ABSTRACT

Title of Dissertation:	EMPOWERING TRAFFIC OPERATIONS AND SAFETY WITH TRANSPORTATION BIG DATA: PRACTICE SCAN, METHODOLOGY, AND APPLICATIONS
	Mofeng Yang, Doctor of Philosophy, 2022
Dissertation directed by:	Professor, Paul Schonfeld, Department of Civil and Environmental Engineering

In the past two decades, along with the technological advancement in mobile sensors and mobile networks, transportation big data, such as probe vehicle data and mobile device location data (MDLD), have been growing dramatically in terms of the spatiotemporal coverage of population and its mobility. These data sources have shown their great potential for large-scale and near real-time transportation applications to support travel behavior analysis, travel demand modeling, traffic operations and safety analyses. The objectives of this dissertation are to (1) comprehensively examine the state-of-the-practice applications and the state-of-the-art models developed based on emerging transportation big data, (2) identify key metrics, and (3) establish a series of big-data driven frameworks to enhance traffic operations and safety. Three main sections are included.

The first section of this dissertation presents a literature review on models, tools, and metrics used for various levels of traffic analysis, and analyzes a survey distributed to

transportation professionals to quantify the importance of these key metrics for improving traffic operations and safety. Based on the literature review and survey insights, two big-data driven frameworks are proposed accordingly to address both traffic operations and safety issues.

In the second section of this dissertation, a big-data driven framework is developed which aims at improving the accuracy and reliability of emergency medical services (EMS) and trauma triage decisions for elderly persons at crash sites. The proposed framework integrates transportation big data sources from both the demand side (such as traffic volumes, and time-dependent vehicle speeds obtained from large-scale probe vehicles) and the supply side (i.e., transportation network features), as well as publicly available statewide crash data with health-related decisions such as EMS and hospital records. Decision tree models are adopted to simulate the decision-making process due to their wide applications, a proven capability in prediction, and interoperability. With records of over 55,000 elderly patients, results demonstrate that the proposed framework contributed to enhanced EMS decision and trauma triage accuracy for the elderly, and saving more lives from severe vehicle crashes.

In the third section of this dissertation, a big-data driven framework is proposed for estimating a critical operational metric, namely vehicle volume, on an all-street network, and further estimating the pedestrian and bicyclist crashes at all intersections. This framework employs a series of cloud-based computational algorithms to extract multimodal trajectories and trip rosters from terabytes of MDLD. A scalable map matching and routing algorithm is then applied to snap and route vehicle trajectories to the roadway network. The observed vehicle counts on each roadway segment are weighted and calibrated against ground truth control totals, i.e., Annual Vehicle Miles of Travel (AVMT), and Annual Average Daily Traffic (AADT). The proposed framework is built on Amazon Web Service (AWS) which leverages cloud computing techniques

to estimate vehicle volumes for all roadway segments in the state of Maryland using MDLD for the entire year 2019. The estimated vehicle volume is further integrated with statewide crash records to estimate the pedestrian and bicyclist crashes at all intersections with statistical models. Results indicate that the proposed framework can produce reliable vehicle volume estimates and estimated pedestrian and bicyclist crashes, while also demonstrating its transferability and generalization ability.

In summary, this dissertation comprehensively examines the literature on transportation big data applications and proposes two big-data driven frameworks demonstrated with two real-world case studies. Results reveal the feasibility and advantages of empowering traffic operations and safety analysis with transportation big data.

EMPOWERING TRAFFIC OPERATIONS AND SAFETY WITH TRANSPORTATION BIG DATA: PRACTICE SCAN, METHODOLOGY AND APPLICATIONS

by

Mofeng Yang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Paul Schonfeld, Chair, Department of Civil and Environmental Engineering
Professor Kathleen Stewart, Dean's Representative, Department of Geographical Sciences
Professor Ali Haghani, Department of Civil and Environmental Engineering
Assistant Professor Taylor Oshan, Department of Geographical Sciences
Assistant Professor Chenfeng Xiong, Department of Civil and Environmental Engineering, Villanova University

Dedication

To my beloved parents Kun Yang and Peifan Li, and my wife Zhiyue Xia.

And to my grandparents in heaven.

Acknowledgements

This dissertation is partially funded by U.S. Department of Transportation (U.S. DOT), Federal Highway Administration (FHWA), Maryland Department of Transportation State Highway Administration (MDOT SHA), and Maryland Transportation Institute (MTI). Opinions herein do not necessarily represent the views of the research sponsors. The author is responsible for the statements in the thesis.

"*Night is now falling.*

So ends this day.

The road is now calling, and I must away."

This is the lyrics from the song "*The Last Goodbye*" from the movie "*The Hobbit*". Just like *Bilbo Baggins*, instead of reclaiming the Lonely Mountain from the dragon Smaug, I was sitting in the plane at Beijing airport, waiting for my unknown journey to the other side of the Pacific Ocean. Now, four years have passed, and I am here writing this acknowledge to summarize this "unexpected journey".

First, I would like to express my sincere gratitude to my advisor, Dr. Paul Schonfeld for his continuous guidance since I joined the program in 2018. Dr. Schonfeld became my advisor in June 2022 and was also the committee member for my master thesis in 2020. I really appreciate all the support and guidance Dr. Schonfeld provides.

I would also like to express my special thanks to Dr. Kathleen Stewart, not only for being the dean's representative of my dissertation committee, but also for the guidance through research collaborations in the past few years. Dr. Stewart's research works have always been my top references when conducting my own research. In the meantime, I would also like to thank all my doctoral dissertation committee members: Dr. Ali Haghani, Dr. Taylor Oshan, and Dr. Chenfeng

Xiong for offering me their valuable comments to improve my research. I am extremely grateful to Dr. Haghani for his support when my previous advisor left the university. I sincerely appreciate Dr. Oshan for the perfect course I took with him at the Department of Geographical Science as well as serving in my master thesis committee. A special thanks to Dr. Chenfeng Xiong. The past four years of working together under Dr. Xiong supervision will be an unforgettable experience that I will always cherish. I would like to thank Dr. Lei Zhang for giving me the opportunity to come to the U.S. and get me involved into research projects.

I want to thank my colleagues at Maryland Transportation: Jina Mahmoudi, Sepehr Ghader, Aref Darzi, Minha Lee, Weiyi Zhou, Aliakbar Kabiri, Songhua Hu, Guangchen Zhao, Weiyu Luo, Mohammad Ashoori, Saeed Saleh, Asal Tabrizi.

Last, I would like to thank my parents, Kun Yang and Peifan Li, for always respecting my opinions and decisions. Thanks to my grandparents, Shizhen Yang and Daohua Chen, who passed away in 2021 and 2022. It was, is and will always be a pity in my life that I couldn't be with you at the very last moment of your life. Thanks for my wife, Zhiyue Xia, who always supports me whenever I meet obstacles.

Though not knowing the journey and where it leads, I embrace it, and I welcome every moment of it. Just like what it says in the song:

"And though where the road then takes me.

I cannot tell.

We came all this way.

But now comes the day.

To bid you farewell"

Mofeng Yang, at Greenbelt, MD

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AADT	Annual Average Daily Traffic
ANN	Artificial Neural Networks
AVMT	Annual Vehicle Miles of Travel
AWS	Amazon Web Service
BI	Bayesian Inference
BMC	Baltimore Metropolitan Council
BN	Bayesian Network
BTS	Bureau of Transportation Statistics
CART	Classification and Regression Tree
CASI	Computer-Assisted Self-Interview
CATI	Computer-Assisted Telephone Interview
CATT	Center for Advanced Transportation Technology
CBSA	Core-based Statistical Area
CDR	Call Detail Record
CNN	Convolutional neural Network
CODES	Crash Outcome Data Evaluation System
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DHMM	Discrete Hidden Markov Model
DNN	Deep Neural Networks
DMV	Washington Metropolitan Area
EMS	Emergency Medical Service
FHWA	Federal Highway Administration
GPS	Global Positioning System
HPMS	Highway Performance Monitoring System
KNN	K-Nearest Neighbors
LBS	Location-based Service
LPR	License Plate Recognition
LRI	Location Recording Interval
MDOT	Maryland Department of Transportation
MDOT SHA	Maryland Department of Transportation State Highway Administration
MLP	Multi-Layer Perceptron
MaaS	Mobility-as-a-Service
MTI	Maryland Transportation Institute
MWCOG	Metropolitan Washington Council of Government
NASS	National Automotive Sampling System
NB	Negative Binomial

NHTS	National Household Travel Survey
NHTSA	National Highway Traffic Safety Administration
NTM	National Transit Map
OD	Origin and Destination
OSM	OpenStreetMap
PAPI	Paper-And-Pencil Interview
PCMDL	Passively Collected Mobile Device Location
RETTS-A	Rapid Emergency Triage and Treatment System
RF	Random Forest
RITIS	Regional Integrated Transportation Information System
SHA	State Highway Administration
SMOTE	Synthetic Minority Over-sampling Technique
SVC	Support Vector Classifier
SVM	Support Vector Machine
TAZ	Traffic Analysis Zone
TPB	Transportation Planning Board
TRBAM	Transportation Research Board Annual Meeting
UMD	University of Maryland
U.S.	the United States
USDOE	the United States Department of Energy
USDOT	the United States Department of Transportation
XGB	eXtreme Gradient Boosting
ZIP	Zero-Inflated Poisson
ZINB	Zero-Inflated Negative Binomial

Chapter 1: Introduction

1.1 Background

The current federal transportation legislation, "Moving Ahead for Progress in the 21st Century" (MAP-21), which was signed into law on July 6, 2012, advances statewide and metropolitan planning processes to incorporate a comprehensive performance-based approach to decision-making. Typical performance metrics in planning for traffic operations and safety include changes in vehicle trips, vehicle miles traveled, emissions reduction, travel time savings, improvements in travel time reliability, energy consumption reduction, noise impacts, safety impacts, monetary values of these changes, and lists of traffic operations equipment and costs. In addition to traditional data sources such as loop detector data, and video-based traffic counts, emerging transportation big data such as probe vehicle data, connected vehicle data, and passively collected mobile device location data have enabled large-scale and near real-time models and methods to support operation, safety, demand, and planning analyses. More specifically, it is now possible to tell which users and which origin-destination pairs are using a particular transportation facility and in turn.

In the past two decades, along with the technological advancement in mobile sensors and mobile networks, transportation big data, such as probe vehicle data and mobile device location data (MDLD), have been growing dramatically in terms of the spatiotemporal coverage of population and its mobility. Initially, these data sources are considered supplements to travel surveys and travel behavior analysis. A series of practices and research studies have demonstrated the effectiveness of such data in enhancing traditional travel surveys as well as revealed its great potential to replace travel surveys [1, 2]. At the same time, obtaining travel statistics solely based

on these data sources are also worth investigating in order to reduce labor and cost compared to travel surveys.

Apart from supporting travel surveys and travel behavior analysis, in recent years, transportation big data have also been leveraged in large-scale and near real-time transportation applications for traffic operations and safety analysis. For instance, vehicle volume, as a critical operational metric, is the fundamental basis for traffic signal control, transportation project prioritization, road maintenance plans, and more. Traditional methods of quantifying vehicle volume rely on manual counting, video cameras, and loop detectors at a limited number of locations. These efforts require significant labor and cost for large-scale implementations. Researchers and private sector companies have explored alternative solutions such as probe vehicle data, which still suffer from low penetration rates. With the introduction of transportation big data, vehicle volume can be estimated from terabytes of movement data for a larger geographical area with larger sample size.

1.2 Objectives

The objective of this dissertation is to comprehensively examine the state-of-the-practice transportation big data applications and to develop state-of-the-art big-data driven frameworks that fully leverage the potential of the transportation big data, cloud computing techniques to improve traffic operations and safety analysis. In order to fulfill the research objective, four tasks are identified: (1) evaluating the state-of-the-practice applications and the state-of-the-art methods based on transportation big data and identifying the key research gap; (2) designing and distributing a survey to transportation professionals to identify key metrics for varying level of traffic analysis focusing on traffic operations and safety; (3) developing a big-data driven framework that leverages the instantaneous vehicle speed estimated nearly in real-time from large-

scale probe vehicle data to enhance Emergency Medical Services (EMS) and trauma triage for elderly persons at crash sites; (4) developing and validating a big-data driven framework that estimates vehicle volume on an all-street network and quantifies the pedestrian and bicyclist crashes at all intersections.

1.3 Contributions

The main contributions of this dissertation can be classified into three aspects: (1) Uniqueness of the data; (2) Methodology innovations; (3) Applications.

1.3.1 Uniqueness of the Data

This dissertation utilizes multiple transportation big data sources and develops in-house data collection methods for model development and validation. Though extensive studies have leveraged transportation big data such as MDLD and probe vehicles data, transportation big data used in this dissertation include both probe vehicle data from the Regional Integrated Transportation Information System (RITIS) developed by the Center for Advanced Transportation Technology (CATT) and large-scale anonymized MDLD from the Maryland Transportation Institute (MTI). Both of the aforementioned data sources cover the entire U.S. and are able to capture high-granularity spatiotemporal traffic and movement across the country. Therefore, the methodology frameworks proposed in this dissertation have the advantage of generalization to expand to the entire U.S. In addition to these external data sources, an in-house data collection method is achieved by developing one of the most advanced Mobility-as-a-Service (MaaS) mobile application, *incenTrip*. *incenTrip* collects travel behavior data with user-confirmed ground truth information including trip origin and destination, departure time, and travel mode. The intermediate locations of each trip are also recorded for model development.

Apart from transportation big data, this dissertation also collects multimodal transportation networks and transportation statistics from multiple authorized sources, such as the OpenStreetMap (OSM), the Maryland Department of Transportation State Highway Administration (MDOT SHA), United States Department of Transportation (USDOT) Bureau of Transportation Statistics (BTS), USDOT National Transit Map (NTM). These publicly available data sources are used together with the transportation big data sources for model developments and validations.

1.3.2 Methodology Innovations

The methodology contribution of this dissertation focuses on the following aspect:

(1) Integrating transportation big data with health-related data: This study is among the first to develop and demonstrate a methodological framework for integrating transportation-sector data with health-related data to support various decision-making scenarios in transportation safety, emergency responses, and trauma-care triage.

(2) Computation algorithms for deriving travel behavior data from large-scale MDLD: The computation algorithms proposed in this study are developed in order to derive travel behavior data, such as activity locations, and travel modes, from large-scale MDLD. They are calibrated and validated against the existing travel surveys and annual vehicle miles of travel in order to serve real-world research needs. Also, these algorithms are further compiled and integrated into a data pipeline in the Amazon Web Service (AWS) platform which fully leverages the computing power and scalability of cloud computing techniques.

(3) *Scalable map matching and routing, weighting, and calibration algorithms that have superior transferability and generalization ability*: The big-data framework proposed to estimate vehicle volume and pedestrian and bicyclist can be applied for every state in the U.S. The

algorithms are scalable, and the data sources proposed for weighting and calibration are generally available across the U.S.

1.3.3 Applications

The frameworks proposed in this dissertation have been calibrated, validated, and ultimately deployed in several real-world applications and have huge potential for boosting future transportation big data applications. These applications cover a wide range of topics, including travel demand management (*incenTrip*, *https://incentrip.org*), traffic operations and safety (*Vulnerable Road User Density Exposure Dashboard project*, *https://mti.umd.edu/sdi*), public health and mobility (*University of Maryland COVID-19 Impact Analysis Platform*, *https://data.covid.umd.edu*), travel behavior analysis and surveys (*Next Generation National Household Travel Survey National Passenger/Truck Origin-Destination Data*, https://nhts.ornl.gov/od).

1.4 Organization

Figure 1-1. Dissertation Outline.

The outline of this dissertation is organized as shown in Figure 1-1. Chapter 2 provides a comprehensive literature review about the state-of-the-practice transportation applications using and the state-of-the-art methodologies applied to transportation big data. Chapter 3 firstly scans the existing tools and metrics for traffic operations and safety analysis. Then a survey is designed and distributed to transportation professionals to identify key metrics for traffic operations and safety analyses. Chapter 4 develops a framework that leverages large-scale probe vehicle data to

improve the accuracy and reliability of emergency medical services (EMS) and trauma triage decision scenarios for the elderly population in crashes. Chapter 5 proposes a big data-driven framework. The proposed framework leverages cloud computing techniques to digest terabytes of transportation big data and produces an important operational metric, vehicle volume, on the all-street network and further estimating the corresponding pedestrian and bicyclist crashes in Chapter 6. Finally, Chapter 7 summarizes the conclusion and suggests future research directions.

Chapter 2: Literature Review

2.1 Transportation Big Data Applications

In this section, existing applications based on transportation big data are reviewed. The transportation big data are categorized into three types: Global Positioning Service (GPS) data, Cellular and sighting data, and Location-based Service (LBS) data, where a majority can be also categorized as Mobile Device Location Data (MDLD). The applications for each type are reviewed and the state-of-the-art methods are summarized.

2.1.1 GPS Data

The earliest and most widely used type of transportation big data is that obtained from GPS technology, where personal longitudinal location data is collected via GPS data loggers. Since the mid-1990s, researchers began investigating the possibility of using GPS data to enhance the quality of travel surveys. The initial version of the GPS data logger could only be installed in a vehicle and charged by the vehicle's battery [3-11]. The vehicle location was recorded in each second when the vehicle was moving [6]. This approach can significantly improve the spatiotemporal accuracy of travel surveys by recording the exact origin and destination as Ill as the trip start and end times, but it only captures vehicle trips. Later, the wearable GPS further allowed respondents to carry them so that trips traveled by non-vehicle travel modes could also be recorded [12-15]. Some travel surveys utilized both in-vehicle and wearable GPS data loggers to take advantage of both technologies [16-18].

Since the GPS data can offer accurate locations of the devices, access to individual-level trajectories is highly restricted. Therefore, the individual-level GPS data are also aggregated by private sector companies to reveal travel demand without raising privacy concerns. For instance,

INRIX Traffic collects GPS probe data from commercial vehicle fleets, connected vehicles, and mobile device applications [19]. RITIS also started to incorporate the probe vehicle data into their commercial products. The data can be further aggregated into link- or corridor- levels to provide a real-time estimation of traffic speed and travel time [20-22]. Nonetheless, the low penetration rate (i.e., 2%-10%) of the commercial probe vehicle data remains the core challenge with respect to drawing the whole picture of travel patterns.

2.1.2 Cellular and Sighting Data

Since mobile devices, such as smartphones and tablets, have gained in popularity, investigations into individual-level mobility patterns have become more practical. The cellular data, which are generated through communication between cellphones and cell towers when a phone call or a text message is made by the phone [23], have shown their great value in supporting large-scale travel demand analysis. In general, the cellular data can be categorized into Call Detail Record (CDR) and sightings [24]. Call Detail Record (CDR) data provide details on calls and messages, such as timestamp, duration, and locations of routing cell towers. Therefore, the location information of CDR data fully depends on the density of the cellular network and does not reflect the actual location of the device [24]. Similarly, sightings are also generated through communication with cell towers, but the actual location of the device is determined via triangular calculation [24]. Both types of cellular data have been widely used in studying human mobility patterns in the past two decades. For instance, Gonzalez, et al. combined two sets of CDRs to explore individual mobility patterns; one is composed of six months of records for 100,000 randomly selected anonymous individuals and the other is a complementary dataset capturing the locations of 206 mobile phone users every two hours for one week [25]. Further studies on human mobility have been conducted based on similar datasets [26-33]. Cellular data have also been widely applied in other research

areas such as social networks, residential location, and socioeconomic level [34-36]. Despite the large volume of data, cellular data are limited by their spatial and temporal resolution, which is determined by the density of cell towers and user cellphone usage [37]. However, on the positive side, cellular data require less advanced phones and should raise less concern about user privacy.

2.1.3 Location-based Service Data

Another type of transportation big data is Location-based Service (LBS) data, in which spatial information is generated when a mobile application updates the device's location with the most accurate sources, based on the existing location sensors such as Wi-Fi, Bluetooth, cellular tower, and GPS [24, 38]. Compared to the CDR data, the LBS data can reflect the exact location of mobile devices and thus provide invaluable location information describing individual-level mobility patterns [24, 25-27, 38, 39]. Many applications have been developed using the LBS data. For instance, a recent smartphone-enhanced travel survey conducted in the U.S. used a mobile application, rMove, developed by Resource Systems Group (RSG), to collect high-frequency location data and let respondents recall their trips by showing the trajectories in rMove [40-43]. Airsage leveraged LBS data to develop a traffic platform that can estimate traffic flow, speed, congestion, and road user sociodemographic for every road and time of day [44].The Maryland Transportation Institute (MTI) at the University of Maryland (UMD) developed the COVID-19 Impact Analysis Platform (*data.covid.umd.edu*) to provide insight on COVID-19's impact on mobility, health, economy, and society across the U.S. [45-47].

In summary, transportation big data used in the literature are different in terms of spatiotemporal coverage of population and its mobility, as well as data quality, e.g., spatial accuracy and location recording interval (LRI) [48, 49]. The GPS data in general have the highest spatial accuracy (e.g., 10 meters) and the lowest LRI (usually 1 second), but usually cover only a

small percentage of the population, and thus cannot reflect population-level travel behavior without a statistical weighting process. Therefore, most of the GPS data are used as supplementary data sources for regional travel surveys. The cellular data and LBS data have significantly higher spatiotemporal coverage of the population than the GPS data because of the large penetration rate of cellphone and mobile devices in the U.S. However, the ground truth information is usually missing, The LRI for both types of data is high and has a larger variation depending on mobile device usage thus also has a larger variation [49]. In addition, although cellular data may have higher coverage, the spatial accuracy of the data and the temporal frequency of the pins are inferior to the LBS data. This is because cellular technology relies on the density of cell towers and does not reflect the actual location of the devices. Also, cellular data are generated based on calls and messages or a network-driven event which might lead to a lower number of events.

2.2 Models and Algorithms for Transportation Big Data

2.2.1 Trip End Identification

The trip end identification algorithm for high-frequency data, i.e., GPS data, has been well-studied and used in practical applications [48]. To obtain accurate trip ends, the traditional way is the rule-based trip end identification method. This type of method designs rules and parameters based on domain knowledge. The trip ends are obtained by applying the rules to location data point by point and at the same time examining the interrelation between two consecutive location points. The parameters used in these rules are mostly defined by domain knowledge, such as dwell time and speed [50-57]. In recent years, some researchers also leveraged supervised machine learning models as a supplement to the rule-based methods, which classify each location point as static or moving [58-60]. Different clustering methods are also applied to obtain trip ends by first

identifying people's activity locations from the location data [61-64]. A recent study utilized a spatiotemporal clustering method with three combined optimization models to detect trip ends [64]. In recent years, there was also a special focus on deriving the trip ends from LBS data. A "Divide, Conquer and Integrate" (DCI) framework was proposed to process the LBS data to extract mobility patterns in the Puget Sound region [39]. The proposed framework combined a rule-based method and incremental clustering method to handle the bi-modally distributed LBS data. The results were aggregated at the census tract level and compared with household travel surveys.

2.2.2 Travel Mode Imputation

After the trip ends are identified, it is also important to impute the travel mode for each trip to obtain multimodal travel patterns. Travel mode imputation can be categorized mainly into two approaches: (1) trip-based approach; and (2) segment/point-based approach. The trip-based approach is based on the already identified trip ends, where each trip has only one travel mode to be imputed. The segment/point-based approach separates the trip into fixed-length segments (time or distance) or a single point and then imputes the travel mode for each segment or point [49]. Then the segments/points with the same travel mode are further merged to form a single-mode trip. Both previous trip-based approaches and segment/point-based approaches have used similar features in order to distinguish between different travel modes.

Table 2-1. Studies on Travel Mode Imputation Methods.

Author	LRI	Model	Main Features	Modes	Acc.
Gong et al. 2012 [54]	/	Rules	Speed, Acceleration, Transit Stations, Transit Network	Drive, Train, Bus, Walk, Bike, Static	82.6%
Stenneth et al. 2011 [65]	30 s	RF	Speed, Acceleration, Heading change, Bus location, Transit Network	Drive, Bus, Train, Walk, Bike, Static	93.7%
Bruunauer et al. 2013 [66]	1-10 s	MLP	Speed, Acceleration, Bendiness	Drive, Bus, Train, Walk, Bike	92.0%
Xiao et al. 2015 [68]	1 s	BN	Speed, Acceleration, Trip Distance	Drive Bus, Walk, Bike, E-Bike	92.0%
Nitsche et al. 2014 [67]	1 s	DHMM	Speed, Acceleration, Direction	Drive, Bus, Motorcycle, Train, Tram, Subway, Walk, Bike	65% - 95%
Dabiri and Heaslip. 2018 [71]	1-5 s	CNN	Speed, Acceleration, Jerk, Bearing Rate	Drive, Bus, Train, Walk, Bike	84.8%
Bachir et al. 2019 [32]	/	BI	Road and Rail Trip Counts	Road, Rail	/
Vaughan et al. 2020 [73]	/	DNN	Speed, Trip Distance, Land Use, Time of Day	Drive, Bus, Active (Walk, Bike)	87%
Burkhard et al. 2020 [49]	1 s subsampled to 5 min	KNN, RF, etc.	Speed, Public Transport Stops, and Lines	Drive, Train, Tram, Bus, Walk, Bike	/
Breyer et al. 2021 [74]	/	KNN etc.	Road and Train Route Geometry	Road, Train	95.5%

** RF: Random Forest; MLP: Multi-Layer Perceptron; BN: Bayesian Network; DHMM: Discrete Hidden Markov Model; CNN: Convolutional neural Network; BI: Bayesian Inference; DNN: Deep Neural Network*

Table 2-1 summarizes typical methods and features that are used for travel mode imputation. According to the literature review done by Huang et al. 2019 and Burkhard et al. 2020, it can be observed that typical features include speed and acceleration [49, 54, 65-73]. Specifically, when the LRI is less than 10 seconds, the speed (speed variation) and acceleration features are more important in differentiating among different travel modes, which can be imputed solely from the data. When the LRI is relatively high, such as 30 s, additional features can be added to maintain the same level of accuracy such as real-time transit information [65], multimodal transportation network [49, 54, 65, 74], and sociodemographic information [70, 73]. However, most of these studies tested the algorithms using the low-LRI GPS data sample, which has frequent observations. Limited efforts have been spent on developing suitable algorithms for cellular data or LBS data that suffer from the high-LRI issue. Burkhard et al. examined the required spatial accuracy and LRI to accurately detect travel mode from the high-LRI MDLDs by subsampling the low-LRI GPS data [49]. They concluded that the LRI should be less than a minute to ensure the travel mode imputation accuracy. Bachir et al. developed a Bayesian Inference (BI) method to separate road and rail modes from the CDR data in the Greater Paris region by leveraging the road and rail trip counts from the travel survey [32]. Vaughan et al. trained a Deep Neural Network (DNN) model to separate drive, bus, and active modes with artificial CDR traces reconstructed from the travel survey data [73]. The model is applied to the real-world CDR data to obtain travel mode shares. Breyer et al. developed multiple classification methods using labeled CDR data to separate only the road and train modes between two OD pairs [74]. The major limitation of these studies is that either the study area is small (e.g., an OD pair or a region) or the method only separates easy-to-detect modes (e.g., Road versus Rail). As Huang et al. 2019 mentioned in their review [48], the supervised machine learning methods have not been fully exploited yet due to the lack of ground

truth labeled data, and might be worth investigating for MDLDs, especially for the cellular data and LBS data. Besides, rather than identifying easy-to-detect modes (e.g., rail versus road), their review suggests including more mode categories

2.3 Transportation Big Data for Traffic Operations and Safety

2.3.1 State-of-the-Practice on Crash Scene Decision Makings

For a long time, older people's limitations in traffic safety research were emphasized in discussing the contributing factors of crash injury severity, such as traffic conditions, roadway geometries, land use type, environmental conditions, and driver characteristics through different statistical analyses [75-79]. Also, numerous methods have been developed to predict the crash injury severity with the crash report data. Although crash severity prediction models have come into common discussion during the last decade in policy, research as well as practice, they are still suffering from a lack of clarity and accuracy regarding their interpretation and data availability. In turn, it limits the capability of applying these methods for real-time decision-making. In reality, two major decisions need to be made at the crash scene:

- whether an EMS is needed: when someone is injured in a vehicle crash, the responding EMS providers must provide emergency care at the scene and then transport the patient to healthcare based on the injury severity [80];
- whether an injured person should be triaged to the trauma centers: If an EMS is dispatched, the EMS providers must not only determine the severity of the injury and initiate medical management, but also identify the most appropriate transport destination facility through a process called "field triage" [81].

Table 2-2. State-of-the-Art Methodologies of Trauma Triage

Authors	Method	Outcome	Factors	Performance
Scheetz et al. 2007 [88]	Decision Tree	Trauma or Non-trauma	Age, Gender, Height, Light, Glasgow coma scale, Injury severity, etc.	95.15% SE* and 76.47% SP * for severe injury, 83.1% SE and 81.5% SP* for moderate injury
Wang et al. 2009 [83]	Rule-based	Trauma or Non-trauma	Glasgow coma scale, Blood pressure, Respiratory rate, Crash characteristics, Estimated traffic speed, etc.	/
Sasser et al. 2012 [84] & Davidson et al. 2014 [85]	Rule-based	Trauma or Non-trauma	Same as Wang et al. 2009	/
Newgard et al. 2016 [89]	Decision Tree	Trauma or Non-trauma	Age, Gender, Glasgow coma scale, Blood pressure, Respiratory rate, Mechanism of injury, etc.	92.1% SE* and 41.5% SP*
AtiksaIdparit et al. 2019 [91]	Statistical model	Severe injury and death	Age, Gender, Body mass index (BMI), Crash characteristics, EMS response time, Mechanism of injury, Physiological status, etc.	90.2% SE* and 75.9% SP* for severe injury, 98.7% SE* and 68.8 SP* for death
Van Rein et al. 2019 [92]	Statistical model	Injury severity score	Age, Glasgow coma scale, Blood pressure, Mechanism criteria, Penetrating injury etc.	88.8% SE* and 50.0* SP*
Van der Sluijs et al. 2019 [90]	Decision Tree	Injury severity score	Age, Gender, Glasgow coma scale, Blood pressure, Mechanism of injury, Injury type etc.	Not reported
Magnusson et al. 2020 [93]	Rule-based	RETTS-A* triage levels	Dispatch medical index (DMI) including Chest pain, Extremity, Respiratory difficulties etc.	81.0% SE* and 64.0% SP*
Shanahan et al. 2021 [94]	Statistical model	Injury severity score	Same as Van Rein et al. 2019	83.0% SE* and 50.0% SP*

** SE: Sensitivity, also called true positive rate or recall.*
** SP: Specificity, also called true negative rate.*
** RETTS-A: The Rapid Emergency Triage and Treatment System.*

After an EMS team arrives at the scene, field triage decisions need to be made by EMS providers to determine whether the injured occupants should be sent to a trauma center. A recent study used the National Automotive Sampling System (NASS) to study the factors affecting triage decisions. Their results indicate that though injury severity and resulting mortality among the older group (age > 60) was higher than for younger counterparts, the older group is less likely to be transported to a trauma center [81, 82]. These findings emphasized that the triage decision significantly saves people's lives, especially old people. With these considerations, several field triage decision guidelines were developed for reference.

