2018

Evaluation and Augmentation of Traffic Data from Private Sector and Bluetooth Detection System on Arterials

Yaobang Gong
University of Central Florida

EVALUATION AND AUGMENTATION OF TRAFFIC DATA FROM PRIVATE SECTOR AND BLUETOOTH DETECTION SYSTEM ON ARTERIALS

by

YAOBANG GONG
B.S. Central South University, 2016

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil, Environmental and Construction Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2018

Major Professor: Mohamed Abdel-Aty

ABSTRACT

Traffic data are essential for public agencies to monitor the traffic condition of the roadway network in real-time. Recently, public agencies have implemented Bluetooth Detection Systems (BDS) on arterials to collect traffic data and purchased data directly from private sector vendors. However, the quality and reliability of the aforementioned two data sources are subject to rigorous evaluation. The thesis presents a study utilizing high-resolution GPS trajectories to evaluate data from HERE, one of the private sector data vendors, and BDS of arterial corridors in Orlando, Florida. The results showed that the accuracy and reliability of BDS data are better than private sector data, which might be credited to a better presentation of the bimodal traffic flow pattern on signalized arterials. In addition, another preliminary study aiming at improving the quality of private sector data was also demonstrated. Information about bimodal traffic flow extracted by a finite mixture model from historical BDS is employed to augment real-time private sector data by a Bayesian inference framework. The evaluation of the augmented data showed that the augmentation framework is effective for the most part of the studied corridor except for segments highly influenced by traffic from or to the expressway ramps.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisor Professor Mohamed Abdel-Aty, for the continuous guidance and support of my master's study and other research, for his patience, motivation and immense knowledge. I could not have imagined having a better advisor!

Secondly, I would like to appreciate the rest of my respectable committee members: Dr. Samiul Hasan and Dr. Qing Cai, for their invaluable comments.

I would like to thank my friends and my colleagues. I would also like to acknowledge Florida Department of Transportation for funding this research and providing data.

Finally, I would like to thank my parents. You gave me bravery to pursue my dream here, ten thousand miles away from home.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Background

Traffic data is essential for traffic agencies in various aspects including traffic system planning and finance, management and control, monitoring and performance evaluation. Especially, the Active Traffic Management (ATM) needs traffic data feed in real time. Various real-time traffic data, such as those collected by Microwave Vehicle Detection System (MVDS) and Automatic Vehicle Identification (AVI), are largely available for Interstate freeways and other limit-access expressways. Yet historically there were limited traffic data of arterials.

In recent years, traffic agencies have begun to implement Bluetooth detection systems (BDS) to collect travel time data on arterials in real time. BDS employed Bluetooth technology to detect the travel time and space mean speed of vehicles carried with a Bluetooth device (e.g. smartphones and hand-free devices). The advantages of BDS are its relatively low-installation and maintenance cost (Singer, Robinson, Krueger, Atkinson, & Myers, 2013) and acceptable data quality (Bhaskar & Chung, 2013; Haghani, Hamedi, Sadabadi, Young, & Tarnoff, 2010).

Meanwhile, traffic agencies have also purchased real-time traffic data directly from private sector traffic service companies such as INRIX, HERE and TomTom as supplementary data sources. These data vendors take advantage of probe vehicle tracking technologies to provide speed and travel time data of the road network including most of the arterials. Therefore, traffic agencies get a more extensive geographic data coverage without installing and maintaining new infrastructure-based traffic detectors, which might be a cost-effective approach.

1.2 Motivations and Objectives

Data from BDS and private sectors provide opportunities to traffic agencies to monitor the traffic condition on arterials in real-time. And then various traffic management strategies could be implemented to release certain problems of arterial network. Therefore, in order to represent the network condition correctly, the validity and the accuracy of such traffic are subject to rigorous evaluation.

In terms of the quality of data from private sectors, there is often a concern that the exact data source and processing algorithms are always proprietary. (Elefteriadou, Kondyli, & George, 2014). As a consequence, there are a lot of studies aiming at evaluating its quality by different kinds of benchmarks. However, most of those studies only validated the data of limit-access facilities. Since the nature of traffic flow on arterials, which are heavily interrupted by intersections, is totally different from that on the limit-access facilities. Hence, it is unreasonable to simply transfer the evaluation results of limit-access facilities. Therefore, it is needed to conduct a comprehensive investigation on the quality of traffic data from private sectors on arterials.

As mentioned in the last section, the quality of data from BDS is acceptable. In fact, traffic data from BDS were frequently used as the benchmark "ground truth" data to evaluate private sector data (Hu, Fontaine, & Ma, 2016; Young et al., 2015; Zhang, Hamedi, & Haghani, 2015). However, several studies have pointed out several systematic errors of BDS, ranging from the detection of the Bluetooth device to the data filtering and cleaning. Thus, carefully evaluating and processing traffic data from BDS is also important.

Furthermore, since traffic data from different sources might have their inherent biases, augmentation of data quality is needed to better represent the actual traffic condition, especially for the data from private sectors which might have large biases. Hence, a data augmentation framework of the data from private sector is also valuable.

In short, there are two major objectives of this thesis: firstly, thoroughly evaluating of traffic data from both private sector and BDS; secondly, proposing a preliminary augmentation framework of the data from private sector.

1.3 Study Methodology

The thesis includes two studies: the evaluation of quality of data from both private sectors and BDS and the augmentation of data from the private sector.

For the data evaluation, space mean speed data from two different BDSs and one private sector, HERE, of arterial corridors in Orlando, Florida are evaluated. Field-collected high-resolution GPS trajectories as utilized as the benchmark "ground truth". Two metrics which are used by various studies, namely Average Absolute Speed Error (AASE) and Speed Error Bias (SEB), are employed to measure the deviation of a specific data source from the ground truth. When the deviation is smaller than a certain level, the data from that certain source is considered as accuracy.

Then an augmentation framework based on Bayesian inference is proposed to enhance the quality of the private sector data. The framework is able to extract the bimodal traffic flow

pattern on signalized arterials using a finite mixture model. And then the data from private sector is enhanced by such information.

1.4 Thesis Structure

The rest of the thesis is organized as follow: Chapter 2 provides a brief review of early studies; Chapter 3 describes the study sites and data preparation process; Chapter 4 elaborates the methodology, the results of the data evaluation and findings from the results; Chapter 5 reports the methodology and the results of the preliminary augmentation framework. Finally, a summary is presented in Chapter 6.

CHAPTER 2: LITERATURE REVIEW

2.1 Evaluation of Data from Private Sectors

The interrupted traffic flow leads to a more challenging environment for probe vehicles to obtain reliable traffic data (Hu et al., 2016; Zhang et al., 2015). Several early studies have used different kinds of benchmark ground truth data to evaluate the accuracy and reliability of travel time data from private sectors on arterials.

An early study conducted by Wang et al. (Y. Wang, Araghi, Malinovskiy, Corey, & Cheng, 2014) evaluates the speed data from one of the private sectors, INRIX, of only one arterial corridor in the State of Washington. The data from automatic license plate reader (ALPR) system were used as the benchmark ground truth. After comparing with the ground truth, they found that data from INRIX was less responsive to traffic changes and tended to have a systematic bias. This is probably due to its conservative estimation of free flow travel time. However, data from only three segments of one arterial corridor is used in this study.

Elefteriadou et al. (Elefteriadou et al., 2014) conducted a study aiming at evaluating travel time (space mean speed) from two different private sectors, INRIX and HERE. An instrumented vehicle equipped with a GPS device was used to collect the benchmark ground truth speed data. The results show that data from neither source is accurate. However, the authors admitted that the sample size used in the study was relatively small, especially during the oversaturated runs.

Zhang et al. (Zhang et al., 2015) utilized travel time data from BDS to validate INRIX travel time data of 2.8-mile arterial segment. They found that the INRIX data have a relatively larger deviation from BDS data for the entire daytime on weekends and peak period on weekdays. In this study, BDS is assumed as the "ground truth", however, as mentioned in the last chapter, data from BDS itself is subject to rigorous validation.

