5-2017

Errors and Truths from Transportation Data Aggregation: Some Implications for Research and Practice

Hyeonsup Lim
University of Tennessee, Knoxville, hlim4@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Hyeonsup Lim entitled "Errors and Truths from Transportation Data Aggregation: Some Implications for Research and Practice." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Civil Engineering.

Lee D. Han, Major Professor

We have read this dissertation and recommend its acceptance:

Shashi Nambisan, Christopher Cherry, Hamparsum Bozdogan

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Errors and Truths from Transportation Data Aggregation: Some Implications for Research and Practice

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Hyeonsup Lim

May 2017

This dissertation is dedicated to
my outrageously loving and supportive wife, Eunjeong Kim,
our sweet little girl and boy, Haram and Han,
and my ever faithful parents, Chayeon Lim and Gyeongsim Kim.

ACKNOWLEDGMENTS

First all, I would like to thank everyone who has encouraged, supported and guided me to during my time at the University of Tennessee. In special, I would like to express my sincere gratitude and appreciation to my advisor, Dr. Lee Han, for his mentorship and friendship. I, just like any other students, was very fortunate to have him as my advisor. He is “the professor” who will definitely be my forever mentor. I also want to thank Dr. Shashi Nambisan, Dr. Christopher Cherry, and Dr. Hamparsum Bozdogan for serving on my committee.

My dissertation would not be possible without support of Dr. Shih Miao Chin and Dr. Ho-ling Hwang, who are my mentors sponsoring my research at Oak Ridge National Laboratory. I was also fortunate to work on a number of research projects sponsored by Tennessee Department of Transportation, Oak Ridge National Laboratory, Southeastern Transportation Center, and Chancellor’s Fellowship at the University of Tennessee. I am very grateful for their support and recognition.

Last but not least, I want to thank all the friends in our transportation program: Bumjoon Bae, Stephanie Hargrove, Jianjiang Yang, Yuandong Liu, Yang Zhang, Brandon Whetsel, Zhihua Zhang, Brandon Worley, Luis Taboada, Hunter McCracken, Ranjit Khatri, Taekwan Yoon, Casey Langford, Dua Abdelqader, Ziwen Ling, Nirbesh Dhakal, Pankaj Dahal, Jun Liu, Kwaku Boakye, Meng Zhang, and many others. They have made this journey memorable and enjoyable.

ABSTRACT

Data aggregation, which is a process to combine information by defined groups for statistical analysis, summary, data size reduction, or other purposes, has fundamental challenges, such as loss of the original information. Improper data aggregation, such as sampling bias or incorrect calculation of average, may cause misreading of information. In first chapter, it is revealed that the harmonic mean, which is used to calculate space mean speed for fixed segment, has a sampling bias, i.e., overestimation with small samples. The several impact analyses show that the sampling bias is affected by sampling rate, time interval, segment length, and distribution type.

If the data aggregation is properly used, it can help us improve analytical efficiency, encounter some of critical problems, or reveal its casualties and other relevant information. Second and third chapters utilize the aggregation of multi-source data to estimate error distributions of data sources and improve accuracy of their measurements. This is a leaping point of evaluating data sources as the proposed model does not require ground truth data. Second chapter focuses more on the methodology, i.e., a modified Approximate Bayesian Computation, incorporated to construct the error distribution with numerous simulations. In the simulated experiment, the proposed model outperformed the alternative approach, which is a conventional way of evaluating data source that is gathering error information by comparing with ground data source. Several sensitivity analyses explore that how the model performance is affected by sample size, number of data sources, and distribution types. The proposed model in chapter II is limited to one dimensional variable, and then the application is expanded to improving the position and distance measurement of connected vehicle environment. The proposed model can be used to further improve the accuracy of vehicle positioning with other existing methods, such as simultaneous

localization and mapping (SLAM). The estimation process can be conducted in real-time operation, and the learning process will try to keep improving the accuracy of estimation. The results show that the proposed model noticeably improves the accuracy of position and distance measurements.

TABLE OF CONTENTS

INTRODUCTION 1
CHAPTER I A Challenge of Using Harmonic Mean as a Calculation of Space Mean Speed on a Fixed-segment: Proof of Bias over Sample Size and the Correction 4
Abstract 5
Introduction 6
Literature Review 8
Proof of Bias 12
Numerical Example 15
Correction of Bias: Analytical Approach 17
Correction of Bias: Simulation-based Approach 19
Impact of Bias in Practice: Vehicle Trajectory Data 25
Impact of Bias in Practice: Multiple Segments in Network 32
Impact of Bias in Practice: Data Comparison 34
Conclusion 37
CHAPTER II Estimation of Error Distribution for Multi-source Data without Ground Truth Data using Modified Approximate Bayesian Computation ... 39
Abstract 40
Introduction 41
Literature Review 43
Methodology 46
Simulation Procedure 49
Results 53
Conclusion 61
CHAPTER III Enhancing Accuracy of Position and Distance Measurements for Connected Vehicles based on Modified Approximate Bayesian Computation Approach 63
Abstract 64
Introduction 65
Literature Review 67
Methodology 71
Simulated Experiments 79
Results 82
Conclusion 88
CONCLUSION 90
LIST OF REFERENCES 92
VITA 102

LIST OF TABLES

Table 1 Range of Parameter Estimates for Simulated Distributions 22
Table 2 Summary of Selected Traffic Data Sources 34
Table 3 Moments of Distributions in the Simulated Experiment 57

LIST OF FIGURES

Figure 1. Exact Expected Values of Harmonic Mean over Sample Size 16
Figure 2. Example of Simulation-based Correction 21
Figure 3. The Impact of Overestimation by Distribution Type (Uniform) 23
Figure 4. The Impact of Overestimation by Distribution Type (Normal) 24
Figure 5. The Impact of Overestimation by Distribution Type (Gamma) 24
Figure 6. Study Area of NGSIM Dataset 26
Figure 7. Expected Space Mean Speed of Sampled Data (NGSIM I-80) 27
Figure 8. SMS Overestimation by Time Interval (3 NGSIM Data Set) 29
Figure 9. SMS Overestimation by Segment Length (3 NGSIM Data Set) 31
Figure 10. Sample Data Generation Procedure for RTMS Data, TN Region 1 ... 32
Figure 11. Impact of Sampling Bias on Region 1, Tennessee 33
Figure 12. The Impact of Overestimation on Comparison of Data Sources 36
Figure 13. Parameter Estimation by Approximate Bayesian Computation 45
Figure 14. Modified Approximate Bayesian Computation (Proposed Model) 48
Figure 15. Example of Estimated Error Distribution over the Iteration 53
Figure 16. Example of Bhattacharyya Distance for This Study 55
Figure 17. Bhattacharyya Distance over Number of Data Sources 55
Figure 18. Bhattacharyya Distance over Sample Size 57
Figure 19. Model Performance by Mean and Standard Deviation 58
Figure 20. Model Performance by Standard Deviation and Skewness 59
Figure 21. Q-Q Plot of Estimated Mean of Error 60
Figure 22. Q-Q Plot of Estimated Standard Deviation of Error 60
Figure 23. Overall Procedure of Proposed Model 73
Figure 24. Candidate Set Generation using Dynamic Grid Search 75
Figure 25. Procedure of Self-Evaluation and Learning for Proposed Model 78
Figure 26. Example of Generated Network 80
Figure 27. Example of Estimated Error Distribution (Distance Measurement) 83
Figure 28. Example of Estimated Error Distribution (Position Measurement) 83
Figure 29. Overall Bhattacharyya Distance over Learned Time Period 84
Figure 30. Example of Estimated Positions vs Measured Positions 85
Figure 31. Overall Mean Absolute Error over Learning Time 87

INTRODUCTION

Albert Einstein said, “Whoever is careless with the truth in small matters cannot be trusted with important matters.” I, as an engineer and as a researcher, see that the efforts to find the truths are often ignored despite its vital importance. For instances, people evaluate models or systems based on benchmark data without justifications of selecting the benchmark data. How can we be sure the benchmark data are true or nearly true? This question could bring more philosophical discussions even in science fields, e.g., uncertainty of quantum mechanics. Having a confidence interval in statistical inferences, rather than providing a deterministic number, could possibly be inherited with considerations of the stochastic nature. Unfortunately, this dissertation does not cover the philosophical matters, but focuses on a specific data aggregation issue and a way to utilize multi-source data to better estimate the truth. Although the most of discussions are limited to transportation field with specific circumstances, implications and applications of the dissertation should not be limited in such area.

Technological advancements have enabled vast volumes of data and information available in both real-time and historical bases. Major traffic information providers have built a data warehouse of multiple petabytes or even more that stores minute-by-minute traffic data across millions of road segments. The unprecedented amounts of data streaming from various sources, so-called “Big Data”, have brought technical challenges for managing the enormous data to handle a variety of problems efficiently and effectively. The size of information that we collect grows exponentially and makes it almost impossible to process all the information at the original level.

Besides the limited data storage or processing time, the data aggregation may be forced for other reasons. One of the major concerns of providing the raw data is confidentiality, which involves ethical or legal issues. However, this should not be misinterpreted as if aggregate data does always guarantee the confidentiality. For instance, for counties that have only one coal-mine company, the county level coal production, employment, and other information will be directly associated with the companies even if it was not intended to be revealed at that level.

In this study, the term “data aggregation” refers to an action or process that combines information by defined groups for statistical analysis, summary, data size reduction, or other purposes. Whether it is necessary or not, data can always be aggregated at a certain level, depending on how we define it. For instance, one’s personal travel speed for a specific segment and time stamp is aggregated data at the defined spatial temporal area and the person. In other words, the data can also be disaggregated by smaller spaces or time intervals, and even the individual may be further separated into different status such as vehicle type or number of passengers. Note that neither the term “raw” nor “aggregated” assures that the data is aggregated or disaggregated at an excessive level.

One of the fundamental challenges of the aggregating data is losing some of the original information, unless the raw data are all preserved in accessible places. Recovering the original information of raw data from the aggregated data is generally unattainable. Even if the aggregated data is enough to provide current benefits in an efficient way, more disaggregated data or the raw data may be an essential requirement to reuse the information for unforeseen purposes. Moreover, improper data aggregation, such as sampling bias or incorrect calculation of average, may cause misreading of information. As an example of this study’s result, the average speed, calculated by harmonic mean, of a fixed

segment tends to be overestimated when only small samples are captured. The expected bias could be ignorable at some places, but the consequences could become noteworthy as the scope gets larger. The sampling bias may cause poor predictive power of modeling analysis, incorrect assessment of projects, and inefficient resource allocation.

However, if the data aggregation is properly used, not to mention the reduction of the required management resources and the protection of confidentiality, it also helps us improve analytical efficiency, encounters some of critical problems, or reveals its casualties and other relevant information, which appears to be elusive at the original level of data.

Although the data aggregation occurs and matters in almost all places with a broad definition of the aggregation, this study aims to reveal a few specific challenges of the data aggregation, provide methods to alleviate the issues, and utilize the aggregation of multi-source data to estimate the true error distribution, in transportation field.

To this end, the dissertation is organized into the following three chapters.

- **Chapter I** identifies and proves a bias of space mean speed over sample size when using harmonic mean, and provides a correction method.
- **Chapter II** estimates the error distribution of multi-source data without information of ground truth, when the data is collected independently and simultaneously.
- **Chapter III** utilizes the second chapter and applies the method to connected vehicle systems, to enhance accuracy of position and distance measurements.

CHAPTER I

A CHALLENGE OF USING HARMONIC MEAN AS A CALCULATION OF SPACE MEAN SPEED ON A FIXED-SEGMENT: PROOF OF BIAS OVER SAMPLE SIZE AND THE CORRECTION

This chapter presents a modified version of a research paper by Hyeonsup Lim, Bumjoon Bae, Lee D. Han, and Hamparsum Bozdogan.

Abstract

Harmonic mean, which is the reciprocal of the arithmetic mean of the reciprocals of observations, is considered to be an appropriate average for rates or ratios. In transportation, the harmonic mean is used to calculate Space Mean Speed, where a designated segment length is fixed or passing vehicles are assumed to complete the segment with given speeds. This study identifies and proves a sampling bias of harmonic mean, which definitively affects the sampling bias of the space mean speed on the fixed segments. The study shows, a mathematical proof and numerical example, that harmonic mean is overestimated when the sample size is smaller than the population. The study also provides both analytical and simulation-based correction approach. From the simulations and the three case studies of investigating the impact of the sampling bias in this study, it is recognized that the sampling bias is affected by sampling rate, time interval, segment length, and distribution type.

Introduction

There will be no argument regarding the sample size if we have the data of entire population. However, the collected data generally do not cover or guarantee the entire population in the field. When an average value of sampled data is used, there is an underlying assumption that the average value can represent the entire population within a certain tolerance (e.g., a confidence interval in statistical inference).

The average value of sampled data, however, may not represent well the population if not appropriately aggregated. One possible reason is a sampling bias, in which some elements of the population are less or more likely to be included in the observation than others. Suppose that a loop detector was installed only in the left most lane on a segment of freeway, while truck drivers are allowed to drive only on the rest of lanes. It is expected that the average travel speed of collected data is likely to be higher than the average of all vehicles passed the segment.

Another possible reason of having a biased average is an inappropriate calculation of averaging the sampled data. This involves unknown parameters used in the calculation, such as a case that a vehicle length is estimated for speed data from a single loop detector. Furthermore, the calculation method itself could be biased over sample size, such as a harmonic mean, which will be discussed throughout this chapter.

Harmonic mean, one of the Pythagorean means, is the reciprocal of the arithmetic mean of the reciprocals of observations. The harmonic mean is considered to be an appropriate average for rates or ratios, e.g., vehicle speeds. As a simple example, when two vehicles travel a certain distance, the two

vehicles' total travel time is the same as the travel time that one travels the whole distance at the harmonic mean of the two speeds. It has been used often, but not always, as a calculation of space mean speed in transportation field.

The statement, "space mean speed is the same as the harmonic mean of observed speeds", is debatable in a sense that the definition of space mean speed differs from researchers. The distinction of definitions has been frequently made whether it specifies only the space, or both the space and time. The first definition allows the calculation of space mean speed as harmonic mean of vehicle speeds, which does not always apply to the second one because some of vehicles within the section may not have completed the crossing of segment. Therefore, summarizing the definitions, the use of harmonic mean for calculating SMS should have, at least, the following conditions:

- The designated segment is fixed, and so is the length of segment.
- There is an assumption that all observed vehicles completed (i.e., traveled the full distance) the segment over a given time.
- If vehicle speeds are measured at a point over time, the use of harmonic mean assumes that speeds of individual vehicles do not vary much over the segment.

With the aforementioned conditions and definitions, many agencies, where infrastructure-based detectors such as loop or radar detectors are employed, often use the harmonic mean to calculate the space mean speed because of either following the first definition or having difficulties of measuring the explicit travel distances of each vehicle in the space-time frame.

The objective of this study is to identify and prove a bias of space mean speed over sample size when using harmonic mean, and provide a correction method. Note that this paper does not elaborate how to estimate SMS. The study will show that space mean speed is overestimated when the sample size is smaller

than the population, and the following sections will describe a mathematical proof, numerical example, and a case study of the bias.

Again, it is vital to consider if the expected value of the sample average is biased over sample size. It is a theoretical fault affecting the calculation of everywhere using harmonic mean of sample data to represent that of the population, even if the bias is minimal in some places. Three different case studies were conducted to investigate the impact of the sampling bias by time interval, segment length, and sampling rate.

Literature Review

Space Mean Speed

In physics, speed is a scalar quantity, which is distinguished from velocity, a vector quantity being aware of direction. Transportation engineers more often use the term “speed” rather than “velocity”, although a direction is often considered technically. The individual vehicle speed is measured through an observation over time and space. When averaging a group of vehicle speeds, there are two ways of calculation, Time Mean Speed (TMS) and Space Mean Speed (SMS). While the TMS is an average of observations of vehicle speeds over a given time, the definition of SMS differs from researchers.

One definition of the SMS is the mean speed of vehicles to travel a given distance [1-4]. Note that this definition does not specify a time domain. With this definition, the harmonic mean can be a correct calculation of the SMS, only when all of the observed vehicles completed the segment with a given time. Furthermore, Wardrop and Edie implicitly allowed the use of harmonic mean of

instant speeds, i.e., the speeds measured at a point over time [1, 4]. This could be an insignificant matter if the speed of individual vehicles does not vary much over the segment. Otherwise, there will be a difference between the SMS over the segment and the harmonic mean of the instant speeds.

The ITE handbooks and HCM define the SMS as the total travel distance divided by the total travel time [5-7]. This definition specifies an explicit rectangular space and time frame as an observation domain of the average. Wohl and Martin defines the SMS as a weighted average associated with the travel time spent traveling a given length of segment [8].

Another definition of SMS takes an average speed of all of the vehicles within a given segment at an instant time [9-12]. The major distinction is made in which they use the arithmetic mean of the vehicle speeds, not the harmonic mean. Specifically, Haight shows that the SMS calculated in this manner is unbiased to the true distribution of speeds, by assuming that each vehicle does not change their speed over the time space diagram [12].

FHWA states that "*Regardless of the particular definition put forward for space mean speed, …, it is necessary to ensure that one has measured space mean speed, rather than time mean speed.*" [13]. Our study focuses on the SMS where the harmonic mean speed is used as a calculation, and hence the term 'space mean speed' or 'SMS' hereinafter is limited to the certain condition.

Estimation of Space Mean Speed

Equation 1 and 2 describe TMS and SMS using the limited definitions, by making them equivalent to arithmetic mean and harmonic mean respectively. In the equations, $\bar{u}_t$ denotes TMS, $\bar{u}_s$ denotes SMS, and u_i is an individual speed of vehicle i.

$$\bar{u}_t = \frac{1}{N}\sum_{i=1}^{N} u_i \tag{1}$$

$$\bar{u}_s = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{u_i}} \tag{2}$$

TMS is more influenced by faster vehicles, thus the average speed can be overestimated for the consideration of macroscopic traffic flow characteristics, and consequently the density can be underestimated, although the extent of the differences is site-specific [14, 15]. Therefore, it is well known that the SMS should be used to estimate a correct density [15, 16]. Compared to TMS, Soriguera and Robusté (2011) addressed the importance of SMS in terms of both modeling of traffic flow theory and practical purposes.

In 1952, Wardrop showed the general relationship between TMS and SMS [17] :

$$\bar{u}_t = \bar{u}_s + \frac{\sigma_s^2}{\bar{u}_s} \tag{3}$$

The difference between TMS and SMS comes exactly from the calculation of average, the arithmetic mean and the harmonic mean. With these definitions, TMS is always greater than or equal to SMS. Although Equation 3, shown by Wardrop (1952), is theoretically evident and popular, it is not common to estimate TMS from SMS and the variance of SMS in practice. Vehicle speed data from detectors in highway are often aggregated for a certain time interval (e.g., 30 sec or 1 min) and transmitted to Traffic Management Center (TMC) due to a technical or cost constraint, in which the aggregated speed is TMS, not SMS [18]. In addition, the variance of SMS is typically unobtainable. Therefore, the counter relationship between TMS and SMS has been suggested based on empirical studies [14, 19, 20].

$$\bar{u}_s = \bar{u}_t - \frac{\sigma_t^2}{\bar{u}_t} \quad (4)$$

Soriguera and Robusté (2011) suggested a probabilistic method to estimate SMS from TMS based on Equation 4 using aggregated double-loop detector data, in which the standard deviation of TMS is estimated by assuming the normality of vehicle speed distribution [18]. However, in the study, the normality assumption is not proven distinctly. There is a research about determining the required penetration or sampling rate to obtain certain confidence interval of SMS in probe data collection [21], but the study does not describe the bias of SMS over sample size.

Estimation of Harmonic Mean

In the field other than traffic engineering, several studies have suggested methods to estimate the harmonic mean. Limbrunner et al. (2000) introduced an Maximum Likelihood Estimate (MLE) for the harmonic mean and showed that their model is more efficient than the previously introduced estimators ([22, 23]) for the harmonic mean, but the studies are limited to lognormal observations [24]. Satagopan et al. (2000) suggested a method to stabilize the harmonic mean estimator which is used for the Bayes factor, based on the approach of reducing the parameter space by modified estimator for the harmonic mean of heavier tailed densities [25]. Although more studies are found in providing analytical evidences of the inequality of harmonic mean compared to, normally, either arithmetic mean or geometric mean, they didn't explicitly describe the estimation of harmonic mean [26-33].

Jensen et al. (1997, 1998) and Limbrunner et al. (2000) calculated the approximations to the bias and variance of harmonic mean estimators where data is lognormally distributed [22-24]. If a variable X follows a lognormal

distribution, then $ln(X)$ follows a normal distribution with mean μ and variance σ^2. Jensen et al. describes, as shown in Equation 5, the bias term, which also can be used to correct the bias for the lognormal distribution.

$$\text{bias} = \exp\left(\mu - \frac{\sigma^2}{2}\right) \cdot \frac{[\exp(\sigma^2) - 1]}{n} \tag{5}$$

However, the proof is strictly limited to the case of lognormal distribution. Based on our literature review, no study has proved that the expected value of harmonic mean decreases as the sample size increases regardless of data distribution.