The universally used Field Triage Decision Scheme was revised by a National Expert Panel organized by the Centers for Disease Control and Prevention, where comprehensive crash and health-related data were used [83]. In 2012, the National Center for Injury Prevention and Control and the Division of Injury Response, in collaboration with the National Highway Traffic Safety Administration (NHTSA), Office of Emergency Medical Services, and in association with the American College of Surgeons, Division of Research and Optimal Patient Care also released the Guidelines for Field Triage of Injured Patients [84, 85]. These represent the latest development on rule-based guidelines for field triage. Several studies were also conducted focusing on validating these guidelines [86, 87]. Apart from these guidelines, data mining and decision tree methods, such as Classification and Regression Tree (CART) [88, 89] and gradient boosting decision tree [90] were also largely used by researchers to predict the trauma triage decisions. These studies are reviewed in Table 2-2.

While several studies have touched on the strong association between crash severity and traffic conditions (e.g., [83]), few studies have linked trauma triage decision-making with transportation domain knowledge and/or transportation-sector data. With the recent engineering

advances in transportation big data and data-driven analytical methods, transportation-sector data becomes increasingly available, in terms of both the data coverage and the timeliness to support real-time or near real-time decisions such as EMS and trauma triage. Motivated by this cross-disciplinary research needs, this study aims at filling the data gap with integrated transportation and health data. It contributes to the existing literature with a methodological framework that integrates relevant transportation-sector data sources including network characteristics, traffic volumes, and historical travel speed at the crash scenes into the health-related decisions. Such integration is also believed to contribute to an enhanced accuracy compared to existing studies (as shown in Table 2-2). This study then empirically tests the framework on EMS and trauma triage decision scenarios using Maryland datasets. Decision tree models are adopted due to their wide applications and proven capability in prediction. Results demonstrate that the integrated transportation and health data contribute to enhanced prediction accuracy, reducing under-triage for the elderly, and saving more lives from vehicle crashes.

2.3.2 Estimating Vehicle Volume based on Transportation Big Data

Traditional methods of quantifying vehicle volume rely on manual counting, video cameras, and loop detectors at a limited number of locations. These efforts require significant labor and cost for expansions. Researchers and private sector companies have also explored alternative solutions such as probe vehicle data, while still suffering from a low penetration rate. In recent years, along with the technological advancement in mobile sensors and mobile networks, Mobile Device Location Data (MDLD) have been growing dramatically in terms of the spatiotemporal coverage of the population and its mobility. Three ways of estimating vehicle volumes are reviewed below.

Loop detectors are widely used to record traffic volumes and occupancy levels. These sensors are usually buried under the pavements to detect the induction change from the presence

of a vehicle. Kwon et al. 2003 developed an algorithm using data from single loop detectors to estimate truck traffic volumes [95]. The results showed a 5.7% error compared with the ground truth highway data. Loop detector data were also applied together with probe vehicle data to estimate queue length [96] and vehicle volume at a city-wide scale [97]. Although proven to be efficient in estimating vehicle volume, the high installation and maintenance cost of loop detectors limit their capability of being scaled up to cover the entire transportation network. Therefore, loop detector datasets are often incomplete and mostly unavailable at minor arterials and local streets.

In the past two decades, MDLD have gained significant attention and have been utilized for estimating various traffic characteristics, including vehicle volumes. With the development of MDLD, estimating vehicle volumes at the city scale became a reality. Probe vehicles can record their trajectory data with high granularity (i.e., 1Hz). Based on the trajectory data obtained from probe vehicles, a wide range of methods can be used by researchers to solve transportation problems. Zhao et al. proposed novel methods to estimate queue length and vehicle volume based on the probability theory without prior information about the penetration rate or queue length distribution [98]. Guo et al. estimated vehicle volume and queue length at signalized intersections and proposed a new framework to optimize traffic signal control operations [99]. Sekuła et al. applied several machine learning and neural networks to estimate historical hourly vehicle volume between sparsely located sensors based on the probe vehicle data [100]. Shockwave theories were also applied to probe vehicle data by a few studies [101, 102].

Many studies have been conducted focusing on estimating traffic flow and detecting congestion using cellular data [103, 104]. Xing et al. utilized CDR with the Time Difference of Arrival (TDOA) positioning technique in order to estimate multimodal traffic volumes on different types of urban roadways by identifying three modes of travel – namely, drive alone, carpooling,

and bus [105]. The results showed that compared with the ground truth vehicle volume obtained from License Plate Recognition (LPR) cameras, the mean relative error was in the range of 17.1% to 25.7%, depending on the roadway type. Despite significant advances in positioning techniques, cellular data still suffer from low accuracy issues, whereas LBS data have a noticeable advantage due to utilizing different sources to accurately locate the user – a feature that has resulted in increased usage of this type of data by researchers and the private sector for estimating vehicle volume. Fan et al. developed a computing framework alongside a heuristic map matching algorithm to estimate Vehicle Miles of Travel (VMT) and AADT for the state of Maryland using INRIX data [106]. The results showed an R^2 of 0.878 when fitting the estimated AADT with the ground truth AADT. Moreover, a number of state agencies conducted rigorous evaluations of vehicle volume obtained through traditional methods as well as from MDLD obtained by private sector companies. They found the latter to be a promising source for supplementing current surveys and traditional methods [107].

2.3.3 Pedestrian and Bicyclist Crashes Estimation Methods

According to the National Highway Traffic Safety Administration's *Traffic Safety Facts 2019 Report* indicates, in 2019, pedestrian and bicyclist fatalities accounted for nearly 20% of all traffic crash-related deaths in the U.S. [108]. In Maryland alone in 2019, 3,136 pedestrian crashes and 848 bicycle crashes occurred, where over 90% of pedestrian crashes and over 80% of bicycle crashes resulted in injuries or fatalities [109, 110]. Approximately one out of every four individuals killed in traffic crashes in Maryland was a pedestrian [109].

Studies on pedestrian and bicyclist safety issues are abundant. They identify key contributing factors to pedestrian- and bicyclist-involved crashes as well as suitable methodologies for crash frequency analysis. To address the fundamental issues typically associated with crash

frequency data, previous research studies have employed various methodologies to analyze pedestrian- and bicyclist-involved crash frequency. Many factors have been suggested to play a role in pedestrian and bicyclist crashes, including those representing pedestrian and bicyclist risk exposure [111-116]. land use and the built environment [113, 117-120], and sociodemographic/socioeconomic status [113, 117, 118, 120, 121]. Among those, one of the important factors is vehicle volume, which significantly correlates with the frequency of pedestrian and bicyclist crashes. In this case, vehicle volume estimated from the MDLD can be integrated into existing traffic safety modelling methods to estimate pedestrian and bicyclist crashes for all intersections to promote traffic safety analysis.

According to Lord and Mannering [122], one of the main issues characterizing crash frequency data is overdispersion, which happens when the standard deviation of the crash counts is considerably larger than the mean. The other issue that usually affects crash frequency data is having excess zeros, which happens when crash counts contain a significant number of zero values [116, 122]. To predict pedestrian and bicyclist crash frequency at intersections, Saad et al. [115] used bicycle crowdsourced data from Strava [123] and developed a negative binomial (NB) model. They found that the frequency of bicycle crashes at intersections was positively associated with intersection size, the intersection being a signalized intersection, the number of intersection legs being four (compared to three-legged intersections), as well as total entering vehicle volume. The study also indicated that the frequency of bicycle crashes at intersections was negatively associated with the presence of a bike lane at those intersections. Raihan et al. [116] used a zero-inflated negative binomial (ZINB) model to develop crash modification factors (CMFs) for bicyclist crashes in Florida's urban areas. They found that road design characteristics such as lane width and speed limit had positive effects on reducing bicycle crashes. Lower bicycle crash probabilities

on segments were associated with increased bicycle activity. However, increased bicycle activity was associated with higher bicycle crash probabilities at intersections. Increased bicycle crash probabilities at intersections were also associated with the number of bus stops within the intersection influence area as well. Ukkusuri et al. [117] examined the role of various built environment, land use, road network, and sociodemographic factors as well as key exposure measures including traffic volume, transit ridership, and proportion of nonmotorized trip-makers in the frequency of total, injury-causing, and fatal pedestrian crashes. The study employed NB and ZINB models to estimate crash frequency and found that increased numbers of total and/or fatal pedestrian crashes were associated with increased proportions of industrial and commercial land use, increased transit ridership, increased numbers of subway stations, increased proportions of intersections with four and five approaches, increased proportions of primary roads without access restriction, and increased number of lanes. Sanders et al. [119] employed Poisson regression to examine the role of various factors in pedestrian exposure at intersections as well as bicycle exposure at various road segments in Seattle, Washington. They found that variables representing population and land use (i.e., number of households, number of commercial properties, and the presence of a university near the intersection) were significantly associated with pedestrian exposure at intersections. Moreover, bicycle exposure was associated with the number of bicycle lanes on the road segment and land use variables such as the presence of a university or a school near the count location. The findings of that study provided insights into the factors affecting pedestrian and bicyclist risk exposure, which is a key contributing factor to pedestrian and bicyclist crashes. Jestico et al. [124] used a crowdsourced bicycling incident dataset for the Capital Regional District in British Columbia, Canada, to identify design attributes associated with unsafe intersections between multi-use trails and roads. NB regression was used to model the links

between the number of bicycle crashes and near-miss incidents and the infrastructure characteristics at multi-use trail-road intersections. The results showed that factors associated with bicycle incident frequency at multi-use trail-road intersections included bicycling volumes, vehicle volumes, and trail sight distance.

Many other studies also investigated factors affecting pedestrian and bicyclist safety risk exposure and modeled pedestrian- and bicyclist-involved crash frequency. The key contributing factors affecting pedestrian/bicyclist safety exposure and crash frequency that emerge from the literature include: sociodemographic and socioeconomic factors such as proportion of the population by race or age group [113, 117, 118, 120, 121]; land use and built environment factors such as population density, employment density, activity diversity, bus stop density, and ratio of residential, industrial, and commercial uses [113, 118-120]; and traffic- and travel-related factors such as vehicle, pedestrian, and bicycle volumes as exposure measures [111-116].

Further, the literature review reveals that the most prominent methodologies that have been applied to pedestrian and bicyclist crash frequency analysis are Poisson regression, negative binomial (NB) regression, zero-inflated Poisson (ZIP) regression, and zero-inflated negative binomial (ZINB) regression [111, 120, 122-125]. The Poisson regression is usually considered the starting point in crash frequency modeling [111]. Moreover, while the ZIP and ZINB regression methodologies have frequently been applied in empirical research to account for the preponderance of zeros observed in crash count data, the ZINB regression is applicable for count data that exhibit both overdispersion and excess zeros issues [116].

Table 2-3. Examples of Past Studies on Pedestrian and Bicyclist Safety Models

Study	Unit of Analysis	Study Area	Safety Measure	Methodology	Key Exposure Measure(s)
Ukkusuri et al. 2012 [117]	Census tract, zip code	New York City (NYC), NY	Total pedestrian crashes, severe crashes, and fatal crashes	NB, ZINB	Traffic volume, pedestrian activity, operating speeds
Hosseinpour et al. 2012 [111]	Road segment	Federal Road Network, Malaysia	Frequency of pedestrian crashes	Poisson, NB, ZIP, ZINB	Motorized traffic volume
Lee et al. 2015 [121]	Zip code	Various locations in FL	Pedestrian crashes per crash location zip code, crash-involved pedestrians per residence zip code	Bayesian Poisson lognormal simultaneous equations spatial error model	Log of population, log of vehicle miles traveled
Sanders et al. 2017 [119]	Intersection, road segment	Seattle, WA	Pedestrian and bicyclist counts	Poisson model	— [a]
Jestico et al. 2017 [124]	Multi-use trail intersection	Capital Regional District, British Columbia, Canada	Frequency of bicyclist crash and near miss incidents	NB	Pedestrian, bicyclist, and vehicle volumes
Xie at al. 2017 [113]	Grid cell (300×300 ft^2)	Manhattan (NYC), NY	Pedestrian crash cost	Tobit model	Vehicle miles traveled, taxi trips, subway ridership
Mansfield et al. 2018 [120]	Census tract	United States	Frequency of pedestrian fatalities	NB, ZINB, ZINB mixed model	Vehicle miles traveled density (thousand VMT/mi^2) by roadway functional class
Saad et al. 2019 [115]	Intersection	Orange County, FL	Frequency of bicycle crashes	NB	Total entering volume, bicycle volume
Raihan et al. 2019 [116]	Intersection, road segment	Urban areas, FL	Bicycle crash modification factors	ZINB	Bicycle activity (Strava volumes) [122]
Lee et al. 2019 [125]	Intersection	Orange and Seminole Counties, FL	Pedestrian crashes	NB, ZINB	Observed and predicted pedestrian trips

Notes: — [a]: This was an exposure study; therefore, the exposure measures were the response variables in the models (i.e., pedestrian and bicyclist counts).

Considering factors and methodologies used in exposure and crash analyses for vulnerable road users, Table 2-3 summarizes a few previous pedestrian and bicyclist safety studies. Overall, the literature review reveals that while pedestrian and bicyclist safety risk analyses are becoming more data-driven, usage of consistent and reliable exposure data such as crowdsourced big data in conducting pedestrian and bicyclist crash analyses remains scarce—particularly with regards to pedestrians. This study aims at addressing that gap in empirical research by utilization of mobile-device location big data in analysis of pedestrian and bicyclist crashes.

Chapter 3: Identification of Metrics Used for Various Levels of Traffic Analysis

3.1 Models, Tools, and Metrics for Various Levels of Traffic operations and Safety Analysis

This section reviews the state-of-the-practice models, tools, and metrics developed by State agencies such as the Department of Transportation (DOT) and Metropolitan Planning Organizations (MPOs) or universities for planning and designing transportation projects while considering systemic feasibility and efficiency, including for traffic operations and safety. Transportation project decisions require cooperative actions across various organizations, offices, and working groups within an organization when the plans cover different municipal areas or techniques governed by multiple authorities. Many different tools and methods are available to support the quantitative analysis of TSM&O and traffic operations strategies in planning and programming. Based on the U.S. Department of Transportation's (USDOT) Federal Highway Administration's (FHWA) Applying Analysis Tools in Planning for Operations report, the following tools can be used for analyzing strategies at various levels of the planning process.

- Sketch planning and prioritization tools for highway needs inventory (e.g., Tool for Operations Benefit-Cost Analysis – TOPS-BC, MOSAIC)
- Travel demand models (e.g., MSTM, BMC InSITE, MWCOG model) with postprocessors (e.g., Intelligent Transportation System (ITS) Deployment Analysis System – IDAS)
- Analytical tools (e.g., Highway Capacity Manual and traffic signal optimization tools).
- Microscopic simulation models (e.g., VISSIM, AIMSUN)
- Mesoscopic simulation models (e.g., DTALite, DynusT)

Table 3-1. State-of-the-Practice Models, Tools, and Metrics for Various Levels of Traffic operations and Safety Analysis

State/Agency	Model	Descriptions
Sketch-Planning Tools		
FHWA	ITS Deployment Analysis System (IDAS)	The objective of IDAS is to estimate the impacts and costs resulting from the deployment of various ITS components.
Northeastern Illinois	IDAS	IDAS is used to evaluate four types of ITS deployment: electric toll collection, freeway variable message signs, electric transit fare collection system, and transit vehicle signal priority.
Ohio-Kentucky-Indiana	IDAS	The components of Advanced Regional Traffic Interactive Management Information System (ARTIMIS) are evaluated using IDAS, including closed-circuit TV cameras, electronic dynamic message signs, traveler advisory telephone service, highway advisory radio, freeway service patrol vans, ramp and reference makers, vehicle detectors, total station electronic surveying equipment and operations control center.
Michigan	IDAS	The components of Temporary Traffic Management System (TTMS) are investigated, including closed-circuit TV cameras, portable dynamic message signs, detection devices for traffic queueing and construction zones, video monitoring stations, telephone/web-based traveler information, and a traffic management center.
Florida DOT	Florida Standard Urban Transportation Model Structure (FSUTMS)	FSUTMS can produce various performance measures including vehicle miles of travel, vehicle hours of travel, average speed, number of accidents, fuel consumption, monetary benefits to users and/or agency, and emissions.
CalTrans	California Life-Cycle Benefit/Cost (Cal-B/C)	Cal-B/C uses a set of spreadsheet-based tools that cover multi-modal analysis of highway, transit, bicycle, pedestrian, ITS, operational improvement, and passenger rail projects.
University of South Florida	Trip Reduction Impacts of Mobility Management Strategies (TRIMMS)	TRIMMS allows quantifying the net social benefits of a wide range of transportation demand management initiatives in terms of emissions reductions, accident reductions, congestion reductions, excess fuel consumption, and adverse global climate change impacts by estimating changes in travel behavior.
New York State DOT	ITS Options Analysis Model (ITSOAM)	ITSOAM has three components including Delay Model, Safety Model, and Environmental Benefits Model.
Post-Processing Analysis		

Florida DOT	Integrated Regional Information Sharing and Decision Support System (IRISDS)	IRISDS is a web-based platform that provides decision support for estimating and predicting system performance using data mining techniques, traffic analysis, and simulation modeling.
Florida DOT	Florida ITS Evaluation (FITSEVAL)	FITSEVAL evaluates the benefits and costs of thirteen different ITS deployment alternatives and can assess the mobility, safety, environmental, and monetary benefits and produces estimates of the present-worth and benefits-cost ratios of ITS.
Florida DOT	ITS Data Capture and Performance Management (ITSDCAP)	ITSDCAP conducts ITS evaluations based on ITS data and four types of ITS can be evaluated including incident management, ramp metering, smart work zone, and road weather information system.
Virginia DOT	Virginia System Operations Performance Reports (VSOPR)	VSOPR assesses four categories of measures including Traffic, Incidents, Traveler information, and ITS device reliability.
Wisconsin DOT	Summary of ITS evaluation methods	The evaluation process consists of nine steps and assesses four types of measures, including Performance metrics, Benefits valuation measures, Net benefits, and B/C ratio.
Multi-Dimensional Models		
Florida DOT	FITSEVAL	FITSEVAL uses the output of the FSUTMS modeling environment under CUBE, which quantified Congestion/Mobility, Safety, Environmental and energy, and Agency and user costs measures.
Oregon DOT	Analysis and modeling tools	Statewide Integrated Model (SWIM), SWIM2, Land Use Scenario DevelopR (LUSDR), DTA, VISSIM, etc.
Maryland DOT	InSITE ABM-DTALite and SILK AgBM-DTALite	InSITE ABM-DTALite is the result of integrating a DTA tool based on an existing DTALite model that covers the InSITE ABM. SILK AgBM-DTALite is an agent-based microsimulation travel demand model.
Maryland DOT	Maryland Integrated Travel Analysis Modeling System (MITAMS)	MITAMS has a special focus on various applications ranging from short-term and long-term applications.
Ohio DOT and Kentucky Transportation Cabinet	ARTIMIS	ARTIMIS aims to optimize freeway system efficiency, improve safety and benefit air quality. It includes over 80 cameras, 57 center-lane miles of fiber-optic cable, approximately 1100 detectors, and numerous freeway message signs in Cincinnati.
University of Florida	Corridor Simulation (CORSIM/TSIS)	The Traffic Software Integrated System (TSIS) integrates with the microscopic TRAF tools of CORSIM, namely FRESIM for freeway simulation and NETSIM for surface arterials and network simulation.

Based on the FHWA's Operations Benefit/Cost Analysis Desk Reference, there are in general three types of tools:

- **Sketch-planning tools** can provide a simple, quick, and low-cost estimation of operational strategy benefits and costs. Examples include spreadsheets that rely on generally available data as Ill as static cause-effect relations between strategies and their impacts. Usually, these are inexpensive to use but have a high inaccuracy or risk.
- **Post-processing analysis tools** seek to link the evaluation of operations with the travel demand, network data, and performance measure outputs from regional travel demand and simulation models. They are often more capable of assessing the impacts of the route, mode, or temporal shifts than sketch-planning methods but tend to cost more.
- **Multi-dimensional models** are the most complex and costly, but typically provide a high level of confidence in the accuracy of the results. They are often used to integrate various analyses (e.g., a travel demand model and a DTA simulation) to estimate the full range of impacts of operations strategies or transportation projects.

Table 3-1 summarizes the state-of-the-practice models, tools, and metrics used by various DOTs and MPOs. These models and tools usually largely rely on traditional transportation data collection methods, such as loop detectors and manual counting.

3.2 Operations Practice Scan Survey

Based on the literature review in Section 3.1, the selection of analysis, modeling, and simulation tools, and the corresponding performance metrics vary during each stage of the transportation planning and operations process and should serve analytical purposes. At the long-range planning stage, it is impractical to apply the most complex tool for each conceived traffic operations project. Sketch planning tools or travel demand model postprocessing tools may be more suitable. At the

Transportation Improvement Program (TIP) and project planning stages, mesoscopic and microscopic traffic simulation tools may be considered for traffic operations project studies. Multi-scenario and multi-resolution tools for estimating travel reliability impact under different weather and accident conditions may also be added at these stages to provide more comprehensive information to support decision-making. Post-project evaluation could rely on existing performance monitoring dashboard tools such as the Regional Integrated Transportation Information System (RITIS).

To obtain a standard workflow for prioritizing tools and metrics for transportation planning and operations, a dedicated survey is designed to collect insights from stakeholders including transportation practitioners from federal, state, and local agencies and other private-sector professionals with experience with performance evaluation of transportation projects. The objective of the survey is to help understand what performance measures are needed to make decisions at the planning, construction, and operations stages of a transportation project. Different types of projects and the level of analyses and metrics required to make reasonable recommendations have been identified and reviewed. A flowchart framework that documents the best practice metrics used in evaluating projects for different stages of planning and operations processes was produced to support transportation planners and engineers in their decision-making.

In this survey, three major stages of the general transportation project are identified, including: Feasibility & Planning, Design & Construction, and Maintenance & Operations. Under each stage, various performance metrics have been used to evaluate the project are listed. Based on these presumptions and categorization, the survey questions focus on understanding (1) best practices in performance metrics chosen for the evaluation of projects at different stages of the traffic operations planning process; (2) the usefulness of the metrics; and (3) challenges and

potential solutions for data and additional metrics that would offer insights. The survey collected 78 usable responses from the web-based survey.

3.3 Survey Results

Table 3-2. States for which the Respondents Work.

State	Number of Respondents	State	Number of Respondents
Georgia	1	Mississippi	1
North Carolina	1	Nebraska	2
Maryland	50	Pennsylvania	3
South Carolina	1	Washington	3
Virginia	3	Wyoming	2
Maryland, Virginia, District of Columbia	3		

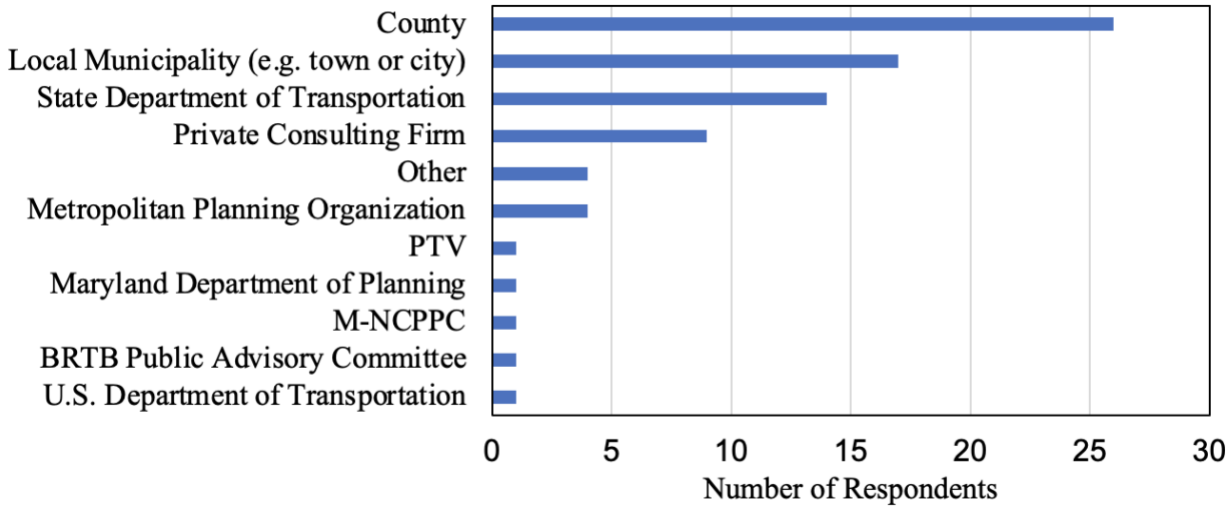


Figure 3-1. Agencies that the Respondents Work at.

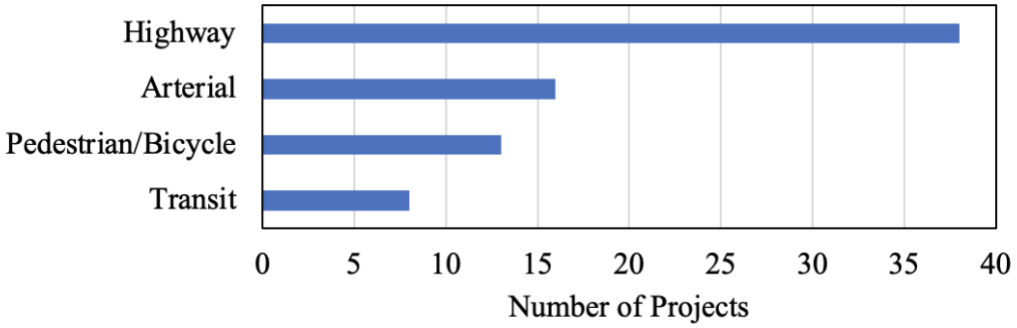


Figure 3-2. Projects that the Respondents Work on.

The detailed survey and results are documented in Appendix I. As shown in Table 3-2, Figure 3-1, and Figure 3-2, in total there are 79 respondents completed the survey, either filled online distributed by Maryland Department of Transportation State Highway Administration (MDOT SHA) or filled in-person using electronic devices during the Transportation Research Board Annual Meeting (TRBAM). Table 3-2 shows the states for which the respondents are currently working (nine of them did not select locations). Figure 3-1 summarizes the agency distribution of the respondents from across the U.S. Most of the respondents are from Counties (26), Local Municipalities (17), State Departments of Transportation (14), and Private Consulting Firms (9), while the remaining (13) are from other organizations. As shown in Figure 3-2, the respondents have mixed backgrounds, with 37 of them working most frequently on highway projects, 16 on arterial projects, 13 on pedestrian and bike projects, and 8 on transit-related projects (four of them did not select projects).

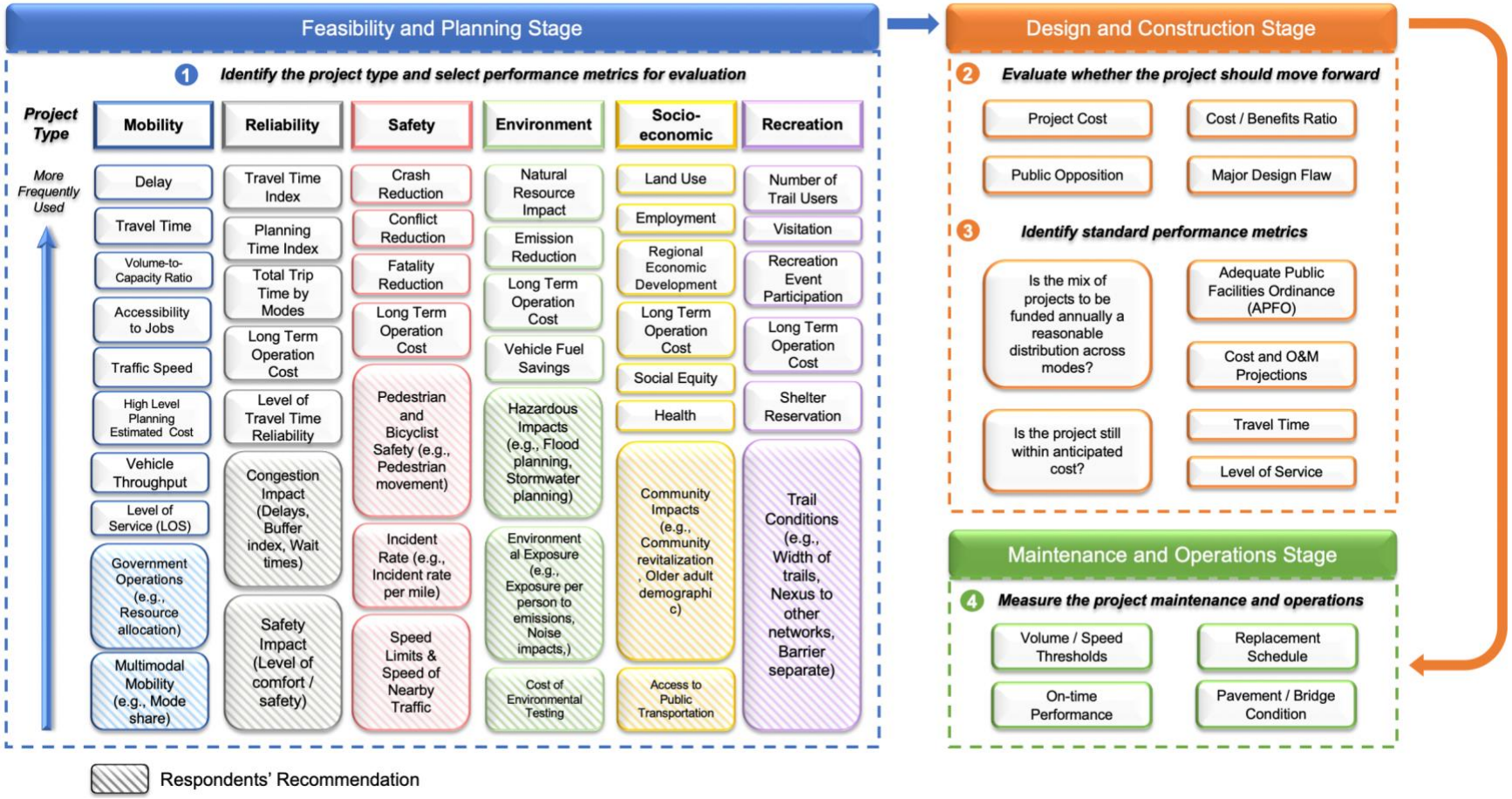


Figure 3-3. Projects that the Respondents Work on.

3.4 Performance Metrics Flowchart

Based on the survey results from 79 respondents, a flowchart (see Figure 3-3) is produced that documents the best practice performance metrics used in prioritizing and evaluating transportation projects. The complete survey questionnaire and results can be found in Appendix I.

In the Feasibility and Planning stage, the project type was further refined to reflect industry needs, knowing that best-practice metrics would differ structurally within different project types. Six types of transportation projects were identified separately:

- Mobility: focused on reducing congestion delays, typically capacity improvements, micro-mobility infrastructure, transit solutions, etc;
- Reliability: focused on maximizing existing operations, such as technology deployments to manage the transportation system more effectively;
- Safety: focused on systematically and holistically promoting safety, using metrics such as the severity of crashes, high rate of crashes, vulnerable user interactions with vehicles, and freight design concerns;
- Environmental: focused on managing environmental impact, sustainability, energy/emissions, and public health. Metrics could also include stream restoration and flooding mitigation;
- Socio-economic: focused on economic revitalization, food desert programs, equity-related, etc.;
- Recreational: focused on trails, visitor rest stops, etc.

