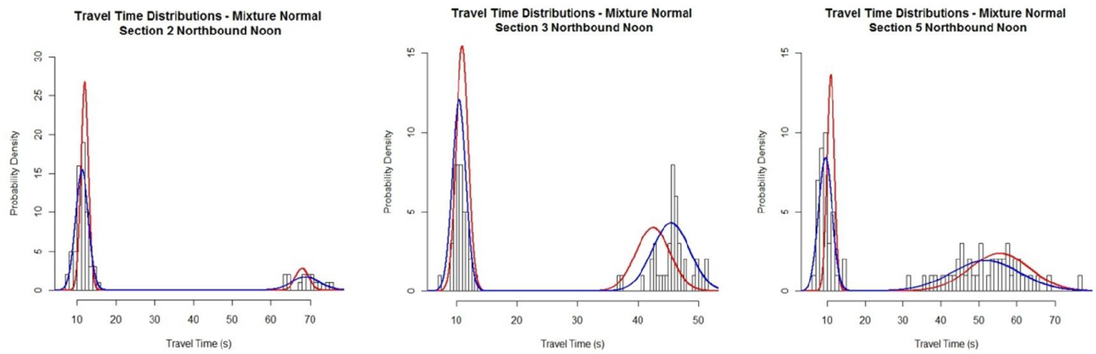


Figure 3.6 Comparison Between EM Algorithm and Estimation from Signal Timing and Geometry

Chapter 4 Mean Route Travel Time Estimation

So far the discussion of travel time estimation has been limited to a single arterial link. However, it is more useful if the route travel time can be estimated. A route is defined as a consecutive sequence of arterial links. General road users are more interested in obtaining the travel time information from their origins to destinations which most likely consists of multiple links. This chapter proposes a method based on Markov Chain to estimate mean route travel time.

Due to the signal control, congestion, and traffic from the side street, a vehicle's travel times between arterial links is not necessarily independent. For example, vehicles that did not stop on an upstream link may have a certain probability of not stopping on the downstream link while vehicles that were stopped on the upstream intersection could have a different probability of not stopping on the downstream link. The simplest case would be that the travel time on the current link depends only on the state encountered on the immediate upstream link. That is, the sequence of traffic states encountered on the arterial links is governed by a Markov Chain (Anderson and Goodman, 1957; Lin et al., 2003; Geroliminis and Skabardonis, 2005). For example, the transition matrix for two consecutive links, with two traffic states is:

$$P_k = \begin{bmatrix} p_{00,k} & 1 - p_{00,k} \\ 1 - p_{11,k} & p_{11,k} \end{bmatrix}$$

Where,

$p_{00,k}$ is the probability of a non-stopped travel time state on link *k* given non-stopped travel time state on link *k-1,* and;

$p_{11,k}$ is the probability of stopped travel time state on link *k* given stopped travel time state on link *k-1*

To use a Markov Chain to describe the movement through a sequence of arterial links, the system states, the transition matrix, and initial state probabilities should be defined properly.

System States: System states are different travel time categories which could be derived from link travel time distribution diagrams. Suppose our interest is focused on through – through vehicles. As stated in Section 2.1 and illustrated in Figure 2.2 (b), four states in the system are defined: non-stopped, non-stopped with delay, stopped, and stopped with delay.

Transition matrix: Four system states results in a four by four transition matrix. The one step transition matrix (two consecutive links) is shown as the following matrix, where p_{ij} indicates the probability of switching from travel time state *i* to *j*.

$$\mathrm{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{14} \\ p_{21} & p_{22} & \cdots & p_{24} \\ \vdots & \vdots & & \vdots \\ p_{41} & p_{42} & \cdots & p_{44} \end{pmatrix}$$

p_{ij} also could be expressed in conditional probability as follows:

$$p_{ij} = P\{S_{n+1} = j | S_n = i\} \quad \text{for } i, j = 1,2, \ldots, 4 \text{ and } n = 1,2,3, \ldots \tag{4.1}$$

Where, $S_{n+1} = j$ indicates travel time state at link *n+1* is *j*.

Initial State Probabilities: Initial state probabilities are defined as the probabilities of each travel time state before vehicles entering the first link of the route.

$$\boldsymbol{\lambda_0} = (\lambda_{01} \quad \lambda_{02} \quad \lambda_{03} \quad \lambda_{04})$$

Where, $\boldsymbol{\lambda_0}$ is the initial state probability vector which is consistent with the travel time states and $\sum_{i=0}^{4} \lambda_{0i} = 1$.

The Markovian property can also be expressed as conditional probability as follows:

$$P\left\{S_{n+1} = j \middle| S_n = i, S_{n-1} = i_{n-1}, S_{n-1} = i_{n-2} \cdots S_1 = i_1\right\} = P\{S_{n+1} = j | S_n = i\} = p_{ij} \tag{4.2}$$

Because of the conditional independence, the joint probability of each step in a Markov Chain is the product of the transition probabilities of each step:

$$P\{S_0 = i_0, S_1 = i_1 \cdots, S_n = i_n\} = \lambda_{0i} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} \tag{4.3}$$

Based on Equation 4.3, given the initial distribution $\boldsymbol{\lambda_0}$ and all the link-link transition matrices $P^{(x)}$, the mean route travel time can be calculated by multiplying the mean travel time of each possible combination of system states and its probability. That is:

$$TT_{route} = \sum_{i_1=1}^{4} \sum_{i_2=1}^{4} \cdots \sum_{i_n=1}^{4} \left(T_{i_1}^{(1)} + T_{i_2}^{(2)} + \cdots + T_{i_n}^{(n)} \right) \times p_{i_0 i_1} \times p_{i_1 i_2} \times \cdots \times p_{i_{n-1} i_n} \tag{4.4}$$

Where, TT_{route} is mean route travel time, and

$T_{i_k}^{(j)}$ means travel time of link j is in state i_k.

Theoretically there are total 4^n possible different combinations of travel time states from origin to destination, where n is the total number of link in the route. However, if the traffic signals along the route are coordinated, the signal control will tend to reduce the possible combinations. In addition, the probabilities of travel time state 2 and state 4 may be small, which means that those states might not be observed in a finite sample.

To check on the validity of this model, predicted mean route travel times are compared to the measured route travel time. Travel times of Noon traffic condition of NGSIM Peachtree St dataset (Figure 3.1) are used to test the model. To apply Equation 4.4, mean travel times of each state in every link need to be calculated first. These are just the means of corresponding distributions. The result is shown in Table 4.1.

Table 4.1 Estimated Mean Travel Times of Different States in Each Link

	State 1	State 2	State 3	State 4
Link 2	11.29	38.12	68.87	88.08
Link 3	10.49	26.02	45.47	75.82
Link 4	9.54	N/A	N/A	N/A
Link 5	9.58	23.47	51.76	84.88

Theoretically, there are four different initial states. However, in almost all cases, the initial state of travel time is either state 1 (non-stopped) or state 3 (stopped), so the initial probabilities of states 2 and 4 were taken to be zero.

Case I: Given the vehicle is a non-stopped vehicle at the entrance

Initial state: λ_0= [1 0 0 0];

The transition matrices estimated from the data are:

$$P^{(1)} = \begin{bmatrix} 14/27 & 0 & 13/27 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} P^{(2)} = \begin{bmatrix} 1/14 & 0 & 13/14 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$P^{(3)} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 12/13 & 0 & 1/13 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Where, $P^{(1)}$ is the transition matrix from initial distribution to travel time states of Link 2,

$P^{(2)}$ is the transition matrix from Link 2 to Link 3,

$P^{(3)}$ is the transition matrix from Link 3 to Link 5, since Link 4 only has one state.

Only 4 combinations of travel time state sequence were observed:

1-1-3 (0.0370), 1-3-1 (0.4444), 1-3-3 (0.0370), and 3-1-3(0.4815)

The value in the bracket is the probability of this particular travel time states combination.

The estimated mean route travel time is:

$$\begin{aligned} TT &= (11.29+10.49+51.76) \times 0.0370 + (11.29+45.47+9.58) \times 0.4444+ \\ &\quad (11.29+45.47+51.76) \times 0.0370 + (68.87+10.49+51.76) \times 0.4815 + 9.54 \\ &= 108.89s \end{aligned}$$

The mean travel time from the data is 110.78s, and an approximate 95% confidence interval is (97.16s; 124.40s).

Case II: Given the vehicle is a stopped vehicle at the entrance

Initial state: λ_0=[0 0 1 0];

The transition matrices estimated from the data are:

$$P^{(1)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 50/55 & 2/55 & 2/55 & 1/55 \\ 0 & 0 & 0 & 0 \end{bmatrix} P^{(2)} = \begin{bmatrix} 24/50 & 0 & 26/50 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$P^{(3)} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 25/28 & 0 & 1/28 & 2/28 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

There are 7 combinations of travel time state sequences observed here:

1-1-3 (0.4364); 1-3-1 (0.4182); 1-3-3 (0.0182); 1-3-4 (0.0364); 3-1-3 (0.0364); 2-3-1 (0.0364); 4-1-3 (0.0182)

The predicted mean route travel time is:

$$\begin{aligned} TT = {} & (11.29+10.49+51.76) \times 0.4364 + (11.29+45.47+9.58) \times 0.4182+ \\ & (11.29+45.47+51.76) \times 0.0182 + (11.29+45.47+84.88) \times 0.0364 + \\ & (68.87+10.49+51.76) \times 0.0364 + (38.12+45.47+9.58) \times 0.0364 + \\ & (88.08+10.49+51.76) \times 0.0182 + 9.54 \\ = {} & 87.40\text{s} \end{aligned}$$

The mean travel time from the data is 86.25s, with an approximate 95% confidence interval of (79.59s; 92.91s).