Proof of Bias

Jensen's Inequality

The Jensen's Inequality, which has been applied in a variety of engineering fields, shows that the convex transformation of expected value of a variable x is less than or equal to the expected value applied after convex transformation as described in Equation 6, where $\varphi(x)$ is a convex function.

$$\varphi(\mathrm{E}(x)) \leq \mathrm{E}(\varphi(x)) \tag{6}$$

The Jensen's inequality, although it is not the only way, can prove that arithmetic mean is greater than or equal to harmonic mean by Equation 7 to 9, where $\mathrm{f}(x) = 1/x$.

$$\mathrm{f}(\mathrm{E}(x)) \leq \mathrm{E}(\mathrm{f}(x)) \tag{7}$$

$$\frac{n}{x_1 + x_2 + \cdots + x_n} \leq \frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}{n} \tag{8}$$

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \tag{9}$$

Expansion of Jensen's Inequality

Suppose that H_{n} refers to harmonic mean of sample size n, where there are N observations in the population and $\mathrm{n} \leq \mathrm{N}$. First, we write an expected value of harmonic mean with two samples, H_2,

$$E(H_2) = \frac{1}{\mathrm{C}(\mathrm{n}, 2)} \times [\frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} + \frac{2}{\frac{1}{x_1} + \frac{1}{x_3}} + \cdots + \frac{2}{\frac{1}{x_{N-1}} + \frac{1}{x_N}}] \tag{10}$$

By Jensen's Inequality,

$$\frac{2}{\frac{1}{x_a} + \frac{1}{x_b}} \leq \frac{x_a + x_b}{2} \tag{11}$$

Therefore,

$$\frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} \leq \frac{x_1 + x_2}{2}, \frac{2}{\frac{1}{x_1} + \frac{1}{x_3}} \leq \frac{x_1 + x_3}{2}, \ldots, \frac{2}{\frac{1}{x_{N-1}} + \frac{1}{x_N}} \leq \frac{x_{N-1} + x_N}{2} \tag{12}$$

Then, we rewrite Equation 10 using the relationship in Equation 12.

$$E(H_2) = \frac{1}{\mathrm{C(N,2)}} \times [\frac{2}{\frac{1}{x_1}+\frac{1}{x_2}} + \frac{2}{\frac{1}{x_1}+\frac{1}{x_3}} + \cdots + \frac{2}{\frac{1}{x_{N-1}}+\frac{1}{x_N}}]$$
$$\leq \frac{1}{\mathrm{C(N,2)}} \times \left[\frac{x_1+x_2}{2} + \frac{x_1+x_3}{2} + \cdots + \frac{x_{N-1}+x_N}{2}\right] \quad (13)$$
$$= \frac{2}{\mathrm{N(N-1)}} \times \left[\frac{(N-1)(x_1+x_2+\cdots+x_N)}{2}\right] = E(x) = E(H_1)$$

This proves,

$$E(H_2) \leq E(H_1) \quad (14)$$

For comparing H_3 and H_2, we first write an expected value of H_3.

$$E(H_3) = \frac{1}{\mathrm{C(N,3)}} \times [\frac{3}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}} + \frac{3}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_4}} + \cdots + \frac{3}{\frac{1}{x_{N-2}}+\frac{1}{x_{N-1}}+\frac{1}{x_N}}] \quad (15)$$

Equation 16 takes the harmonic mean of all possible combinations of the two samples, from a group with three samples.

$$3/(\frac{1}{x_a}+\frac{1}{x_b}+\frac{1}{x_c}) = 3/(\frac{\frac{1}{x_a}+\frac{1}{x_b}}{2} + \frac{\frac{1}{x_a}+\frac{1}{x_c}}{2} + \frac{\frac{1}{x_b}+\frac{1}{x_c}}{2}) \leq (\frac{2}{\frac{1}{x_a}+\frac{1}{x_b}} + \frac{2}{\frac{1}{x_a}+\frac{1}{x_c}} + \frac{2}{\frac{1}{x_b}+\frac{1}{x_c}})/3 \quad (16)$$

Thus, Equation 15 can be rewritten as:

$$E(H_3) = \frac{1}{\mathrm{C(N,3)}} \times \left[\frac{3}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}} + \cdots + \frac{3}{\frac{1}{x_{N-2}}+\frac{1}{x_{N-1}}+\frac{1}{x_N}}\right] \quad (17)$$

$$\leq \frac{1}{\mathrm{C(N,3)}} \times \left[\frac{\frac{2}{1/x_1+1/x_2} + \frac{2}{1/x_1+1/x_3} + \frac{2}{1/x_2+1/x_3}}{3} + \cdots + \frac{\frac{2}{1/x_{N-2}+1/x_{N-1}} + \frac{2}{1/x_{N-2}+1/x_N} + \frac{2}{1/x_{N-1}+1/x_N}}{3}\right]$$

$$= \frac{1}{\mathrm{C}(\mathrm{N},3)} \times \left[\frac{E(H_2) \cdot (\mathrm{N}-2) \cdot \mathrm{C}(\mathrm{N},2)}{3}\right] = E(H_2)$$

Therefore, it proves $E(H_3) \leq E(H_2)$.

Likewise, we can compare H_4 and H_3,

$$E(H_4) \leq \frac{1}{\mathrm{C}(\mathrm{N},4)} \times \left[\frac{E(H_3) \cdot (\mathrm{N}-3) \cdot \mathrm{C}(\mathrm{N},3)}{4}\right] = E(H_3) \tag{18}$$

By following the same step, we get the inequality over the sample size n and $n+1$:

$$E(H_{n+1}) \leq E(H_n) \tag{19}$$

By generalizing the relationship, an expected value of harmonic means is decreased as the sample size is increased.

$$E(H_m) \leq E(H_n), if\ m \geq n \tag{20}$$

Numerical Example

Suppose that we have a population of 20 data points from 1 to 20 with unit increment (1, 2, …, 19, 20). We assume that the observation follows a discrete uniform distribution, where each data point is equally likely to be observed with a probability of 1/20. Then, the unbiased expected value for sample size N can be

calculated by simply averaging all possible combinations of picking N-samples among the population, which of number of cases is ${}_{20}C_N$.

For $N = 1$ as an example, the number of all possible combination cases is ${}_{20}C_1$ (= 20), i.e., 20 groups of one sample size. In this case, the harmonic mean of each group is identical to the value of the data point in the group. Thus, the expected value of harmonic mean of $N = 1$ is same as the arithmetic mean of the population, which is 10.5.

$$E(H_1) = \frac{H(1) + H(2) + \cdots + H(19) + H(20)}{20} = \frac{1 + 2 + \cdots + 19 + 20}{20} = 10.5 \quad (21)$$

For $N = 3$, the number of all possible combination cases is ${}_{20}C_3$ (=1,140). By calculating the harmonic mean of all of the combinations, we obtain the expected value of 7.83, as shown in Equation 22.

$$E(H_3) = \frac{\frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}} + \frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{4}} + \cdots + \frac{3}{\frac{1}{18}+\frac{1}{19}+\frac{1}{20}}}{1{,}140} = 7.82 \quad (22)$$

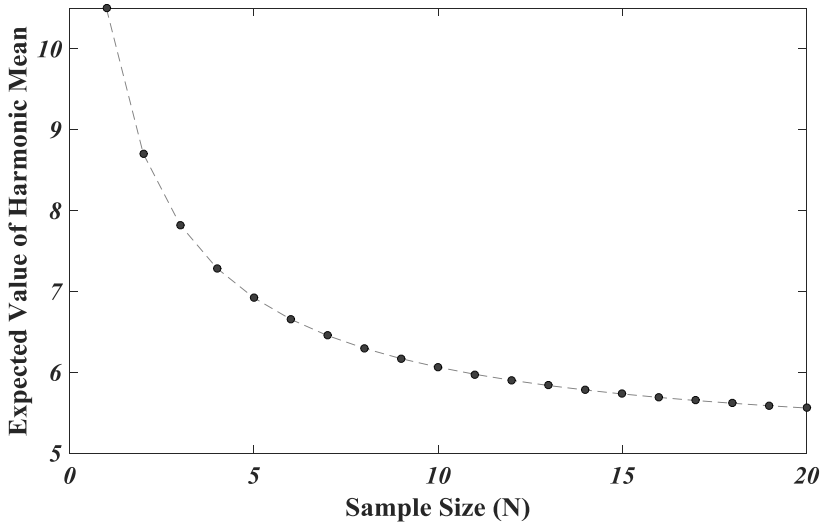


Figure 1. Exact Expected Values of Harmonic Mean over Sample Size

Accordingly, for $N = 20$, the expected value of harmonic mean is same as the harmonic mean of population, which is 5.56. Figure 1 shows the changes of expected value of harmonic mean over sample size for this example. It is very clear that the expected value of harmonic mean decreases as the sample size increases.

Correction of Bias: Analytical Approach

To provide an analytical example of the correction of bias, we assume that the travel time of vehicle i, t_i, follows a gamma distribution.

Sum of Independent Gamma Random Variables

Let t_i be a travel rate, an inverse of vehicle speed v_i, and then the harmonic mean of n vehicle speeds can be calculated as Equation 23.

$$H_{\mathrm{n}} = \frac{n}{\frac{1}{v_1} + \frac{1}{v_2} + \cdots + \frac{1}{v_n}} = \frac{1}{\frac{t_1}{n} + \frac{t_2}{n} + \cdots + \frac{t_n}{n}} \tag{23}$$

Suppose that t_i follows a gamma distribution, i.e., v_i follows an inverse gamma distribution, with a shape parameter α and a rate parameter β. Then, the corresponding probability density function can be calculated as Equation 24, with the expected value α/β and variance α/β^2.

$$f(t_i) = \frac{\beta^{\alpha} e^{-\beta t_i} {t_i}^{\alpha-1}}{\Gamma(\alpha)}, \qquad for\ t_i \geq 0 \tag{24}$$

Then, t_i/n follows a gamma distribution with a shape parameter α and a rate parameter $\mathrm{n}\beta$, since multiplying the $1/\mathrm{n}$ changes only the ratio.

A sum of two independent gamma random variables, with a same rate parameter, also follows a gamma distribution. Then, the shape parameter can be obtained as a sum of the two shape parameters of the two independent variables and the rate parameter remains the same [34]. Therefore, the sum of t_1/n to t_n/n, which is a denominator of right hand side of Equation 23, follows a gamma distribution with a shape parameter $\mathrm{n}\alpha$ and a rate parameter $\mathrm{n}\beta$.

$$\frac{t_1}{n}+\frac{t_2}{n}+\cdots+\frac{t_n}{n}\ \sim\mathrm{Gamma}(\mathrm{n}\alpha,\mathrm{n}\beta) \tag{25}$$

Then, the expected value and variance can be calculated as the followings:

$$\begin{aligned}\mathrm{E}(t_1/\mathrm{n}\ +\cdots t_n/\mathrm{n}) &= (\mathrm{n}\alpha)/(\mathrm{n}\beta)=\alpha/\beta,\\ \mathrm{Var}(t_1/\mathrm{n}+\ldots t_n/\mathrm{n}) &= (\mathrm{n}\alpha)/(\mathrm{n}\beta)^2=\alpha/(\mathrm{n}\cdot\beta^2)\end{aligned} \tag{26}$$

Inverse Gamma Distribution

When x is a gamma random variable with the parameters α' and β', $1/x$ has an inverse gamma distribution with the moments of x as described in Equation 27 [35].

$$\mathrm{E}(x^n)=\frac{(\beta')^n}{(\alpha'-1)\cdots(\alpha'-\mathrm{n})},\qquad if\ \alpha'>n \tag{27}$$

For $n=1$,

$$\mathrm{E}(x) = \frac{\beta'}{\alpha' - 1}, \qquad \text{for } \alpha' > 1 \tag{28}$$

For $n = 2$,

$$\mathrm{E}(x^2) = \frac{(\beta')^2}{(\alpha' - 1)(\alpha' - 2)}, \qquad \text{for } \alpha' > 2 \tag{29}$$

Thus, the variance can be calculated as Equation 30.

$$\mathrm{Var}(x) = \mathrm{E}(x^2) - \mathrm{E}(x)^2 = \beta'^2/(\alpha' - 1)^2(\alpha' - 2) \tag{30}$$

Using the expected value and variance of inverse gamma distribution, the expected value and variance of harmonic mean of n vehicle speeds can be calculated as shown in Equation 31.

$$\begin{aligned} \mathrm{E}(1/(\mathrm{t}_1/\mathrm{n} + \cdots \mathrm{t}_\mathrm{n}/\mathrm{n})) &= \mathrm{n}\beta/(\mathrm{n}\alpha - 1) \\ \mathrm{Var}(1/(\mathrm{t}_1/\mathrm{n} + \cdots \mathrm{t}_\mathrm{n}/\mathrm{n})) &= (\mathrm{n}\beta)^2/(\mathrm{n}\alpha - 1)^2(\mathrm{n}\alpha - 2) \end{aligned} \tag{31}$$

Therefore, the corrected expected value of harmonic mean, with sampling rate γ and collected sample size $\mathrm{m} = \mathrm{n} \cdot \gamma$, can be estimated by Equation 32.

$$\begin{aligned} \mathrm{E}(H_\mathrm{n}) &= \mathrm{E}\left(H_{\mathrm{m},\gamma}\right) = (\mathrm{m}/\gamma) \cdot \beta/((\mathrm{m}/\gamma) \cdot \alpha - 1) \\ \mathrm{Var}(H_\mathrm{n}) &= \mathrm{Var}\left(H_{\mathrm{m},\gamma}\right) = ((\mathrm{m}/\gamma) \cdot \beta)^2/((\mathrm{m}/\gamma) \cdot \alpha - 1)^2((\mathrm{m}/\gamma) \cdot \alpha - 2) \end{aligned} \tag{32}$$

Correction of Bias: Simulation-based Approach

The analytical approach to correct the bias is useful only when it is obtainable. However, this is not feasible for many cases since the solutions could be too complex or unobtainable. More fundamentally, defining distribution of such variables to a certain type with parametric estimates may not be reasonable for

some cases. To this end, simulation-based approach for the correct of bias might be more useful and applicable for most of cases.

Our concern is to know the expected bias from the collected sample data to population data. This can be implemented by generating numerous simulations with given data distribution and comparing the harmonic mean of two different sample sizes of data, one for the population and the other for the sample. The ratio of the harmonic mean of simulated population to the harmonic mean of simulated sample is defined as an 'adjustment factor' in this study, as shown in Equation 33. Then, the adjustment can be simply done by multiplying the adjustment factor to the harmonic mean of actual sample data. The basic idea here is running Monte-Carlo simulations to generate enough data to build the reliable adjustment factors.

$$\begin{aligned} \mathrm{Adj}\left(H_{\mathrm{m},\gamma}\right) &= \widehat{\mathrm{E}(H_{\mathrm{n}})}/\widehat{\mathrm{E}(H_{\mathrm{m}})} \\ \widehat{H_{\mathrm{n}}} &= H_{\mathrm{m}} \cdot \mathrm{Adj}\left(H_{\mathrm{m},\gamma}\right) \end{aligned} \tag{33}$$

where

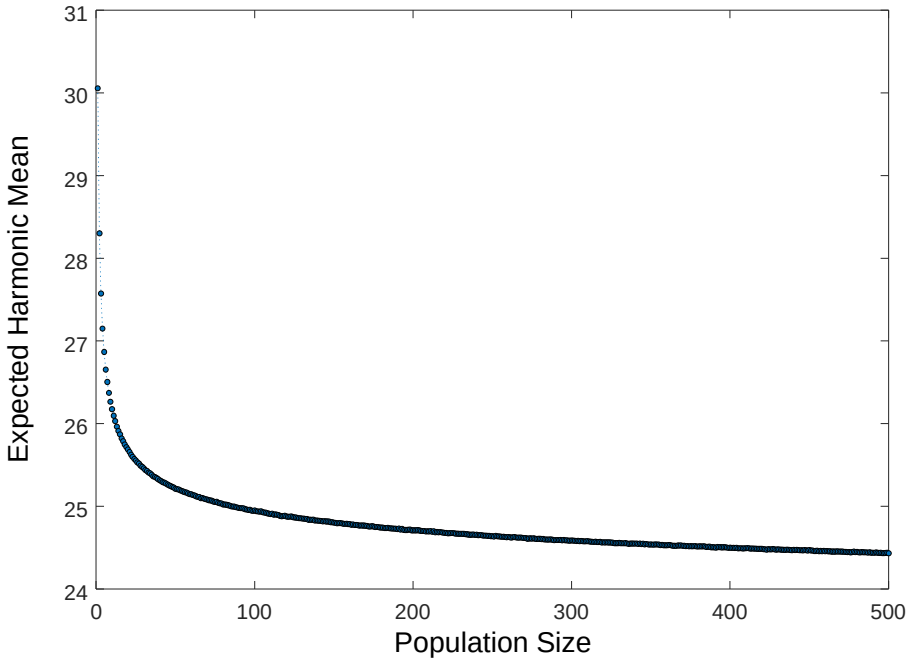
$\mathrm{Adj}\left(H_{\mathrm{m},\gamma}\right)$ is an adjustment factor with sample size m and sampling rate γ,

$\widehat{H_{\mathrm{n}}}$ is an estimated harmonic mean of population size n, and

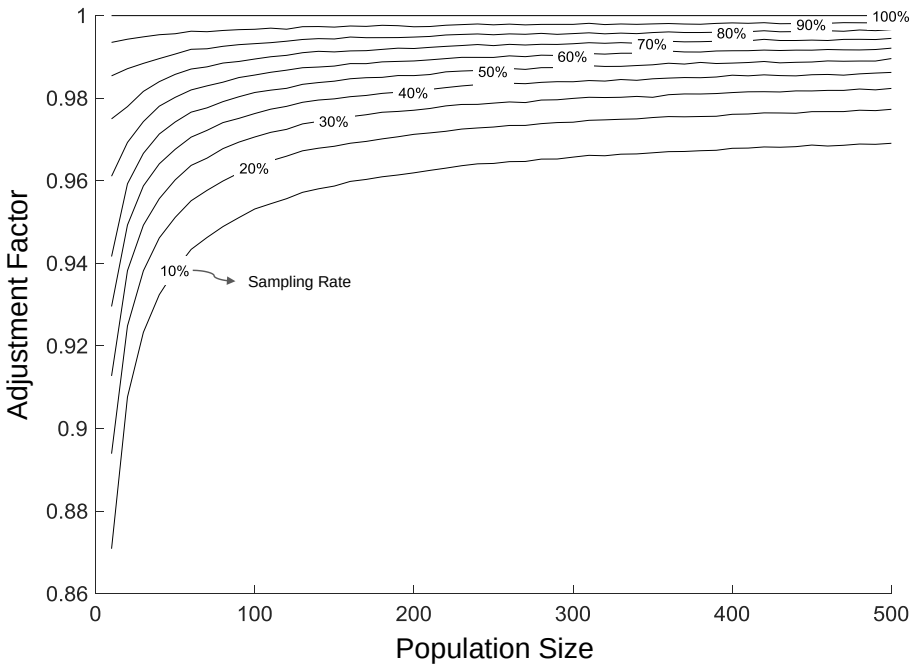
$\widehat{\mathrm{E}(H_{\mathrm{n}})}$ and $\widehat{\mathrm{E}(H_{\mathrm{m}})}$ are estimated expected values of harmonic mean for samples size n and m, from the Monte-Carlo simulations.

Example of Simulation-based Correction

Figure 2 shows a simulation-based correction for the bias of harmonic mean over sample size. The example assumes a normally distributed random variable with mean of 30 and standard deviation of 10. Then, the expected value of harmonic mean over sample size can be estimated by numerous simulations, e.g., a million times, as shown in Figure 2 *(a)*.



(a) Expected Harmonic Mean over Population Size



(b) Adjustment Factor over Sampling Rate

Figure 2. Example of Simulation-based Correction

Based on the estimated expected harmonic mean, the adjustment factors are calculated for various sampling rates and population sizes. Suppose that we captured only 10% out of 100 vehicles in the field, i.e., the sample size of the collected data is only 10 while the population size is 100. In Figure 2 *(a)*, the estimated expected harmonic means for population size of 10 and 100 are 26.18 and 24.95, respectively. The adjustment factor will be 0.95 = 24.95/26.18. The correction can simply be conducted by multiplying this adjustment factor to the harmonic mean of the collected data. This adjustment factor can be obtained at the population size of 100 and 10% sampling rate in Figure 2 (b).

Adjustment Factor by Distribution Type

To provide some implications on the impact of the bias by different distribution types, the adjustment factors are calculated for uniform, normal, and gamma distributions with randomly generated parameter estimates and numerous simulated data sets. This analysis should enable us to see how the parameter estimates of the distributions could affect the adjustment factors. The ranges of the parameter estimates tested for each distribution are described in Table 1. The parameter estimates were generated within the ranges, and 100,000 different sets of data were obtained to calculate average adjustment factors.

Table 1 Range of Parameter Estimates for Simulated Distributions

Distribution Type	Parameter	Range	
		Min	Max
Uniform	Lower Bound	10	55
	Upper Bound	55	100
Normal	Mean	20	100
	Standard Deviation	5	50
Gamma	Shape Parameter	2	12
	Scale Parameter	5	105

Figure 3, 4, and 5 represent the adjustment factors of 10% sampling rate, calculated based on the simulations, for the three distribution types. As shown in Figure 3, the adjustments factor of uniform distribution decreases i.e., more biased, as the upper bound increases and the lower bound decreases. This implies that the adjustment factor of uniform distribution decreases as their variance increases, since the variance of uniform distribution is proportional to the square of difference between the upper bound and the lower bound.

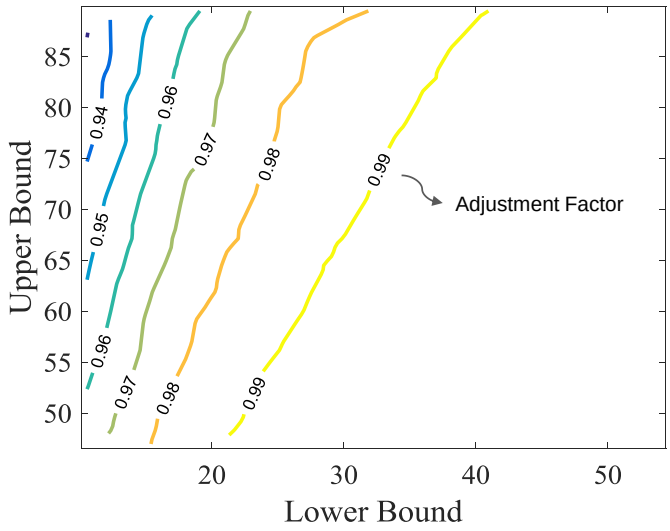


Figure 3. The Impact of Overestimation by Distribution Type (Uniform)

Figure 4 illustrates the impact of the overestimation for normal distribution. Like uniform distribution, the adjustment factor decreases as the variance increases. Also, normally distributed data with larger means tend to have smaller adjustment factors, i.e., more bias.

Figure 5 displays the overestimation impact for gamma distribution. Based on the simulated tests for gamma distribution, the harmonic mean of sample data tends to be more biased when the shape parameter is small. As compared to uniform and normal distributions where the both two parameters affect the adjustment factor obviously, the impact of overestimation for gamma distribution is very sensitive to the shape parameter while the scale parameter relatively affects little.

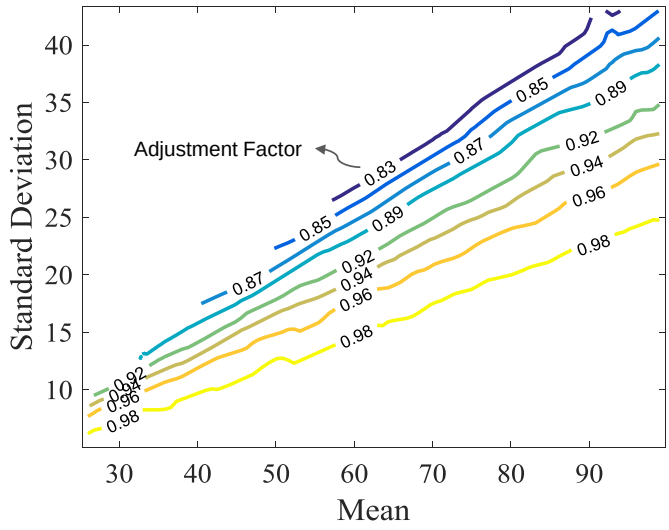


Figure 4. The Impact of Overestimation by Distribution Type (Normal)

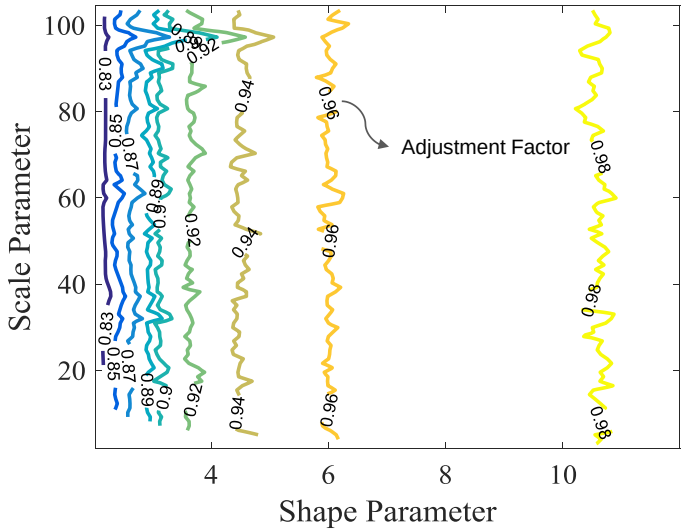


Figure 5. The Impact of Overestimation by Distribution Type (Gamma)

Impact of Bias in Practice: Vehicle Trajectory Data

As shown in the numerical example, the effect of sample size on expected value of harmonic mean for certain cases can be roughly seen with simulated data. To simplify the simulation, we may consider making an assumption that each observation is independent. However, the vehicle-to-vehicle variation of speed is not independent, unless completely free flow. Therefore, this study will use real vehicle trajectory data, which can be aggregated into different levels of space and time interval.