Different from aforementioned studies which only focus on the accuracy of data in real-time, Hu et al. (Hu et al., 2016) evaluated both the ability of private sector data to track real-time conditions and to identify long-term traffic state changes. Similar to the study conducted by Zhang et al. (Zhang et al., 2015), data from BDS are employed as the ground truth as well. They concluded that private sector data were not suitable to provide real-time information since they were not able to capture speed reductions and they tended to underestimate the variability of travel time. Yet, they may be able to show the changes of traffic state in long-term

Beginning in 2013, Vehicle Probe Project (VPP) validation team developed a comprehensive analysis of data quality from the private sector on signalized arterials (Young et al., 2015). Similar to the previous studies, they also found that data from the private sector failed to detect speed reductions during congested periods. In addition, complex traffic flow patterns on signalized arterials such as bimodal flow could not to be observed from the private sector's data. As a consequence, they recommend the use of private sector data when there is only one signalized intersection per mile or less on an arterial corridor.

In summary, the aforementioned studies came up with a conclusion that there is highly likely a systemic bias of traffic data on arterials from the private sector's data, but data from

private sectors are reliable under certain circumstances. Agencies are not recommended to utilize the traffic data on arterials from the private sector without a rigorous site-specific evaluation. However, studies (Hu et al., 2016; Young et al., 2015) also pointed out that the data quality might be improved by subsequent enhancement of data stream and processing algorithm.

2.2 Systematic Biases of data from BDS

Although traffic data from BDS were frequently used as the benchmark "ground truth" data to evaluate private sector data (Hu et al., 2016; Young et al., 2015; Zhang et al., 2015), researchers have pointed out several systematic biases and proposed possible approaches to eliminate them.

BDS employed Bluetooth detectors which utilize Bluetooth communication technology to detect vehicle carrying a "discoverable" Bluetooth device within their large detection zone. Theoretically, it takes up to 10.24s for a Bluetooth device to be "discovered". This communication delay might lead to some error especially when the travel time is relatively short (Bhaskar & Chung, 2013; José, Díaz, Belén, González, & Wilby, 2015).

Another issue is called "multiple detection". Typically, in order to save budget, ensure the power supply and increase detection rate, on the arterials, Bluetooth detectors are installed within or close to the signal cabinet near the intersection. Thus, the detectors are typically configured to be able to detect all the "discoverable" vehicles within the intersection. Therefore, the detection range is typically 300 to 400 feet. Therefore, a single vehicle could be detected by a specific detector more than once within a short period of time due to this large detection range.

This leads to a location ambiguity and increases the error of travel time estimation (Araghi, Christensen, Krisnan, Olesen, & Lahrman, 2013; José et al., 2015). In order to overcome this issue, Araghi et al. (Araghi et al., 2013) suggested using the detection with highest Bluetooth signal strength among "all the multiple detections".

Additionally, researchers and practitioners found that there are a number of outliers in raw BDS travel time data caused by transportation modes other than car, such as public transit and non-motorized modes. (Bhaskar & Chung, 2013). Vehicles to or from drive ways might also cause extremely long travel time. Filtering algorithms such as Moving Median filter, Median Absolute Deviation (MAD) filter, Box-and-Whisker filter etc. (Bhaskar & Chung, 2013), were suggested to filter out those outliers.

It is worth noting that the penetration rate of BDS data is very critical for the data reliability. Previous researchers stated that the overall penetration rate of BDS data are higher than 3%, which is sufficient for a 95% confidence level in travel time and speed estimates (Yuan et al., 2018; Yuan, Abdel-Aty, Gong, & Cai, 2019; Yuan & Abdel-aty, 2018)

In short, due to those aforementioned systematic biases, carefully evaluating and pre-processing traffic data from BDS is also important.

2.3 Augmentation of Traffic Data

Since traffic data from different data sources have their inherent biases, augmentation of data quality is necessary to provide better representation of the real time condition of the traffic

network. A number of researchers have employed data fusion techniques to combine traffic data from different sources for higher accuracy and resolution.

Most of the early studies fused traffic data from inductive loop detectors (ILD) and GPS probe vehicles on arterials. A lot of methodologies were employed such as, weighted average (Zheng, Ma, Wu, Wang, & Cai, 2014; Zhu, Guo, Polak, & Krishnan, 2017), fuzzy regression and Bayesian pooling (Choi & Chung, 2002), iterative Bayesian estimation (Liu, Cui, Cao, & Wang, 2016), artificial neural networks (ANN) (Zhu et al., 2017) and traffic flow model based approaches (Bhaskar, Chung, & Dumont, 2011; Cai, Wang, Zheng, Wu, & Wang, 2014; Nantes, Ngoduy, Bhaskar, Miska, & Chung, 2016; Z. Wang, Cai, Wu, Zheng, & Wang, 2016). Zhu et al. (Zhu et al., 2017) even further fused ILD data and GPS probe data with traffic data from mobile phones.

Bhaskar et al. (Bhaskar, Tsubota, Kieu, & Chung, 2014) proposed a traffic flow based model to integrate speed and density data from ILD and travel time from BDS to conduct traffic flow estimation. The model is validated by both real and simulated benchmark data.

To best of the author's knowledge, only Zhang et al. (Zhang, Hamedi, & Haghani, 2017) attempted to fuse traffic data from private sector with other data sources. They proposed a context-dependent fusion framework to fuse data from INRIX and BDS to enhance the data quality. However, they did not conduct the quantitative evaluation for the fusion framework due to the lack of "ground truth" data.

2.4 Summary

The literature review clearly showed that:

1) In the context of the arterials, the data from private sectors might have some biases but it might be accurate under certain circumstances. A site-specific evaluation is necessary.

2) There are a lot of potential biases of data from BDS. A rigorous evaluation of its accuracy should be conducted.

3) There is few study about of the augmentation of data from private sectors. Thus, a quantitative data augmentation framework is valuable.

CHAPTER 3: STUDY SITE AND DATA PREPARATION

3.1 Study Site

Two arterial corridors in Orlando, Florida are selected to evaluate the accuracy of travel time (space mean speed) data from private sector and BDS. The first one is a 5.4-mile segment of Alafaya Trail, which is a principal arterial corridor. The segment runs northbound from McCulloch Road to Curry Ford Road. And its posted speed limit of the segment is 45 mph. The map of the corridor segment is presented in Figure 1. The vendor of the BDS implemented on the corridor is BlueMAC.

The other corridor is a 1.7-mile segment of an arterial corridor, Mitchell Hammock Rd. The segment runs westbound from Alafaya Trail to Lookwood Blvd with the posted speed limit of 45 mph. Figure 2 illustrates the map of the segment. Different from Alafaya Trail, the vendor of the BDS installed here is BlueTOAD. Although both vendors provides traffic time data, their format and granularity are different. The details will be illustrated in the section 3.2 and section 3.3.

3.2 Data Collection

Although data from both source are travel time and/or equivalent space mean speed of a specific roadway segment, private sector company, HERE Technologies, provides one-minute aggregated space mean speed of their probe vehicles at Traffic Message Channel (TMC) codes.