Results show in both cases, estimated mean route travel times from the Markov Chain model are close to those measured from data. That means this Markovian relationship could approximate the dependence of consecutive travel time states.

Chapter 5 Applications of Travel Time Distribution

Travel time distributions provide a range of performance measures, such as mean travel time, the standard deviation, and the 95th percentile travel time. Meanwhile, GPS data offer us information of individual vehicle in real-time. Useful information can be derived combining travel time distribution and GPS data. This Chapter describes how real-time traffic conditions can be identified, and how to update the parameters of travel time distributions.

5.1 Real-Time Traffic Condition Identification

Figure 3.1 and Figure 3.2 show that the travel time distributions of each link under different traffic conditions are different. In consequence, transition matrices between links may also be different. Therefore, different traffic conditions should be characterized by travel time distributions of each link, in each state, and the sequence of link-link transition matrices. When a GPS probe vehicle goes through a certain route, it records travel time of each individual link. The probabilities that this travel time sequence belongs to each existing traffic condition are also different. If the probability that this travel time sequence belongs to traffic condition A is higher than all other traffic conditions, it possibly suggests that the current traffic condition is A. Then how to calculate these posterior probabilities becomes a key question. One approach using Bayes Theorem is proposed next.

Either the travel time distribution of the whole route or the travel time distribution of each link can be used to identify a traffic condition. Using the travel time of the whole route is a more direct way. However, it is more feasible to use travel time distribution of every single link because then the route could be constructed freely as long as the relationship between links can be found. In addition, under different traffic conditions, the travel time distribution of the whole route could be similar, even though the travel time distributions of single links differ. Therefore, in this study, link travel time distributions are used.

Suppose an arterial route consists of n consecutive links, and a sequence of travel time data of each link is collected from a GPS vehicle running along this route. Let S_i and T_i denote travel time state and travel time of link i respectively. The joint probability of travel time states could be expressed by the following equation.

$$P(S_1 = s_1, S_2 = s_2, \dots, S_n = s_n) \\ = P(S_n = s_n | S_{n-1} = s_{n-1}, \dots, S_2 = s_2, S_1 = s_1) \times P(S_{n-1} = s_{n-1} | S_{n-2} \\ = s_{n-2}, \dots, S_2 = s_2, S_1 = s_1) \times \cdots \times P(S_2 = s_2 | S_1 = s_1) \times P(S_1 = s_1) \tag{5.1}$$

Assuming the travel time of a single link is conditionally independent of travel time states on other links:

$$\mathrm{P}(\mathrm{T_i} = \mathrm{t_i} | \mathrm{S_1} = \mathrm{s_1}, \mathrm{S_2} = \mathrm{s_2}, \dots, \mathrm{S_i} = \mathrm{s_i}, \dots, \mathrm{S_n} = \mathrm{s_n}) = \mathrm{P}(\mathrm{T_i} = \mathrm{t_i} | \mathrm{S_i} = \mathrm{s_i}) \tag{5.2}$$

Combined with Equation 5.1 and Equation 5.2, the join distribution of travel time and travel time states can be seen in Equation 5.3.

$$P(T_1 = t_1, \dots, T_n = t_n, S_1 = s_1, \dots, S_n = s_n) = \left(\prod_{i=1}^{n} P(T_i = t_i | S_i = s_i)\right) P(S_1 \\ = s_1, \dots, S_n = s_n) \tag{5.3}$$

So the marginal distribution for travel times becomes

$$P(T_1 = t_1, \dots, T_n = t_n) = \sum_{S_1 = s_1, \dots, S_n = s_n} \left(\prod_{i=1}^{n} P(T_i = t_i | S_i = s_i)\right) P(S_1 = s_1, \dots, S_n = s_n) \tag{5.4}$$

Assuming there are m different traffic conditions, the probability that a given travel time sequence belongs to traffic condition C_i can be calculated by Bayes theorem (DeGroot, and Schervish, 2002)

$$P(C_i|T_1 = t_1, \dots, T_n = t_n) = \frac{P(T_1=t_1,\dots,T_n=t_n|C_i)\times P(C_i)}{\sum_{j=1}^{m} P(T_1=t_1,\dots,T_n=t_n|C_j)\times P(C_j)} \tag{5.5}$$

Since the travel time distributions under every traffic condition are known, the likelihood $P(T_1 = t_1, \dots, T_n = t_n|C_j)$ can be obtained from Equation 5.4. The prior probability $P(C_j)$ can be flat or calculated from the traffic volume of each traffic condition. For example, in the NGSIM Peachtree St Dataset, there are two traffic conditions: Noon and PM. 82 vehicles went through the route during the Noon period and 67 vehicles during the PM period. If a vehicle is randomly picked, the probability it belongs to noon data set, which is P(Noon)=82/(82+67)=0.55. Consequently, P(PM)=1-0.55=0.45. Then the marginal distribution of travel time under different traffic conditions $P(T_1 = t_1, T_2 = t_2, T_3 = t_3|Noon)$ and $P(T_1 = t_1, T_2 = t_2, T_3 = t_3|PM)$ can be then calculated from travel time distributions.

Because currently GPS data of that arterial segment are not available, travel time data from the NGSIM Peachtree St Dataset are used to test the method. If it works well, vehicles from the Noon data set should have a high posterior probability of belonging to the Noon traffic condition and a low probability of belonging to the PM traffic condition, and vice versa. Figure 5.1 shows the results.

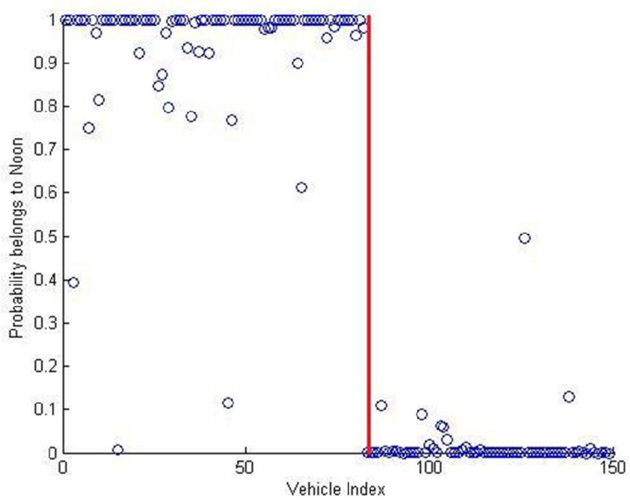


Figure 5.1 Traffic Condition Identification

The dots on the left side of the vertical line in Figure 5.1 are the vehicles from the Noon data set and right side are the vehicles from the PM data set. One dot represents one vehicle. The figure shows that generally, data from a single GPS-equipped vehicle can discriminate between the two traffic conditions. Although the probabilities of a few dots on the left are low, most of them located around one. It is worth pointing out that the identification could be wrong (the other side) or vague (the probability is around 0.5). In that case, more than one GPS vehicle is needed to identify the traffic condition.

One limitation of this approach is that it could only discriminate GPS data from existing traffic conditions. The previous example assumed the data are from either Noon or PM. When a GPS travel time sequence is from a new traffic condition, the Bayes' approach couldn't tell because the posterior probability is actually a ratio, among all existing traffic conditions. One way to identify new traffic condition is to calculate the probabilities this travel time sequence belongs to each traffic condition separately, which is $P(T_1 = t_1, \dots, T_n = t_n | C_i)$. If all the probabilities are very small, it could be considered as evidence that this GPS vehicle is from a new traffic condition. According to Equation 5.4, the marginal distribution of travel times is the product of the probabilities of each link travel time belongs to the corresponding travel time distribution given the travel time state. If a travel time falls out of the 99th percentile of the travel time distribution, which the corresponding probability is less than 0.01, then it can be considered not come from this travel time distribution. So as a rule of thumb, under each traffic condition, if the product of the probability is less than 0.01^N, where N is the number of links, then this GPS data doesn't belong to this traffic condition. For example, in Peachtree St example, the threshold could be set to $0.01^3=10^{-6}$.

5.2 Travel Time Distribution Parameter Update

This section proposes an approach to update the parameters of the travel time distribution under a Bayesian scheme. To apply a Bayesian approach, a prior distribution must be chosen first. When no prior information is available, it is common to use a noninformative prior. A noninformative prior is usually a uniform distribution or a normal distribution with a large variance which has minimal impact on the posterior distribution. Many statisticians are in favor of noninformative priors because they appear to be more objective. But in some cases, noninformative priors can lead to improper posteriors. More importantly, in our case, even if the noninformative prior is a proper prior, given the density of the GPS probe vehicles is very low and spread into different traffic conditions, it might take a very long time to update to a good posterior distribution. Especially under the traffic conditions such as off-peak hours when there are fewer GPS probes, the time would be even longer. Consequently, we considered travel time distributions estimated from signal timing and geometric structure of the arterial (Chapter 3.2) as prior distributions when no prior travel time data are available. Combining with GPS data, posterior distributions can then be calculated.

