The Thesis Committee for Amanda Marie Deering
Certifies that this is the approved version of the following thesis:

A Framework for Processing Connected Vehicle Data in Transportation Planning Applications

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Chandra Bhat

Jennifer Duthie

A Framework for Processing Connected Vehicle Data in Transportation Planning Applications

by

Amanda Marie Deering, B.S.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

December 2016

Acknowledgements

I would like to thank Natalia Ruiz-Juri for all her assistance, suggestions and support while I was working on this research study. Thanks to Weijia Xu and Amit Gupta at TACC for their technical assistance on computing methods.

I am grateful to my family for their unfailing support and encouragement throughout my graduate career, and especially while writing this thesis.

Finally, I am grateful to Dr. Jen Duthie for her constructive comments on this thesis, Dr. Chandra Bhat for his advice on all my work, and to them both for the opportunities I received at CTR during my graduate studies.

Abstract

A Framework for Processing Connected Vehicle Data in Transportation Planning Applications

Amanda Marie Deering, M.S.E.

The University of Texas at Austin, 2016

Supervisor: Chandra Bhat

This thesis presents a framework to process connected vehicle data into a format that is practical for implementation in the transportation planning field. Whereas prior research on connected vehicles has used theoretical models or small data samples for analysis, this study uses the largest public connected vehicle dataset currently available – the Sample Data Environment from the Safety Pilot Model Deployment project out of Ann Arbor, Michigan. This data includes basic safety messages and driving data for 2800 vehicles over two months. An algorithm to process basic safety message data into a trip level dataset is presented. This thesis also includes a process for spatial aggregation of trips into origin and destination zones using a hexagonal grid. These processes are implemented through the combination of a variety of open-source tools including Hadoop and PostgreSQL. Excerpts from the processed data are provided as well as sample analysis applications for the trip and spatial data. Recommendations and guidance are provided on handling the details of such an immense dataset. Since similar future vehicle-

to-vehicle communications datasets are likely, it is imperative to develop methods to process and analyze this rich data effectively.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Connected vehicle technology has been around for over a decade but it is still not familiar to most American consumers. Part of connected vehicle (CV) technology is vehicle-to-vehicle (V2V) communication, which consists of transmitting and receiving basic safety messages (BSM) over a dedicated short range communications (DSRC) wavelength. These BSMs carry safety information such as vehicle location, speed, acceleration, and time. Extended V2V technologies also have the capacity to monitor and report brake system status, headlight usage, windshield wiper usage, lane detection and more. While crash avoidance is a primary benefit of such technologies, they also provide avenues for improvement in other areas of transportation. Real time data can be used to improve traffic operations, decrease crash frequency, increase mobility, and mitigate environmental impacts (ITS JPO, 2016). This thesis provides a framework for processing and aggregating large datasets produced by this CV technology, facilitating their use in the aforementioned transportation planning applications.

Connected vehicles produce a large amount of detailed data that can be used on the travel demand and planning side of transportation applications as well. Planners and transportation engineers have long been searching for accurate and detailed location-based travel data with which to create, validate, and improve their models. Up until now the primary data sources used for travel demand modeling were household travel surveys, active transportation studies, and passive data sources, such as Bluetooth roadway sensors or GPS probe data (Bohte and Maat, 2009; Carpenter et al., 2012). Passive data sources have become increasingly popular due to their low cost and increased levels of detail. They still have shortcomings though in terms of coverage area, long term travel implications, and travel behavior. Detailed location and vehicle data from connected

vehicles may be able to close this gap. In academia, much research has been done anticipating a shift in the modeling paradigm due to the advent of connected and autonomous vehicles (Bagheri et al., 2015; Li et al., 2013). Since vehicles with V2V communication technology aren't easily accessible yet, most of this research has been conducted using theoretical models. While this is useful to anticipate general changes, it doesn't explain the practical details of how this data source can be used by the average transportation management center, or what aspects are valuable for transportation studies. Recently a large real-life dataset on 2800 connected vehicles with BSMs and other recorded driving data was released as a part of the Safety Pilot Model Deployment (SPMD) project in Ann Arbor, Michigan. This detailed dataset offers a plethora of new research opportunities using real-life travel data from fully connected passenger vehicles, buses, and freight trucks.

This thesis explores the potential applications of vehicle and location data produced by connected vehicles in the planning field using actual data produced in the SPMD project. A framework to process the initial data into a useable format for origin-destination (OD) analysis, trip-level evaluation, and system performance measures is presented. Additionally, an extensive comparison of CV data to other commonly used passive data sources is provided. Challenges and lessons learned from investigating and evaluating such a large, expansive dataset are also discussed in detail.

Chapter 2: Literature Review

Many disciplines are involved in the area of connected vehicle research. From electrical and mechanical engineers, to computer scientists, to urban planners, the coordination of multiple groups is vital to the exploration of CV applications. In the transportation realm, the data that are produced by connected vehicles could prove to be incredibly valuable, but this source is not without its own challenges. Location data from common sources such as mobile phones, Bluetooth sensors, or probe vehicles can be used to improve model inputs, validate model results, and hone network assumptions. The high level of detail of data on location, vehicle status, and acceleration offered by connected vehicle technology is poised to become an invaluable source of information in the transportation field. This section reviews the travel demand modeling process, transportation performance measures, commonly used passive data sources, connected vehicle technology, U.S. policy, and related research.

2.1 Travel Demand Modeling and Performance Measures

Planning agencies can use historical connected vehicle data to improve the entire planning process. This includes travel demand modeling, congestion management, system operations, freight movement, and safety (USDOT et al., 2007). Though there are many potential applications for such information, this research will focus on travel demand modeling and system operations.

Traditionally, planning agencies follow a four-step process for travel demand estimation. The four steps are trip generation, trip distribution, mode choice, and route assignment (USDOT et al., 2007). Historically, planning agencies have used travel surveys of a population sample to construct trip tables for step one (Stopher and Greaves, 2007). Identifying trip purpose from these surveys is instrumental in developing origin-

destination matrices for step two (Oliveira et al., 2014). These are considered to be traffic assignment model inputs. Mode choice can be determined from personal GPS units or survey responses (Bohte and Maat, 2009). Route assignment, the fourth step, is the model output which can be validated against historical data using performance measures such as travel time or volume counts. Location and trip information from connected vehicles has the potential to be used as inputs in steps one and two of the modeling process. Additionally, travel time and speed data can be used to more accurately validate step four of the modeling process. Analysis of travel behaviors such as route choice, driving patterns, or the presence of trip chaining are a way to verify the assumptions made in the planning process (Dhakar and Srinivasan, 2014; Hunter et al., 2006; Ma et al., 2016). CV data records are uniquely equipped for specific user behavior studies due to the ability to track trip-making behavior over a longer period of time. However, since the current scope of connectivity is limited to trucks and personal vehicles, there is no additional contribution to the third step of mode choice.

An alternative to trip-based modeling, as described above, is activity-based modeling (ABM). This approach focuses on the demand side of predicting travel patterns by using tours (i.e. a sequence of trips) rather than individual trips. ABM provides a framework to improve how temporal and behavioral aspects of individual travel patterns are generated (Lin et al., 2008). Vovsha et al. (2011) gives an overview of the design features of ABM and Bhat et al. (2013) discusses how ABM models the interactions between households and how they influence travel patterns. Connected vehicle data has the potential to be of use in studying travel patterns at a household level due to its level of detail. Additionally, chained trips are easily captured in the way location and trip data are reported in V2V communications.

Monitoring system performance of specific corridors is also a crucial part of the planning process. The most common metrics used to manage arterial and freeway systems are travel speed, volume counts, and travel time. These measures are a good proxy for congestion and route choice, which influence demand. However, the accuracy and frequency of these measures often depend on the information source. There has been a shift towards using passive data sources for performance measures, where transportation professionals don't manually collect data, but rather focus on extracting meaning from automatically recorded data, making collection easier (Carpenter et al., 2012). A review of the many passive data sources used is provided in the next section, but these performance measures can be extracted from CV data as well.

2.2 PASSIVE DATA SOURCES

The most common passive data sources for transportation applications are mobile phone data, Bluetooth sensor data, GPS probe vehicle data, Automatic Vehicle Location (AVL) data, and other roadside sensor data. Each source offers different benefits in cost, ease of implementation, and data processing effort. The associated challenges and advantages for each source are discussed in the following sections.

2.2.1 Mobile Phone Data

Mobile phone data in the form of call data records (CDR) or GPS trajectories have become a common method of passively collecting trip information (Toole et al., 2015). There is an abundance of literature on using phone trajectories to construct trip tables and OD matrices for metropolitan areas (Herrera et al., 2010; Nitsche et al., 2014; Wang et al., 2013). Typically, GPS location data points available at aggregation levels ranging from three seconds to five minutes are pieced together using statistical methods to form origin-destination matrices that are used in the modeling process (Herrera et al.,

2010). An alternate method uses CDR data to form trip tables by identifying trip purposes and then aggregating at the TAZ level (Çolak et al., 2015). Mobile phone datasets have the advantages of high samples sizes and easy obtainability. However, it is not always easy to follow user behavior over time. Many studies don't keep track of which calls belong to which device for a period longer than 24 hours, usually for privacy reasons (Schlaich, 2010). Another potential weakness in cell phone data is that due to the nature of its large sample size and ties to industry, data are often aggregated and preprocessed by a third party company (e.g., AirSage) before they are given to researchers (Huntsinger and Ward, 2015). This limits the amount of control researchers have regarding the level of detail they can reproduce in transportations studies. Additionally, GPS locations from cell phone data are not always precise, depending on how close the nearest cell towers are located (Herrera et al., 2010). Overall, mobile phone data sources are beneficial when large sample sizes are desired and most useful for creation of aggregated OD matrices. Mobile phone data are not well-suited for corridor studies, system performance or speed studies.

2.2.2 Bluetooth Sensor Data

Bluetooth sensor data are similar to CDR in that they make use of the abundance of mobile phones, but the data collected are more targeted in scope and are commonly used for travel time analysis (Carpenter et al., 2012). Bluetooth sensors are placed along a corridor which detect Bluetooth-enabled devices and record the timestamp and device ID. Grone et al. (2011) covers the practical aspect of sensor placement for multiple sensor types under varying conditions. Although Bluetooth is the main type of sensor discussed in this review, many other non-intrusive infrastructure sensors can be applied to produce similar results (Grone et al., 2011). Information acquired from these sensors can be

pieced together to obtain corridor travel times, speeds, and sometimes OD tables, depending on the data size (Carpenter et al., 2012). Bluetooth sensors can pick up multiple modes and often record only part of a vehicle's trip through the corridor. Thus it is necessary to perform statistical cleaning before attempting to analyze the data (Kieu et al., 2012). Advantages of this data collection method include the ability to narrow the scope of a transportation study, and low costs, as mentioned before. It is easy to set up sensors along a corridor to assess traffic conditions there, but it is less simple to canvas a larger region, such as an entire city or metropolitan area. This limited scope and the need for a larger sample size in order to validate OD matrices are challenges faced by this method of transportation data collection.

2.2.3 Probe Vehicles and Automatic Vehicle Location

An alternative to the previous methods is removing the driver variability aspect by using probe vehicles. These vehicles are equipped with GPS technology and driven by researchers through corridors of interest in order to record location information and travel times. This data collection method is best suited for studying traffic signal coordination, corridor travel time, and origin-destination patterns (Remias et al., 2013). Probe data often don't cover a large travel area and are best for arterial analysis. When handling the data, fewer assumptions need to be made about the vehicle's trip origin and destination as these are directly known. Even though the data is clear and easy to work with, it is laborious and expensive to repeat this technique on many corridors. Instead, transit buses can be used as proxies for probe vehicles when they are equipped with Automatic Vehicle Location (AVL) technology. AVL units track the location, time, and stops a bus makes on its route (Cathey and Dailey, 2003). This data source is very similar to general probe vehicle data but it has additional transit specific data. Often statistical cleaning is

needed to identify a buffer around the study area and interpolate the data to reflect the more disaggregate data points needed for arterial studies (Coifman and Kim, 2009). It has been used with some success to estimate corridor travel times and speeds, but these are often underestimations due to the frequent stops buses make (Cathey and Dailey, 2003). Additionally, the data points are often reported at large intervals, such as five minutes, which don't always match up with transit stop locations (Coifman and Kim, 2009). AVL datasets have also been used to track paths taken by riders throughout the day (Munizaga and Palma, 2012). On the whole, passenger vehicle and transit probes are better suited for specific corridor or transit studies. Their travel demand planning applications are limited due to frequency and accuracy of reported data.

2.2.4 Global Positioning System Devices

Both of the above probe methods are smaller in scope than a full GPS instrumented study. GPS studies regularly involve outfitting a large sample of personal vehicles with a device that records GPS location and timestamp at a specified interval, usually five seconds (Glick et al., 2015). GPS travel surveys also make use of handheld GPS devices that participants carry with them on all trips made throughout the day (Bohte and Maat, 2009). These personal units have the benefit of being able to capture trips made by multiple modes. Such studies are becoming more common due to the rather simple method of collection which results in a large amount of detailed data. Previous research has used GPS data to create trip tables, estimate travel times, and analyze route choice (Dhakar and Srinivasan, 2014; Hunter et al., 2006; Oliveira et al., 2014). Such outfitted vehicles are also commonly used in freight logistics to improve route choice and monitor performance (Liao, 2014; Ma et al., 2016). The high level of detail offered by this GPS data may be preferred to self-reported data often used in planning travel surveys

due to increased accuracy and decreased cost (Stopher and Greaves, 2007). Another advantage of this method of data collection is its ability to be used at all scales of the transportation planning process. Additionally, all GPS datasets have similar structure no matter the scope of the experiment, which leads to useful, reproducible processing methodologies.

