

11-6-2015

Development of Truck Route Choice Data Using Truck GPS

Mohammadreza Kamali

University of South Florida, mkamali@mail.usf.edu

Follow this and additional works at: <http://scholarcommons.usf.edu/etd>

 Part of the [Urban Studies and Planning Commons](#)

Scholar Commons Citation

Kamali, Mohammadreza, "Development of Truck Route Choice Data Using Truck GPS" (2015). *Graduate Theses and Dissertations*.
<http://scholarcommons.usf.edu/etd/5968>

This Thesis is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.

Development of Truck Route Choice Data Using Truck GPS

by

Mohammadreza Kamali

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Civil Engineering
Department of Civil and Environmental Engineering
College of Engineering
University of South Florida

Major Professor: Abdul R. Pinjari, Ph.D.
Pei-Sung Lin, Ph.D.
Seckin Ozkul, Ph.D.

Date of Approval:
October 29, 2015

Keywords: Probe Data, Map-matching, Route Generation, Route Choice, Route Variability

Copyright © 2015, Mohammadreza Kamali

DEDICATION

I dedicate this work to my family. My mom, Ameneh Chalahaghghi, whose constant love and sacrifice through all stages of my life have gotten me where I am now. My dad, Jafar Kamali, who has been not only my unwavering support, but also my hero and my best friend. And last but not least, my brother, Sepehr Kamali, who has never abandoned my side and always remained a true source of inspiration for me.

ACKNOWLEDGEMENTS

I express my gratitude to my advisor, Dr. Abdul Pinjari, from whom I have learned a bunch during my research. This thesis would not be possible without his insightful supervision and support. I would also like to thank the rest of my thesis committee, Dr. Seckin Ozkul and Dr. Pei-Sung Lin, for their constructive comments and encouragement.

I would like to acknowledge the financial support of the Florida Department of Transportation (FDOT) and FDOT District 4 during my graduate studies at USF. However, “the opinions, findings, and conclusions expressed in this publication are those of the author and not necessarily of the Florida Department of Transportation or Florida Department of Transportation District 4 or the U.S Department of Transportation”.

I want to specially thank two of our research team members, Akbar Bakhshi Zanjani whose true friendship and support helped me get through grad school, and Trang Luong whose dedication and enthusiasm inspired me. I would also like to thank the rest of the research team members: Tatok Raharjo, Spencer Wong, Bertho Agustin, and Sashikanth Gurram for their help and support when I needed it the most. I also want to thank officemates and friends: Petr, George, Lukai, Jorge, Robert and Melanie for all the fun we have had in the last couple of years.

Last but not least I want to thank my family whose love and support always inspired me during my studies.

TABLE OF CONTENTS

LIST OF TABLES	iv
LIST OF FIGURES	vi
ABSTRACT.....	x
CHAPTER 1 : INTRODUCTION	1
1.1 Background.....	1
1.2 Motivation.....	2
1.3 Objectives	4
1.3.1 First Part.....	4
1.3.1.1 Processing and Cleaning ATRI GPS Data.....	5
1.3.1.2 Procedure to Generate Routes from Raw GPS Data.....	6
1.3.1.3 Variability Measure	6
1.3.2 Second Part	6
1.3.2.1 Objective 1: Gather ATRI’s Truck-GPS Data.....	7
1.3.2.2 Objective 2: Identify & Separate Petroleum Tanker Trucks	8
1.3.2.3 Objective 3: Derive Trip Chains of Trucks.....	8
1.3.2.4 Objective 4: Derive Truck Travel Paths	8
1.4 Organization of the Thesis.....	8
CHAPTER 2 : LITERATURE REVIEW	10
2.1 Introduction.....	10
2.2 Previous Studies on Map-matching Methods	10
2.3 Previous Studies on Route Validation	11
2.4 Gaps in the Literature.....	12
CHAPTER 3 : DATA AND MEASUREMENT	14
3.1 Introduction.....	14
3.2 Characteristics of ATRI GPS Data	14
3.3 Anomalies in GPS Data	18
CHAPTER 4 : PROCEDURE TO GENERATE ROUTES FROM RAW GPS DATA	20
4.1 Introduction.....	20
4.2 Raw GPS Data to Trips.....	20
4.3 Map-matching Dataset Preparation.....	22
4.4 Map-matching	24
4.5 Route Generation	27

4.6 Route Feasibility and Validation	28
4.7 Route Variability Measure	30
4.8 Validation.....	33
CHAPTER 5 : CONCLUSION OF THE FIRST PART	41
5.1 Conclusions.....	41
5.2 Map-matching and Route Generation.....	41
5.3 Opportunities for Future Research.....	42
CHAPTER 6 : AN OVERVIEW OF THE DATA USED IN THE PROJECT	43
6.1 Introduction.....	43
6.2 ATRI GPS Data	43
6.2.1 Data Coverage.....	44
6.2.2 Separating Tanker Trucks from Other Trucks.....	47
6.3 Fuel Recipient Data.....	48
6.3.1 DOR Data.....	49
6.3.2 HERE Data.....	50
6.3.3 Comparison between DOR and HERE Data	50
CHAPTER 7 : DERIVING TRIP CHAINS FOR TANKER TRUCKS	52
7.1 Introduction.....	52
7.2 Algorithm Description	52
7.2.1 Validation Checks.....	54
7.3 Characteristics of Truck Trips Derived from ATRI Data.....	56
7.4 Identifying Trip Ends' Location Descriptions	59
7.5 Cleaning the Trip Dataset	63
7.6 Deriving Trip Chains	64
7.6.1 Procedure for Deriving Trip Chains	66
CHAPTER 8 : DERIVING TRIP ROUTES.....	76
8.1 Introduction.....	76
8.2 Map-matching Algorithm	76
8.3 Route Generation	77
8.4 Validation.....	77
CHAPTER 9 : CONCLUSION OF THE SECOND PART	82
9.1 Conclusion	82
9.2 Gathering ATRI's Truck GPS Data and Separating Petroleum Tanker Trucks	82
9.3 Derive Trip Chains of Trucks Originating at PEV	82
9.3.1 Modifying Trip Extraction Algorithm and Extracting Truck Trips.....	83
9.3.2 Identifying Characteristics of Truck Trips.....	83
9.3.3 Rectifying and Enriching Existing Data on Fuel Recipients	83
9.4 Deriving Trip Routes	84
9.5 Opportunities for Future Research.....	84

REFERENCES	86
APPENDIX A: POLYGONS AROUND PORT EVERGLADES FUEL TERMINALS	88
APPENDIX B: TRUCK TRIP CHARACTERISTICS	94
APPENDIX C: TRIP FILE CLEANING TASKS.....	98
C.1 Removing Non-Fuel Delivery Trips	98
C.2 Solving Trip Ends on the Road	99
C.3 Joining/Replacing Trip Ends on the Roadway.....	100
C.4 Joining Truck IDs from D1 Data.....	101

LIST OF TABLES

Table 2.1 Summary of literature on map-matching and route validation	12
Table 3.1 Cross-tabulation between largest ping rate and its corresponding spatial difference for GPS data with spot speed	16
Table 3.2 Cross-tabulation between largest ping rate and its corresponding spatial difference for GPS data without spot speed	18
Table 4.1 General trip information for 80 routes selected for feasibility and consistency checks	34
Table 4.2 Spatial gap vs. ping rate for selected 80 trips	36
Table 4.3 Route variation measurement results	37
Table 6.1 Attributes of the September 2014 and March 2015 GPS data	45
Table 6.2 Cross-tabulation of spatial gap against ping rate for September 2014 – D1 data	46
Table 6.3 Cross-tabulation of spatial gap against ping rate for September 2014 – D2 data	46
Table 6.4 Cross-tabulation of spatial gap against ping rate for March 2015 – D1 data	47
Table 6.5 Cross-tabulation of spatial gap against ping rate for March 2015 – D2 data	47
Table 7.1 General trip statistics for 14,598 extracted trips	54
Table 7.2 Summary of trip end location description of 100 trips selected for validation checks	55
Table 7.3 Land use description distribution for the trip origins and destinations of the total 14,598 trips	63
Table 7.4 Land use description distribution for the trip origins and destinations for 12,649 trips after cleaning the trip file	64

Table 7.5 Location distribution for starting and ending points of 807 trip chains that have visited PEV at least once	73
Table 7.6 Location distribution for starting and ending points of 513 trip chains that do not visit PEV	73
Table 8.1 General trip information for 50 routes selected for feasibility, consistency, and connectivity checks	79
Table C.1 Land use description distribution for the trip origins and destinations of the total 14,598 trips	103
Table C.2 Land use description distribution for the trip origins and destinations of 12,649 trips	105

LIST OF FIGURES

Figure 4.1 Example of a journey containing two trips.....	22
Figure 4.2 Map-matching algorithm.....	26
Figure 4.3 (a) All generated routes (left image), (b) All generated routes between an OD pair (right image).....	28
Figure 4.4 Route Length distribution of all 78,381 generated trips.....	28
Figure 4.5 Route length distribution for selected 80 trips	34
Figure 4.6 Consistency and feasibility check for one trip	37
Figure 4.7 Routes with and without time sampling	38
Figure 4.8 90th percentile and unique routes.....	39
Figure 6.1 An example of raw D1 data.....	44
Figure 6.2 An example of raw D2 data.....	44
Figure 6.3 Location of the terminals (yellow points) in PEV and their associated number (in red)	49
Figure 6.4 Terminal 1 (in yellow circle) and the red polygon used for GPS data extraction.....	49
Figure 6.5 Overlapping points from HERE (red) and DOR (blue).....	51
Figure 7.1 Trip length distribution for 100 validation trips.....	55
Figure 7.2 Trip time distribution for 100 validation trips.....	56
Figure 7.3 Trip speed distribution for 100 validation trips.....	56
Figure 7.4 12-county region in south Florida, sample of ATRI GPS points (red dots), and study boundary (large blue polygon).....	58

Figure 7.5 Trip length distribution of all 14,598 trips	58
Figure 7.6 Trip time distribution of all 14,598 trips	59
Figure 7.7 Average trip speed distribution of all 14,598 trips	59
Figure 7.8 Spatial gap distribution for non-matching trip pairs (N = 1,744 trip pairs)	68
Figure 7.9 Distribution of trip chains among number of trips (N = 1,320 trip chains)	69
Figure 7.10 Distribution of trip chains among number of trips (number of trips =< 5) (n=822 trip chains)	69
Figure 7.11 3-D histogram of spatial gap vs. temporal gap (N = 12,538 trip pairs).....	70
Figure 7.12 Trip chain length distribution (N=1,320 trip chains)	71
Figure 7.13 Trip chain time distribution (N = 1,320 trip chains)	71
Figure 7.14 Profile of starting time of the trip chains (N = 1,320 trip chains)	72
Figure 7.15 Profile of ending time of the trip chains (N = 1,320 trip chains)	72
Figure 7.16 Distribution of number of trips per trip chain for trip chains that visit PEV at least once (N = 807 trip chains).....	74
Figure 7.17 Distribution of number of trips per trip chain for trip chains that visit PEV at least once (N = 328 trip chains).....	74
Figure 7.18 Trip chain length distribution for trip chains that visit PEV at least once (N=807 trip chains)	75
Figure 7.19 Trip chain time distribution for trip chains that visit PEV at least once (N = 807 trip chains).....	75
Figure 8.1 Final 11,907 derived routes	79
Figure 8.2 Trip length distribution of 50 trips selected for validation checks.....	79
Figure 8.3 An example of following a route for feasibility and consistency check	81
Figure A.1 Terminal 1 (in yellow circle) and the red polygon used for GPS data extraction	88
Figure A.2 Terminal 2 (in yellow circle) and the red polygon used for GPS data extraction	89