Unfortunately, the posterior distribution of a mixture normal density is analytically intractable (McLachlan and Peel, 2000). As a result, Markov Chain Monte Carlo (MCMC) simulation (Berg, 2004) is applied to estimate the posterior distribution. One particular sampling technique of Markov Chain called Gibbs sampling (Diebolt and Robert, 1994) is used. In this paper, WINBUGS ("BUGS" stands for Bayesian inference using Gibbs sampling) software is used to compute the posterior distributions.

Notice that the mixture normal distribution consists of two unimodal normal distributions and the two normal components are connected by the ratio *p*. *p* follows a Bernoulli random variable which captures the variations of the portions of the two components. For a unimodal normal distribution with unknown mean and variance, there is conjugate prior for the posterior distribution, given the data are iid normal.

The conjugate prior for two parameter normal is the normal-inverse gamma distribution (Bernardo and Smith, 1993) with density

$$f(\mu, \sigma^2) = f(\mu|\sigma^2) f(\sigma^2)$$

When $f(\mu|\sigma^2) = N(\alpha_0, \beta_0)$ and $f(\sigma^2) = IG(\gamma_0, \delta_0)$, $IG(\gamma_0, \delta_0)$ is the Inverse-gamma distribution with parameter γ_0, δ_0. Note that if $X \sim IG(\gamma_0, \delta_0)$, then $X^{-1} \sim \mathrm{Gamma}(\gamma_0, \delta_0)$.

Therefore, we replace σ^2 with precision $\lambda = 1/\sigma^2$. With this approach, the prior for the parameter (μ, λ) is a normal-gamma distribution, denoted by

$$f(\mu, \lambda) \sim NG(\alpha_0, \beta_0, \gamma_0, \delta_0)$$

Suppose the data (likelihood) are iid normal, $f(Y_i|\mu, \lambda) \sim N(\mu, \lambda)$ $i = 1,2, \ldots, n$. Because of independence, the total likelihood is the product of each likelihood, denoted by $f(Y|\mu, \lambda) = \prod_{i=1}^{n} f(Y_i|\mu, \lambda)$. Suppose the sample mean is $\bar{y}$ and the sample variance is V_n.Then, the corresponding posterior distribution is also a normal-gamma distribution with parameters $(\alpha_1, \beta_1, \gamma_1, \delta_1)$.

$$f(\mu, \lambda|y) = f(\boldsymbol{Y}|\mu, \lambda) \times f(\mu, \lambda) = NG(\alpha_1, \beta_1, \gamma_1, \delta_1)$$

Where,

$$\alpha_1 = \alpha_0 + \frac{n_1}{2}, \beta_1 = \beta_0 + \frac{n_1}{2}\left(V_n + \frac{\delta_0(\bar{y} - \gamma_0)^2}{\delta_0 + n_1}\right), \gamma_1 = \frac{\gamma_0\delta_0 + n_1\bar{y}}{\delta_0 + n_1}, \delta_1 = \delta_0 + n_1$$

Given the above structure, a hierarchical Bayesian model could be constructed in WINBUGS. The structure of the hierarchical model can be specified as below:

$Y_i | P_i, \lambda_{(Pi)}, \mu_{(Pi)} \sim Normal(\mu_{(Pi)}, \lambda_{(Pi)})$

$\mu_{(Pi)} | P_i, \lambda_{(Pi)} \sim Normal(\varepsilon_{(Pi)}, \delta_{(Pi)})$

$\lambda_{(Pi)} | P_i \sim Gamma(\alpha_{(Pi)}, \beta_{(Pi)})$

$P_i \sim Bern(r)$ *r: determined by ratio of non-stopped and stopped vehicles*

Some simulation experiments are run in WINBUGS to test the model. The number of Markov Chains is set to one. One simulation scenario includes 21,000 samples, in which the first 1,000 samples are burn-in samples and the following 20,000 samples are used to create the posterior distribution. The posterior means of each parameter are selected to be the new parameters of the travel time distribution. Figure 5.2 shows an example of the comparison between the posterior travel time distribution from Bayesian update (red curve) and travel time distribution directly estimated from EM algorithm (blue curve). In this example, same travel time data are used. One can see the two approaches lead to similar results because asymptotically Bayesian inference and classical maximum likelihood estimation should converge to the same distribution when the sample size is big enough.

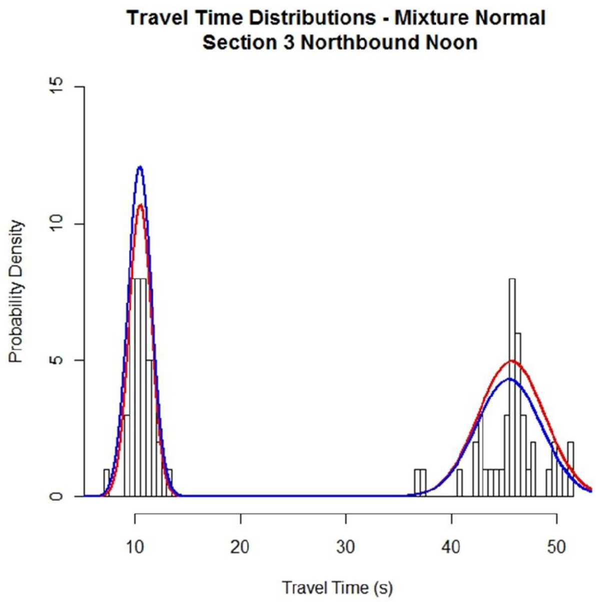


Figure 5.2 Posterior Travel Time Distribution

Chapter 6 A Case Study – NGSIM Peachtree Dataset

In previous chapters, methods regarding for constructing travel time distribution and their applications were discussed. This chapter aims to connect these models together and demonstrate a comprehensive case study based on the NGSIM Peachtree dataset.

Suppose we are interested in estimating the travel time distributions of a given arterial under different traffic conditions. Usually the signal timing plans and the geometric structure of the arterial are known, but the travel time data are not available in the first place. As a result, instead of starting with flat prior, travel time distributions are estimated based on signal timing plans and the geometric structure of the arterial. (Chapter 3.2) When GPS data are collected for the targeted route, traffic condition identification process (Chapter 5.1) is run and the probability one particular data point belongs to every traffic condition is calculated. Data are then classified according to the posterior probabilities. Finally, a Bayesian update (Chapter 5.2) is run to calculate posterior distributions under each traffic condition combining with classified data. Furthermore, when new GPS data are available, the posterior distributions from the previous update are considered as new priors and the update process is repeated. As long as more and more GPS data become available, the posterior distribution would be more and more consistent with reality. Figure 6.1 illustrates the process stated above.

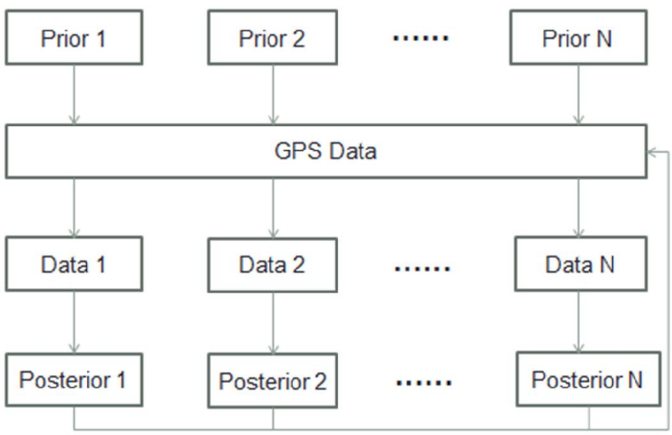


Figure 6.1 Travel Time Distribution Update Process

The NGSIM Peachtree Street dataset are used to test the scenario. There are two traffic conditions in the dataset: Noon and PM. We assume there is no prior travel time data under both traffic conditions and consider travel time data from vehicle trajectories as incoming GPS data. The total sample size of GPS data in two traffic conditions is 149 which are further divided into two parts in order to perform the update process iteratively. One GPS datum is one travel time sequence including the three link travel times in Link 2, Link 3 and Link 5 respectively.

- **Prior distributions**

First, prior distributions are estimated from signal timing and geometric structure of the arterial. Table 6.1 and Table 6.2 list the distribution parameters under traffic condition Noon and PM respectively.

Table 6.1 Prior Travel Time Distribution Parameters of Each Section at Noon

	μ_1	σ_1	μ_2	σ_2	p
Section 2	12	1.1	68	1.667	0.852
Section 3	11	1	42.5	2.833	0.577
Section 5	11	1	56.5	7.833	0.423

Table 6.2 Prior Travel Time Distribution Parameters of Each Section at PM

	μ_1	σ_1	μ_2	σ_2	p
Section 2	12	1.1	18	2.33	0.682
Section 3	11	1	43	3	0.343
Section 5	11	1	55.5	6.83	0.324

- **Data Classification**

The traffic condition identification process is then performed on the first half of the data (75 travel time sequences). Figure 6.2 shows the probabilities that each GPS vehicle belongs to Noon.