2.3 CONNECTED VEHICLE DATA

The review of the above methods of data collection in transportation suggests that there are some weaknesses that could be strengthened with the addition of an alternate source. The gaps that currently exist in popular collection methods include data aggregation level, survey area size, and ability to track travel patterns. Messages transmitted by connected vehicles have the potential to fill this void in transportation research. Following is a review of connected vehicle technology, policy and practice, previous related research, and potential applications in transportation planning.

2.3.1 Connected Vehicle Technology

Connected vehicle technology allows vehicles to 'communicate' with one another by transmitting messages to other vehicles and to the surrounding infrastructure using dedicated short-range communications (DSRC) (ITS JPO, 2014). Such communications occur on the band of 5.9 GHz spectrum that was set aside by the FCC for ITS use (NHTSA, 2014). These messages can generally be sent and received at about a maximum distance of 300 meters and contain information on speed, location, acceleration, and vehicle status (Argote-Cabañero et al., 2015). Vehicle-to-vehicle communication is when these messages are transmitted to vehicles, while vehicle-to-infrastructure (V2I) communication involves transmitting messages to the traffic signals, roadside devices, and other transportation infrastructure. As far as V2V goes, there are a few different types

of devices with levels of connectivity that can be installed in a vehicle during manufacture or in an aftermarket scenario. The specifics on these units are discussed in the section on data background. However some of the capabilities of V2V include intersection movement assist, left turn assist, and emergency electronic brake light (NHTSA, 2014). These are just a few current safety applications that are a part of a large system goal of crash avoidance through vehicle connectivity. Connected vehicles do have their limits, though, and should not be confused with autonomous vehicles. Unlike fully autonomous vehicles, connected vehicles will not stop a vehicle or take over control from a driver. However, connected vehicle communications will continue provide value to the transportation planning process even with the advent of autonomous vehicles in the future.

2.3.2 Policy and Practice in the United States

The U.S. Department of Transportation and the National Highway Traffic Safety Administration (NHTSA) have been heavily involved in rulemaking regarding the advent of connected and autonomous vehicles. In August of 2014 NHTSA released an advanced notice of proposed rulemaking (NPRM) that detailed findings on feasibility, privacy, and security to support eventual regulations on connected vehicles (ITS JPO, 2015). The NPRM was presented to the Office of Management and Budget (OMB) in January 2016 and is still under review as of June (Anderson, 2016). It is projected that V2V equipment will be required in all new vehicles, with phase-in starting in 2019 (ITS JPO, 2015). It is estimated that this mandate would initially add $350 per vehicle in 2020, which would decrease to closer to $200 over time (NHTSA, 2014). The benefits of this technology are projected to be considerable. V2V and V2I safety applications have the potential to prevent 80% of non-drunk driving related crashes (NHTSA, 2014). For all the benefits,

privacy protection and security remain significant concerns that need to be addressed by policy and regulation. The NPRM is anticipated to have provisions on the layers of security measures that will be required in V2V systems. All devices will need to have valid certificates so that messages between vehicles can be trusted (NHTSA, 2014). It is also expected that the private sector will be in charge of managing the security side of CV communications (ITS JPO, 2015). V2V systems are intentionally designed to not be equipped to collect and share any personal information about vehicles. Additional devices would have to be installed, and permission granted, to conduct any related transportation research.

2.3.3 Transportation Planning Applications

When recorded, information transmitted by connected vehicles has high potential for a variety of transportation measurements including travel time estimation, average speeds, turning percentages, trip tables, and queue length (Argote-Cabañero et al., 2015; Li et al., 2013). The latter two attributes are not easily ascertained from traditional passive data sources. Often such system measurements require active data collection efforts that can only cover a narrow study area. Historical data that mimics naturalistic driving studies can also provide insights into crash or near crash incidents (Liu and Khattak, 2016). Liu and Khattak use such information to assess "instantaneous driving decisions… and identify critical events" in drivers' trips. This kind of detailed data on driver behavior has not been previously available at this level of sampling. Additionally, there are a multitude of real time CV data applications including dynamic signal timing, congestion monitoring, and safety applications (Bagheri et al., 2015; Talebpour et al., 2014). Advantages of CV system information include its large scale survey potential, high level of detail, and ease of collection. Unlike Bluetooth sensors and probe vehicles,

CVs can collect detailed system performance measures on an entire urban transportation system. Unlike GPS units and mobile phone data, CV records can be used to analyze specific travel behavior and driving actions. However, the biggest challenge of this new source is how to process and validate such an expansive amount of data to make it useful for the transportation planner and practitioner.

Chapter 3: Data Background

Most previous studies of connected vehicle technology and its potential benefits have been undertaken using network simulations or very small sample sizes. The first large scale data collection of connected vehicle data is the Safety Pilot Model Deployment (SPMD) project. It provides invaluable information regarding the practical methods needed to deploy a connected vehicle fleet for research purposes while informing future policy decisions (Bezzina and Sayer, 2015). The Safety Pilot Model Deployment project was conducted by The University of Michigan Transportation Research Institute (UMTRI) in partnership with the National Highway Traffic Safety Administration (NHTSA) under the auspices of the United States Department of Transportation (USDOT) (Bezzina and Sayer, 2015). Data were collected from the testbed in Ann Arbor, Michigan from 2012 to 2013 from over 2800 vehicles, two months of which are available on the USDOT Research Data Exchange website (Research Data Exchange, 2016). This project sought to advance V2V technology and examine vehicle-to-infrastructure operations, as well as safety applications and security implementations (Bezzina and Sayer, 2015). The following sections detail the necessary equipment, data acquisition process, and the resulting datasets produced by the project.

3.1 Equipment and Deployment

The breakdown of the technology that the vehicles were equipped with during the study is as follows:

- 2429 vehicles were equipped with Vehicle Awareness Devices (VADs), which only transmit basic safety messages. They do not receive messages or communicate with the driver.

- 294 vehicles were equipped with Aftermarket Safety Devices (ASD) which transmit and receive BSMs. They relay safety messages to the driver regarding curve speed, emergency electronic braking, and potential forward collisions.
- 16 trucks and 3 buses were equipped with Retrofit Safety Devices (RSD) which perform very similarly to ASDs but are designed for freight and transit operations.
- 64 cars and 3 trucks were equipped with Integrated Safety Devices (ISD) which have full integration with connected vehicle technology and safety applications as well as both sending and receiving BSMs (Bezzina and Sayer, 2015).

All vehicles were also equipped with data loggers to record all BSM transmissions. In addition, data acquisition systems (DASs) of two main types (described further in Section 3.3) were equipped in a portion of the vehicles to record more specific driving behavior data. DAS1 was developed by UMTRI and 118 of these devices were installed on ASD or RSD vehicles. 64 DAS2 devices, developed by the Virginia Tech Transportation Institute (VTTI), were installed on the integrated light vehicles (ILV) with ISDs. The study also included 75 miles of instrumented roadway for communication to the connected vehicles (V2I) (Henclewood and Rajiwade, 2015).

3.2 Data Acquisition Methods

Participants for this study were recruited from the entire community. UMTRI recruited parents from the area school system, as well as employees and students from the University of Michigan Health System and the College of Engineering. A diverse sample of participants was desired in order to get the most natural distribution of vehicle interactions possible. During the data collection period of the study, the health of the collection and transmission devices was frequently monitored for errors. Every time a

vehicle passed a roadside device, the device would record the transmitted BSM. This log was checked periodically by UMTRI for absent vehicles or other reporting issues.

Data acquisition fell into four categories: driving, message, contextual, and subjective. The focus of this research is on the driving data (DAS) and message data (BSMs), but it should be noted that other information such as weather conditions and participant surveys were considered. BSMs were transmitted at 10Hz or 100ms frequencies and were composed of two parts as defined by the SAE J2735 standard. Part one is the main part of the message, which includes data on time, GPS location, speed, acceleration, yaw rate, and associated accuracy measurements. Part two provides supplementary information on vehicle system statuses and safety extensions that flag events such as hard braking, anti-lock braking system engagement, and stability control activation. The archive of the BSM data is extensive, containing over 69 billion records detailing close to four million trips covering 25 million miles (Bezzina and Sayer, 2015). The naturalistic driving data that was captured by the DASs is quite expansive as well. The DAS device in each vehicle records time and location data similar to the BSM but it also records events such as lane detection, changing lanes, wiper activation, headlight activation, number of braking applications, and more. Overall the two DAS archives contain 15 billion rows of data covering 221,000 trips and 1.7 million miles (Bezzina and Sayer, 2015).

3.3 RESULTING DATASETS

The data that was archived during the study went through some preprocessing and validation by a third party consultant before its release. Trip IDs were assigned to the raw data and trip summary files were created which catalogued each trip along with a selection of summary statistics such as average speed, total travel distance, and trip

duration. In addition, specific location data and other information may have been altered in order to remove personally identifiable information (PII). Due to the immense size of the data, only a selection of it has been publicly released for research purposes on the USDOT Research Data Exchange website. The available datasets cover the time period of October 1, 2012 to October 31, 2012 and April 1, 2013 to April 30, 2013.

For this study, BSM data was used for trip-level processing while DAS2 data was used for spatial aggregation at the origin-destination zone level. A sample of the disaggregate BSM data table is available in the Appendix. The full data table is 98.7 GB in size. A description of the column characteristics can be seen below in Table 3.1.

Table 3.1 Basic Safety Message Data Column Characteristics

Field Name	**Type**	**Units**	**Description**
RxDevice	Integer	None	ID number of the device that logs a BSM
FileID	Integer	None	Reference number to locate the source of the data in its original file
TxDevice	Integer	None	ID number of the device that transmits a BSM
Gentime	Integer	Microseconds	A more secure form of Epoch time, measuring the number of microseconds elapsed since midnight, January 1, 2004
TxRandom	Integer	None	Randomly assigned ID to mask the device ID of the transmitting device for security purposes
MsgCount	Integer	None	Message ID that gets incremented by one with each BSM
DSecond	Integer	Deciseconds	Time in deciseconds since ignition started
Latitude	Float	Degrees	Current latitude of the vehicle
Longitude	Float	Degrees	Current longitude of the vehicle
Elevation	Float	Meters	Current elevation of vehicle according to GPS
Speed	Real	m/sec	Vehicle speed
Heading	Real	Degrees	Vehicle heading/direction
Ax	Real	m/sec^2	Longitudinal acceleration
Ay	Real	m/sec^2	Lateral acceleration
Az	Real	m/sec^2	“Vertical” acceleration
Yawrate	Real	deg/sec	Vehicle yaw rate
PathCount	Integer	None	Number, between 1 and 23, representing a group of points that communicate a vehicle’s position and motion. Each group of points is of non-uniform size.

Table 3.1 (Continued)

RadiusOfCurve	Float	Centimeters	Estimate of the radius of a curve being negotiated, which is derived from a number of systems and sensors. Positive and negative values reflect right and left turns, respectively, and +/- 32767 for straight paths.
Confidence	Integer	Percent	Signals the accuracy and non-steady state and steady state of curvature estimate. In steady state (straight roadways or curves with constant radius of curvature), a high confidence value is reported.

Note that the time is given in 'gentime' which reports the number of microseconds elapsed since January 1, 2004 UTC, so this will need to be converted to a timestamp that is more familiar to the average viewer. Another column of interest is the file id which corresponds with each trip for a vehicle, designated by 'rxdevice'.

Two initial data tables of driving data were used in the spatial aggregation process: a trip-level summary file and a driving log dataset with the recorded GPS driving information. DAS2 was chosen over DAS1 for this study because it corresponded to the 64 vehicles with ISDs that were fully connected. The disaggregate driving data tables provide entries for each vehicle and trip at every tenth of a second throughout the trip. Each entry includes a timestamp, GPS latitude and longitude, and various system statuses described in Table 3.2.