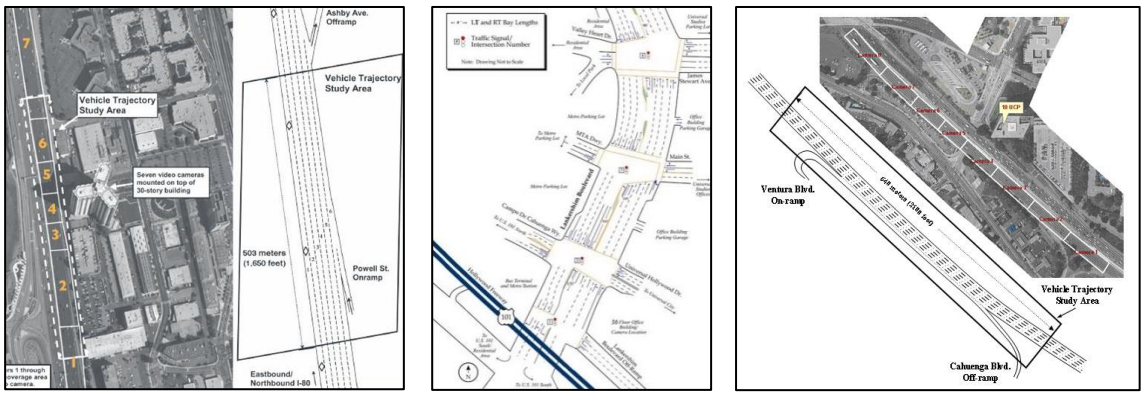
Data Description

Next Generation SIMulation (NGSIM) is a public-private partnership program, provided by the Federal Highway Administration (FHWA), to develop open behavioral algorithms for microscopic traffic simulations. The program provides data sets for three segments: 1) I-80 in the San Francisco Bay area in Emeryville, CA, on April 13, 2005 (45 minutes), 2) Lankershim Boulevard in the Universal City neighborhood of Los Angeles, CA, on June 16, 2005 (30 minutes), and 3) southbound US 101, also known as the Hollywood Freeway, in Los Angeles, CA, on June 15th, 2005 (45 minutes). The vehicle trajectory data, processed from videos, is provided at one-tenth of second with detailed information including lane positions and relative locations to other vehicles.

Case Study Result

Figure 7 displays the simulation results of expected SMS of sampled data, using the NGSIM data on I-80 in the San Francisco Bay area. The study area was virtually segmented by a certain distance, e.g., every 100 ft. All observed vehicles passed the segment were assigned as a population set, with a fixed time interval, e.g., 1-minute. Then, the population set was sampled randomly in

simulation by different sampling rate, e.g., 10%. In Figure 7, the result *(a)* shows the result of a single simulation run, while the result of *(b)* is based on 100 simulations.

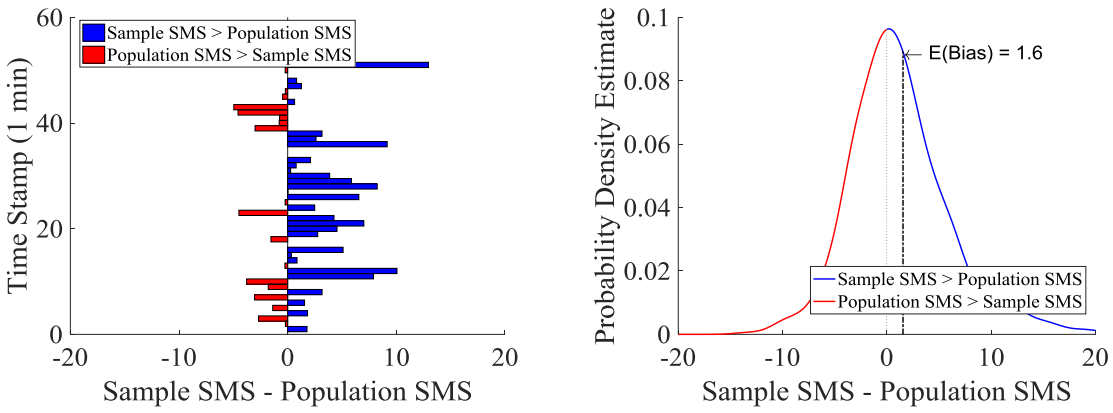


(a) I-80 *(b) Lankershim* *(c) US 101*

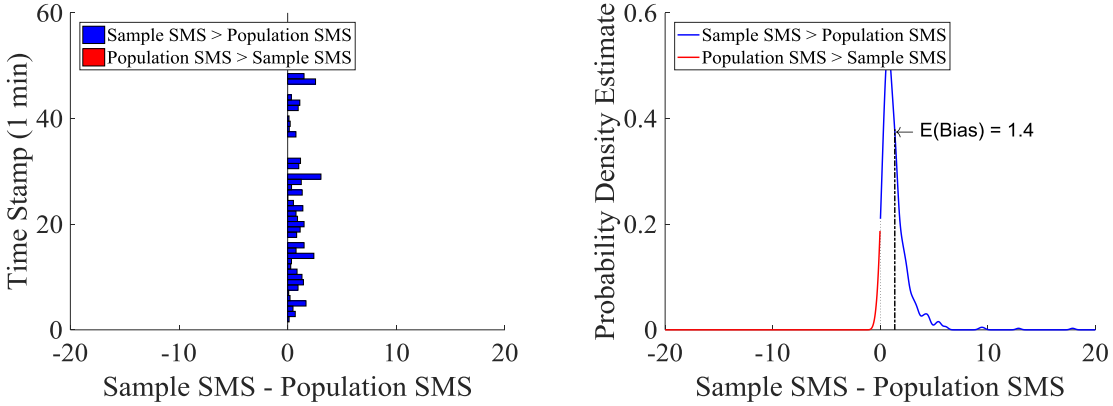
Figure 6. Study Area of NGSIM Dataset

(Source: http://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm, [36])

In Figure 7 *(a)*, SMS of the sample data is often greater than that of the population set and the positive bias (sample SMS > population SMS) is more likely to happen than the negative bias. Figure 7 *(b)* more clearly represents that there is a positive bias. Note that the expected value of SMS of the sampled data is greater than that of the population data set in both cases.



(a) Space Mean Speed of Sampled Data (1 Simulation)



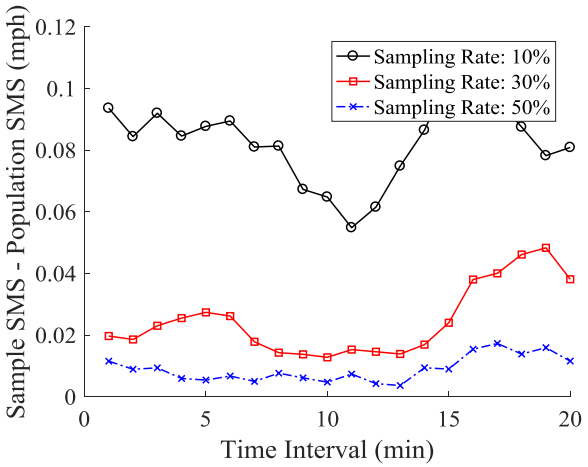
(b) Expected Space Mean Speed of Sampled Data (100 Simulations)

Figure 7. Expected Space Mean Speed of Sampled Data (NGSIM I-80)

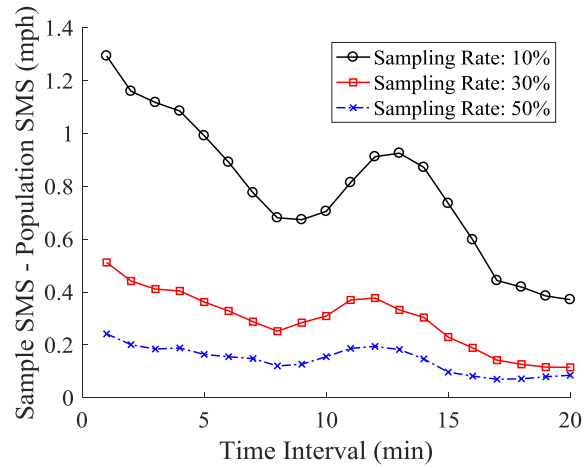
Impact of Time Interval

Figure 8 shows how the time interval could impact on the overestimation of SMS. There are three lines with different marks by three sampling rates, 10%, 30%, 50%. To see the impact of only time interval, the segments for sites were set to be a complete section of collected data, i.e., one segment for each site. Overall, the sampling bias tends to be larger where the time interval is relatively small. This is not surprising because the adjustment factors get close to 1 as population size of single time interval gets larger with longer time interval. This is consistent with the example shown in Figure 2. However, the trend of having larger bias with small time interval seems not very clear in the I-80 and US-101 data sets. This is possibly due to the limited time period of the data collection, which is less than an hour. Furthermore, having different time interval affects not only traffic counts of each time interval, but also the distribution of vehicle speeds within the time interval.

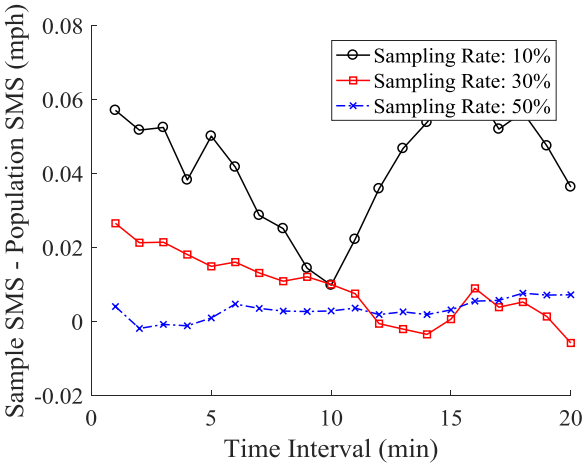
Although the data sets are very limited both temporarily and spatially, it is recognizable that the bias of SMS is also different by the sites. With 10% sampling rate and 1-minute time interval, the bias of SMS is larger than 1 mph in the Lankershim data set while the other two sites have smaller than 0.1 mph of bias.



(a) NGSIM I-80 Data



(b) NGSIM Lankershim Data



(c) NGSIM US-101 Data

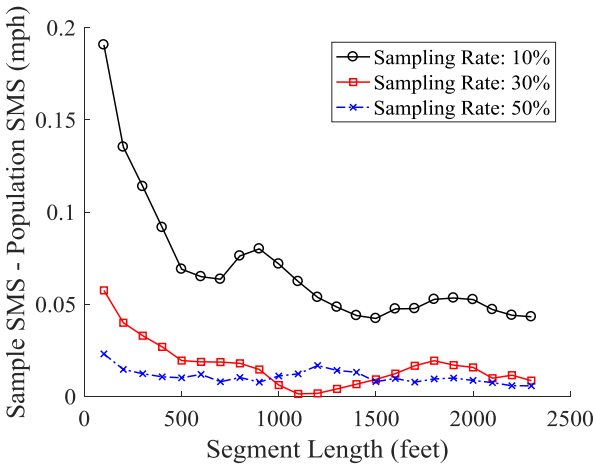
Figure 8. SMS Overestimation by Time Interval (3 NGSIM Data Set)

Impact of Segment Length

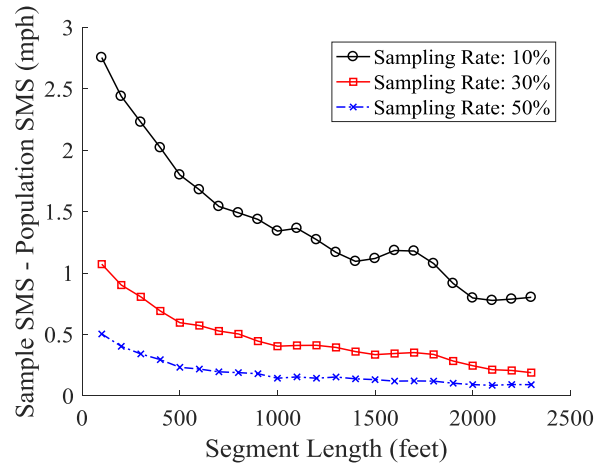
While Figure 8 shows the impact of overestimation by time interval, Figure 9 represents the impact of segment length, with controlling the time interval of 5 minute. Since NGSIM data sets have no definite segment information but actual trajectories of individual vehicles, the segmentation was done virtually by the vehicle travel distances and a fixed segment length. For the each virtually generated segment, individual vehicle speed is obtained by travel distance over travel time within the segment.

Overall, the impact of bias of sample SMS gets smaller as the segment length increases. This trend was expected since longer segment will likely to have less variation of speed if traffic condition of roadway is consistent. Like the results in the impact of time interval, the biases in Lankershim are larger than the other two sites. With 100ft segment length and 10% sampling rate, the SMS of sample data is overestimated by almost 3 mph.

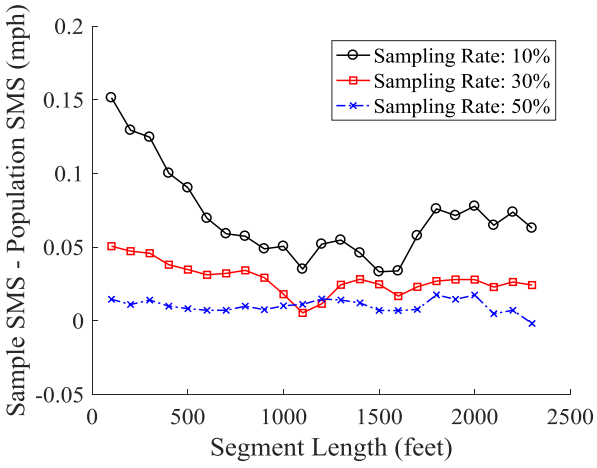
Note that the results are based on very limited vehicle trajectory data. Investigating those trajectory data to see the impact of sampling bias by sampling rate, time interval, and segment length, could help traffic operation agencies on determining technologies and data sources to produce reliable traffic speed information. However, trajectory data are often unavailable yet although new traffic data sources, such as GPS on mobile phone, will become more accessible to operators and users.



(a) NGSIM I-80 Data



(b) NGSIM Lankershim Data



(c) NGSIM US-101 Data

Figure 9. SMS Overestimation by Segment Length (3 NGSIM Data Set)

Impact of Bias in Practice: Multiple Segments in Network

This section investigates the impact of sampling bias using typical travel speed data that are aggregated by segments in network.

Data Description

Tennessee Department of Transportation (TDOT) has deployed Remote Traffic Microwave Sensors (RTMS) to collect traffic count, speed, and occupancy every 30 seconds, in Tennessee. The data used in this study cover 204 stations in region 1 of Tennessee, as shown in Figure 10 (a), which is mainly Knoxville and its sub-urban area, from August 2016 to November 2016.

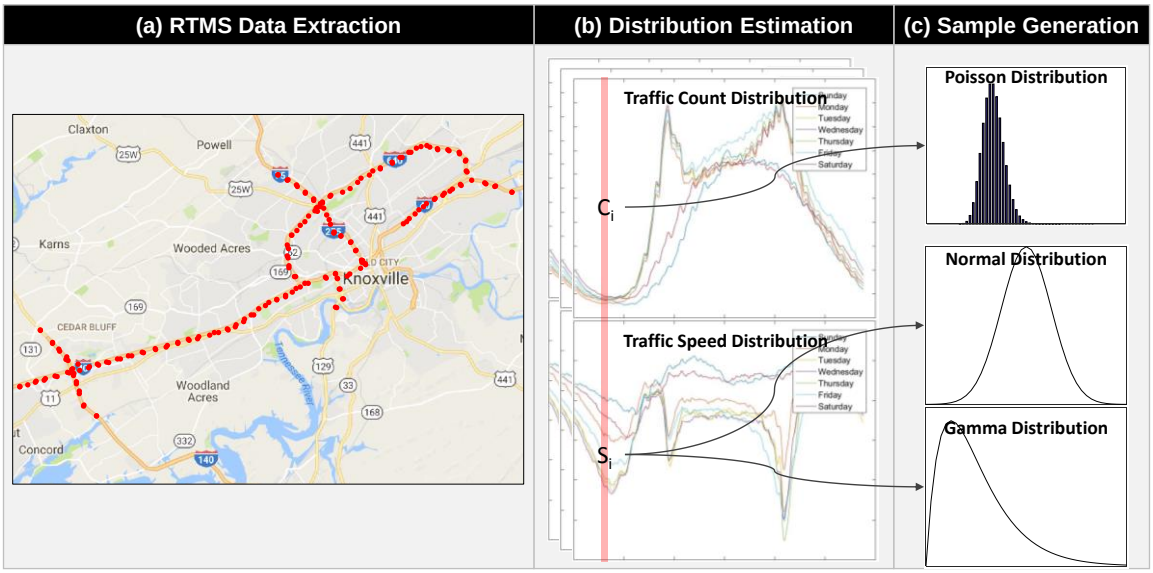


Figure 10. Sample Data Generation Procedure for RTMS Data, TN Region 1

Based on three time period sets (5, 15, and 30 minutes), the 30-seconds RTMS data were re-organized by station and day. For each time stamp, the distribution of traffic count and speed were estimated using Maximum Likelihood Estimates (MLE). As shown in Figure 10, the traffic count was assumed to follow Poisson

distribution, and the speed data was assumed to be following either normal or gamma distribution. The distribution type of speed, among the two, was selected by Bozdogan's Information Complexity [37]. Using the defined distributions with their parameter estimates, traffic count and speed of each in population data were generated. Then, sample data (or collected data) were captured by sampling rate from 0.01% to 100%, to calculate the difference between SMS of the sample data and the generated population data.

Impact of Sampling Rate in Segmented Network

Figure 11 shows that the sampling bias gets larger as the sampling rate decreases. The impact of sampling bias is more significant when the time interval is small, but the difference of the impact between the time intervals is relatively smaller than that of sampling rate. With sampling rate of 40%, the average sampling bias of SMS in this network is less than 0.2 mph.

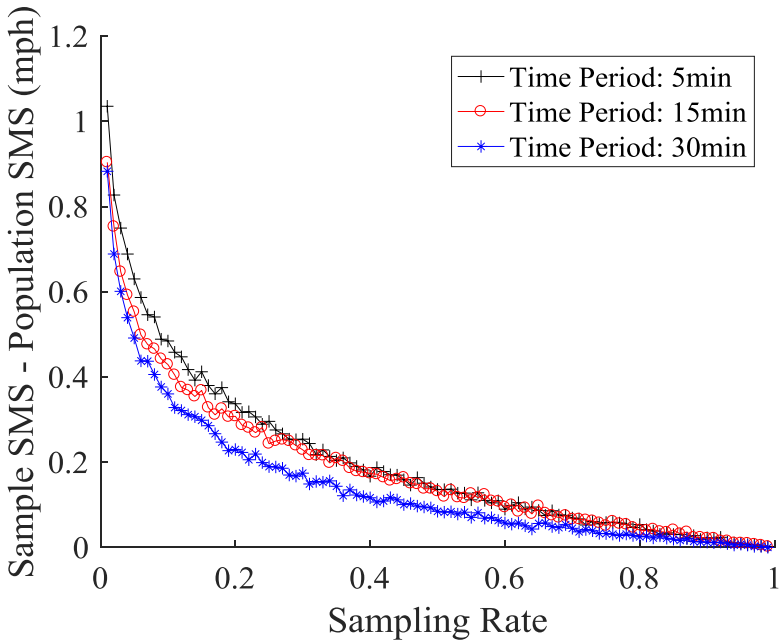


Figure 11. Impact of Sampling Bias on Region 1, Tennessee

Impact of Bias in Practice: Data Comparison

This section analyzes the impact of sampling bias where multiple data sources are compared.

Data Description

To evaluate accuracy of real-time speed data, Hargrove et al. collected traffic data in Nashville, TN and compared four different data sources, Bluetooth, RTMS, and two private traffic information providers, by using License Plate Recognition (LPR) data as ground truth [38]. Table 2 summarizes the data sources used in their study.

Table 2 Summary of Selected Traffic Data Sources

	Bluetooth	Data Provider 1	Data Provider 2	RTMS
Data Type	Time, Signal Strength	Speed, Travel Time	Speed, Travel Time	Volume, Occupancy, Speed, Vehicle Classification
Aggregation & Time Resolution	Each MAC Address, All Lanes	60-sec, All Lanes	60-sec, All Lanes	30-sec, per Lane
Data Source	Cellular and in-vehicle Bluetooth devices	State installed sensors, probe vehicles, GPS, cellphone	State installed sensors, probe vehicles, GPS.	Roadside detectors
Accuracy Checks Performed	Post collection processing with filters.	Independently verified in large-scale testing.	Data checks prior to map matching.	Post collection processing with filters.

(Source: Hargrove et al., *Empirical Evaluation of the Accuracy of Technologies for Measuring Average Speed in Real Time, 2016* [38])

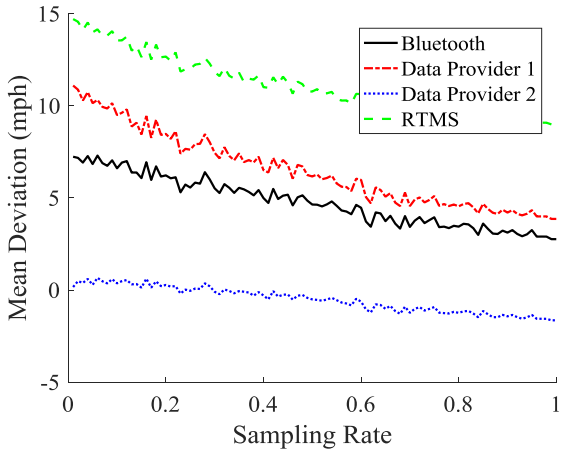
In this study, the same data set was used to see the impact of sampling rate on the comparison results. The challenge here is that LPR technology does not capture all vehicles and the license plates read by the technology is not

completely accurate, which also cause reducing the sampling rate. Although their enhanced LPR matching algorithm improved the matching ratio significantly, 98% matching rate with less than 1% of false matching, number of matches from LPR data are only about 20% of traffic count from RTMS.

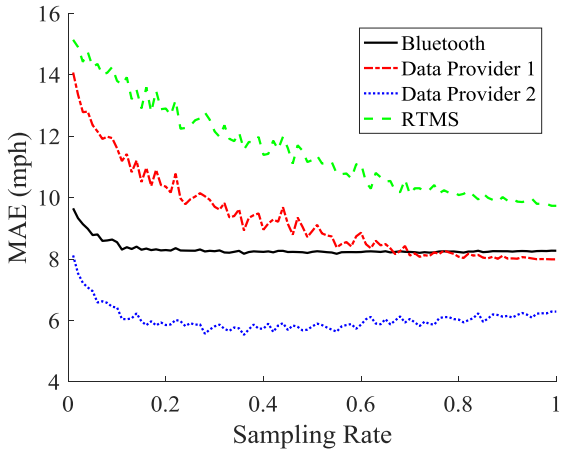
Impact on Comparison of Data Sources

Figure 12 shows the impact of sampling bias on three performance measurements of data accuracy, mean deviation, mean absolute error, and root mean square error. The three accuracy measurements are decreased as the sampling rate is increased, except mean absolute error and root mean square error of the Data Provider 2.