Figure A.3 Terminal 5 (in yellow circle) and the red polygon used for GPS data extraction	89
Figure A.4 Terminal 7 (in yellow circle) and the red polygon used for GPS data extraction	90
Figure A.5 Terminal 11 (in yellow circle) and the red polygon used for GPS data extraction	90
Figure A.6 Terminal 12 (in yellow circle) and the red polygon used for GPS data extraction	91
Figure A.7 Terminal 13 (in yellow circle) and the red polygon used for GPS data extraction	91
Figure A.8 Terminal 14 (in yellow circle) and the red polygon used for GPS data extraction	92
Figure A.9 Terminal 15 (in yellow circle) and the red polygon used for GPS data extraction	92
Figure A.10 Terminals 3, 4, and 10 (in yellow circle) and the red polygon used for GPS data extraction	93
Figure A.11 Terminals 6 and 9 (in yellow circle) and the red polygon used for GPS data extraction	93
Figure B.1 Trip length distribution for all trips extracted from D2 data (N = 14,162 trips)	94
Figure B.2 Trip length distribution for all trips extracted from D1 data (N = 436 trips)	95
Figure B.3 Trip time distribution for all trips extracted from D2 data (N = 14,162 trips)	95
Figure B.4 Trip time distribution for all trips extracted from D1 data (N = 436 trips)	96
Figure B.5 Trip average speed distribution for all trips extracted from D2 data (N = 14,162 trips)	96
Figure B.6 Trip average speed distribution for all trips extracted from D1 data (N = 436 trips)	97
Figure C.1 Trip length distribution of 27 truck IDs removed from D1 and D2 data (September 2014 and March 2015 combined) (N = 1,879 trips)	102

Figure C.2 All 1,273 on-the-road trip destinations shown in blue dots	103
Figure C.3 All of the 378 on-the-road trip origins	104
Figure C.4 Distribution of spatial gap between consecutive truck IDs (N = 44 truck ID pairs).....	104
Figure C.5 Distribution of temporal gap between consecutive truck IDs (N = 44 truck ID pairs).....	105

ABSTRACT

Over the past few decades, the value and weight of freight shipments have grown steadily in both developed and developing countries. A recent statistic in the U.S. reveals that weight of shipments increased from 18,879 to 19,662 million tons between 2007 and 2012 (1). It is also expected that this amount will increase to 28,520 million tons by 2040 (1). It is worth mentioning that 67 percent of shipments are shipped by truck mode in 2012. The monetary value of freight is expected to escalate even faster than weight. This value is estimated to rise from US\$ 882 per ton in 2007 to US\$ 1,377 per ton in 2040. As a result, freight transportation management and modeling has aroused the interest of both public sector and groups of firms to improve the efficiency of the business operations. Traffic assignment plays a central role in the current freight modeling, and freight route analysis is of fundamental importance in understanding the truck flows explicitly.

In the first part of this thesis, large streams of truck-GPS data from the American Transportation Research Institute (ATRI) are cleaned, processed, and analyzed using easy to implement and practical procedures to study the diversity of observed truck routes between a given origin-destination (OD) pair. This is because, for any given OD pair, the analyst could observe and compare the route choices of a large number of trips, as opposed to observing only one or a few trips. Doing so helps in quantifying the number of different routes taken by trucks between an OD pair and paves the way for a systematic analysis of the “diversity” in route choices between any OD pair. This thesis develops methods to measure the diversity of routes

between a given OD pair and identifies unique routes used between the given OD pair. From a practical standpoint, such analysis of the diversity in observed route choices helps in improving the existing route choice set generation algorithms.

In the second part of the thesis, the methodologies developed in the first part are implemented in an FDOT sponsored project entitled “GPS Data for Truck-Route Choice Analysis of Port Everglades Petroleum Commodity Flows”. This project aims to use truck-GPS data from ATRI to derive petroleum tanker trucks’ travel path (or route) information, describing the routes that the tanker trucks take to travel from Port Everglades to their final delivery points.

CHAPTER 1 : INTRODUCTION

1.1 Background

Understanding freight movement and planning infrastructure policy responses to manage this movement is critical for a well-functioning economy. In the United States, data from the Freight Analysis Framework (FAF) shows that total freight movements are expected to grow from 19.7 billion tons in 2012 to 28.5 billion tons in 2040 (an overall growth of 45 percent or 1.3 percent annually). The value of freight is expected to grow at an even faster rate from \$17.4 Trillion in 2012 to \$39.3 Trillion in 2040 (a growth of 126 percent or three percent annually) (1). Further, the dominance of truck is expected to continue with around 70 percent of all commodities will be shipped (by weight) by trucks (1). All this freight movement by trucks contributes to congestion and causes extensive wear and tear to the infrastructure. Therefore, knowing how trucks travel and the paths they take will help design policy responses that allow for maintenance of infrastructure, improved reliability, and congestion mitigation.

One way to understand the paths trucks take is to make use of the data from advanced vehicle monitoring (AVM) systems that allow remote monitoring of truck fleets using Geographical Positioning Systems (GPS) technology-based Automatic Vehicle Location (AVL) systems. The availability of this GPS data provides the means to develop a deeper understanding of the paths trucks take when traveling over long distances. Using GPS data for studying truck paths imposes several challenges such as digesting large stream of GPS data points, converting GPS data into truck paths, and investigating truck paths in terms of similarity or variability. As

traditional methods for dealing with the above-mentioned challenges are either outdated or impractical, new methodologies have to be developed to deal with such challenges effectively.

1.2 Motivation

Freight transportation management and modeling has aroused the interest of both public sector and groups of firms to improve the efficiency of the freight business operations. Traffic assignment plays a central role in the current freight modeling, and freight route analysis is of fundamental importance in understanding the truck flows explicitly.

Following the more advances in freight transportation modeling, data collection and calibration processes have drawn a notable attention among planners and practitioners. In 2000, President Clinton announced the termination of the selective availability of GPS data, which significantly improved GPS accuracy and made it a viable option to monitor the freight travel behavior. Freight firms, consequently, use GPS to manage their equipment and capture truck data. Availability of such detailed data to the public sector has opened a new gate for freight route choice analysis. Improvements in data gathering and modeling capabilities have attenuated erroneous predictions in freight transportation modeling. Recent studies benefit from GPS information to explore and predict more accurate essential trip data elements. Little is known, however, about the accuracy of extracted route elements when using the less frequent GPS points.

The current study is an attempt to investigate truck route generation and variability analyses by using probe data drawn from GPS devices installed on trucks. Unprecedented partnership between private-sector truck data providers and freight carriers has opened up an opportunity to collect GPS data and provide it to public agencies in recent years. A joint venture between ATRI and the Federal Highway Administration (FHWA) is a good example of such

partnership that aim at developing a national system for monitoring freight performance measures (FPM) in the U.S. This FPM data contains GPS data collected from trucking companies that use GPS-based AVM technologies to keep track of their fleet. The FPS data contains large traces of GPS for trucks that travel on major corridors in the country (and Florida). This type of data provides professionals, freight stakeholders, and transportation researchers with an excellent opportunity to understand and measure freight behavior ranging from county-wide to nation-wide scale.

The first part of this thesis aims to introduce a general and practical framework for data cleaning, processing, and map-matching that enables both researchers and practitioners to deal with less frequent, but large number of data over a long period of time. The framework is quite distinct from previous studies in a couple of ways. First, the GPS data used for this study is less frequent as opposed to other similar studies. High frequency GPS data includes coordinates every one or two seconds while this study proposes a framework that enables us to generate routes for data with frequency of five to twenty minutes. Second, the number of data used in this study is significantly larger than other studies that focus on route generation methods. In terms of map-matched routes, particularly, this thesis utilizes a framework to generate the routes for more than 78,000 trips while similar efforts have usually been made for less than 50,000 trips. Third, the geographical scale of the data is large. The data includes the trucks that crossed the border or moved within the state of Florida for four months in 2010. As a result, the route generation problem needs to be solved on a statewide level, taking into account urban and rural geographies. Most route generation methods investigate the issue in an urban setting where the roadway network is dense. In this case, however, we mainly deal with routes that stretch throughout rural

areas, which have been overlooked in the current literature. The findings of this study are the building block for route choice generation and selection analyses.

The second part of the thesis is the implementation of the methodology developed in the first part within the context of a Florida Department of Transportation (FDOT) funded project. All the steps taken to complete the project will also be explained. FDOT District 4 is currently conducting the "Port Everglades Petroleum Commodity Flow Pilot Study". This is a proof-of-concept data collection pilot project jointly sponsored by the FHWA through its SHRP2 C20 program. The purpose of the project is to find an innovative methodology to collect and analyze petroleum flow data in and out of Port Everglades to better understand the supply-demand dynamics of the petroleum commodities in South Florida.

The information needed for the above-mentioned project includes the petroleum origin and destination data describing the supply side and demand side of the petroleum products, preferably at the Traffic Analysis Zone (TAZ) and Micro Analysis Zone (MAZ) level. Also needed is the truck travel path (or route choice) information of the petroleum tanker trucks for their travel between Port Everglades (PEV) and the final delivery points.

1.3 Objectives

1.3.1 First Part

The overarching goal of this thesis in the first part is to develop a methodology for generating and investigating trucks' route choices using GPS data. The proposed methodology should be an easy to implement and practical procedure that can digest large streams of GPS points with low frequency. The large number of GPS data points provides an unprecedented opportunity to develop rich observed truck route choice sets that can be useful for improving route choice set generation algorithms. The framework presented in the first part of this thesis is

meant to achieve two goals, 1) to generate truck routes from a large stream of GPS data; 2) to measure the variability of routes between an OD pair. Findings from this effort can help improve route choice set generation algorithms.

1.3.1.1 Processing and Cleaning ATRI GPS Data

The first part of the thesis is done based on more than 145 million raw truck GPS data points gathered by ATRI between March and June in 2010 for Florida. These GPS points correspond to a sample of trucks that traveled within, into, and out of state of Florida. Subsequently, an algorithm developed by Thakur et al. (2) is used to convert the GPS data into 1.2 million truck trips.

Considering the main objective of this thesis, characteristics of the data such as data type, data frequency, and data coverage have to be investigated so that the proper portion of data is selected for further analysis. This task involves measuring the spatial gap and temporal gap (hereafter, ping-rate) between consecutive GPS points and comparing spatial gap and ping-rate between different types of data. This is an important step because the main goal in the first part of this thesis is to design a technique that can convert GPS data into truck route on a roadway network. As a result, insights into nature of the GPS data define the path towards building such techniques.

Moreover, the coverage of the data has to be determined. This is done through observing the geographical distributions of truck trips. An algorithm developed by Thakur et al. (2) with some minor changes has been used to convert the raw GPS data into truck trips. Then, distributions of truck trips between OD pairs inside and outside of Florida are obtained to better understand the spatial characteristics of truck trips. This will help devise a process to select trips that are suitable for further analysis.

The last task in this objective is to detect possible anomalies in GPS data that can negatively impact the final results of this thesis. These anomalies exist due to systematic errors in GPS receivers or devices. Erroneous time stamps, incorrectly recorded latitudes or longitudes are examples of such anomalies. Therefore, a procedure has to be put in place to clean the GPS data (and resulted truck trips) from data anomalies.

1.3.1.2 Procedure to Generate Routes from Raw GPS Data

The first objective of this thesis is to develop a method for extracting the route taken by a truck on a roadway network using raw GPS data. This task consists of two steps, namely, map-matching and route generation. Quddus et al. (3) defines map-matching as a technique that uses a combination of GPS data and roadway network data to identify the correct link that has been traversed by the vehicle on the network. Map-matching is the first step towards generating the route taken by trucks on the network.