Based on these predefined project types, typical performance metrics used for each type of project were reviewed and listed in the survey questions to facilitate the post-processing of responses. Respondents were then asked to rank the frequency of these performance metrics when

performing a planning-level analysis or feasibility assessment. In case metrics were missing from the list, the respondents were asked to fill in an open-ended section with the metrics they felt were relevant to the question. As shown in the flowchart, during the feasibility and planning stage, more frequently used performance metrics were identified based on the respondents' responses. Below is a summary of the major findings:

- For mobility-related projects, "Delay", "Travel Time" and "Volume-to-Capacity Ratio" are the three most frequently used performance metrics. Respondents also suggest metrics such as "Government Operations (e.g., Resource allocation, Master plan conformance, Equipment availability)" and "Multimodal Mobility (e.g., Mode share, Transfer time, Bicycle network).
- For reliability-related projects, in addition to the commonly used "Travel Time Index", the "Planning Time Index" and "Total Trip Time by Modes" are also frequently used. Respondents also suggest metrics such as "Congestion Impact (e.g., Delays, Buffer index, Wait times)" and "Safety Impact (e.g., Level of comfort/safety)".
- For safety-related projects, the typical performance metrics include "Crash Reduction", "Conflict Reduction" and "Fatality Reduction". Respondents emphasized the importance of "Pedestrian and Bicyclist Safety (e.g., Pedestrian movement)". Some other frequently used performance metrics mentioned by the respondents include "Incident Rate (e.g., Incident rate per mile)" and "Speed Limits & Speed of Nearby Traffic".
- For environment-related projects, "Natural Resource Impact" and "Emission Reduction" are deemed frequently used. Respondents also suggest "Hazardous Impacts (e.g., Flood planning, stormwater planning)", "Environmental Impacts (e.g., Exposure per person to emission, Noise impact)", and "Cost of Environmental Testing".

- For socio-economic-related projects, “Land Use”, “Employment”, and “Regional Economic Development” are deemed frequently used. Respondents also suggest “Community Impacts (e.g., Community revitalization, Older adult demographic)” and “Access to Public Transportation”.
- For recreation-related projects, “Number of Trail Users”, “Visitation” and “Recreation Event Participation” are deemed necessary for decision-making. Respondents also suggest “Trail Conditions (e.g., Width of trails, Nexus to other network, Barrier separation)”.

In the Design and Construction stage, the actual design and construction plan for the project should be the main consideration. Therefore, when the project moves to the Design and Construction stage, most performance metrics are used to determine whether the project has critical failure points. In the survey, respondents are asked to rank the performance metrics used to examine project performance. The survey results helped identify four major performance metrics to support the decision-making, including “Project Cost”, “Cost/Benefits Ratio”, “Public Opposition”, and “Major Design Flaw”. Furthermore, a list of standard performance metrics was also suggested by respondents based on their own experience. These standard performance metrics included “Is the mix of projects to be funded annually a reasonable distribution across modes?”, “Is the project still within anticipated cost?”, “Adequate Public Facilities Ordinance (APFO)”, “Cost and O&M Projects”, “Travel Time”, and “Level of Service”.

During the Maintenance and Operation stage, the focus shifts to how to maintain and operate the project at the expected levels. Based on the results, these can be measured with “On-time Performance”, “Alternative Routes”, “Bridge Condition”, “State of Good Repair”, “Age of Transit Flee”, “Surface Condition”, “Signage Availability”, “Sufficient Funding”, “Clear Making (e.g. marking for crosswalks, travel lanes)”, “Reporting Issues” and “Priority Lists”.

3.4 Summary

In summary, this chapter presents a dedicated survey to help understand what performance measures are needed to make decisions at the planning, construction, and operations stages of a transportation project. Different types of projects and the level of analyses and metrics required to make reasonable recommendations are identified and reviewed. A flowchart framework that documents the best practice metrics used in evaluating projects for different stages of planning and operations processes is produced to support transportation planners and engineers in their decision-making.

Among six types of projects, "Mobility" and "Safety" projects are the two types that are closely related to traffic operations and safety research which aims at congestion reduction and safety improvement. Based on the flowchart, it can be seen that for "Mobility" projects, delay, travel time, and volume-to-capacity ratio metrics are deemed most frequently used. For "Safety" projects, crash/fatality rate and crash/fatality reduction are of the greatest concerns to transportation professionals when planning transportation projects. In order to quantify the importance of these performance metrics, mainly three types of upstream data are needed, namely vehicle volume, travel time (or traffic speed), and crash rate (or crash count). As indicated in the literature, instead of collecting data manually or using traditional technologies, transportation big data can be used to estimate these upstream data. In this dissertation, probe vehicle data in RITIS and large-scale MDLD are used to develop big-data driven frameworks to support crash-related decisions, estimate vehicle volumes, and pedestrian and bicyclist crashes, which ultimately support "Mobility" and "Safety" transportation projects.

Chapter 4: Supporting Triage Decisions for High-Risk Trauma Patients at Crash Sites with Location Data

4.1 Introduction

As identified in Chapter 3, when planning for "Safety"-related transportation projects, the most considered, and important metrics are crash/fatality rate and crash/fatality reduction. This chapter focuses on incorporating one type of transportation big data, probe vehicle data, as well as the annual average daily traffic data to support crash-related decisions. More specifically, to estimate the emergency medical services (EMS) and trauma triage decisions at crash scene in order to reduce fatality rate caused by severe injuries.

Although the research on improving EMS efficiency marks contemporary healthcare service and traffic safety discussions, its roots can be traced back to the 2000s, when epidemiologists, health researchers, and practitioners stressed the vulnerability of the old age persons who are involved in traffic crashes [126-128]. When a vehicle crash occurs, the decision for EMS and field trauma triage must be made in a timely manner to save lives, especially for elder persons [129]. According to a Population Bulletin, "Aging in the United States," _the number of US seniors (age 65+) is projected to nearly double from 52 million in 2018 to 95 million by 2060, and the 65-and-older age group's share of the total population will rise from 16 percent to 23 percent. As they have passed through each major stage of life, baby boomers have brought both risks and challenges to the transportation, infrastructure, and healthcare institutions.

"Under-triage" often occurs in the medical decision process, where a large proportion of seriously injured older patients are transported to non-trauma hospitals or fail to be transported at all [89, 130-134]. This leads to a significant mismatch between the supply side from hospitals and

the demand side from those patients. Therefore, the gap not only degrades the health outcomes but also imparts the whole Illbeing from a longer perspective [75, 76]. On the other side, transferring the under-triaged patients to trauma centers inappropriately, might also waste time and public resources which could be used to help other crash victims. Thus, it is important for policymakers, health-related practitioners and scholars to be equipped with a tool for better evaluating and analyzing the efficiency of the current EMS system in the era of data-driven problem-solving.

4.2 The Big-Data Driven Framework Integrating Transportation and Health Data

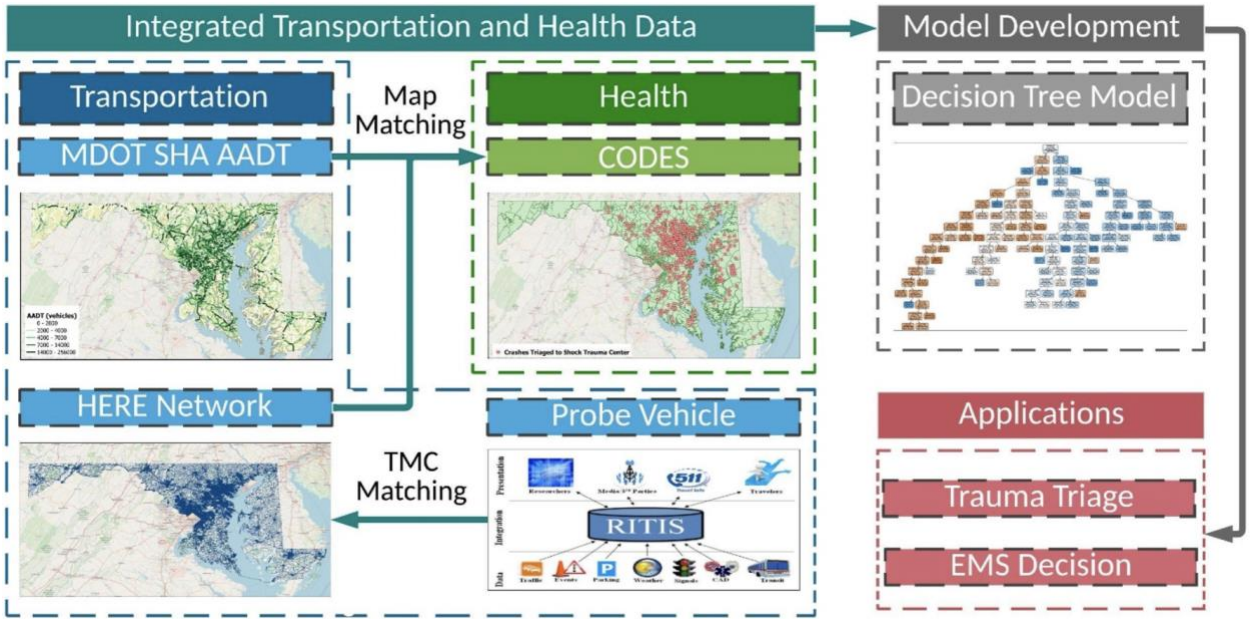


Figure 4-1. The Big-Data Driven Framework for Integrating Transportation and Health Data.

The overall big-data driven framework is illustrated in Figure 4-1. It consists of two main pillars. Pillar 1 is the integration of the transportation big data and health datasets (shown on the left of Figure 4-1). Health and safety-related data, such as crash, EMS, and hospital triage records, is typically stored in Crash Outcomes Data Evaluation System (CODES) in Maryland. These CODES data are mapped to the roadway network (HERE network is used for its availability) using their geolocation information and then joined with datasets from the transportation sector through

a spatiotemporal map matching. The two most important transportation datasets, i.e., the traffic volumes and the historical time-dependent travel speed, are obtained at the roadway segment level from the Annual Average Daily Traffic dataset (AADT) and large-scale probe vehicle data sources (available from RITIS, the Regional Integrated Transportation Information System). These two transportation datasets are then matched to the network using the Traffic Message Channel (TMC).

With the integrated dataset, information beyond crashes themselves becomes available, including the EMS involvement, hospital triage, as well as traffic and roadway conditions at the crash scene (volumes, travel speed, weather, road surface conditions, etc.). Pillar 2 of the framework employs the integrated dataset to evaluate decision-making. Classification models of decision trees are developed to model EMS and trauma decisions. This big-data driven framework enables the analysis of a wide spectrum of modeling methods in various application contexts.

4.2.1 Integrated Transportation and Health Data

In this research, CODES dataset was with transportation data, including the roadway network, link level volume information (i.e. Annual Average Daily Traffic, AADT), and link level observed travel speed information obtained via probe vehicle information. In this section, the three major data sources are described.

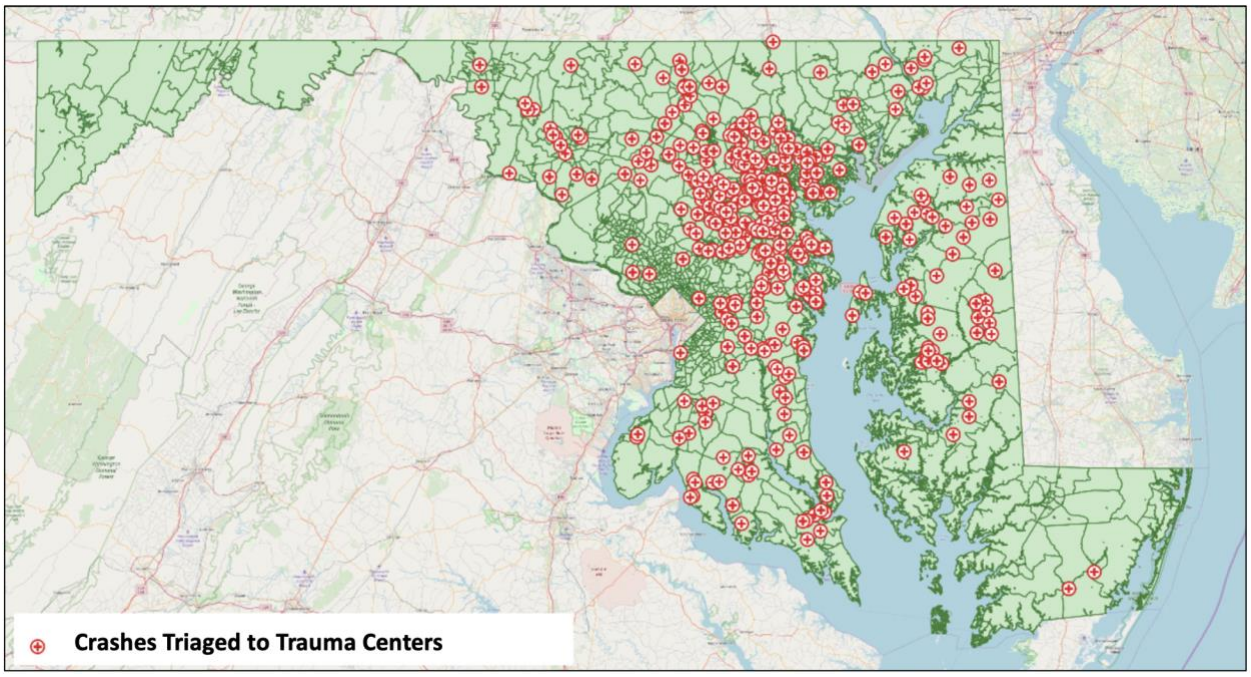


Figure 4-2. CODES Data in the state of Maryland.

Table 4-1. Descriptive Statistics of the CODES Data

Variable Name	Type	Description
sex	Binary	Gender 1 = female (53.5%); 2 = male (45.8%); 99 = unknown (0.7%)
speed_limi	Categorical	Speed limit 5mph (2.5%); 10mph (2.3%); 15mph (3.3%); 20mph (0.9%); 25mph (16.2%); 30mph (13.9%); 35mph (16.8%); 40mph (13%); 45mph (8.2%); 50mph (7.2%); 55mph (11.9%); 60mph (0.3%) 65mph (2.8%); 70mph (0.5%)
age	Numeric	Age min=65; max=109; mean=73; standard deviation=6.9
eldveh	Binary	If the person shared a vehicle with an elderly person A constant value of 1
damage	Binary	If the vehicle is disabled or destroyed due to the crash 1=yes (43.8%); 0=no (56.2%)
eject	Binary	If the person is ejected 1=yes (0.4%); 0=no (99.6%)
notbelted	Binary	If the belt is used improperly 1=yes (2.7%); 0=no (97.3%)
light_code	Categorical	Light condition 0=Not Applicable (1.3%); 1=Daylight (77%); 3=Dark Lights On (13%) 4=Dark No Lights (3.9%); 5=Dawn (1.4%); 6=Dusk (2.4%) 7=Dark- Unknown Lighting (0.7%); 88=Other (0.2%); 99=Unknown (0.2%)
collision_	Categorical	Collision type 0= Not Applicable (1.3%); 1=Head On(2.6%); 2= Head On Left

		Turn(7.2%); 3=Same Direction Rear End(29.7%); 4= Same Direction Rear End Right Turn(3.9%); 5= Same Direction Rear End Left Turn(1.4%); 6= Opposite Direction Sideswipe(1.3%); 7= Same Direction Sideswipe(6.5%); 8= Same Direction Right Turn(2.3%); 9= Same Direction Left Turn(2.5%); 10= Same Direction Both Left Turn(0.4%); 11= Same Movement Angle(20.3%); 12= Angle Meets Right Turn(0.6%); 13= Angle Meets Left Turn(0.7%); 14= Angle Meets Left Turn Head On(0.4%); 15= Opposite Direction Both Left Turn(0.2%); 17= Single Vehicle (11.8%); 88=Other (10.8%); 99=Unknown (0.3%)
harm_event	Categorical	Subjects involved in accidents 0=Not Applicable (0.9%); 1=Other Vehicle (78.1%); 2=Parked Vehicle (4.5%); 3=Pedestrian (3.9%); 4=Bicycle (0.6%); 5=Other Pedalcycle (0.05%); 6=Other Conveyance (0.05%); 7=Railway Train (0.004%); 8=Animal (0.8%); 9=Fixed Object (7.2%); 10=Other Object (0.5%) ; 11=Overturn (0.1%); 12=Spilled Cargo (0.03%); 13=Jackknife (0.01%); 14=Units Separated (0.01%); 15=Other Non-Collision (0.2%); 16=Off Road (1.5%); 17=Downhill Roadway (0.01%); 18=Explosion or Fire (0.1%); 19=Backing (0.09%); 20=U-turn (0.01%); 21.15=Immersion (0.01%); 22.15=Fell Jumped from Motor Vehicle (0.04%); 23.15=Thrown or Falling Object (0.06%); 88=Other (0.3%); 99=Unknown (0.08%)
belt_use	Categorical	Type of belts 1=Combined lab-shoulder protection (83.1%); 2= shoulder only (7.9%); 8= lap only (7.7%); 0=no restraint use (1.2%)
emstrans (Output)	Binary	EMS decision 1=sent via EMS (18%); 0=not transported via EMS (82%)
trauma (Output)	Binary	Trauma triage decision 1=sent to trauma center (0.8%); 0=not sent to trauma center (99.2%)

1) CODES data: The crash data are collected from Crash Outcome Data Evaluation System (CODES). The dataset includes crash scene information of car crashes that involve people 65 years old or older. Figure 4-2 illustrates a key CODES data studied by this research, i.e., crashes triaged to a trauma center. This decision variable is later on employed in the decision tree modeling practices using the integrated transportation and health data. A previous study showed that CODES offers great promise as a mechanism to guide triage decisions [135]. Table 4-2 summarizes the CODES data used in this study, which has 54,826 observations from 2015 to 2017. The CODES data is provided by National Study Center for Trauma/EMS at the University of Maryland

Baltimore (UMB) through the National Highway Traffic Safety Administration (NHTSA) CODES. The provided data has been filtered by UMB to only include accidents involving person who is more than 65 years old to focus on elder population. This study establishes models to support the triage decisions as well as EMS transport decisions (summarized in the last two rows in Table 4-2). It aims to support short-term EMS/triage decisions even before police or medical personnel arrive at the scene. Therefore, features such as police-reported injury severity and fatality have been removed to comply with data privacy laws and regulations. Studied features include age, gender, local speed limit, light condition, collision type, harm events, usage of the belt, whether the vehicle carries elder people, whether the vehicle is damaged, and whether the person is ejected.

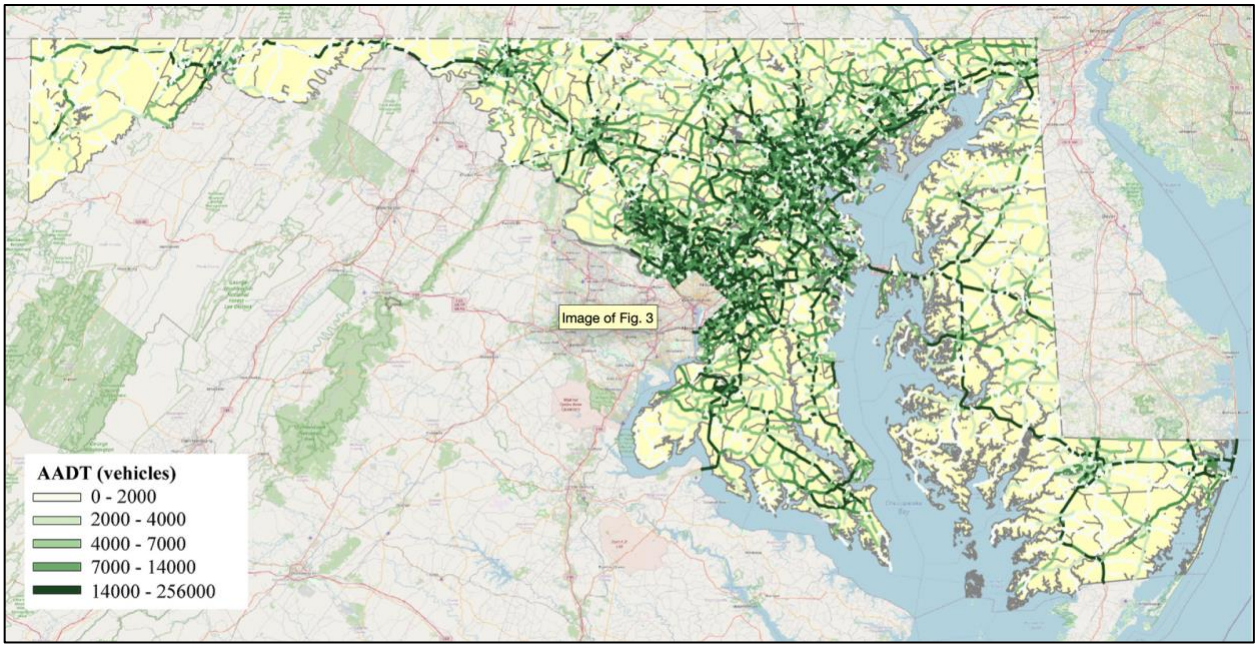


Figure 4-3. Annual Average Daily Traffic in the state of Maryland.

2) Annual average daily traffic (AADT) data: Annual average daily traffic (AADT) is a measure used primarily in transportation planning, representing the total volume of vehicle traffic of a highway or road for a year divided by 365 days. In transportation planning applications, AADT can serve as a critical criterion for project selection, pavement design, capacity analysis, and safety analysis. In this study, ground truth Annual Average Daily Traffic (AADT) data is collected from

the Maryland Department of Transportation (MDOT) State Highway Administration (SHA). Figure 4-3 illustrates the link-level AADT in the state of Maryland. Usually, major interstate freeway and highway segments carry the highest AADT values, followed by major US and state routes. Local streets and neighborhood roads usually accommodate lower AADT.

3) Large-scale probe vehicle data: With increased vehicle connectivity through in-vehicle GPS and sensors, automobiles on the road provide real-time traffic and travel speed information about the links and intersections they just traversed [136]. These probe vehicle data are made available via several data vendors (e.g., HERE, TomTom, and INRIX) and validated to estimate link speeds and travel times. This data source includes useful information about the historical speed at the scene when the crash occurred.

4.2.2 Modeling Method

In this chapter, decision tree models are trained, cross-validated, and then used to predict the outcomes of EMS and trauma triage decisions [137]. A decision tree predicts by using yes-or-no splits at each tree node and then splits samples until a certain criterion is met. Its nonlinearity can identify hidden patterns in the dataset, and its nonparametric nature promises increasing complexity as more data become available. To compare the effectiveness, four models are built and evaluated:

- Benchmark 1: EMS decision tree with CODES information only.
- Model 1: EMS decision tree using integrated transportation and health data.
- Benchmark 2: Trauma triage decision tree with CODES information only.
- Model 2: Trauma triage decision tree using integrated transportation and health data.

The entire process has four components, including data cleaning, feature selection/transformation, parameter finetuning, model selection, model valuation, and prediction

deployment. Building a pipeline makes it easy to reproduce the same procedures for a different dataset. In this case, the major purpose of using a pipeline is to allow the implementation of target encoding without data leakage. This process will greatly simplify the cross-validation process, which is detailed below. The entire process includes data cleaning, feature selection/transformation, parameter finetuning, model selection, model valuation, and prediction deployment. Building a pipeline makes it easy to reproduce the same procedures for a different dataset. In our case, the major purpose of using a pipeline is to allow the implementation of target encoding without data leakage. This process will greatly simplify the cross-validation process.

1) Class Imbalance: For each target variable (EMS decision and trauma decision), a benchmark model is built using data from CODES, and then another model is built with additional transportation-sector data for comparison. Since the data have demonstrated a medium to severe level of class imbalance (EMS to non-EMS ratio is about 1:4, and trauma triage to non-trauma ratio is about 1:100), minority classes are oversampled, and majority classes are down-sampled to counteract the class imbalance. The over/down-sample of the minority/majority classes is achieved by doing randomized sampling from each group with replacement. Based on the original class distribution, a class ratio of 1:1 is achieved after the over/down-sampling process for each modeling practice. This approach can ensure that the parameter finetuning process of the decision tree modeling will not lean towards the majority class(es) and consequently affect the prediction accuracy for minority class(es) [138].

2) Feature Transformation: The dataset involves many categorical features of high cardinality, to which decision tree models give less preference over continuous features. Thus, feature transformation is needed. In this study, a handful of features are selected as mentioned previously. These features include several categorical features, which only bring a mild increase

in dimensionality. In this case, standardized one-hot encoding (i.e., 1-of-*K* scheme) can be used [139]. However, one-hot encoding, as a general method for encoding, fails to give reasonable results due to high cardinality. The solution adopted here is target-encoding [140]. By this method, categorical features are replaced by a mixture of the posterior probability of the target given categorical values and the prior probability of the target in the training set. Note that such a feature transformation needs to first fit on the training set and then the same transformation has to be applied on the validation/test set to avoid data leakage.

3) Parameter Finetuning: Four major hyper-parameters of decision trees are fine-tuned, including the maximum depth to which a tree can grow, the maximum number of features considered when splitting a tree leaf node, the minimum number of samples required for splitting, and the minimum number of samples required to remain at a leaf node. They are tuned by grid search cross-validation and then post-pruning is conducted to avoid overfitting and to increase interpretability. Finally, the model is re-evaluated by cross-validation on the entire dataset and average cross-validation results are reported.

4) Model Selection/Evaluation: The performance is evaluated using standard metrics for machine learning models, including accuracy, AUROC (area under the receiver operating characteristic curve), F1 score, precision, and recall [141]. Each metric evaluates the model from a different perspective. For our case, reported precision, recall, and F1 score are the most important metrics to evaluate performance in classifying the positive class (EMS and trauma). Precision measures how confident the classifier is when it classifies a crash accident as positive, while recall measures the classifier's ability to detect the positive class. The F1 score is the harmonic product of precision and recall, representing the general performance regarding the positive class.

The entire process includes data cleaning, feature selection/transformation, parameter finetuning, model selection, model valuation, and prediction deployment. Building a pipeline makes it easy to reproduce the same procedures for different dataset. In our case, the major purpose of using pipeline is to allow the implementation of target encoding without data leakage. This process will greatly simplify the cross-validation process.

4.3 Results and Discussions

The field triage process should aim at differentiating, as much as possible, crash victims who need and do not need EMS/trauma care. Different decision tree evaluations are presented in Table 2. In the statistics for machine learning, proper triage is reflected by the True Positive (TP) and True Negative (TN). The rate of under-triage is represented by the False Negative (FN), indicating that samples that need EMS/trauma care have been under triage. *Recall* (or sensitivity in clinical statistics) is used to assess under triage. A higher recall rate of a model indicates fewer under-triage situations. Similarly, False Positive (FP) represents an over-triage situation that provides unnecessary EMS/trauma to patient. *Precision* is used to measure over triage. A higher precision rate of a model indicates less over-triage situations.

Table 4-2. Decision Tree Model Evaluations

Decision Tree Model Validation	**Predicted EMS/Trauma**	**Predicted Non-EMS/Non-Trauma**	
Observed EMS/ Trauma	True Positive (TP); Proper Triage Reflects *Sensitivity*	False Negative (FN); Under-Triage	$Recall = \frac{TP}{TP + FN}$
Observed Non-EMS/Non-Trauma	False Positive (FP); Over-Triage	True Negative (TN); Proper Triage	
	$Precision = \frac{TP}{TP + FP}$		

Table 4-3. Model Performance Measures and Comparison

	EMS Decision		Trauma Triage Decision	
	Benchmark 1	**Model 1**	**Benchmark 2**	**Model 2**
Accuracy[1]	0.789	0.884	0.891	0.915
AUROC[2]	0.853	0.898	0.942	0.948
F1	0.816	0.900	0.847	0.880
Precision	0.752	0.823	0.798	0.830
Recall	0.891	0.995	0.901	0.937

1. All performance measures are averaged across a 5-fold cross-validation
2. Area Under Receiver Operating Characteristic Curve

There is often a tradeoff between recall and precision. Improving recall and mitigating under triage essentially save lives. On the other hand, reducing over-triage (i.e. higher precision) leads to more effective usage of hospital resources. In this analysis, by using the integrated transportation and health data, both the recall and precision of the decision tree models are significantly improved. Table 4-2 summarizes all performance measures of the decision tree models. As demonstrated in Table 4-3, both the EMS model and the trauma triage model have been significantly improved after transportation-sector data has been integrated. All selected performance measures show positive change, compared to the benchmarks. The recall rate of EMS decisions has been improved by over 10% to 0.995, indicating that 99.5% of the instances who need EMS service are classified correctly by our model after the integration. While EMS decisions enjoy more significant improvement, the trauma triage decision model performance exhibits moderate enhancement, with its recall rate increased from 0.901 to 0.937. Nevertheless, that indicates 3.6% of crash victims who could be under triaged would be sent to trauma centers appropriately with the updated triage decision model. Meanwhile, the precision rates have also increased to above 0.8 for both cases.