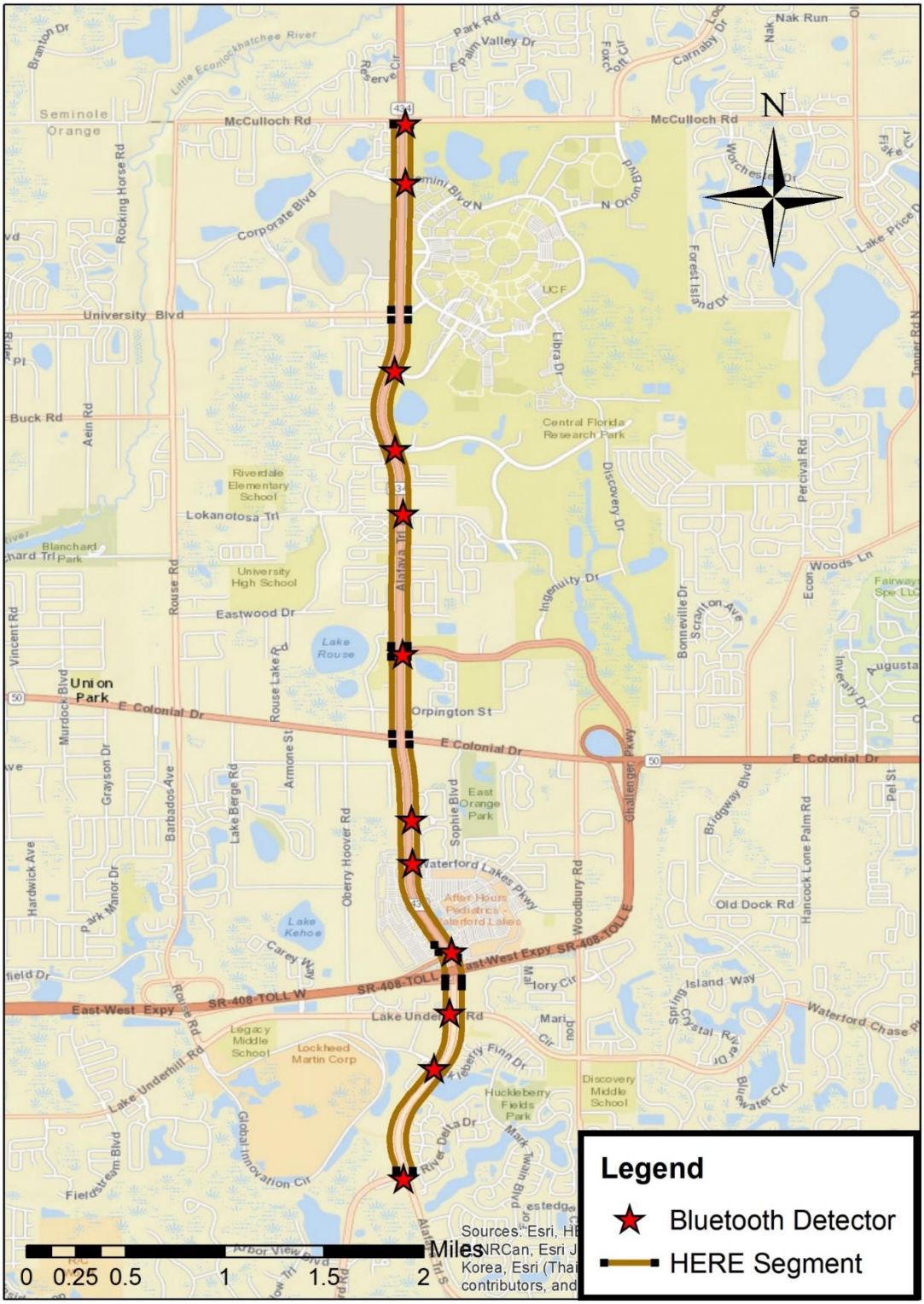

Figure 1 Here Links and Bluetooth Detectors of Alafaya Trail

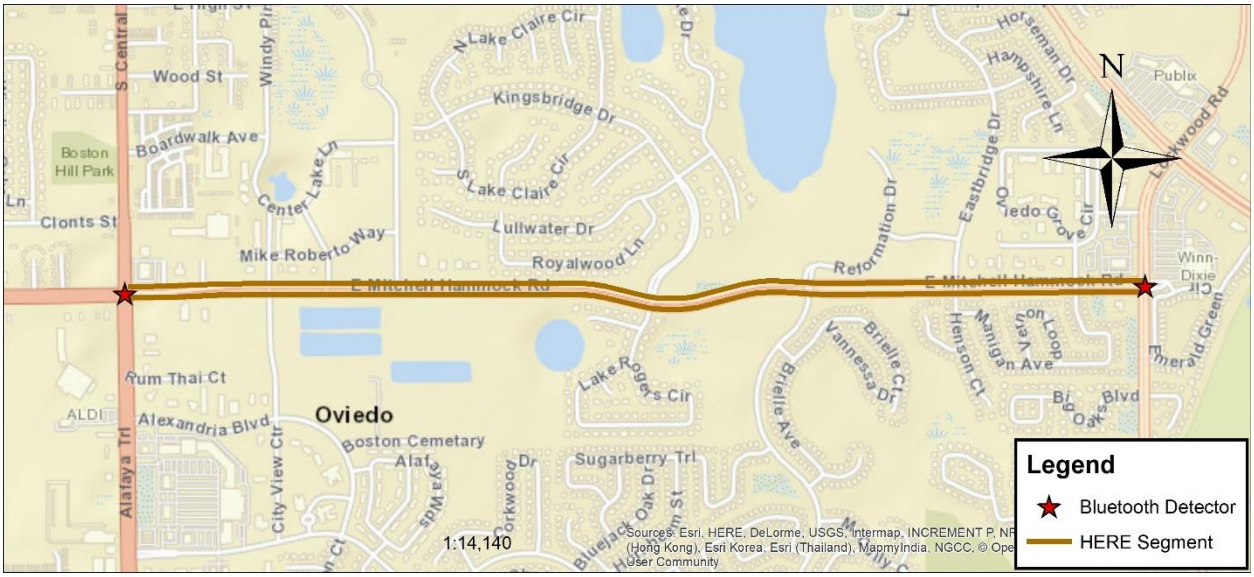

Figure 2 Here Links and Bluetooth Detectors of Mitchell Hammock Road

TMC refers to a widely used standard roadway segment referencing system. On arterials, TMC segments break primarily at major signalized intersections, whereas minor intersections are contained within a single TMC segment (Young et al., 2015). Take the Alafaya Trail as an example. Figure 1 shows the TMC segments (HERE Segments) of the study corridor. It is noted that some TMC segments refer to the isolated intersections, whose length is too short (about 50 feet) for segment travel time to be estimated by other data sources. Thus, these segments were excluded in this study and not displayed in the Figure 1 and Figure 2. Hence, data from totally fourteen TMC segments for both directions of both corridors were utilized in this study.

Data from different BDS have different formats. For the Alayrafa Trail, BlueMAC detectors implemented by the Orange County Traffic Engineering provides raw detection readings, which are the unique MAC address of the Bluetooth device carried by the vehicle and corresponding time stamps. Then the raw detections are further processed by the author to get the travel time and space mean speed. Other than BlueMAC, BlueTOAD systems implemented in

Mitchell Hammock Road does not have the functionality to provide raw detection readings. Instead, the BlueTOAD system is only able to provide processed travel time of individual vehicles. Then the details about data processing will be elaborated in Section 3.3. The locations of Bluetooth detectors are presented in Figure 1 and Figure 2. There are 12 BlueMAC detectors monitoring traffic in both directions, which are placed on the top of signal cabinets. And there are 2 BlueTOAD detectors, which are places close to the signal cabinet.

High-resolution trajectory data collected by GPS devices was utilized as the "Ground truth" benchmark to evaluate the aforementioned two data sources. Six trained volunteers were asked to record their trajectories when they drove along the whole or part of the corridors by smart phone GPS tracking applications. The applications are able to record accurate geographical coordinates of the vehicle every one to two seconds, which provide a high-resolution trajectory data. The trajectories are used to calculate the ground truth travel time and space mean speed of the roadway segments.

HERE data, BDS data and ground truth trajectory data were collected from April 2017 to June 2017.

3.3 BDS Data Processing and Filtering

In terms of data from BlueMAC, travel time and space mean speed of an individual vehicle from upstream Bluetooth detector *a* to downstream detector *b* is obtained by matching the raw detections, which are the unique MAC address of the Bluetooth device carried by the vehicles and corresponding time stamps.

$$TT(i,a,b) = T(i,b) - T(i,a) \quad (1)$$

$$v(i,a.b) = \frac{D(a,b)}{TT(i,a,b)} \quad (2)$$

where $TT(i,a,b)$ is the travel time from detector *a* to detector *b* of vehicle *i*; $v(i,a,b)$ is the space mean speed of the vehicle *i* traveling along the corresponding segment; $T(i,k)$ is the time stamp when the vehicle *i* is observed at the detector *k*; $D(a,b)$ is the length of roadway segment between detector *a* and detector *b*.