1.3.1.3 Variability Measure

This objective is geared towards creating a tool for measuring the variability of derived routes from GPS data. Being able to measure similarities or differences between truck routes on a network is an important step towards understanding truck route choice behavior. The large number of GPS data points provides an unprecedented opportunity to develop rich observed truck route choice sets that can be useful for improving route choice set generation algorithms. This objective is meant to measure the variability of routes generated between a given OD pair that can help improve route choice set generation algorithms.

1.3.2 Second Part

The second part of the thesis describes the implementation of the proposed methodology within the context of an FDOT District 4 project entitled “GPS Data for Truck-Route Choice

Analysis of Port Everglades Petroleum Commodity Flows”. This project aims to use truck-GPS data from ATRI to derive petroleum tanker trucks’ travel path (or route) information, describing the routes that the tanker trucks take to travel from Port Everglades to their final delivery points. To this end, following goals are investigated in detail.

1.3.2.1 Objective 1: Gather ATRI’s Truck-GPS Data

This task established a non-disclosure agreement (NDA) between ATRI and USF to protect the confidentiality of the GPS data that ATRI shared with USF. The agreement allowed for the aggregate results and data products from the research to be delivered. However, the agreement did not allow either the raw GPS data or individual GPS data points to be shared with anyone outside the research team at USF.

Once the NDA was in place, ATRI extracted and shared the relevant truck-GPS data with USF. This included eight-weeks of GPS data of trucks in the months of September 2014 and March 2015 for the 12-county region served by the Port Everglades –Miami-Dade, Broward, Palm Beach, Monroe, Martin, St. Lucie, Indian River, Okeechobee, Glades, Hendry, Lee, and Collier Counties. ATRI extracted and provide to USF raw GPS data on trucks originating in the Port Everglades (PEV) and traveling in the 12-county region.

In addition to the truck-GPS data, the following other data were needed for this work:

- 1) A shape file of TAZs or MAZs in the 12-county region,
- 2) A shape file of a detailed highway network in the 12-county region,
- 3) A shape file of the gas stations in the 12-county region, and
- 4) A shape file of PEV, identifying specific locations within the port where petroleum tanker trucks might originate from.

The research team relied on FDOT District 4 and their consulting team to obtain the information above.

1.3.2.2 Objective 2: Identify & Separate Petroleum Tanker Trucks

ATRI provided to USF raw GPS data on trucks originating at PEV and traveling in the 12-county region identified above. However, it was not necessary that all those trucks carry petroleum products. Therefore, this task developed simple rules or heuristics to identify and separate petroleum tanker trucks originating at PEV based on the land-uses (particularly gas terminals at PEV) of the locations visited by the trucks.

1.3.2.3 Objective 3: Derive Trip Chains of Trucks

The raw GPS data was converted into a database of truck trip chains. The algorithms developed previously by Thakur et al. (2) were utilized in this task. However, the algorithms were developed primarily for the purpose of deriving individual trips, as opposed to deriving trip chains. As part of this project, such algorithms were modified to derive trip chains from the raw-GPS data.

1.3.2.4 Objective 4: Derive Truck Travel Paths

This task derived the travel paths for tanker trucks traveling between PEV and gas stations. For each truck trip between PEV and a gas station, the travel route was derived in the form of a GIS shapefile.

1.4 Organization of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 provides a review of the literature on map-matching methods and route variability measurements. Chapter 3 describes the GPS data used for developing the methodology in the first part of the thesis. Chapter 4 defines the algorithms for data preparation, route generation, and route variability measurement. Chapter

5 presents the conclusion of the first part of the thesis and recommendations for future research. Chapter 6 is the beginning of the second part of the thesis and presents an overview of the data used in the FDOT project. Chapter 7 describes characteristics of tanker truck trips and steps taken to derive their trip chains. Chapter 8 presents implementation of the methodology developed in the first part of the thesis to derive tanker trucks' routes. Chapter 9 summarizes the findings in the second part of the thesis and identifies opportunities for future research

CHAPTER 2 : LITERATURE REVIEW

2.1 Introduction

This section of the study discusses the literature on how the roadway network performs for trucks, followed by a review on the studies that review both map-matching and validation processes along with route variability. This research does not aim to introduce a new method, rather to borrow efficient solutions to build the desired algorithm for truck route analysis. Therefore, the review of both trip selection and map-matching processes are essential.

2.2 Previous Studies on Map-matching Methods

Map-matching technique may date back to 1996, in which Kim et al. (4) introduced a simple algorithm that mapped the GPS points to the closest node or shape point in the network. Ever since, a mushrooming literature has evolved varying from simple methods to complex mathematical techniques. Ochieng et al. (5) discussed comprehensively the pros and cons of each common method. From the methodology side, developed algorithms fall into four major categories, namely, geometric based, geometric and topologic based, probabilistic based, and advanced algorithms. The geometric based algorithm uses the distance of either point-to-curve or curve-to-curve, or the angle of curve-to-curve for map-matching. While the geometric and topologic based algorithm diminishes the incorrect candidate points by considering the connectivity of the network elements. Ochieng et al. (5) pioneered the probabilistic based algorithm that uses a confidence region defined around each GPS point. Then, the confidence region is imposed on the road network to understand the road segments. The choosing of

appropriate segments, finally, is carried out by closeness, connectivity, and heading criteria. Following the Kalman filter method, several complex algorithms such as hybrid Bayesian network, fuzzy logical model, Belief function, and Dempster-Shafer theory of evidence have growingly emerged in the field of traffic network analysis. These methods shaped the kernel of advanced map-matching algorithms. Table 2.1 summarizes previous efforts for map-matching analysis with a wide diversity in analysis methods.

2.3 Previous Studies on Route Validation

From the validation side, studies might be divided into three major categories, namely, site based methods, comparison methods, and analytical methods. In site based methods a field test is implemented. A vehicle carrying a probe system then traverses a pre-chosen route. The points from the probe system are map-matched using the algorithm and finally, the pre-chosen route and the produced route are compared. Ochieng et al. (5), Yang et al. (6), and Dhakar (7) have effectively implemented site based methods for route validation. The advantage of site-based approach is that it truly measures the accuracy of the map-matching algorithm. On the other hand, the involved costs limit its implementation. Xu et al. (8) and Chen et al. (9) have utilized comparison methods to validate the map-matching algorithm. The former compares the results with an already validated map-matched data while the latter proposes to time-sample the data and compare the results with the original data. While comparison methods overcome some of the difficulties of site based methods, they demand for either larger datasets or already validated data. Feasibility and continuity analysis done by Hess et al. (10), and correct road matching ratio implemented by Jagadeesh et al. (11) are considered analytical methods that are successfully implemented. Analytical methods are more frugal in terms of cost of

implementation but they still need an already validated dataset serving as the base of comparison.

Table 2.1 Summary of literature on map-matching and route validation

Author	Year	Place	Map Match	Frequency	Validation Method	Accuracy
Miwa	2012	Japan	nearest link	90 sec	ARR and IARR	10-85%
YANG	2005	South Korea	nearest link	2-5 min	Field test	100%
Chen	2013	China	MDP-MM	2 sec	ARP	92%
Jagadeesh	2004	Singapore	Fuzzy Logic Model	0-30 m	Route matching ratio	96%
Dhakar	2012	N/A	N/A	N/A	Field test	78%
Hess	2015	U.K	N/A	N/A	Link removal	91%
Wang	2011	U.S.	Fuzzy Logic Model	5-15 min	N/A	N/A
Rahmani	2012	Sweden	N/A	30-180 s	N/A	N/A

2.4 Gaps in the Literature

The current literature of map-matching algorithms has certain gaps that preclude the author from applying them on the less frequent GPS point data. First, as shown in Table 2.1, previous empirical analyses have proposed methods that are valid only for high frequency GPS points. Consequently, employing these methods where consecutive temporal gap between GPS points is more than 10 minutes may demolish the accuracy of results. In the truck route choice analysis, the extracted data from in-vehicle GPS devices presents less frequent GPS points. Hence, building the results on the previous map-matching analysis hinders a fine-grained analysis of truck movements on the road network. Second, most of the map-matching techniques that are summarized here deal with very dense urban networks and in turn, are very complicated to match the GPS points to the right links as precisely as possible. The data used in this thesis, on the other hand, belongs to long-haul trucks that usually traverse major highways and arterials as

they travel, and do not appear in dense urban areas for the most part of their trip. Hence, the fact that trucks usually appear on major highways demands for a less complicated map-matching approach.

To the best of the author's knowledge, there are a few studies on truck route generation and analysis that propose a practical algorithm for map-matching and validation of large streams of GPS data. The current research, therefore, is an attempt to bridge the above-mentioned gaps by shedding some light on how to turn truck GPS data into truck trip routes so that they can be used to understand truck movement behavior. This thesis proposes a simple, yet effective algorithm for turning truck GPS data into truck trip routes and their respective links. The main idea of this approach is rooted in the nearest link and second nearest link algorithm introduced by Yang in 2005.

CHAPTER 3 : DATA AND MEASUREMENT

3.1 Introduction

The first part of the thesis is done based on more than 145 million raw truck GPS data points gathered by ATRI between March and June in 2010 for Florida. Characteristics of the data such as data type, data frequency, and data coverage are investigated in this chapter. This task involves measuring the spatial gap and ping-rate between consecutive GPS points and comparing spatial gap and ping-rate between different types of data. This is an important step because it leads to design a technique that can convert GPS data into truck route on the roadway network.

3.2 Characteristics of ATRI GPS Data

ATRI's truck GPS data represent a sample of truck flows within, coming into, and going out of Florida. This sample is not a census of all trucks traveling in the state. Also, it is unknown what proportion of heavy truck flows in the state is represented by this data sample. To address this question, truck traffic flows implied by one-week of ATRI's truck GPS data were compared with truck counts data from more than 200 Telemetered Traffic Monitoring Sites (TTMS) in the state. The results from this analysis suggest that, at an aggregate level, the ATRI data provides 10.1 percent coverage of heavy truck flows observed in Florida. When the coverage was examined separately for different highway facilities (based on functional classification), the results suggest that the data provide a representative coverage of truck flows through different types of highway facilities in the state (6).

The final data includes a unique ID number assigned to each truck (hereafter truck ID), spatial characteristics such as latitude and longitude of the GPS points, and temporal characteristics such as date and time. “Unique Truck ID” is a random number assigned to each vehicle and cannot be used to trace back the actual vehicle from the trucking company. Truck ID however, can be used to distinguish between different trucks in the database for trip measurement purposes. A subset of the data has instantaneous speed of the corresponding truck for each GPS record (henceforth called data with spot speed) and the remaining portion of the data does not have such information (henceforth called data without spot speed). The spatial and temporal characteristics of the data play an important role in the accuracy and feasibility of the route generation. The frequency and spatial gap have a positive correlation with the accuracy of the final generated routes. Higher ping-rates in the data result in routes that are more accurate. However, it should be kept in mind that in some cases while the ping-rate is small, the spatial gap between two consecutive GPS points can be quite large resulting in errors. Therefore it is necessary to consider the spatial gap between consecutive GPS points to increase the accuracy of data. While being mindful of these spatial and temporal gaps, it is also necessary to understand that the feasibility of truck route generation is dependent on the amount of GPS data available for use. Therefore, selecting a sufficient number of GPS observations while minimizing the spatial and temporal gaps is critical to obtaining a meaningful dataset.

The goal of obtaining a meaningful dataset is achieved by a two-step process. First the data is compared with and without spot speed. Tables 3.1 and 3.2 show, respectively, the cross-tabulation of the data by spatial gap and ping rate with and without spot speed. Comparing the two tables reveal that data without spot speed is coarser than data with spot speed. While Table

3.1 shows that the 25 percent of observations with spot speeds have a ping-rate of 15 minutes or less and spatial gap of 15 miles or less.