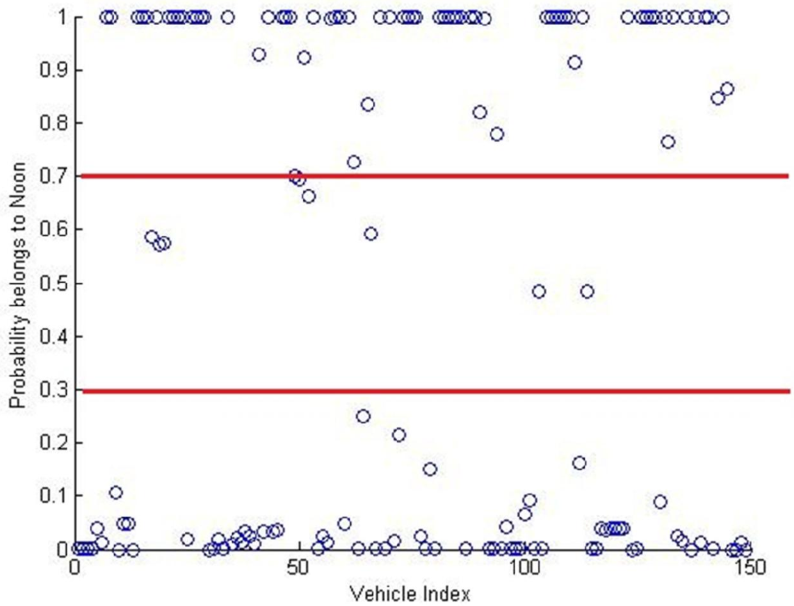


Figure 6.2 Data Classification Based on Prior Distribution

The following criteria are applied while classifying the data:

If P(belong to noon)>=0.7 -> use this data to update Noon travel time distribution

If P(belong to noon)<=0.3 -> use this data to update PM travel time distribution

If 0.3<P(belong to noon)<0.7 -> discard the data (because it is too vague to tell which traffic condition this data belongs to)

- **Bayesian update**

Applying the Bayesian update, the posterior means of each parameter under different traffic conditions, after updating with 75 observations are shown in Table 6.3 and Table 6.4.

Table 6.3 Posterior Travel Time Distribution Parameters of Each Section at Noon after Updating Half of Data

	μ_1	σ_1	μ_2	σ_2	p
Section 2	11.87	1.848	70.59	2.821	0.7318
Section 3	10.47	1.108	45.21	3.457	0.4574
Section 5	10.16	1.472	54.59	8.707	0.5248

Table 6.4 Posterior Travel Time Distribution Parameters of Each Section at PM after Updating Half of Data

	μ_1	σ_1	μ_2	σ_2	p
Section 2	13.21	1.371	19.52	2.818	0.6406
Section 3	10.96	1	45.28	4.561	0.2433
Section 5	12.03	1.183	52.54	6.2	0.1478

- Repetition

Considering the posterior distributions in Table 6.3 and Table 6.4 as new prior distributions, the traffic condition identification process is performed again on the second half of the data (74 samples). Figure 6.3 shows the identification results.

The same criteria are applied to classify the data and then the corresponding posterior distributions are updated. Table 6.5 and Table 6.6 show the posterior means of each parameter under different traffic conditions (Noon and PM) after updating with all 149 observations.

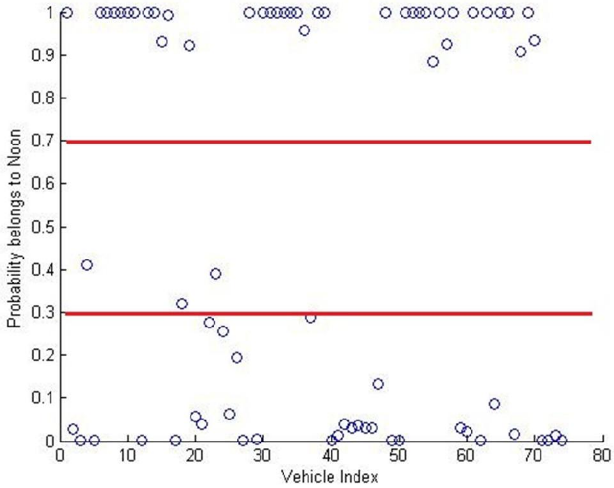


Figure 6.3 Data Classification Based on First Posterior

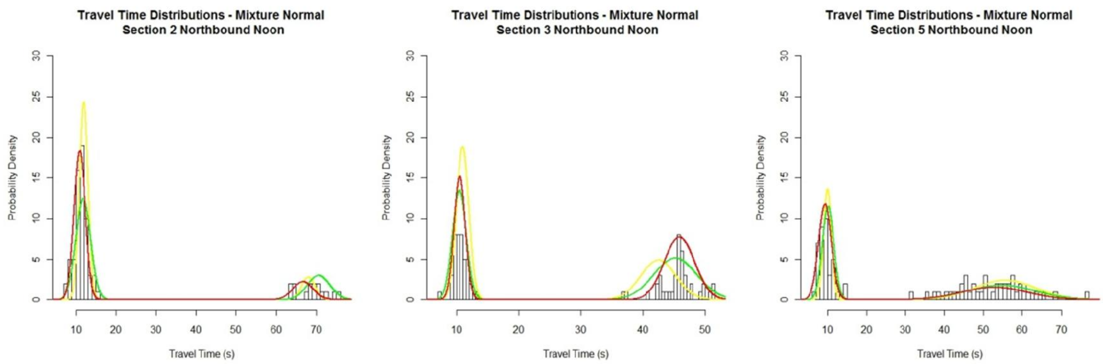
Table 6.5 Posterior Travel Time Distribution Parameters of Each Section at Noon After Updating All Samples

	μ_1	σ_1	μ_2	σ_2	p
Section 2	10.99	1.408	66.68	2.541	0.8204
Section 3	10.52	0.9083	45.88	2.4368	0.4221
Section 5	9.369	1.6507	52.79	8.4215	0.6021

Table 6.6 Posterior Travel Time Distribution Parameters of Each Section at PM After Updating All Samples

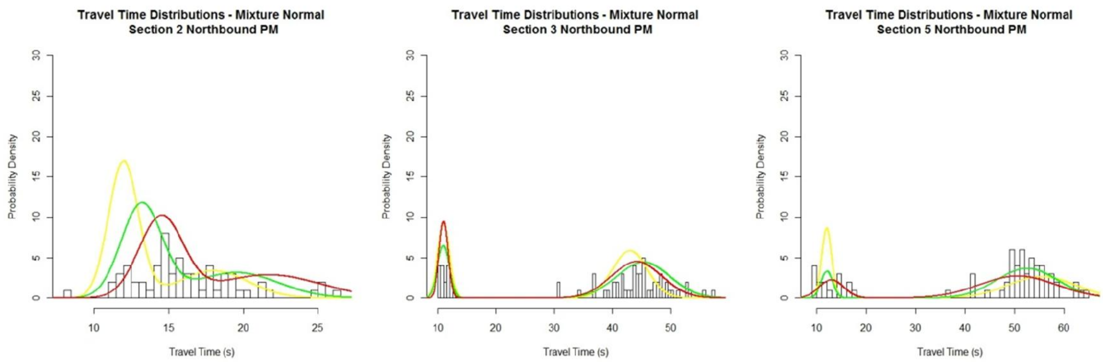
	μ_1	σ_1	μ_2	σ_2	p
Section 2	14.51	1.4767	21.68	3.5132	0.5918
Section 3	11	0.7638	44.14	4.3315	0.2715
Section 5	12.92	2.4884	50.65	7.7452	0.2104

Figure 6.4 and Figure 6.5 show the comparison among prior distributions (yellow), posterior distributions after updating half of the data (green) and posterior distributions after updating all the data (red) under two traffic conditions. The histograms represent the original GPS data.



(a) (b) (c)

Figure 6.4 Comparison Among Prior and Posterior Distributions at Noon

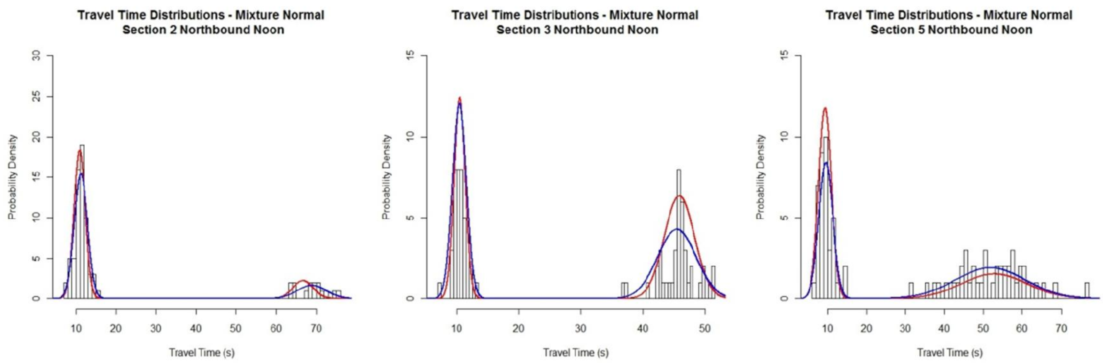


(a) (b) (c)

Figure 6.5 Comparison Among Prior and Posterior Distributions at PM

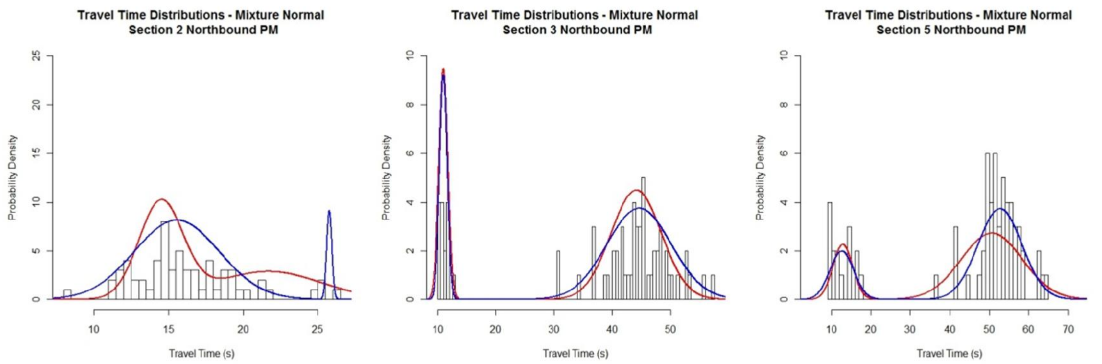
It can be seen from the figures that if the estimation of prior distribution is quite accurate such as Figure 6.4 (c), the posterior distribution doesn't differ very much from the prior. If the estimation of prior distribution is not good such as Figure 6.5 (a), the posterior distributions tend to pull the prior distribution in the right direction. If more data are used, then the posterior distribution should become more accurate.