Due to the multiple detections issue, as suggested by Araghi et al. (Araghi et al., 2013), in this study, the detection with the highest Received Signal Strength Indicator (RSSI) which indicates the closest location to the Bluetooth detector of the vehicle was selected as the $T(i,k)$.

The travel times and space mean speeds calculated by the aforementioned method and those directly from BlueTOAD system have some noise as discussed in the Chapter 2. Thus, a Moving Median Absolute Deviation (MAD) filtering algorithm was adopted to eliminate the noises. First, the median of all space mean speed readings within a nine-minute moving window is calculated (Bhaskar, Qu, & Chung, 2015). And the upper bound value (UBV) and lower bound value (LBV) are defined as

$$v_{ub} = Med + \hat{\sigma}f \quad (3)$$

$$v_{lb} = Med - \hat{\sigma}f \quad (4)$$

$$\hat{\sigma} = 1.4826 * MAD \quad (5)$$

$$MAD = median(|v_{kj} - median(v_{ij})|) \quad (6)$$

where v_{ub} and v_{lb} is the UBV and LBV; Med is the median space mean speed of the speed observations among the moving time window; MAD is the median absolute deviation from the Med; $\hat{\sigma}$ is the standard deviation from MAD. f value was decided considering the variance level of speed data. A smaller f gives a higher confidence of the estimation, yet several valid data might be filtered out if the variance is relatively large. In this study, 1.5 is selected as the f value.

3.4 Spatial-Temporal Alignment

Evaluating traffic data from one data source by another requires spatial and temporal alignment of both data sources when they have different segmentation convention and temporal aggregation levels.

Both HERE and BDS traffic data are reported at the segment level, however, according to Figure 1 and 2, the segmentations of HERE data and BDS data are not always consistent with each other. Figure 3 illustrates the methodology of spatial alignment. In this study, the space mean speed from BDS corresponding to a single HERE segment was estimated by a weighted average value by length of the speed of the same time period from all BDS segments fully and partially within the HERE segment. For example, the space mean speed of TMC1 from BDS was estimated as:

$$v_1 = v_{ab}\frac{L_{ab}}{L_1} + v_{bc}\frac{L_{bc}}{L_1} + v_{cd}\frac{L_{cd1}}{L_1} \quad (7)$$

where v_1 is the space mean speed from BDS of segment TMC1; v_{ij} is the space mean speed of the roadway segment between Bluetooth detectors i and j; L_1 is the length of segment TMC1. Note that L_{cd1} is not the distance between detectors c and d. It is the length of the overlaying part of TMC1 and segment between Bluetooth detectors c and d.

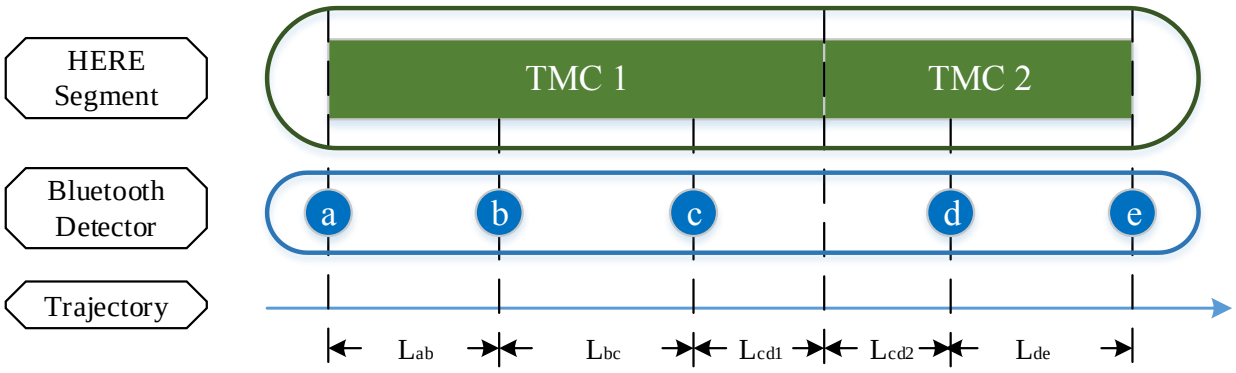

Figure 3 Spatial Matching of Three Data Sources

The travel time and space mean speed from the "ground truth" benchmark dataset, the data from the trajectories of probe vehicles, in HERE segment level were calculated directly by the similar method as BDS travel time estimation (Functions (1) and (2)).

The space mean speed, which represents the travel time, is reported in one-minute aggregation level for HERE data and in individual vehicle level for BDS and ground truth data. Therefore, data from three sources need to be aggregated in the same temporal aggregation level. Thus, individual space mean speed observations in a five-minute time interval were aggregated to generate the average speed corresponding to that time interval. HERE data were aggregated based on its reported time stamp, whereas BDS and GPS data were aggregated based on the time stamp when the individual vehicle finished the trip ($T(i,b)$), which also indicated the "reported"

time stamps. Then the average space mean speeds from the three data sources were matched with each other to get the master dataset.

After the spatial-temporal alignment, there are roughly 30 GPS ground truth space mean speed readings for each HERE segment.

3.5 Summary

This chapter describes the study sites used for the data evaluation and further augmentation and the detailed preparation process of three data sources. Figure 4 summarized the whole data preparation process.