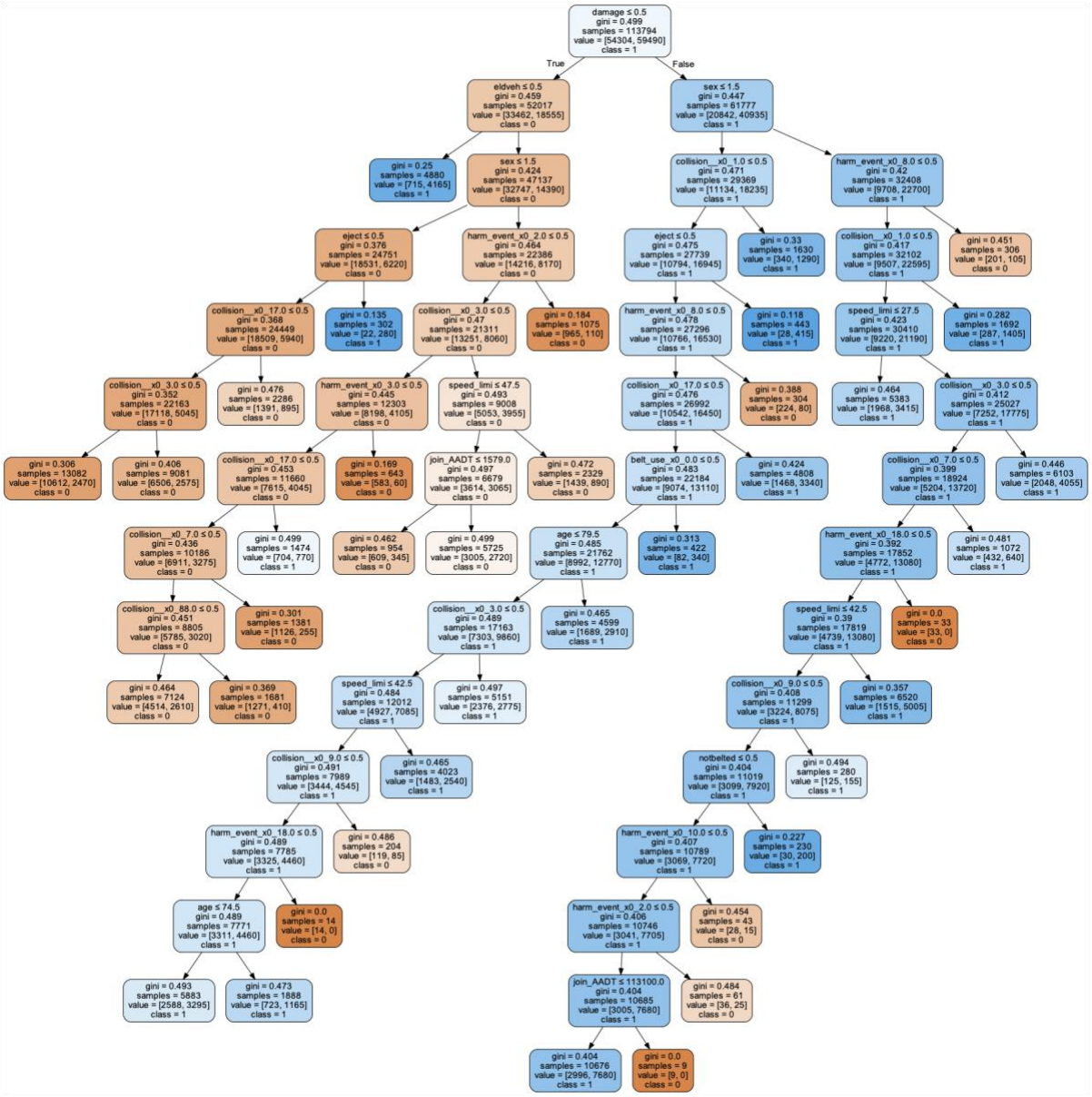


Figure 4-4. The Decision Tree Model of EMS Triage Using the Integrated Data

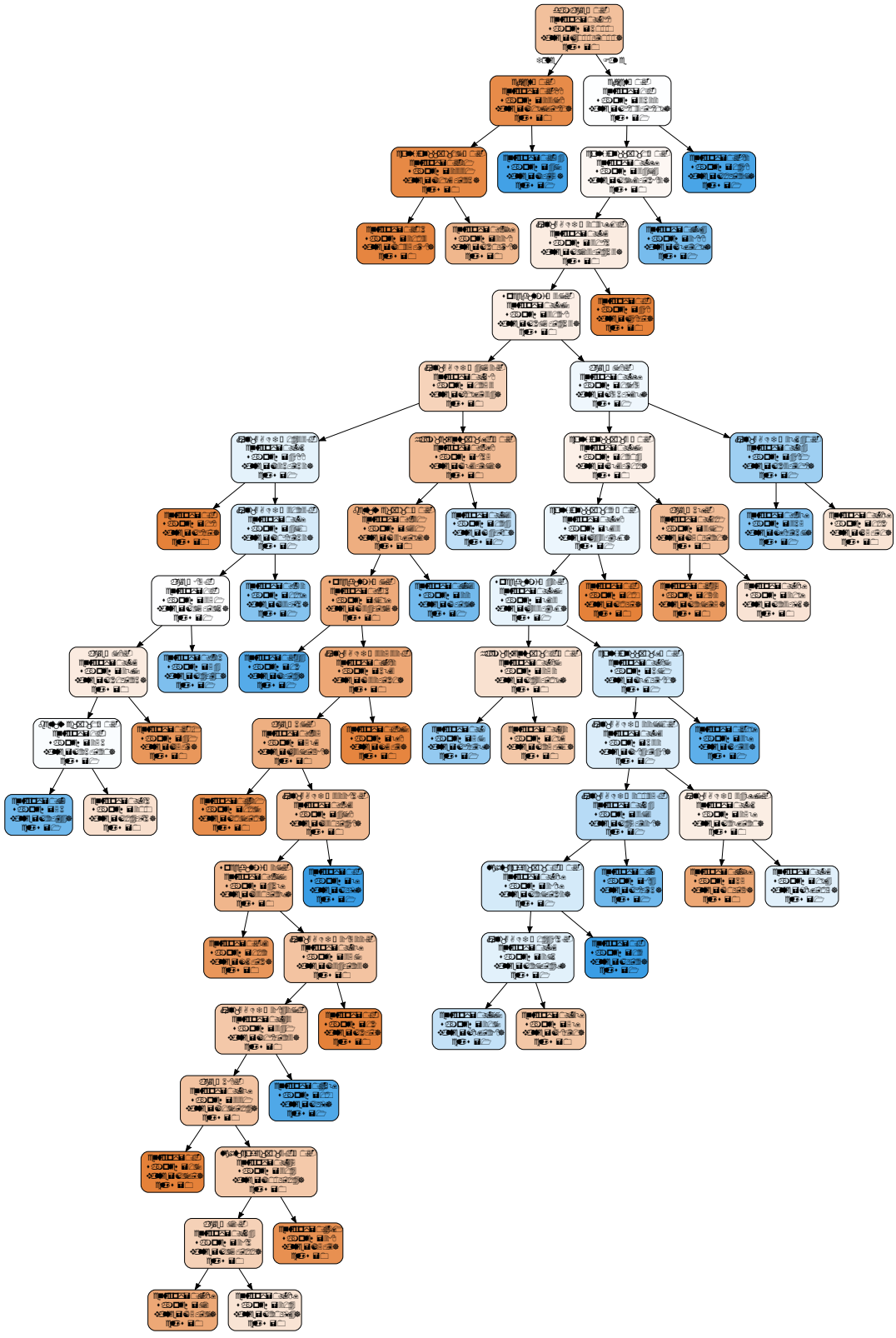


Figure 4-5. The Decision Tree Model of Trauma Triage Using the Integrated Data

The two decision tree models are visualized in Figure 4-4 and Figure 4-5. Each box in the figures represents a leaf node in the decision tree branches. The color of the leaf node denotes the probabilistic classification. Blue denotes EMS in Figure 4-4 and triage to trauma centers in Figure 4-5, while red indicate decisions of non-EMS and non-trauma. A darker color indicates the probabilistic classification provides a higher confidence that the node should belong to a certain class. For instance, the first branch on the left of the tree in Figure 4 has several nodes colored with darker red, indicating that the instances belonging to this branch have a higher chance to be triaged to non-trauma hospitals.

Each node in the decision tree contains the following information:

- The splitting variable. For instance, the first splitting variable selected by the decision tree model is *damage*. If the vehicle is not disabled or damaged in the crash, the instance is more likely to belong to the left tree branch.
- Entropy. It is a measure of the mixture of the instances belonging to different classes. For example, a node where all instances should be triaged to trauma hospitals would have an entropy value of 0. And a node with 50% trauma instances has an entropy value of 1. This measurement is also employed to visualize the color of the nodes.
- Samples, indicating the number of crash instances in this leaf node. In these two trees, the total numbers of samples reflect the oversampling and down-sampling data processes described in Section 2.2.
- Values, a two-dimensional vector depicting the classification of the samples.
- Class, 0 indicates decision of non-EMS transport (Figure 4-4) and triage to non-trauma hospitals (Figure 4-5); 1 indicates decisions of EMS transport (Figure 4-4) and triage to trauma hospitals (Figure 4-5).

The decision tree models enable the interpretation of the prediction rules. For instance, two models all have selected damage as the first splitting variable. If the vehicle shows no damage after the crash, the instance is more likely to be classified to the left side of the decision trees, where more leaf nodes indicate a relatively "safer" status (i.e. class 0, no EMS or trauma is needed). This is a reasonable interpretation. The results also indicate that the AADT volume information tends to be more decisive than the other transportation-sector data as AADT has been selected as the splitting variable at many leaf nodes. These decision trees can be further pruned and re-evaluated for field triage implementations.

4.4 Conclusions

This chapter develops two decision tree models to improve the EMS and trauma triage decision processes respectively. First of all, transportation-related data sources, such as traffic volumes and time-dependent vehicle speeds, have been integrated with EMS and hospital records, offering measurements of exposure to crash risks. The critical transportation information is typically missing from the field triage process but becomes readily available in real-time to support decision-making. Then the integrated data has been employed to construct machine learning models with the new predictors. Predictions of the need for EMS transport and triage to trauma centers have been analyzed using a Maryland dataset with records of over 54,000 elderly patients. Compared to benchmark models without transportation information as predictors, the models trained with integrated data exhibit superior prediction accuracy.

This is one of the first studies that integrates transportation-sector data with health data for safety big-data analytics. The usage of this integrated dataset to support decision making is demonstrated. With two sets of decision tree models, the results indicate that the under triage, a major issue for high-risk senior crash victims, can be significantly addressed. At the same time,

the models also achieve slight improvement in over triage, which indicates that the decision mechanism, once fully developed and implemented, can potentially improve the effectiveness and efficiency of hospital resource allocations.

Chapter 5: A Big-Data Driven Framework for Estimating Vehicle Volume on Mobile Device Location Data

5.1 Problem Statement

As identified in Chapter 3, when planning for "Mobility"-related transportation projects, the important metrics, such as delay, volume-to-capacity ratio, are all directly related to or dependent on vehicle volume estimates. In the meantime, according to the literature, incoming vehicle volume is one of the most important features for pedestrian and bicyclist crashes. Therefore, this chapter uses on one type of transportation big data, MDLD, to estimate vehicle volume for all roads and further estimate pedestrian and bicyclist crashes at intersection based on vehicle volume estimates. More specifically, to estimate the emergency medical services (EMS) and trauma triage decisions at crash scene in order to reduce fatality rate caused by severe injuries.

Vehicle volume serves as a critical metric and the fundamental basis for traffic signal control, transportation project prioritization, road maintenance plans, and more. Traditional methods of quantifying vehicle volume rely on manual counting, video cameras, and loop detectors at a limited number of locations. These efforts require significant labor and cost for large-scale implementations. Researchers and private sector companies have also explored alternative solutions such as probe vehicle data, while still suffering from a low penetration rate. In recent years, along with the technological advancement in mobile sensors and mobile networks, Mobile Device Location Data (MDLD) have been growing dramatically in terms of the spatiotemporal coverage of the population and its mobility. This chapter presents a big-data driven framework that can ingest terabytes of MDLD and estimate vehicle volume at a larger geographical area with larger sample size. The proposed framework first employs a series of cloud-based computational

algorithms to extract multimodal trajectories and trip rosters. A scalable map matching and routing algorithm is then applied to snap and route vehicle trajectories to the roadway network. The observed vehicle counts on each roadway segment are weighted and calibrated against ground truth control totals, i.e., Annual Vehicle-Miles of Travel (AVMT), and Annual Average Daily Traffic (AADT). The proposed framework is implemented on the all-street network in the state of Maryland using MDLD for the entire year of 2019. Results indicate that our proposed framework produces reliable vehicle volume estimates and also demonstrates its transferability and generalization ability.

5.2 The Big-Data Driven Framework for Estimating Vehicle Volume and Pedestrian and Bicyclist Crashes

5.2.1 The Framework

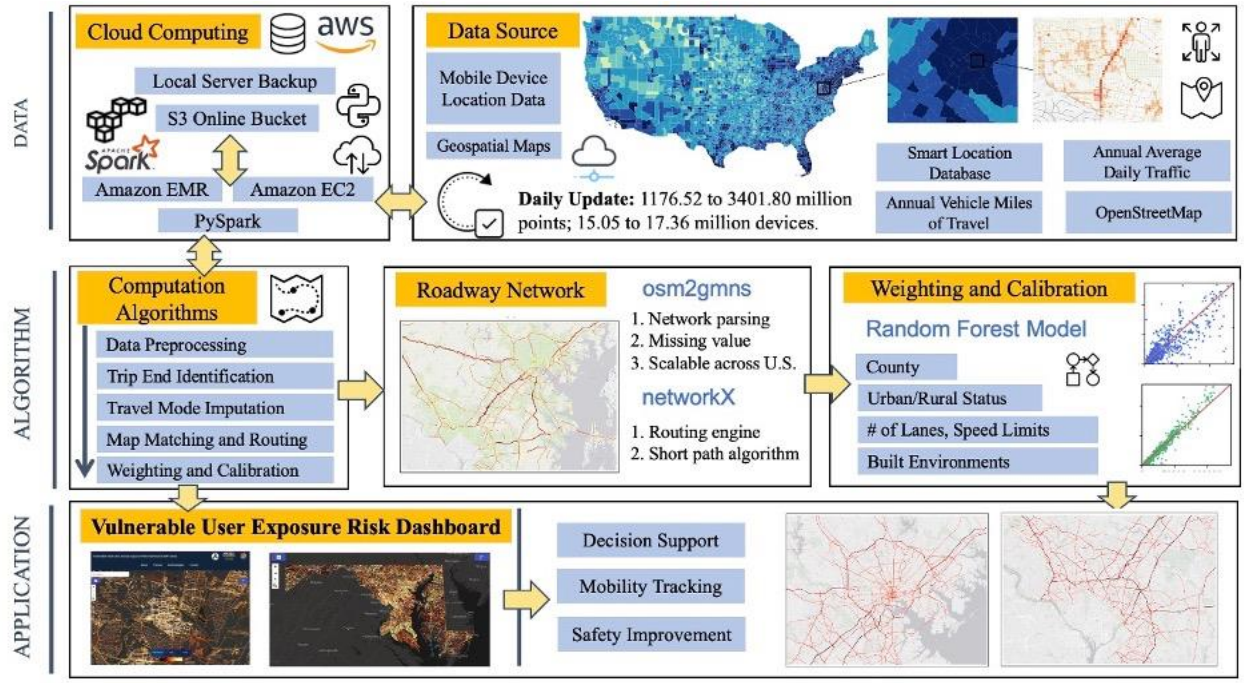


Figure 5-1. The Big-Data Driven Framework for Estimating Vehicle Volume and Pedestrian and Bicyclist Crashes

In this chapter, a big-data driven vehicle volume estimation framework is proposed, which offers the capability of efficiently estimating vehicle volume ingested from terabytes of MDLD. Figure 5-1 shows the proposed framework. The proposed framework is built on Amazon Web Services (AWS). MDLD and all supporting data are stored in Simple Cloud Storage (S3). All algorithms are developed based on Apache Spark, which uses Resilient Distributed Datasets (RDD), and are coded in PySpark using the Elastic MapReduce (EMR) services. In the cloud environment, MDLD are spliced into RDDs given the number of executors [142, 143]. At the same time, all external data sources (i.e., K-D Tree, network, routing engine) are broadcasted into all executors for master and core nodes. The same algorithms are applied to each RDD along with the broadcasted variables, and the results are aggregated and outputted into S3.

5.2.2 Trip Identification and Travel Mode Imputation

A trip is the basic unit of analysis for almost all transportation applications. However, MDLD usually does not contain any trip-related information. Therefore, in this chapter, a trip end identification algorithm is used to extract trip-level information from the MDLD, including trip start location, trip end location, departure time, and arrival time. Then, a travel mode imputation model is further applied to infer four travel modes–namely, airline, vehicle, rail, and nonmotorized modes, based on heuristic rules and a random forest model.

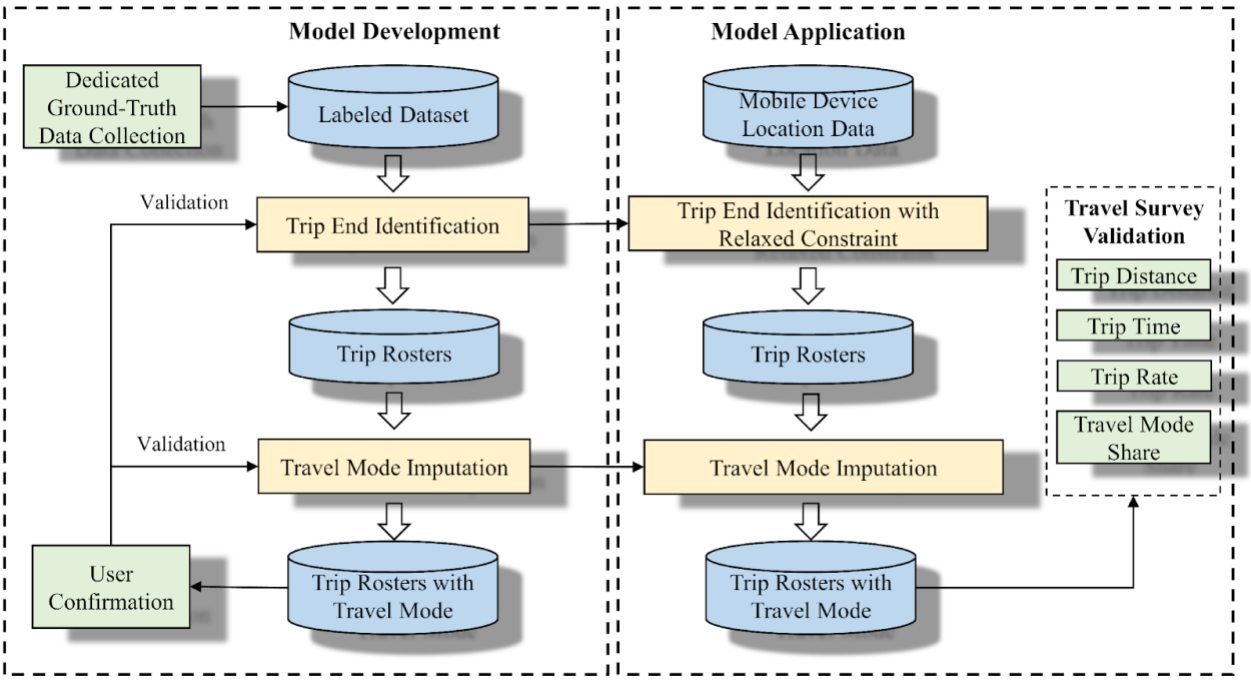


Figure 5-2. The Data-Driven Travel Mode Share Estimation Framework

Figure 5-2 shows the data-driven framework for the trip identification and travel mode imputation algorithms. On the left is the Model Development pillar, wherein a dedicated ground-truth data collection of labeled and mode-specific trips and trajectories is conducted in order to train the trip end identification algorithm and travel mode imputation model. These trained models are then applied to the Model Application pillar on the right. The Model Application generates trip rosters with imputed travel modes for the unlabeled MDLD datasets in the application contexts. Finally, a validation process compares the aggregated mode share, as well as other statistics, with travel surveys before the data products are deemed useful and applicable for any transportation planning applications. Detailed descriptions of the trip end identification algorithm and the travel mode imputation model can be found in references [2, 144].

5.2.3 Scalable Map Matching and Routing via Cloud Computing

To ensure flexibility and scalability of our map matching and routing method across the entire U.S., the drivable network is extracted from OpenStreetMap (OSM) using the latest open-source Python package osm2gmns. This package can parse roadway network data from OSM and output networks to the General Modeling Network Specification (GMNS) format. It provides customized and practical functions to facilitate traffic modeling. Functions include complex intersection consolidation, movement generation, traffic zone creation, short link combination, and network visualization. More details about osm2gmns can be found at https://osm2gmns.readthedocs.io/en/latest/

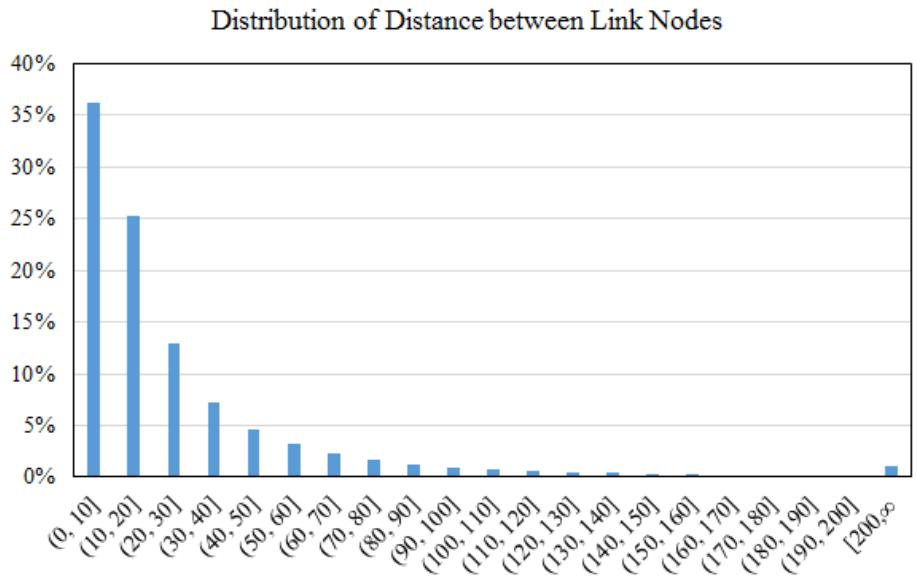


Figure 5-3. Distribution of Distance between Link Nodes in the OSM Network

To match each location sighting to our OSM network, the OSM network is firstly parsed and converted into the routable formats, where roadway segments are represented by links and nodes. With the network topology, networkX package is used to build a shortest path-based routing engine. The latitude and longitude of the start node and end node for each link are then transformed

to the plane coordinate (in meters), and then calculate link direction (degree) using the arctan value between the two nodes. The travel direction between consecutive sightings is also calculated. Similar to the method for link direction calculation, the coordinates of each sighting are converted to plane coordinates, and then the degree is calculated using the arctan value between consecutive sightings. A spatial index structure, K-Dimensional Tree (K-D Tree), is built using the link geometric nodes (i.e., link nodes). Then, for each sighting, all links within 100 meters are searched. The 100-meter threshold is selected to balance the algorithm efficacy and computing speed. If the value is increased, more candidate links will be considered but this will require more computing resources. If the value is decreased, it might not be able to find a candidate link when the sighting is sparse. To validate, the distance between consecutive link nodes is calculated using the Maryland OSM network as an example. Results indicate that more than 95% of the link nodes are within 100 meters of their neighbors, as shown in Figure 5-3. Therefore, using the 100-meter value as the radius for searching candidate nodes is reasonable.

As the next step, each sighting's travel direction is compared to all candidate links' travel directions. The closest link with an absolute travel direction difference smaller than 30 degrees will be selected as a valid matched link for the sighting. This 30-degree threshold is selected mainly to avoid the sighting being matched to the link in the opposite direction. In common cases, the degree difference between the travel direction and the link direction should be approximately 0. Here, a 30-degree threshold is used to consider the uncertainty of location accuracy in MDLD. After the matched link for each sighting is found, given the observed link sequence, the routing engine can fill the gap between consecutively observed links and retrieve the complete route. Another layer of reasonable checks is conducted at the routing stage. For each pair of consecutive

sightings that are snapped to links, the routed distance is calculated by summing the link length of all the links traveled between the two sightings. Two reasonableness checks are carried out:

- If the routed distance is greater than the cumulative distance between the two sightings snapped to links by 2,000 meters or more, I consider the route invalid.
- The travel time on these links will be calculated based on the timestamp difference between the two sightings. With the routed distance and travel time, the average travel speed on these links can be calculated. If the speed exceeds 50 m/s (i.e., 112 mph or 180 km/h), one of the two sightings is assumed to match the wrong link.

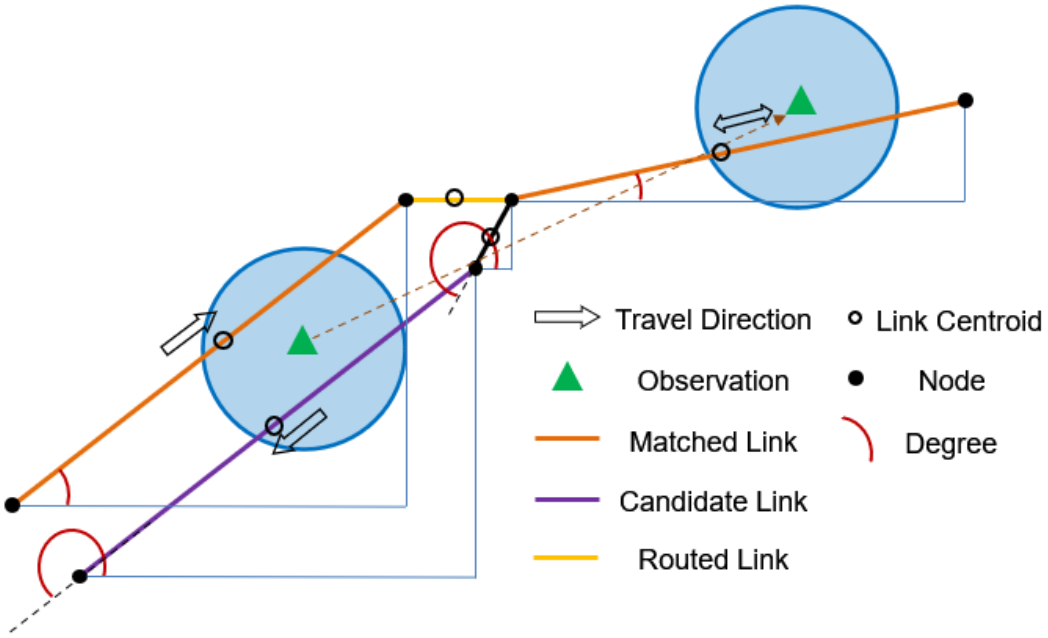


Figure 5-4. Example of Map Matching and Routing.

If either of these two violations is observed, a trial-and-error process is applied by removing the latter sighting and performing the routing using the next sighting snapped to the network until it does not violate the 2,000-meter threshold or the 50 m/s threshold [145]. A simple example of the map matching and routing method is illustrated in Figure 5-4. This algorithm can also be applied to pedestrian and bicyclist trips using non-motorized OSM network.

5.2.4 Weighting

After map matching and routing, the routes for all vehicle trips are collected and aggregated by links to obtain the observed vehicle volume for each link. Afterward, a link-based weighting method is developed to match the AVMT in the region. Each link is classified by county, urban/rural status, and functional classes, and calculate the link weight using the formula below:

$$w_{C,u,f} = \frac{AVMT_{C,u,f}}{\sum_{N_C} O_{C,u,f,i}} \quad (1)$$

where $w_{C,u,f}$ represents the weight for links in county *C*, with urban/rural status of *u*, and with functional class *f*; $AVMT_{c,u,f}$ represent the AVMT; and $O_{C,u,f,i}$ represents the observed vehicle volume on link *i*; N_C represents the total number of links in county *C*. For instance, if the study area has 20 counties, 2 urban/rural statuses, and 6 functional classes, then a total of 240 link-based weights are generated. Subsequently, the weighted vehicle volume for each link can be calculated as:

$$V_{c,u,f,i} = w_{C,u,f} \times O_{c,u,f,i} \quad (2)$$

where $V_{c,u,f,i}$ represents the weighted vehicle volume on link *i*.

5.2.4 Volume Calibration

The weighted vehicle volume is further calibrated to match the ground truth AADT collected from loop detectors at a limited number of locations. In this study, the random forest regression is used to calibrate the weighted vehicle volume against the AADT to obtain the final vehicle volume. During the calibration process, a 10-fold cross-validation (CV) process is used to fine-tune the

random forest regression hyperparameters with 70% training data. The fine-tuned models are then applied to the 30% testing data.

5.3 Vehicle Volume Estimation Case Study: the State of Maryland

5.3.1 Data

5.3.1.1 Mobile Device Location Data and the Study Area

This case study used MDLD data obtained from Maryland Transportation Institute (MTI). MTI integrated and cleaned the raw MDLD from multiple data vendors and built a national MDLD data panel that consists of more than 270,000,000 Monthly Active Users (MAU) and represents movements across the U.S.

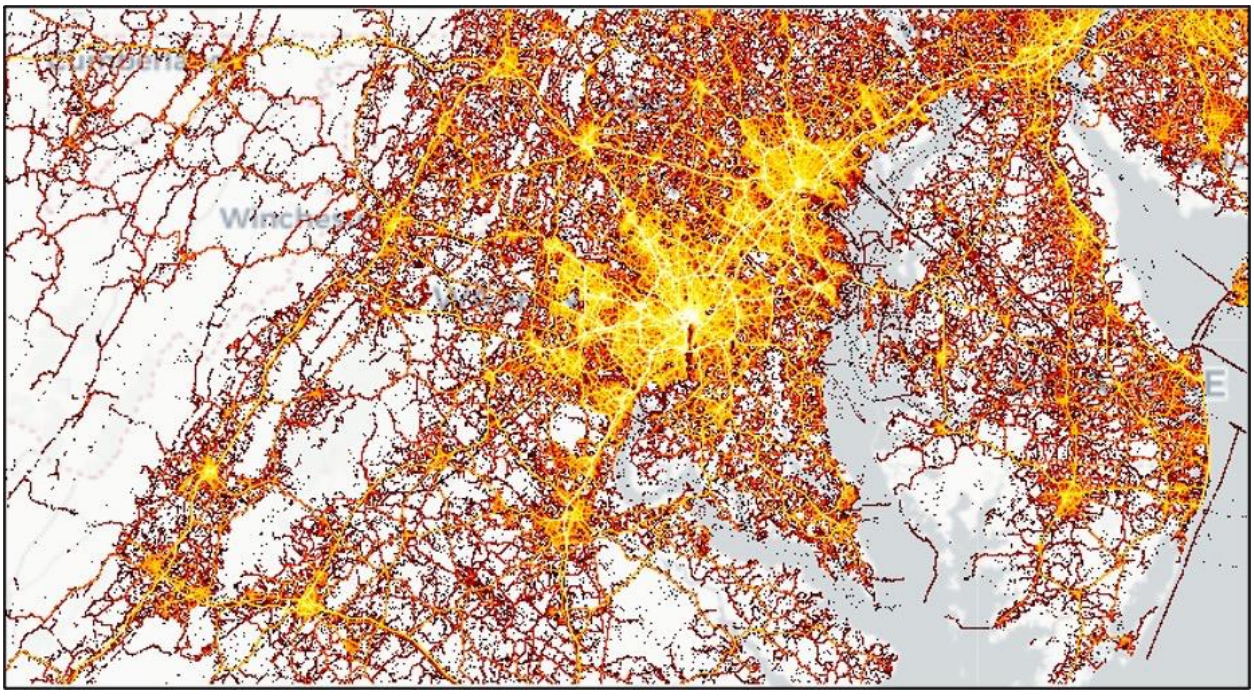


Figure 5-5. Mobile Device Location Data around the State of Maryland.

Figure 5-5 shows the density of location sightings covering locations within and outside of the boundaries of the state of Maryland. In this study, all MAU observed in the state of Maryland

for the entire year of 2019 are used. The MDLD is processed on a daily basis and the results are aggregated to produce an annual total result.

5.3.1.2 OpenStreetMap Network

Using the osm2gmns package, a total of 634,516 drivable roadway segments are extracted within the state of Maryland. Information about the number of lanes and speed limits was recorded for only 111,835 roadway segments (17.6%) and 84,728 roadway segments (13.4%), respectively.

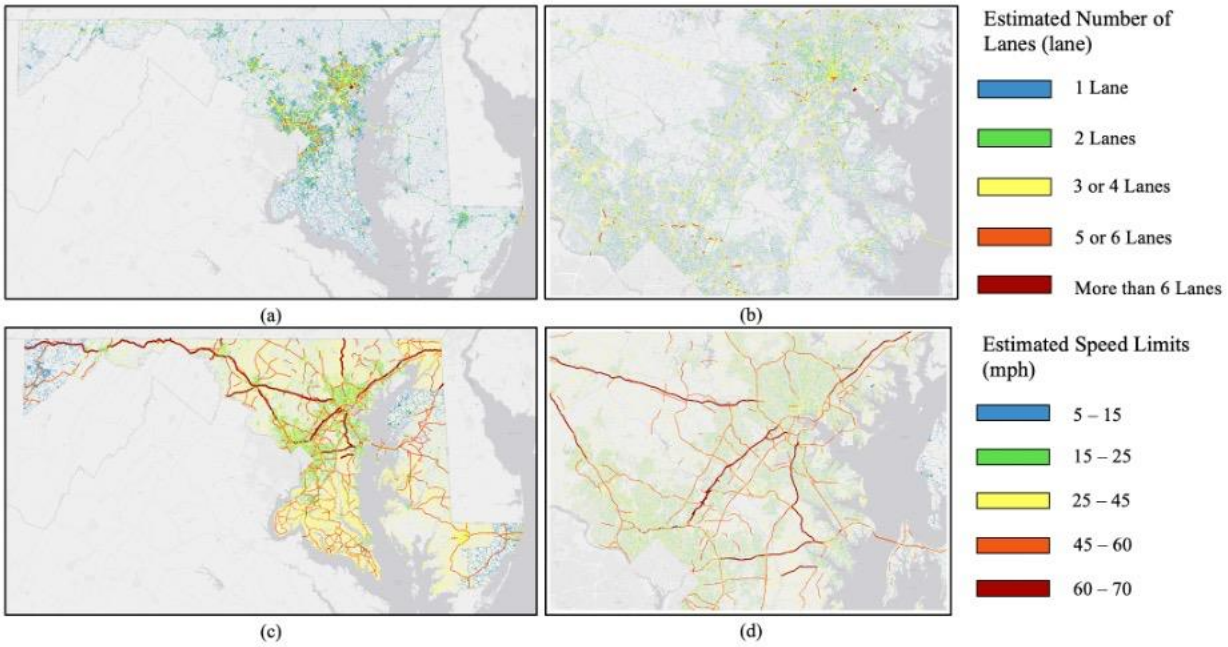


Figure 5-6. Number of Lanes and Speed Limits in OSM

As shown on the left-hand side in Figure 5-6, the missing values for the number of lanes and speed limits were estimated based on the corresponding values on nearby roadways in the same county, and with the same urban/rural status, and road functional classes. These two variables are further used as features in the vehicle volume calibration model.