Table 3.2 DAS2 Message Data Column Characteristics

Field Name	**Type**	**Units**	**Description**
DeviceID	Integer	None	A unique numeric ID assigned to each DAS
Trip	Integer	None	Count of ignition cycles—each ignition cycle commences when the ignition is in the on position and ends when it is in the off position
Time	Integer	Centiseconds	Time in centiseconds since DAS started, which starts when the ignition is in the on position
GPS Elevation	Float	Meters	Elevation of vehicle according to GPS
GPS Fix Quality	Integer	None	Quality of GPS information
GPS Hdop	Float	None	Horizontal Dilution of Precision, used to determine position accuracy

Table 3.2 (Continued)

GPS Heading	Float	Degrees	Heading of vehicle according to GPS
GPS Latitude	Float	Degrees	Latitude of vehicle according to GPS
GPS Longitude	Float	Degrees	Longitude of vehicle according to GPS
GPS Number Satellites	Integer	None	Number of satellites used in GPS solution
GPS Pdop	Float	None	Positional Dilution of Precision, used to determine position accuracy
GPS Speed	Float	Meters/second	Speed of vehicle according to GPS
GPS UTC Time	Integer	Milliseconds	UTC Time of vehicle according to GPS
GPS Valid	Integer	None	Validity of GPS data
DAS Pitch Rate	Float	Degrees/second	Vehicle angular velocity around the lateral axis
DAS Roll Rate	Float	Degrees/second	Vehicle angular velocity around the longitudinal axis
InVehicle ABS State	Character	None	Provides ABS state of the vehicle
InVehicle Brake Status	Character	None	Provides brake status of the vehicle
InVehicle Headlight Status	Integer	None	Provides status if headlights are currently in use
InVehicle Longitudinal Accel	Float	Meters/second^2	Vehicle acceleration in the longitudinal direction
InVehicle Longitudinal Speed	Float	Meters/second	Vehicle speed sampled from the vehicle network
InVehicle PRNDL	Integer	None	Vehicle transmission state
InVehicle Stability Control Status	Integer	None	Vehicle stability control status
InVehicle Steering Position	Float	Degrees	Vehicle steering wheel position in degrees
InVehicle Throttle Position	Float	None	Vehicle throttle position
InVehicle Traction Control Status	Character	None	Vehicle traction control status
InVehicle Turn Signal Left	Integer	None	Vehicle left turn signal status
InVehicle Turn Signal Right	Integer	None	Vehicle right turn signal status
InVehicle Wiper Status	Integer	None	Vehicle wiper status
InVehicle Yaw Rate	Float	Degrees/second	Vehicle yaw rate
LaneTrack Crossing Left	BIT	None	There is an exit on the left side of the road
LaneTrack Crossing Right	BIT	None	There is an exit on the right side of the road
LaneTrack Distance Left Marker	Float	Millimeters	Distance from vehicle centerline to inside of left-side lane marker based on vehicle-based machine vision
LaneTrack Distance Right Marker	Float	Millimeters	Distance from vehicle centerline to inside of right-side lane marker based on vehicle-based machine vision

Table 3.2 (Continued)

LaneTrack Lane Width	Float	Millimeters	Distance between the inside edge of the innermost lane marking to the left and right of the vehicle
LaneTrack Probability Left Exist	Integer	Percent	Probability that vehicle-based machine vision lane marking evaluation is providing correct data for the left-side lane markings
LaneTrack Probability Right Exists	Integer	Percent	Probability that vehicle-based machine vision lane marking evaluation is providing correct data for the right-side lane markings
LaneTrack Shift Aborted	BIT	None	Driver aborted crossing a line
LaneTrack Shift Left	BIT	None	Vehicle is crossing a line on the left
LaneTrack Shift Right	BIT	None	Vehicle is crossing a line on the right
LaneTrack Shift Successful	BIT	None	Vehicle lies in the lane between the painted lines
LaneTrack Type LeftLane LeftMarker	Integer	None	Type of left-most marker toward the left of the vehicle
LaneTrack Type LeftLane RightMarker	Integer	None	Type of right-most marker toward the left of the vehicle
LaneTrack Type RightLane LeftMarker	Integer	None	Type of left-most marker toward the right of the vehicle
LaneTrack Type RightLane RightMarker	Integer	None	Type of right-most marker toward the right of the vehicle

The DAS2 logs report many driving events, thus providing a wealth of detail on driving behavior that is ripe for analysis.

The DAS2 trip level summary files give trip information such as trip distance, average speed, travel time, brake counts, and more. However, as can be seen in Table 3.3, they are not associated with any location data. Samples of the DAS2 data tables are provided in the Appendix. The full data tables are 25 GB in size, collectively.

Table 3.3 DAS2 Summary Table Column Characteristics

Field Name	**Type**	**Units**	**Description**
DeviceID	Integer	None	A unique numeric ID assigned to each DAS
TripID	String	None	Count of ignition cycles—each ignition cycle commences when the ignition is in the on position and ends when it is in the off position

Table 3.3 (Continued)

Epoch Start Time	Integer	Seconds	Epoch time -- also known as Unix time, is the number of seconds that has elapsed since midnight January 1, 1970 -- at the start of a trip
Start Date	Date	mm/dd/yyyy	Date on which the trips started
Start Time	Time	hh:mm:ss	24h time stamp of the start of a trip
Epoch End Time	Integer	Seconds	Epoch time -- also known as Unix time, is the number of seconds that has elapsed since midnight January 1, 1970 -- at the end of a trip
End Date	Date	mm/dd/yyyy	Date on which the trips ended
End Time	Time	hh:mm:ss	24h time stamp of the end of a trip
Total Trip Distance	Integer	Kilometers	Total distance traveled in a trip
Distance Over 25mph	Real	Kilometers	Distance traveled in a trip when the vehicle's speed is greater than or equal to 25 mph
Distance Over 55mph	Real	Kilometers	Distance traveled in a trip when the vehicle's speed is greater than or equal to 55 mph
Trip Duration	Real	Seconds	Total time duration of a trip
Average Speed	Real	m/s	Average speed over the entire length of the trip
Maximum Speed	Real	m/s	Maximum speed reached during a trip
Brake Count	Integer	None	Number of times the driver applies the brake during a trip
Wiper Activated	String	None	Indicates whether or not the wipers were activated during a trip

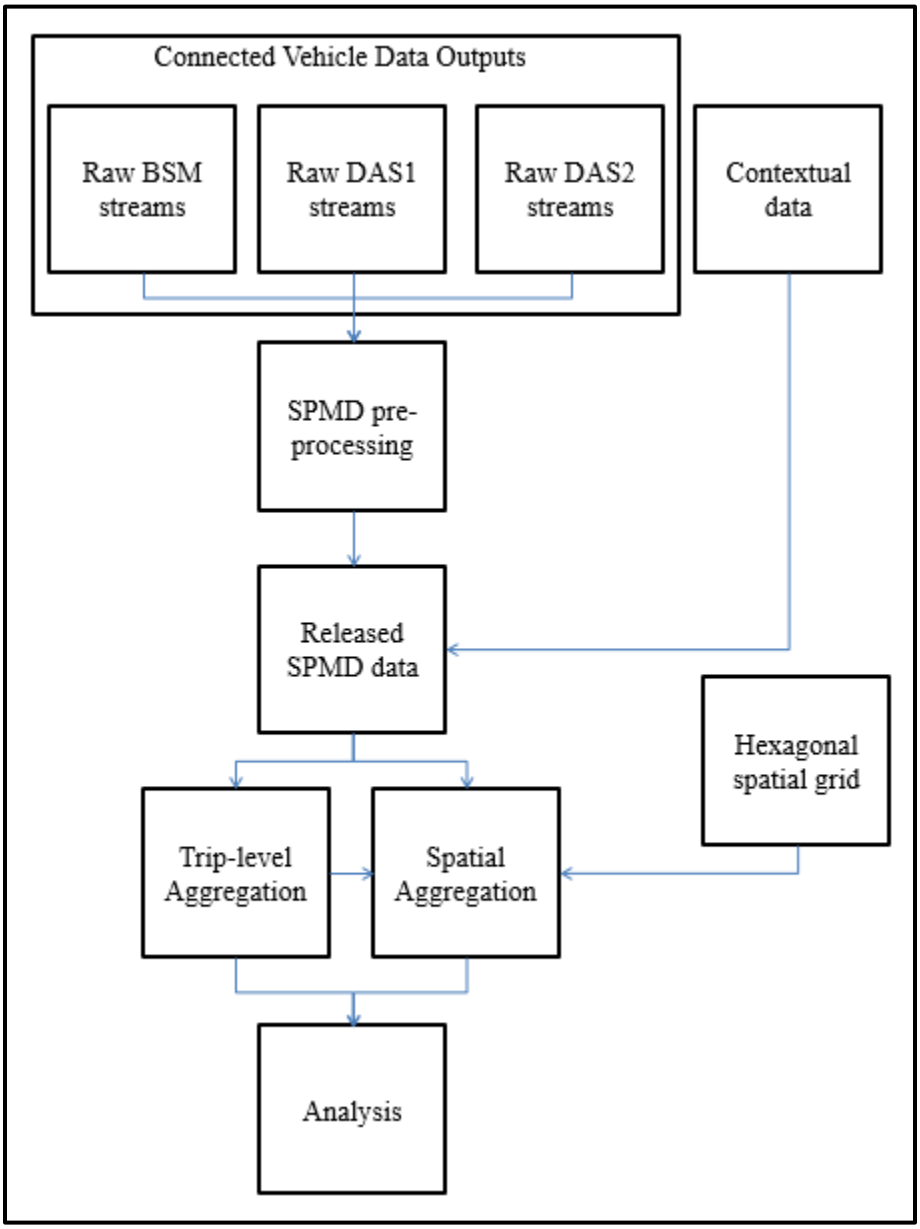
Chapter 4: Methodology

Proper processing of the initial CV data is a crucial step to obtaining usable trip information. Since the raw data tables are very large in size – some over 700 million rows and 100 GB – initial aggregation is necessary to make the data usable for any type of analysis. Additional challenges include understanding the initial data format and dealing with null values and outliers. The methodology section is broken down into five parts as follows. Part one gives an overview of the process and defines the objectives, while part two details the database structure implemented to achieve those objectives. The third and fourth parts deal with the trip level and spatial aggregation, respectively, which are performed on the data. Finally, part five provides sample analyses that can be performed on the aggregated travel data.

4.1 Conceptual Framework and Objectives

The processing methodology itself consists of four key components: inputs, objectives, database structure, and aggregation algorithms. An overview of the process is presented in Figure 4.1. It can be seen from the flow chart that the basic inputs are the three data streams from the connected vehicles as well as contextual data provided by the SPMD project. These inputs and the transportation planning context were used to develop research objectives for this data. These objectives are important to consider before beginning the aggregation process. The aggregation algorithms and the database structure that enabled them to be met are discussed in more depth in following sections. The rest of this section is focused on the initial establishment of scope and objectives. Since the scope of this research endeavor is to aid in the improvement of travel demand models and their performance measures, the specific objectives chosen reflect this.

Figure 4.1 Conceptual Overview Flowchart

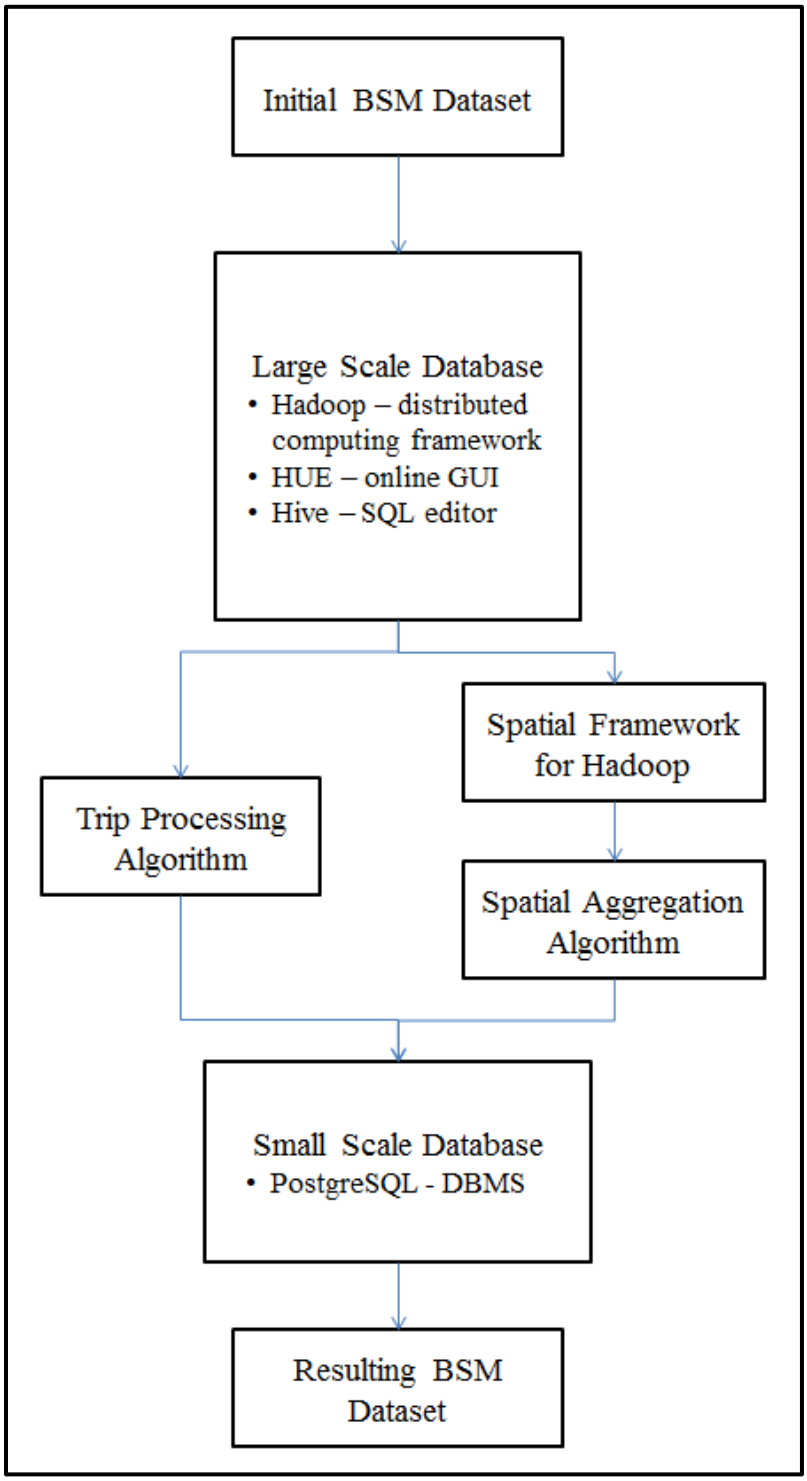


The primary objective is to aggregate the raw data at the trip level. This involves establishing parameters such as timestamps and GPS coordinates for the beginning and end of each trip, in addition to summary statistics for each trip. The summary statistics that were utilized here are trip distance, average trip speed, maximum trip speed, and travel time. These metrics were chosen because of their common usage as performance measures in transportation planning. The secondary objective is to aggregate spatially at the TAZ level in order to perform origin-destination (OD) analysis. This spatial aggregation process uses a large grid with hexagonal cells that are one mile across, with an area of 0.68 mi^2. The resulting parameters are origin and destination zones for all trips made, which can be analyzed for travel demand estimation and travel patterns. The following sections detail how these objectives were implemented and achieved.