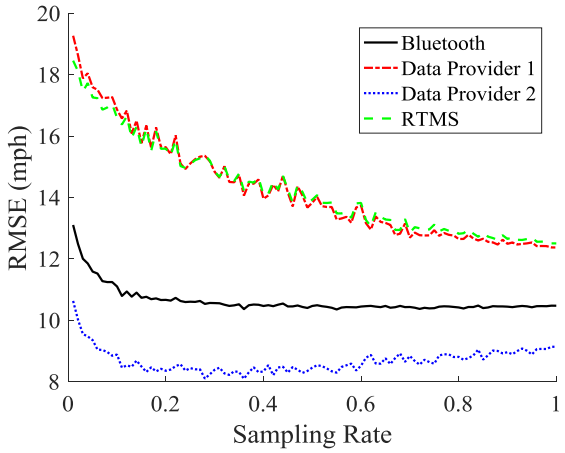
Note that there is a crossing point in mean absolute error, Figure 12 (b), between Bluetooth and Data Provider 1. This implies that the sampling rate could be critical to determine a more accurate data source for some cases, although the changes of performance measurements may seem ignorable for other cases. In practical data comparison or evaluation, it is important to know that low sampling rate could bring not only increased variation of performance measurements, but also the bias of the measurements.



(a) Mean Deviation



(b) Mean Absolute Error



(c) Root Mean Square Error

Figure 12. The Impact of Overestimation on Comparison of Data Sources

Conclusion

Our study shows that the expected value of harmonic mean decreases as the sample size increases, regardless of its distribution. In other words, the harmonic mean of population is overestimated when we have the sample smaller than the entire population. This indicates that using the harmonic mean with sample data needs extra cautions. Especially for the traffic data, the sample size implies not only number of vehicles, but also segment lengths. Thus, to calculate the SMS of a segment, the collected data should cover enough of both the length of segment and the number of vehicles passing the segment.

This study covers only the three sites to see the impact of sampling bias. Therefore, not to mention that the analysis results are limited to those sites, the impact of sampling bias in practice may be more significant, depending on their sampling rate, time interval, and segment length. Therefore, the impact of the bias needs to be investigated on a case-by-case basis.

Both the analytical correction and the simulation-based correction approach are provided. Since the analytical correction approach is limited to certain distribution types, the simulation-based correction approach is recommended for most of cases. The simulated experiments for collected data is also important to see the significance of the sampling bias impact, and the correction could be unnecessary (or the bias is ignorable) depending on purpose of its use. More importantly, it is recommended to consider using data sources with a larger sample size before considering the corrections, since a smaller sample size not only brings the sampling bias, but also increases its variations.

It is important to know that the harmonic mean is used in many places, although our study focused on the SMS. These area or examples, of where the harmonic

mean is used and therefore the sampling bias should be considered, include calculating the followings [39]:

- a fuel economy measurement, an average MPG of a group,
- an average of multiples, such as price-earnings ratio, in finance,
- an aggregated performance score for algorithms and systems in computer science,
- an average contribution per component, such as parallel resistance and parallel inductance, in electronics,
- the fluctuation effects in generation size of effective breeding population in population genetics, and
- other aggregated measurements in geometry, hydrology, sabermetrics, chemistry, and so on.

To understand and solve the issue of sampling bias of harmonic mean with small sample size, further discussions and studies are needed. This includes the impact of sampling bias on weighted harmonic mean and more empirical analysis to investigate the sampling bias of harmonic mean in the fields, which of data are not necessarily independent and identically distributed. Furthermore, there is a remaining question of how to determine required sample size considering both the sampling bias and its variation.

CHAPTER II

ESTIMATION OF ERROR DISTRIBUTION FOR MULTI-SOURCE DATA WITHOUT GROUND TRUTH DATA USING MODIFIED APPROXIMATE BAYESIAN COMPUTATION

This chapter presents a modified version of a research paper by Hyeonsup Lim, Lee D. Han, Shih Miao Chin, and Ho-ling Hwang.

Abstract

One of the challenges in measuring accuracy of multi-source data, before this study, is a requirement of ground truth data (or baseline data), since the accuracy of each data source is defined as the difference between the truth and the measurements of the data source. Determining the ground truth data source is another challenge since measuring the accuracy of the ground truth involves additional requirement of more accurate baseline data. This study proposes a methodology to estimate error distributions of data sources by aggregating measurements from multi-source data. Approximate Bayesian Computation was adopted and modified to construct the error distribution based on simulations. In the simulated experiment, the proposed model outperformed the alternative approach, which is a conventional way of evaluating data source that is gathering error information by using the benchmark data. The sensitivity analysis is also provided to explore the model performance by sample size, number of data sources, and distribution types. The proposed model is limited to one dimensional variable with an assumption of independence between the data sources, but the basic approach provided in this study might be easily expanded in other applications.

Introduction

With technology advancements, data are exploding in respect to its size and variety. Major traffic information providers have built a data warehouse of multiple petabytes or even more that stores minute-by-minute traffic data across millions of road segments. The large and complex data, so called “Big Data”, have brought a new era of variety of data-driven modeling and applications.

Specifically, multi-source data have brought the immense and innovative benefits in many places. Medical scientists utilize multiple sources of data, including patient records, physician reports, medical malpractice claims data, journal articles, and other databases, to identify and assess diagnostic error in medicine. In ecology, researchers have been reconciling multiple data sources to improve prediction of forest disease incidence. In addition, multiple GPS receivers are used where highly accurate location information is needed.

One of the big challenges of handling the multi-source data is how to measure the reliability and accuracy of each data source. Although the multi-source data seem to be promising the better prediction and estimation, use of inaccurate data source or improper data aggregation could erroneous estimation results and end up with misleading conclusions. Therefore, data accuracy must be considered enough and each data source should be treated carefully in the modeling and analysis, involving a decision whether the data should be included in the analysis.

The information of data accuracy is, often, not available or provided at an insufficient level. Sometimes, the accuracy of data may not be important for their originated purposes. Even if they do provide some measurements, the information could be based on highly controlled conditions such as laboratory

experiments. For instance, actual gas mileage reported by drivers is often different than the MPG provided by car manufacturers. Moreover, organizations may require concealing data and the data accuracy for confidentiality reasons. As a result, the prior knowledge of the data accuracy information might be inaccurate.

Even if we all agree on that we must provide reliable accuracy information, measuring the accuracy is still a big challenge. Even before all of discussions of post-processing of data for measuring the accuracy, such as filtering, aggregating, and smoothing, one of critical point of measuring the accuracy is a requirement of ground-truth data (or baseline data), since the accuracy of each data source is defined as the difference between the ground-truth and the compared data source. Therefore, the error of ground-truth data should be near zero or within certain criteria. However, the resources to obtain the ground-truth data are often expensive.

When we have enough data sources that are not extremely biased to one-side in average, the distribution of ground-truth can be estimated based on the prior knowledge of accuracy of each data source even if some of information is wrong. This is similar to applying a democratic decision making in a sense that we believe that the decision made by majority vote is better or at least fair. The difference is that we listen to all of opinions, make a combined decision, and remember who were wrong and how much. The recorded information is used to estimate how much wrong that person could be in a next discussion, as well as making a better decision.

The objective of this study is to estimate an error distribution of multi-source data in one dimension when we do not have the ground truth data. We assume that each data source works independently. This study utilizes multi-source data to estimate the distribution of ground-truth and then estimate the error distribution of

each data source. The proposed algorithm, a modified Approximate Bayesian Computation, starts with the prior knowledge of each data source, but keeps updating the knowledge as more data are gathered with a simulated set of ground-truth.

Literature Review

Most of studies in transportation evaluating data accuracy, traditionally, have adopted a third source of data as ground truth or reference point, assuming that the measurement error of the data is close to zero. To measure the accuracy of speed (or travel time) data, previous studies used different types of sensors as a benchmark data source, that is assumed to be the ground truth data in their analysis. The typical data sources include License Plate Recognition (LPR) [40-44], probe vehicle[45, 46], Bluetooth[47-52], and Radio Frequency Identification (RFID) [53, 54]. Ribeiro et al. have used cartography data as a ground truth source to compare the accuracy of several alternate low-cost methods, Google Earth, a digital inclinometer, and a laser distance measurer, to measure road gradient for cycling infrastructures [55]. They concluded that any of the tools presented in the study was unreliable as a single source, but a combination of the three tools could be useful to conduct a preliminary assessment of the geomorphologic suitability and audit the urban road environment for pedestrians and cyclists.

The Bayesian theorem has been applied to combine multi-source data in many area, including signal/sensor data [56-58], spatial data [59-61], and audio/visual data [62-64]. Likewise, there has been a proliferation of Bayesian theorem based studies and applications in transportation field: traffic flow forecasting [65-70], travel time/speed estimation [71-76], and traffic crash analysis [77-82].

Specifically, Choi et al. proposed a data fusion algorithm using Bayesian polling technique [73].

The typical incarnation of Bayesian theorem can be written as Equation 42, where θ denotes a particular parameter value given data D [83].

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{34}$$

In Equation 42, the likelihood $P(D|\theta)$ defines the probability of the observed data under the particular statistical model parameter value θ. The likelihood, typically, can be calculated from an analytical formula for simple models. However, the analytical formula for complex models is often elusive.

Approximate Bayesian Computation (ABC), proposed by Mark et al., overcomes this issue by approximating the likelihood with systematic simulations, where the likelihood function is analytically intractable [84]. Actually, before the term ABC was established by Mark et al., the idea of ABC started in 1980s. Diggle and Gratton, in 1984, proposed a simulation based method to approximate the likelihood by defining a grid of parameter space and simulating each grid point [85]. Also, Donald stated a hypothetical sampling mechanism, which coincides with the ABC-rejection scheme [86].

While the Diggle and Gratton's approach aimed at approximating the likelihood rather than the posterior, Tavaré et al. described computational methods, i.e., ABC algorithm, for posterior inference of the coalescence time (time since the most recent common ancestor) of DNA sequence data [87]. Toni et al. combined ABC and Sequential Monte Carlo (SMC) to estimate parameters of dynamic models and provide a better statistical inference of the model parameters and its sensitivity [88]. The proposed model can also be used for the standard Bayesian

model selection. Furthermore, Mark et al. specifically described the ABC-approach and its suitability for problems in population genetics, and the application of ABC has spread to epidemiology, systems biology, and etc. [89]. However, based on our literature review, the authors could not find any study using the ABC approach in transportation field.

As shown in Figure 13, ABC performs numerous simulations based on prior distribution of model parameter value, and compares the summary statistic of simulated data with the observed data to determine acceptance/rejection of each simulation [89]. Since the probability that simulated data exactly coincides with the observed data is extremely low in most of cases, the rejection rule should not be too strict. Finally, the posterior distribution of model parameter θ can be obtained from the accepted simulations.

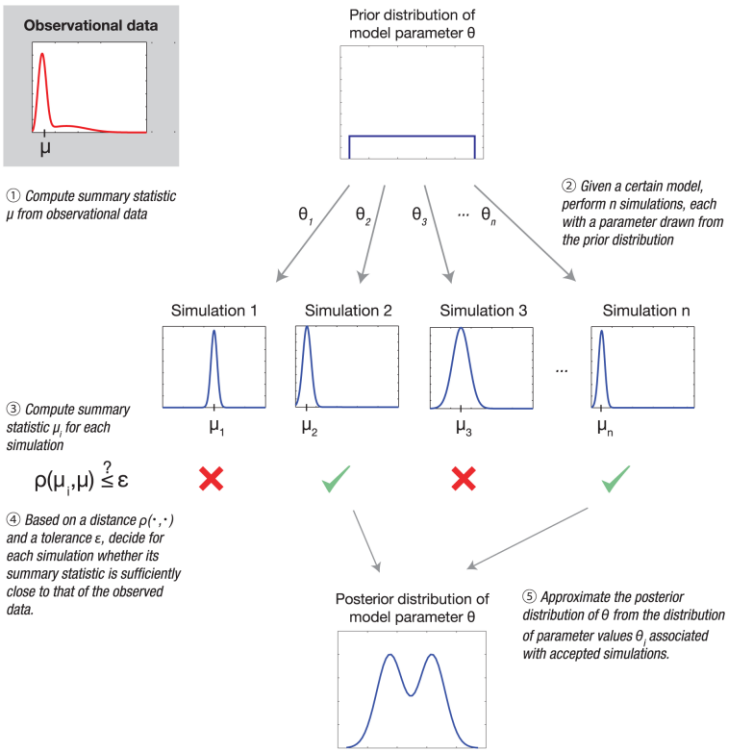


Figure 13. Parameter Estimation by Approximate Bayesian Computation
(Source: Mark et al., *Approximate Bayesian Computation, 2013* [89])

Methodology

Modified Approximate Bayesian Computation

Suppose that there are three data sources, a, b, and c, and their observations, x_a, x_b, and x_c. Unlike the aforementioned ABC, our concern is to know the error distribution of each data source. If we know the truth x_T, the error of each data source for this observation will be $x_a - x_T$, $x_b - x_T$, and $x_c - x_T$.

$$P(Err(x_a) = x_a - x_T | x_a, x_b, x_c) = P(x_T | x_a, x_b, x_c) \tag{35}$$

However, the ground truth x_T is often unavailable or very costly to obtain, and hence a set of candidates of ground truth $\hat{x}_{T1}, \dots, \hat{x}_{Tn}$ is constructed. This makes our interim target to estimate the probability of each candidate of ground truth.

$$P(\hat{x}_{Tn} | x_a, x_b, x_c) = \frac{P(x_a, x_b, x_c | \hat{x}_{Tn}) P(\hat{x}_{Tn})}{P(x_a, x_b, x_c)} \tag{36}$$

Since the relative scale of probability should be obtained for updating the error distribution, the probability of the observation $P(x_a, x_b, x_c)$ and the probability of each ground truth candidate $P(\hat{x}_{Tn})$ can be ignored if we construct an enough set of candidates of ground truth, i.e., almost all possible values of truth and the observation of each data source is independent.

$$P(\hat{x}_{Tn} | x_a, x_b, x_c) \propto P(x_a, x_b, x_c | \hat{x}_{Tn}) \tag{37}$$

$$\begin{aligned} P(\hat{x}_{Tn} | x_a, x_b, x_c) &= \alpha \cdot P(x_a | \hat{x}_{Tn}) \cdot P(x_b | \hat{x}_{Tn}) \cdot P(x_c | \hat{x}_{Tn}) \\ &= \alpha \cdot P(Err(x_a)) \cdot P(Err(x_b)) \cdot P(Err(x_c)) \end{aligned} \tag{38}$$

Then, the estimated error distribution of each data source for the observation can be calculated from the probability of the set of candidates of ground truth. For instance, the error distribution of data source a for the observation is shown in Equation 47.

$$P(\widehat{Err}(x_a) = x_a - \hat{x}_{Tn} | x_a, x_b, x_c) = P(\hat{x}_{Tn} | x_a, x_b, x_c) \quad (39)$$

Figure 14 describes the procedure of proposed algorithm to estimate the error distribution of each data source. First, before running the simulation, the observed data from multiple sources and the prior knowledge of error distribution of each data source will be gathered to calculate the likelihood.

The key point of the proposed algorithm is constructing a candidate set of true data points $\hat{x}_{Tn}$. This is similar to model parameter value to be estimated in ABC described in the previous chapter, but no rejection criterion is used because the candidate set should cover most of possible points for ground truth. For each simulated ground truth point, the likelihood will be calculated based on the observed data and the prior error distribution of each data source. As a result, a set of likelihoods will be obtained for each observation.

For each data source, the difference between the observed value and each data point in the set of ground truth represents the error of the data source. Therefore, a set of each data source error value, with the its estimated likelihood, will also be obtained, associated with the set of ground truth. Finally, the posterior distribution, which is the error distribution of each data source for the observation, will be used to update our knowledge on the error distribution of the data sources for upcoming observations.

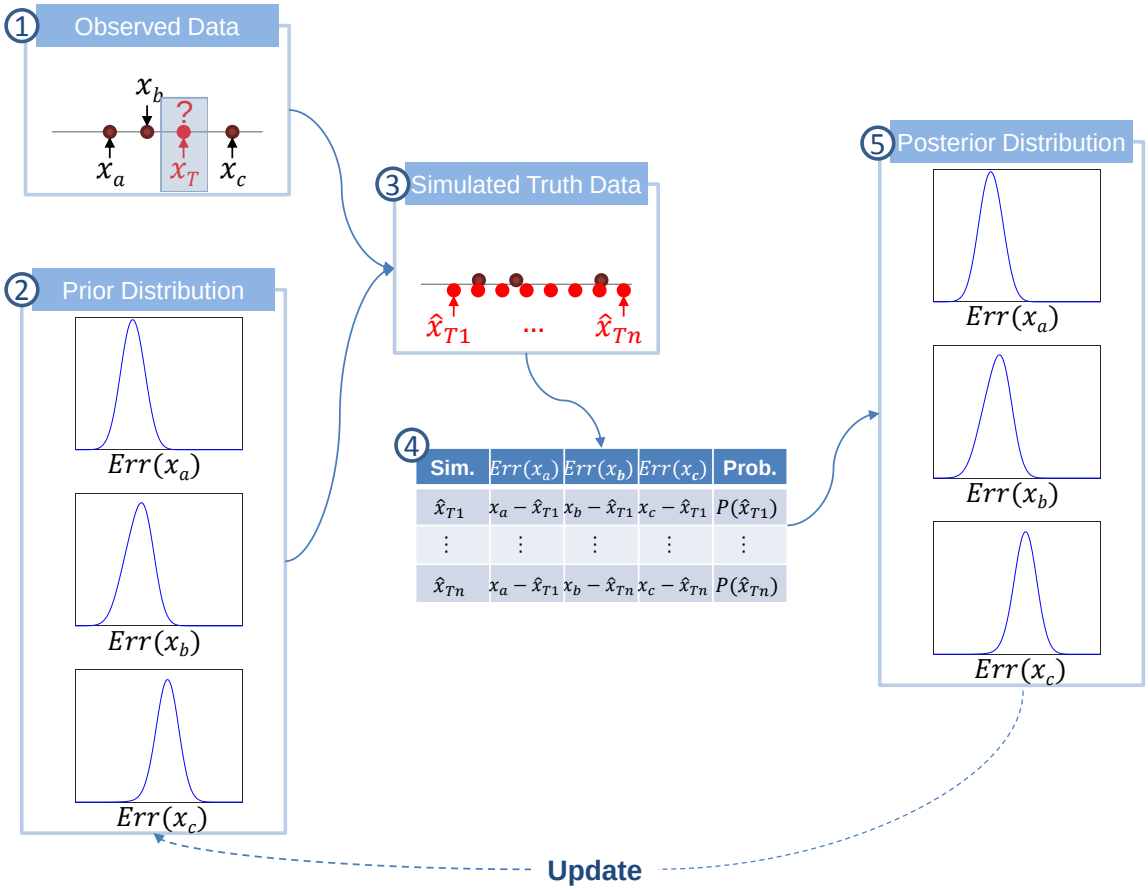


Figure 14. Modified Approximate Bayesian Computation (Proposed Model)

Simulation Procedure

To accurately assess the performance of the proposed model, the ground truth data and the true error distributions of data sources must be known. Such information is, generally, not obtainable in practice. Therefore, to investigate theoretical aspects of usefulness and limitations of the proposed approach, a simulation-based case study was conducted rather than observed data in field. In the simulated experiment, the ground truth and the true error distribution of data sources are generated by defined conditions, and therefore controlled. The overall procedures of the simulated experiment and their analysis results are described in the following sections.

Case Description

To explicitly demonstrate the approach of this study and make it easy to understand, the study uses a simple case that there are three detectors measuring a distance between two target objects. The detectors have prior information of their error distribution, which may or may not be accurate. The following assumptions are made for this case of simulation:

- The measured distance by each detector has an error due to only the defined distribution of the detector error, which is independent to any endogenous or exogenous factors including other detectors.
- The distribution of detector error remains unchanged during the experiment.
- The detector error follows a gamma distribution.

The distance measurer procures the distance between the target objects. Note that the distance measurement described here does not include a direction, and

hence it is one dimension. The true distance is defined as the actual distance between the target locations while the error is zero. The detector error represents the difference between the true distance and measured distance. The procedure 1 and 2 describes how the detector errors and ground truth data set was constructed, and the estimation is conducted by repeating the procedure 3 to 6, which is also associated to the steps in Figure 14.

Procedure 1: Generating Distribution of Detector Errors

The first procedure of simulation is generating the detector errors. It is assumed that each detector works independently and does not affect the accuracy of other detectors. It is also assumed that the detector error is independent with magnitude of the true distance. This may not be realistic since distance measurer could be designed for short or long range of distance and perform better on those ranges.

Procedure 2: Generating Ground Truth Data Set and Observed Data

The distance can be measured without knowing the location of objects, although it could be necessary in some practical applications. Therefore, the simulation generates the set of distances between the objects as ground truth data, without concerning the movements of objects. Accordingly, the observed data for each detector is generated based on the distribution of detector errors defined in Procedure 1. Assuming no missing observation, the number of observations will be the number of ground truth data multiplied by the number of detectors (data sources). Then, the rest of procedures assume that only the observed data, but no ground truth data, is available.

Procedure 3: Constructing a Set of Candidates for Ground Truth

Procedure 3 begins the proposed algorithm to estimate the distribution of multi-source detector errors. For each observation with multiple measurements of the detectors, a set of candidates for ground truth is constructed based on the observation. The candidate set will cover a certain confidence interval of the true distance, which is estimated based on the prior knowledge of detector errors. Then, a vector of n, the number of candidates of ground truth for each observation, evenly spaced points in the feasible range will be generated.

Procedure 4: Calculating Probability

Now, the observation is associated with each of candidate ground truth points. For each detector, the estimated error of the detector with the candidate point is a difference between the observed value and the candidate value. The probability of obtaining the observation for the detector can be obtained from the prior knowledge of distribution of detector error. Then, the likelihood of the true distance is equal to the candidate value can be calculated by multiplying the probability of obtaining the observation of each detector. This calculation is repeated for all set of candidates of ground truth points.

Procedure 5: Estimating a Distribution of Ground Truth

After calculating the probability of each candidate points, the distribution of ground truth is estimated by kernel density estimate, which is a non-parametric method to estimate the probability density function. The calculated probability of each candidate points is used as a weighting factor. The point estimate of the true value can also be obtained by using maximum likelihood estimate (MLE).

Procedure 6: Updating an Error Distribution of Each Detector Error

If the true distance is known, the error of each detector can simply be calculated by subtracting the observed value from the true value. Therefore, estimating a distribution of ground truth is identical to obtaining a distribution of detector error for the observation, and the output of estimated distribution of ground truth will be directly used to update a distribution of each detector error. In the simulated experiment, the distribution of detector error updated after the distributions of ground truth data are estimated for all observations. This can be modified to update the error distribution more frequently in real-time operations. Then, Procedures 3 to 6 are repeated until the difference of estimated distribution of detector errors between the previous iteration and the current iteration is very small.