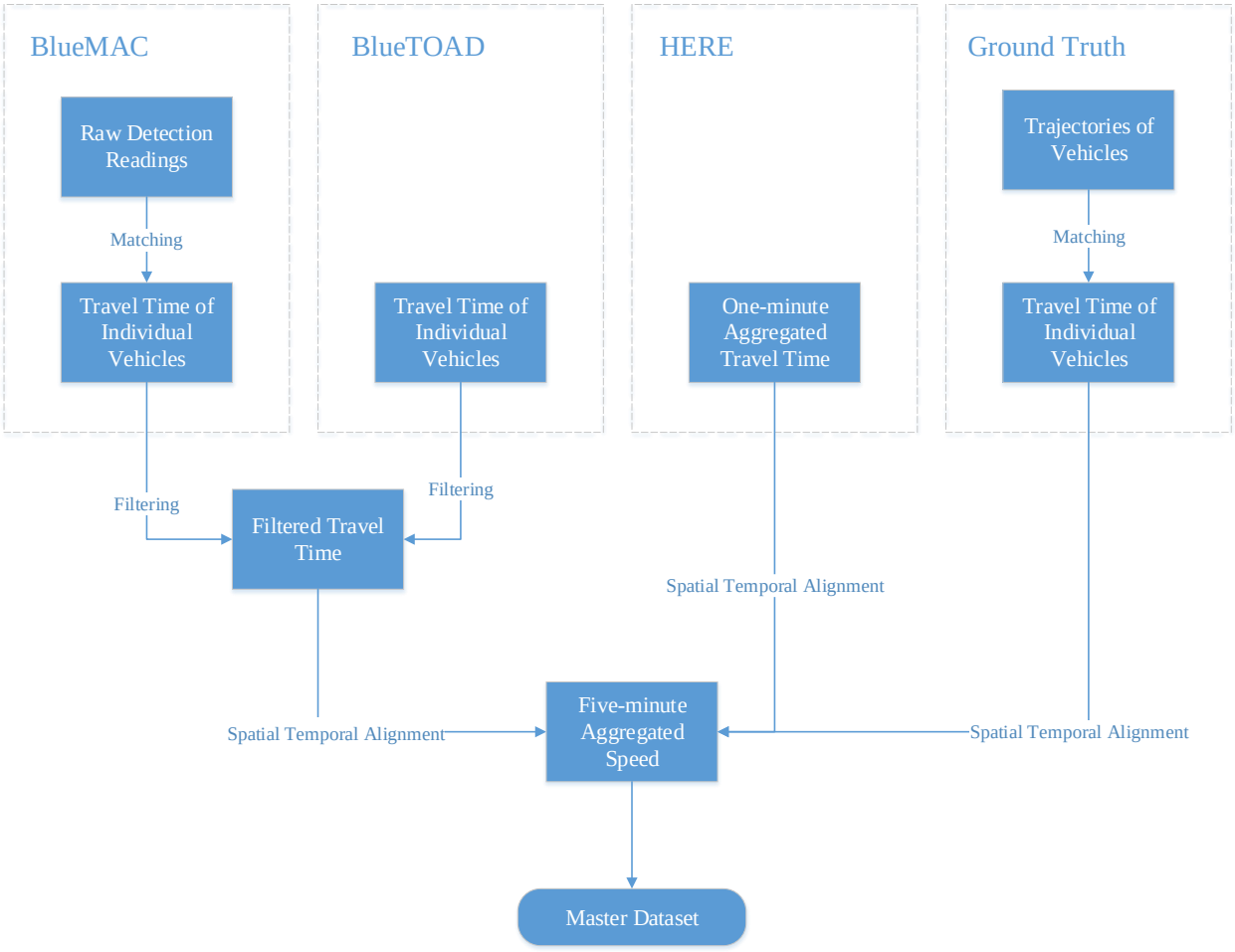

Figure 4 The Process of the Data Preparation

CHAPTER 4: DATA EVALUATION

4.1 Methodology

The basic logic of data evaluation in this study is to assess whether a specific data source is able to provide credible travel time (space mean speed) information to individual travelers. Thus, the average speed of five-minute intervals from HERE and BDS were compared with the corresponding ground truth data, which is the space mean speed (travel time) of the "actual" trips. Although the size of the validation data sets is not large, it is reasonable to represent the travel times of the actual trips

Two different metrics were utilized as the quantitative measurement of the deviation of a specific data source from the ground truth. One is the Average Absolute Speed Error (AASE), which reflects the absolute difference and the other is Speed Error Bias (SEB), which is aiming to examine the direction of the difference.

AASE is calculated by function (8).

$$AASE = \frac{1}{n}\sum_{j=1}^{n}\left|v_{ki} - v_{gi}\right| \tag{8}$$

where n is the number of speed observations of a specific HERE segment; v_{ki} is the ith speed observation of data source k; v_{gi} is the correspondent speed observation of ground truth data.

SEB is calculated by function (9). These metrics show the existence and magnitude of a consistent positive or negative deviation. The notations have the same meanings as function (8).

$$SEB = \frac{1}{n}\sum_{j=1}^{n}(v_{ki} - v_{gi}) \quad (9)$$

4.2 Results

Table 1 shows the AASE and SEB of the two data sources compared with the GPS ground truth for each HERE segment. According to Table 1, in terms of AASE, data from BDS are always closer to the ground truth than data from HERE, which indicates that the BDS data have the better accuracy. The only segment whose AASE of BDS data is greater than 10 mph is segment "102N10190". A possible reason is that this segment is the part of arterial between two ramps of a limited-access expressway exit. The traffic on the segment includes the vehicles from and to the expressway through ramps, which might cause a lot of weavings. A careful examination of the volunteers' trajectories showed that no drivers had ever travelled from or to the expressway through this specific exit. Therefore, the validation data of that specific segment might not represent all kinds of traffic travelling along the segment. Another possible reason is that data from BDS cannot represent traffic pattern due to the same issue.

The signs of SEB shows that there is no consistent positive or negative deviation from the ground truth. Interestingly, for most of the segments (11 out of 14), the speed from HERE is less than the corresponding speed from BDS.

Table 1 Deviation of Speed from HERE and BDS from the Ground Truth

HERE_ID	HERE		BDS	
	AASE(mph)	SEB(mph)	AASE(mph)	SEB(mph)
Alafaya Trail				
102+10187	11.60	2.08	**9.13**	**-2.88**
102+10188	**7.56**	**-1.85**	**6.38**	**-0.01**
102+10189	**6.05**	**0.02**	**1.59**	**-0.82**
102+10190	**7.26**	**-3.74**	**6.55**	**1.86**
102+10191	**9.05**	**-3.32**	**7.37**	**-2.60**
102-10186	9.85	7.04	7.70	-5.07
102-10187	11.57	-8.70	**8.05**	**0.41**
102-10188	10.77	-0.81	**6.96**	**1.85**
102-10189	13.21	-9.89	**8.66**	**3.44**
102-10190	9.41	-7.14	**8.03**	**-1.66**
102N10190	12.82	-11.04	10.70	6.45
102P10190	11.13	6.50	**6.49**	**1.14**
Mitchell Hammock Road				
102+10361	**7.14**	**-0.50**	**5.19**	**2.42**
102-10360	**9.31**	-8.13	**7.69**	**-1.26**
Note: The bold value indicates the data of the segment meets the reliability criteria.				

4.3 Discussion

According to the literature (Young et al., 2015), the reliability of speed data is acceptable if its AASE is less than 10 mph and SEB is less than ±5 mph in each of four speed ranges: the low speed, median speed, high speed and speed above the speed limit. However, the ground truth data set of this study is not large enough to validate each of the aforementioned four speed ranges. Therefore, if only the overall AASE and SEB is considered, the BDS data is reliable for most of the segments (12 out of 14, and the SEB of segment "102-10186" is -5.07, which is slightly higher than the criteria), while the HERE data is reliable for only six segments. As a result, if two data source are both available for a specific arterial corridor, the data from BDS is recommended. Interestingly, the accuracy of the data from HERE on Mitchell Hammock Road is better than Alafaya Trail, especially in terms of AASE. This might credit to the different traffic

flow conditions. For example, the AADT in 2017 of Alafaya Trail is 52500 while that of Mitchell Hammock Road is only 34500, which infers that the Alafaya Trail is more congested than Mitchell Hammock Road. This result also implied that site-specific evaluation of private data is necessary.

However, one of the advantages of the private sector data is its large geographical coverage compared with infrastructure based data like BDS. For instance, in Orlando Metropolitan Area, BDS data are only available for several arterial corridors, whereas HERE data are available for all the principal arterials and most of the minor arterials. Thus, for an area where only HERE data are available, a proper augmentation framework is needed.

A possible augmentation approach is using the known traffic follow pattern. The pattern could be estimated extra by a more reliable data source. Figure 5 illustrates the distributions of raw data of a HERE segment and a BDS segment (Note that the magnitude of percentage is different for better illustration). As shown in Figure 5 and discussed in the introduction, while the BDS data clearly presents the bimodal flow caused by traffic signals, which is a “slower mode” representing vehicle platoons forced to stop and waiting for the green phase and a “faster mode” representing those traveling through the corridor without stopping, HERE data failed to provide that information. Interestingly, the speed from HERE does not always reflect either “faster mode” or “slower mode”, it illustrates a pseudo mode whose expected value of the speed is between that of the aforementioned two modes. This probably due to HERE company utilized proprietary algorithms to adjust the speed observations. Unfortunately, the evaluation results showed that the effectiveness of the adjustment is not optimal on the study corridor.