Table 3.1 Cross-tabulation between largest ping rate and its corresponding spatial difference for GPS data with spot speed

Spatial gap (miles) \ Ping-rate (minutes)								
	< 1	1-5	5-15	15-20	20-25	25-30	30 <	Sum
< 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1-2	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%
2-5	2.7%	14.2%	0.3%	0.0%	0.0%	0.0%	0.0%	17.2%
5-14	4.3%	11.2%	3.6%	0.0%	0.0%	0.0%	0.0%	19.2%
14-15	0.8%	4.1%	24.5%	3.9%	0.0%	0.0%	0.0%	33.2%
15-20	4.5%	0.7%	1.6%	0.1%	0.0%	0.0%	0.0%	6.8%
20-25	3.4%	0.1%	0.3%	0.1%	0.0%	0.1%	0.0%	4.0%
25-30	2.4%	0.2%	0.4%	0.1%	0.1%	0.1%	0.1%	3.4%
30-45	2.1%	0.1%	0.2%	0.0%	0.1%	0.1%	0.1%	2.6%
45 <	12.0%	0.2%	0.3%	0.0%	0.0%	0.0%	0.5%	13.0%
Sum	32.6%	31.0%	31.0%	4.2%	0.2%	0.2%	0.6%	100.0%

Table 3.2 shows that 44 percent of observations without spot speeds have ping-rates greater than 45 minutes and a spatial gap greater than 30 miles. Such large spatial gaps and ping rates impose practical difficulties on route generation efforts. The main problem with large spatial gaps and ping rates is that the location of the truck is unknown between two consecutive GPS points. Additionally, there is no other source of information that can help identify the location of the truck during large spatial gaps (or ping rates). For example, there is no other information on trucking companies, the usual routes their fleet take, or type of commodities that they carry. Therefore, large scale simplifying assumptions have to be made regarding the route choice of the trucks in order to generate routes. This will result in generated routes that can significantly be different from the real routes taken by those trucks. Therefore, it is better not to use data without spot speed and select data with spot speed that is more frequent for further analysis.

Next step is to impose some spatial gap and ping rate limitations on data with spot speed in order to select the final portion of data for route generation. Even though data with spot speed is more frequent than data without spot speed, there are some rare streams of GPS points with spot speed that have large spatial gaps or ping rates. Therefore, such streams of data must be removed while a significant portion of data is remained in order to make meaningful analysis of generated routes in future. To this end, observations of relationships between spatial gap and ping rate in data with spot speed revealed that maximum spatial gap of 20 miles and maximum ping rate of 20 minutes is ideal. This means that when GPS points of a trip is observed if the largest ping rate amongst those GPS points is less than 20 minutes and the spatial gap corresponding to the largest ping rate is less than 20 miles, then that trip and its GPS points are kept for future analysis. To save more data in this process, those trips whose largest ping rate is greater than 20 minutes but the corresponding spatial gap is less than 5 miles are also kept. This is because in route generation the spatial gap between consecutive GPS points matter the most. Therefore, streams of GPS points that have small spatial gaps must be retained regardless of their corresponding ping rates.

To recap, the final data includes two portions of data with spot speed: (1) GPS points with a spatial gap of less than 20 miles and a ping-rate of less than 20 minutes; and (2) GPS points with a ping-rate of more than 20 minutes but with a spatial gap of less than 5 miles. These criteria result in more than 97 percent of the data with spot speed being retained which is an acceptable amount for further analysis.

Table 3.2 Cross-tabulation between largest ping rate and its corresponding spatial difference for GPS data without spot speed

Spatial gap (miles) \ Ping-rate (minutes)		Spatial gap (miles)							Sum
		< 1	1-5	5-15	15-20	20-25	25-30	30 <	
< 1		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1-2		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
2-5		0.2%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%
5-14		0.3%	3.3%	0.6%	0.0%	0.0%	0.0%	0.0%	4.2%
14-15		0.0%	0.4%	0.2%	0.0%	0.0%	0.0%	0.0%	0.6%
15-20		0.1%	1.4%	1.7%	0.2%	0.0%	0.0%	0.0%	3.5%
20-25		0.1%	0.8%	1.8%	0.2%	0.1%	0.0%	0.0%	3.1%
25-30		0.1%	0.5%	1.9%	0.5%	0.3%	0.2%	0.1%	3.6%
30-45		0.3%	0.6%	3.2%	1.8%	1.5%	1.3%	1.4%	10.1%
45 <		18.1%	1.0%	3.7%	1.9%	2.4%	3.3%	44.0%	74.3%
Sum		19.2%	8.3%	13.2%	4.7%	4.3%	4.7%	45.5%	100.0%

3.3 Anomalies in GPS Data

Even though GPS data is usually of high quality in terms of consistency and accuracy, some rare anomalies can still be found in stream of GPS data. Such anomalies are inevitable due to systematic errors of GPS satellites and GPS receiver devices. It is important to detected and properly handle these anomalies to produce valuable results in future steps.

There are two main issues that are found during this research in stream of GPS data that can be problematic for route generation practices. First, there might be a loss of data signal for a period of time during a trip. This means that for a considerable amount of distance and time during a trip there are no GPS records in the data. This is a problem with regard to route generation because it is not clear what route alternatives have been taken by the truck during the loss of signal. Therefore, any estimation during this time interval will impose a significant error on the final predicted route for that trip. Consequently, it is reasonable to remove such trips from the dataset in order to avoid problematic trips. Second, some consecutive GPS records show very high average speeds. For example, the average speed between two consecutive GPS points is

more than 100 mph. It is common knowledge that trucks usually travel around 60 mph and therefore, very high average speeds during the trip is not reasonable. The reason to observing such high average speeds is systematic errors in GPS systems that cause wrong records of time stamps, latitudes, or longitudes. Having trips with irrational average speeds will also impose the danger of wrong route estimation in future analysis. As a result such trips should also be removed from the data set.

To protect the final dataset from such anomalies conditions (b) and (c) are added to Stage 3 in Section 4.3 of Chapter 4. These conditions remove trips that have unreasonably large spatial gaps or average speeds between consecutive GPS points.

CHAPTER 4 : PROCEDURE TO GENERATE ROUTES FROM RAW GPS DATA

4.1 Introduction

In order to investigate truck route generation and variability a two-phase framework was developed. The first phase predominantly was dedicated to data processing, data preparation, and map-matching. Map-matching is a process during which GPS points are snapped to their correct links on a roadway network. Before this step, GPS data has to be converted into trips, then processed and be ready for map-matching. Then route generation procedure is implemented to generate routes from map-matched GPS points. In the second phase, the framework for the variability of routes between OD pairs is laid out. A measurement is introduced to identify different routes between a given OD pair in order to better understand truckers' route choice behavior.

4.2 Raw GPS Data to Trips

As a first step, the raw GPS data needed to be converted to truck trips in order to be ready for route generation. The process is summarized below (7):

- 1) Identify stops based on spatial movement and speed between consecutive GPS points (<5mph)
- 2) Derive a preliminary set of trips based on a minimum dwell-time buffer of 5 min (eliminate stops of duration < 5 min)
- 3) Eliminate rest stops
 - a) Used a rest-areas land-use file (very useful but not exhaustive of all rest areas)

- b) Eliminated stops in close proximity of interstates (< 800 ft.)
- c) Join consecutive trips ending and beginning at rest stops
- 4) Find circular trips (with ratio of air distance to network distance < 0.7)
- 5) Break circular trips into shorter (valid) trips by allowing smaller stop dwell-time buffers at the destinations (redo Steps 3 and 4)
- 6) Join insignificant (< 1 mile) trips to a preceding long trip or eliminate them

The procedure above has been developed and discussed by Thakur et al. (2) with a few changes. A trip in this thesis is defined as a displacement between a starting point and a stopping point. That means a journey with multiple stops is broken into multiple trips. Suppose a truck travels between origin A and destination B (Figure 4.1). Suppose that the truck stops at C between A and B for 30 minutes to make a small delivery. Therefore, the journey between A and B is two trips, one between A and C and another between C and B. Considering the route choice behavior of the truck, the route that has been taken in the trip from A to C affects the route choice between C and B. Consequently, a correct interpretation cannot be made regarding the truck's route choice behavior from A to B if C is disregarded. Subsequently, the two trips discussed here should be considered individually exclusive in order to correctly understand the route choice behavior. Moreover, observations that have been done during this research showed that the journey between A and B is not necessarily the shortest path because the truck had to stop at an intermediate point (i.e., C in this example). As a result the route taken by the truck between A and B is counter-intuitive. This issue is even more complex when there is no other source of information on the decision maker's side (i.e., truckers). Therefore, journeys that include multiple stops have to be broken into trips. That is why a minimum dwell-time of 5 minutes is used in the algorithm above.

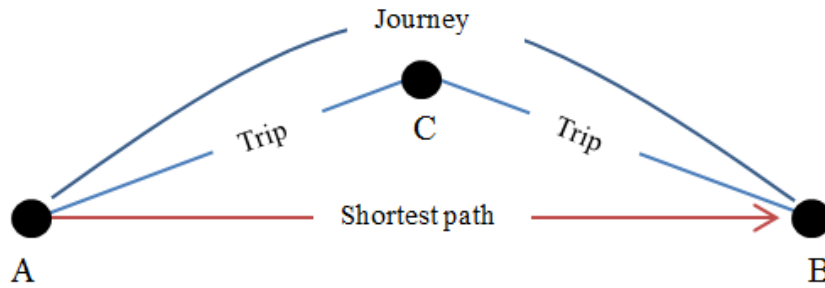


Figure 4.1 Example of a journey containing two trips

Having the GPS data converted to truck trips, some data filtering is needed before map-matching and route generation. This data filtering is aimed at removing GPS points that are difficult to map-match, or are not of high value when route generation procedure is implemented. Removing such GPS points is an important step toward developing an efficient route generation and map-matching procedure because this thesis deals with big GPS data. Technical competency is always of high importance when dealing with big data because the data size can cause very time-consuming processes. Therefore, it is imperative to reduce the size of the data to keep the processes efficient.

4.3 Map-matching Dataset Preparation

Following the conversion of raw GPS data to trips, the data was prepared further to obtain a dataset for map-matching. The stages in this process are summarized below.

- 1) Stage 1: Select trips that have all the following criteria:
 - a) Both ends in FL since the available network (i.e., Navteq) data covered only Florida;
 - b) Don't start and end in the same TAZ (traffic analysis zone) since we are interested in capturing the route variability between different origin-destination pairs;
 - c) $\frac{\text{direct OD distance}}{\text{trip length}} > 0.7$ to avoid circuitous routes that are typically undertaken by short haul trucks and is not of interest in this study; and

- d) Belonging to OD pairs that have equal to or greater than 20 trips to avoid low trip frequencies as they don't have variability of routes for a given OD pair.
 - e) Length > 5 miles to exclude urban/short-haul trips.
- 2) Stage 2: Get the GPS points belonging to the output of stage 1
- 3) Stage 3: Keep the trips and corresponding GPS points that satisfy the following conditions. The first two criteria below are explained in Sections 3.2 and 3.3 of Chapter 3. The third criterion eliminates cases where there are very small (or even zero) ping rates while the corresponding spatial gap is not zero. Such cases happen due to systematic errors in GPS receiver devices.
- a) Ping rate criteria:
 - i. maximum ping rate ≤ 20 minutes and corresponding spatial gap ≤ 20 miles, or
 - ii. Maximum ping rate > 20 minutes and corresponding spatial gap ≤ 5 miles
 - b) spatial gap criteria : maximum spatial gap ≤ 20 miles
 - c) average velocity criteria: maximum average velocity (i.e., $\frac{\text{spatial gap}}{\text{ping rate}}$) between consecutive GPS points ≤ 100 mph
- 4) Stage 4: Time sample GPS points of each trip every 5 minutes to reduce computation costs.
- 5) Stage 5: Remove GPS points within one mile radius of origin/destination for each trip. Doing so eliminates wrong route estimations near points of origin/destination. A lot of times the network is not fine enough within the one mile buffer of origin/destination and that leads to loops (irrational circular maneuvers) ingenerated routes close to trip ends.