Procedure 7: Adjusting Estimated Error Distribution over Iteration

Probability calculated in Procedure 4 can be obtained by log-sum of the probability of the estimated error of each detector. At the end, to update the estimated error distribution, the log-sum should be re-converted to a probability by taking an exponential. In this calculation, multiplication of very small probabilities, i.e., a large negative number of the log-sum, could be calculated as zero in a computer, due to its limited decimal fractions. This might be encountered more seriously as number of data sources is increased, which means more number of multiplications. The number of iterations could also affect increasing this issue, although it might not be a critical issue where those small probabilities are ignorable. To encounter the zero probability issues, a minimum positive probability among the candidate set is added when probability of zero exists.

Results

Example of Estimated Error Distribution

Figure 15 compares the estimated error distribution and true error distribution for one data source, over the iteration. As the iteration increases, the line of estimated error distribution gets closer to the true error distribution. The estimated error distribution could capture that the true error distribution is slightly skewed to right and their variance, as well as the mean of error. This is very important point to show flexibility of the proposed model, because it does not specify a certain type of distribution. This flexibility comes from the approach of the proposed model heavily rely on the collected data sets. With this power of estimating error distribution, the proposed method can also be used to estimate the distribution of ground truth for each data point, if the error distribution of detector is sufficiently learned.

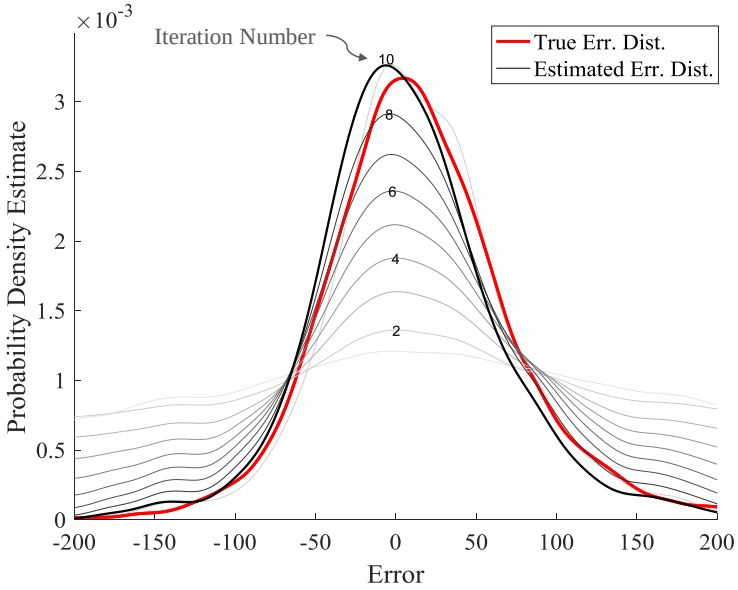


Figure 15. Example of Estimated Error Distribution over the Iteration

Evaluation Criteria

The Bhattacharyya distance, which is a measurement to account a similarity between two probability distributions, was used to evaluate the model performance [90]. The Bhattacharyya distance is calculated by taking a negative logarithm of Bhattacharyya coefficient, which approximately measures the amount of overlap between two statistical samples, as described in Equation 40.

$$D_B(p,q) = -\ln(\mathrm{BC}(p,q))$$
$$\mathrm{BC}(p,q) = \int \sqrt{p(x)q(x)}\,dx \tag{40}$$

where

$D_B(p,q)$ is the Bhattacharyya distance between two distributions, p and q,

$\mathrm{BC}(p,q)$ is the Bhattacharyya coefficient of the two distributions.

To visually provide inferences of the Bhattacharyya distance on this study, Figure 16 displays four examples of estimated error distributions along with their Bhattacharyya distances and Bhattacharyya coefficients. For instance, the Bhattacharyya distance of estimated error distribution on the left-top of Figure 16 can be calculated as $-\ln(0.8) = 0.22$. The given examples are actual estimation results from the simulated experiments.

Impact of Number of Data Sources

To analyze the impact of number of data sources on the model performance, 100 simulation runs were conducted for each number of data source, with sample size of 300. Figure 17 represents the 90% confidence interval, i.e., an interval from 5% percentile to 95% percentile, the sample estimate, i.e., the center of the confidence interval, and the individual results of Bhattacharyya distance. As shown in Figure 17, the Bhattacharyya distance decreases as the number of data sources increases. In other words, it is expected that the estimated error

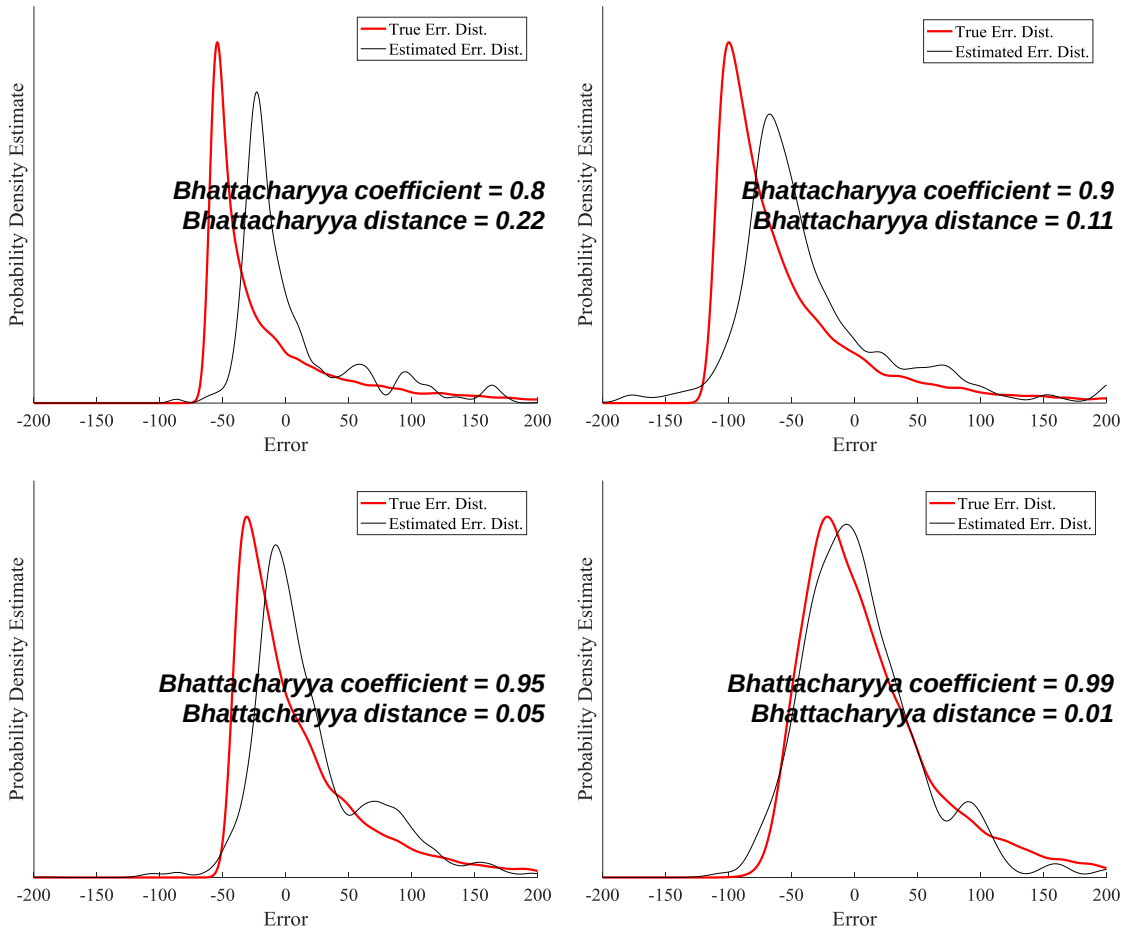


Figure 16. Example of Bhattacharyya Distance for This Study

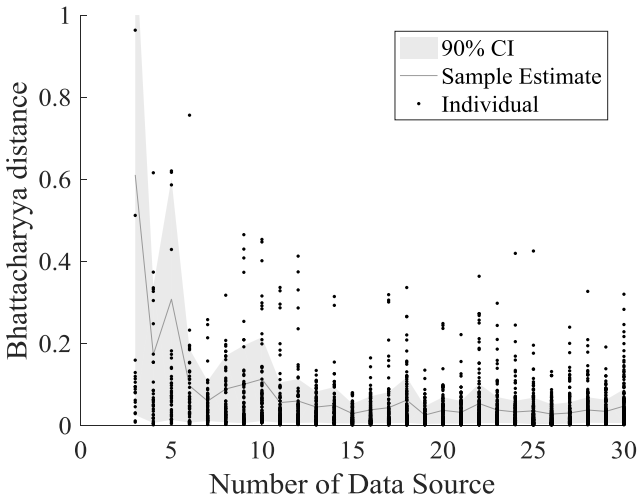


Figure 17. Bhattacharyya Distance over Number of Data Sources

distributions are likely to be closer to the true error distributions when more data sources are available. This seems intuitively reasonable since the estimated distributions of true value for each data point should become more accurate as more information is available. In this simulated experiment, more than 90% of estimated error distributions, with seven or more data sources, have the Bhattacharyya distance of less than 0.2. The marginal gain we obtain for having closer error estimation from additional data source seems to be decreasing as the number of data sources increases. In other words, a single additional data source could better improve the estimation of error distribution when only a few data sources are currently available.

Impact of Sample Size

Figure 18 illustrates the performance of the estimation of error distribution over sample size. In this impact analysis, the number of data sources was set to be 10 for all cases. It is obvious that larger sample size better improves the model performance, i.e., decreasing the Bhattacharyya distance. With sample size of less than 200, several estimated error distributions have the Bhattacharyya distance of larger than 0.8, which of the Bhattacharyya coefficient is smaller than 0.45.

Sensitivity Analysis by Distribution Type

This section was conducted to test the validity of the proposed model on different distribution types, i.e., mean, standard deviation, and skewness of the error distribution. To reflect flexible families of distributions, the Pearson system, a family of continuous probability distributions, was used with randomly generated distribution moments, i.e., mean, standard deviation, skewness, and kurtosis. The ranges of the moments are described in Table 3.

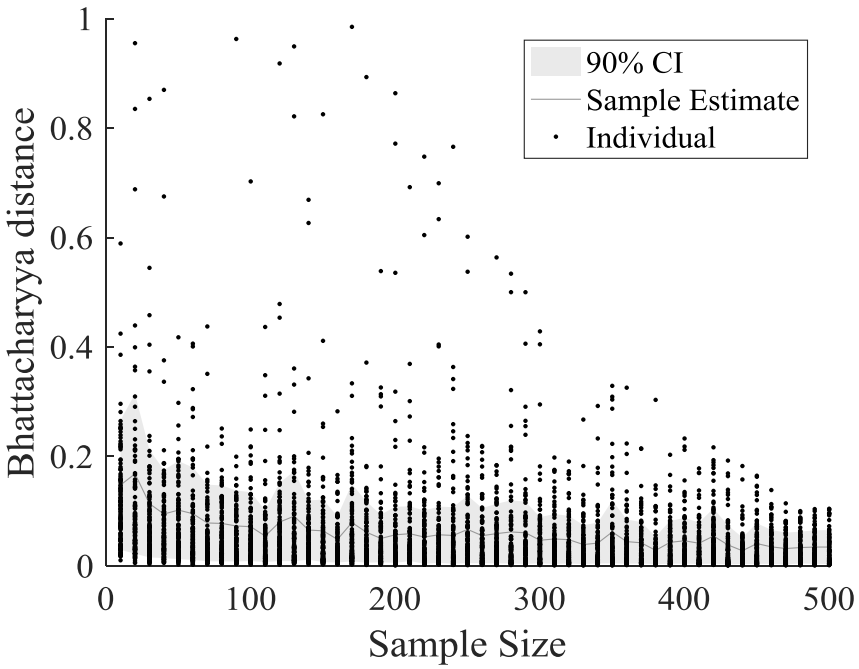


Figure 18. Bhattacharyya Distance over Sample Size

Table 3 Moments of Distributions in the Simulated Experiment

	Mean	Standard Deviation	Skewness	Kurtosis
Min	-50	10	-5	3
Max	50	50	5	1000

In Figure 19, the model performance is displayed by mean on a horizontal axis and standard deviation on a vertical axis. To clearly show where the proposed model performed better, the Bhattacharyya coefficient, the larger the better, was displayed here instead of the Bhattacharyya distance. The two line-graphs alongside the scatter plot represent the average Bhattacharyya coefficient. As shown in Figure 19, the model performance seems to be more sensitive to standard deviation than mean. As the standard deviation decreases, the Bhattacharyya coefficient also decreases.

In Figure 20, the model performance is displayed by standard deviation on a horizontal axis and skewness on a vertical axis. The Bhattacharyya coefficient is higher in average where the skewness is close to zero, but the difference seems not very clear relatively, as compared to the impact of standard deviation. Overall, the proposed model better estimates error distributions than the alternative by more than 84%.

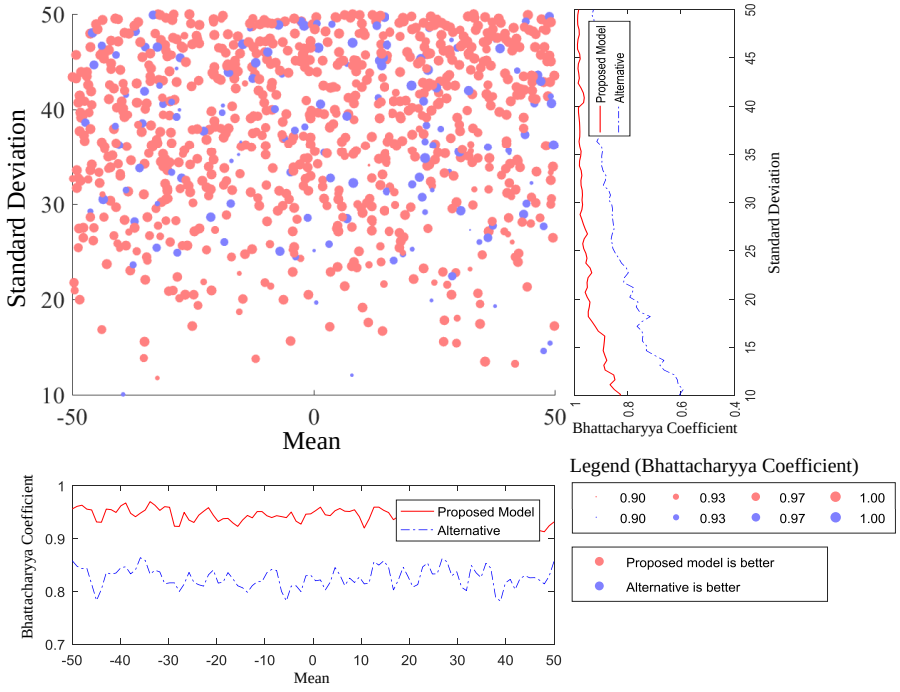


Figure 19. Model Performance by Mean and Standard Deviation

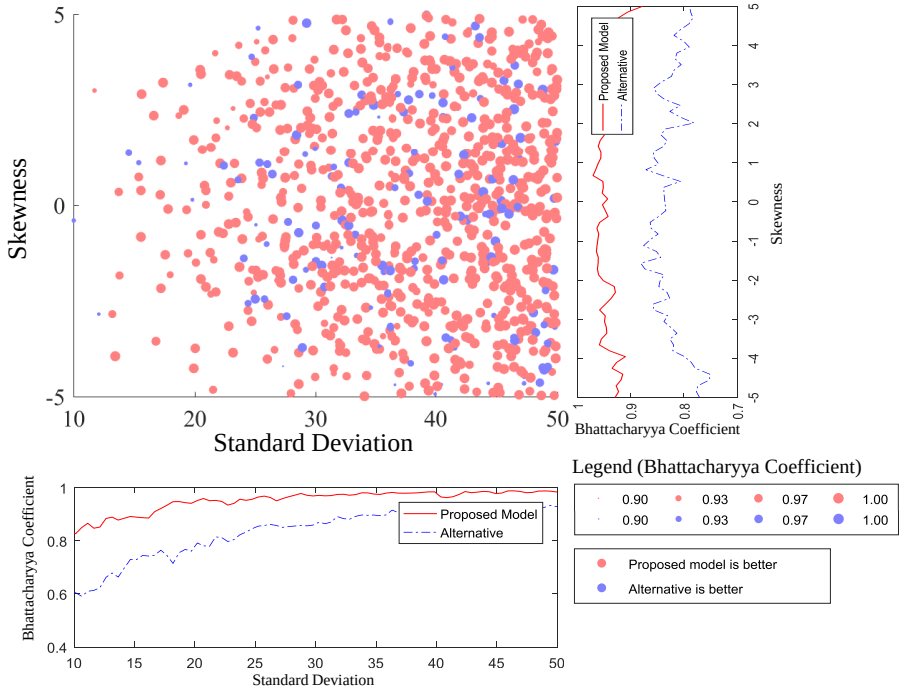


Figure 20. Model Performance by Standard Deviation and Skewness

Parameter Estimates of Error Distribution

Although this study mainly focused on estimating error distribution itself, the proposed model could be used to estimate parameters of distributions, e.g., mean and standard deviation. In Figure 21, the mean of estimated error distribution was compared with the mean of the true error distribution. The estimated mean from the proposed model is much closer to the true mean of error, as compared to the alternative.

Likewise, Figure 22 is a Q-Q plot of the estimated standard deviation of error versus the true standard deviation of error. The estimated standard deviation of error by alternative approach tends to be larger than the true standard deviation, while the estimated standard deviation by the proposed model is relatively closer to the true standard deviation. This is because using a single source of benchmark data among the data sources brings extra variations to estimated true values by its own variation, even if the benchmark data source is unbiased.

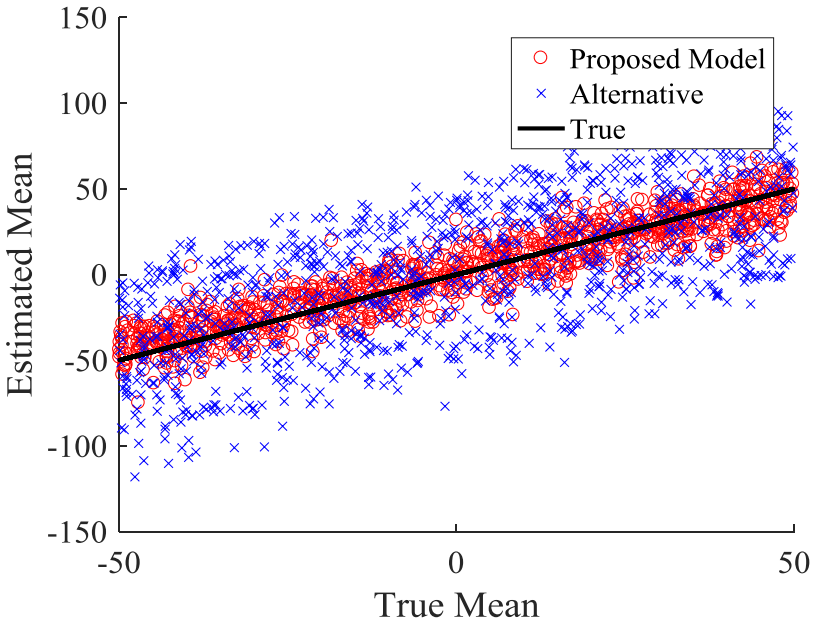


Figure 21. Q-Q Plot of Estimated Mean of Error

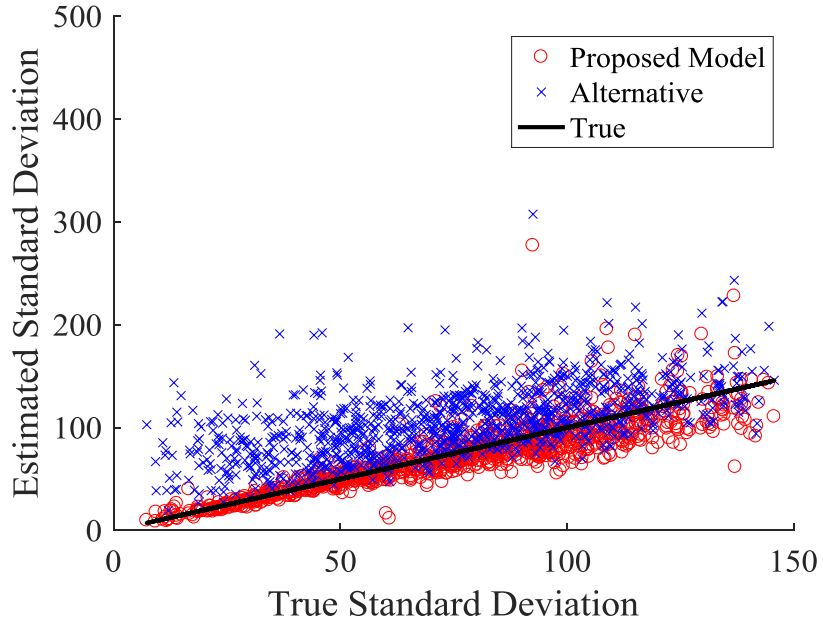


Figure 22. Q-Q Plot of Estimated Standard Deviation of Error

Conclusion

This study enabled estimating error distribution of data source without the need of ground truth data, by incorporating modified Approximate Bayesian Computation. In the simulated experiment, the results show that the proposed model outperforms the alternative approach, which is a conventional way of evaluating data source that is gathering error information by using the benchmark data. The sensitivity of the model performance was conducted by sample size, number of data sources, and distribution types.

The key benefits of proposed model include the followings:

- The estimation process does not require a process of determining the benchmark (or ground truth) data source,
- Error estimation of the proposed model does not require parametric estimates,
- Given prior knowledge of each data source might be useful to improve the model performance, but the proposed model still can be applied without the prior knowledge,
- Final output of the proposed model entails the actual shape of error distribution, which enables user to visually analyze the error distribution and make their own decisions, in addition to estimated parameter estimates such as mean and standard deviation of the errors.

The proposed model was evaluated on the simulated experiment of distance measurements, which is a one-dimensional continuous variable. In practice, it is expected that the proposed approach can be applied in various places. For instance, when travel speeds are gathered simultaneously from many vehicles, the vehicle speed distribution of individual can be estimated using the proposed approach. In this case, the estimated distributions represent the characteristics of

individual, not the errors. Furthermore, the estimation of error distribution can be applied to determining a best data source, which is preferred to be unbiased with less variation of error.

To validate and expand applications of the proposed model, further discussions and studies are needed, including, but not limited to:

- Additional simulation-based experiments with two-dimensional problems and different numbers of data sources,
- Consideration of estimating the error distributions where the accuracy of data sources is associated with other factors,
- Considering weighted estimation using a prior knowledge where a certain data source is known to be more accurate,
- Building an association matrix or chain to consider error distributions of where variables are not continuous, i.e., discrete, ordinal, nominal, or other non-continuous types,
- Enhancing the estimation speed, e.g., modifying number of iterations or number of candidates, to apply the proposed model in real-time operation

CHAPTER III

ENHANCING ACCURACY OF POSITION AND DISTANCE MEASUREMENTS FOR CONNECTED VEHICLES BASED ON MODIFIED APPROXIMATE BAYESIAN COMPUTATION APPROACH

This chapter presents a modified version of a research paper by Hyeonsup Lim, Lee D. Han, Shih Miao Chin, and Ho-ling Hwang.