4.2 DATABASE STRUCTURE

In order to deal with the sizeable input tables, a variety of open-source database and computing tools were used to implement the aggregation algorithms. Figure 4.2 presents an overview of this implementation framework, which is hosted on the Texas Advanced Computing Center's (TACC) servers. The initial SPMD data tables were imported into HUE (Hadoop User Experience), which is a web-based graphical user interface (GUI) for Apache Hadoop (Cloudera, Inc, 2016). Hadoop is an open-source distributed computing framework, which means that it is able to run computationally intense queries by breaking them up and distributing the work across clusters in order to return results with increased speed (Apache Software Foundation, 2014). Hadoop has been previously used in big data applications as well as for CV analysis (Nkenyereye and Jang, 2015; Patel et al., 2012).

Figure 4.2 Implementation Framework

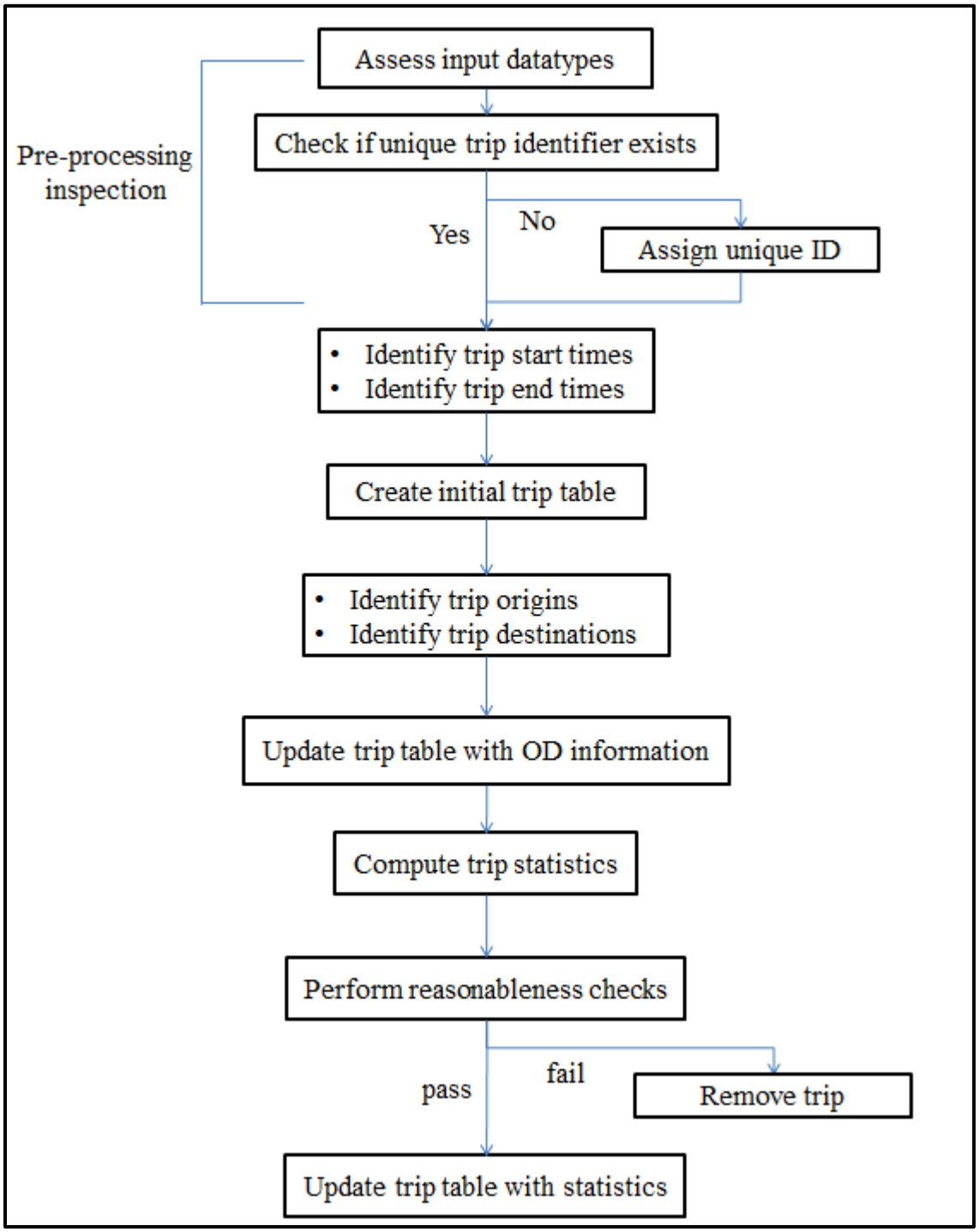


HUE also houses the Apache Hive project, which allows the user to edit and manage the datasets using SQL queries that are compatible to run on the Hadoop framework in the background (Apache Software Foundation, 2016). These three components unite to facilitate manipulation of large, often cumbersome data tables. The trip processing algorithm (Section 4.3) is run from this database setup using SQL queries. However, spatial applications require an additional component. A spatial framework for Hadoop, available on GitHub, provides User-Defined Functions (UDF) for spatial analysis in Hive (ESRI, 2016). This extension is used to facilitate the spatial aggregation algorithms (Section 4.4). Once all aggregations have been completed, the compact summary data tables can be exported to any preferred relational database management system (DBMS). In this case, PostgreSQL was used for any further data manipulation (The PostgreSQL Global Development Group, 2016). The open-source statistical software, R, was used for visualization of possible analysis applications in Section 4.5 (The R Foundation, 2016).

4.3 TRIP-LEVEL AGGREGATION

The first phase in creating useful data tables from connected vehicle data is processing the data at the trip level. Figure 4.3 provides an overview of the trip processing algorithm that was developed for this dataset and can be used for similar future datasets. The BSM data table for October will be used as input in this explanation of the trip level aggregation. Before beginning the aggregation process, it is important to discuss a few things about the input data. First, consider the input datatypes and units. Whether the numbers represent values or enumerate the presence of certain conditions is an important distinction that will affect how aggregation is implemented. This is especially important for timestamps.

Figure 4.3 Trip Processing Algorithm



In the datasets for this study, time is given in three formats: gentime, epoch time (a.k.a. Unix time), and timestamp in Central Daylight Savings time (UTC-4). Time zone is important to note because built in functions that convert epoch time to a timestamp often convert it to either local time or universal time (UTC). In this case, BSM time data is in gentime which will be converted to epoch time, then converted to local time (UTC-5). More details will be provided on how this is achieved later, but it is beneficial to acknowledge this discrepancy early on to prevent future problems. Second, check for key columns in the data tables that identify each unique trip. If no such keys exist, they will need to be added before aggregation can proceed. In this case, each trip ID is not unique across vehicles (device ID), thus a distinct trip is defined as a unique device-trip combination. If the above inspections don't identify any issues, then it is time to proceed with the algorithm.

Step one: Identify trip start and end times. The first action is to identify the trip start and end times in the BSM data table, *bsm_october*. Trip start times are defined as the minimum gentime for that unique trip. Similarly, trip end times are defined as the maximum gentime for each trip. The table *bsm_mintimes* is created for start times with columns for device ID, trip ID, start date, and start time. The query to produce this result in Hive is:

```
select rxdevice, file_id, min(gentime) as startgentime,
to_date(from_unixtime(int((min(gentime)/1000000)+1072915165
))) as startdate,
from_unixtime(int((min(gentime)/1000000)+1072915165)) as
mintime from bsm_october group by rxdevice, file_id
```

Gentime is converted to epoch time by adding the proper number of seconds between January 1st, 1970 and January 1st, 2004. This is then converted to local time by the from_unixtime function. Next, a similar table for trip end times (*bsm_maxtimes*) is created using the following query:

```
select rxdevice, file_id, max(gentime) as endgentime,
to_date(from_unixtime(int((max(gentime)/1000000)+1072915165
))) as enddate,
from_unixtime(int(((max(gentime)/1000000)+1072915165))) as
maxtime from bsm_october group by rxdevice, file_id
```

Step two: create initial trip table. Next, join the start time and end time tables together to create the initial trip table (*bsm_trip*). This is done using a simple join where unique trip identifiers trip ID and device ID are the same between the two tables.

```
select a.rxdevice as deviceid, a.file_id as tripid,
a.startgentime, a.startdate, a.mintime as starttime,
b.endgentime, b.enddate, b.maxtime as endtime from
bsm_mintimes a, bsm_maxtimes b where a.rxdevice=b.rxdevice
and a.file_id=b.file_id
```

Step three: identify the GPS coordinates of the origin and destination of each trip. The longitude and latitude of a trip origin are the GPS coordinates of the data entry where gentime is at its minimum for that trip. To find this entry, select the row from *bsm_october* where the gentime is the same as the gentime in the *bsm_mintimes* table for that trip. Since the resolution of gentime is in microseconds and messages are transmitted every decisecond, no two gentimes will be the same between the same trips. The latitude and longitude for a trip's destination is found in the same way, using the maximum gentime as the end time for each trip, found in the *bsm_maxtimes* table.

Step four: update trip table with OD information. Once these origin and destination coordinates have been identified create a trip table with longitude and latitude columns (*bsm_trip_od*). This is achieved by joining the *bsm_trip* table to the newly identified origin and destination longitude and latitude columns from *bsm_october*.

```
select a.deviceid, a.tripid, a.startgentime, a.startdate,
a.starttime, a.endgentime, a.enddate, a.endtime, b.lat as
o_lat, b.long as o_long, c.lat as d_lat, c.long as d_long
from bsm_trip a, bsm_october b, bsm_october c where
a.deviceid=b.rxdevice and a.tripid=b.file_id and
a.startgentime=b.gentime and a.deviceid=c.rxdevice and
a.tripid=c.file_id and a.endgentime=c.gentime
```

Step five: compute the selected statistics for each trip. As discussed in Section 4.1, the chosen summary statistics for this study are average speed, maximum speed, trip distance, and trip duration. Trip distance, in kilometers, is calculated by assuming each message during the trip is transmitted every 0.100 seconds and then multiplying this value by the speed reported in each message. It is assumed that the speed remains constant for the 0.100 seconds between messages. Since time is recorded in microseconds, the distance calculation is still accurate to six decimal places for kilometers. Average and maximum speeds, in meters per second, are calculated by taking the average and maximum of all speeds reported for that trip, respectively. Trip duration, in seconds, is calculated by subtracting the max gentime from the min gentime for that trip. The query used in Hive to perform these calculations and create table *bsm_stats* from the results is excerpted below:

```
select rxdevice, file_id, sum(speed*0.1/1000) as
tripdistancekm, avg(speed) as avgspeedms, max(speed) as
maxspeedms, (max(gentime/1000000)-min(gentime/1000000)) as
tripdurations from bsm_october group by rxdevice, file_id
```

Step six: perform reasonableness checks on the summary statistics. Various checks are performed at this stage in order to avoid reporting incorrect values for trip characteristics. Of course, validation has been occurring throughout the entire process, making sure that all queries work as expected. Specific checks for this step include: (1) check for trips with trip duration less than one second; (2) make sure average speeds are not greater than maximum speeds; (3) check that all GPS coordinates are reasonable and not null.