Removing GPS points within one-mile buffer around origin/destination helps avoid such loops. This step also removes origin and destination GPS points. These points are later added at the last step of map-matching algorithm.

- 6) Stage 6: Remove GPS points with spot speed < 20 mph. This helps eliminate situations when the truck reduced its speed to stop during the trip. Such stops would cause detours in the generated routes which would have led to false interpretations of route variability.
- 7) Stage 7: Remove trips that have less than 3 GPS points. Such trips were removed to avoid false route generation that might have resulted from lack of GPS data.

1,583,164 trips (corresponding to 53,185,413 GPS points with spot speed) existed in the dataset before implementing the data preparation stages. 84,236 trips (corresponding to 725,483 GPS points with spot speed) were retained after implementing all the data preparation stages.

4.4 Map-matching

The next step is to apply the map-matching algorithm to the 84,236 trips from Step 2. This algorithm is a modified version of an algorithm introduced by Yang et al. (6). The map-matching algorithm is as follows:

- 1) Step 1: Find the closest and second closest link to each GPS point. D_1 and D_2 denote the distance from each GPS point to closest link and second closest link, respectively.
- 2) Step 2: If $D_1 > 1000$ ft. then remove the GPS point. GPS points that have no links within their 1000 ft. buffer are very difficult to map-match. This step eliminates such GPS points to avoid matching them to the wrong link.
- 3) Step 3: If $\frac{D_2}{D_1} > 2$ then go to Step 4, else go to Step 5.
- 4) Step 4: If $D_1 + D_2 > 35$ ft. then match the GPS point to the closest link. Otherwise, remove the GPS point. This step has been implemented to avoid matching GPS points to

the wrong link at interchanges or near ramps. Since links are very close to each other at such places, $D1$ and $D2$ might be smaller than GPS maximum accuracy that can lead to matching the GPS point to a wrong link. Therefore, there should be a lower bound on $D1+D2$ to make sure $D1+D2$ is greater than twice of GPS maximum accuracy. This lower bound has set to be 35 ft. because GPS maximum accuracy is 5 meters (16.4 ft.) according to Department of Defense report (12).

- 5) Step 5: Make a 65 ft. buffer around each GPS point that did not satisfy the ratio in “Step 3”. If there is only one intersection node falling in that buffer, then match the point to the intersection. Otherwise, remove the GPS point. This step deals with situations where a GPS point is close to an intersection. If the GPS point is near an intersection and only one intersection node falls in the 65 ft. buffer then the GPS point is matched to the intersection node. This is because some intersections have more than one node in Navteq. Consequently, two or more nodes might fall inside the 65 ft. buffer around a GPS point. Since it is difficult to decide to which node the GPS point should be matched to, it was decided to remove GPS points that have two or more intersection nodes falling inside their buffers.
- 6) Step 6: Add the origin and destination GPS points to the data for each trip.
- 7) Step 7: Remove any trip that has less than 5 GPS points. Some trips lose most of their GPS points after the map-matching algorithm is implemented. Therefore, generating the routes for such trips will impose high chances of errors. To avoid such routes, trips with less than 5 GPS points are removed.

Figure 4.2 shows the algorithm for the map-matching process. After the map-matching process was implemented the dataset had 78,381 trips for which routes were generated. This

map-matching algorithm provides a good balance on the tradeoff between accuracy of results and the relative size of the data. Most map-matching methods that result in very high accuracy outputs utilize complicated algorithms that are costly in terms of replication and implementation. Furthermore, such algorithms are not tested against large GPS datasets. In addition, such complicated algorithms are not available in the public domain making implementation difficult. The proposed method in this study on the other hand, benefits from a much less complicated algorithm that can easily handle a large GPS dataset while maintaining a satisfactory level of accuracy. Equally important, it can be implemented using widely used software packages such as ArcMap thereby helping reach a wider audience which results in better data being available to all.

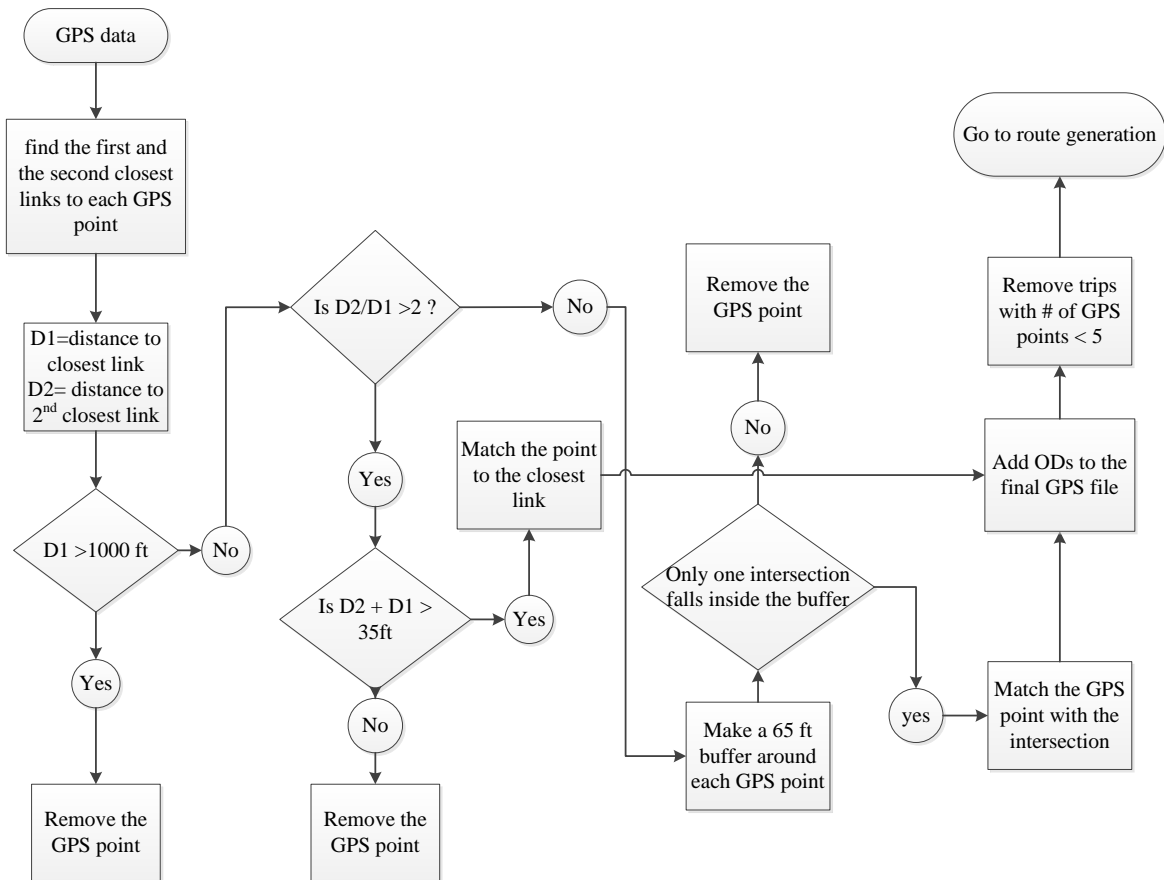


Figure 4.2 Map-matching algorithm

4.5 Route Generation

Due to the infrequent nature of the GPS data in this study, the map-matching algorithm detailed in Section 4.4 does not capture all links in a trip. Consequently, missing links need to be found so that a route can be generated for each trip. To this end, ArcMap 10.3 Network Analyst extension was employed to generate the routes for map-matched GPS points. Network Analyst utilizes a modified version of Dijkstra's algorithm to find shortest paths between two points. For each trip, the shortest path between consecutive GPS points was found based on minimizing travel time.

The final output of route generation is a GIS shapefile in which each feature is a network link that contains network information as well as trip information. Figure 4.3 (a) shows an overall view of the generated routes for 78,381 trips that belong to 2,237 OD pairs in Florida. Figure 4.3 (b) is an example of generated routes for one specific OD pair with 218 trips. In this example the origin TAZ is in Polk County (in central Florida) and the destination TAZ is in Miami-Dade County (in south east of Florida). Figure 4.4 shows the route length distribution for 78,381 trips. The resulting distribution is intuitive; there are few trips whose lengths are greater than 500 miles because longer trips usually stretch out of Florida. In addition, if a truck stops more than 5 minutes during its trip, that stop is called a destination resulting in breaking the trip. This reduces the probability of capturing trips longer than 500 miles. Trips that are 5 miles or shorter do not exist in the final dataset since such trips were eliminated during the procedure in Section 4.3.

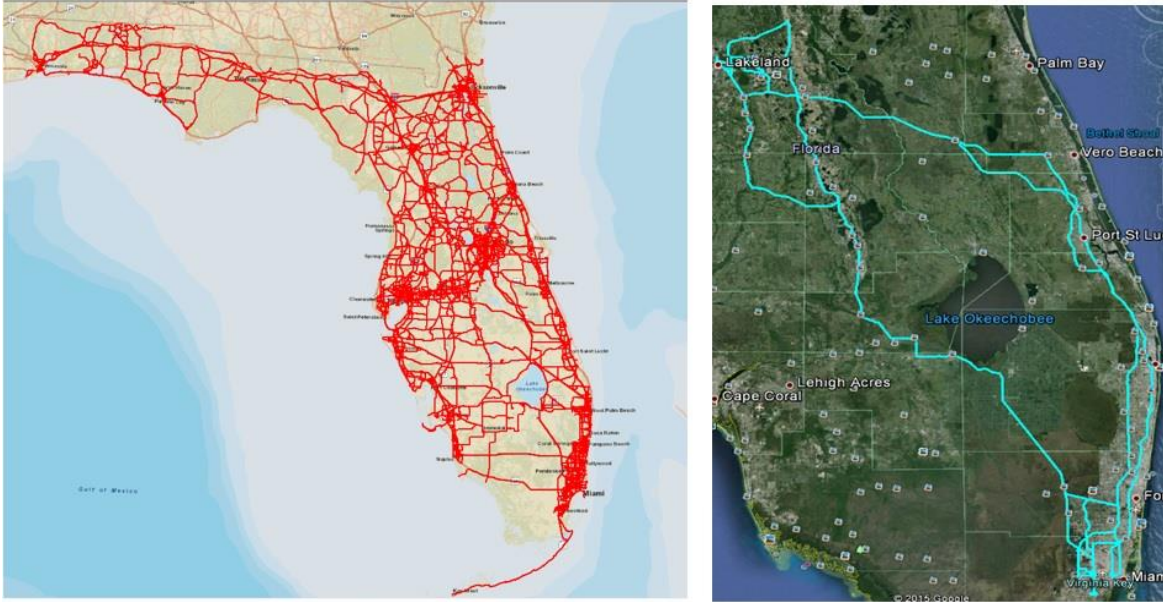


Figure 4.3 (a) All generated routes (left image), (b) All generated routes between an OD pair (right image)