Abstract

Accurate positioning of vehicles is a critical element of autonomous and connected vehicle systems. Most of other studies heavily focused on enhancing simultaneous localization and mapping (SLAM) methods, i.e., computationally constructing or updating a map of an unknown environment and tracking an object within the map. This paper provides a method that can, in addition to existing SLAM or relevant methods, enhance the raw measurements of position and distance and therefore. The basic idea of this study is to identify and update error distribution of multi-source raw data measurements by combining all available information. A modified Approximate Bayesian Computation method was incorporated. The estimation is conducted real-time based, and the learning process will try to keep improving the accuracy of estimation. The results show that the proposed model noticeably improves the accuracy of position and distance measurements. The estimated error distribution can also be used for improving results of other post-processing techniques which require assumptions of certain type of error distributions. A similar approach can also be utilized to enhance accuracy of other sensors or measurements in connected vehicle or relevant systems, where multi-data sources are available.

Introduction

Autonomous Vehicles (AVs) and/or Connected Vehicles (CVs) will become available in near future, most likely will be both as CVs will add cost of only several hundred dollars or less [91]. There are a variety of types of sensors that could be embedded to AVs and CVs, but one of inevitable functionality is positioning and measuring distances between vehicles and objects. The sensors and their measurements will affect the accuracy and the reliability of overall connected vehicle systems, potentially to road safety as well.

This importance of the measurement accuracy brought attention of industry and government agencies that the sensor error needs to be allowed in a certain range. Although there are regulations that define range of errors allowed in application, how we can measure the errors is a remaining question. To measure the error, we often conduct experiments in limited circumstances. Of course, a wide range of experiments and a benchmark data sources with high precision will lead to more trustful results. Then, how can we be sure they are enough or not, with consideration of additional resources needed?

Another challenge is determining the benchmark source to evaluate the designated measurements or data source. This requires a method to evaluate the benchmark data source or at least to justify why it is used as an alternative of the ground truth. Also, even if we conducted the test & field experiments enough to define the error distribution of each data source, it may be changed over the operations in real field, due to their installation, maintenance, geographical circumstance, intervene with other sensors, and so on. These are often hard to expect accurately before the implementation.

The motivation of this paper is to identify and update error distribution of multi-source raw data measurements by combining all available information. Then the learned distribution is applied, as we operate them, to improve accuracy of the measurements in real-time AVs and CVs operations. Before this study, most of other studies heavily focused on enhancing simultaneous localization and mapping (SLAM) methods, i.e., computationally constructing or updating a map of an unknown environment and tracking an object within the map. The estimation of error distribution is not an alternative of those existing methods, but an enhancement method that can be used in addition to other methods.

The benefit of proposed method also involves low cost of gathering information, which will be near to zero except a little bit of additional computational workloads. Potentially, it can be used to identify what caused error measurements if measurements of those other factors can also be collected. The paper made a considerable effort on applying a proposed method in connected vehicles' position and distance measurements, as providing detailed procedures that can be applied in practice. Potentially, a similar approach can also be applied where multi-source data are available.

Although this paper assumed using raw data of measurements without any localization and further post-processing methods, but the proposed method may also be used before or after other post-processing techniques. In other words, the paper represents how much the proposed method can improve raw data of position and distance measurements, and the uses of other existing smoothing and filtering techniques are encouraged to further improve the estimation results in practice. To this end, there is no comparison between the proposed method and other methods, but the improvements from raw data measurements are described.

Literature Review

In recent years, smart driving technologies, i.e., connected vehicles (CVs) and autonomous vehicles (AVs), has gotten so much attentions as a promising future that will improve our mobility and safety. NHSTA reported that these technologies may reduce car crashes drastically, almost 90% of all the crashes that are caused by human error [92].

Although there will be a significant need for discussions on policies and regulations related to CVs and AVs, we will still need to confront and address technological challenges related to different kinds of the safety and mobility issues, such as hacking, malfunctions of computer systems, and low accuracy of sensor measurements. The impact of the potential outcomes from implementing the new systems under unreliable circumstances could be far more serious than types of the car crashes with conventional human-driving systems, since vehicles in the new systems are “connected”.

Positioning sensors and telecommunications systems, cameras, and automatic transmissions, with a variety of other special sensors, navigation and security services, are the key elements for CVs and AVs [93]. It is obvious that highly accurate position and distance measurements are essential for CVs and AVs, as they will require super reliable navigation capabilities.

Global Positioning Systems (GPS), which is a constant position tracking method based on global location and time references of objects from satellites, is an indispensable element of CVs and AVs. Typically, the accuracy of position measurements of GPS-enabled smartphone, under open sky, is within 5 meter, and the accuracy can be improved by using dual-frequency receivers and/or augmentation systems [94]. Recently, centimeter-accurate (or even millimeter-

accurate) GPS has been proposed and developed with its increasing demand, but they typically require additional modules or infrastructures, which are used for local referencing points to improve the accuracy, and thereby extra costs are necessary [95-98].

Another key sensor of CVs and AVs is a Laser Illuminated Detection and Ranging (LiDAR), which is a laser detection sensor to identify surrounding objects and precisely measure distances to the objects. The accuracy of LiDAR varies a lot depending on their cost and environment, like GPS. Bowen and Waltermire stated that the accuracy for LIDAR data ranges from root mean square error of 1 to 2 meter horizontally and 15 to 20 centimeter vertically, in published evaluations [99]. Csanyi and Toth said "*State-of-the-art lidar systems can achieve 2 to 3 cm ranging accuracy under ideal conditions*", but they also pointed out that the accuracy range is not realistic for typical navigation-based direct sensor platform orientations [100].

To improve accuracy of vehicle position measurements, most papers have focused on localization methods of tracking objects. Simultaneous Localization and Mapping (SLAM) is a method commonly used for improving position estimations by using sensor measurements, landmarks, map, and an estimator, such as extended Kalman filter and particle filter [101-120]. The extended Kalman filter has been used widely, especially in robotics, to address the limitation of linearity assumption of Kalman filter for an estimate of the current mean and covariance [112, 113, 116, 117]. However, Huang and Gamini have addressed convergence and consistency issues of the extended Kalman filter [121, 122]. Also, the extended Kalman filter has limitations on data association problem and an assumption of Gaussian distribution for sensor measurement noises. To overcome this issue, the particle filter has been introduced by estimating state from processing raw data without feature detection [104]. However, each particle in the particle filter represents a trajectory point, and thus

increases computational load. Then, Montemerlo et al. applied Rao-Blackwellized particle filter (RBPF), which reduces memory usage by sharing a map between particles [123]. Although RBPF requires predetermined landmarks, it has become more popular recently [101-104, 119, 120]. Unfortunately, the performance of both methods varies much depending on those assumptions and limitations, study sites and sensor errors, and by far no agreement has been made to which method is better in general.

Instead of using all features of localization, Lee et al. proposed a localization method based on GPS and DR error estimation that is from a lane detection with curved lane models, stop line detection, and curve matching, from [124]. The result of their experimental site shows that the error of estimated position stayed within a meter. There are decent number of researches on MonoSLAM, localization and mapping using a singular data source, mostly visual data [125-129].

Several papers have utilized multi-source data to further improve sensor measurement accuracy [130-134]. In 2006, Mahlisch et al. provided an approach to cross-calibrate vision and ranging sensors by a spatio-temporal alignment [134]. They have shown the proposed model could be applicable for real-time operation by using low level fusion of multibeam LiDAR and vision sensor measurements. Although most of studies in SLAM also use multi-source data, focuses on SLAM are more on localization and mapping rather than data source type and their error distributions.

One of the main challenges of the smoothing and approximation techniques, used in SLAM, is that error distributions of raw measurements are assumed to be a certain type, e.g., Gaussian, or the performance is affected by the error distributions and other assumptions. The authors indicate that the error estimation seems not considered enough in most of studies. Lee et al.’s paper,

relatively, focused more on the error estimation, but their approach relies on the accuracy of lane detection [124].

More fundamentally, almost all proposed SLAM methods in this literature review will need a certain degree of accuracy from raw data measurements, although the impact of such accuracy could be different by the smoothing techniques and purpose of uses.

This study uses Bayesian approach, which has been used in estimating parameters and states for decades, to improve raw data measurements of position and distance by estimating error distribution of each data source. Even the Kalman filter and particle filter are based on Bayesian statistical inference, estimating a joint probability distribution of unknown variables or a conditional probability of the states of some processes. In transportation, it has become available and more often used than before, not just because their benefits of performance, but also due to the introduction of easier approaches such as Markov Chain Monte Carlo (MCMC) based Bayesian approach [135].

The modified Approximate Bayesian Computation, described in Chapter II, was also applied in this chapter. Most of times, focuses of those Bayesian approaches are on estimation of designated parameters or states, not in estimating error distributions of those parameters. This is a significant difference point of this paper, which of the real-time estimation and learning process is based on continuous self-evaluation and updates of the error distributions.

Methodology

Modified Approximate Bayesian Computation

The modified Approximate Bayesian Computation in Chapter II was also used in this study. Instead of considering only one-dimensional variable in Chapter II, this study involves a mixture of one- and two-dimensional variables, the distance and position measurements. For each time stamp, we assume that a vehicle will have the following information:

- Distance measured from the designated vehicle to the most adjacent front vehicle (source: a sensor in the designated vehicle)
- Distance measured from the designated vehicle to the most adjacent rear vehicle (source: a sensor in the designated vehicle)
- Position of the designated vehicle (source: a sensor in the designated vehicle)
- Distance measured from the most adjacent front vehicle to the designated vehicle (source: a sensor in the most adjacent front vehicle)
- Position of the most adjacent front vehicle (source: a sensor in the most adjacent front vehicle)
- Distance measured from the most adjacent rear vehicle to the designated vehicle (source: a sensor in the most adjacent rear vehicle)
- Position of the most adjacent rear vehicle (source: a sensor in the most adjacent rear vehicle)

If there is no error on all the measurements, the distance and position measurements must be consistent. For instance, the calculated distance from the measured positions of between the front and the designated vehicle should be equal to the distance measured from both the designated vehicle and the front vehicle. In this case, the number of data sources at a single time frame will be seven, and the estimated error distribution of each will contribute to calculate

probability of candidate set of true vehicle positions and update its own error distributions.

Overall flowchart (diagram) of estimation process

Figure 23 describes the overall flow of estimation process from gaining sensor measurement to updating the estimated error distribution and estimating the true position, which is conducted simultaneously but can also be implemented as a separate module. In this study, we assume that each vehicle has a such module described here and get the sensor measurements from two adjacent vehicles (front and rear) as they transmit such information with CV environment. The circumstances of data transmission may be different depending on technology developments, regions and their policies. Although the scope of this study does not cover data missing, the estimation of error distribution process may still work on those interruptible situations since the learning process can just skip those missing time stamps. In actual applications using the proposed approach, the module can be modified to still conduct the process with the limited information if only several measurements are missing.

If initial prior distribution is unknown or not setup because of uncertainty, one can gather decent number of samples to estimate the initial error distribution. Either the prior distribution is manually setup from other information or created by the firstly collected sample data, the estimated distribution will be used in the rest of process to generate candidates, calculate log-likelihood of the candidates, and therefore update the estimated error distribution and estimate the true position of vehicles.

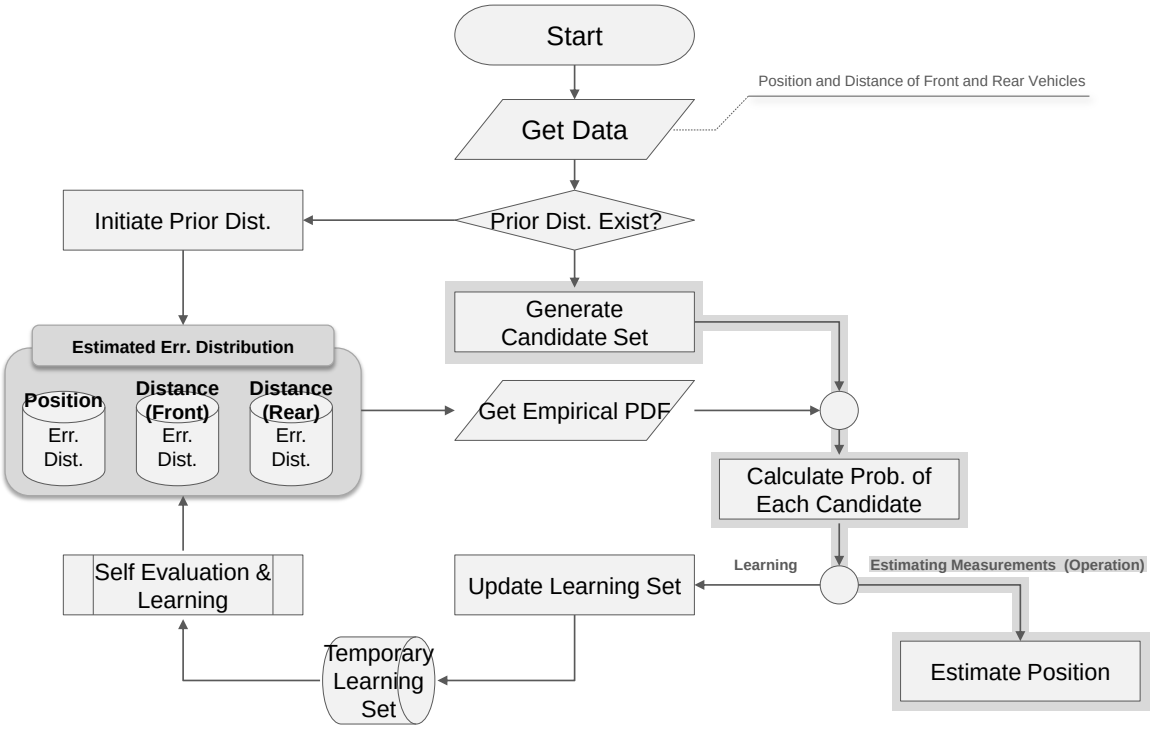


Figure 23. Overall Procedure of Proposed Model

Generating Prior Distribution (Initial Learning)

At the beginning of the proposed model, we may or may not have enough information (or prior knowledge) about error distribution of each sensor measurement. The approach described here is to be applied only where the given prior knowledge is considered to be not enough.

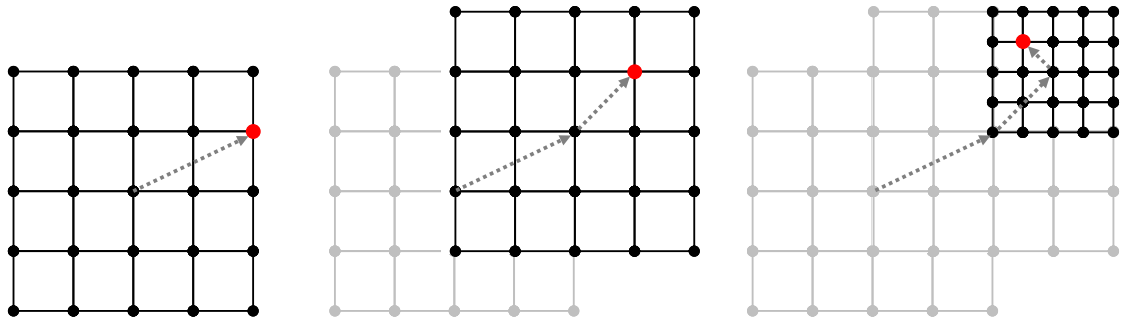
The basic idea of building a prior distribution of measurement error is to consider external information as benchmark data sources and calculate measurement errors as compared to the estimated true measurements based on the benchmark data. For instance, distance measurements from front and rear vehicles to a designated vehicle are used as the benchmark data, i.e., the estimated true measurements, and the estimated measurement errors will be a difference between the estimated true measurements and the collected measurements of the designated vehicle. The assumption of using external data sources as benchmark data would be definitely not true, but it could be enough to build a prior distribution if we use various external data sources, i.e., here, distance measurements from many vehicles, and enough sample size.

Generating Candidate Set

Basically, the grid search is used to generate a local candidate set. The estimated true position will be determined by a candidate point that has maximum likelihood among the all trial set. The detailed calculation of this process is explained in the following section '*Calculating Log-likelihood of Candidates*'.

To estimate the true position more precisely, i.e., to have a higher resolution for a local optimum, the size of grid search is reduced as shown in Figure 24, by a condition, where a local MLE position is inside of boundary of the searching area,

not on the boundary. This is to consider the cases where the local MLE position is far from the starting point of the search. Each candidate point will have a probability calculated by the given information and will be used to update learning set, which is explained in the following section.



(a) *(b)* *(c)*

Figure 24. Candidate Set Generation using Dynamic Grid Search

If we have N consecutively connected vehicles and test all combinations of M candidates for each vehicles' position, the number of computations should be made M^N times. To reduce the computational load, we only look at local MLE positions of front and rear vehicles to estimate a position of designated vehicle and update its learning set. Since local MLE position of a vehicle affects the estimation of adjacent vehicles, estimating the local MLE positions could be done iteratively by using MLE positions of previous iteration each time. This reduces the computation load from $O(M^N)$ to $O(M \times N \times K)$, where K is a number of iterations.

Calculating Log-likelihood of Candidates

Seven measurements are used to calculate log-likelihood of each candidate: the position and the two distance measurements of designated vehicle, and the position and the distance measurement of the two adjacent vehicles from the adjacent vehicles to the designated vehicle.

The log-likelihood of candidate position with a given position measurement information can be calculated by Equation 41.

$$LnL_{position} = Log(P_{position}(err_x, err_y)) \tag{41}$$

where,

$$\begin{aligned} err_x &= measurement_x - candidate_x, \\ err_y &= measurement_y - candidate_y \end{aligned} \tag{42}$$

Likewise, the log-likelihood of having a certain distance measurement error for each candidate, which is based on the Euclidean distance between the candidate positions of the vehicles, can be calculated as the following:

$$LnL_{distance} = Log(P_{distance}(err_{dist.})) \tag{43}$$

where,

$$\begin{aligned} err_{dist.} &= measurement_{dist.} - candidate_{dist.}, \\ candidate_{dist.} &= \sqrt{candidate_{dif\_x}{}^2 + candidate_{dif\_y}{}^2} \end{aligned} \tag{44}$$

For both the position and distance measurement, the probability of having a certain error is determined based on the most recently updated (or learned) error distribution.

In practical applications, the log-likelihood could be too small (too large negative value), which results in having a zero value for most of candidates due to computational limitations when it is converted into a probability at the end. To avoid this issue, researchers might add a constant value to the log-likelihood, which is equivalent to multiplying a constant value to probability.

$$adjusted\ LnL = original\ LnL + \alpha \tag{45}$$

$$adjusted\ Likelihood = Exp\left(\sum original\ LnL\right) \times Exp(\alpha) = original\ Likelihood \times \beta \quad (46)$$

These two constant values, α and β, are used only for the avoiding a zero value of probability issue, but should not impact overall learning experiences. In other words, theoretically, the difference between the original likelihood without the adjustment and the adjusted likelihood is only a multiplication of a constant value, but any additional difference is made by the computational limitations in the calculation of the original likelihood.

Self-Evaluation and Learning

We could update the estimated error distribution of measurements very frequently, in extreme case, every time when a new data set is collected. However, too frequent updates might result in poor prediction of error distribution for each update, which will also affect the accuracy of posterior distribution and the calculation of log-likelihood in each process. Therefore, we need to setup a criterion to make system wait the updating of the estimated error distribution until it's learned enough. The remaining question is how we determine "enough".

Figure 25 illustrates the procedure of the evaluation and update of error distribution. In the proposed methodology, the evaluation is conducted to see whether the updated error distribution with a newly learned set can better explain the collected data measurements. The evaluation criterion here is the sum of log-likelihood between the one based on the most currently updated error distribution and a temporarily estimated error distribution based on the newly learned set. Since the learned set is generated based on the most currently updated error distribution, the temporarily estimated error distribution will tend to have a lower log-likelihood, like having a penalty, if there is no improvement on the current learning set. In other words, the update of error distribution is conducted only

when we confidently expect it improves the estimation of error distribution. Also, a minimum 30 of sample size is used to avoid some randomness, when determining whether the temporarily estimated error distribution is better than the current one.

One of the most benefits of using the suggested evaluation approach is that it does not require additional training data set to evaluate the model performance.

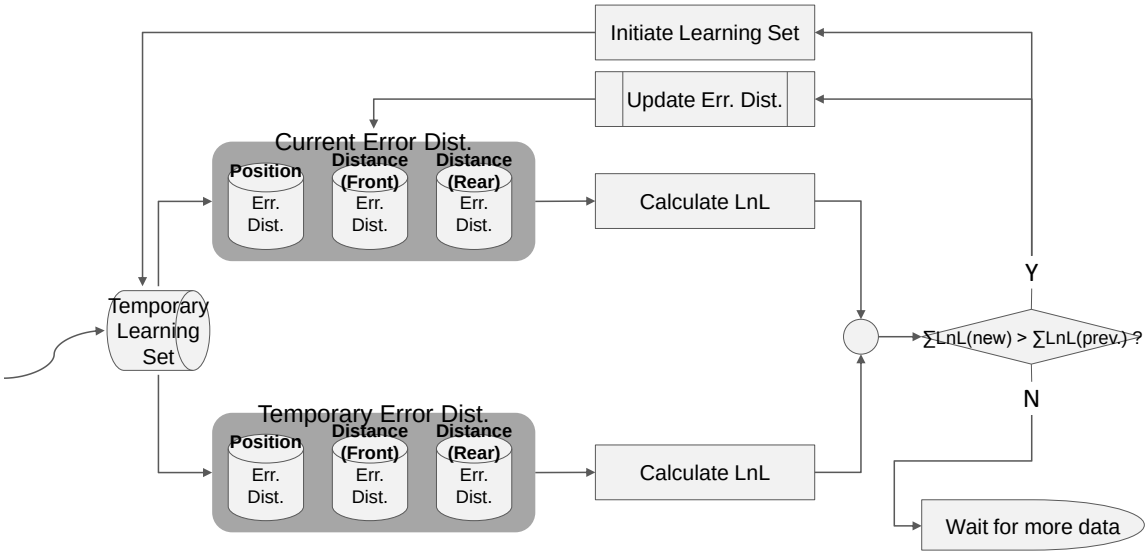


Figure 25. Procedure of Self-Evaluation and Learning for Proposed Model

The self-evaluation will ensure that numerous updates would direct to improve the estimation of error distribution and so does on the estimation of true positions. However, too strict evaluation criteria, e.g., too large minimum sample size or having a too large threshold on the improvement of the log-likelihoods, will slow down the update frequency and may reduce the accuracy of estimation in a short period of time of learning.

Estimating Positions and Distances

Real-time estimation of position and distance can be conducted in one of two different ways: one that uses an exactly same approach used in the learning process based on the seven measurements, and the other one to use just the three measurements, position and distances to a front/rear vehicle of a designated vehicle.