Step seven: create the final trips summary table. Discard or fix any trips that do not pass the above reasonableness checks. Add the summary statistics to the trip

summary table (*bsm_trip_summary*) by joining the tables *bsm_stats* and *bsm_trip_od* by device ID and trip ID as shown below:

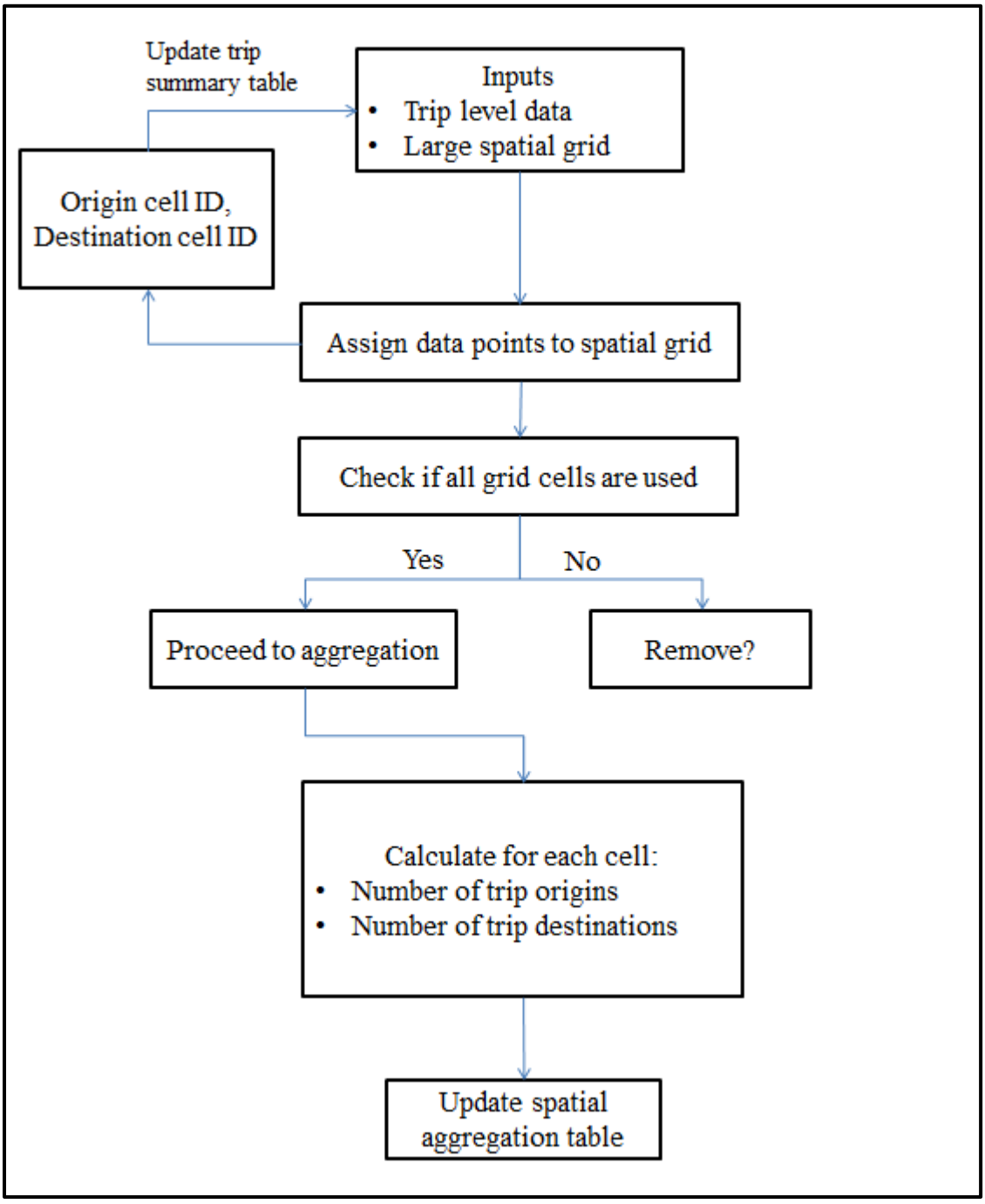
```
select a.deviceid, a.tripid, a.startgentime, a.startdate,
a.starttime, a.endgentime, a.enddate, a.endtime, a.o_lat,
a.o_long, a.d_lat, a.d_long, b.tripdistancekm,
b.avgspeedms, b.maxspeedms, b.tripdurations from
bsm_trip_od a, bsm_stats b where a.deviceid=b.rxdevice and
a.tripid=b.file_id
```

This completes the trip level aggregation by producing a trip summary table for the BSM data. The final table has 16 columns and one row for each of the 112,095 trips and is excerpted in the Appendix. The size of this file is 22.95 MB, down from the initial 98.7 GB in *bsm_october*.

4.4 SPATIAL AGGREGATION

Once the trip level aggregation has been completed, it can be used for the spatial aggregation procedures. For spatial processing at the OD level, the two inputs are the trip level data table and the preferred size grid. For this application a hexagonal grid (*grid_json*) with cells that measure one mile across and yield an area of 0.68 mi^2 was used. The grid was created in CartoDB, an open source GIS software tool, sized by GPS coordinate extents, and exported as a JSON file. A hexagonal grid was chosen because of its lack of sampling bias due to edge effects (Birch et al., 2007). The JSON file format is necessary for compatibility with the spatial extension for Hadoop. This spatial aggregation process can be performed with any trip summary table. In this case the input table was the preprocessed summary table for the DAS2 dataset, modified to add origin and destination coordinates much like in Section 4.3. This table has 13,153 trips and will be referred to as *trip_od_das2*. The spatial aggregation framework itself is visually presented in Figure 4.4 below, followed by a discussion of the steps.

Figure 4.4 Spatial Aggregation Algorithm



Step one: Assign trips to cells in the spatial grid. Using the spatial extension to Hadoop, a query is run to join the grid and feature IDs of the hexagonal grid to each trip ID in the trip table (*trip_od_das2*).

```
create temporary function ST_Point as
'com.esri.hadoop.hive.ST_Point';
create temporary function ST_Intersects as
'com.esri.hadoop.hive.ST_Intersects';

create table device_trip_OD_grid as select distinct
deviceid, trip, o_fid, o_gid, g2.fid as d_fid, g2.gid as
d_gid
from
(select distinct deviceid, trip, g.fid as o_fid, g.gid as
o_gid, dest
from (
  select distinct deviceid, trip, ST_Point(o_long, o_lat)
as ori, ST_Point(d_long, d_lat) as dest
         from trip_od_das2 )sub1, grid_json as g
       where ST_Intersects(g.geometry, ori)
)sub2, grid_json as g2
where ST_Intersects(g2.geometry, dest)
```

This step results in a table, *device_trip_OD_grid*, which contains trip ID, feature ID (fid) and geometry ID (gid) for each origin and destination for each trip in the input table (See Appendix for excerpt). The geometry ID is the number that coordinates with the grid cell ID; the feature ID is an index that is irrelevant to the larger process. At this point, it is crucial to check that all trips have been properly assigned to a grid cell for both the origin and destination. Possible errors can occur if: (a) the origin or destination coordinates are outside the grid range, or (b) the origin or destination lies on a grid cell border. These issues can be resolved by altering the grid bounds and remove a double entry for a trip that falls on a border.

Step two: check for unused grid cells. If the grid is very large, spatial manipulation and visualization will become quite cumbersome. At this point it is appropriate to assess the boundary of the desired analysis area. If there are outlier trips that originate or end outside of the study area, this can cause a very large grid to be created. In order to avoid this, external zones can be created for the designation of these

outlier trips. Then the rest of the grid cells will be dedicated to the main study area. Alternatively, these trips could be removed from the study as outliers. In this case, there were nine trips outside of the grid area, so they were assigned to the external zone. A number of grid cells that were outside the determined study area and had no trip origins or destinations assigned to them were removed from the table for easier computation and visualization. It is up to the researcher to decide which method is best suited for their work.

Step three: formulate OD table. In the previous step the table is indexed by trip ID. The objective of this step is to create an OD table that contains each cell ID (gid) followed by the number of trips with their origin in the cell and the number of trips with their destination in the cell. This format will be more useful for visualization and OD analysis. The OD table (*das2_od*), excerpted in the Appendix, is created by the following query:

```
select o_gid, d_gid, count(trip) as trips into das2_od from
device_trip_OD_grid group by o_gid, d_gid;
```

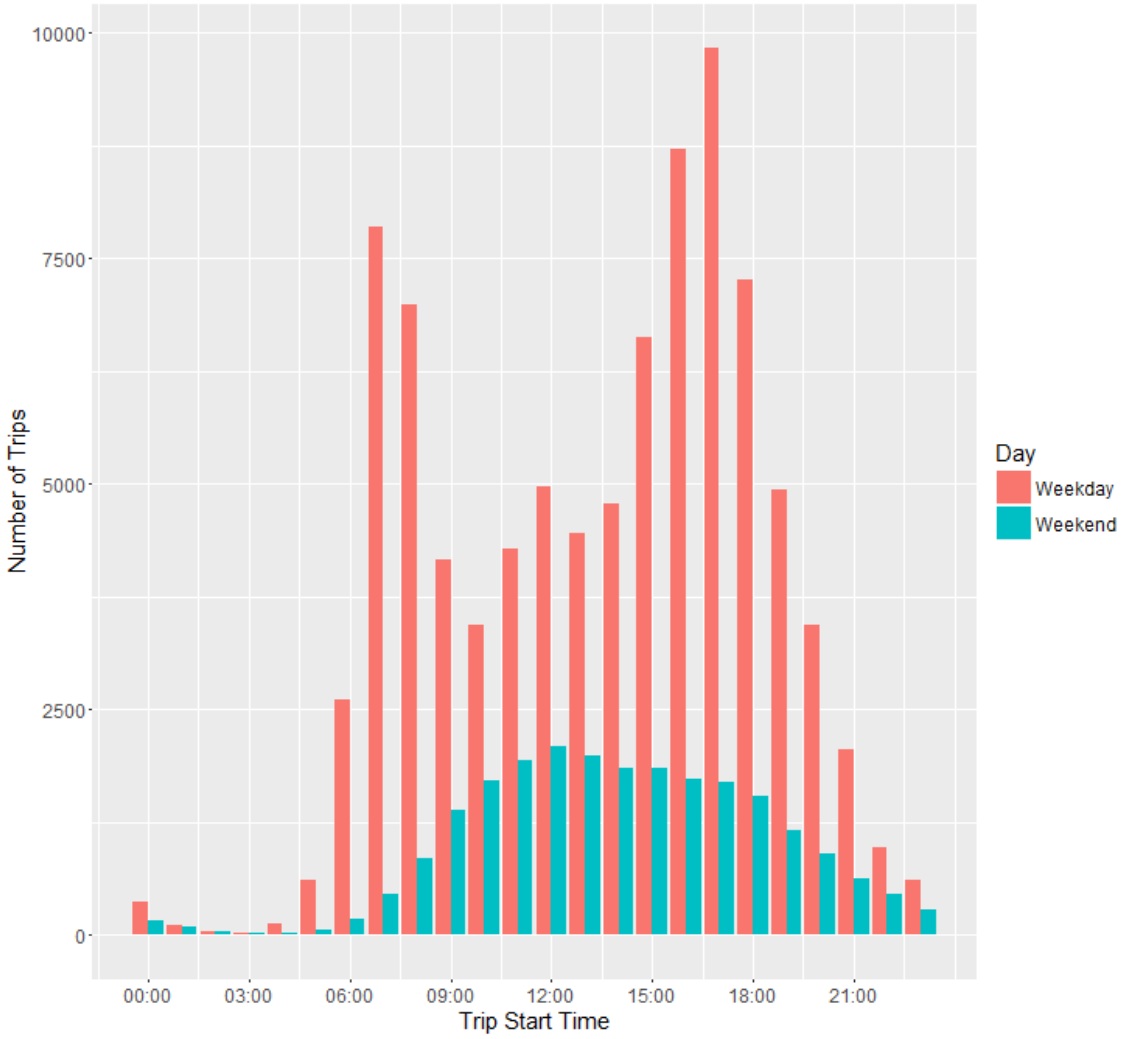
4.5 Sample Analysis Outcomes

This section provides some brief insight into possible analysis outcomes when working with the processed data tables presented above. These are by no means a complete listing of applications or results. Additionally, no deep analysis is made of travel demand patterns or system performance for the SPMD dataset. Rather, the figures presented here represent areas of analysis that are facilitated by practical data processing covered in this thesis.

Transportation models often rely on assumptions about how demand changes by time of day. It is common to look at the pattern of trips made across the 24-hour cycle.

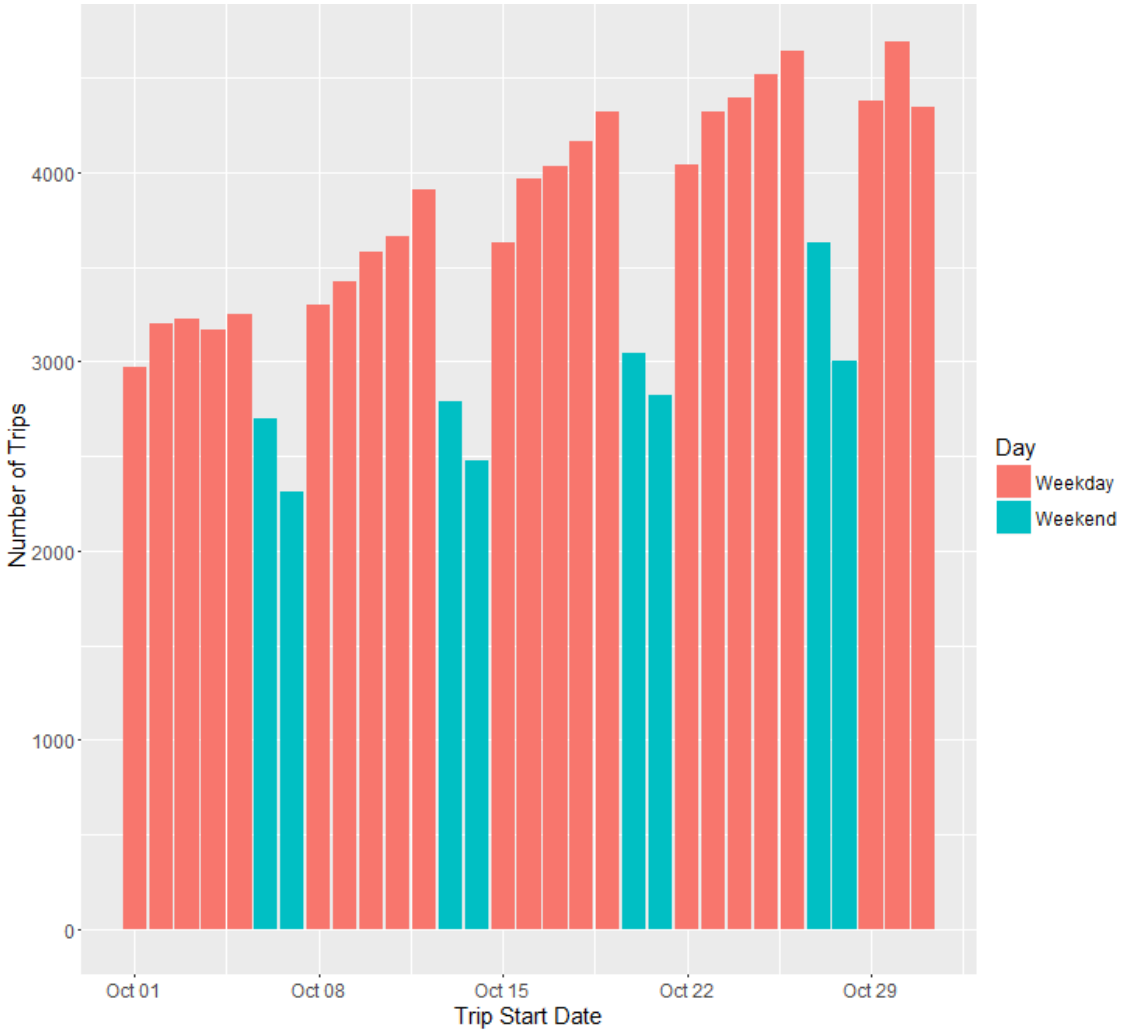
The following figure shows this distribution for the BSM trip data, grouped by weekday or weekend.