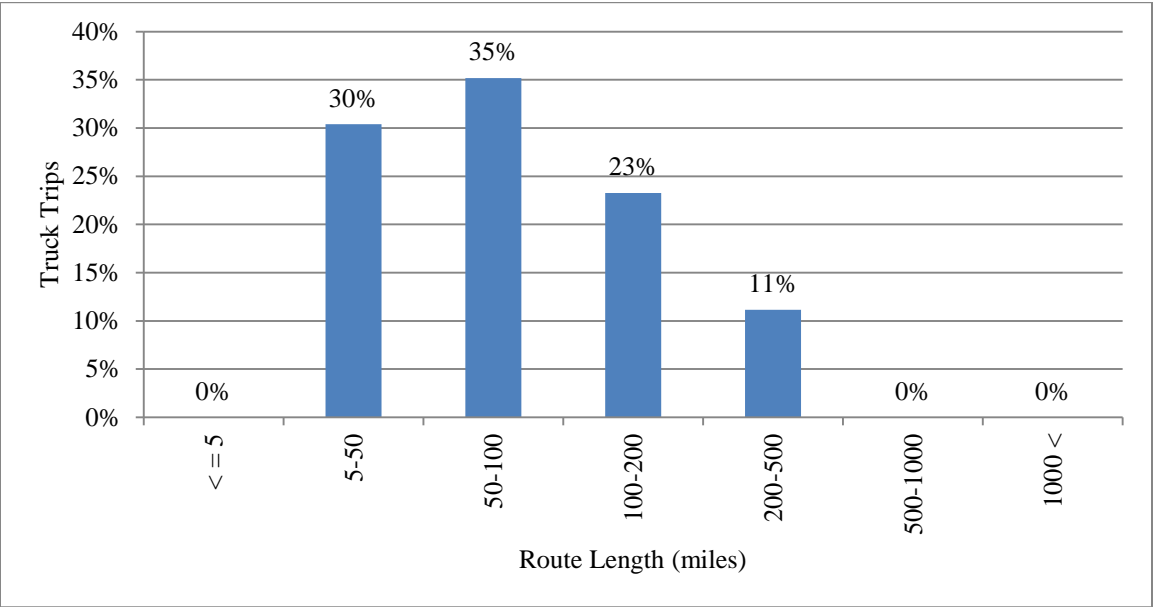


Figure 4.4 Route Length distribution of all 78,381 generated trips

4.6 Route Feasibility and Validation

The generated routes were validated in terms of feasibility and consistency. Routes are consistent if:

- 1) the direction of the travel is consistent throughout the entire route
- 2) there are no loops throughout the entire route

And feasible if and only if:

- 3) there are no impossible maneuvers throughout the entire route (impossible maneuvers such as jumping off a bridge)

80 trips were selected and then followed on Google Earth for validation checks. Table 4.1 lists all the 80 trips that were selected for feasibility and consistency checks with their corresponding trip time and trip length information. The last column from left in this table illustrates the status of each trip with regard to feasibility and consistency checks. As can be observed, all the 80 routes are marked as “Ok” which means they are all feasible and consistent. Figure 4.5 shows the route length distribution for these 80 trips. Two in five trips are between 200 and 500 miles. This is because such trips have a higher chance of inconsistency or infeasibility. Table 4.2 illustrates the cross-tabulation of the data by spatial gap and ping rate for the 80 trips. This cross-tabulation shows that the 80 trips are a good representative of the population in terms of ping-rate and spatial gap.

Consistency and feasibility checks are done simultaneously when the route is followed on Google Earth. To check the consistency, each generated route was compared to the route from Google Earth to determine if the generated route shows the same direction through the entire trip. For example if the truck has to take the north bound direction on the highway to get to the destination, the generated route should show that the truck has maintained that direction through the entire trip.

For feasibility check, each trip is observed at interchanges or overpasses or ramp junctions to see if the generated route shows any impossible maneuvers at such locations. Figure

4.6 shows an example of consistency and feasibility check for one trip. The yellow arrow shows the direction of travel depicted by the generated route. It is consistent through the entire trip. Moreover, there are no impossible maneuvers at interchanges or overpasses, meaning that the route is feasible. All 80 routes passed the consistency and feasibility checks.

Another validation concern was to verify if time-sampling the GPS data changes the original routes of trips. To examine this issue 45 randomly chosen trips were map-matched without time-sampling. Next, the same 45 trips were map-matched using time-sampling. Routes for both sets of trips were generated and then compared. It was found that generated routes for both sets of trips were similar. This shows that time-sampling the GPS data at a 5-minute rate can reduce the computation time for map-matching while not damaging the original map of a route. Figure 4.7 shows an example comparing routes before time-sampling and after time-sampling. The routes do not change by implementing time-sampling.

4.7 Route Variability Measure

Generating truck route choice sets is the first step towards building truck route choice models and the first five steps detail a process that can be an improvement to generating truck route choice sets. There are three main approaches for generating choice sets (11). The first approach is modelling the membership of each route alternative in the final choice set explicitly. This approach is too costly in terms of computation complexity and therefore, cannot be used even for medium sized problems. The second approach which is based on heuristic approximations of the explicit choice set models, is tricky to implement. This approach is based on the assumption that the universal choice set is known to the observer. The third approach is based on establishing the master set for all route alternatives and then reducing the master set for

each individual route to obtain the individual choice set. Attractiveness, plausibility and route similarity are factors to be considered when the master set is reduced for each individual route.

Route similarity is usually interpreted as route overlap in a sense that the more two routes overlap, the more similar they are. Routes with little overlap are considered as unique routes. In a network as complex as a statewide road network there is an extensive number of overlapping route alternatives that exist between any given OD pair. Only unique routes need to be kept in the final choice set for future route choice modeling. In order to generate route choice sets we use a Path Size similarity measurement approach for identifying unique routes between OD pairs. The reason for choosing this approach is twofold. First, it is capable of identifying routes that are partly shared with one distinct route and partly shared with another distinct route. For example, if route *i* shares 40 percent of its length with distinct route A and the other 50 percent of its length with distinct route B, the proposed algorithm identifies route *i* as a distinct route. Second and equally important, it is easy to implement and not computationally costly.

The approach based on the total length of shared links and the algorithm is as follows:

- 1) Step 1: Identify the first route (in the dataset) and consider it as a unique route.
- 2) Step 2: Get the next route and find its shared links with each unique route.
- 3) Step 3: Compute “shared link length ratio” between the current route and each of the unique routes.
- 4) Step 4: If any of the computed ratios is greater than 0.75 then dismiss the route.

Otherwise, add it to the unique routes.

- 5) Step 5: Go to Step 2

$$\text{shared link length ratio} = \frac{\sum_{i=1}^n l_i^k}{\sum_{j=1}^N l_j^k}$$

l_i^k = length of link *i* in route *k*

l_j^k = length of link j in route k

n = number of shared links for route k

N = number of all links in route k

If this ratio for route k is less than 0.75 comparing to each unique route, then route k is considered a unique route. This series of comparisons continue until all the unique routes are found. The proposed algorithm for identifying unique routes is implemented on 10 different OD pairs. All the OD pairs selected for unique route identification were at least 50 miles apart and had more than 50 trips. To better analyze the issue of route variability detours are not taken into consideration. Detours are significantly longer than the majority of routes between an OD pair. One cannot draw a concrete conclusion regarding the reason behind occurrence of detours based solely on GPS data. However, one viable assumption in the context of this study could be that detours happen due to some minor deliveries along the main trip. After a series of experiments, it was found that most detours are longer than the 90th percentile of the longest route for 10 OD pairs. Therefore, for each OD pair all the trips falling under 90 percentile of the longest route were selected in order to exclude possible detours. Subsequently, unique routes were identified using the algorithm that is discussed earlier. Figure 4.8 shows an example of identified unique routes for five OD pairs.

The results suggest that one of the key factors that impact truck route choice variability is the network structure between the OD pair. The more competitive routes are available, the more different routes are observed. Case 2 in Figure 4.8 illustrates this phenomenon. On the other hand, where there are only one or two viable route options available, less variability is observed in the chosen routes between the OD pair. Case 3 is an example of this situation.

4.8 Validation

The results were validated in two separate phases. In the first phase, the identified unique routes for each OD pair were manually compared to all the 90th percentile routes to check if the identified routes cover most of the variations. In the second phase, within each OD pair identified routes were compared to each other to check they do not overlap more than 75% of the route length. This validation was done for all the 10 OD pairs and the results show that the performance of the algorithm is satisfactory. Table 4.3 illustrates the results for route variation measurement.

The diversity between the routes for OD pairs primarily depends on the network structure and availability of competitive route alternatives. For example cases one and two in Table 4.3 are quite different in number of different routes while their OD distance and number of routes are very close. The fact that for longer routes the diversity of routes is low can be explained in the context of study region. In Florida, interstates I-75 and I-95 are two major interstates in north-south direction. Both interstates are stretched along the longer side of the state. I-10 is another major interstate that is in east-west direction connecting the panhandle to the east coast of Florida. Therefore, most trips longer than 200 miles end up on these three options and as a result, route diversity for trips longer than 200 miles is low.

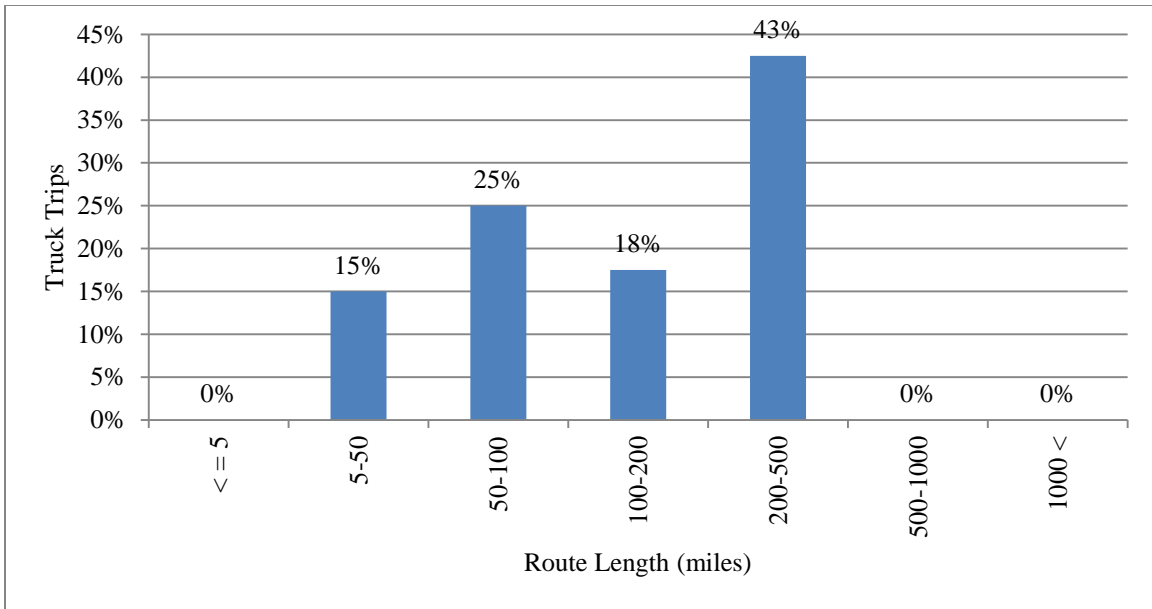


Figure 4.5 Route length distribution for selected 80 trips

Table 4.1 General trip information for 80 routes selected for feasibility and consistency checks

Case Number	Route number	Trip Time(minutes)	Trip Length(miles)	Status
1	10718	75	82	Ok
2	11042	81	89	Ok
3	11113	90	96	Ok
4	114552	379	459	Ok
5	13	200	226	Ok
6	14648	200	229	Ok
7	14773	247	299	Ok
8	14795	247	300	Ok
9	14797	70	68	Ok
10	14807	282	343	Ok
11	14839	247	302	Ok
12	14894	84	85	Ok
13	14939	65	69	Ok
14	15724	40	38	Ok
15	15760	40	38	Ok
16	15780	41	38	Ok
17	16202	63	67	Ok
18	16257	75	82	Ok
19	1626	187	203	Ok

Table 4.1 (Continued)