The posterior distributions after updating with all data are compared with results from EM algorithm as shown in Figure 6.6 and Figure 6.7. The red curve represents posterior distribution from Bayesian update and the blue curve represents distributions estimated from EM algorithm.



(a) (b) (c)

Figure 6.6 Comparison Between Bayesian Update and EM Algorithm at Noon



(a) (b) (c)

Figure 6.7 Comparison Between Bayesian Update and EM Algorithm at PM

From the previous two figures, it can be seen that the Bayesian update and EM algorithm achieved similar results except Section 2 at PM (Figure 6.7 c). In this case, the EM algorithm indicates that the two components are separated because there is no data in between. However, the Bayesian posterior distribution says the two components are mixed and the high probabilities between the two components (around 21 to 24 second) imply some data are missing. If we take a look at the time space diagram of Section 2 at PM as shown in Figure 6.8, the Bayesian result is more convincing.

Figure 6.8 Time Space Diagram of Section 2 at PM

The arterial link between Intersection 1 and Intersection 2 are noted as section 2. If a vehicle passes Intersection 1 at the beginning of green phase at time point a and arrives Intersection 2 at time point b, it will experience signal delay from time point b to time point d, the end of red phase. If a vehicle passes Intersection 1 at time point c and arrives Intersection 2 at time point d when the green phase of Intersection 2 just begins, it becomes a non-stopped vehicle. Vehicles passing Intersection 1 between time point a and c experience signal delay which gradually decreases from a to c. In other words, the signal delay time gradually decreases to zero from d-b and thus there should be no obvious gap between non-stopped and stopped vehicles. The posterior distribution could capture this phenomenon because the Bayesian update combines information from data, and from the prior distribution which is estimated from signal time information.

In addition, Kolmogorov-Smirnov test (Ross, 2004) is run to test the goodness of fit to both EM and Bayesian models. The results are shown in Table 6.7 and Table 6.8. D statistic expresses the distance between the empirical cdf and theoretical cdf and P-value expresses the probability to accept the null hypothesis that the data are drawn from the theoretical distribution. Both the D statistic and p-value show in all circumstances EM fits the data better than Bayesian update. That is because EM is purely based on the data, while the Bayes estimate also incorporates prior information. In addition, the traffic condition identification or data classification process may put the data into wrong traffic condition. Although the error rate is low, still some data are wrongly updated. However, as stated in the previous paragraph, sometime Bayesian approach could reflect the real world situation when some data is missing while frequentist (EM) did not.

Table 6.7 D-statistic of Each Section under Different Traffic Condition

K-S test D statistic	**Section 2 Noon**	**Section 3 Noon**	**Section 5 Noon**	**Section 2 PM**	**Section 3 PM**	**Section 5 PM**
EM	0.0649	0.0797	0.0613	0.0695	0.0745	0.0807
Bayesian	0.1356	0.0926	0.1303	0.1770	0.1065	0.8079

Table 6.8 P-value of Each Section under Different Traffic Condition

p-value	**Section 2 Noon**	**Section 3 Noon**	**Section 5 Noon**	**Section 2 PM**	**Section 3 PM**	**Section 5 PM**
EM	0.917	0.7211	0.9407	0.9405	0.8772	0.2037
Bayesian	0.1353	0.5344	0.1571	0.0516	0.475	0.0110

Chapter 7 Conclusions and Further Research

7.1 Conclusions

The primary objectives of this thesis were to characterize arterial travel time patterns by travel time distributions and then utilize those distributions to obtain useful information, such as mean route travel time and real-time traffic condition, by combining with GPS data.

First of all, travel time patterns of signalized arterial links were analyzed based on travel time histograms. The thesis mainly focused on through-through vehicles. Four travel time states were defined: non-stopped, non-stopped with delay, stopped and stopped with delay.

Empirical studies show link travel time distributions can be approximated using mixtures of normal densities. According to the definition of travel time states, two approaches to estimate travel time distributions were proposed. If prior travel time data is available, travel time distributions can be estimated empirically, using the EM algorithm. Otherwise, travel time distribution can be estimated based on signal timing information and geometric structure of the arterial.

Link travel time was then extended to route travel time. Because travel times on successive links are not independent, a Markov Chain based model was proposed to capture the potential relationship of travel time between consecutive links. Empirical examples showed the estimated route travel time by the Markov Chain model was very accurate with errors less than 2%.

When new GPS travel time data is collected, they could be used to predict the travel time by identifying the traffic condition. The traffic condition identification approach was developed based on Bayes Theorem. Results showed that in most cases, one GPS sample was enough to discriminate between the two traffic conditions. Compared to other approaches, the proposed method needs much less real-time information. Moreover, these GPS data could also be used to update the parameters of the travel time distributions using a Bayesian update. The iterative update process makes the posterior travel time distributions more and more accurate.

Finally, a comprehensive case study using the NGSIM Peachtree Street dataset was conducted. The case study first estimated prior travel time distributions based on signal timing and geometric structure under different traffic conditions. Then, travel time data were classified into different traffic conditions and corresponding distributions were updated. In addition, results from the Bayesian update and EM algorithm were compared. Overall, the EM algorithm fit the data better than Bayesian update. However, in some scenarios the Bayesian approach could reflect the real world situation when some data were missing.

7.2 Further research

This thesis lays a foundation for the characterization of travel time distributions of urban arterial. Future research will need to address the following issues:

- The current research focusing on through-through vehicles. Turning vehicles have more complicated travel time states. Therefore, turning vehicles must necessarily be further divided into sub categories and each category represents only a special scenario of turning. For example, right turn from side streets which go through at the downstream intersection.
- Estimating travel time distributions under oversaturated traffic conditions. When the traffic is in congested condition and queues can't be fully discharged within one cycle, the bimodal pattern may not be applicable and more travel time states are expected. For example, the vehicles which stopped twice for red signal would generate a new state. Meanwhile, the estimation of travel time based on signal time and geometric structure needs to be modified accordingly.
- Criteria to classify the GPS data. The 0.7 and 0.3 thresholds in data classification (Figure 6.2 and Figure 6.3) are selected randomly in the case study. It is necessary to classify the travel times based on certain data classification techniques. If the

interval between the two thresholds is wider, then more data will be discarded. If the interval is narrower, the probability that assign the data into wrong traffic condition is increased. A cost function should be constructed and minimized or maximized to determine the threshold.

References

A. Skabardonis and R. Dowling, “Improved Speed-Flow Relationship for Planning Applications”. Transportation Research Record: Journal of the Transportation Research Board, No. 1572, National Research Council, Washington, D.C., 1997, pp. 18-23.

Alexander M. Hainen, Jason S. Wasson, Sarah M. L. Hubbard, Stephen M. Remias, and Grant D. Farnsworth. Estimating Route Choice and Travel Time Reliability using Field Observations of Bluetooth Probe Vehicles. 90th Annual Conference of Transportation Research Board. 2011

Bernardo J.M, and Smith A.F.M. Bayesian Theory. John Wiley & Sons, Inc., New York, 1993.

Bernd A. Berg. Markov Chain Monte Carlo Simulations and Their Statistical Analysis. WSPC. 2004

Bohnke, P., and E. Pfannerstill. A system for the Automatic Surveillance of Traffic Situations. ITE Journal, Vol. 56, No.1, pp. 41-45. 1986

Box, P.C., and J.C. Oppenlander. Manual of Traffic Engineering Studies, 4th edition. Institute of Transportation Engineers, Washington, D.C., USA, 1976.

Bureau of Public Roads. *Traffic Assignment Manual*. Washington DC: U.S. Department of Commerce. June 1964.

C. Xie, R. Cheu, and D. Lee, “Calibration-Free Arterial Link Speed Estimation Model Using Loop Data”. ASCE J. of Transportation Engineering, Nov/Dec 2001, pp. 507-514.