Figure 4.5 BSM Trip Distribution by Time of Day



The figure shows a typical pattern of peaks in weekday travel demand during the AM (6-9 am) and PM (3-7 pm). Trips can also be viewed by day of month. It is often assumed that all weekdays have the same level of demand, but this assumption can be examined in Figure 4.6, also created from the BSM trip data.

Figure 4.6 BSM Trip Distribution by Day



This figure shows increasing demand throughout the weekdays and a larger number of trips made on Saturday than Sunday. Further analysis of the data may lead to altered conclusions about daily demand distribution. The spatial data results lend themselves to visualization of different traits. In the example below a map of the spatial results was created to show which zones are the most popular origins. Figure 4.7 is a zoomed out view of the DAS2 zones and Figure 4.8 is zoomed in to show detail for Ann Arbor.

Figure 4.7 Full View of Study Area

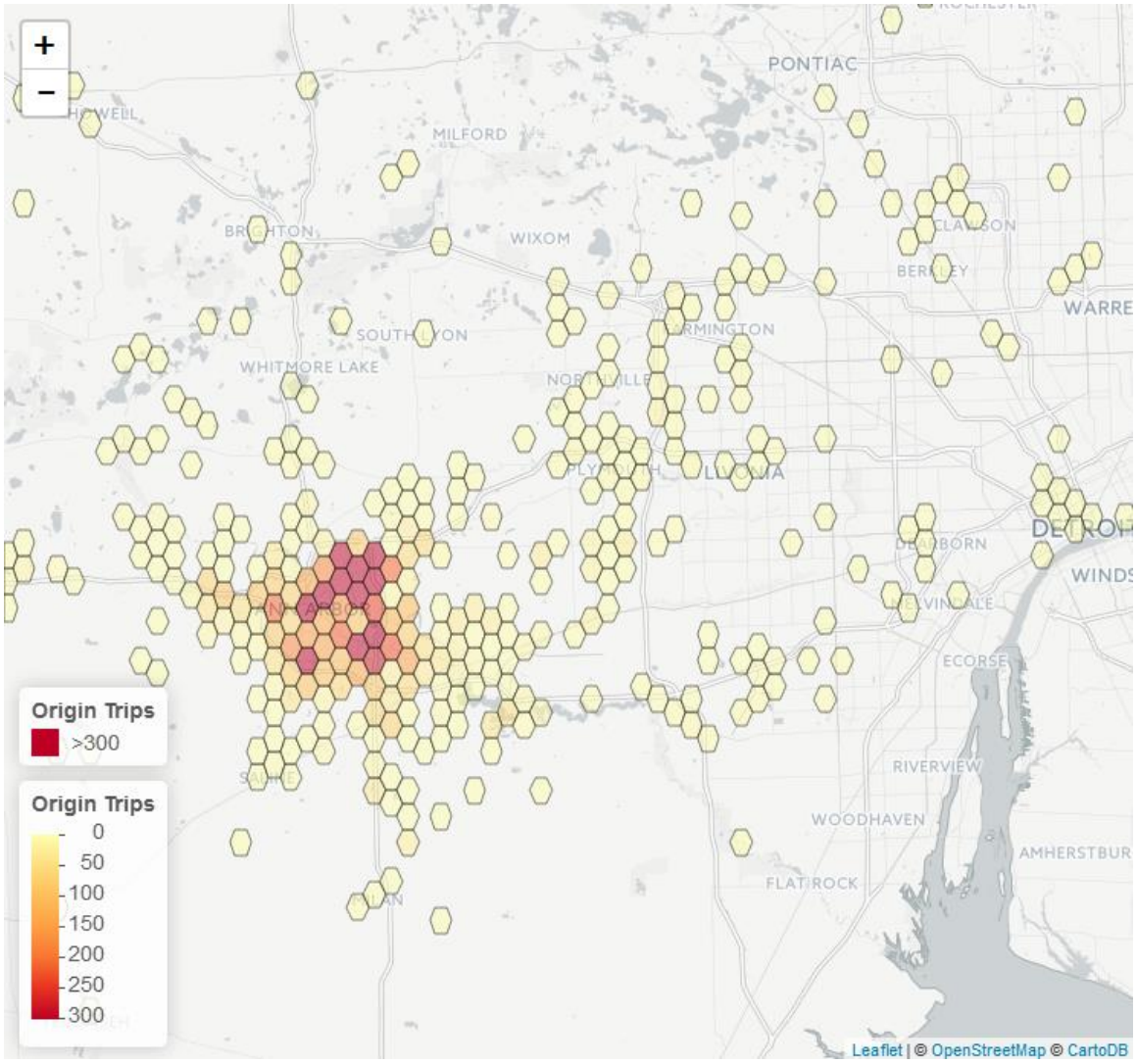
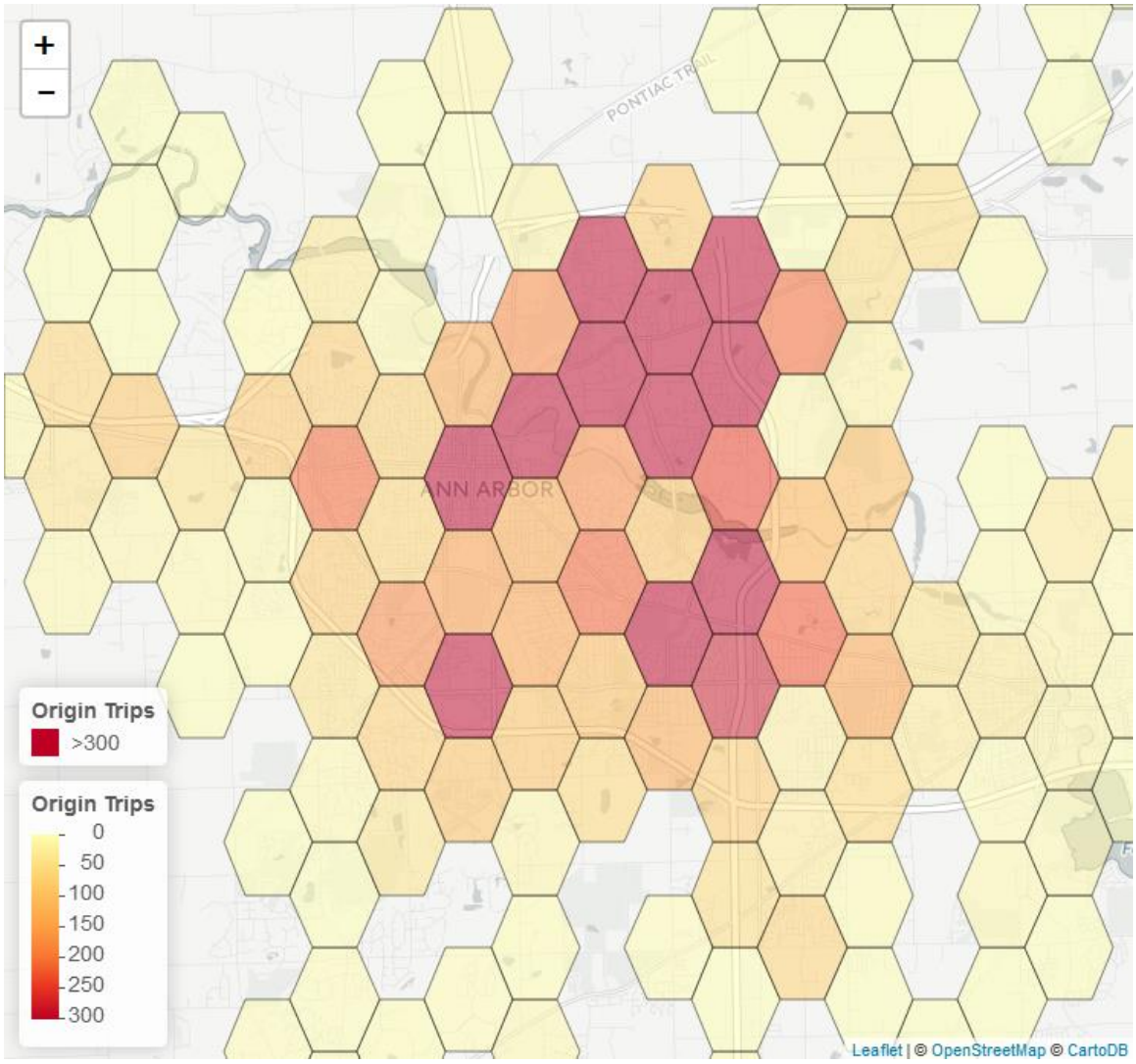


Figure 4.8 Popularity of Origin Zones in Ann Arbor



The above map shows the more popular origin zones in red and less popular zones in yellow. This information could be used to support and validate travel demand models and OD matrices. It also gives a clear visual of where the most trips are being produced. The above figures are just a few examples of the possible applications of the processed data that was created using the proposed framework.

Chapter 5: Conclusions

There are many transportation analysis possibilities in the future with connected vehicle data. This thesis covers just a few applications with the largest CV dataset currently available. The following sections serve to summarize the above findings, as well as provide recommendations for future CV data processing efforts, and explore future areas of exploration.

5.1 SUMMARY

Connected vehicle technology is becoming increasingly popular and, due to future safety requirements, will most likely become commonplace in the next decade. This new technology offers new transportation data that opens a world of possibilities for applications in travel demand planning, system performance measures, and more. CV data offers easily accessible, detailed driving and trip data that has the potential to fill in the gaps left by traditional data acquisition methods. However, the greatest challenge in using this data source is how to process into a manageable size and comprehensible format. This thesis explored two methods of aggregation, trip level and spatial, to produce CV data in a practical format. Trip level aggregation involved processing raw message data into trips with time and location characteristics, in addition to summary statistics on speed, distance, and duration. Potential issues and troubleshooting techniques were discussed to retain as much value in the data as possible. Spatial aggregation processing included the creation of a hexagonal grid, then implementing spatial joins to group trip origins and destinations into zones. These aggregation procedures were facilitated by the distributing computing framework, Hadoop, as well as various other database management tools and spatial extensions. The focus of this study was on creating a replicable processing framework for CV data; however a few snapshots of

possible analyses were also presented. The necessity of practical processing methodologies will grow as the availability of CV data continues to increase in the future.

5.2 Recommendations and Lessons Learned

This section provides recommendations for successful manipulations of similar CV data sets in future studies. Among the lessons learned during the course of this study, one of the more crucial ones is to always inspect and fully understand the inputs before beginning to process the data. In this study many incongruities had been created in the documentation that could have caused problems later on, if not caught upon initial inspection. These included time zone issues, trip determination discrepancies, and lack of unique identifiers. The analyst should make sure all units are properly defined and are consistent across data tables. This includes understanding how epoch time is converted to universal time, and the time zone difference between the collection and processing locations. It is important to differentiate between quantitative and enumerated variables for proper aggregation. The integer 15 in a braking column may represent a single event occurrence or 15 events depending on how the table is organized. All of these aspects can be condensed into the singular idea that one should always question the data source, and any manipulation or cleaning that was performed before its release.

As far as the actual processing procedures, the implementation framework makes a huge difference in computation time and database management. Using a distributed computing framework, like Hadoop, will become essential as CV datasets increase in the future. The importance of computing resources when working with this data opens up the possibility for interdisciplinary cooperation. This is especially relevant for spatial queries over a larger study area. The city of Ann Arbor only covers 30 mi^2, and many future endeavors will cover more expansive regions. Additionally, the choice of the grid cell

size and coverage area will impact analysis results, so it is important consider why a certain zone size is chosen. Overall, carefully considered inputs and a consistent processing methodology can lead to useful, practical results.