5.3.1.3 Annual Vehicle Miles of Travel Data

Vehicle miles traveled data from the Maryland Department of Transportation State Highway Administration (MDOT SHA) is used as a control total number to weight observed vehicle volume. Every year, MDOT SHA publishes an annual vehicle miles of travel (AVMT) report by county and functional classification for the state, county, and municipal highway systems. This AVMT report features the current FHWA Functional Classification Codes (1-7) and provides additional classifications (i.e., Urban, Rural, Principal Arterial and Other Freeways and Expressways, and Minor Collector). As discussed in the methodology section, the weights are generated based on county, urban/rural status and functional classes. Here, 23 Maryland counties plus Baltimore City, urban or rural, and two function classes (highway and non-highway) are considered. The OSM link types are mapped to FHWA Functional Classification Codes and generated the highway and non-highway classes. More specifically, "motorway", "trunk" and "ramp" are classified as highway (i.e., 1, 2 in FHWA class), and the other types are classified as non-highway (i.e., 3,4,5,6,7 in FHWA class). More details about the AVMT data can be found here: https://www.roads.maryland.gov/mdotsha/Pages/index.aspx?PageId=302

5.3.1.4 Annual Average Daily Traffic Data

AADT also from MDOT SHA to is used calibrate weighted vehicle volume against the ground truth at a limited number of locations. The AADT data consists of linear and point geometric features which represent the geographic locations and segments of roadway throughout the state of Maryland that include traffic volume metrics such as AADT. More details about the AADT can be found here:

https://data.imap.maryland.gov/maps/77010abe7558425997b4fcdab02e2b64/about

5.3.1.5 Smart Location Database and Features for Volume Calibration

The Smart Location Database (SLD) is a nationwide geographic data resource for measuring location efficiency [147, 148]. The SLD is produced by the U.S. Environmental Protection Agency (EPA)'s Smart Growth Program. It provides more than 90 variables on land use and built environment characteristics such as population and employment densities, land use diversity, urban design attributes, destination accessibility, transit accessibility, and socioeconomic/sociodemographic characteristics at the census block group level. Most attributes are available for every census block group in the United States. In this study, SLD variables are used as features in the random forest regression to calibrate weighted vehicle volume to account for the effects of the built environment. The SLD variables used in this study include "TotEMP", "Pct_AO0", "D1A", "D1C", "D3AAO", "D3B", and "D5AR":

- TotEMP = total employment;
- Pct_AO0 = percent of zero-car households;
- D1A = gross residential density (housing units per acre) on unprotected land;
- D1C = gross employment density (jobs per acre) on unprotected land;
- D3AAO = network density in terms of facility miles of auto-oriented links per square miles;
- D3B = street intersection density (weighted, auto-oriented intersections eliminated);
- D5AR = jobs within 45 minutes auto travel time, time decay (network travel time) weighted

Urban/rural status, county code, link type, number of lanes, and speed limits are also included as features in the calibration process.

5.3.2 Vehicle Volume Estimation Results

5.3.2.1. Overall Comparison

As mentioned above, a series of variables are used as features in the vehicle volume calibration process, including “TotEMP”, “Pct_AO0”, “D1A”, “D1C”, “D3AAO”, “D3B”, and “D5AR” from SLD, urban/rural status, county code, link type, number of lanes, and speed limits.

Using the AADT data in 2019 from MDOT SHA, a random forest regression model is built to map weighted vehicle volume into observed AADT at each segment. A total of 13,359 AADT records were obtained for the model development, including 806 interstate highway and highways, 1,575 primary roads, 2,867 secondary roads, 3,352 tertiary roads, 2,055 local roads, and 2,704 ramps.

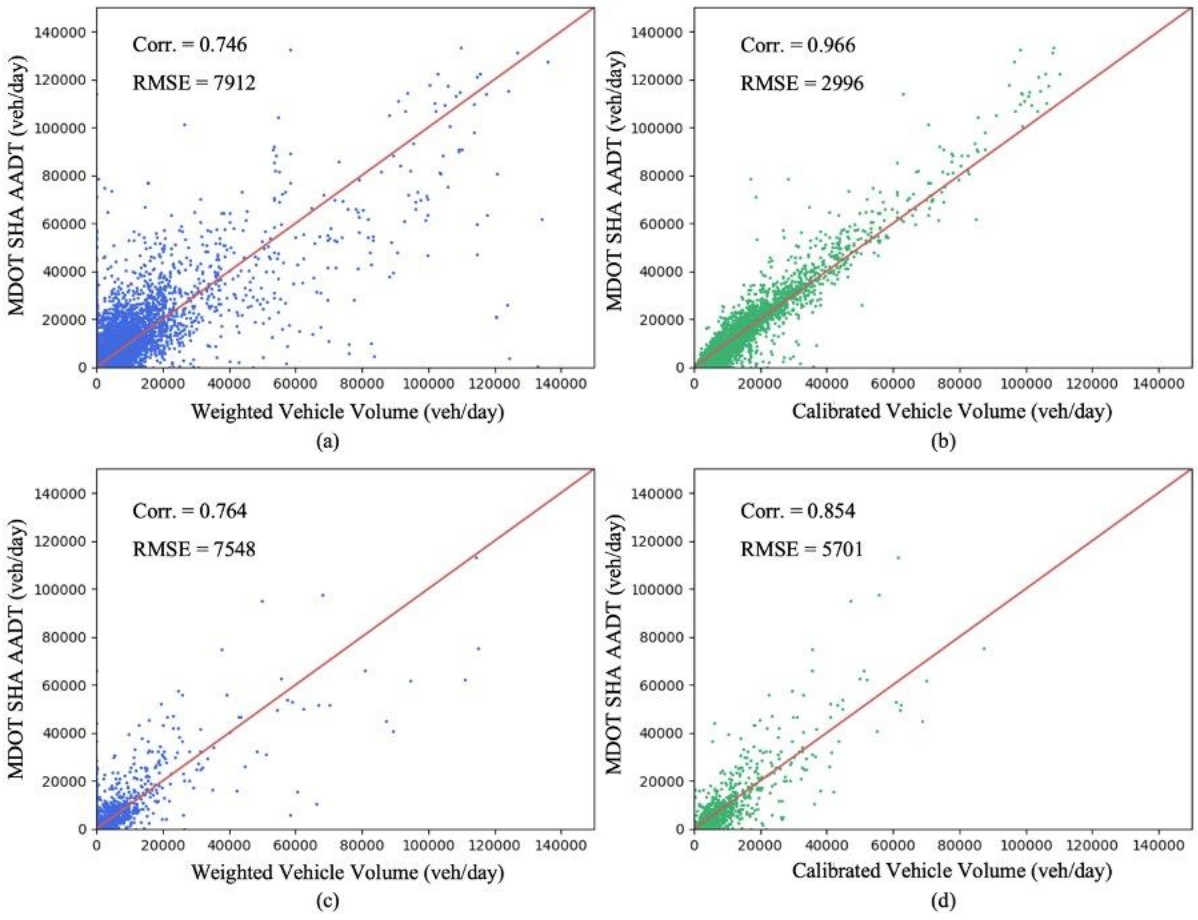


Figure 5-7. (a) Weighted Vehicle Volume in Training Set; (b) Calibrated Vehicle Volume in Training Set; (c) Weighted Vehicle Volume in Testing Set; (d) Calibrated Vehicle Volume in Testing Set.

Figure 5-7 shows the weighting and calibration results for both training and testing sets. The blue dots represent weighted volume comparisons and the green dots represent calibrated vehicle volume comparisons with MDOT SHA AADT. Figure 5-7 (a) and (b) compares the weighted vehicle volume and calibrated vehicle volume with the MDOT SHA AADT in the training set respectively; Figure 5-7 (c) and (d) compares the weighted vehicle volume and calibrated vehicle volume with the MDOT SHA AADT in the testing set respectively. It can be seen from Figure 5-7 (a), for the training set, the Pearson correlation value and the Root Mean Square Error (RMSE) between the weighted vehicle volume and the ground truth AADT are 0.746 and 7,912, respectively. These values are improved to 0.966 and 2,996 after calibration, as shown in Figure 5-7 (b). Similarly, for the testing set, the Pearson correlation and RMSE are improved from 0.764 and 7,548, to 0.854 and 5,701, respectively, after calibration.

5.3.2.2 Vehicle Volume Validation by Link Types and Urban/Rural Status

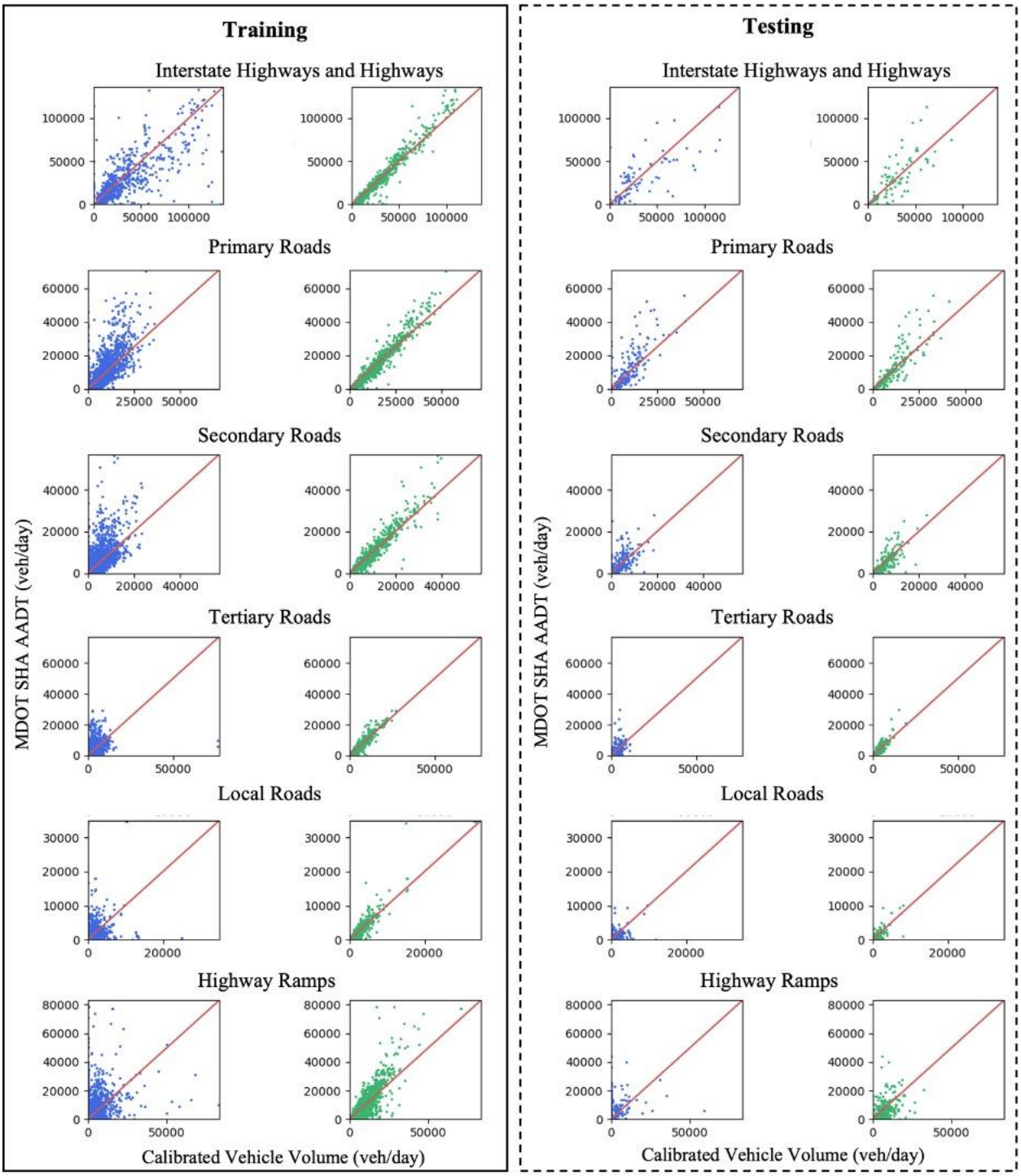


Figure 5-8. Volume Calibration Results Comparison by Link Type.

Table 5-1. Volume Calibration Results Comparison by Link Type

Link Type	Training Set				Testing Set			
	Corr.		RMSE		Corr.		RMSE	
	Before	After	Before	After	Before	After	Before	After
All	0.746	0.966	7912	2996	0.764	0.854	7548	5701
Interstate Highways and Highways	0.752	0.975	20081	6559	0.712	0.775	19633	15246
Primary Roads	0.699	0.971	7909	2695	0.721	0.846	8665	6509
Secondary Roads	0.627	0.960	4899	1776	0.617	0.813	3667	2667
Tertiary Roads	0.414	0.959	3486	994	0.511	0.869	3090	1877
Local Roads	0.374	0.944	2474	853	0.426	0.742	1701	1083
Highway Ramps	0.242	0.866	10426	4722	0.182	0.402	9119	6846

Figure 5-8 and Table 5-1 show the calibrated vehicle volume by link types for both the training and testing sets. For all link types, a good correlation (i.e., over 0.80) can be observed between the calibrated vehicle volume and the ground truth AADT, except for Local Roads and Highway Ramps in the testing set. The results indicate that our proposed framework can accurately estimate vehicle volume on higher-level roadways (i.e., Interstate Highways and Highways, Primary Roads, Secondary Roads), while concurrently maintaining high correlations for lower-level roadways (i.e., Tertiary Roads, Local Roads, Highway Ramps). The relatively weaker performance for the case of lower-level roadways can be attributed to limitations in technology. The MDLD only capture part of the daily trips of a device within the area with mobile network connections and higher-level roadways usually have a better coverage compared to lower-level ones. This variability might also result in capturing more travelers on highways and major arterials. In addition, the LBS data sample is more likely to include the active travelers that make more trips and/or longer-duration trips, such as long-distance travel for leisure or business purposes or long-distance commute which usually happen on interstate highways.

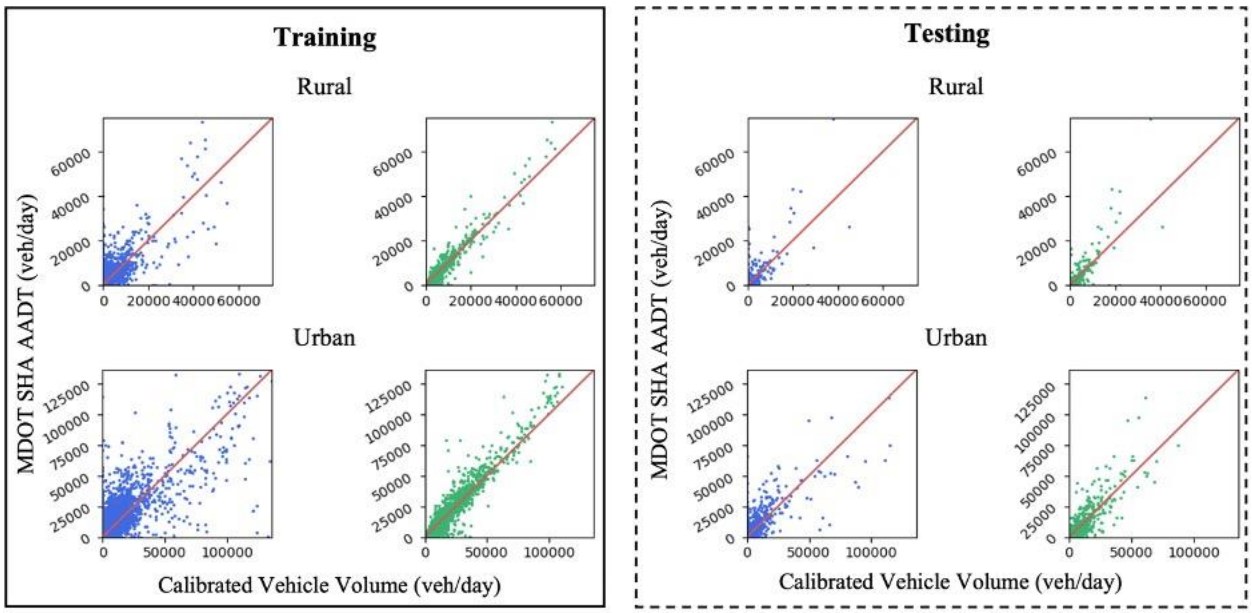


Figure 5-9. Volume Calibration Results Comparison by Urban/Rural Status.

Table 5-2. Volume Calibration Results by Urban/Rural Status.

Link Type	Training Set				Testing Set			
	Corr.		RMSE		Corr.		RMSE	
	Before	After	Before	After	Before	After	Before	After
All	0.746	0.966	7912	2996	0.764	0.854	7548	5701
Rural	0.769	0.967	3583	1442	0.727	0.826	4810	4075
Urban	0.738	0.964	8913	3363	0.764	0.853	8311	6179

Figure 5-9 and Table 5-2 show the calibration of vehicle volume by urban/rural status for both the training and testing sets. In summary, for both urban and rural roads, a good correlation (i.e., over 0.80) can be observed between the calibrated vehicle volume and the ground truth AADT, whereas a higher correlation can be observed for urban roads. The relatively weaker performance in rural roadways can also be attributed to the technology limitation mentioned above.

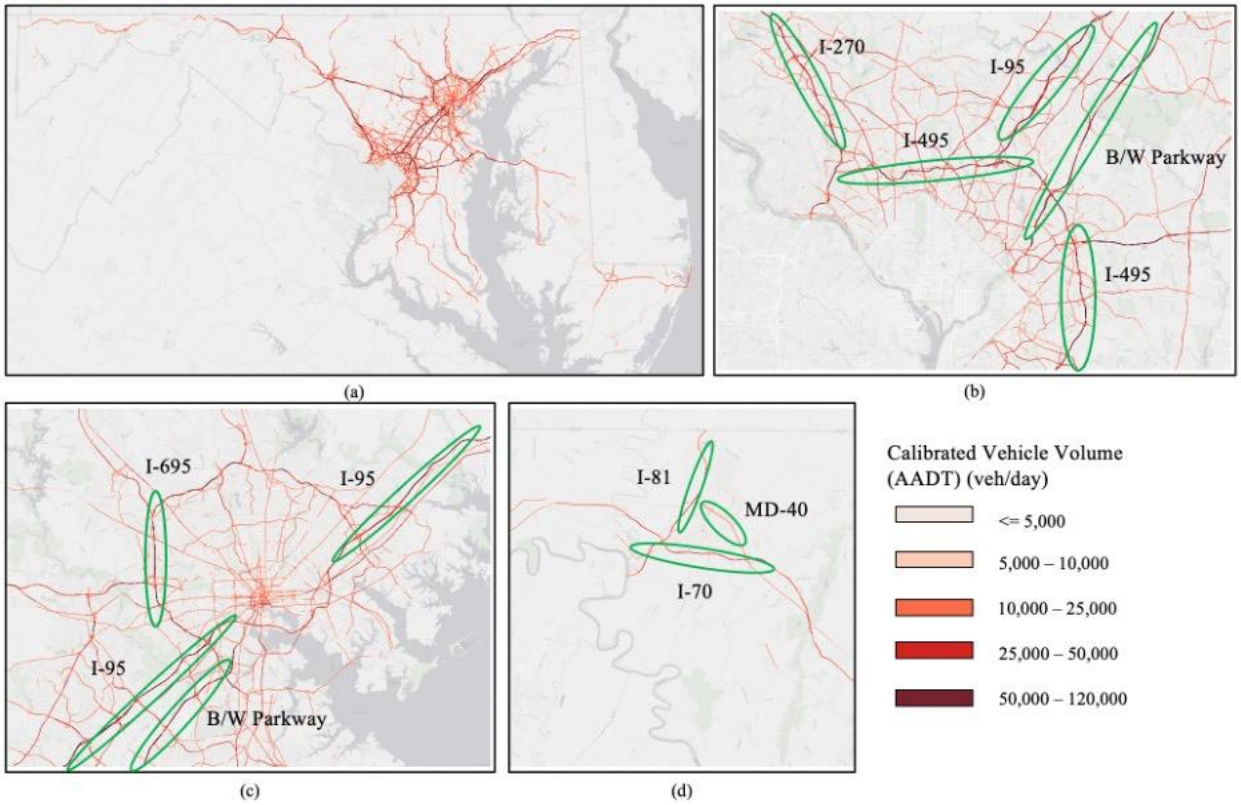


Figure 5-10. Visualization of Calibrated Vehicle Volume. (a) the State of Maryland; (b) Washington D.C.; (c) Baltimore City; (d) Hagerstown, MD.

Figure 5-10 (a) visualizes the calibrated vehicle volume averaged from the entire year of 2019 (represented as AADT) on the all-street network in the state of Maryland. It can be seen that the interstate highway and the highway skeletons can be clearly identified from the map. Major arterials also stand out from the map. Figure 5-10 (b) zooms into the Washington D.C. area, where routes I-495, I-270, I-95 and the Baltimore/Washington Parkway are clearly seen. Figure 5-10 (c) zooms into the Baltimore area, where I-395, I-695, I-795, I-95, and I-70 are all captured. Figure 5-10 (d) zooms into Hagerstown, MD, which is a city in Washington County, MD near the border of Pennsylvania. The I-70, I-81, and MD-40 are all captured, demonstrating the ability of the proposed framework to produce reliable results in rural areas.

5.4 Conclusion

This chapter presents a big-data driven framework that is able to ingest terabytes of MDLD and estimate vehicle volume based on MDLD. The proposed framework first employs a series of cloud-based computational algorithms to extract vehicle trajectories. A map-matching and routing algorithm is then applied to snap and route vehicle trajectories to the road network. The observed vehicle counts on each road segment are weighted and calibrated against the control total, i.e., annual vehicle miles traveled (VMT), and data collected from real-world loop detectors. The proposed framework is implemented and validated on the all-street network in the state of Maryland using MDLD data from 2019. After weighting and calibration processes, high correlation and low RMSE values are observed between our vehicle volume estimates and the ground truth data.

Chapter 6: Modeling Pedestrians and Bicyclist Crashes with Transportation Big Data

6.1 Pedestrian and Bicyclist Crashes Estimation

After the volume calibration, one use case is to integrate the calibrated vehicle volume into crash estimation models. This chapter focuses on pedestrian and bicyclist crashes specifically. Different statistical models were developed and estimated to examine the role of various key contributing factors including safety risk exposure factors in pedestrian and bicyclist crashes that occurred at Maryland intersections, including Poisson, NB, ZIP, and ZINB. Using Washington et al. [146] as a reference, a brief review of these models is described below:

6.1.1 Poisson and NB Models

Poisson regression models are estimated by specifying the Poisson parameter λ_i (i.e., the expected number of events per period) as a function of independent variables [146]. The probability of y_i crashes occurring at an intersection *i* during a particular year is given by:

$$P(y_i) = \frac{EXP\ (-\lambda_i)\ \lambda_i^{y_i}}{y_i!} \qquad (3)$$

where y_i is a nonnegative integer representing the number of crashes at intersection *i* during a particular year; $P(y_i)$ is the probability of y_i crashes occurring at intersection *i* during a particular year; λ_i is the Poisson parameter for intersection *i*, which is equal to the expected number of crashes per a particular year at intersection *i*, $E[y_i]$.

The most common relationship between independent variables and the Poisson parameter is the log-linear model formulated as (see *34*):

$$\lambda_i = EXP\ (\beta X_i) \qquad (4)$$

where X_i is a vector of independent variables; and β is a vector of model parameters (i.e., estimable coefficients).

The main requirement for applicability of the Poisson regression model is that the mean of the count variable equals its variance. If the variance of the count data is larger than the mean, the data are considered to be overdispersed—a condition that can lead to biased results should a Poisson regression model is applied to such count data. In the case of overdispersion, the NB model is a more suitable modeling method.

The NB model is derived by reformulating equation (1) as:

$$\lambda_i = EXP\ (\beta X_i + \varepsilon_i) \qquad (5)$$

where $EXP\ (\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α^2; and α is the dispersion parameter.

The NB model formulation allows the variance to differ from the mean as:

$$VAR[y_i] = E[y_i] + \alpha E[y_i]^2 \qquad (6)$$

The selection between the Poisson model and the NB model is determined based on the value of the estimated dispersion parameter α. If parameter α equals zero, the negative binomial model reduces to the Poisson model. If parameter α is greater than zero, the negative binomial model is a more appropriate model to be used on the data.

With regards to crash data, a common issue that arises is preponderance of zeros, which is a result of observing zero crashes during the observation period. In general, observing zero events during the observation period can be due to two different processes: *i)* a normal count-process state—i.e., an event not occurring during the observation period; and *ii)* a zero-count state—i.e., an inability ever to experience an event [146]. Considering the case of pedestrian and bicyclist crashes at an intersection as an example, a normal count-process state can be the case when no

pedestrian/bicyclist crashes occurred during the observation period. A zero-count state in that case would be when no pedestrian/bicyclist crashes could occur at the intersection (i.e., the likelihood of a pedestrian/bicyclist crash occurring is extremely rare) owing to reasons such as the intersection not having any pedestrian or bicyclist volumes (e.g., rural intersections). The two different processes generating observations of zero in crash data lead to excess zeros in the data. In addition, data obtained from two-state processes (i.e., the normal-count and the zero-count states) often suffer from overdispersion due to the number of zeros being inflated by the zero-count state *(34)*. Models that account for the two-state processes are referred to as zero-inflated models. These include the ZIP and ZINB regression models.

6.1.1 ZIP and ZINB Models

Applying the theories discussed in Washington et al. [146] to pedestrian/bicyclist crashes at an intersection *i*, the ZIP model assumes that the crashes, $Y = (y_1, y_2, \ldots . y_n)$ are independent and the probability density function for the model can be formulated as:

$$P(Y = y_i) = \begin{cases} P_i + (1 - P_i)\ EXP\ (-\lambda_i) & y_i = 0 \\ \frac{(1-P_i)\ EXP\ (-\lambda_i)\ \lambda_i^y}{y!} & y_i = y \end{cases} \quad (7)$$

The ZINB model has the same assumption, and its probability density function can be formulated as:

$$P(Y = y_i) = \begin{cases} P_i + (1 - P_i) \left[\frac{\frac{1}{\alpha}}{\left(\frac{1}{\alpha}\right) + \lambda_i}\right]^{1/\alpha} & y_i = 0 \\ (1 - P_i) \left[\frac{\Gamma\left(\left(\frac{1}{\alpha}\right)+y\right) {u_i}^{1/\alpha}\ (1-u_i)^y}{\Gamma\left(\frac{1}{\alpha}\right) y!}\right], y = 1, 2, 3, \ldots & y_i = y \end{cases} \quad (8)$$

where y is the number of pedestrian/bicyclist crashes occurring at an intersection i during a particular year; α is the dispersion parameter; and

$$u_i = \frac{\left(\frac{1}{\alpha}\right)}{\left(\frac{1}{\alpha}\right) + \lambda_i} \qquad (9)$$

Parameters in ZIP and ZINB models are estimated using the maximum likelihood methods.

The results of the Poisson model, the negative binomial model, the zero-inflated Poisson model, and the zero-inflated negative binomial model as applied to data on crashes involving pedestrian/bicyclist at Maryland intersections were compared in this study to identify the most suitable model that best fits the data.

6.2 Pedestrian and Bicyclist Crashes Estimation Case Study: the State of Maryland

6.2.1 Data

6.2.1.1 Calibrated Vehicle Volume and Pedestrian and Bicyclist Volume

The calibrated vehicle volume estimated in Section 5.3 are used as one critical variable to support the pedestrian and bicyclist crash estimations.

In addition to the vehicle volume, the same algorithm has also been applied to the imputed pedestrian and bicyclist trips in the MDLD and map-matched to non-motorized network in OSM. Similar to the vehicle volume calibration, the pedestrian and bicyclist volume calibration aims at minimizing the difference between our estimates and the ground truth pedestrian and bicyclists counts. The RF model is also applied for the calibration. The pedestrian and bicyclist volume were calibrated and validated at intersection-level using the same variables against the real probe data of pedestrians and bikes from the MDOT SHA at 845 intersections in Maryland. For details, please refer to [149].

6.2.1.2 American Community Survey

The American Community Survey (ACS) is an annual survey program conducted by the United States Census Bureau. Data collected through this survey provide information about the population (e.g., socioeconomic/sociodemographic characteristics, means of commuting to work) as well as housing (e.g., financial and physical characteristics for housing units) at many geographical scales.

The 2019 five-year ACS estimates at the census block group were used in this study as the source of socioeconomic and sociodemographic data in development of pedestrian/bicyclist crash frequency models.

6.2.1.3 Level of Traffic Stress (LTS)

Figure 6-1. LTS Examples (Source: http://www.northeastern.edu/peter.furth/research/level-of-traffic-stress)

Developed in 2012 [150], Level of Traffic Stress (LTS) is a scale that rates a road segment based on the traffic stress it imposes on bicyclists. The LTS is defined based on traffic characteristics such as number of lanes and speed limit, and ranges from 1 (for the lowest level of traffic stress)

to 4 (for highest level of traffic stress). As shown in Figure 6-1, the four levels of LTS are described in more detail by Furth [151]:

- LTS 1: Strong separation from all except low speed, low volume traffic. Simple crossings. Suitable for children;
- LTS 2: Except in low speed/low volume traffic situations, cyclists have their own place to ride that keeps them from having to interact with traffic except at formal crossings. Physical separation from higher speed and multilane traffic. Crossings that are easy for an adult to negotiate. Limits traffic stress to what the mainstream adult population can tolerate, those who are "interested but concerned" in the classification popularized by Portland, Oregon's bike program. Criteria for this level correspond to design criteria for Dutch bicycle route facilities;
- LTS 3: Involves interaction with moderate speed or multilane traffic, or close proximity to higher speed traffic. A level of traffic stress acceptable to those classified as "enthused and confident.";
- LTS 4: Involves being forced to mix with moderate speed traffic or close proximity to high-speed traffic. A level of stress acceptable only to the "strong and fearless."