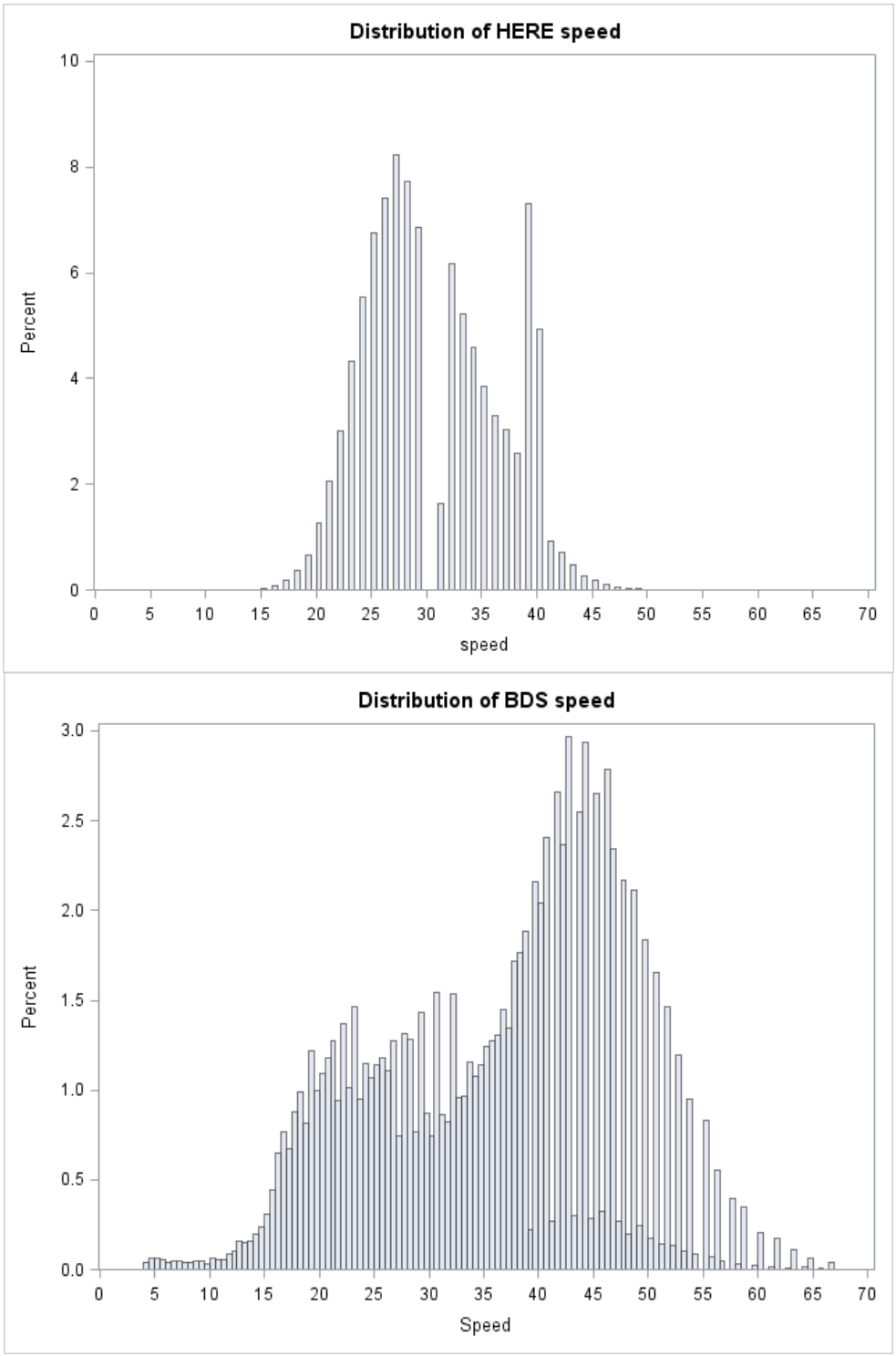

Figure 5 The Distribution of Speed from HERE (up) and BDS (down)

Another issue is that for most of the segments (11 out of 14), the speed from HERE is less than the corresponding speed from BDS. One possible reason is that two data sources have different estimations of compliance of the posted speed limit (or travel time of the free-flow condition). According to the *Florida Statute* section 318.18 (3), there is no fine if a person is cited for exceeding the speed limit by up to 5 mph. (except for a citation in a legally posted school zone or a legally posted enhanced penalty zone, but there is no legally posted school zone or posted enhanced penalty zone in the study corridors) Figure 5 shows that the percentage of the observations whose speed is greater than 50 mph (45+5) is less than 0.01% in the HERE data set, it is reasonable to infer that HERE data optimistically estimate the compliance of the speed limit. Since HERE speed data is collected by probe vehicles, it is very unlikely for a qualified probe vehicle driver to violate the speed limit. Therefore, the HERE data might optimistically estimate the compliance of the posted speed limit of all other vehicles. A better understanding of the compliance of the speed limit would benefit the estimation of the "faster mode". Although the information about the compliance of speed limit is unnecessary for traveller information dissemination, it is still valuable for traffic system planning and evaluation, especially for traffic safety planning and evaluation and the driving behavior studies.

4.4 Summary

This chapters presents the evaluation results of the space mean speed data from HERE and BDS. The evaluation results show that speed data from BDS are more accurate since it is able to provide a better understanding of the bimodal flow on signalized arterials. This bimodal traffic flow pattern could be utilized for augmentation of private sector data. In addition, the

accuracy of data from HERE is different for different corridors, which indicates that a site-specific evaluation of private sector data is needed.

CHAPTER 5: AUGMENTATION OF PRIVATE SECTOR DATA

5.1 Methodology

In order to augment the quality and reliability of the private sector data, HERE, information of the bimodal flow pattern was combined or "fused" with the original private sector data. This information is able to be estimated by traffic flow theory or data-driven approaches. Since the traffic signals on the study corridor are controlled by an adaptive traffic signal control system. Their signal timing is changing in respond to real-time traffic demand. Thus, it is difficult to estimate the traffic flow pattern without knowing the real-time traffic demand and their proprietary controlling algorithm. As a consequence, in this study, the bimodal information is estimated by a data-driven approach, finite mixture model, based on historical BDS data. In addition, since the sample size of data from BDS of Mitchell Hammock Road is not sufficient to estimate the traffic flow pattern. Therefore, the proposed augmentation framework is only tested using data from Alafaya Trail.

5.1.1 Estimation of Bimodal Traffic Flow Pattern by Finite Mixture Model

Assume the distribution of space mean speed of a segment is the summation of two normal distributions: a distribution of "slower" speed $N(\mu_s, \sigma_s^2)$ and a distribution of "faster" speed $N(\mu_f, \sigma_f^2)$, where$\mu_i, \sigma_i^2, i \in \{s, f\}$ denote the mean and variance of the distribution i. The μ_f represents the expected space mean speed (travel time) of vehicles traveling through the corridors without stopping and the σ_f^2 represents the speed variance of individual vehicles. The μ_s represents the expected space mean speed (travel time) of vehicles forced to stop and waiting

for green phase and the σ_s^2 indicates the variance of delay. The probability density function (PDF) of mixture model $P(v)$ is defined as

$$P(v) = w_s P(v_s) + w_f P(v_f) \tag{10}$$

where $P(v_i),\ i \in \{s, f\}$ is the PDF of the distribution $N(\mu_i, \sigma_i^2),\ i \in \{s, f\}$; $w_i,\ i \in \{s, f\}$ is the weight of the distribution *i*.

The parameters of the finite mixture models were estimated by SAS® software based on all data from BDS during the study period. In order to provide more precise information, the model was estimated separately for 12 (segments) × 2 (weekend/weekday) × 18 (hours, from 5 am to 9 pm) scenarios. Note that the bimodal flow pattern was not observed during the night time (10 pm to 4 am), thus the mixture model collapses to a model with the only "faster" mode.

5.1.2 Augmentation based on Bayesian Inference

The space mean speed (travel time) of a segment follows a mixture of "faster" distribution and "slower" distribution. Assume that a specific HERE speed observation was taken from one of the two distributions. A Bayesian inference framework was proposed to estimate average space mean speed of 5-minute time interval considering the bimodal traffic flow pattern. The estimated finite mixture model parameters serves as the "prior" information of the bimodal pattern, and the posterior distribution $P(v|v_H)$ is given as

$$P(v|v_H) = w_s \frac{P(v_H|v_f)P(v_f)}{P(v_H)} + w_f \frac{P(v_H|v_s)P(v_s)}{P(v_H)} \tag{11}$$

where w_s, $P(v_f)$, w_f, $P(v_s)$ is given by function (10); v_H is the speed observation from HERE;

the likelihood function $P(v_H|v_f)$ and $P(v_H|v_s)$ is given as

$$(v_H|v_{s/f}) = \frac{1}{\sqrt{2\pi\sigma_H^2}}\exp(-\frac{(v_H - v_{s/f})^2}{2\sigma_H^2}) \quad (12)$$

where the σ_H^2 denotes the variance of HERE speed data, which is not able to be observed. Thus it is estimated by all HERE data during the study period for different scenarios developed by the same manner as the bimodal information estimation process.