Determining an approach of these two is mainly depending on required minimum time lag of providing such estimations and the given data communication environments. In other words, the first approach will likely to have a better estimation but need more time to process since it requires gathering information from other vehicles, and the decision will be made by whether the extra process time is worthwhile to have the certain improvement on the estimation. In this study, the authors used the first approach, which utilizes all given information.

Simulated Experiments

The main purpose of simulated experiments is to evaluate the model performance and their limitations. The greatest difference compared to using field data is having actual ground truth data. In the field data, there are always limitations, e.g., precision, accuracy, and sample size, to collect ground truth (or benchmark) data.

Network

Each simulation run uses a randomly generated network, so that the impact of using a certain network, e.g., a straight line, could be minimal to overall model performance evaluation.

$$Node(x_i, y_i) = Node(x_{i-1} + \delta\cos(\theta_{i-1}), y_{i-1} + \delta\sin(\theta_{i-1})) \quad (47)$$

$$\theta_i = \theta_{i-1} + \Delta\theta_{i-1} \quad (48)$$

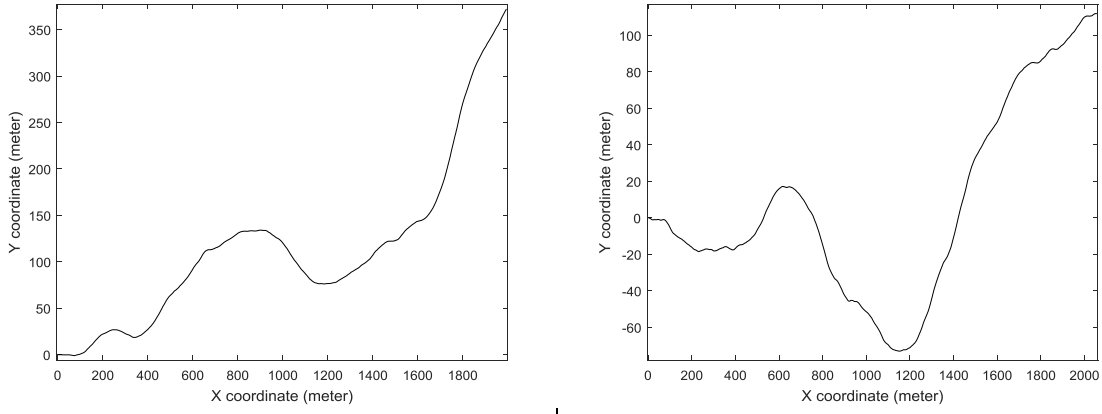
$$\Delta\theta \sim U(\Delta\theta_{min}, \Delta\theta_{max}) \quad (49)$$

where,

δ is a unit distance from a node to the nearest node.

$Node(x_i, y_i)$ is a x/y coordinate of node i, starting from (0,0).

Although the authors do not specify the roadway type into freeway or arterial, it is considered to be similar to freeway than arterial since the created networks have no signalized intersection.



Generated Network Example 1 | Generated Network Example 2

Figure 26. Example of Generated Network

Generating True Positions: Gipps car-following model

The Gipps car-following model was used to generate true positions of vehicles, with uniform random parameter values in the given ranges:

- Desired speed: 60~100 km/h (17 ~ 28 m/s),

- Maximum acceleration: 0.6~1.4 m/s^2,
- Most severe braking that the follower wishes to undertake: 1.5~3.5 m/s^2,
- Follower's estimate of the leader's most severe braking capability: 1.5~3.5 m/s^2,
- The leader's real length + the follower's desired inter-vehicle spacing: 5~10 m

Although a certain car following model was used, the performance of the proposed model should not be affected by the type of car following model or their parameter values, since the proposed model only deals with collected position and distance measurement, but not considering their sequences, i.e., time series. However, when the proposed model is applied along with other smoothing techniques to further improve the measurements, the results would be affected by a chosen car following model. In practical applications, the authors suggest to use actual field data in the interested area, not limited to a certain car following model, as they can better reflect driver behaviors and the characteristics of the local traffic flow.

Performance Measurements (Model Evaluation)

The performance of the proposed model is evaluated in two aspects:

- Is the estimated error distribution similar to the true error distribution?
- Does the estimation of position and distance significantly improve the collected measurements?

To answer the first question, the Bhattacharyya distance, which is a measurement to account a similarity between two probability distributions, was used, like Chapter II. See Equation 40 in Chapter II.

For the second question, Mean Absolute Error (MAE) was used to compare estimated measurements to true values.

$$MAE_i = \frac{\sum_{j=1}^{N_i} |e_{ij}|}{N_i} \tag{50}$$

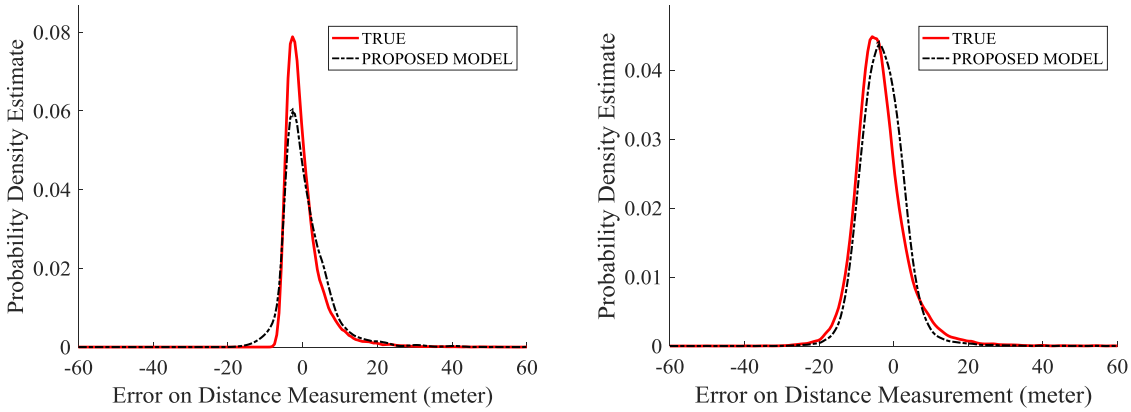
$$e_{ij} = \widehat{D}_{ij} - D_{ij}\ for\ distance, \tag{51}$$

Results

Estimation of Error Distribution

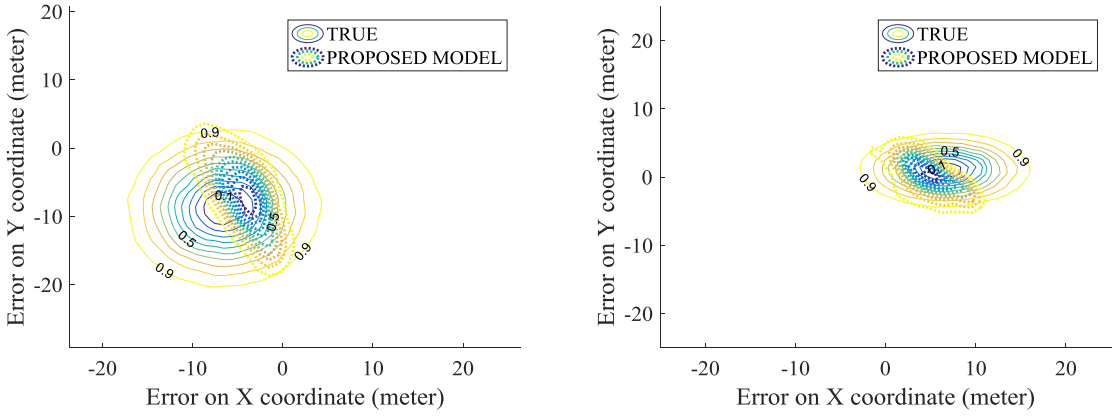
Figure 27 illustrates how close the estimated error distribution is to the true error distribution of distance measurements. In Figure 27 (a), both the estimated and true error distributions are skewed to right. In Figure 27 (b), the distance measurements tend to be underestimated, i.e., a negative error, and the estimated error distribution captures this bias quite well. Capturing the bias is so important since the estimated error distribution could be used to calibrate the distance measurement even when other external data sources are not available.

Figure 28 is a contour plot of estimated errors on position measurement. The value of contour plot represents the probability that the error is within the boundary. For instance, the most outward boundary, i.e., a largest one, has a value of 0.9, meaning that 90% of position measurement errors are within the boundary. Errors of both the x coordinates and y coordinates of position measurements in Figure 28 (a) are negative for most of times, while the errors on x coordinates of Figure 28 (b) are likely to be positive. In Figure 28 (b), the absolute error range of x coordinate is much larger than that of y coordinate. The estimated error distribution for both cases have smaller ranges than the true error distribution, but the bias (negative or positive) seems to be captured well.



(a) *(b)*

Figure 27. Example of Estimated Error Distribution (Distance Measurement)



(a) *(b)*

Figure 28. Example of Estimated Error Distribution (Position Measurement)

Figure 29 represents the average Bhattacharyya distance of all estimated error distributions over learning time. The average Bhattacharyya distances of the distance measurements begin with less than 0.4 and drop to below 0.03 in 24 hours. Considering a data frequency of 1 second in the simulated experiment, we can achieve equivalent performance results within 3 hours if the data is collected every 0.1 second. The average Bhattacharyya distances of the position measurements is larger than the distance measurements over the given time period.

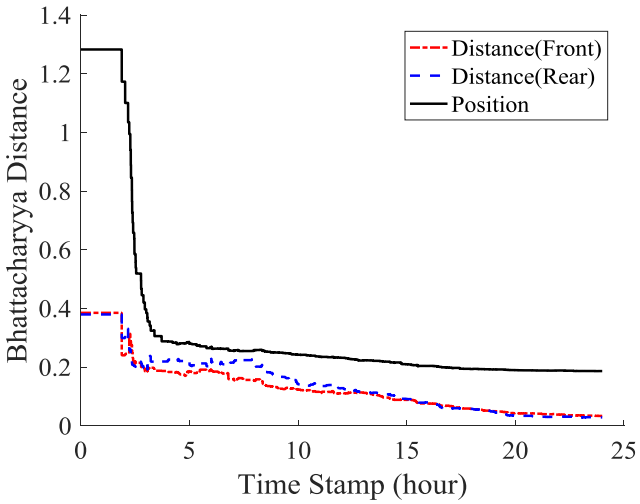


Figure 29. Overall Bhattacharyya Distance over Learned Time Period

Estimation of Vehicle Positions

In Figure 30, the examples of estimated vehicle positions are displayed for single time period. As shown in Figure 30, the estimated positions (‘x’ marker) are much closer to the true position (‘circle’ marker), as compared to the measured positions (‘square’ marker). The improved position estimation can also help provide more accurate estimated distance measurements, as it is simply the Euclidian distance between the adjacent vehicle positions.

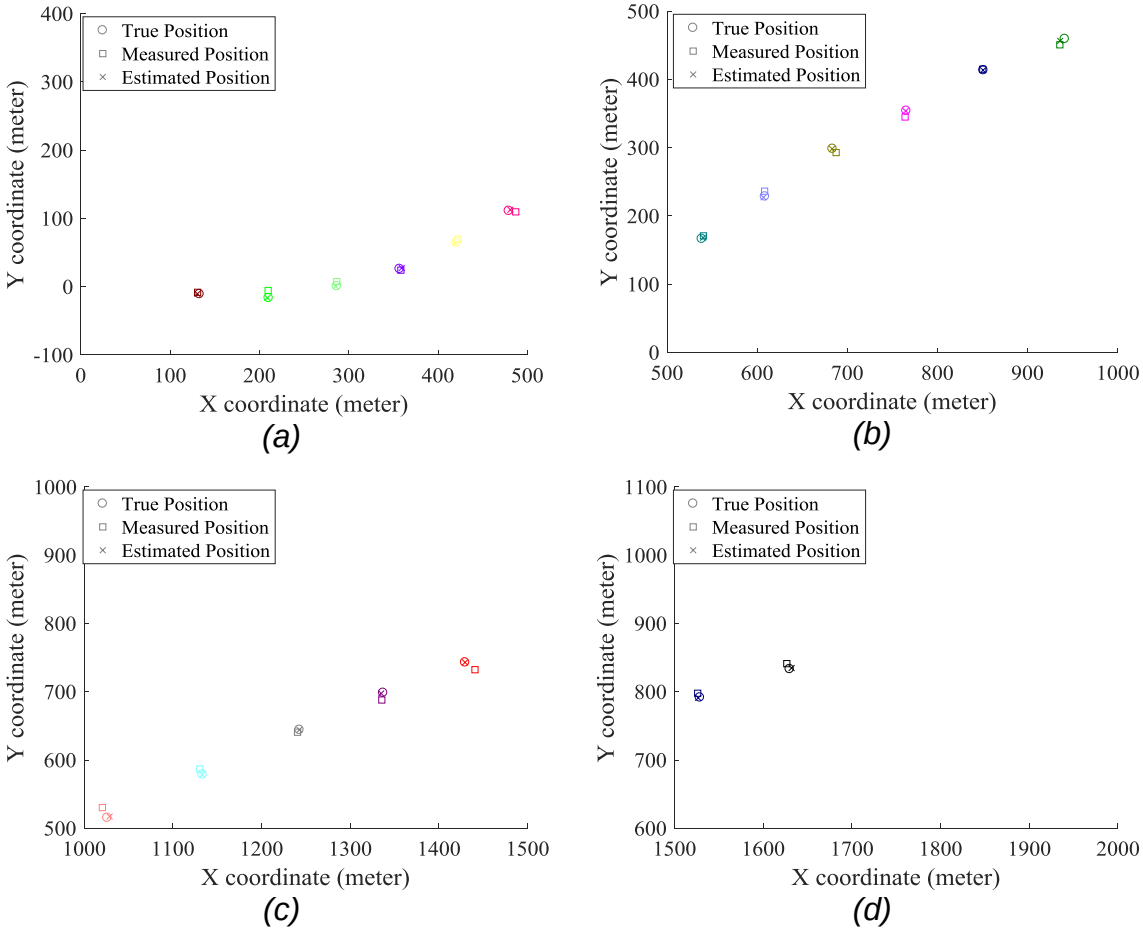
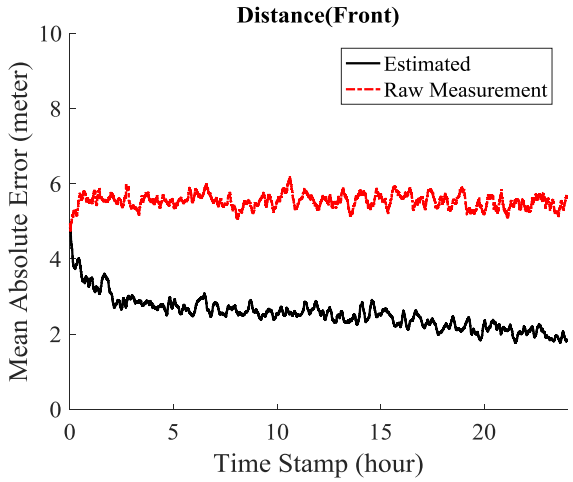
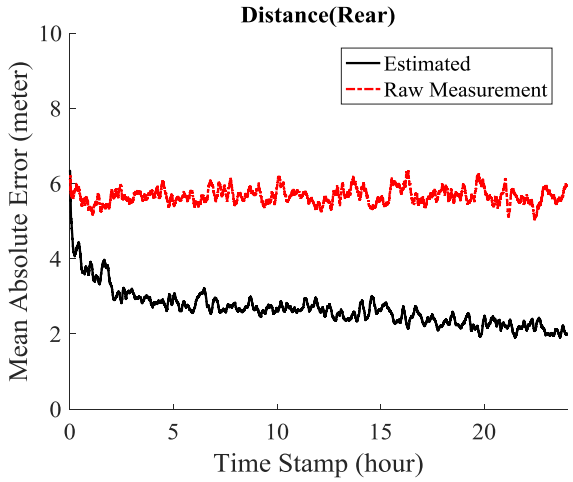


Figure 30. Example of Estimated Positions vs Measured Positions

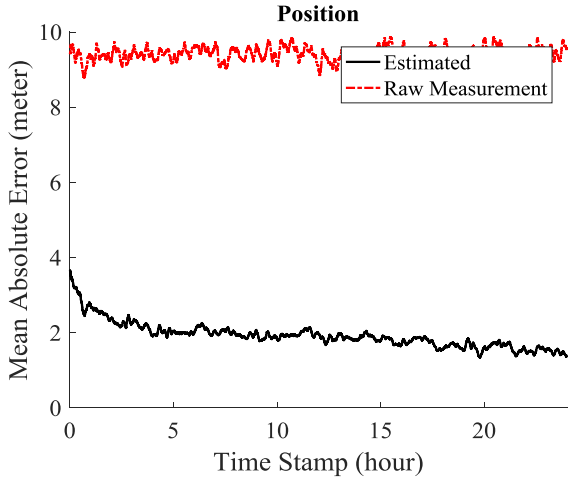
Figure 31 shows the overall performance of those position and distance estimation results, as compared to raw measurement data. The MAE of position measurement here is the Euclidean distance between the true position and the estimated position. It is obvious that the proposed method improves the accuracy of both the distance and position measurements significantly. After 24 hours of learning, the estimated distance is more accurate than the raw data by 3.8 meter (about 70% of original MAE with the raw data), based on MAE. The proposed model also improves the MAE of the position measurements by 8.1 meter (more than 80% of original MAE with the raw data).



(a) Distance (Front)



(b) Distance (Rear)



(c) Position

Figure 31. Overall Mean Absolute Error over Learning Time

Conclusion

This study incorporated the modified Approximate Bayesian Computation and estimated error distributions of position and distance measurements in connected vehicle environment. The results show that the proposed model noticeably improves the accuracy of position and distance measurements.
The key to improve the accuracy is estimating the error distribution of each data source, especially where the shapes of the distributions are not defined. The estimated error distribution can also be used for improving results of other post-processing techniques which require assumptions of certain type of error distributions.

The result shows that the estimated position and distance measurements are more accurate than raw measurements from the initial time of learning, and it becomes more accurate as more data are captured. It is expected that the proposed model could improve further than what is shown in this study, as more data will be available in the fields.

A similar approach can also be utilized to enhance accuracy of other sensors or measurements in connected vehicle or relevant systems, where multi-data sources are available. For instance, vehicle speed can be obtained from multiple sources, e.g., GPS, odometer, Bluetooth, and roadside detectors. The integration of such information could improve the accuracy of the speed and it can be further enhanced by knowing the error distribution of the data sources.

Another possible benefit of proposed approach is that it enables to update the estimated error distribution as new set of data is gathered. This could be critically important to where the error distribution of the sensor measurements is likely to be changed over time, affected by its local environments.

However, it is also important to know that the proposed model is not able to estimate the exact true values of measurements even if the learning time goes infinite. It does provide a likelihood of potential true values, and therefore the true values can be estimated using maximum likelihood, but the variances of error between the true values and the estimated values still exist. In other words, the proposed model makes an effort to reduce uncertainties of measurements using multi-source information, but those inherent variations will not be completely eliminated.

Another limitation of the proposed model is that it assumes overall error distributions of all sensors tend to be unbiased. After decent amount of learning process, the bias of the estimated error distribution of each data source will be significantly contributed by the average error of all the data sources. Therefore, if the most (or all) of data sources are biased to one direction, the estimated error distributions might be biased as well. If the bias of population is known, the estimation can be calibrated by the known bias and the shape of the distribution should be captured by the relative difference between the data sources.

CONCLUSION

This dissertation combined several issues and utilizations of data aggregation in transportation field. These studies were conducted to investigate the sampling bias of harmonic mean, propose a methodology to estimate error distributions of multi-source data, and apply the methodology for improving the accuracy of position and distance measurements in connected vehicle environments.

First, the sampling bias of harmonic mean was shown with a mathematical proof and a numerical example, as well as their analytical and simulation-based corrections, and the impacts of the sampling bias were investigated. The results of the impact analysis show that the sampling bias of harmonic mean is affected by time interval, segment length, and sampling rate. It is important to know that the sampling bias and its corrections, as well as determining required sample size, should be considered differently by purpose of its use and their local traffic conditions.

Second, aggregating multi-source data was utilized to better estimate error distribution of each data source, by incorporating the modified Approximate Bayesian Computation. The proposed model eliminates the need for determining a benchmark data source (or ground truth). The proposed model outperformed the alternative approach, which is a conventional way of evaluating data sources by comparing them with the benchmark data. Numerous simulations were conducted for sensitivity analysis of sample size, number of data sources, and distribution types on its model performance.

Finally, the modified Approximate Bayesian Computation was applied for improving the accuracy of the distance and position measurement in connected vehicle systems. The results show that the proposed method can enhance the

accuracy of the raw measurements. The proposed approach can be easily expanded to other measurements in connected vehicle systems or other relevant systems, where multi-source data are available.

There are still many remaining challenges on data aggregation in the transportation field. Although the subjects covered in the dissertation are very limited to the travel speed data and the position and distance measurements, the implications and potential applications can be expanded in other fields.

LIST OF REFERENCES

1. Edie, L.C., *Flow theories.* Traffic Science, 1974: p. 9-16.
2. Gazis, D.C. and L.C. Edie, *Traffic flow theory.* Proceedings of the IEEE, 1968. 56(4): p. 458-471.
3. Lighthill, M.J. and G.B. Whitham, *On kinematic waves. II. a theory of traffic flow on long crowded roads.* Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 1955. 229(1178): p. 317-345.
4. Wardrop, J.G. and G. Charlesworth, *A method of estimating speed and flow of traffic from a moving vehicle.* Proceedings of the Institution of Civil Engineers, 1954. 3(1): p. 158-171.
5. *Highway capacity manual.* Special Report in Transportation Research Board, 1985. 209.
6. *Transportation and traffic engineering handbook, 3rd edition.* Institute of Transportation Engineers, 1976.
7. *Traffic engineering handbook, 4th edition.* Institute of Transportation Engineers, 1992.
8. Wohl, M. and B.V. Martin, *Traffic system analysis for engineers and planners.* McGraw-Hill Inc.,US, 1967.
9. Ardekani, S. and R. Herman, *Urban network-wide traffic variables and their relations.* Transportation Science, 1987. 21(1): p. 1-16.
10. Leutzbach, W., *Introduction to the theory of traffic flow.* Springer-Verlag Berlin Heidelberg, 1972.
11. Kennedy, N., *Fundamentals of traffic engineering.* University of California, 1966.
12. Haight, F.A., *Mathematical theories of traffic flow.* Academic Press, 1963.
13. Hall, F.L., *Traffic stream characteristics.* In N.H. Gartner, et al., Traffic Flow Theory. US Federal Highway Administration, 1996.
14. Rakha, H. and W. Zhang, *Estimating traffic stream space mean speed and reliability from dual-and single-loop detectors.* Transportation Research Record: Journal of the Transportation Research Board, 2005(1925): p. 38-47.
15. Knoop, V., S. Hoogendoorn, and H. Zuylen, *Empirical differences between time mean speed and space mean speed*, in *Traffic and Granular Flow '07*, C. Appert-Rolland, et al., Editors. 2007, Springer. p. 351-356.
16. Hoogendoorn, S., G. Hegeman, and T. Dijker, *Traffic flow theory and simulation*. 2006: TU Delft.
17. Wardrop, J.G., *Some theoretical aspects of road traffic research.* Road Paper, 1952. 36: p. 325-362.
18. Soriguera, F. and F. Robusté, *Estimation of traffic stream space mean speed from time aggregations of double loop detector data.* Transportation Research Part C: Emerging Technologies, 2011. 19(1): p. 115-129.
19. Gerlough, D.L. and M.J. Huber, *Traffic flow theory: a monograph.* Washington, D.C.: Transportation Research Board, 1975.