Table 6-1. Level of Traffic Stress Correspondence Table

Level	**OSM Criteria**
LTS 1	Trail
	T1
	T2
	T3
	T13
	T24
	S8
	S7

	S1 or auto;bike auto;walk;bike etc, lane <= 2, mph < 25
LTS 2	M1, lane < 2, mph < 35
	M2a, lane < 2, mph < 35
	M2b, lane < 2, mph < 35
	M2ab, lane < 2, mph < 35
	M3b, lane < 2, mph < 35
	M4, lane < 2, mph < 35
	L1a, lane < 4, mph < 35
	S1 or auto;bike auto;walk;bike etc, 2< lane <= 3, mph < 25
LTS 3	M1, lane >= 2, 35<= mph < 40
	M2a, lane >= 2, 35<= mph < 40
	M2b, lane >= 2, 35<= mph < 40
	M2ab, lane >= 2, 35<= mph < 40
	M3b, lane >= 2, 35<= mph < 40
	M4, lane >= 2, 35<= mph < 40
	L1a, lane >= 2, 35<= mph < 40
	S1 or auto;bike auto;walk;bike etc, 4<= lane <= 5, mph < 25
	S1 or auto;bike auto;walk;bike etc, 2<= lane <= 3, mph < 35
LTS 4	S1 or auto;bike auto;walk;bike etc, lane >=6, mph < 25
	S1 or auto;bike auto;walk;bike etc, lane >=4, mph < 35
	S1 or auto;bike auto;walk;bike etc, lane >=2, mph >= 35
	mph >= 40

As a surrogate measure of pedestrian and bicyclist safety, the LTS was included in this analysis to quantify the traffic stress imposed on vulnerable road users at Maryland intersections. Based on the LTS definition and OSM network attributes, this study develops a mapping of OSM network features into LTS (see Table 6-1 above). Data used in OSM include Bike Lane Existence, Speed Limits, Number of Lanes, and Authorized Travel Modes.

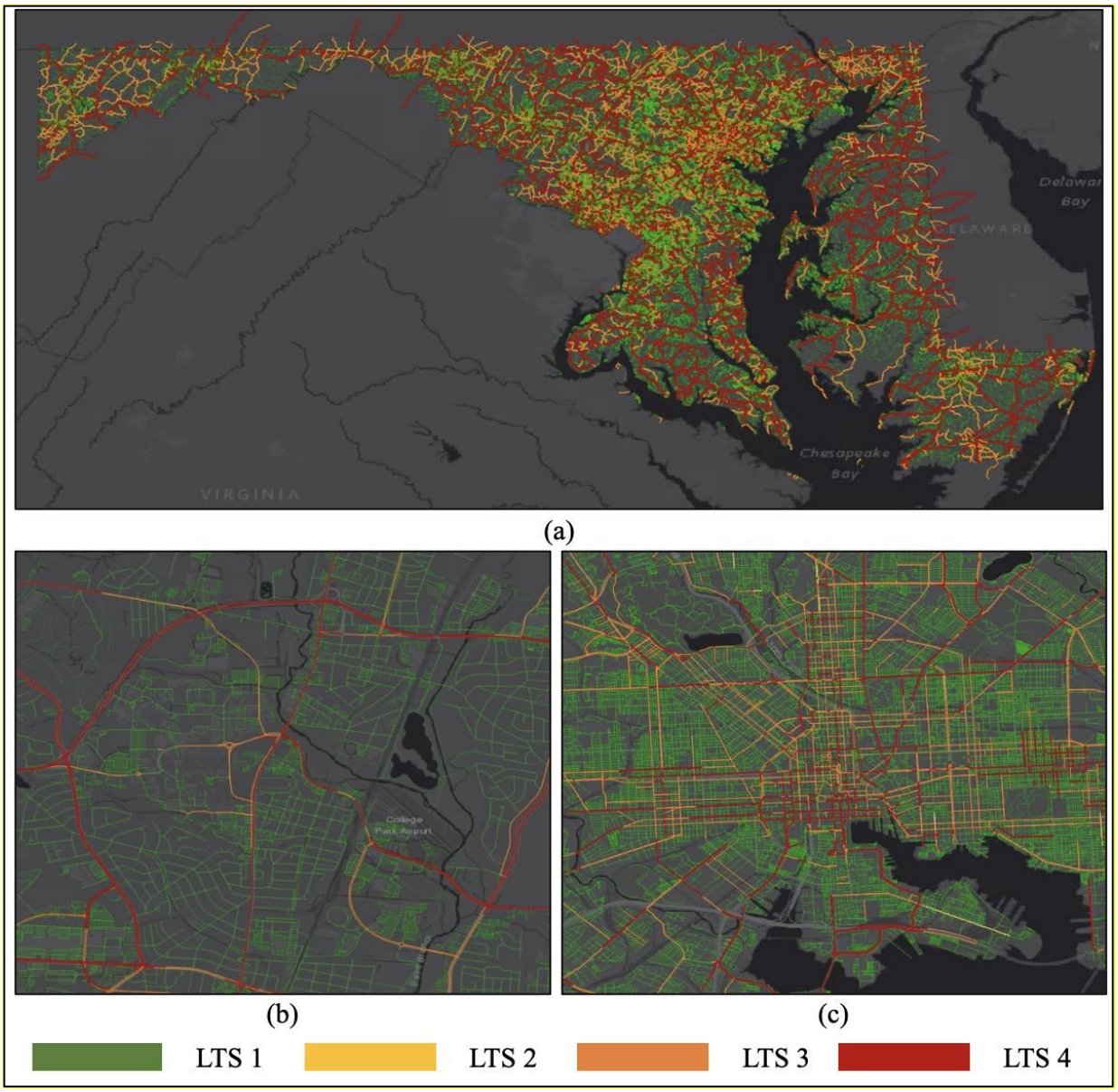


Figure 6-2. LTS Estimates for: (a) the state of Maryland; (b) University of Maryland College Park Campus; (c) Baltimore City

Figure 6-2 (a) shows the estimated LTS for all roads in the state of Maryland based on the correspondence table developed in this study. As shown in Figure 6-2 (b), it can be seen that most roads in the University of Maryland (UMD) campus have the lowest LTS, indicating that campus is more bicyclist-friendly. In the meantime, the entrances and exits of the campus, and the

surrounding major arterials (MD-193 and Route 1) all show highest LTS, indicating that these roads are neither safe nor friendly for bicyclists. Similarly, in Figure 6-2 (c), in the Inner Harbor area of Baltimore City, it can be seen that major roads all have high LTS although some of them have bicycle lanes. The Mount Vernon area with bike lanes shows a lower LTS value, indicating a bicyclists-friendly environment.

Using the estimated LTS at roadway segment level, the intersection LTS is calculated by averaging the LTS value for all incoming roadway segments, weighted equally for all approaches.

6.2.1.4 National Transit Map (NTM)

Initially released in 2016, the National Transit Map (NTM) database is a product of the U.S. Department of Transportation's Bureau of Transportation Statistics. This nationwide database provides information on fixed-guideway and fixed-route transit service across the entire U.S. including transit agencies' stops, routes, and schedules [152]. In this analysis, the NTM has been used to obtain information about transit stop locations.

6.2.1.5 National Walkability Index (NWI)

The National Walkability Index is a nationwide spatial data resource provided by the U.S. Environmental Protection Agency's Smart Growth Program. The National Walkability Index dataset ranks each U.S. Census block group according to its relative walkability [147]. The present study utilizes walkability scores from this national dataset in the analysis of pedestrian and bicyclist crashes that occurred at Maryland intersections.

6.2.1.6 OpenStreetMap (OSM) Intersections

The same OSM network used for vehicle volume estimation is also applied here. The intersection is extracted directly from the network, with features including incoming legs, signalizations, and bike-lane existence.

6.2.1.7 Smart Location Database (SLD)

The same SLD is also used for modeling pedestrian and bicyclist crashes to account for land use and built environment characteristics such as population and employment densities, land use diversity, urban design attributes, destination accessibility, transit accessibility, and socioeconomic/sociodemographic characteristics at the census block group level. These characteristics of SLD make it a suitable dataset for examining the role of land use and the built environment in the frequency of pedestrian and bicyclist crashes.

6.2.1.8 Vulnerable Road User (i.e., Pedestrian and Bicyclist) Crash Data

Data from the Maryland State Government's open data portal [153] were utilized in this study to obtain all pedestrian and bicyclist-involved crash data from 2019.

6.2.2 Model Development

6.2.2.1 Model Dependent Variables

The dependent variable for the statistical models is the frequency of pedestrian and bicyclist crashes at a particular intersection. The frequency of pedestrian and bicyclist crashes is a nonnegative count.

Table 6-2. Frequency of Pedestrian/Bicyclist Crashes at Maryland Intersections in 2019

Number of Pedestrian and Bicyclist Crashes at the Intersection	**Frequency (Number of Intersections)**						**Percent**
	Severity Level			***Signalization Status***			
	Non-injury-causing Crashes	Injury-causing Crashes	Fatal Crashes	Signalized Intersection	Non-Signalized Intersection	Total	Total Percentage
0	0	0	0	6,714	183,698	190,412	98.92

1	173	1,696	68	589	1,348	1,937	1.01
2	3	114	3	62	58	120	0.06
3	0	19	1	12	8	20	0.01
4	0	5	0	5	0	5	0.00
5	0	3	0	2	1	3	0.00
Total	176	1,837	72	7,384	185,113	192,497	100.00

Table 6-2 tabulates the frequency of pedestrian and bicyclist crashes in 2019 at Maryland intersections as extracted from the Maryland State Government's open data portal [153]. The table also provides information about the pedestrian and bicyclist crashes with respect to signalization status of intersections as well as the severity level of crashes. The table shows that pedestrian/bicyclist crashes did not occur at a large fraction of intersections, leading to a data distribution that is positively skewed with many observations being zero. The preponderance of zeros is a common characteristic of count data that represent occurrence of an event (in this case, occurrence of a pedestrian/bicyclist crash at an intersection).

6.2.2.2 Independent Variables in Models

The independent variables included in the models were selected based on the literature review and engineering judgement. These independent variables represent the key contributing factors that affect the occurrence of pedestrian and bicyclist crashes, including sociodemographic and socioeconomic factors, land use and built environment factors, and design-, traffic-, and travel-related factors—a few of which characterize the safety risk exposure for pedestrians and bicyclists at intersections.

Table 6-3. Independent Variables for Pedestrian/Bicyclist Crash Frequency Models

Variable	Description	Mean	SD	Data Source
Intersection Design- and Traffic-related Attributes				

Legs	Number of intersection approaches	3.46	1.26	OSM
Traffic Signal	Intersection is signalized – 1: yes, 0: no	0.04	0.19	OSM
Average Level of Traffic Stress (LTS)	Average of LTS rating for all intersection approaches	1.62	0.99	LTS (OSM)
Average Daily Pedestrian/Bicyclist Volume	Average daily pedestrian/bicyclist volume passing through the intersection	84.93	210.40	MDLD (MTI)
Average Daily Vehicle Volume	Average daily vehicle volume passing through the intersection	376.31	797.99	MDLD (MTI)
Travel-related Attributes				
Automobile Mode Share	Automobile commute mode share for CBG	84.61	13.22	ACS
Public Transportation Mode Share	Public transit commute mode share for CBG	6.12	8.60	ACS
Nonmotorized Mode Share	Walk/Bike commute mode share for CBG	2.40	5.44	ACS
Land Use and Built Environment Attributes				
Road Network Density	Total road network density for CBG	11.20	8.35	SLD
Pedestrian-oriented Network Density	Network density in terms of facility miles of pedestrian-oriented links per mile^2 of CBG	8.18	6.73	SLD
Multimodal Network Density	Network density in terms of facility miles of multimodal links per mile^2 of CBG	2.01	2.35	SLD
Intersection Density	Intersection density in terms of automobile-oriented intersections per mile^2 of CBG	1.21	2.83	SLD
Residential Density	Gross residential density (housing units/acres) for CBG	2.23	3.46	SLD
Employment Density	Gross employment density (jobs/acres) for CBG	2.34	10.38	SLD
Activity Density	Gross activity density [(employment + housing units)/acres] for CBG	4.58	11.64	SLD
Land Use Diversity	Employment and household entropy for CBG	0.50	0.22	SLD
National Walkability Index	Walkability index score for CBG	9.45	4.20	NWI
Number of Transit Stops	Count of bus stops within CBG	2.64	5.41	NTM
Sociodemographic and Socioeconomic Attributes				
Population Over 65	Percent of population ≥ 65 years old in CBG	16.63	9.27	ACS
Population Under 18	Percent of population < 18 years old in CBG	21.32	7.45	ACS
Male Population	Percent of the male population in CBG	48.58	6.46	ACS
African American Population	Percent of African American population in CBG	24.92	27.62	ACS
Enrolled in School	Percent of CBG population enrolled in school	25.01	8.83	ACS
Unemployed	Percent of unemployed population in CBG	3.15	2.96	ACS
Low-wage Workers	Percent of CBG workers earning ≤ $ 1250/month	21.61	4.75	SLD

Households with No Cars	Percent of zero-car households in CBG	6.46	10.45	SLD

***Notes**: CBG = Census Block Group; SD = Standard Deviation; MTI = Maryland Transportation Institute.*

Table 6-3 provides information on the original independent variables considered for inclusion in the models. Pearson pairwise correlation coefficients were computed to examine the correlations between all original independent variables. To reduce the risk of multicollinearity, highly correlated variables were not simultaneously included in the models. The final independent variables/models were selected by comparing the estimated models based on their Akaike's information criterion (AIC) and Bayesian information criterion (BIC). The model with the smallest AIC and/or BIC is considered a more appropriate model among a set of candidate models. These information criteria have been formulated in statistical and econometric analysis reference books [154, 155] as follows:

$$AIC = -2logL(\hat{\theta}) + 2k \qquad (10)$$

$$BIC = -2logL(\hat{\theta}) + klog(n) \qquad (11)$$

where, θ is the vector of model parameters; $L(\hat{\theta})$ is the likelihood of the candidate model given the data when evaluated at the maximum likelihood estimate of θ; k is the number of estimated parameters in the candidate model; n is the number of observations.

6.2.3 Pedestrian and Bicyclist Crash Estimation Results

6.2.3.1 Model Selection and Estimation

Table 6-4. Results of the Pedestrian/Bicyclist Crash Frequency Models

Dependent Variable: Frequency of Pedestrian/Bicyclist Crashes at the Intersection				
Independent Variable	**Type of Model**			
	Poisson	**NB**	**ZIP**	**ZINB**
Intersection Design- and Traffic-related Attributes				

Legs (*Reference: Number of Intersection Legs = 3 Legs*)				
Number of Intersection Legs = 4	0.0604	0.0585	0.0093	0.0028
Number of Intersection Legs ≥ 5	0.3396***	0.3441***	0.2390***	0.2281***
Traffic Signal (*1: Signalized Intersection, 0: Otherwise*)	1.0143***	1.0241***	0.9123***	0.9351***
Average Level of Traffic Stress (LTS)	0.4517***	0.4331***	0.2239***	0.2217***
Average Daily Pedestrian/Bicyclist Volume	0.0004***	0.0006***	0.0003***	0.0004***
Average Daily Vehicle Volume	0.0002***	0.0002***	0.0001***	0.0001***
Travel-related Attributes				
Automobile Mode Share	0.0035	0.0035	0.0036	0.0036
Public Transportation Mode Share	0.0058*	0.0070**	0.0061*	0.0068**
Nonmotorized Mode Share	0.0058	0.0045	0.0078**	0.0071*
Land Use and Built Environment Attributes				
Road Network Density	0.0255***	0.0262***	0.0168***	0.0162***
Multimodal Network Density	-0.0163**	-0.0148*	-0.0091	-0.0083
Intersection Density	-0.0102	-0.0143**	-0.0072	-0.0086
Activity Density	0.0038***	0.0047***	0.0038***	0.0047***
Land Use Diversity	-0.1385	-0.0571	-0.0451	-0.0385
National Walkability Index	0.1157***	0.1080***	0.0688***	0.0687***
Number of Transit Stops	-0.0042	-0.0062	-0.0030	-0.0054
Sociodemographic and Socioeconomic Attributes				
Population Over 65 (%)	-0.0056*	-0.0060**	-0.0054**	-0.0053*
Population Under 18 (%)	-0.0033	-0.0047	-0.0046	-0.0051
Male Population (%)	-0.0044	-0.0040	-0.0034	-0.0031
African American Population (%)	0.0022**	0.0018*	-0.0002	-0.0002
Enrolled in School (%)	-0.0014	-0.0002	-0.0010	-0.0004
Unemployed (%)	0.0032	0.0050	0.0028	0.0035
Low-wage Workers (%)	0.0166***	0.0195***	0.0162***	0.0177***
Households with No Cars (%)	0.0066***	0.0076***	0.0067***	0.0072***
Model Goodness of Fit/Information Criteria				
Pseudo R^2	0.1968	0.1809	—	—
Akaike's Information Criterion (AIC)	20141.06	19952.39	19574.20	19466.10
Bayesian Information Criterion (BIC)	20395.25	20216.75	19848.73	19750.79
Dispersion Parameter (Alpha)	—	1.9938	1.3203	1.3203
Likelihood Ratio Chi^2 Test of Alpha = 0	—	chibar2(01) = 190.67***	—	—
Number of Observations (i.e., Intersections) = 192,497				

***Notes**: *, **, *** = Coefficient is significant at the 10%, 5% and 1% significance level, respectively; — = N/A.*

Table 6-4 provides model estimation results for the models developed to relate the frequency of pedestrian and bicyclist crashes to various key contributing factors that affect crash occurrence at intersections.

As introduced above, four models are used in this analysis, including Poisson, NB, ZIP, and ZINB. The main assumption of the Poisson model, which requires the mean of the count variable to be equal to its variance, has been checked. The mean and variance of the frequency of pedestrian and bicyclist crashes frequency are 0.0118028 and 0.0141571, respectively. Therefore, the variance is larger than the mean, indicating presence of overdispersion in data. The likelihood ratio Chi^2 test (a post-estimation test for the NB model indicating if the dispersion parameter, alpha, is equal to zero) has been performed. The result of this test is statistically significant (p-value < 0.0001), which suggests that the dependent variable is overdispersed and is not adequately estimated by the Poisson model. The AIC and BIC for all the models have been computed and compared with each other. The comparison reveals that the ZINB model has the smallest AICs and BICs among all four models.

Consideration has been given to the dispersion parameter (alpha). The dispersion parameter has been estimated by the NB, ZIP, and ZINB models to be greater than zero (1.9938 in the NB model and 1.3203 in the ZIP and ZINB models). Therefore, due to the data being overdispersed, the model developed using these data is better estimated with the ZINB rather than ZIP modeling methodology. Thus, the ZINB model is the most suitable model for estimating the frequency of pedestrian/bicyclist crashes at Maryland intersections.

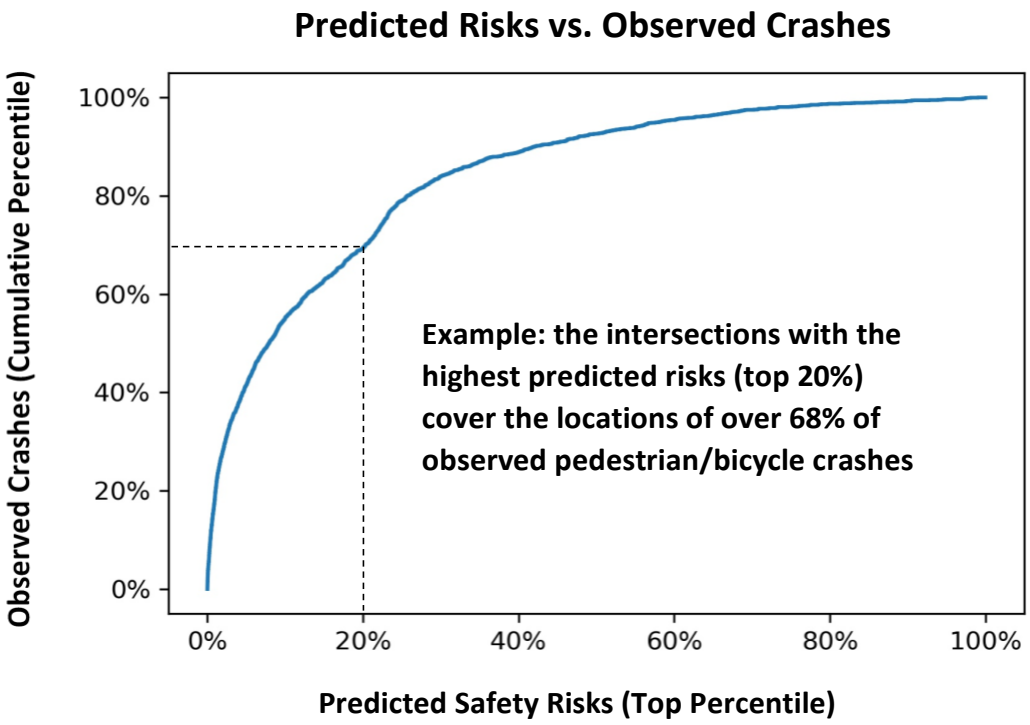


Figure 6-3. ZINB model performance.

The ZINB model is subsequently used to predict the number of pedestrian and bicycle crashes at each intersection to assess crash risk. The model is highly capable of capturing the pedestrian/bicycle high-crash-risk intersections. As depicted in Figure 6-3, the intersections with the highest predicted risks cover major locations where the observed crashes occurred. The following sections interpret the result for each category of independent variables.

6.2.3.2 Intersection Design and Traffic Attributes

The model estimation results indicate that the frequency of pedestrian and bicyclist crashes at an intersection is associated with intersection design- and traffic-related characteristics such as number of intersection legs, presence of a traffic signal at the intersection, average level of traffic

stress (LTS) for the intersection, and average daily pedestrian/bicyclist and vehicle volumes at the intersection.

More specifically, frequency of pedestrian and bicyclist crashes at the intersection is positively associated with the intersection having a larger number of approaches (i.e., number of intersection legs ≥ 5), as well as the presence of a traffic signal. This result is consistent with the literature suggesting that higher numbers of pedestrian and bicycle crashes as well as higher injury risk for bicyclists are associated with higher numbers of intersection approaches [115, 117], and signalized intersections [112, 115]. On one hand, intersections with fewer approaches create fewer turning conflicts [114]. On the other hand, intersections with more approaches may have higher vehicular and pedestrian/bicyclist volumes, and thereby are more prone to crashes. With respect to pedestrians, Tiwari et al. [156] found that as signal waiting time increased at signalized intersections, pedestrians became impatient and violated the traffic signal. Therefore, attempting to cross the intersection prematurely by pedestrians at signalized intersections may have contributed to the higher frequency of pedestrian-involved crashes at signalized intersections within the study area.

Average LTS for the intersection shows a significant positive correlation with frequency of pedestrian and bicyclist crashes at the intersection. This is a reasonable result considering that higher LTS ratings represent conditions that impose higher traffic-related stress on bicyclists including interaction with higher speed traffic, close proximity to high speed traffic, and multilane traffic [151]—all of which can contribute to bicyclist-involved crashes.

As expected, frequency of pedestrian and bicyclist crashes is also positively associated with the average daily pedestrian/bicyclist and vehicle volume, which represent safety risk exposure for pedestrians and bicyclists. The estimated coefficients of the two variables are

statistically significant at the 1% significance level in all four models, indicating the key role of traffic volumes in crashes involving vulnerable road users. The results of the models corroborate past findings suggesting that increased frequencies of pedestrian- and bicyclist-involved crashes at intersections are associated with increased levels of risk exposure measures such as vehicle volumes, pedestrian volumes, and bicycle volumes [115, 121, 124].

6.2.3.3 Travel Attributes

The travel-related characteristics also play a key role. The results show that increased frequencies of pedestrian/bicyclist crashes at intersections are observed at census block groups with higher public transportation and nonmotorized mode shares. These alternative modes of travel may lead to increased risk exposure for pedestrians and bicyclists, which can in turn lead to increased crash frequency for such vulnerable road users. This aligns with Ukkusuri et al. [117] which found that the proportion of commuters who travel to work by transit or nonmotorized modes were among the contributing factors to the number of pedestrian crashes.

6.2.3.4 Land Use and Built Environment Attributes

The impacts of land use and built environment characteristics on the frequency of pedestrian and bicyclist crashes are summarized below:

Total road network density has a statistically significant and positive association with the frequency of pedestrian and bicyclist crashes at intersections, whereas multimodal network density and intersection density have negative associations with those crashes. The negative association between multimodal network density and frequency of pedestrian and bicyclist crashes (only statistically significant in Poisson and NB models) highlights the role of multimodal network designs in increasing safety of vulnerable road users. by reducing the number of pedestrian and bicyclist crashes at those intersections. Increased activity density is also observed to be associated

with increased numbers of pedestrian and bicyclist crashes at the intersection, which is in line with the literature [118]. The National Walkability Index also exhibits a positive and statistically significant coefficient in all four models, where increased frequencies of pedestrian and bicyclist crashes at intersections are associated with increased walkability. This is an expected result since higher levels of walkability could increase pedestrian activity, which can in turn increase risk exposure for pedestrians at intersections.

6.2.3.5 Sociodemographic and Socioeconomic Attributes

Based on the model results, higher numbers of pedestrian and bicyclist crashes at the intersection are associated with lower percentages of the senior population (i.e., population over 65 years old), which is in line with Ukkusuri et al [117]. Further, higher frequencies of pedestrian and bicyclist crashes at intersections are associated with higher percentages of low-wage workers and higher percentages of households with no private vehicle. These findings may indicate higher usage levels of nonmotorized and public transit modes by individuals with a lower socioeconomic status.

6.2.4 Assessment of Contribution of the Vehicle Volume and Pedestrian and Bicyclist Volume Estimated by MDLD to Model Performance

To evaluate whether inclusion of vehicle volume and pedestrian and bicyclist volume estimated by MDLD can improve the crash estimation models, AIC and BIC of all four models have been computed under with and without MDLD-estimated volume scenarios.

Table 6-5. Model Improvement Assessment Based on LBS Variables

Type of Model / **MDLD Variable (s) Included**	**Poisson Model**	**NB Model**	**ZIP Model**	**ZINB Model**
	Information Criteria (AIC; BIC)			

Average Daily Pedestrian/Bicyclist Volume & Average Daily Vehicle Volume	20141; 20395	19952; 20217	19574; 19849	**19466; 19751**
No Vehicle Volume and Pedestrian and Bicyclist Volume	20407; 20641	20238; 20482	19651; 19905	19559; 19823

Table 6-5 summarizes these information criteria for all four models. Comparison of the AICs and BICs indicates that for all four model types, the model that includes the vehicle volume and pedestrian and bicyclist volume has the smaller AIC and BIC values. This means that the addition of these two variables is an improvement to the models. Moreover, the appropriateness of the ZINB model with vehicle volume and pedestrian and bicyclist volume is further emphasized by having the smallest AIC and BIC values among all model types.

6.3 Conclusions and Discussions

This chapter used the estimated vehicle volume in Chapter 5 to support pedestrian and bicyclist crash estimation. It found out that estimated vehicle volume based on MDLD can be leveraged in safety risk exposure analysis demonstrated via a case study on estimating pedestrian and bicyclist crashes in Maryland. In particular, the proposed estimation method can particularly be beneficial for safety risk exposure and crash analysis with respect to vulnerable road users (e.g., pedestrians and bicyclists). Pedestrian and bicyclist exposure data have traditionally been collected through surveys or count collections at sample locations, which is limited and suffer from high labor cost [119, 157]. In addition to being costly and labor-intensive, these conventional data collection methods are susceptible to subjectivity and may yield inaccurate data. Consequently, high-quality and readily available pedestrian and bicyclist exposure data are considered as limitation in safety analysis [158]. As exposure data are crucial for contextualization of crash analysis and prioritization of safety countermeasures [119], utilization of high-quality and consistent exposure

data is imperative. When it comes to safety analysis, using MDLD for volume estimation—as performed in this study—provides a tremendous advantage over using data obtained from traditional volume estimation methods. This is due to the potential of the MDLD to produce more reliable exposure data. Employment of such high-fidelity exposure data (i.e., MDLD-estimated volumes) as input for safety and crash analyses can lead to more accurate results and guide data-driven, evidence-based policy decision-making to improve the safety of all road users, including the most vulnerable ones.

The main advantage of using such location big data instead of traditional exposure data is that having exposure measurements for more locations enable an analyst to identify the potential safety hotspots that may otherwise be overlooked. Using classical crash estimation statistical models, including Poisson, NB, ZIP, and ZINB, the results show that inclusion of vehicle volume and pedestrian and bicyclist volume in the models improves the performance of pedestrian and bicyclist crash estimation. As consistent and high-quality pedestrian and bicyclist exposure data is often regarded as a limitation in safety analysis [115, 119, 158], this finding highlights the critical role of transportation big data in contextualization of pedestrian/bicyclist crash analysis, where the main contribution of this chapter also lies.

Chapter 7: Conclusions, Limitations and Future Works

7.1 Conclusions

This dissertation comprehensively examines the state-of-the-practice applications and the state-of-the-art models developed based on emerging transportation big data, identifies key metrics, and develops two big-data driven frameworks to enhance traffic operations and safety.

A literature review on models, tools, and metrics used for various levels of traffic analysis was conducted, and a survey was distributed to transportation professionals to quantify the importance of these key metrics for improving traffic operations and safety. Based on the literature review and survey insights, three types of upstream data were identified for supporting transportation projects, namely vehicle volume, travel time (or traffic speed), and crash rate (or crash count). This dissertation further proposed two big-data driven frameworks to produce these data and address both traffic operations and safety issues.

The first big-data driven framework integrates transportation big data sources from both the demand and the supply sides to improve the accuracy and reliability of emergency medical services (EMS) and trauma triage decisions for elderly persons at crash sites. Using decision tree models, with records of over 55,000 elderly patients, results demonstrate that the proposed framework contribute to enhanced EMS decision and trauma triage accuracy for the elderly, and saving more lives from severe vehicle crashes.

The second big-data driven framework employs a series of cloud-based computational algorithms to extract multimodal trajectories and trip rosters from terabytes of MDLD to further estimate vehicle and pedestrian and bicyclist volume at all roads. The outputs of the framework are further used to support pedestrian and bicyclist crashes at intersection-level. Results indicate

that the proposed framework can produce reliable vehicle volume estimates and estimated pedestrian and bicyclist crashes, while also demonstrating its transferability and generalization ability.

In summary, this dissertation comprehensively examines the literature on transportation big data applications and proposes two big-data driven frameworks demonstrated with two real-world case studies. Results reveal the feasibility and advantages of empowering traffic operations and safety analysis with transportation big data.

7.2 Limitations

The limitations of this dissertation are listed as follows:

For Chapter 4, as a first attempt, only short-term crash outcomes data are integrated and employed. Verifications from hospitals, such as in-patient/out-patient records, hospital transfers, and long-term recovery and health outcomes, can and should be used to validate and calibrate the models and further improve the model usefulness. Secondly, the modeling tools developed in this chapter would be more valuable if assessed in real-time. With increased availability of real-time traffic data, location-based service data streams, and potential collaboration with hospitals and emergency response teams, the modeling algorithms can be further developed to produce time-dependent predictions, supporting stakeholders in making timely decisions.

For Chapters 5 and 6, first, various other factors with the potential to impact the frequency of pedestrian and bicyclist crashes were not included in the analysis, such as vehicle speeds and pavement conditions. For instance, although the LTS measure included in this analysis is a function of the speed limit—and can therefore serve as a proxy for vehicle speeds, a more direct measure of speed may capture the effect of vehicle speeds at the intersection on the frequency of pedestrian and bicyclist crashes more accurately. Further, a similar crash frequency

analysis can also be performed for road segments. In addition, future research can conduct safety risk analyses for other vulnerable road users such as e-scooter users. Finally, unobserved heterogeneity, a potential source of bias in aggregate count models, has not been accounted for in this initial study—a limitation that can be addressed in future related research. Nonetheless, the findings of this study contribute to the body of knowledge on road user safety by providing evidence on the role of big LBS exposure data in pedestrian/bicyclist safety analysis. These findings highlight the tremendous potential of this emerging source of big data, which offers many advantages, including elimination of costly and resource-intensive surveys and count collections and potential of generalization to other areas by providing data for the entire U.S.