Cambridge Systematics. Inc. NGSIM Peachtree Street (Atlanta) Data Analysis. 2007

Dailey, D.J., Cathey, F.W., “AVL-Equipped Vehicles as Traffic Probe Sensors”. Final Report, Publication WA-RD534.1. Washington State Transportation Center, University of Washington., 2002

Diebolt J., and C. Robert. Estimation of Finite Mixture Distributions through Bayesian Sampling. Journal of the Royal Statistical Society: Series B (Methodological), Vol. 56, No. 2, 1994, pp. 363–375.

Gary A. Davis and Hui Xiong. Access to Destinations: Travel Time Estimation on Arterials. 2007. MnDot Report MN/RC 2007-35.

Gault, H. E. An On-line Measure of Delay in Road Traffic Computer Controlled Systems. *Traffic Engineering and Control*. Vol. 22, No. 7 pp. 384-389. 1981

Geroliminis, N., Skabardonis, A. Real-time vehicle re-identification and performance measures on signalized arterials. Proceedings of the Ninth International IEEE Conference on Intelligent Transportation Systems, Toronto, Canada. 2006

H. Spiess, “Conical volume-delay functions”. Transportation Science, Vol. 24, No.2, 1990.

H.M. Zhang. Link-Journey-Speed Model for Arterial Traffic. Transportation Research Record 1676. pp 109-115. 1999

Henry X. Liu, Wenteng Ma. A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials. Transportation Research Part C 17 (2009) 11–26

Hui Xiong and Gary A. Davis. Field Evaluation of Model-Based Estimation of Arterial Link Travel Times. In Transportation Research Record: Journal of the Transportation Research Board, No. 2130. Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 149-157.

J.-C. Herrera, D. Work, X. Ban, R. Herring, Q. Jacobson and A. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: the *Mobile Century* field experiment, *Transportation Research C*, 18, pp. 568–583, 2010.

Jing Dong, Hani S. Mahmassani. 2009. Flow breakdown and travel time reliability. Transportation Research Record. No. 2124 pp. 203-212.

Jiyoun Yeon, Lily Elefteriadou, Siriphong Lawphongpanich. 2008. Travel time estimation on a freeway using discrete time Markov chains. Transportation Research Part B. 42 (2008) 325-338

Juan Du. Combined Algorithms for Constrained Estimation of Finite Mixture Distributions with Grouped Data and Conditional Data. Master Thesis. McMaster University. 2002

Klayut Jintanakul, Lianyu Chu, and R. Jayakrishnan. Bayesian Mixture Model for Estimating Freeway Travel Time Distributions from Small Probe Samples from Multiple Days. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2136, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 37–44.

McLachlan G., and D. Peel. Finite Mixture Models. John Wiley & Sons, Inc., New York, 2000.

McLachlan, G. and Krishnan, T. The EM Algorithm and Extensions. Wiley series in probability and statistics. John Wiley & Sons. 1997

Minnesota DOT: Statewide Mileage and Land Miles.

http://www.dot.state.mn.us/roadway/data/reports/2009/mileage_lanemiles/fzstmccs.xls. Accessed July 23th, 2010

Minnesota DOT. Metropolitan Freeway System 2009 Congestion Report. Metro District Office of Operations and Maintenance.

Morris H. DeGroot, Mark J. Schervish. Probability and Statistics (3rd Edition). Addison Wesley Longman, 2002.

Nezamuddin, N, Crunkleton, Joshua, and Tarnoff, Philip John. Speed Distribution Profile of Traffic Data and Sample Size Estimation. TRB 88th Annual Meeting Compendium of Papers DVD. 2009

NGSIM Community Home http://ngsim-community.org. Accessed Apr 10th, 2011.

N. Geroliminis and A. Skabardonis. Prediction of arrival profiles and queue lengths along signalized arterials, by using a Markov decision process. *Transportation Research Record*, 1934:116-124, 2005.

Oppenlander, J.C. Sample Size Determination for Spot-Speed Studies at Rural, Intermediate, and Urban Locations. In Highway Research Record: Journal of the Highway Research Board, No. 35, HRB, Washington, D.C., 1963, pp. 78–80.

R. Herring, A. Hofleitner, S. Amin, T. Nasr, A. Khalek, P. Abbeel, A. Bayen. Using Mobile Phones to Forecast Arterial Traffic Through Statistical Learning. *Transportation Research Board 89th Annual Meeting*, Washington D.C., January 10-14, 2010

Sheldon Ross. Introduction to probability and statistics for engineers and scientists. Third Edition. Elsevier Academic Press. 2004.

Strobel, H. Traffic Control Systems Analysis by Means of Dynamic State and Input-Output Models. A-2361. Laxenburg, Austria: International Institute for Applied Systems Analysis. 1977

T. Hunter, R. Herring, P. Abbeel, A. Bayen, Path and travel time inference from GPS probe vehicle data. *Neural Information Processing Systems foundation (NIPS)*, Vancouver, Canada, December 2009

T.W. Anderson, L.A. Goodman. Statistical inference about Markov chains. Ann. Math. Stat., Vol 28 (1957), pp. 89-109

Transportation Research Record, Special Report 209: Highway Capacity Manual. National Research Council, Washington, D.C., 2000.

Usami, T., K. Ikenoue, and T. Miyasako. Travel Time Prediction Algorithm and Signal Operation at Critical Intersections for Controlling Travel Time. In *Second International Conference on Road Traffic Control, Institute of Electrical and Electronics Engineers*. pp. 205-208. 1986

W. N. Venables, D. M. Smith and the R Development Core Team, An Introduction to R. 2005

Wei-Hua Lin, Amit Kulkarni, Pitu Mirchandani. Arterial travel time estimation for advanced traveler information systems. Transportation Research Board 2003 Annual Meeting

Young, C.P. A Relationship Between Vehicle Detector Occupancy and Delay at Signal-controlled Junctions. *Traffic Engineering and Control*, Vol. 29, pp. 131-134. 1988

Zhang, H.M. Wu, T.Q. et.al. Arterial Travel Time Estimation Using Loop Detector Data. Interim Technical Report to Minnesota Department of Transportation, April 1997

Appendix A Travel Time Histograms of NGSIM Peachtree Street Dataset

Refer to Figure 2.1 to find the location of each origin. The colors of the origins are defined as:

1	2	3	4	5	6	7	8	9	10	11
114	115	113	112	121	106	122	103	123	102	101
black	blue	green	cyan	red	pink	brown	yellow	grey	drab	olivine

The figure on the left is the histogram of all vehicles and the figure on the right is the histogram of vehicles going through at downstream intersection.