5.3 FUTURE WORK

Once the data has been processed and aggregated into a usable manner, many doors are opened in the transportation planning process. Trips can be used for OD analysis and travel demand model inputs. Raw message data can be used to analyze congestion, using speeds on roadway segments, or can be formed into trajectories to study route choice. This information can be useful on an arterial or corridor level as well. Performance measures such as corridor travel time and speed can be calculated from the detailed records that CVs produce. The driving data that was released along with the BSM data can be used in place of naturalistic driving studies to analyze driver behavior. Driver behavior is relevant to the planning perspective in terms of modeling assumptions such as the user equilibrium principle and temporal travel patterns. CV records have the potential to answer this question. Most sources do not provide long term historical data that would be necessary to assess such travel trends that are often taken for granted. Furthermore, many vehicle system statuses have the potential to be monitored with CV technology. These features could aid tremendously in conflict and collision studies. The SPMD dataset could have many implications in the transportation planning field, but these findings remain hidden if the data format isn't functional.

Additionally, there are future studies in the works that benefit from the practical operability contributions of the SPMD project. Three USDOT CV pilot deployment programs are currently in progress in Wyoming, New York City, and Tampa, Florida. In Wyoming, V2I applications are being implemented along its I-80 corridor to prevent

weather-related incidents. This highway is a heavy trucking corridor and experiences a majority of accidents due to snow-covered roads, high winds, and poor visibility. Specific V2V and V2I applications include Forward Collision Warning (FCW), Spot Weather Impact Warning, and Distress Notification (USDOT, 2015). In New York City, the primary objective is to use CV technology to decrease pedestrian accidents and increase safety of all travelers. A multitude of V2V safety and pedestrian applications are planned including Red Light Violation Warning (RLVW), Pedestrian in Signalized Crosswalk, Intersection Movement Assist (IMA), and Emergency Electronic Brake Lights (EEBL) (USDOT, 2015). Up to 8,000 equipped vehicles are expected to be part of the deployment, which will be a chance to study CV technology in a dense urban environment. In Tampa, the goal is to minimize accidents and peak hour delays by increased efficiency in signal operations and mobility through different ITS applications. These applications include Intelligent Traffic Signal System, Curve Speed Warning, FCW, EEBL, and IMA, as defined above (USDOT, 2015). The Florida test spot is unique because it involves a blend of transportation environments that range from a commuter expressway to arterials with transit services and dense pedestrian traffic. These three V2V and V2I studies are surely the first among many future experiments to provide comprehensive data on our transportation systems and their performance. The New York study will have almost three times the amount of vehicles in the Ann Arbor SPMD project, which only makes it more critical to establish practical analysis methods for the produced datasets.

Not all future CV systems are likely to produce large datasets due to the privacy concerns such studies incur. The balance between protecting privacy and data collection needs is difficult to achieve and it is still very uncertain how it will be achieved. Often the protection of identifying information falls to the data owners, which are also unknown in

the CV market. On one hand, private third parties could be in charge of data collection systems and would perhaps sell this data in aggregate forms to researchers. On the other hand, if public entities were in charge of monitoring this data, it is unlikely very much of it would be stored long term. Either way it is unlikely that personally identifiable information from DSRC messages would be released to the public. The USDOT has said that V2V systems will not be enabled to collect and store personal data, and that the system will be operated by private entities (ITS JPO, 2015). Concerns over privacy will influence how useful CV data is and how easy it will be to obtain. Large scale aggregation protects identity but it removes a lot of the defining level of detail CV technology can provide. Privacy protection is a good example of an area of CV technology that has yet to be finalized and might make the difference in the future of DSRC data in transportation planning. The SPMD project in Ann Arbor, MI provided an opportunity to explore the possibilities of such data with the hope that it will influence policy by demonstrating what can be accomplished with this new frontier.

Appendix

Table A.1 BSM Data Sample

rxdevice	file_id	txdevice	gentime	txrandom	msgcount	dsecond	lat
10	13750	10	276176188142815	0	20	2000	42.238522
10	13750	10	276176188242789	0	21	2100	42.238529
10	13750	10	276176188342827	0	22	2200	42.238548
10	13750	10	276176188442792	0	23	2300	42.23856
10	13750	10	276176188542827	0	24	2400	42.238571
10	13750	10	276176188642791	0	25	2500	42.238579
10	13750	10	276176188742793	0	26	2600	42.23859
10	13750	10	276176188842795	0	27	2700	42.238609
10	13750	10	276176188942850	0	28	2800	42.238621
10	13750	10	276176189042797	0	29	2900	42.238628

Table A.1 (Continued)

long	elevation	speed	heading	Ax	Ay	Az
-83.6519928	213.8000031	13.14000034	3.424999952	-0.140000001	0.01	-10
-83.65197754	213.8000031	13.14000034	3.037499905	-0.140000001	0.01	-10
-83.65197754	213.8000031	13.14000034	3.212500095	-0.140000001	0.01	-10
-83.65197754	213.8000031	13.10000038	3.125	-0.209999993	0.01	-10
-83.65197754	213.8000031	13.14000034	3.025000095	-0.209999993	0.01	-10
-83.65197754	213.8000031	13.18000031	3.287499905	-0.289999992	0.01	-10
-83.65197754	213.8000031	13.06000042	2.875	-0.209999993	0.01	-10
-83.65197754	213.8999939	13.10000038	3.887500048	-0.209999993	0.01	-10
-83.65197754	213.8000031	13	4.050000191	-0.209999993	0.01	-10
-83.65196991	213.8000031	12.89999962	3.700000048	-0.289999992	0.01	-10

Table A.1 (Continued)

Yawrate	PathCount	RadiusofCurve	Confidence
0.800000012	12	3276.699951	68
0.689999998	12	3276.699951	71
0.689999998	12	3276.699951	76
0.689999998	12	3276.699951	82
0.689999998	12	3276.699951	86
0.889999986	12	3276.699951	84
0.400000006	12	3276.699951	99
0.689999998	12	3276.699951	98

Table A.1 (Continued)

0.689999998	12	2232	97
1.200000048	12	1909.5	82

Table A.2 DAS2 Message Data Sample

deviceid	trip	time	gps elevation	gps fix quality	gps hdop	gps heading	gps latitude
10	412198	0	207.393005	1	0.83	140.720001	42.245967
10	412198	100	207.393005	1	0.83	140.720001	42.245956
10	412198	200	207.292999	1	0.83	141.009995	42.245944
10	412198	300	207.393005	1	0.83	140.759995	42.245932
10	412198	400	207.393005	1	0.83	140.729996	42.245921
10	412198	500	207.393005	1	0.83	140.619995	42.245909
10	412198	600	207.393005	1	0.83	140.699997	42.245897
10	412198	700	207.393005	1	0.83	140.839996	42.245886
10	412198	800	207.292999	1	0.83	140.789993	42.245874
10	412198	900	207.292999	1	0.83	140.639999	42.245863

Table A.2 (Continued)

gps longitude	gps number satellites	gps pdop	gps speed	gps utc time	gps valid	das pitch rate	das roll rate
-83.650319	9	1.5	16.14125	1350150235300	1	-1.951172	5.52832
-83.650306	9	1.5	16.061083	1350150235400	1	-0.325195	-4.552734
-83.650294	9	1.5	15.884306	1350150235500	1	-0.650391	-5.853516
-83.650281	9	1.5	15.909486	1350150235600	1	-0.325195	-3.577148
-83.650268	9	1.5	15.971153	1350150235700	1	-1.951172	-1.300781
-83.650255	9	1.5	15.922333	1350150235800	1	1.951172	3.251953
-83.650242	9	1.5	15.980403	1350150235900	1	2.276367	6.829102
-83.65023	9	1.5	15.726028	1350150236000	1	2.276367	1.300781
-83.650217	9	1.5	15.638667	1350150236100	1	0.975586	-3.577148
-83.650205	9	1.5	15.722944	1350150236200	1	0.650391	-1.300781

Table A.2 (Continued)

invehicle abs state	invehicle brake status	invehicle headlight status	invehicle longitudinal accel	invehicle longitudinal speed	invehicle prndl	invehicle stability control status	invehicle steering position
2	0	0	-0.297	16.655556	2	1	-2.78125
2	0	0	-0.297	16.588888	2	1	-3.59375
2	0	0	-0.3746	16.533333	2	1	-3.6875
2	0	0	-0.3746	16.533333	2	1	-3.6875

Table A.2 (Continued)

2	0	0	-0.3746	16.458334	2	1	-3.375
2	0	0	-0.5298	16.283333	2	1	-3.28125
2	0	0	-0.3746	16.38611	2	1	-3.28125
2	0	0	-0.5298	16.391666	2	1	-2.78125
2	0	0	-0.4522	16.494444	2	1	-2.375
2	15	0	-0.8402	16.433332	2	1	-2.5

Table A.2 (Continued)

invehicle throttle position	invehicle traction control status	invehicle turn signal left	invehicle turn signal right	invehicle wiper status	invehicle yaw rate	lanetrack crossing left
0	2	0	0	0	0.031525	0
0	2	0	0	0	-0.368475	0
0	2	0	0	0	-0.168475	0
0	2	0	0	0	-0.168475	0
0	2	0	0	0	-0.168475	0
0	2	0	0	0	0.131525	0
0	2	0	0	0	0.031525	0
0	2	0	0	0	-0.168475	0
0	2	0	0	0	-0.368475	0
0	2	0	0	0	0.331525	0

Table A.2 (Continued)

lanetrack crossing right	lanetrack distance left marker	lanetrack distance right marker	lanetrack lane width	lanetrack probability left exist	lanetrack probability right exists	lanetrack shift aborted
0	-1483	2077	3561.078613	39	227	0
0	-858	2397	3256.27832	39	240	0
0	-502	2722	3225.796143	39	244	0
0	-548	2788	3337.55542	39	257	0
0	91	3215	3124.193115	39	261	0
0	462	3307	2844.789551	37	279	0
0	751	2987	2235.188721	23	304	0
0	-1534	2971	3657.599854	23	300	0
0	-1940	2240	3657.599854	23	320	0
0	-1951	2173	3657.599854	23	323	0

Table A.2 (Continued)

lanetrack shift left	lanetrack shift right	lanetrack shift successful	lanetrack type leftlane leftmarker	lanetrack type leftlane rightmarker	lanetrack type rightlane leftmarker	lanetrack type rightlane rightmarker
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0

Table A.3 DAS2 Trip Summary Data Sample

deviceid	tripid	epochstarttime	startdate	starttime	epochendtime	enddate
23	1541756	1367352191600	4/30/2013	20:03:12	1367352693800	4/30/2013
36	1468324	1365359330500	4/7/2013	18:28:50	1365359457000	4/7/2013
32	1522912	1367168623800	4/28/2013	17:03:44	1367169130700	4/28/2013
87	1616489	1366110780400	4/16/2013	11:13:00	1366112314000	4/16/2013
36	1460272	1364938650000	4/2/2013	21:37:30	1364941211600	4/2/2013
27	1443012	1365503611700	4/9/2013	10:33:32	1365504267200	4/9/2013
14	1474724	1366568344500	4/21/2013	18:19:04	1366568726400	4/21/2013
35	1514042	1364941544100	4/2/2013	22:25:44	1364942168800	4/2/2013
37	1429645	1366893043700	4/25/2013	12:30:44	1366893740200	4/25/2013
75	1497697	1365315684000	4/7/2013	6:21:24	1365316476500	4/7/2013

Table A.3 (Continued)

endtime	totaltripdistance	distanceover25mph	distanceover55mph	tripduration	averagespeed
20:11:34	5316.955125	5082.539423	101.0010114	502200	16.87952423
18:30:57	4192.276454	4192.276454	4192.276454	126500	35.07694244
17:12:11	5576.463881	5172.342529	1443.330402	506900	13.04009819
11:38:34	8508.451587	7009.928312	2754.700024	1533600	12.79083252
22:20:12	78541.51709	78460.23687	76969.0977	2561600	33.92284775
10:44:27	15883.60832	15555.70448	12609.35655	655500	25.35387039
18:25:26	5960.107321	5864.966913	74.35819197	381900	15.5573225
22:36:09	7945.283655	7274.10498	99.49617743	624700	12.9056139

Table A.3 (Continued)

12:42:20	8716.98514	7158.711641	306.3503472	696500	12.33102608
6:34:37	20800.48312	20679.47215	14340.70929	792500	26.80497932

Table A.3 (Continued)

maximumspeed	brakecount	wiperactivated
25.38199997	1	
40.12649918	1	
27.15542984	1	
38.31658173	1	1
39.15987396	1	
36.28466797	1	
24.79308319	1	
25.02176476	1	
26.18726349	1	
34.26354218	1	

Table A.4 BSM Trip Summary Excerpt

deviceid	tripid	startgentime	startdate	starttime	endgentime	enddate
50	33482	276630836303729	10/6/2012	12:53:21	276633406403299	10/6/2012
50	33484	276646397186938	10/6/2012	17:12:42	276646618187057	10/6/2012
50	33486	276649179006612	10/6/2012	17:59:04	276649445206753	10/6/2012
50	33487	276669942853180	10/6/2012	23:45:07	276670883053063	10/7/2012
50	33488	276695841487467	10/7/2012	6:56:46	276695871687534	10/7/2012
50	33489	276702469579421	10/7/2012	8:47:14	276702620579328	10/7/2012
50	33490	276704075264791	10/7/2012	9:14:00	276704187764589	10/7/2012
50	33492	276712229457270	10/7/2012	11:29:54	276712391257209	10/7/2012
50	33493	276727901604049	10/7/2012	15:51:06	276728262003946	10/7/2012
50	33494	276731491440116	10/7/2012	16:50:56	276731493840073	10/7/2012