In general, one major challenge in leveraging the transportation big data, either probe vehicle data or MDLD, is the sample representativeness and data quality. Due to the technology limitation, probe vehicle data and MDLD both only capture a part of the daily trips of a device when the location service is active. This variability in the LRI or, observation frequency, might also result in capturing more long distance/duration trips from active travelers, such as long-distance travel for leisure or business purposes or long-distance commute, which adds biases into the data. Also, for MDLD, it is difficult to accurately impute travel modes from MDLD, due to lack of both comprehensive training data and validation data. For instance, because of the signal loss and urban canyon effect, especially in major cities and metropolitan areas, the bus and rail travels might be hard to capture from MDLD, resulting in an overall underestimation.

7.3 Future Works

Future studies can improve on the following aspects:

Survey Design: Both proposed frameworks are calibrated and trained based on samples collected in the Washington Metropolitan Area and the state of Maryland. In the real world, the

safety characteristics and mobility characteristics might differ from region to region. In future research, enriching the training dataset or enlarging the study area might help improve the performance of the proposed frameworks by considering such heterogeneity.

Sample Bias: As mentioned above in limitations, both proposed frameworks, either use probe vehicle data or MDLD, are not able to represent the population-level travel characteristics. Also, these transportation big data cannot capture the travel behaviors of the population without mobile devices or vehicles, which might yield an underrepresentation of the younger, elder, and low-income persons.

To address these two problems, an additional weighting and validation process can be done on top of the sample results using land use and sociodemographic information to account for various dimensions of device characteristics. For instance, the weighting can be conducted based on income group, age groups, gender, race, and household characteristics. In the meantime, it is also important to conduct dedicated data collection, labeling, training, and validation process to enhance the training data. Rather than collecting data such as the traditional GPS data, this process should focus on assimilating the MDLD provided by data providers so that the models and algorithms developed based on the collected data are transferrable.

Data Coverage: Both proposed frameworks are calibrated and trained based on samples collected in the Washington Metropolitan Area and the state of Maryland. In the real world, the safety characteristics and mobility characteristics might differ among regions. In future research, enriching the training dataset that could cover different regions, such as the GPS-enhanced travel survey dataset available in Transportation Secure Data Center (TSDC) at National Renewable Energy Lab (NREL), might help improve the performance of the proposed frameworks by considering the travel behavior heterogeneity. In addition, the proposed framework relies on

multimodal transportation networks, including drive, rail and bus. For regions without well-maintained transportation networks, it could be hard to capture the rail/bus travel. To decrease the dependence on multimodal transportation networks, additional information such as acceleration and stop time can be potentially considered.

Study Scope: Both proposed frameworks can be improved regarding their study scopes.

For the first proposed framework in Chapter 4, the current study scope only covers the vulnerable population (people older than 65) while not considering the other population. The health condition and driving style heterogeneity among these age groups might result in different decision scenarios for both EMS and trauma triage. In the future, this framework can be extended to include each age group and produce an EMS and trauma triage decision process correspondingly.

For the second proposed framework in Chapter 5 and 6, the current study scope mainly focuses on vehicle volume estimation and applies the same methodology to estimate pedestrian and bicyclist volume at the same time. Though this is mainly due to lack of training data, separate models may be developed for pedestrian and bicyclist volume estimation only. This may account for the inherent difference between vehicle activity generation and pedestrian and bicyclist activity generation, where pedestrian and bicyclist activities are generated mostly in dense and populated urban areas while vehicle activities are along the highway system.

Appendix I. MDOT SHA Operations Practice Scan Survey

I.1 Survey

SECTION 1: INTRODUCTION & BASELINE ASSUMPTIONS TO SURVEY

Q1 Identification of Best Practice Metrics for Varying Levels of Traffic Operations Analysis

This survey is conducted by the Maryland Transportation Institute at the University of Maryland (UMD) in support of a Maryland Department of Transportation State Highway Administration (MDOT SHA)'s research project, titled "Best Practice Metrics for Varying Levels of Traffic Operations Analysis". The objectives of this survey are to:

Identify various performance metrics and evaluation tools developed and used by state or local jurisdictions across the nation during different stages of a transportation project.

Evaluate the usefulness of these performance metrics and evaluation tools in supporting the decision-making processes, learn what performance metrics and evaluation tools supported decision-making, and what performance metrics and evaluation tools did not seem to be very useful in decision-making.

Summarize the survey and pursue targeted in-depth interviews with additional stakeholders, share the findings with all who participated in the survey and/or interview, and develop a best practice summary of performance metrics and evaluation tools used for different levels of traffic operations analysis.

Q2 Which states do you currently work for? You may select multiple states if your organization crosses state lines (e.g. private consulting firm or Metropolitan Planning Organization).

☐ Alabama (1)……

Q3 Which agency do you work for?

- ○ U.S. Department of Transportation
- ○ State Department of Transportation
- ○ County
- ○ Local Municipality (e.g. town or city)
- ○ Metropolitan Planning Organization
- ○ Private Consulting Firm
- ○ Other _____

Q4 What projects do you most frequently work on?

- ○ Transit
- ○ Highway
- ○ Arterial
- ○ Pedestrian/Bicycle

Q5 Given the goals of this survey, would you: (Optional)

☐ Wish to see the results of this survey

☐ Volunteer to be a possible subject matter expert for further discussion

Q6 Please provide your name and contact information

○ Name _____

○ Email _____

Q7 Our initial practice scan shows that agencies have generally advanced seven types of transportation projects:

- ***Mobility***: reducing delays, typical capacity improvements, micro-mobility infrastructure, transit solutions, etc.
- ***Reliability***: technology deployments to manage transportation system more effectively.
- ***Safety***: severity of crashes, high rate of crashes, vulnerable user interactions with vehicles, freight design concerns, etc.
- ***Environmental***: stream restoration, flooding mitigation.
- ***Socio-economic***: economic revitalization, food desert programs, equity-related, etc.
- ***Recreational***: trails, visitor rest stops.
- ***Political***: for the purpose of this survey, we will not be delving into this category.

Of those, transportation projects experience three major timelines: ***feasibility & planning***, ***design & construction***, and ***maintenance & operations***, where various performance metrics may be used to evaluate the need for the project. The following questions will categorize the performance metrics identified in our practice scan for you to rank as those your agency uses in the decision-making process (i.e. to select the alternative moving forward).

SECTION 2: FEASIBILITY AND PLANNING PERFORMANCE METRICS

Q8 Please rank the use of the following Mobility performance metrics when performing a feasibility or planning level analysis:

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all	Not sure
Travel Time	○	○	○	○	○	○
Delay	○	○	○	○	○	○
Traffic Speed	○	○	○	○	○	○
Volume-to-Capacity Ratio (v/c)	○	○	○	○	○	○
Level of Service (LOS)	○	○	○	○	○	○
Vehicle Throughput	○	○	○	○	○	○
Person Throughput	○	○	○	○	○	○
Accessibility to Jobs	○	○	○	○	○	○
High Level Planning Estimated Cost (not design costs)	○	○	○	○	○	○
Pedestrian or Bicycle Level of Comfort/Stress	○	○	○	○	○	○
Bus Ridership	○	○	○	○	○	○
Accessibility for Persons with Disabilities Compliance	○	○	○	○	○	○

Connectivity of System (e.g. no dead-end sidewalks)	○	○	○	○	○	○
Number of Mode Shift Transfers	○	○	○	○	○	○
Long Term Operational Cost	○	○	○	○	○	○

Q9 What other ***Mobility*** performance metrics do you consider during the ***feasibility and planning*** stage and how would you rate them? Please fill the performance metrics' names and rate them. (Optional)

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all
(1)_____	○	○	○	○	○
(2)_____	○	○	○	○	○
(3) _____	○	○	○	○	○

Q10 Please rank the use of the following ***Reliability*** performance metrics when performing a ***feasibility or planning level analysis***:

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all	Not sure
Travel Time Index (TTI)	○	○	○	○	○	○
Planning Time Index (PTI)	○	○	○	○	○	○
Level of Travel Time Reliability (LOTTR)	○	○	○	○	○	○
Total Trip Time by Modes	○	○	○	○	○	○
Long Term Operation Cost	○	○	○	○	○	○

Q11 What other ***Reliability*** performance metrics do you consider during the ***feasibility and planning*** stage and how would you rate them? Please fill the performance metrics' names and rate them. (Optional)

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all
(1) _____	○	○	○	○	○
(2) _____	○	○	○	○	○
(3) _____	○	○	○	○	○

Q12 Please rank the use of the following ***Safety*** performance metrics when performing a ***feasibility or planning level analysis***:

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all	Not sure
Crash Reduction	○	○	○	○	○	○
Fatality Reduction	○	○	○	○	○	○
Conflict Reduction	○	○	○	○	○	○
Long Term Operation Cost	○	○	○	○	○	○

Q13 What other ***Safety*** performance metrics do you consider during the ***feasibility and planning*** stage and how would you rate them? Please fill the performance metrics' names and rate them. (Optional)

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all
(1) _____	○	○	○	○	○
(2) _____	○	○	○	○	○
(3) _____	○	○	○	○	○

Q14 Please rank the use of the following ***Environmental*** performance metrics when performing a ***feasibility or planning level analysis***:

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all	Not Sure
Emission Reduction	○	○	○	○	○	○
Vehicle Fuel Savings	○	○	○	○	○	○
Natural Resource Impact	○	○	○	○	○	○
Long Term Operation Cost	○	○	○	○	○	○

Q15 What other ***Environmental*** performance metrics do you consider during the ***feasibility and planning*** stage and how would you rate them? Please fill the performance metrics' names and rate them. (Optional)

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all
(1) _____	○	○	○	○	○
(2) _____	○	○	○	○	○
(3) _____	○	○	○	○	○

Q16 Please rank the use of the following ***Socio-economic*** performance metrics when performing a ***feasibility or planning level analysis***:

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all	Not sure
Employment	○	○	○	○	○	○
Health	○	○	○	○	○	○
Land Use	○	○	○	○	○	○
Regional Economic Development	○	○	○	○	○	○
Social Equity	○	○	○	○	○	○
Long Term Operation Cost	○	○	○	○	○	○

Q17 What other ***Socio-economic*** performance metrics do you consider during the ***feasibility and planning*** stage and how would you rate them? Please fill the performance metrics' names and rate them. (Optional)

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all
(1) _____	○	○	○	○	○
(2) _____	○	○	○	○	○
(3) _____	○	○	○	○	○

Q18 Please rank the use of the following ***Recreational*** performance metrics when performing a ***feasibility or planning level analysis***:

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all	Not Sure
Number of Trail Users	○	○	○	○	○	○
Visitation	○	○	○	○	○	○
Shelter Reservations	○	○	○	○	○	○
Recreation Event Participation	○	○	○	○	○	○
Long Term Operation Cost	○	○	○	○	○	○

Q19 What other ***Recreational*** performance metrics do you consider during the ***feasibility and planning*** stage and how would you rate them? Please fill the performance metrics' names and rate them. (Optional)

	Always used	Frequently used	Occasionally used	Used once or twice before (ad-hoc)	Not used at all
(1) _____	○	○	○	○	○
(2) _____	○	○	○	○	○
(3) _____	○	○	○	○	○

SECTION 3: DESIGN AND CONSTRUCTION

Q20 Given all types of projects, please rank the following in the most frequent reason why a project *did not* move forward. (Rank the reasons from 1 to 6 vertically, where 1 represents the most frequent reason, 6 represents the least frequent reason)

	Mobility	Reliability	Safety	Environmental	Socio-economic	Recreational
Cost generally too high	1	1	1	1	1	1
Cost Benefit Analysis too low	2	2	2	2	2	2
Public opposition	3	3	3	3	3	3
Major design flaw identified (e.g. materials, hazards etc)	4	4	4	4	4	4
Lack of Confidence in the Projected Benefits	5	5	5	5	5	5
Other, please specify_____	6	6	6	6	6	6

Q21 Does your agency use standard performance metrics during design and construction to determine if a project should keep moving forward?

- ○ Yes
- ○ No

Q22 What are these standard performance metrics during design and construction? Please describe below

_____

_____

_____

_____

_____

SECTION 4: MAINTENANCE AND OPERATIONS

Q23 During maintenance and operations, which of the following metrics are used to keep the system performing to its desired level of performance:

- ☐ Pavement condition
- ☐ Threshold of volume or speeds that must be met
- ☐ Others, please specify _____

Q24 What performance metrics **should** be used in maintenance and operations?

_____

_____

_____

_____

_____

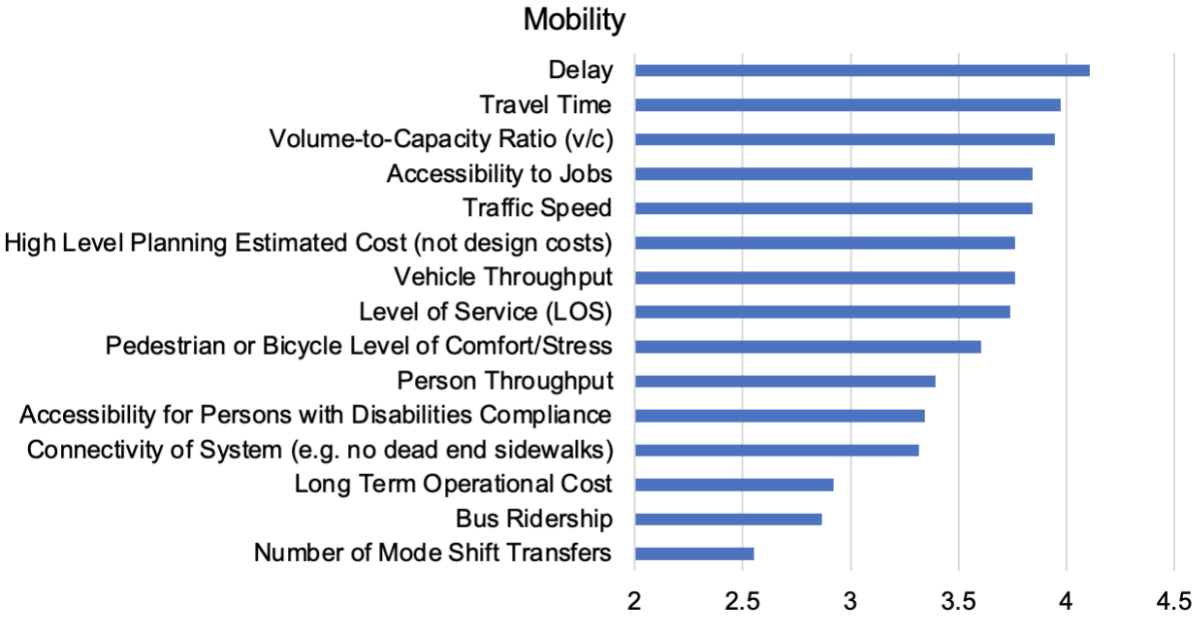
Please proceed to the next page to complete the survey.

I.2 Survey Results

The scores for each performance metric from each correspondent are summed as the total score. The total score for each performance metric is divided by the number of respondents in order to perform ranking. In addition, the respondents also recommended other metrics that were not included in our design and provided the scores for them. These performance metrics should also be taken into account based on the specific project needs.

For Mobility-related projects (see Figure A3), the top three metrics that have been frequently used are Delay (4.10), Travel Time (3.97), and Volume-to-Capacity Ratio (v/c) (3.95).

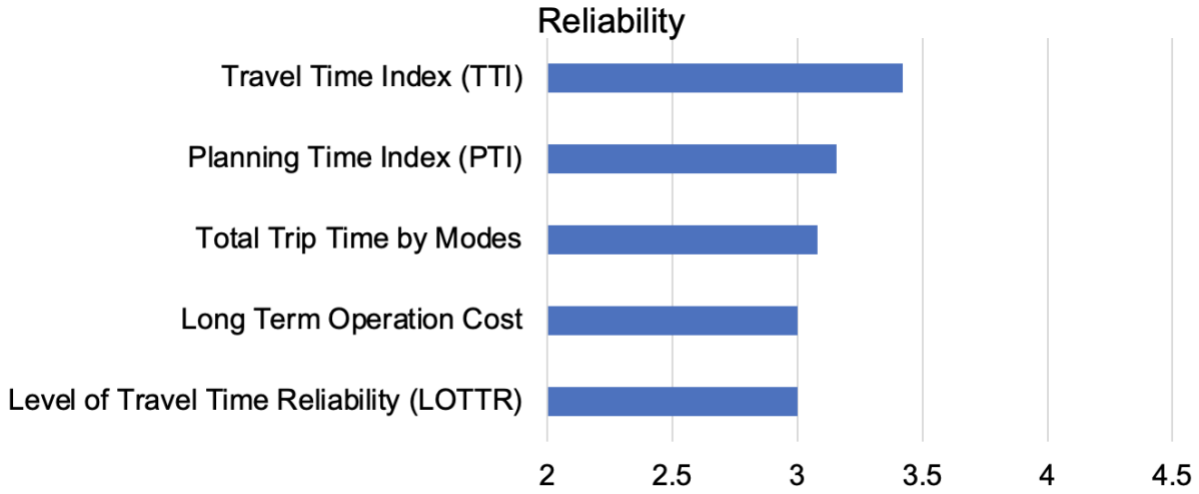
The three least used metrics are Number of Mode Shift Transfers (2.55), Bus Ridership (2.87), and Long-Term Operational Cost (2.92).



Respondents also recommended the mobility metrics not mentioned in the survey:

Total Score	Metrics	# of Respondents
16	Government Operations (e.g., Equipment availability, Master plan comformance, Resource allocation)	4
15	Multimodal Mobility (e.g., Mode share, Transfer Time, Bicycling network, Frequency at key stops, Accommodation of Amish buggies)	5
4	Average Vehicle Occupancy Destinations	1
4	Driver Expectancy	1
4	Diversion Off Local Facilities	1
3	Business Opening Times	1

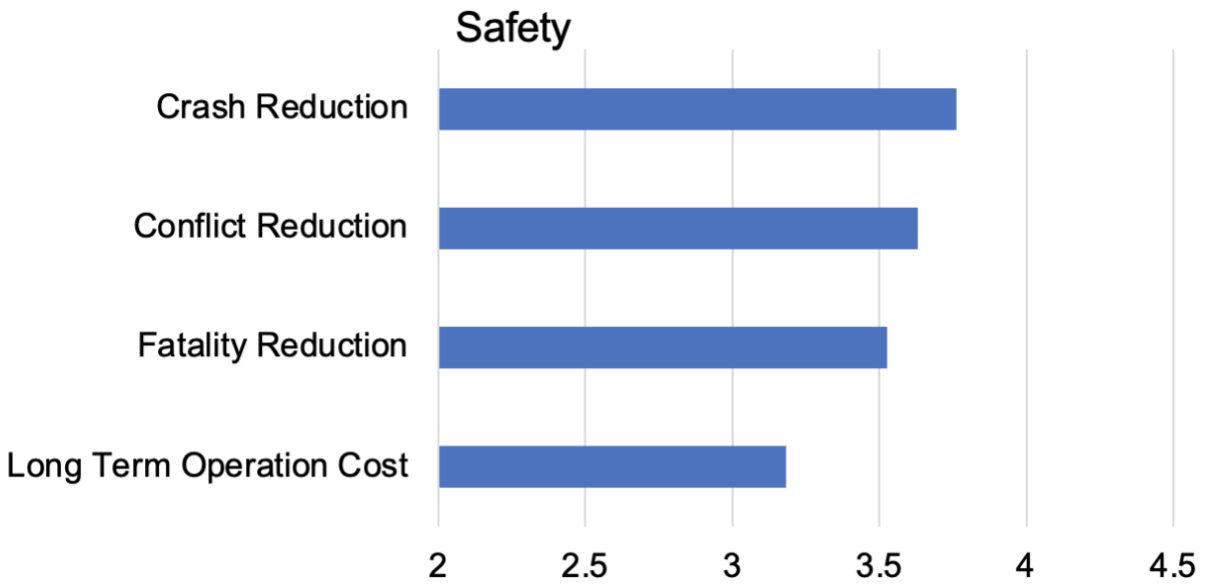
For Reliability related projects, the top metric that has been frequently used is Travel Time Index (3.42) and the least used metric is the level of travel time reliability (3.16).



Respondents also recommended the reliability metrics not mentioned in the survey:

Total Score	Metrics	# of Respondents
16	Congestion Impact (e.g., Duration of congestion (compared to similar periods, Transit on-time performance, Wait times, delays, buffer index)	5
9	Safety Impact (e.g., Level of comfort/safety, % days with incidents (subdivided by impact))	2
5	Average Travel Speed	1
4	Interconnectedness	1

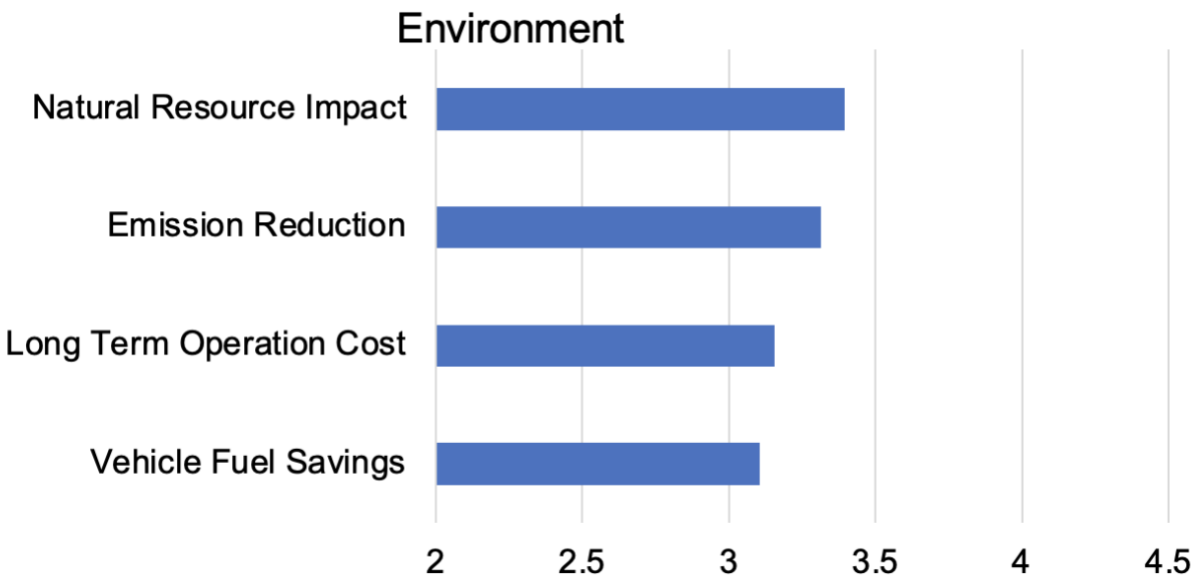
For Safety related projects, the top metric that has been frequently used is Crash Reduction (3.76) and the least used metric is Long Term Operational Cost (3.18). The scores of the four safety metrics are very close, indicating that the survey respondents rated these metrics with a similar level of importance.



Respondents also recommended the safety metrics not mentioned in the survey:

Total Score	Metrics	# of Respondents
12	Pedestrian and Bicyclist Safety (e.g., Location of trails/bike lanes, pedestrian movement, safety parameters per mode)	3
8	Incident Rate (e.g., Incident rate per mile, Goal of "no longer significantly higher than statewide average")	2
8	Speed Limits & Speed of Nearby Traffic	2
4	Visibility / Lighting	1

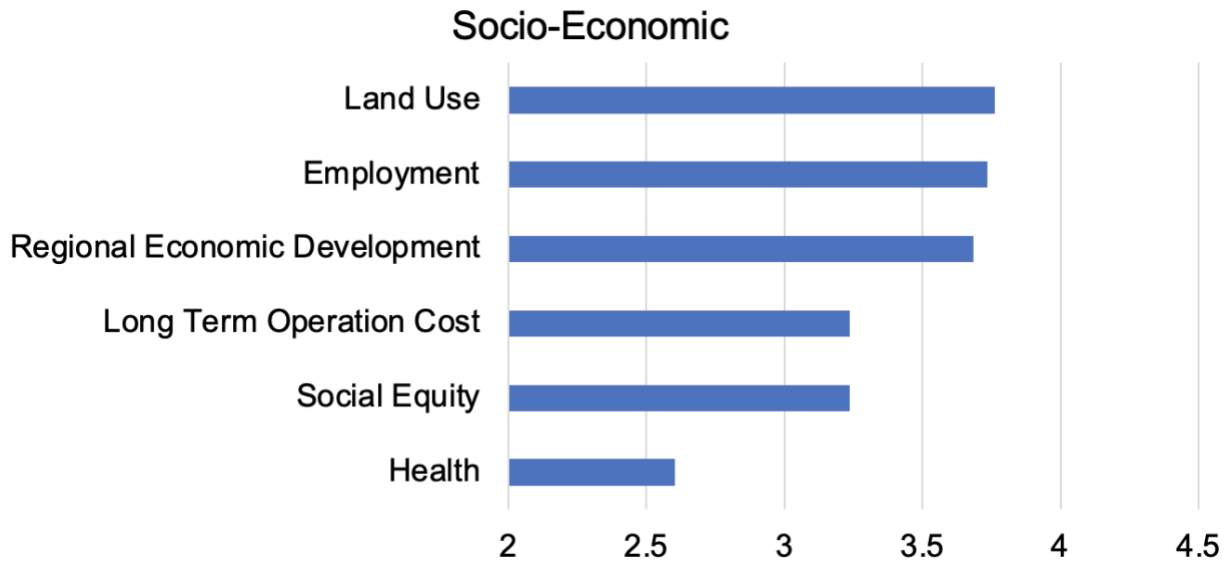
For Environment-related projects, the top metric that has been frequently used is Natural Resource Impact (3.39) and the least used metric is Vehicle Fuel Savings (3.10).



Respondents also recommended the environment metrics not mentioned in the survey:

Total Score	Metrics	# of Respondents
16	Hazardous Impacts (e.g., Flood planning, Stormwater planning, Draining)	5
11	Environmental Exposure (e.g., Noise impacts, Exposure per person to emissions)	3
8	Cost of Environmental Testing (e.g. Rural preservation)	2
5	Indirect/Cumulative Effects	1
3	Development Control	1

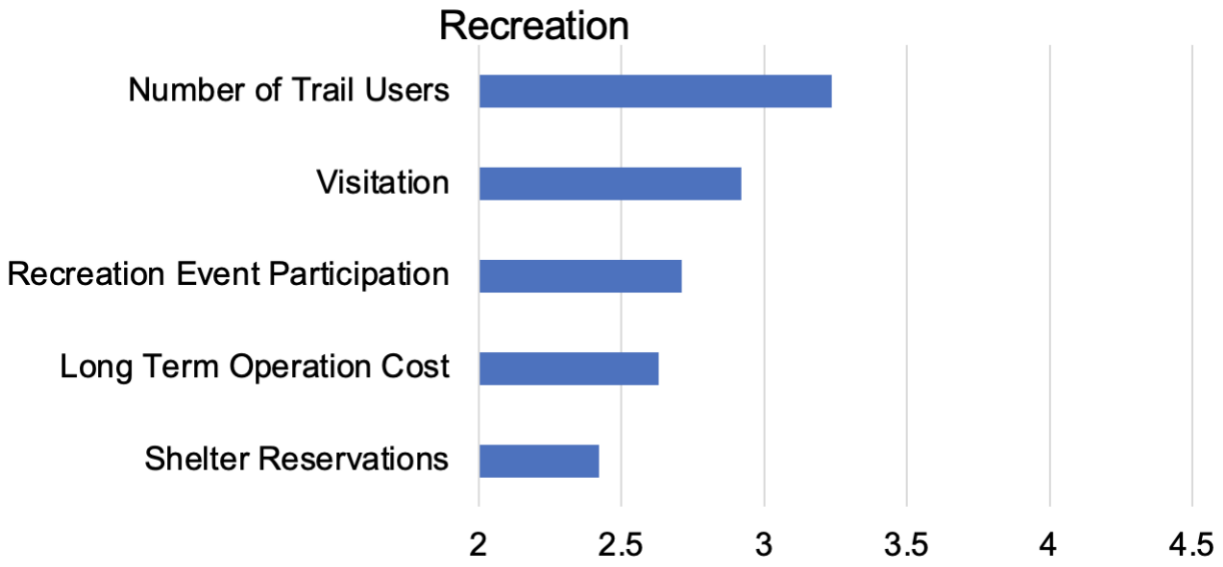
For Socio-economic-related projects, the top metric that has been frequently used is Land Use (3.76) and the least used metric is Health (2.60).



Respondents also recommended the socio-economic metrics not mentioned in the survey:

Score	Metrics	# of Respondents
13	Community Impact (e.g. Community revitalization, Community/Historic resources, Older adult demographic)	3
5	Access to Public Transportation	1
5	Noise	1
5	Indirect/Cumulative Effects	1

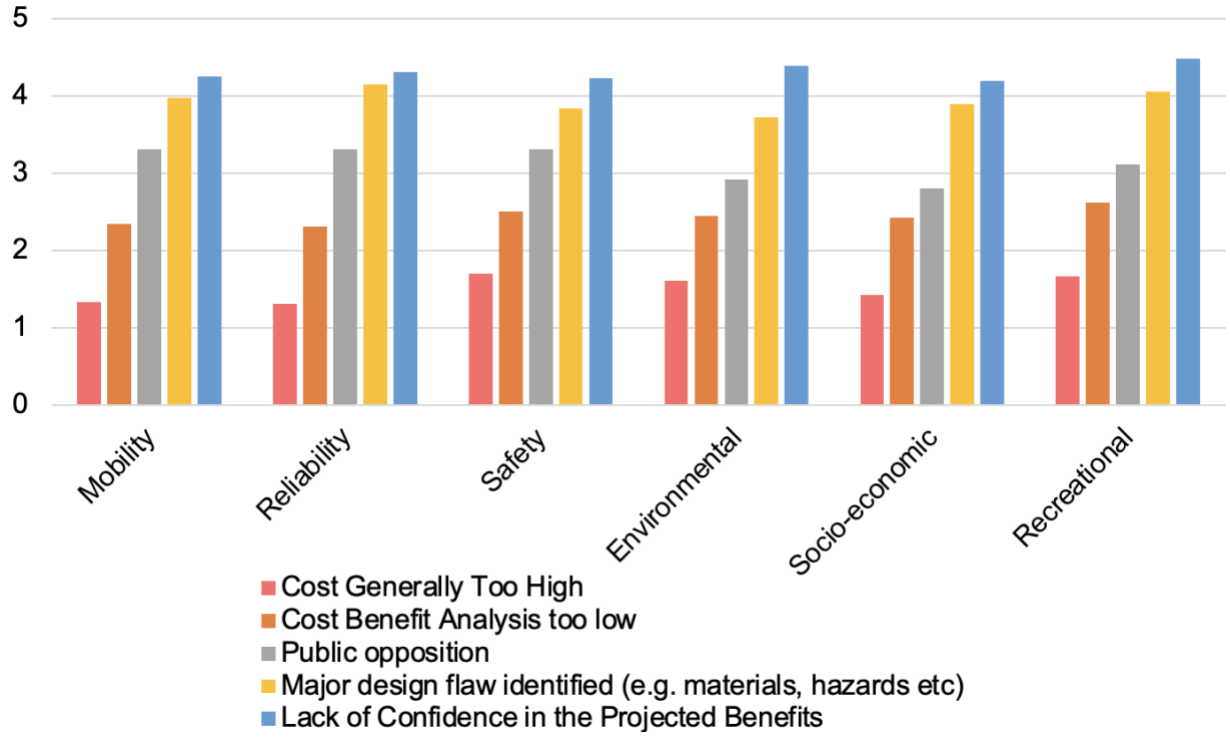
For Recreation-related projects, the top metric that has been frequently used is Number of Trail Users (3.24) and the least used metrics are Shelter Reservations (2.42).