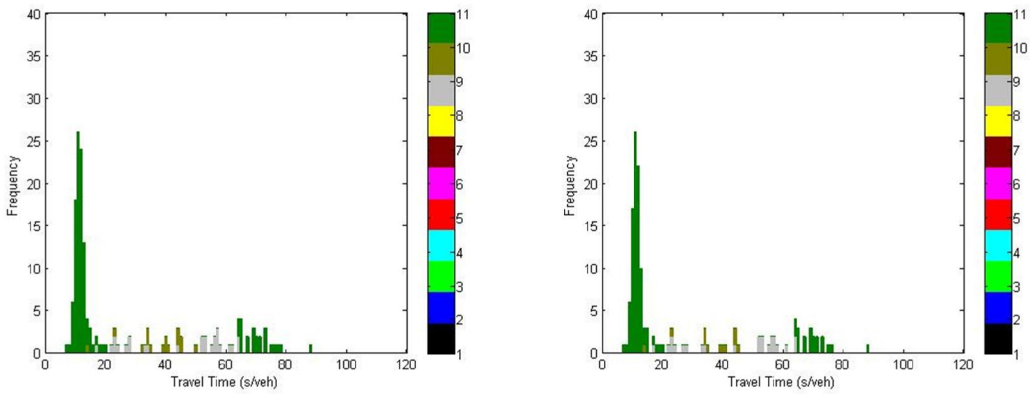


Figure A.1 Section 2 Northbound 12:45-1:00

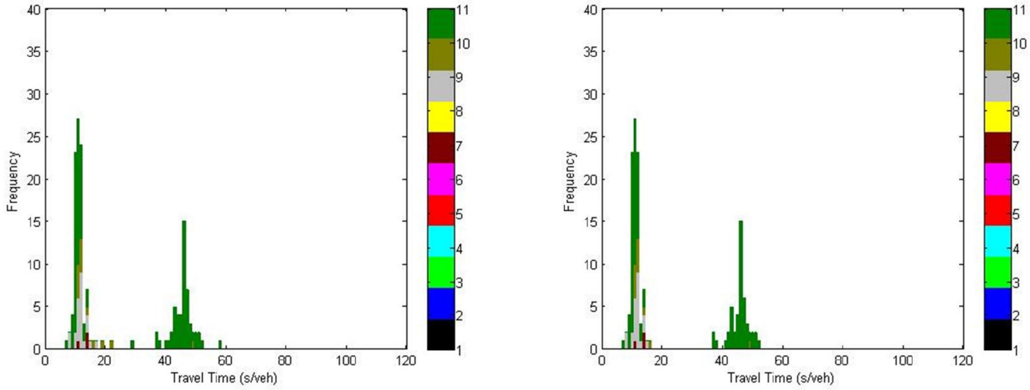


Figure A.2 Section 3 Northbound 12:45-1:00

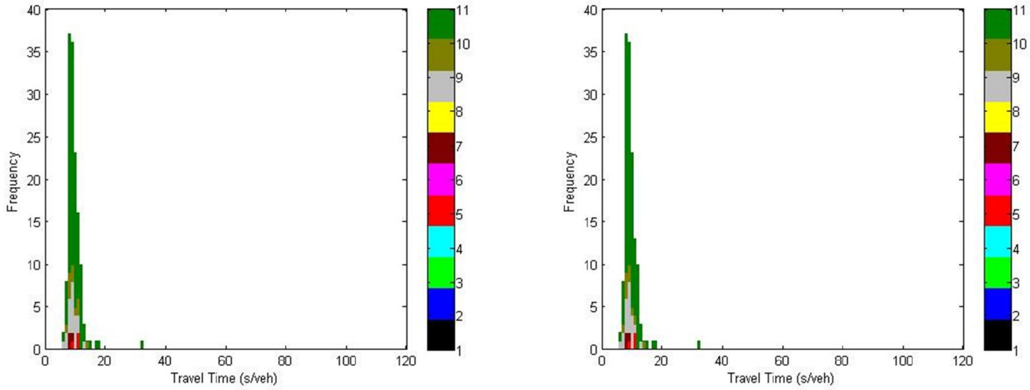


Figure A.3 Section 4 Northbound 12:45-1:00

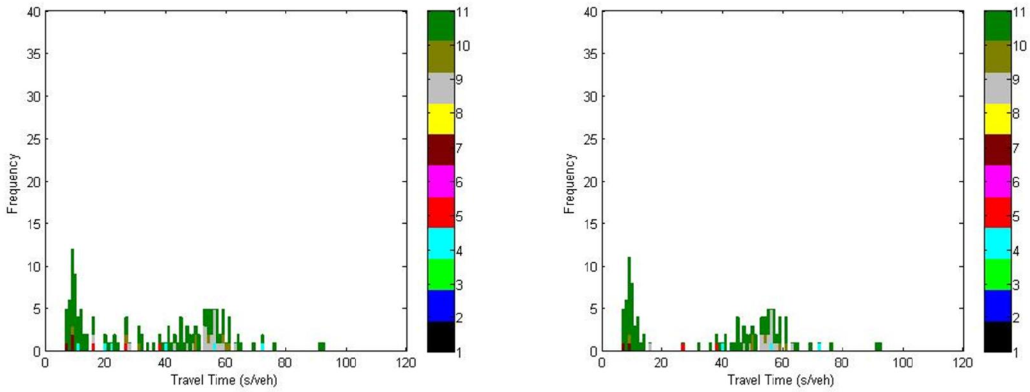


Figure A.4 Section 5 Northbound 12:45-1:00

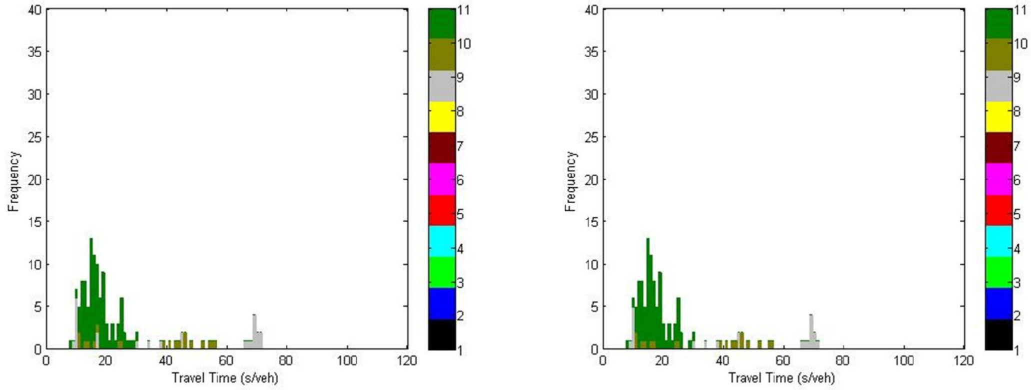


Figure A.5 Section 2 Northbound 4:00-4:15

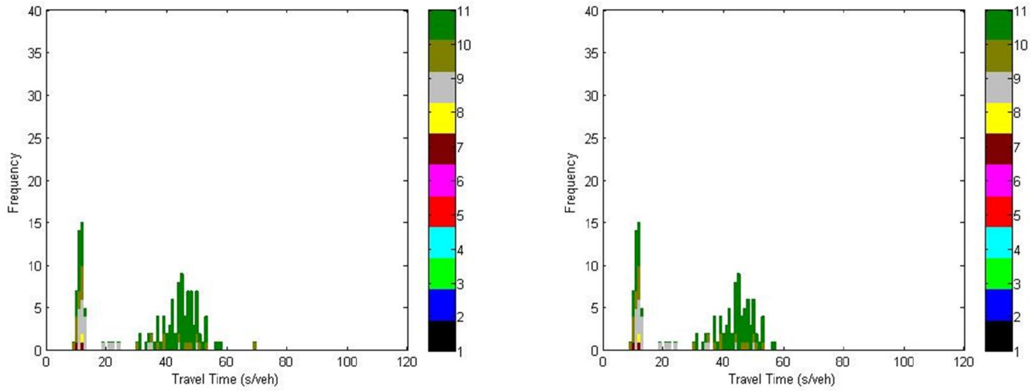


Figure A.6 Section 3 Northbound 4:00-4:15

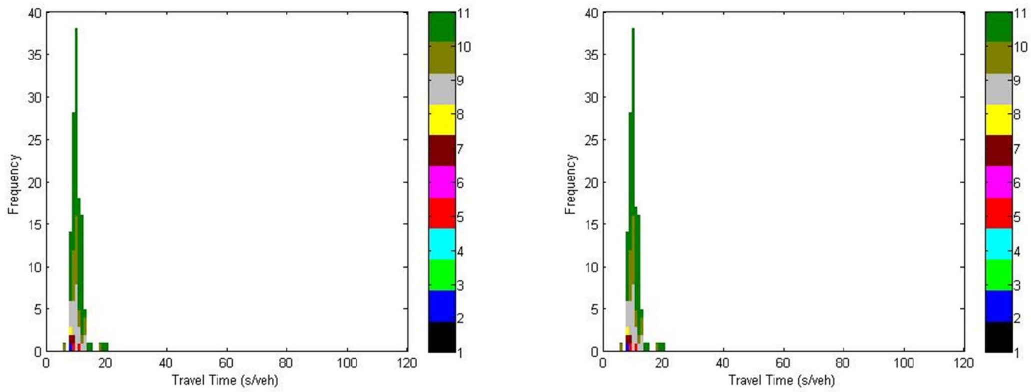


Figure A.7 Section 4 Northbound 4:00-4:15

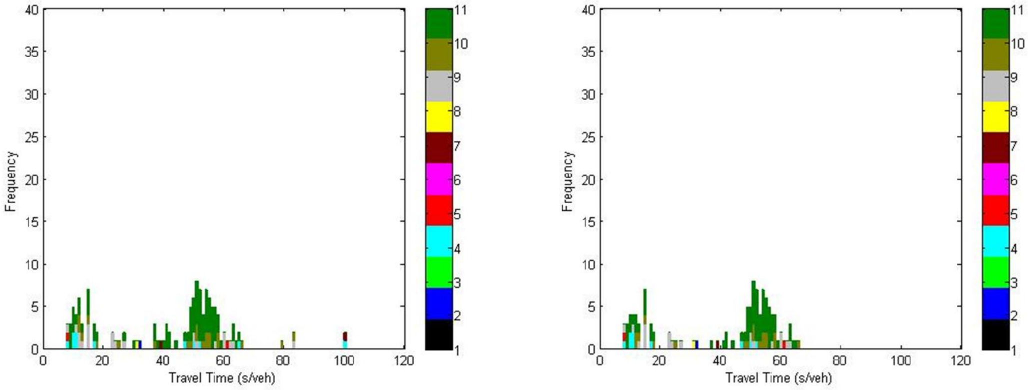


Figure A.8 Section 5 Northbound 4:00-4:15