Table A.4 (Continued)

endtime	o_lat	o_long	d_lat	d_long	tripdistancekm
13:36:11	42.49966812	-83.16877747	42.29072189	-83.69242096	67.10096415
17:16:23	42.28789139	-83.70266724	42.27085114	-83.69724274	2.354433994
18:03:30	42.27088928	-83.69711304	42.28815079	-83.69010925	3.081628002
0:00:48	42.28813171	-83.69975281	42.28815842	-83.69011688	7.836409993
6:57:16	42.28808975	-83.69954681	42.28807068	-83.70352936	0.346181999
8:49:45	42.27767181	-83.69901276	42.28818893	-83.69011688	2.327034002

Table A.4 (Continued)

9:15:52	42.2878685	-83.7026825	42.29032898	-83.71304321	1.433474002
11:32:36	42.2939682	-83.71248627	42.29198837	-83.69246674	2.055895998
15:57:07	42.28712082	-83.68250275	42.21976089	-83.68547821	8.519499991
16:50:58	42.22064972	-83.67978668	42.22113037	-83.67980957	0.055888

Table A.4 (Continued)

avgspeedms	maxspeedms	tripdurations
26.1144052	36.90000153	2570.09957
10.64872906	21.04000092	221.000119
11.57201653	20.31999969	266.200141
10.75248353	19.76000023	940.199883
11.38756574	17.31999969	30.20006698
15.34982851	22.68000031	150.999907
12.69684678	20.07999992	112.499798
12.69071604	20.02000046	161.799939
23.63901218	32.88000107	360.399897
22.35519989	22.65999985	2.399957001

Table A.5 Device Trip OD Grid Excerpt

deviceid	trip	o_fid	o_gid	d_fid	d_gid
10	412198	33984	33985	33984	33985
10	412244	33617	33618	33617	33618
10	412511	33615	33616	33433	33434
10	412584	33984	33985	33984	33985
10	412745	33984	33985	33617	33618
10	412758	33617	33618	33800	33801
10	413027	33617	33618	33984	33985
10	413133	33798	33799	33613	33614
10	413279	34169	34170	32873	32874
10	413337	32873	32874	33984	33985

Table A.6 DAS2 OD Table Excerpt

o_gid	d_gid	trips
33432	33432	24
32876	32876	23
33249	33248	23

Table A.6 (Continued)

33615	33433	23
33804	32875	23
32691	32875	22
32875	32691	22
32877	33433	22
33061	33061	22
33248	33064	22

References

Anderson, B., 2016. NHTSA V2V NPRM Update.

Apache Software Foundation, 2016. Apache Hive TM [WWW Document]. Hive. URL http://hive.apache.org/ (accessed 10.23.16).

Apache Software Foundation, 2014. Welcome to Apache™ Hadoop®! [WWW Document]. Hadoop. URL http://hadoop.apache.org/ (accessed 10.23.16).

Argote-Cabañero, J., Christofa, E., Skabardonis, A., 2015. Connected vehicle penetration rate for estimation of arterial measures of effectiveness. Transportation Research Part C: Emerging Technologies 60, 298–312. doi:10.1016/j.trc.2015.08.013

Bagheri, E., Mehran, B., Hellinga, B., 2015. Real-Time Estimation of Saturation Flow Rates for Dynamic Traffic Signal Control Using Connected-Vehicle Data. Transportation Research Record: Journal of the Transportation Research Board 2487, 69–77. doi:10.3141/2487-06

Bezzina, D., Sayer, J., 2015. Safety Pilot Model Deployment: Test Conductor Team Report (No. DOT HS 812 171). National Highway Traffic Safety Administration, Washington, DC.

Bhat, C.R., Goulias, K.G., Pendyala, R.M., Paleti, R., Sidharthan, R., Schmitt, L., Hu, H.-H., 2013. A household-level activity pattern generation model with an application for Southern California. Transportation 40, 1063–1086. doi:10.1007/s11116-013-9452-y

Birch, C.P.D., Oom, S.P., Beecham, J.A., 2007. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. Ecological Modelling 206, 347–359. doi:10.1016/j.ecolmodel.2007.03.041

Bohte, W., Maat, K., 2009. 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, 285–297. doi:10.1016/j.trc.2008.11.004

Carpenter, C., Fowler, M., Adler, T., 2012. Generating Route-Specific Origin-Destination Tables Using Bluetooth Technology. Transportation Research Record: Journal of the Transportation Research Board 2308, 96–102. doi:10.3141/2308-10

Cathey, F., Dailey, D., 2003. Estimating Corridor Travel Time by Using Transit Vehicles as Probes. Transportation Research Record: Journal of the Transportation Research Board 1855, 60–65. doi:10.3141/1855-07

Cloudera, Inc, 2016. Hue - Hadoop User Experience - The Apache Hadoop UI [WWW Document]. Hadoop User Experience. URL http://gethue.com/ (accessed 10.23.16).

Coifman, B., Kim, S., 2009. Measuring Freeway Traffic Conditions with Transit Vehicles. Transportation Research Record: Journal of the Transportation Research Board 2121, 90–101. doi:10.3141/2121-10

Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C., 2015. Analyzing Cell Phone Location Data for Urban Travel. Transportation Research

Record: Journal of the Transportation Research Board 2526, 126–135. doi:10.3141/2526-14
Dhakar, N., Srinivasan, S., 2014. Route Choice Modeling Using GPS-Based Travel Surveys. Transportation Research Record: Journal of the Transportation Research Board 2413, 65–73. doi:10.3141/2413-07
ESRI, 2016. Spatial Framework for Hadoop [WWW Document]. GitHub. URL https://github.com/Esri/spatial-framework-for-hadoop (accessed 10.23.16).
Glick, T.B., Feng, W., Bertini, R.L., Figliozzi, M.A., 2015. Exploring Applications of Second-Generation Archived Transit Data for Estimating Performance Measures and Arterial Travel Speeds. Transportation Research Record: Journal of the Transportation Research Board 2538, 44–53. doi:10.3141/2538-06
Grone, B., Appiah, J., Rilett, L., 2011. A Methodology for Comparing Non-Intrusive Traffic Detectors under Different Operating Conditions, in: ICTIS 2011. American Society of Civil Engineers, pp. 1627–1633.
Henclewood, D., Rajiwade, S., 2015. Safety Pilot Model Deployment - Sample Data Environment Data Handbook. US Department of Transportation.
Herrera, J.C., Work, D.B., Herring, R., Ban, X. (Jeff), Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. Transportation Research Part C: Emerging Technologies 18, 568–583. doi:10.1016/j.trc.2009.10.006
Hunter, M., Wu, S., Kim, H., 2006. Practical Procedure to Collect Arterial Travel Time Data Using GPS-Instrumented Test Vehicles. Transportation Research Record: Journal of the Transportation Research Board 1978, 160–168. doi:10.3141/1978-21
Huntsinger, L.F., Ward, K., 2015. Using Mobile Phone Location Data to Develop External Trip Models. Transportation Research Record: Journal of the Transportation Research Board 2499, 25–32. doi:10.3141/2499-04
ITS JPO, 2016. CV Pilot Deployment Program - Connected Vehicle Applications [WWW Document]. Intelligent Transportation Systems. URL http://www.its.dot.gov/pilots/cv_pilot_apps.htm (accessed 10.23.16).
ITS JPO, 2015. Intelligent Transportation Systems - Connected Vehicle Basics [WWW Document]. Intelligent Transportation Systems. URL http://www.its.dot.gov/cv_basics/cv_basics_20qs.htm (accessed 10.17.16).
ITS JPO, 2014. Intelligent Transportation Systems - Model Deployment Technical Fact Sheet [WWW Document]. Intelligent Transportation Systems. URL http://www.its.dot.gov/factsheets/technical_fs_model_deployment.htm (accessed 10.17.16).
Kieu, L.M., Bhaskar, A., Chung, E., 2012. Bus and Car Travel Time on Urban Networks: Integrating Bluetooth and Bus Vehicle Identification Data, in: 25th ARRB Conference : Shaping the Future: Linking Policy, Research and Outcomes. Perth, WA, Australia, pp. 23–26.

Li, J.-Q., Zhou, K., Shladover, S., Skabardonis, A., 2013. Estimating Queue Length Under Connected Vehicle Technology. Transportation Research Record: Journal of the Transportation Research Board 2356, 17–22. doi:10.3141/2356-03

Liao, C.-F., 2014. Generating Reliable Freight Performance Measures with Truck GPS Data. Transportation Research Record: Journal of the Transportation Research Board 2410, 21–30. doi:10.3141/2410-03

Lin, D.-Y., Eluru, N., Waller, S., Bhat, C., 2008. Integration of Activity-Based Modeling and Dynamic Traffic Assignment. Transportation Research Record: Journal of the Transportation Research Board 2076, 52–61. doi:10.3141/2076-06

Liu, J., Khattak, A.J., 2016. Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles. Transportation Research Part C: Emerging Technologies 68, 83–100. doi:10.1016/j.trc.2016.03.009

Ma, X., Wang, Y., McCormack, E., Wang, Y., 2016. Understanding Freight Trip-Chaining Behavior Using a Spatial Data-Mining Approach with GPS Data. Transportation Research Record: Journal of the Transportation Research Board 2596, 44–54. doi:10.3141/2596-06

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies 24, 9–18. doi:10.1016/j.trc.2012.01.007

NHTSA, 2014. NHTSA V2V Communications Fact Sheet [WWW Document]. Safer Car. URL https://www.safercar.gov/v2v/index.html (accessed 10.17.16).

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P., 2014. Supporting large-scale travel surveys with smartphones – A practical approach. Transportation Research Part C: Emerging Technologies, Special Issue with Selected Papers from Transport Research Arena 43, Part 2, 212–221. doi:10.1016/j.trc.2013.11.005

Nkenyereye, L., Jang, J.-W., 2015. Addressing Big Data solution enabled Connected Vehicle services using Hadoop [WWW Document]. URL http://www.dbpia.co.kr/Journal/PDFView?id=NODE06235026 (accessed 6.21.16).

Oliveira, M., Vovsha, P., Wolf, J., Mitchell, M., 2014. Evaluation of Two Methods for Identifying Trip Purpose in GPS-Based Household Travel Surveys. Transportation Research Record: Journal of the Transportation Research Board 2405, 33–41. doi:10.3141/2405-05

Patel, A.B., Birla, M., Nair, U., 2012. Addressing big data problem using Hadoop and Map Reduce, in: 2012 Nirma University International Conference on Engineering (NUiCONE). Presented at the 2012 Nirma University International Conference on Engineering (NUiCONE), pp. 1–5. doi:10.1109/NUICONE.2012.6493198

Remias, S., Hainen, A., Day, C., Brennan, T., Li, H., Rivera-Hernandez, E., Sturdevant, J., Young, S., Bullock, D., 2013. Performance Characterization of Arterial Traffic

Flow with Probe Vehicle Data. Transportation Research Record: Journal of the Transportation Research Board 2380, 10–21. doi:10.3141/2380-02
Research Data Exchange, 2016. Safety Pilot Model Deployment Data [WWW Document]. Research Data Exchange. URL https://www.its-rde.net/index.php/data/explore-rde-data/10018-safety-pilot-model-deployment-data (accessed 10.23.16).
Schlaich, J., 2010. Analyzing Route Choice Behavior with Mobile Phone Trajectories. Transportation Research Record: Journal of the Transportation Research Board 2157, 78–85. doi:10.3141/2157-10
Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice, Bridging Research and Practice: A Synthesis of Best Practices in Travel Demand Modeling 41, 367–381. doi:10.1016/j.tra.2006.09.005
Talebpour, A., Mahmassani, H., Mete, F., Hamdar, S., 2014. Near-Crash Identification in a Connected Vehicle Environment. Transportation Research Record: Journal of the Transportation Research Board 2424, 20–28. doi:10.3141/2424-03
The PostgreSQL Global Development Group, 2016. PostgreSQL: Downloads [WWW Document]. PostgreSQL. URL https://www.postgresql.org/download/ (accessed 10.28.16).
The R Foundation, 2016. R: The R Project for Statistical Computing [WWW Document]. R. URL https://www.r-project.org/ (accessed 10.28.16).
Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. Transportation Research Part C: Emerging Technologies, Big Data in Transportation and Traffic Engineering 58, Part B, 162–177. doi:10.1016/j.trc.2015.04.022
USDOT, 2015. CV Pilot Deployment Program [WWW Document]. Intelligent Transportation Systems. URL http://www.its.dot.gov/pilots/wave1.htm (accessed 10.21.16).
USDOT, Transportation Planning Capacity Building Program, Federal Highway Administration, Federal Transit Administration, 2007. The Transportation Planning Process - Key Issues (No. FHWA-HEP-07-039). US Department of Transportation.
Vovsha, P., Freedman, J., Livshits, V., Sun, W., 2011. Design Features of Activity-Based Models in Practice. Transportation Research Record: Journal of the Transportation Research Board 2254, 19–27. doi:10.3141/2254-03
Wang, M.-H., Schrock, S.D., Broek, N.V., Mulinazzi, T., 2013. Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data. Int. J. ITS Res. 11, 76–86. doi:10.1007/s13177-013-0058-8