20. Khisty, C.J. and B.K. Lall, *Transportation engineering: an Introduction.* Prentice Hall, 2003.
21. Long Cheu, R., C. Xie, and D.H. Lee, *Probe vehicle population and sample size for arterial speed estimation.* Computer-Aided Civil and Infrastructure Engineering, 2002. 17(1): p. 53-60.
22. Jensen, J.L., S.D. Thomas, and P.W. Corbett, *On the bias and sampling variation of the harmonic average.* Mathematical Geology, 1997. 29(2): p. 267-276.
23. Jensen, J.L., *Some statistical properties of power averages for lognormal samples.* Water Resources Research, 1998. 34(9): p. 2415-2418.
24. Limbrunner, J.F., R.M. Vogel, and L.C. Brown, *Estimation of harmonic mean of a lognormal variable.* Journal of Hydrologic Engineering, 2000. 5(1): p. 59-66.
25. Satagopan, J.M., M.A. Newton, and A.E. Raftery, *Easy estimation of normalizing constants and Bayes factors from posterior simulation: stabilizing the harmonic mean estimator.* Department of Statistics, University of Washington, 2000.
26. Alzer, H., *An inequality for arithmetic and harmonic means.* Aequationes Mathematicae, 1993. 46(3): p. 257-263.
27. Alić, M., et al., *The arithmetic-geometric-harmonic-mean and related matrix inequalities.* Linear Algebra and Its Applications, 1997. 264: p. 55-62.
28. Alzer, H., *A harmonic mean inequality for the gamma function.* Journal of Computational and Applied Mathematics, 1997. 87(2): p. 195-198.
29. Mond, B. and J. Pečarić, *A mixed arithmetic-mean-harmonic-mean matrix inequality.* Linear Algebra and Its Applications, 1996. 237: p. 449-454.
30. Chu, Y.-M. and W.-F. Xia, *Two sharp inequalities for power mean, geometric mean, and harmonic Mean.* Journal of Inequalities and Applications, 2009. 2: p. 3.
31. Ando, T., *On the arithmetic-geometric-harmonic-mean inequalities for positive definite matrices.* Linear Algebra and Its Applications, 1983. 52: p. 31-37.
32. Mathias, R., *An arithmetic-geometric-harmonic mean inequality involving Hadamard products.* Linear Algebra and Its Applications, 1993. 184: p. 71-78.
33. Gautschi, W., *A Harmonic mean inequality for the gamma function.* SIAM Journal on Mathematical Analysis, 1974. 5(2): p. 278-281.
34. Moschopoulos, P.G., *The distribution of the sum of independent gamma random variables.* Annals of the Institute of Statistical Mathematics, 1985. 37(1): p. 541-544.
35. Cook, J.D. *Inverse gamma distribution* 2008 [accessed 2016 Oct 1]. https://pdfs.semanticscholar.org/d06b/1ba4a5ddbe0e950069b4aab7c4ba7ffe9e29.pdf.

36. *NGSIM–Next Generation SIMulation*. FHWA, U.S. Department of Transportation, [accessed 2016 Oct 1]. http://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm.
37. Bozdogan, H., *Akaike's information criterion and recent developments in information complexity.* Journal of Mathematical Psychology, 2000. 44(1): p. 62-91.
38. Hargrove, S.R., et al., *Empirical evaluation of the accuracy of technologies for measuring average speed in real time.* Transportation Research Record: Journal of the Transportation Research Board, 2016(2594): p. 73-82.
39. *Harmonic mean*. Wikipedia, [accessed 2016 Oct 1]. https://en.wikipedia.org/wiki/Harmonic_mean.
40. Du, S., et al., *Automatic license plate recognition (ALPR): A state-of-the-art review.* IEEE Transactions on Circuits and Systems for Video Technology, 2013. 23(2): p. 311-325.
41. Gilly, D. and K. Raimond, *A survey on license plate recognition systems.* International Journal of Computer Applications, 2013. 61(6): p. 34-40.
42. Wang, J.-X., et al., *The research and realization of vehicle license plate character segmentation and recognition technology.* Proceedings of Wavelet Analysis and Pattern Recognition (ICWAPR), 2010: p. 101-104.
43. Treiber, M. and A. Kesting, *Travel time estimation*, in *Traffic flow dynamics*. 2013, Springer. p. 367-377.
44. Bertini, R.L., M. Lasky, and C.M. Monsere, *Validating predicted rural corridor travel times from an automated license plate recognition system: Oregon's frontier project.* Proceedings of Intelligent Transportation Systems, IEEE, 2005: p. 296-301.
45. Herrera, J.C., et al., *Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment.* Transportation Research Part C: Emerging Technologies, 2010. 18(4): p. 568-583.
46. Yim, Y. and R. Cayford, *Investigation of vehicles as probes using global positioning system and cellular phone tracking: field operational test.* California Partners for Advanced Transit and Highways (PATH), 2001.
47. Haghani, A., et al., *Data collection of freeway travel time ground truth with bluetooth sensors.* Transportation Research Record: Journal of the Transportation Research Board, 2010. 2160(1): p. 60-68.
48. Porter, J.D., et al., *Antenna characterization for Bluetooth-based travel time data collection.* Journal of Intelligent Transportation Systems, 2013. 17(2): p. 142-151.
49. Saeedi, A., et al., *Improving accuracy and precision of travel time samples collected at signalized arterial roads with bluetooth sensors.* Transportation Research Record: Journal of the Transportation Research Board, 2013. 2380(-1): p. 90-98.

50. Wasson, J.S., J.R. Sturdevant, and D.M. Bullock, *Real-time travel time estimates using media access control address matching.* ITE Journal, 2008. 78(6).
51. Quayle, S.M., et al., *Arterial performance measures with media access control readers.* Transportation Research Record: Journal of the Transportation Research Board, 2010. 2192(1): p. 185-193.
52. Malinovskiy, Y., et al., *Field experiments on bluetooth-based travel time data collection.* Proceedings of Transportation Research Board 89th Annual Meeting, 2010(10-3134).
53. Wright, J. and J. Dahlgren, *Using vehicles equipped with toll tags as probes for providing travel times.* California PATH Program, Institute of Transportation Studies, University of California at Berkeley, 2001.
54. Ban, X., et al., *Performance evaluation of travel time methods for real time traffic applications.* Proceedings of 11th World Conference on Transport Research, 2007.
55. Ribeiro, P., D.S. Rodrigues, and E. Taniguchi, *Comparing standard and low-cost tools for gradient evaluation along potential cycling paths.* WSEAS Transactions on Environment and Development, 2015. 11: p. 29-40.
56. Solberg, A.H.S., A.K. Jain, and T. Taxt, *Multisource classification of remotely sensed data: fusion of Landsat TM and SAR images.* IEEE transactions on Geoscience and Remote Sensing, 1994. 32(4): p. 768-778.
57. Mascarenhas, N., G. Banon, and A. Candeias, *Multispectral image data fusion under a Bayesian approach.* International Journal of Remote Sensing, 1996. 17(8): p. 1457-1471.
58. Couillet, R. and M. Debbah, *A Bayesian framework for collaborative multi-source signal sensing.* IEEE Transactions on Signal Processing, 2010. 58(10): p. 5186-5195.
59. Elfes, A., *Multi-source spatial data fusion using Bayesian reasoning.* Data Fusion in Robotics and Machine Intelligence, 1992: p. 137-163.
60. Finley, A.O., S. Banerjee, and R.E. McRoberts, *A Bayesian approach to multi-source forest area estimation.* Environmental and Ecological Statistics, 2008. 15(2): p. 241-258.
61. Ran, Y., et al., *Large-scale land cover mapping with the integration of multi-source information based on the Dempster–Shafer theory.* International Journal of Geographical Information Science, 2012. 26(1): p. 169-191.
62. Aarabi, P. and S. Zaky, *Robust sound localization using multi-source audiovisual information fusion.* Information Fusion, 2001. 2(3): p. 209-223.
63. Khoshelham, K., et al., *Performance evaluation of automated approaches to building detection in multi-source aerial data.* ISPRS Journal of Photogrammetry and Remote Sensing, 2010. 65(1): p. 123-133.

64. Bai, Y., H. Wang, and C. Zaniolo, *Load shedding in classifying multi-source streaming data: A Bayes risk approach.* Proceedings of the 2007 SIAM International Conference on Data Mining, 2007: p. 425-430.
65. Sun, S., C. Zhang, and Y. Zhang, *Traffic flow forecasting using a spatio-temporal bayesian network predictor.* Proceedings of International Conference on Artificial Neural Networks, 2005: p. 273-278.
66. Ghosh, B., B. Basu, and M. O'Mahony, *Bayesian time-series model for short-term traffic flow forecasting.* Journal of Transportation Engineering, 2007. 133(3): p. 180-189.
67. Castillo, E., J.M. Menéndez, and S. Sánchez-Cambronero, *Predicting traffic flow using Bayesian networks.* Transportation Research Part B: Methodological, 2008. 42(5): p. 482-509.
68. Zheng, W., D.-H. Lee, and Q. Shi, *Short-term freeway traffic flow prediction: Bayesian combined neural network approach.* Journal of Transportation Engineering, 2006. 132(2): p. 114-121.
69. Sun, S., C. Zhang, and G. Yu, *A Bayesian network approach to traffic flow forecasting.* IEEE Transactions on Intelligent Transportation Systems, 2006. 7(1): p. 124-132.
70. Tebaldi, C. and M. West, *Bayesian inference on network traffic using link count data.* Journal of the American Statistical Association, 1998. 93(442): p. 557-573.
71. Jintanakul, K., L. Chu, and R. Jayakrishnan, *Bayesian mixture model for estimating freeway travel time distributions from small probe samples from multiple days.* Transportation Research Record: Journal of the Transportation Research Board, 2009(2136): p. 37-44.
72. Van Hinsbergen, C. and J. van Lint, *Bayesian combination of travel time prediction models.* Transportation Research Record: Journal of the Transportation Research Board, 2008(2064): p. 73-80.
73. Choi, K. and Y. Chung, *A data fusion algorithm for estimating link travel time.* ITS Journal, 2002. 7(3-4): p. 235-260.
74. Fei, X., C.-C. Lu, and K. Liu, *A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction.* Transportation Research Part C: Emerging Technologies, 2011. 19(6): p. 1306-1318.
75. Westgate, B.S., et al., *Travel time estimation for ambulances using Bayesian data augmentation.* The Annals of Applied Statistics, 2013. 7(2): p. 1139-1161.
76. Park, T. and S. Lee, *A Bayesian approach for estimating link travel time on urban arterial road network.* Proceedings of International Conference on Computational Science and Its Applications, 2004: p. 1017-1025.
77. Yu, R., M. Abdel-Aty, and M. Ahmed, *Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors.* Accident Analysis & Prevention, 2013. 50: p. 371-376.

78. Huang, H. and M. Abdel-Aty, *Multilevel data and Bayesian analysis in traffic safety.* Accident Analysis & Prevention, 2010. 42(6): p. 1556-1565.
79. Ma, J., K.M. Kockelman, and P. Damien, *A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods.* Accident Analysis & Prevention, 2008. 40(3): p. 964-975.
80. Huang, H., H.C. Chin, and M.M. Haque, *Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis.* Accident Analysis & Prevention, 2008. 40(1): p. 45-54.
81. Song, J.J., et al., *Bayesian multivariate spatial models for roadway traffic crash mapping.* Journal of Multivariate Analysis, 2006. 97(1): p. 246-273.
82. Miaou, S.-P. and D. Lord, *Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods.* Transportation Research Record: Journal of the Transportation Research Board, 2003(1840): p. 31-40.
83. Kendall, M.G., A. Stuart, and J. Ord, *The advanced theory of statistics.* London, 1968. 3.
84. Beaumont, M.A., W. Zhang, and D.J. Balding, *Approximate Bayesian computation in population genetics.* Genetics, 2002. 162(4): p. 2025-2035.
85. Diggle, P.J. and R.J. Gratton, *Monte Carlo methods of inference for implicit statistical models.* Journal of the Royal Statistical Society. Series B (Methodological), 1984: p. 193-227.
86. Rubin, D.B., *Bayesianly justifiable and relevant frequency calculations for the applied statistician.* The Annals of Statistics, 1984. 12(4): p. 1151-1172.
87. Tavaré, S., et al., *Inferring coalescence times from DNA sequence data.* Genetics, 1997. 145(2): p. 505-518.
88. Toni, T., et al., *Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems.* Journal of the Royal Society Interface, 2009. 6(31): p. 187-202.
89. Sunnåker, M., et al., *Approximate bayesian computation.* PLoS Comput Biol, 2013. 9(1): p. e1002803.
90. Bhattachayya, A., *On a measure of divergence between two statistical population defined by their population distributions.* Bulletin Calcutta Mathematical Society, 1943. 35(99-109): p. 28.
91. *Connected vehicle technology industry Delphi study.* Center for Automotive Research, Michigan Department of Transportation, pp 11-18, 2012.
92. Administration, N.H.T.S., *National motor vehicle crash causation survey: Report to congress.* National Highway Traffic Safety Administration Technical Report DOT HS, 2008. 811: p. 059.
93. Litman, T., *Autonomous vehicle implementation predictions.* Victoria Transport Policy Institute, 2014. 28.

94. *Global positioning system standard positioning service performance standard, 4th edition.* DoD Official Document, Washington, pp 21–22, 2008.
95. Li, X., et al., *Accuracy and reliability of multi-GNSS real-time precise positioning: GPS, GLONASS, BeiDou, and Galileo.* Journal of Geodesy, 2015. 89(6): p. 607-635.
96. Talbot, N.C., M.T. Allison, and M.E. Nichols, *Centimeter accurate global positioning system receiver for on-the-fly real-time kinematic measurement and control.* U.S. Patent No. 5,890,091. 30 Mar. 1999, 1999.
97. Zumberge, J., et al., *Precise point positioning for the efficient and robust analysis of GPS data from large networks.* Journal of Geophysical Research: Solid Earth, 1997. 102(B3): p. 5005-5017.
98. Wang, G., *Millimeter-accuracy GPS landslide monitoring using Precise Point Positioning with Single Receiver Phase Ambiguity (PPP-SRPA) resolution: a case study in Puerto Rico.* Journal of Geodetic Science, 2013. 3(1): p. 22-31.
99. Bowen, Z.H. and R.G. Waltermire, *Evaluation of light detection and ranging (LIDAR) for measuring river corridor topography.* JAWRA Journal of the American Water Resources Association, 2002. 38(1): p. 33-41.
100. Csanyi, N. and C.K. Toth, *Improvement of lidar data accuracy using lidar-specific ground targets.* Photogrammetric Engineering & Remote Sensing, 2007. 73(4): p. 385-396.
101. Gouveia, B.D., D. Portugal, and L. Marques, *Speeding up Rao-Blackwellized particle filter SLAM with a multithreaded architecture.* Proceedings of Intelligent Robots and Systems (IROS 2014), IEEE/RSJ, 2014: p. 1583-1588.
102. Fu, M., et al., *A navigation map building algorithm using refined RBPF-SLAM.* Proceedings of Guidance, Navigation and Control Conference (CGNCC), 2016 IEEE Chinese, 2016: p. 2483-2487.
103. Yatim, N.M. and N. Buniyamin, *Development of Rao-Blackwellized Particle Filter (RBPF) SLAM algorithm using low proximity infrared sensors.* Proceedings of Robotic, Vision, Signal Processing and Power Applications, 2017: p. 395-405.
104. Yatim, N.M. and N. Buniyamin, *Particle filter in simulataneous localisation and mapping (SLAM) using differential drive mobile robot.* Jurnal Teknologi (Sci Eng), 2015. 77(20): p. 91-97.
105. Grisetti, G., et al., *Fast and accurate SLAM with Rao–Blackwellized particle filters.* Robotics and Autonomous Systems, 2007. 55(1): p. 30-38.
106. Gil, A., et al., *Multi-robot visual SLAM using a Rao-Blackwellized particle filter.* Robotics and Autonomous Systems, 2010. 58(1): p. 68-80.
107. Sim, R., P. Elinas, and J.J. Little, *A study of the Rao-Blackwellised particle filter for efficient and accurate vision-based SLAM.* International Journal of Computer Vision, 2007. 74(3): p. 303-318.

108. Grisettiyz, G., C. Stachniss, and W. Burgard, *Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling.* Proceedings of Robotics and Automation, 2005: p. 2432-2437.
109. Törnqvist, D., et al., *Particle filter SLAM with high dimensional vehicle model.* Journal of Intelligent & Robotic Systems, 2009. 55(4): p. 249-266.
110. Fairfield, N., G. Kantor, and D. Wettergreen, *Towards particle filter SLAM with three dimensional evidence grids in a flooded subterranean environment.* Proceedings of Robotics and Automation, 2006: p. 3575-3580.
111. Sim, R., et al., *Vision-based SLAM using the Rao-Blackwellised particle filter.* Proceedings of IJCAI Workshop on Reasoning with Uncertainty in Robotics, 2005. 14(1): p. 9-16.
112. Bonato, V., E. Marques, and G.A. Constantinides, *A floating-point extended kalman filter implementation for autonomous mobile robots.* Journal of Signal Processing Systems, 2009. 56(1): p. 41-50.
113. Choi, K.-S. and S.-G. Lee, *Enhanced SLAM for a mobile robot using extended Kalman filter and neural networks.* International Journal of Precision Engineering and Manufacturing, 2010. 11(2): p. 255-264.
114. Bailey, T., et al., *Consistency of the EKF-SLAM algorithm.* Proceedings of Intelligent Robots and Systems, IEEE/RSJ, 2006: p. 3562-3568.
115. Castellanos, J.A., J. Neira, and J.D. Tardós, *Limits to the consistency of EKF-based SLAM.* Proceedings of IFAC Symposium on Intelligent Autonomous Vehicles, 2004.
116. Choi, M., R. Sakthivel, and W.K. Chung, *Neural network-aided extended Kalman filter for SLAM problem.* Proceedings of Robotics and Automation, 2007: p. 1686-1690.
117. Chatterjee, A. and F. Matsuno, *A neuro-fuzzy assisted extended Kalman filter-based approach for simultaneous localization and mapping (SLAM) problems.* IEEE Transactions on Fuzzy Systems, 2007. 15(5): p. 984-997.
118. Yatim, N.M. and N. Buniyamin. *Development of Rao-Blackwellized Particle Filter (RBPF) SLAM Algorithm Using Low Proximity Infrared Sensors*. in *9th International Conference on Robotic, Vision, Signal Processing and Power Applications*. 2017. Springer.
119. Bejuri, W.M.Y.W., et al., *Optimization of Rao-Blackwellized Particle Filter in Activity Pedestrian Simultaneously Localization and Mapping (SLAM): An Initial Proposal.* International Journal of Security and Its Applications, 2015. 9(11): p. 377-390.
120. Choi, J. and M. Maurer, *Hybrid map-based SLAM with Rao-Blackwellized particle filters.* Proceedings of Information Fusion (FUSION), 2014: p. 1-6.
121. Huang, G.P., A.I. Mourikis, and S.I. Roumeliotis, *Analysis and improvement of the consistency of extended Kalman filter based SLAM.* Proceedings of Robotics and Automation, 2008: p. 473-479.

122. Huang, S. and G. Dissanayake, *Convergence and consistency analysis for extended Kalman filter based SLAM.* IEEE Transactions on robotics, 2007. 23(5): p. 1036-1049.
123. Montemerlo, M., et al., *FastSLAM: A factored solution to the simultaneous localization and mapping problem.* Proceedings of AAAI, 2002: p. 593-598.
124. Lee, B.-H., et al., *GPS/DR error estimation for autonomous vehicle localization.* Sensors, 2015. 15(8): p. 20779-20798.
125. Atashgah, M.A. and S. Malaek, *An integrated virtual environment for feasibility studies and implementation of aerial MonoSLAM.* Virtual Reality, 2012. 16(3): p. 215-232.
126. Sadat, S.A., et al., *Feature-rich path planning for robust navigation of mavs with mono-slam.* Proceedings of Robotics and Automation, 2014: p. 3870-3875.
127. Sunderhauf, N., S. Lange, and P. Protzel, *Using the unscented kalman filter in mono-SLAM with inverse depth parametrization for autonomous airship control.* Proceedings of Safety, Security and Rescue Robotics, SSRR/IEEE, 2007: p. 1-6.
128. Holmes, S., G. Klein, and D.W. Murray, *A square root unscented Kalman filter for visual monoSLAM.* Proceedings of Robotics and Automation, 2008: p. 3710-3716.
129. Davison, A.J., et al., *MonoSLAM: Real-time single camera SLAM.* IEEE transactions on pattern analysis and machine intelligence, 2007. 29(6).
130. Ryu, J.H., G. Gankhuyag, and K.T. Chong, *Navigation system heading and position accuracy improvement through GPS and INS data fusion.* Journal of Sensors, 2016. 2016.
131. Sabatini, R., et al., *Low-cost sensors data fusion for small size unmanned aerial vehicles navigation and guidance.* International Journal of Unmanned Systems Engineering., 2013. 1(3): p. 16.
132. Mercado, D., et al., *GPS/INS/optic flow data fusion for position and velocity estimation.* Proceedings of Unmanned Aircraft Systems (ICUAS), 2013: p. 486-491.
133. Hoang, G.M., et al., *Distributed link selection and data fusion for cooperative positioning in GPS-aided IEEE 802.11 p VANETs.* Proc. WPNC, 2015. 15.
134. Mahlisch, M., et al., *Sensorfusion using spatio-temporal aligned video and lidar for improved vehicle detection.* Proceedings of Intelligent Vehicles Symposium, IEEE, 2006: p. 424-429.
135. Washington, S.P., M.G. Karlaftis, and F. Mannering, *Statistical and econometric methods for transportation data analysis, 2nd edition.* Chapman and Hall/CRC, 2010: p. 387-401.

VITA

Hyeonsup Lim was born and raised in Suncheon, a small village in Jeollanam-do province of S. Korea. He earned Bachelor's degree in transportation system engineering from Ajou University in 2008, with honors, and Master's degree in urban planning from Seoul National University in 2010. Before he joined the University of Tennessee (UT) in 2013, he worked at Seoul Institute, a government agency in transportation research, and Smart Card Co., Ltd., a private company in implementing transit-oriented ITS. In 2017, he was granted a doctoral degree in Civil Engineering with concentration in Transportation Engineering and his Master's degree in Statistics at UT.

For his four years of Ph.D. program, Hyeonsup was recipient of the Chancellor's scholarship in UT. He worked on multiple projects in UT, developing integrated traffic simulation algorithm and real-time traffic data extraction tool, and enhancing self-learning license plate matching algorithm. He also has been working on building public information of nationwide freight movements at Oak Ridge National Laboratory. During his graduate studies, Hyeonsup received several scholarships and awards, including Intelligent Transportation Society of Tennessee's annual scholarship award (first recipient in UT), T. Darcy Sullivan TSITE student scholarship, TSITE Student paper competition award, KOTAA Excellent paper awards, and Graduate Student Senate travel awards. He also served as the president of ITE student chapter at UT. His research interests include connected/autonomous vehicle, machine learning, traffic flow theory, and freight logistics.