20	16623	79	73	Ok
21	1735	189	221	Ok
22	1788	100	115	Ok
23	1844	96	111	Ok
24	185	240	267	Ok
25	1871	248	279	Ok
26	1907	173	196	Ok
27	2259	190	217	Ok
28	2805	182	212	Ok
29	2806	174	205	Ok
30	2866	261	306	Ok
31	2883	184	209	Ok
32	2944	150	163	Ok
33	2953	45	35	Ok
34	2960	47	50	Ok
35	2961	149	163	Ok
36	2997	91	99	Ok
37	3737	55	58	Ok
38	421	186	213	Ok
39	44	151	179	Ok
40	4977	39	44	Ok
41	5041	31	25	Ok
42	5197	57	50	Ok
43	58	147	180	Ok
44	7434	45	45	Ok
45	7793	24	21	Ok
46	7818	51	50	Ok
47	7831	55	61	Ok
48	8474	72	80	Ok
49	8675	59	57	Ok
50	92	129	152	Ok
51	759	52	51	Ok
52	244	60	61	Ok
53	756	61	58	Ok
54	601	250	306	Ok
55	801	289	353	Ok
56	318	117	107	Ok
57	831	301	355	Ok

Table 4.1 (Continued)

58	113	244	297	Ok
59	128	234	291	Ok
60	100	239	293	Ok
61	85	73	81	Ok
62	687	61	57	Ok
63	720	61	59	Ok
64	68	131	143	Ok
65	165	18	15	Ok
66	326	116	112	Ok
67	13511	130	118	Ok
68	14353	259	253	Ok
69	16070	325	302	Ok
70	174	196	183	Ok
71	17999	321	286	Ok
72	23240	251	225	Ok
73	50345	270	231	Ok
74	52905	187	197	Ok
75	111147	236	243	Ok
76	115020	311	285	Ok
77	117122	329	325	Ok
78	119952	250	241	Ok
79	138428	229	211	Ok
80	91341	252	262	Ok

Table 4.2 Spatial gap vs. ping rate for selected 80 trips

Spatial gap (miles) \ Ping rate (minutes)	Spatial gap (miles)							Sum
	< 1	1-5	5-15	15-20	20-25	25-30	30 <	
< 1	13.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	13.6%
1-2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2-5	0.2%	12.6%	21.6%	0.0%	0.0%	0.0%	0.0%	34.4%
5-14	0.2%	3.9%	14.5%	0.0%	0.0%	0.0%	0.0%	18.6%
14-15	0.1%	0.3%	19.2%	12.9%	0.0%	0.0%	0.0%	32.5%
15-20	0.1%	0.0%	0.6%	0.0%	0.0%	0.0%	0.0%	0.6%
20-25	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%
25-30	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%
30-45	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
45 <	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Sum	14.5%	16.8%	55.8%	12.9%	0.0%	0.0%	0.0%	100.0%

Table 4.3 Route variation measurement results

Case	OD Distance (miles)	Number of Trips	Number of Unique Routes
1	55	197	6
2	57	197	2
3	91	94	6
4	117	237	10
5	156	91	4
6	189	86	2
7	219	196	2
8	224	72	1
9	348	48	2
10	375	100	1

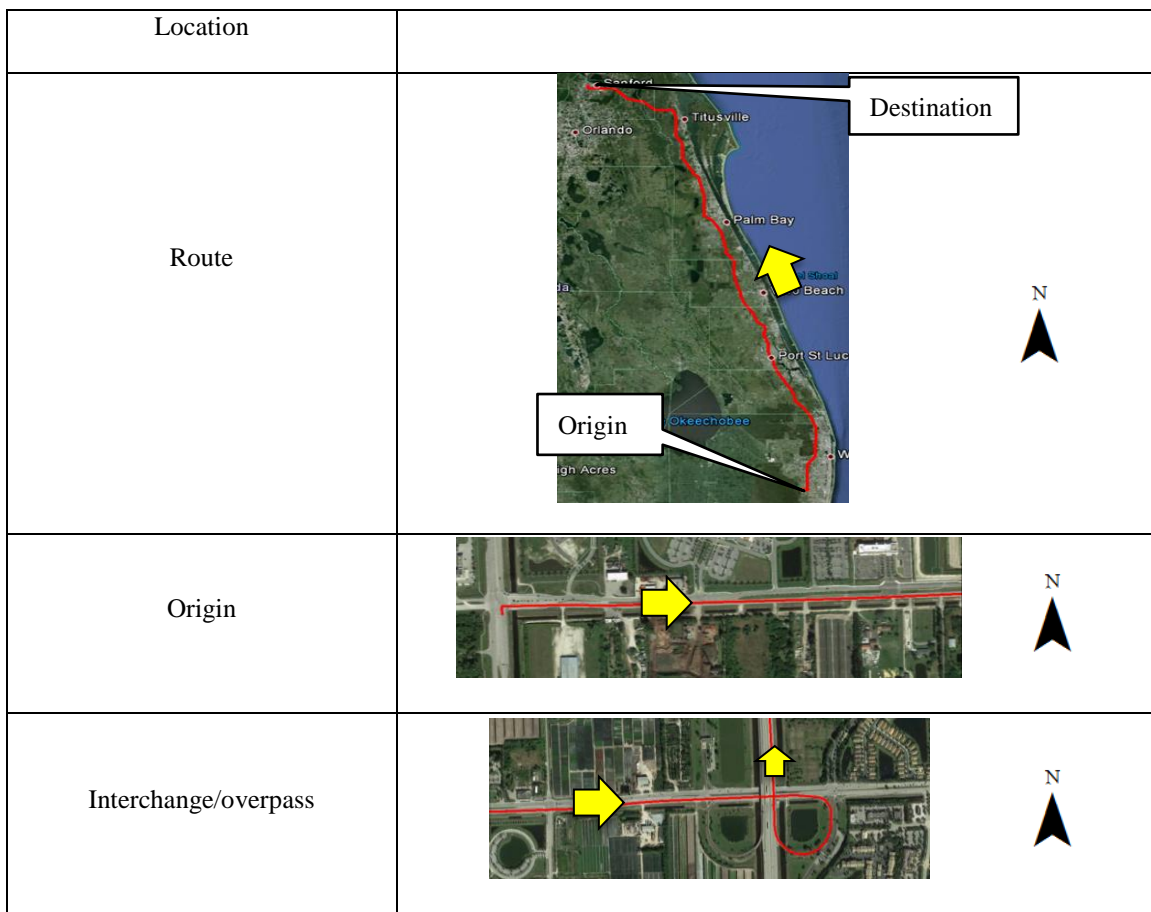


Figure 4.6 Consistency and feasibility check for one trip

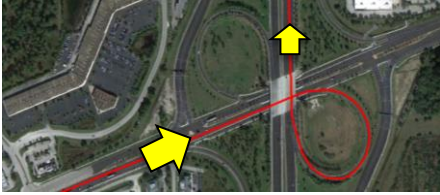

Interchange/overpass	
Destination	

Figure 4.6 (Continued)

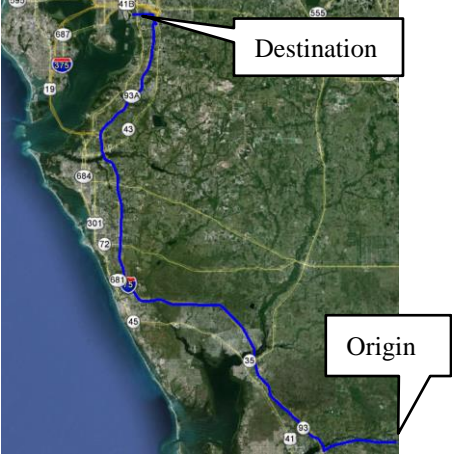

Case number	Before time-sampling	After time-sampling	Are the routes similar?
1			Yes

Figure 4.7 Routes with and without time sampling

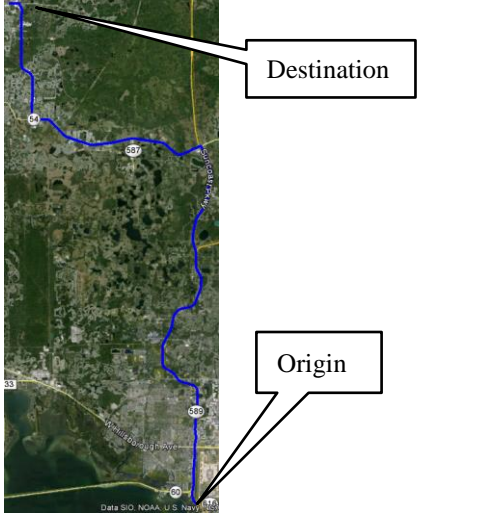
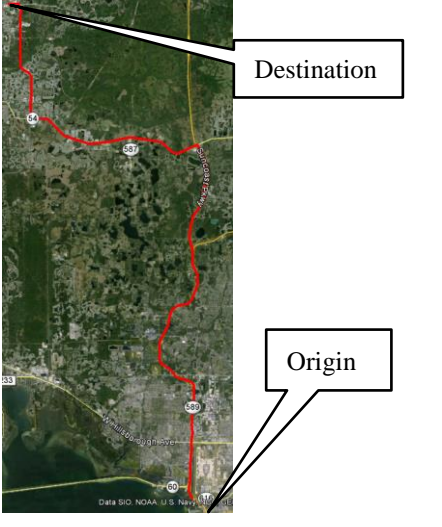
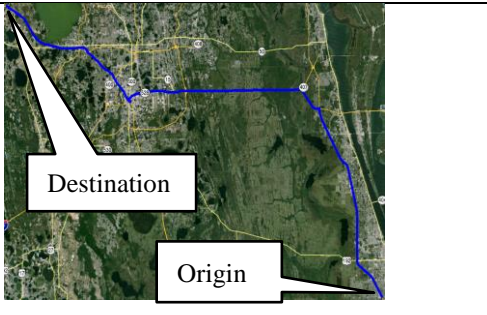
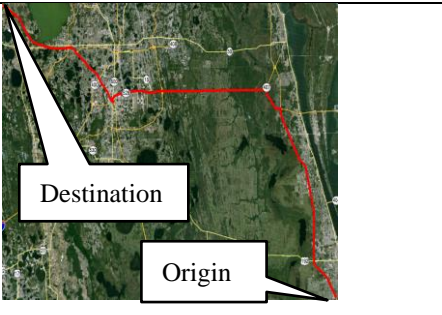
2			Yes
3			Yes

Figure 4.7 (Continued)

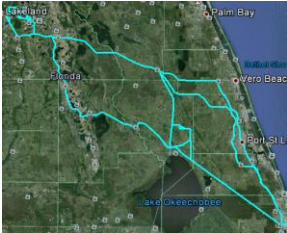
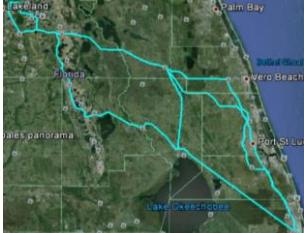
Case number	90 percentile	Unique Routes
1: Minimum OD Distance = 156 miles	 <p data-bbox="581 1570 688 1602">91 routes</p>	 <p data-bbox="1117 1570 1208 1602">4 Routes</p>

Figure 4.8 90th percentile and unique routes


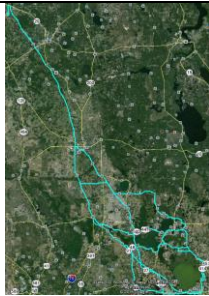
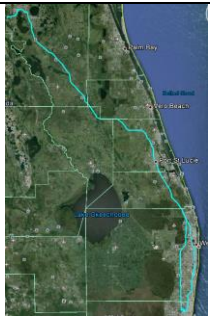
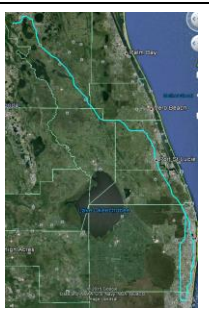
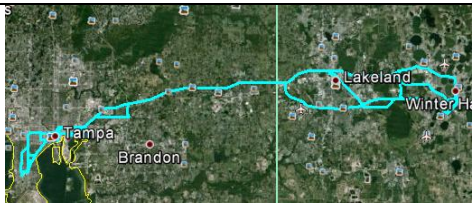
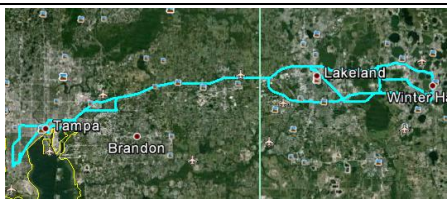