Since the space mean speed v_H, v_f and v_s all follow the normal distribution, the three distributions are conjugated with each other. Therefore, the mean value of the posterior distribution, which is the augmented HERE speed, is given as

$$\mu = w_s(\mu_s \frac{\sigma_H^2}{\sigma_H^2+\sigma_s^2} + \mu_H \frac{\sigma_s^2}{\sigma_H^2+\sigma_s^2}) + w_f(\mu_f \frac{\sigma_H^2}{\sigma_H^2+\sigma_f^2} + \mu_H \frac{\sigma_f^2}{\sigma_H^2+\sigma_f^2}) \quad (13)$$

5.2 Results and Discussion

Figure 6 illustrates the distributions of original and augmented data of the same HERE segment showed in Figure 5 (Also note that the magnitude of percentage is different for better illustration). The distribution of the augmented HERE data clearly shows the bi-modal traffic flow pattern rather than one single normal distribution. The augmented data also provides a more realistic proportion of speed above the posted speed limit, which gives a better understanding of the free flow condition. It is noted that the abnormal high percentage of speed data around 40mph is not correct by the proposed algorithm, although it at least reduced the percentage of abnormal data from around Sin The reason might be the processing algorithm of HERE might

intendedly generate speed readings around 40mph (5mph less than the posted speed limit.) This distorts the normal distribution assumption of the speed data, which might further detrimental to the effectiveness of the augmentation framework. The solution might be simply removing the abnormal data. However, since the exact processing algorithm is still proprietary, this kind of data cleaning might lead to the information loss.

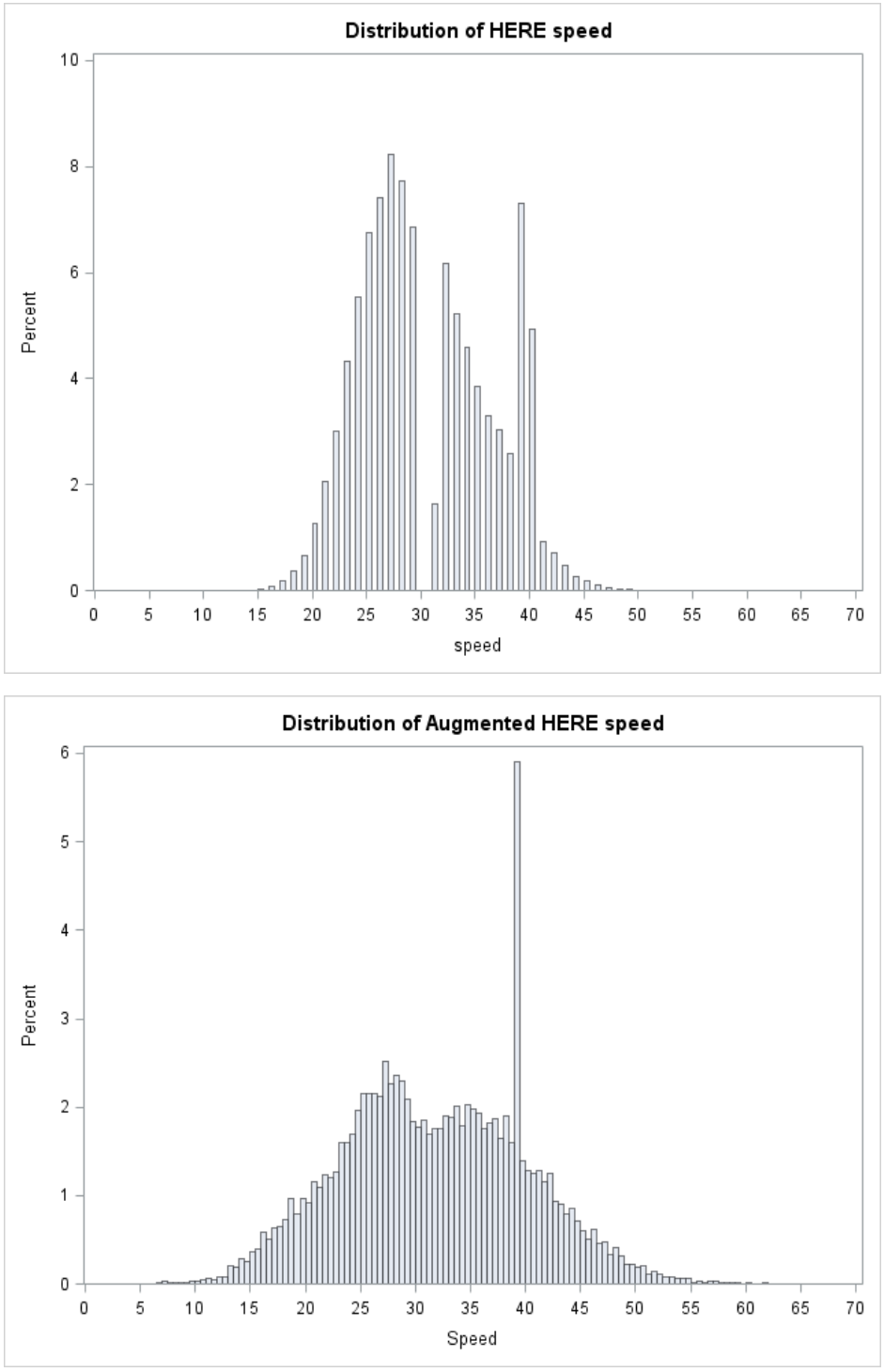

Figure 6 The Distribution of Speed from original (up) and augmented (down) HERE data

Table 2 shows the quantitative measures, AASE and SEB, of the original and augmented HERE speed compared with the GPS ground truth for each segment. According to Table 3, the accuracy of most of the segments (10 of 12) has improved in terms of AASE. And 2 more segments ("102-10186" and "102P10190") turned acceptable (AASE ≤ 10 mph and SEB ≤ ±5 mph) after the augmentation.

Table 2 Deviation of Speed from Original and Augmented HERE from the Ground Truth

HERE_ID	Original		Augmented	
	AASE(mph)	SEB(mph)	AASE(mph)	SEB(mph)
102+10187	11.60	2.08	11.06	1.30
102+10188	**7.56**	**-1.85**	**7.09**	**-1.47**
102+10189	**6.05**	**0.02**	**5.57**	**-1.51**
102+10190	**7.26**	**-3.74**	**7.40**	**-4.46**
102+10191	**9.05**	**-3.32**	**8.94**	**-2.83**
102-10186	9.85	7.04	**8.71**	**4.66**
102-10187	11.57	-8.70	11.21	-8.57
102-10188	10.77	-0.81	10.45	-5.67
102-10189	13.21	-9.89	12.76	-9.96
102-10190	9.41	-7.14	8.88	-5.23
102N10190	12.82	-11.04	13.73	-11.91
102P10190	11.13	6.50	**9.34**	**2.51**
Note: The bold value indicates the data of the segment meets the reliability criteria.				

A possible reason for the accuracy of segment "102N10190" and "102+10190" is not improved is that these two segments are connected with expressway ramps. As discussed above, the traffic on the segment includes the vehicles from and to the expressway through ramps. The dataset used as the validation or BDS data which used to augment the private sector data might not be representative.

5.3 Summary

This chapter presents a preliminary study about the augmentation framework of data from private sectors. In the framework, the bimodal traffic follow pattern estimated by finite mixture model and then combined with the original data by Bayesian inference. The result shows that the augmentation algorithm is able to improve the quality of the private sector data.