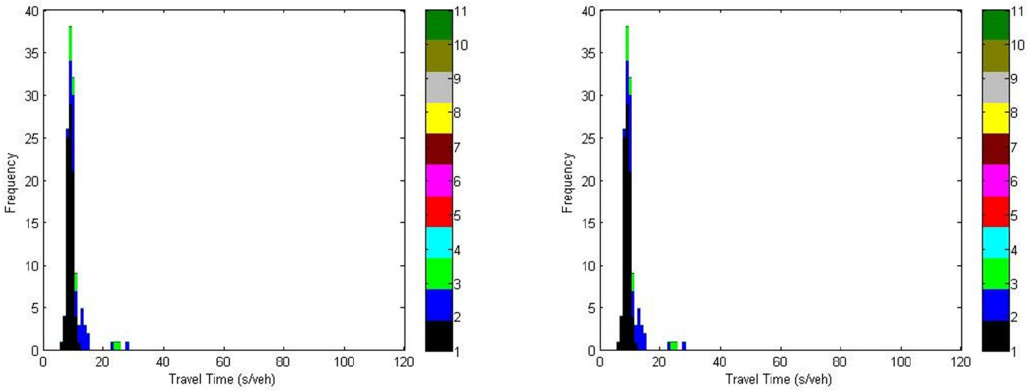


Figure A.9 Section 5 Southbound 12:45-1:00

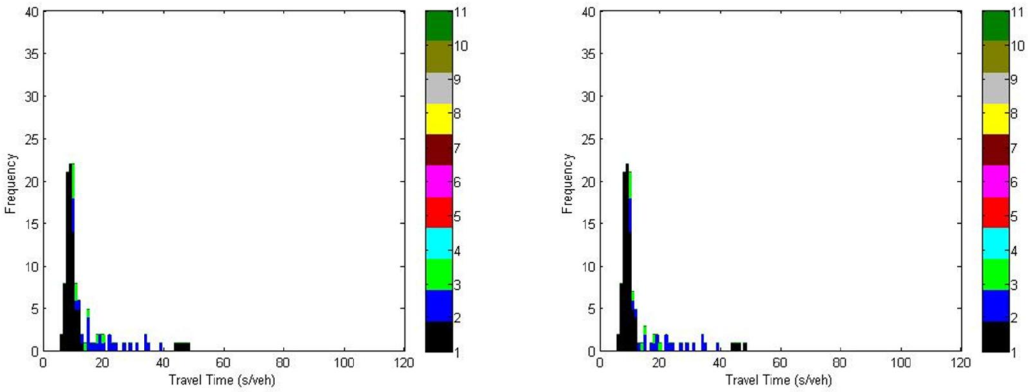


Figure A.10 Section 4 Southbound 12:45-1:00

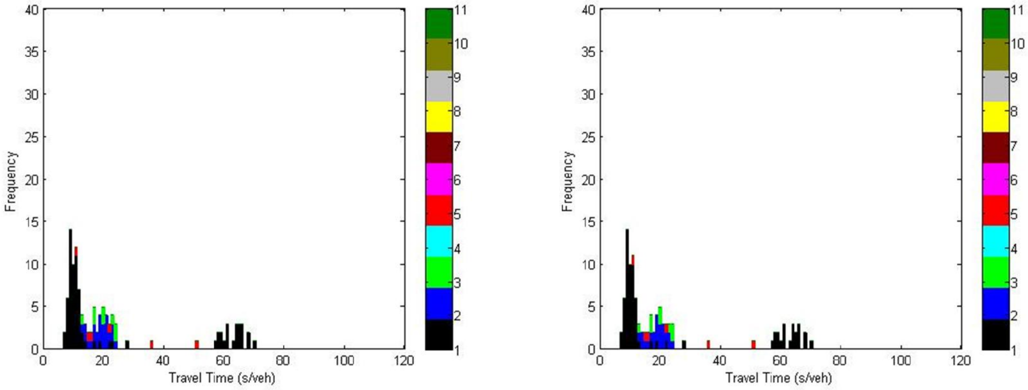


Figure A.11 Section 3 Southbound 12:45-1:00

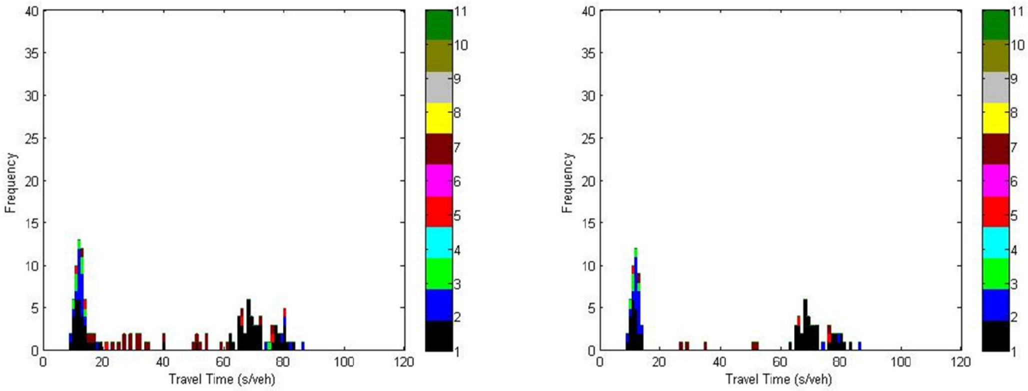


Figure A.12 Section 2 Southbound 12:45-1:00

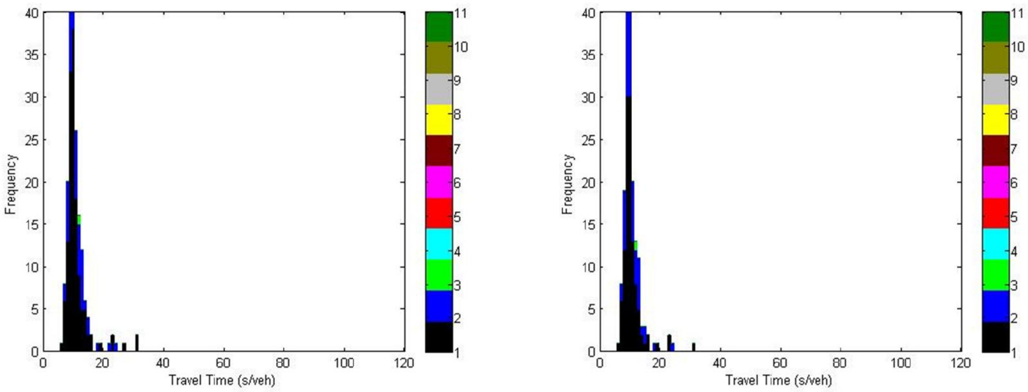


Figure A.13 Section 5 Southbound 4:00-4:15

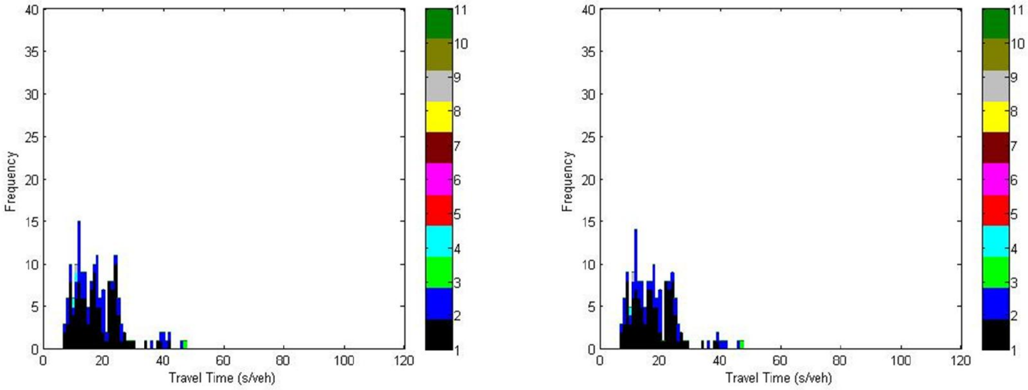


Figure A.14 Section 4 Southbound 4:00-4:15

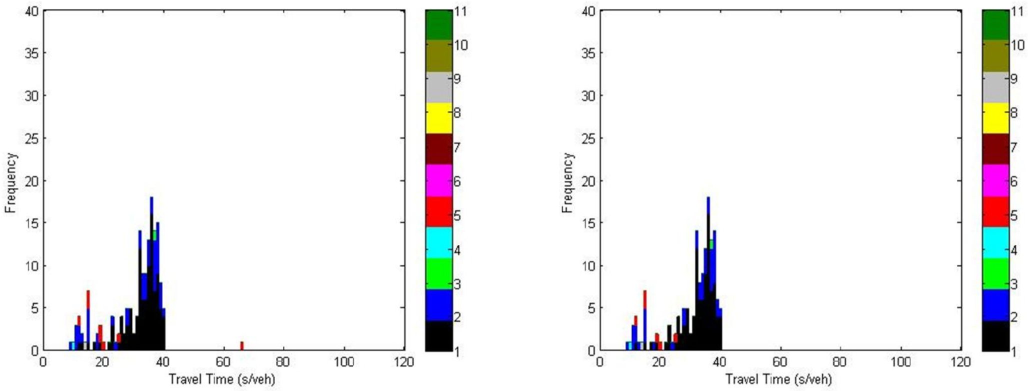


Figure A.15 Section 3 Southbound 4:00-4:15

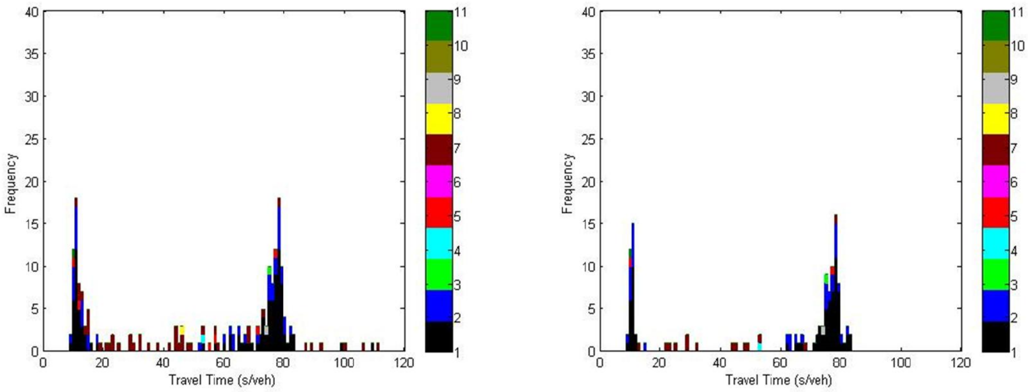


Figure A.16 Section 2 Southbound 4:00-4:15