Respondents also recommended the recreation metrics not mentioned in the survey:

Total Score	Metrics	# of Respondents
18	Trail Conditions (e.g., Widths of crossing/sidewalks/trails, Nexus to other trails / bike networks, Barrier separate)	4
5	ADA Purpose Built	1
5	High Accident Locations	1
3	Open Space Program Implementation	1
3	Social Equity	1

The score in the following figure shows the rank of different reasons of why the different kinds of projects were not moved forward. The lower the score the more important is the reason. For most projects, it can be seen that the “Cost Generally Too High” is the number one reason.



Respondents also provided some standard performance metrics during design and construction state:

Metrics
Is the project still within the anticipated cost?
Is the mix of projects to be funded annually a reasonable distribution across modes?
Department of Public Works staff/engineers reviews such metrics. Adequate Public Facilities Ordinance APFO is also in place which requires such an assessment.
Travel Time, Volume/Capacity, Level of Service (LOS), Delay, 95% queue, Throughput, Hours of congestion, Bike and Ped Connectivity, etc.
Cost and O&M projections versus use

During the maintenance and operations phase, respondents recommended “on-time performance”, “alternative routes”, “bridge condition”, “state of good repair”, “age of transit flee”, “surface

condition", "Signage Availability", "Sufficient Funding", "Clear Marking (e.g. marking for crosswalks, travel lanes)", "Reporting Issues", "Priority Lists".

Bibliography

1. Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*. 68, 285-299, (2016).
2. Yang, M., Pan, Y., Darzi, A., Ghader, S., Xiong, C. and Zhang, L. A data-driven travel mode share estimation framework based on mobile device location data. *Transportation*, pp.1-45 (2021).
3. Battelle. Global Positioning Systems for Personal Travel Surveys: Lexington Area Travel Data Collection Test. Final Report. FHWA, U.S. Department of Transportation, (1997).
4. 2000–2001 California Statewide Household Travel Survey. Final Report. NuStats, Austin, Tex (2002).
5. Kansas City Regional Travel Survey. Final Report. NuStats, Austin, Tex, (2004).
6. Ojah, M. and Pearson, D. F.. 2006 Austin/San Antonio GPS-Enhanced Household Travel Survey. Technical Summary. Texas Department of Transportation, (2008).
7. J. Wolf, and Lee, M. Synthesis of and Statistics for Recent GPS-Enhanced Travel Surveys. Proc., *International Conference on Survey Methods in Transport: Harmonization and Data Comparability*, International Steering Committee for Travel Survey Conferences. Annecy, France (2008).
8. Houston-Galveston Area Council of Governments. Draft Summary Report: 2008-09 Regional Household Activity/Travel Survey. ETC Institute, (2009).
9. Abilene Urban Transportation Study. Summary Report: 2010-11 Regional Household Activity/Travel Survey. ETC Institute, (2011a).

10. El Paso Urban Transportation Study. Summary Report: 2010-11 Regional Household Activity/Travel Survey. ETC Institute, (2011b).
11. Wichita Falls Urban Transportation Study. *Summary Report: 2010-11 Regional Household Activity/Travel Survey*. ETC Institute, (2011c).
12. 2010–2012 Minneapolis – St. Paul Travel Behavior Inventory. Twin Cities Metropolitan Council, (2012).
13. 2012–2013 Delaware Valley Household Travel Survey. Delaware Valley Regional Planning Commission, (2013).
14. Mid-Region Council of Governments 2013 Household Travel Survey. Final Report. Westat, Rockville, Md, (2014).
15. 2014 Southern Nevada Household Travel Survey. Final Report. Westat, Rockville, Md, (2015).
16. Chicago Regional Household Travel Inventory. Draft Final Report. NuStats, Austin, Tex., and GeoStats, Atlanta, Ga, (2007).
17. 2011 Atlanta, Georgia, Regional Travel Survey. Final Report. NuStats, Austin, Tex, (2011).
18. 2010--2012 California Household Travel Survey. Final Report Version 1.0. NuStats, Austin, Tex, (2013).
19. INRIX Traffic. http://www.inrix.com/, (2022).
20. Haghani, A, Hamedi, M. and Farokhi Sadabadi, K. I-95 Corridor coalition vehicle probe project: Validation of INRIX data. I-95 Corridor Coalition 9, (2009).
21. Schrank, D., Eisele, B., and Lomax, T.. 2014 Urban mobility report: powered by Inrix Traffic Data (No. SWUTC/15/161302-1), (2015).

22. Cui, Z., Ke, R., Pu, Z., & Wang, Y.. Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. *arXiv preprint* arXiv:1801.02143, (2018).
23. Horak, Ray. *Telecommunications and data communications handbook*. John Wiley & Sons, (2007).
24. Chen, C., Ma, J., Susilo, Y., Liu, Y., and Wang, M.. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*. 68, 285-299, (2016).
25. Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A. L.. Understanding individual human mobility patterns. *Nature*, 453(7196), 779-782, (2008).
26. Kang, C., Liu, Y., Ma, X., and Wu, L.. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, 19(4), 3-21, (2012).
27. Kang, C., Ma, X., Tong, D., and Liu, Y.. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, 391(4), 1702-1717, (2012).
28. Pappalardo, L., F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti and A.-L. Barabási. Returners and Explorers Dichotomy in Human Mobility. *Nature communications*. Vol. 6,pp. 8166. (2015).
29. Song, C., T. Koren, P. Wang and A.-L. Barabási. Modelling the Scaling Properties of Human Mobility. *Nature Physics*. Vol. 6, No. 10, pp. 818. (2010).
30. Song, C., Z. Qu, N. Blumm and A.-L. Barabási. Limits of Predictability in Human Mobility. *Science*. Vol. 327, No. 5968, pp. 1018-102. (2010).

31. Çolak, S., A. Lima and M. C. González. Understanding Congested Travel in Urban Areas. *Nature communications*. Vol. 7, pp. 10793. (2016).
32. Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M. and Puchinger, J. Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101, pp.254-275. (2019).
33. Fekih, M., Bellemans, T., Smoreda, Z., Bonnel, P., Furno, A. and Galland, S.. A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France). *Transportation*, pp.1-32. (2020).
34. Eagle, N., M. Macy and R. Claxton. Network Diversity and Economic Development. *Science* Vol. 328, No. 5981, pp. 1029-1031. (2010).
35. Frias-Martinez, V., J. Virseda, A. Rubio and E. Frias-Martinez. Towards Large Scale Technology Impact Analyses: Automatic Residential Localization from Mobile Phone-Call Data. *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*, ACM. (2010).
36. Soto, V., V. Frias-Martinez, J. Virseda and E. Frias-Martinez. Prediction of Socioeconomic Levels Using Cell Phone Records. *International Conference on User Modeling, Adaptation, and Personalization*, Springer. (2010).
37. Landmark, A.D., Arnesen, P., Södersten, C.J. and Hjelkrem, O.A.. Mobile phone data in transportation research: methods for benchmarking against other data sources. *Transportation*, pp.1-23. (2021).
38. Wang, F., & Chen, C.. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*. 87, 58-74, (2018).

39. Wang, F., Wang, J., Cao, J., Chen, C., and Ban, X. J.. Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. *Transportation Research Part C: Emerging Technologies*. 105, 183-202, (2019).
40. Puget Sound Regional Travel Study. Report: Spring 2014 Household Travel Survey. RSG. (2014).
41. In-The-Moment Travel Study. Revised Report. RSG. (2015a).
42. Puget Sound Regional Travel Study. Report: 2015 Household Travel Survey. RSG, (2015b).
43. 2017 Puget Sound Regional Travel Study. Draft Final Report. RSG, (2017).
44. Airsage. https://www.airsage.com/, (2022).
45. Zhang, L., Darzi, A., Ghader, S., Pack, M.L., Xiong, C., Yang, M., Sun, Q., Kabiri, A. and Hu, S.. Interactive covid-19 mobility impact and social distancing analysis platform. *Transportation Research Record*, p.03611981211043813 (2020).
46. Xiong, C., Hu, S., Yang, M., Luo, W., and Zhang, L.. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences*, 117(44), 27087-27089 (2020a).
47. Xiong, C., Hu, S., Yang, M., Younes, H., Luo, W., Ghader, S. and Zhang, L.. Mobile device location data reveal human mobility response to state-level stay-at-home orders during the COVID-19 pandemic in the USA. *Journal of the Royal Society Interface*, 17(173), p.20200344. (2020b).
48. Huang, H., Cheng, Y. and Weibel, R.. Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, 101, pp.297-312. (2019).

49. Burkhard, O., Becker, H., Weibel, R. and Axhausen, K.W.. On the requirements on spatial accuracy and sampling rate for transport mode detection in view of a shift to passive signalling data. *Transportation Research Part C: Emerging Technologies*, 114, pp.99-117. (2020).
50. Axhausen, K. W., Schönfelder, S., Wolf, J., Oliveira, M., and Samaga, U.. Eighty weeks of GPS-traces: approaches to enriching the trip information. *Presented at 83rd Annual Meeting of the Transportation Research Board*, Washington, D.C., (2003).
51. Stopher, P., FitzGerald, C., and Zhang, J.. Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*. 16(3), 350-369, (2008).
52. Schuessler, N., & Axhausen, K. W.. Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*. 2105(1), 28-36, (2009).
53. Bohte, W., & Maat, K..n Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*. 17(3), pp.285-297, (2009).
54. Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T.. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 2012. 36(2), 131-139, (2012).
55. Gong, L., Morikawa, T., Yamamoto, T., & Sato, H.. Deriving personal trip data from GPS data: A literature review on the existing methodologies. *Procedia-Social and Behavioral Sciences*. 138, 557-565, (2014).
56. Safi, H., Assemi, B., Mesbah, M., Fereira, L., and Hickman, M.. Design and implementation of a smartphone-based system for personal travel survey: Case study from New Zealand.

Transportation Research Record: Journal of the Transportation Research Board. vol. 2526, pp. 99–107, (2015).

57. Patterson, Z., & Fitzsimmons, K.. Datamobile: Smartphone travel survey experiment. *Transportation Research Record: Journal of the Transportation Research Board*. 2594(1), 35-43, (2016).
58. Gong, L., Sato, H., Yamamoto, T., Miwa, T., & Morikawa, T.. Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. *Journal of Modern Transportation*. 23(3), 202-213, (2015).
59. Zhou, C., Jia, H., Juan, Z., Fu, X., & Xiao, G.. A data-driven method for trip ends identification using large-scale smartphone-based GPS tracking data. *IEEE Transactions on Intelligent Transportation Systems*. 18(8), 2096-2110, (2016).
60. Gong, L., Yamamoto, T., & Morikawa, T.. Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines. *Transportation Research Procedia*. 32, 146-154, (2018).
61. Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L.. Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)*. 25(3), 12, (2007).
62. Ye, Y., Zheng, Y., Chen, Y., Feng, J., and Xie, X.. Mining individual life pattern based on location history. *2009 tenth international conference on mobile data management: Systems, services and middleware*. pp. 1-10, (2009).
63. Chen, W., Ji, M., & Wang, J.. T-DBSCAN: A spatiotemporal density clustering for GPS trajectory segmentation. *International Journal of Online Engineering (iJOE).* 10(6), 19-24, (2014).

64. Yao, Z., Zhou, J., Jin, P. J., & Yang, F.. Trip End Identification based on Spatial-Temporal Clustering Algorithm using Smartphone GPS Data (No. 19-01097), *Presented at 98th Annual Meeting of the Transportation Research Board*, Washington, D.C., (2019).
65. Stenneth, Leon, et al. Transportation mode detection using mobile phones and GIS information. *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*. (2011).
66. Brunauer, R., Hufnagl, M., Rehrl, K., & Wagner, A.. Motion pattern analysis enabling accurate travel mode detection from GPS data only. In 16th International *IEEE Conference on Intelligent Transportation Systems (ITSC 2013)* pp. 404-411. IEEE, (2013).
67. Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., & Maurer, P.. Supporting large-scale travel surveys with smartphones–A practical approach. *Transportation Research Part C: Emerging Technologies*, 43, 212-221. (2014).
68. Xiao, G., Juan, Z., and Zhang, C.. Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems* 54: 14-22, (2015).
69. Shafique, M. A., & Hato, E.. Travel mode detection with varying smartphone data collection frequencies. *Sensors*, 16(5), 716, (2016).
70. Wang, B., Gao, L., & Juan, Z.. Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier. *IEEE Transactions on Intelligent Transportation Systems*, 19(5), 1547-1558, (2017).
71. Dabiri, S. and Heaslip, K.. Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation research part C: emerging technologies*, *86*, pp.360-371. (2018).

72. Broach, Joseph, Jennifer Dill, and Nathan Winslow McNeil. Travel mode imputation using GPS and accelerometer data from a multi-day travel survey. *Journal of Transport Geography* 78: 194-204, (2019).
73. Vaughan, J., Imani, A.F., Yusuf, B. and Miller, E.J.. Modelling cellphone trace travel mode with neural networks using transit smartcard and home interview survey data. *European Journal of Transport and Infrastructure Research*, 20(4), pp.269-285 (2020).
74. Breyer, N., Gundlegård, D. and Rydergren, C.. Travel mode classification of intercity trips using cellular network data. *Transportation Research Procedia*, 52, pp.211-218. (2021).
75. Li, Z., Liu, P., Wang, W., and Xu, C.. Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention*, 2012. 45, 478-486.
76. Kim, K., Nitz, L., Richardson, J., and Li, L.. Personal and behavioral predictors of automobile crash and injury severity. *Accident Analysis & Prevention*, 1995. 27(4), 469-481.
77. Haleem, K., and Abdel-Aty, M.. Examining traffic crash injury severity at unsignalized intersections. *Journal of safety research*, 2010. 41(4), 347-357.
78. Xie, Y., Zhang, Y., and Liang, F.. Crash injury severity analysis using Bayesian ordered probit models. *Journal of Transportation Engineering*, 2009. 135(1), 18-25.
79. Scheetz, L. J., Zhang, J., and Kolassa, J. . Classification tree modeling to identify severe and moderate vehicular injuries in young and middle-aged adults. *Artificial intelligence in medicine*, 2009. *45*(1), 1-10.
80. Lerner, E. B., Cushman, J. T., Blatt, A., Lawrence, R. D., Shah, M. N., Swor, R. A., ... and Jurkovich, G. J.. EMS provider assessment of vehicle damage compared with assessment by a professional crash reconstructionist. *Prehospital emergency care*, 2011. 15(4), 483-489.

81. Ryb, G. E., and Dischinger, P. C.. Disparities in trauma center access of older injured motor vehicular crash occupants. *Journal of Trauma and Acute Care Surgery*, 2011. *71*(3), 742-747.

82. Meagher, A. D., Lin, A., Mandell, S. P., Bulger, E., and Newgard, C. . A Comparison of Scoring Systems for Predicting Short-and Long-term Survival After Trauma in Older Adults. *Academic emergency medicine*, 2019.

83. Wang, S. W., Sasser, S. M., and Jurkovich, G. J.. Universal Medical Rescue Protocol Changed:"High Speed Auto Crash" Changed to "High Risk Auto Crash" in the Field Triage Decision Scheme. *Proceedings of the 21st (ESV) International Technical Conference on the Enhanced Safety of Vehicles*, 2009.

84. Sasser, S. M., Hunt, R. C., Faul, M., Sugerman, D., Pearson, W. S., Dulski, T., ... and Cooper, A.. Guidelines for field triage of injured patients: recommendations of the National Expert Panel on Field Triage, 2011. *Morbidity and Mortality Weekly Report: Recommendations and Reports*, 2012. *61*(1), 1-20.

85. Davidson, G. H., Rivara, F. P., Mack, C. D., Kaufman, R., Jurkovich, G. J., and Bulger, E. M. . Validation of prehospital trauma triage criteria for motor vehicle collisions. *Journal of Trauma and Acute Care Surgery*, 2014. *76*(3), 755-761.

86. Miller, R. T., Nazir, N., McDonald, T., and Cannon, C. M. . The modified rapid emergency medicine score: A novel trauma triage tool to predict in-hospital mortality. *Injury*, 2017. *48*(9), 1870-1877.

87. Scheetz, L. J. . Effectiveness of prehospital trauma triage guidelines for the identification of major trauma in elderly motor vehicle crash victims. *Journal of emergency nursing*, 2003. *29*(2), 109-115.

88. Scheetz, L. J., Zhang, J., and Kolassa, J. E.. Using crash scene variables to predict the need for trauma center care in older persons. *Research in nursing & health*, 2007. *30*(4), 399-412.
89. Newgard C.D.. Improving early identification of the high-risk elderly trauma patient by emergency medical services. *Injury*, 2016. 47. 19-25.
90. van der Sluijs, R., Debray, T., Poeze, M., Leenen, L.P. and van Heijl, M.. Development and validation of a novel prediction model to identify patients in need of specialized trauma care during field triage: design and rationale of the GOAT study. Diagnostic and prognostic research, 3(1), pp.1-8 (2019).
91. Atiksawedparit, P., Rattanasiri, S., Sittichanbuncha, Y., McEvoy, M., Suriyawongpaisal, P., Attia, J. and Thakkinstian, A.. Prehospital prediction of severe injury in road traffic injuries: a multicenter cross-sectional study. *Injury*, *50*(9), pp.1499-1506 (2019).
92. van Rein, E.A., van der Sluijs, R., Voskens, F.J., Lansink, K.W., Houwert, R.M., Lichtveld, R.A., de Jongh, M.A., Dijkgraaf, M.G., Champion, H.R., Beeres, F.J. and Leenen, L.P.. Development and validation of a prediction model for prehospital triage of trauma patients. *JAMA surgery*, *154*(5), pp.421-429 (2019).
93. Magnusson, C., Herlitz, J. and Axelsson, C.. Pre-hospital triage performance and emergency medical services nurse's field assessment in an unselected patient population attended to by the emergency medical services: a prospective observational study. *Scandinavian journal of trauma, resuscitation and emergency medicine*, *28*(1), pp.1-15 (2020).
94. Shanahan, T.A., Fuller, G.W., Sheldon, T., Turton, E., Quilty, F.M.A. and Marincowitz, C.. External validation of the Dutch prediction model for prehospital triage of trauma patients in South West region of England, United Kingdom. *Injury*, *52*(5), pp.1108-1116 (2021).

95. Kwon, J., Varaiya, P. and Skabardonis, A.. Estimation of truck traffic volume from single loop detectors with lane-to-lane speed correlation. *Transportation Research Record*, *1856*(1), pp.106-117 (2003).

96. Li, J.Q., Zhou, K., Shladover, S.E. and Skabardonis, A.. Estimating queue length under connected vehicle technology: Using probe vehicle, loop detector, and fused data. *Transportation research record*, *2356*(1), pp.17-22 (2013).

97. Meng, C., Yi, X., Su, L., Gao, J. and Zheng, Y.. City-wide traffic volume inference with loop detector data and taxi trajectories. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 1-10) (2017).

98. Zhao, Y., Zheng, J., Wong, W., Wang, X., Meng, Y., and Liu, H. X.. Various methods for queue length and traffic volume estimation using probe vehicle trajectories. Transportation Research Part C: Emerging Technologies, 107, 70–91 (2019). https://doi.org/10.1016/j.trc.2019.07.008

99. Guo, Q., Li, L., and (Jeff) Ban, X.. Urban traffic signal control with connected and automated vehicles: A survey. In Transportation Research Part C: Emerging Technologies (Vol. 101, pp. 313–334) (2019). Elsevier Ltd. https://doi.org/10.1016/j.trc.2019.01.026

100. Sekuła, P., Marković, N., vander Laan, Z., and Sadabadi, K. F.. Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A Maryland case study. Transportation Research Part C: Emerging Technologies, 97, 147–158 (2018). https://doi.org/10.1016/j.trc.2018.10.012

101. Anuar, K., and Cetin, M.. Estimating Freeway Traffic Volume Using Shockwaves and Probe Vehicle Trajectory Data. Transportation Research Procedia, 22, 183–192 (2017). https://doi.org/10.1016/j.trpro.2017.03.025

102. Li, F., Tang, K., Yao, J., and Li, K.. Real-Time Queue Length Estimation for Signalized Intersections Using Vehicle Trajectory Data. Transportation Research Record, 2623(1), 49–59 (2017). https://doi.org/10.3141/2623-06

103. Caceres, N., Romero, L. M., Benitez, F. G. and del Castillo, J. M.. Traffic Flow Estimation Models Using Cellular Phone Data. *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1430-1441 (2012). doi: 10.1109/TITS.2012.2189006

104. Janecek, A., Valerio, D., Hummel, K. A., Ricciato, F. and Hlavacs, H.. The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring. *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2551-2572, doi: 10.1109/TITS.2015.2413215 (2015).

105. Xing, J., Liu, Z., Wu, C., and Chen, S.. Traffic volume estimation in multimodal urban networks using cell phone location data. *IEEE Intelligent Transportation Systems Magazine*, 11(3), 93-104 (2019).

106. Fan, J., Fu, C., Stewart, K., and Zhang, L.. Using big GPS trajectory data analytics for vehicle miles traveled estimation. *Transportation research part C: emerging technologies* 103 (2019): 298-307.

107. Codjoe, Julius, Grace Ashley, and William Saunders. Evaluating cell phone data for AADT estimation. No. FHWA/LA. 18/591, LTRC Project Number: 16-3SA, State Project Number: DOTLT1000110. Louisiana Transportation Research Center, (2018).

108. *Traffic Safety Facts 2019: A Compilation of Motor Vehicle Crash Data (Report No. DOT HS 813 141)*. National Center for Statistics and Analysis. National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813141. Accessed April 06, 2022 (2019).

109. *2020 Maryland Pedestrian Safety Program Area Brief*. https://zerodeathsmd.gov/wp-content/uploads/2021/05/FFY21_Ped_ProgramAreaBrief_FINAL.pdf. Accessed July 9, 2021 (2020).

110. *2020 Maryland Bicycle Safety Program Area Brief*. https://zerodeathsmd.gov/wp-content/uploads/2021/05/FFY21_Bicycle_ProgramAreaBrief_FINAL.pdf. Accessed July 9, 2021 (2020).

111. Hosseinpour, M.H., Prasetijo, J., Yahaya, A.S., and Ghadiri, S.M.R. Modeling Vehicle-Pedestrian Crashes with Excess Zero along Malaysia Federal Roads. *Procedia-Social and Behavioral Sciences*, 53, pp. 1216–1225 (2012).

112. Strauss, J., Miranda-Moreno, L.F., and Morency, P. Mapping Cyclist Activity and Injury Risk in a Network Combining Smartphone GPS Data and Bicycle Counts. *Accident Analysis & Prevention*, 83, pp.132–142 (2015).

113. Xie, K., Ozbay, K., Kurkcu, A., and Yang, H. Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots. *Risk Analysis*, 37(8), pp.1459–1476 (2017).

114. *Guidebook on Identification of High Pedestrian Crash Locations.* FHWA-HRT-17-106. FHWA, U.S. Department of Transportation, (2018).

115. Saad, M., Abdel-Aty, M., Lee, J., and Cai, Q. Bicycle Safety Analysis at Intersections from Crowdsourced Data. *Transportation Research Record*, 2673(4), pp.1–14 (2019).

116. Raihan, M.A., Alluri, P., Wu, W., and Gan, A. Estimation of Bicycle Crash Modification Factors (CMFs) on Urban Facilities Using Zero Inflated Negative Binomial Models. *Accident Analysis & Prevention*, 123, pp.303–313 (2019).

117. Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., and Isa-Tavarez, J. The Role of Built Environment on Pedestrian Crash Frequency. *Safety Science*, 50(4), pp.1141–1151 (2012).

118. Jiang, X., Abdel-Aty, M., Hu, J., and Lee, J. Investigating Macro-level Hotzone Identification and Variable Importance Using Big Data: A Random Forest Models Approach. *Neurocomputing*, 181, pp.53–63 (2016).

119. Sanders, R.L., Frackelton, A., Gardner, S., Schneider, R., Hintze, and M. Ballpark. Method for Estimating Pedestrian and Bicyclist Exposure in Seattle, Washington: Potential Option for Resource-constrained Cities in an Age of Big Data. *Transportation Research Record*, 2605(1), pp.32–44 (2017).

120. Mansfield, T.J., Peck, D., Morgan, D., McCann, B., and Teicher, P. The Effects of Roadway and Built Environment Characteristics on Pedestrian Fatality Risk: A National Assessment at the Neighborhood Scale. *Accident Analysis & Prevention*, 121, pp.166–176 (2018).

121. Lee, J., Abdel-Aty, M., Choi, K., and Huang, H. Multi-level Hot Zone Identification for Pedestrian Safety. *Accident Analysis & Prevention*, 76, pp.64–73 (2015).

122. Lord, D., and Mannering, F. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), pp.291–305 (2010).

123. https://www.strava.com/

124. Jestico, B., Nelson, T.A., Potter, J., and Winters, M. Multiuse Trail Intersection Safety Analysis: A Crowdsourced Data Perspective. *Accident Analysis & Prevention*, 103, pp.65–71 (2017).

125. Lee, J., Abdel-Aty, and M., Shah, I. Evaluation of Surrogate Measures for Pedestrian Trips at Intersections and Crash Modeling. *Accident Analysis & Prevention*, 130, pp.91–98 (2019).

126. Wang HE, Mann, NC., Jacobson, and KE. Et al. National characteristics of emergency medical services responses in the United States. Prehospital Emergency Care 2013. 17(1): 8-14 (2013).

127. Albert M. Emergency department visits by persons aged 65 and over: United States, 2009-2010. NCHS data brief, No. 130. Hyattsville, MD: National Center for Health Statistics; (2013).

128. Mafi, S., AbdelRazig, Y., Amirinia, G., Kocatepe, A., Ulak, M. B., and Ozguven, E. E.. Investigating exposure of the population to crash injury using a spatiotemporal analysis: A case study in Florida. *Applied geography*, 104, 42-55 (2019).

129. Khreis, H., Warsow, K. M., Verlinghieri, E., Guzman, A., Pellecuer, L., Ferreira, A., ... and Schepers, P.. The health impacts of traffic-related exposures in urban areas: Understanding real effects, underlying driving forces and co-producing future directions. *Journal of Transport & Health*. 3(3), 249-267 (2016).

130. Nakamura Y., Daya, M., Bulger EM., and Schreiber, M., et al.. Evaluating age in the field triage of injured persons. *Ann Emerg Med*. 60(3): 335-345 (2012).

131. Chang D.C., Bass R.R., Cornwell, E.E., and Mackenzie, E. Undertriage of elderly trauma patients to state-designated trauma centers. *Arch Surg*. 143(8) 776-781 (2008).

132. Lehmann, R., Beekley, A., Casey, L., Salim A., and Martin M. The impact of advanced age on trauma triage decisions and outcomes: a statewide analysis. *AM J Surg*. 197(5): 571-574 (2009).

133. Cox S., Morrison C., Cameron P., S and mith, K. Advancing age and trauma: triage destination compliance and mortality in Victoria, Australia. *Injury*. 45(9): 1312-1319 (2014).

134. Xiang H., Wheeler, K., Groner J., Shi, J., and Haley, K. Under-triage of major trauma patients in the US emergency departments. *American Journal of Emergency Medicare*. 32(9): 997-1004 (2014).

135. Scheetz, L. J., Zhang, J., Kolassa, J. E., Allen, P., and Allen, M.. Evaluation of injury databases as a preliminary step to developing a triage decision rule. *Journal of nursing scholarship*. 40(2), 144-150 (2008).

136. Zhang, X., Hamedi, M. and Haghani, A.. Arterial travel time validation and augmentation with two independent data sources. *Transportation Research Record*, *2526*(1), pp.79-89 (2015).

137. Steinberg, D. and Colla, P.. CART: tree-structured non-parametric data analysis. *San Diego, CA: Salford Systems* (1995).

138. Menardi, G. and Torelli, N.. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, *28*(1), pp.92-122 (2014).

139. Bishop, C.M. and Nasrabadi, N.M.. *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer (2006).

140. Micci-Barreca, D.. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, *3*(1), pp.27-32 (2001).

141. Witten, I.H. and Frank, E.. Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), pp.76-77 (2002).

142. Hu, S., Xiong, C., Yang, M., Younes, H., Luo, W. and Zhang, L., 2021. A big-data driven approach to analyzing and modeling human mobility trend under non-pharmaceutical

interventions during COVID-19 pandemic. *Transportation Research Part C: Emerging Technologies, 124*, p.102955.

143. Fan, J., Fu, C., Stewart, K., and Zhang, L.. Using big GPS trajectory data analytics for vehicle miles traveled estimation. *Transportation research part C: emerging technologies* 103 (2019): 298-307.

144. Zhang, L., Ghader, S., Darzi, A., Pan, Y., Yang, M., Sun, Q., Kabiri, A. and Zhao, G.. Data analytics and modeling methods for tracking and predicting origin-destination travel trends based on mobile device data. *Federal Highway Administration Exploratory Advanced Research Program* (2020).

145. Newson, P. and Krumm, J.. Hidden Markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 336-343) (2009).

146. Washington, S.P., Karlaftis, M.G., Mannering, F. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC. (2003).

147. *Smart Location Mapping*. United States Environmental Protection Agency. https://www.epa.gov/smartgrowth/smart-location-mapping#SLD. Accessed July 10, 2021.

148. *Smart Location Database Version 2.0 User Guide*. https://www.epa.gov/sites/production/files/2014-03/documents/sld_userguide.pdf. Accessed July 10, 2021.

149. A Data-Driven Safety Dashboard Assessing Maryland Statewide Density Exposure of Pedestrians, Bicycles, and E-Scooters. (2021). https://rosap.ntl.bts.gov/view/dot/61320

150. *Low-stress Bicycling and Network Connectivity*. Mineta Transportation Institute Publications.

https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1073&context=mti_publications. Accessed March 17, 2022.

151. Furth, P.G. *Level of Traffic Stress*. https://peterfurth.sites.northeastern.edu/2014/05/21/criteria-for-level-of-traffic-stress. Accessed March 17, 2022.

152. *National Transit Map*. Bureau of Transportation Statistics. U.S. Department of Transportation. https://www.bts.gov/content/national-transit-map. Accessed July 9, 2021.

153. *Maryland Statewide Vehicle Crashes*. https://opendata.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crashes/65du-s3qu. Accessed July 10, 2021.

154. Fabozzi, F.J., Focardi, S.M., Rachev, S.T., and Arshanapalli, B.G. *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*. John Wiley & Sons. 2014.

155. Greene, W.H. *Econometric Analysis Fifth Edition.* Prentice Hall. 2003.

156. Tiwari, G., Bangdiwala, S., Saraswat, A., and Gaurav, S. Survival Analysis: Pedestrian Risk Exposure at Signalized Intersections. *Transportation Research Part F: Traffic Psychology and Behaviour*, 10(2), pp.77–89 (2007).

157. Lee, K., and Sener, I.N. *Emerging Data Mining for Pedestrian and Bicyclist Monitoring: A Literature Review Report*. Safety through Disruption, National University Transportation Center (UTC) Program. (2017). https://safed.vtti.vt.edu/wp-content/uploads/2020/07/UTC-Safe-D_Emerging-Data-Mining-for-PedBike_TTI-Report_26Sep17_final.pdf.

158. *Synthesis of Methods of Estimating Pedestrian and Bicyclist Exposure to Risk at Areawide Levels and on Specific Transportation Facilities*. FHWA-SA-17-041. FHWA, U.S. Department of Transportation, (2017).