CHAPTER 6: CONCLUSION

This thesis evaluated the accuracy and reliability of traffic data from private sector and Bluetooth Detection System data of arterial corridors in Orlando, Florida by the high-resolution GPS trajectories collected from April to June 2017. Two metrics, the AASE and SEB were employed to measure the deviation of space mean speed of the two data sources from the ground truth. The results showed that the BDS data have better accuracy and reliability than the private sector data on the study corridors. A possible reason is that BDS data is able to present the bimodal traffic flow pattern and provides a more realistic compliance of speed limit, which a better understanding of the free flow condition. In addition, the results showed data from private sectors have the different accuracy levels of different arterial corridors, which confirms that data from private sectors requires a site-specific validation.

In order to improve the quality of the private sector data, information of bimodal traffic flow pattern on signalized arterials was estimated by a finite mixture model from historical BDS data. Then, the information is utilized as an informative prior in a Bayesian inference framework to generate enhanced space mean speed. The same evaluation methodology was utilized to evaluate the enhanced data and the results showed that the proposed data augmentation framework is effective for most of the roadway segments except for one segment where the GPS validation datasets were not able to present the influence by the traffic to or from expressway ramps.

For possible future work, the data sources could be evaluated by different scenarios such as different traffic conditions and weather conditions by a larger validation data set for several

corridors. The bimodal traffic flow information derived from traffic flow theory and the combination of model-driven and data-driven approaches to test the proposed augmentation framework.

REFERENCES

Araghi, B. N., Christensen, L. T., Krisnan, R., Olesen, J. H., & Lahrman, H. (2013). Improving The Accuracy Of Bluetooth Based Travel Time Estimation Using Low-Level Sensor Data. *Transportation Research Board 92th Annual Meeting*, *1750*(January), 1–11. https://doi.org/10.3141/2338-04

Bhaskar, A., & Chung, E. (2013). Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. *Transportation Research Part C: Emerging Technologies*, *37*, 42–72. https://doi.org/10.1016/j.trc.2013.09.013

Bhaskar, A., Chung, E., & Dumont, A. G. (2011). Fusing Loop Detector and Probe Vehicle Data to Estimate Travel Time Statistics on Signalized Urban Networks. *Computer-Aided Civil and Infrastructure Engineering*, *26*(6), 433–450. https://doi.org/10.1111/j.1467-8667.2010.00697.x

Bhaskar, A., Qu, M., & Chung, E. (2015). Bluetooth vehicle trajectory by fusing bluetooth and loops: Motorway travel time statistics. *IEEE Transactions on Intelligent Transportation Systems*, *16*(1), 113–122. https://doi.org/10.1109/TITS.2014.2328373

Bhaskar, A., Tsubota, T., Kieu, L. M., & Chung, E. (2014). Urban traffic state estimation: Fusing point and zone based data. *Transportation Research Part C: Emerging Technologies*, *48*, 120–142. https://doi.org/10.1016/j.trc.2014.08.015

Cai, Q., Wang, Z., Zheng, L., Wu, B., & Wang, Y. (2014). Shock Wave Approach for Estimating Queue Length at Signalized Intersections by Fusing Data from Point and Mobile Sensors.

Transportation Research Record: Journal of the Transportation Research Board, *2422*(1), 79–87. https://doi.org/10.3141/2422-09

Choi, K., & Chung, Y. (2002). A Data Fusion Algorithm for Estimating Link Travel Time. *Journal of Intelligent Transportation Systems*, *7*(February 2015), 235–260. https://doi.org/10.1080/714040818

Elefteriadou, L., Kondyli, A., & George, B. St. (2014). *Comparison of Methods for Measuring Travel Time at Florida Freeways and Arterials*. *Transportation Research Center, University of Florida*. https://doi.org/10.1007/s13398-014-0173-7.2

Haghani, A., Hamedi, M., Sadabadi, K., Young, S., & Tarnoff, P. (2010). Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, *2160*(2160), 60–68. https://doi.org/10.3141/2160-07

Hu, J., Fontaine, M., & Ma, J. (2016). Quality of Private Sector Travel-Time Data on Arterials. *Journal of Transportation Engineering*, *142*(4), 1–11. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000815.

José, J., Díaz, V., Belén, A., González, R., & Wilby, M. R. (2015). Bluetooth Traffic Monitoring Systems for Travel Time Estimation on Freeways, *17*(1), 1–10.

Liu, K., Cui, M.-Y., Cao, P., & Wang, J.-B. (2016). Iterative Bayesian Estimation of Travel Times on Urban Arterials: Fusing Loop Detector and Probe Vehicle Data. *PloS One*, *11*(6), e0158123. https://doi.org/10.1371/journal.pone.0158123

Nantes, A., Ngoduy, D., Bhaskar, A., Miska, M., & Chung, E. (2016). Real-time traffic state estimation in urban corridors from heterogeneous data. *Transportation Research Part C: Emerging Technologies*, *66*, 99–118. https://doi.org/10.1016/j.trc.2015.07.005

Singer, J., Robinson, A. E., Krueger, J., Atkinson, J. E., & Myers, M. C. (2013). Travel Time on Arterials and Rural Highways: State-of-the-Practice Synthesis on Arterial Data Collection Technology, FHWA-HOP-13-028, (April), 60. Retrieved from http://www.ops.fhwa.dot.gov/publications/fhwahop13028/index.htm

Wang, Y., Araghi, B. N., Malinovskiy, Y., Corey, J., & Cheng, T. (2014). Error Assessment for Emerging Traffic Data Collection Devices, (June), 106. Retrieved from http://www.wsdot.wa.gov/research/reports/fullreports/810.1.pdf

Wang, Z., Cai, Q., Wu, B., Zheng, L., & Wang, Y. (2016). Shockwave-based queue estimation approach for undersaturated and oversaturated signalized intersections using multi-source detection data. *Journal of Intelligent Transportation Systems*, *21*(3), 1–12. https://doi.org/10.1080/15472450.2016.1254046

Young, S., Hamedi, M., Sharifi, E., Juster, R. M., Kaushik, K., & Eshragh, S. (2015). I-95 Corridor Coalition Vehicle Probe Project : Validation of Arterial Probe Data. Retrieved January 1, 2017, from http://i95coalition.org/wp-content/uploads/2015/02/I-95_Arterial_Validation_Report_July2015-FINAL.pdf?652af7

Yuan, J., & Abdel-aty, M. (2018). Approach-level real-time crash risk analysis for signalized intersections. *Accident Analysis and Prevention*, *119*(April), 274–289. https://doi.org/10.1016/j.aap.2018.07.031

Yuan, J., Abdel-Aty, M., Gong, Y., & Cai, Q. (2019). Real-Time Intersection Crash Risk Prediction Using Long Short-Term Memory Recurrent Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*.

Yuan, J., Abdel-aty, M., Wang, L., Lee, J., Yu, R., & Wang, X. (2018). Utilizing bluetooth and adaptive signal control data for real-time safety analysis on urban arterials. *Transportation Research Part C*, *97*(October), 114–127. https://doi.org/10.1016/j.trc.2018.10.009

Zhang, X., Hamedi, M., & Haghani, A. (2015). Arterial travel time validation and augmentation with two independent data sources. *Transportation Research Record: Journal of the Transportation Research Board*, *2526*(Cv), 79–89. https://doi.org/10.3141/2526-09

Zhang, X., Hamedi, M., & Haghani, A. (2017). A Real-Time Data Fusion Framework for Corridor Travel Time. In *Transportation Research Board 96th Annual Meeting*.

Zheng, L., Ma, H., Wu, B., Wang, Z., & Cai, Q. (2014). Estimation of Travel Time of Different Vehicle Types at Urban Streets Based on Data Fusion of Multisource Data. In *CICTP 2014: Safe, Smart, and Sustainable Multimodal transportation System, ASCE 2014* (pp. 3743–3751).

Zhu, L., Guo, F., Polak, J. W., & Krishnan, R. (2017). Multi-Sensor Fusion Based on the Data From Bus Gps , Mobile Phone and Loop Detectors in Travel Time Estimation, *44*(0).