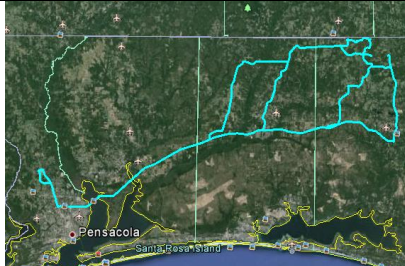
<p>2: Minimum OD Distance = 117 miles</p>	 <p>237 routes</p>	 <p>10 routes</p>
<p>3: Minimum OD Distance 189 miles</p>	 <p>86 routes</p>	 <p>2 routes</p>
<p>4: Minimum OD Distance = 55 miles</p>	 <p>197 routes</p>	 <p>6 routes</p>
<p>5: Minimum OD Distance = 91 miles</p>	 <p>94 routes</p>	 <p>6 routes</p>

Figure 4.8 (Continued)

CHAPTER 5 : CONCLUSION OF THE FIRST PART

5.1 Conclusions

Trucks play a pivotal role to meet the ever increasing demand for freight movement. In this context, the availability of GPS data in recent years has attracted the attention of both researchers and practitioners. As a result, a considerable number of studies have been dedicated to investigate this subject during the past years. The existing literature does not expand on route generation algorithms for low frequency (5 to 20 minutes) GPS data. Moreover, the previous route generation algorithms are mainly geared toward dense urban networks rather than rural networks. Considering route variability analysis, few studies attempt to introduce a practical method for identifying different routes between an OD pair. This study is an attempt to fill gaps in literature with regard to route generation algorithms as well as route variation quantification.

5.2 Map-matching and Route Generation

From a methodological point of view, map-matching based on closest link and second closest link along with modified Dijkstra's method for generating routes have shown satisfactory results. Validation checks have been done by randomly selecting routes and following them on Google Earth to evaluate their feasibility and consistency. As far as route variation analysis is concerned, comparing each route with each of previous unique routes resulted in capturing most of the possible observed route variations. To check the findings from this part, for each OD, identified unique routes were manually compared to all of the observed routes to make sure that all the possible variations are captured in the unique routes quota. Furthermore, between each

OD pair unique routes were compared with each other to make sure that their overlap is less than 75% of the route length. These validation checks confirmed the proposed method for identifying unique routes in this thesis.

From a practical standpoint, the proposed methods are easy to implement with a satisfactory level of accuracy. Most studies use complex methods to map-match GPS data and generate routes. The methods developed here are more practical and can be immediately put into practice and help agencies address policy concerns. Moreover, the methodology is frugal in terms of time considering the size of data and network. The total time needed for map-matching and generating routes proposed in this study is less than 5 hours. This is an advantage over previous works that usually deal with relatively much smaller GPS datasets. More importantly, an effective, yet simple method for identifying unique truck routes is introduced that can be used in choice set generation practices.

5.3 Opportunities for Future Research

While this thesis opens an avenue to explore truck route choice analysis, relying on GPS data has some limitations that need to be addressed in future research. One of the main challenges is lack of information on the decision makers' side. The availability of such information can be used to better estimate the travel path of a truck where the frequency of GPS data is low. Another limitation in this study is considerably low-frequency GPS data compared to similar studies. The low frequency nature of the data has imposed filtering criteria that leads to eliminating portions of data. As far as map-matching is concerned, it would be interesting to compare the performance of proposed algorithm with that of other algorithms, namely, probabilistic based, and geometric and topologic based algorithms. Such comparisons may lead to upgrades to the current proposed algorithms.

CHAPTER 6 : AN OVERVIEW OF THE DATA USED IN THE PROJECT

6.1 Introduction

This chapter is the beginning of the second part of this thesis. In this part, the methodology developed for route generation in previous chapters is implemented in a project sponsored by FDOT. The goal of the project is to derive routes of tanker trucks that deliver fuel commodities from Port Everglades (PEV) to 12 counties in southern Florida. This chapter provides a background on ATRI's truck GPS data and other data that were of use in the project. This chapter also introduces a method that has been used to separate tanker trucks' GPS data from other trucks' GPS data.

6.2 ATRI GPS Data

In this project the GPS data for two months, September 2014 and March 2015, was obtained from ATRI. This data covers tanker trucks visiting 12 counties in Florida: Miami-Dade, Broward, Palm Beach, Monroe, Martin, St. Lucie, Indian River, Okeechobee, Glades, Hendry, Lee, and Collier. At a minimum each record of data from ATRI has the following information:

- 1) Unit information: a specific ID (truck ID henceforth) number belonging to the truck
- 2) Temporal information: time stamp of the time when the position of the truck was recorded in the following format: MM/DD/YYYY HH:MM:SS
- 3) Geographical information : the latitude and longitude that locates the location of the truck

Data provided by ATRI was divided into two formats, D1 and D2, for each month. The characteristics of D1 and D2 data are listed below:

- 1) D1 :
 - a) Includes the truck instantaneous speed (spot speed) in addition to the above-mentioned information
 - b) Truck ID rotates every 24 hours
- 2) D2:
 - a) Does not include spot speed of the truck
 - b) Truck ID is static

A sample record of D1 data looks as below.

Unique Truck ID	Time/Date Stamp	Speed	Latitude	Longitude
absdefghi12232123	05/03/2011 01:55:55	25	33.915932	-84.494760

Figure 6.1 An example of raw D1 data

And a sample record of D2 data looks as below.

Unique Truck ID	Time/Date Stamp	Latitude	Longitude
12232123	05/03/2011 01:55:55	33.915932	-84.494760

Figure 6.2 An example of raw D2 data

It must be noted that “Unique Truck ID” is a random number assigned to each vehicle and cannot be used to trace back the actual vehicle from the trucking company. Truck ID however, can be used to distinguish between different trucks in the database for trip measurement purposes. Moreover, a truck cannot be tracked for more than 24 hours in D1 data. For example, if there are two days of data for one truck in D1 dataset, the truck ID in the first day is different from the truck ID in the second day for that truck. In D2 data on the other hand, truck IDs are static throughout the dataset and therefore, a truck can be tracked for several days.

6.2.1 Data Coverage

Main characteristics of the data such as number of days of data, number of trucks, and number of GPS points were investigated. For both months of data, D2 data was significantly

richer than D1 data in terms of number of days the data was available for, number of GPS records (points), and number of trucks .Table 6.1 shows the data coverage for the two months.

One of the important attributes of GPS data is the ping-rate between consecutive GPS points. The higher the ping rate, the more frequent the GPS data. Another attribute of GPS data that has importance is the spatial gap between consecutive GPS points. Spatial gap and ping rate are indirectly related in a sense that the higher the ping rate, the smaller the spatial gap. Tables 6.2, 6.3, 6.4 and 6.5 illustrate this relationship for each type of data separately. These tables show that D1 data is more frequent than D2 data for both months. In addition, a comparison between table 6.3 (or 6.5) and table 3.1 reveals that D2 data is very similar to the data used in the first part of the thesis in terms of ping rate and spatial gap. Moreover, the majority of data in this project is D2 and therefore, it is reasonable to implement the developed methodologies in previous chapters for this project.

Table 6.1 Attributes of the September 2014 and March 2015 GPS data

	September 2014		March 2015		Total
	D1	D2	D1	D2	
# Days of data	10	30	22	31	94
# Truck IDs	11	44	34	52	141
# GPS Points	8,621	86,606	35,182	112,009	242,418

Table 6.1 shows that there were 141 total truck IDs in the dataset. It must be noted that this number is only the sum of all the truck IDs existing in both months of data, and it is not the number of total unique truck IDs.

Table 6.2 Cross-tabulation of spatial gap against ping rate for September 2014 – D1 data

Spatial gap (miles) \ Ping rate (minutes)		Spatial gap (miles)				Sum
		< 1	1-5	5-15	15 <	
< 1	< 1	45.1%	9.4%	0.0%	0.0%	54.5%
	1-2	22.1%	21.1%	0.0%	0.0%	43.2%
	2-5	0.4%	0.3%	0.0%	0.0%	0.7%
	5-15	0.8%	0.1%	0.0%	0.0%	1.0%
	15-20	0.2%	0.0%	0.0%	0.0%	0.2%
	20-25	0.0%	0.0%	0.0%	0.0%	0.0%
	25-30	0.1%	0.0%	0.0%	0.0%	0.1%
	30-45	0.0%	0.0%	0.0%	0.0%	0.0%
	45-60	0.2%	0.0%	0.0%	0.0%	0.2%
	60-75	0.1%	0.0%	0.0%	0.0%	0.1%
	> 75	0.0%	0.0%	0.0%	0.0%	0.0%
	Sum	68.9%	31.0%	0.1%	0.0%	100.0%

Table 6.3 Cross-tabulation of spatial gap against ping rate for September 2014 – D2 data

Spatial gap (miles) \ Ping rate (minutes)		Spatial gap (miles)				Sum
		< 1	1-5	5-15	15 <	
< 1	< 1	17.1%	0.2%	0.0%	0.0%	17.2%
	1-2	8.2%	0.7%	0.0%	0.0%	9.0%
	2-5	11.3%	9.4%	10.0%	0.0%	30.8%
	5-15	10.4%	7.7%	11.8%	1.6%	31.6%
	15-20	2.8%	0.1%	1.0%	0.7%	4.6%
	20-25	1.6%	0.0%	0.0%	0.0%	1.6%
	25-30	1.5%	0.0%	0.0%	0.0%	1.5%
	30-45	1.5%	0.0%	0.0%	0.0%	1.5%
	45-60	0.5%	0.0%	0.0%	0.0%	0.5%
	60-75	0.2%	0.0%	0.0%	0.0%	0.2%
	> 75	1.0%	0.3%	0.2%	0.1%	1.6%
	Sum	56.2%	18.4%	23.0%	2.4%	100.0%

In fact, there were 19 truck IDs that were shared between two months of September 2014 and March 2015 and therefore, total number of unique truck IDs is 122.

Table 6.4 Cross-tabulation of spatial gap against ping rate for March 2015 – D1 data

Spatial gap (miles) \ Ping rate (minutes)		Spatial gap (miles)				Sum
		< 1	1-5	5-15	15 <	
< 1		42.7%	10.0%	0.0%	0.0%	52.7%
1-2		21.1%	22.4%	0.0%	0.0%	43.5%
2-5		0.9%	0.4%	0.0%	0.0%	1.2%
5-15		1.6%	0.1%	0.1%	0.0%	1.8%
15-20		0.3%	0.0%	0.0%	0.0%	0.3%
20-25		0.1%	0.0%	0.0%	0.0%	0.1%
25-30		0.1%	0.0%	0.0%	0.0%	0.1%
30-45		0.1%	0.0%	0.0%	0.0%	0.1%
45-60		0.1%	0.0%	0.0%	0.0%	0.1%
60-75		0.0%	0.0%	0.0%	0.0%	0.0%
> 75		0.1%	0.0%	0.0%	0.0%	0.1%
Sum		67.0%	32.9%	0.1%	0.0%	100.0%

Table 6.5 Cross-tabulation of spatial gap against ping rate for March 2015 – D2 data

Spatial gap (miles) \ Ping rate (minutes)		Spatial gap (miles)				Sum
		< 1	1-5	5-15	15 <	
< 1		19.0%	0.2%	0.0%	0.0%	19.2%
1-2		7.8%	1.2%	0.0%	0.0%	9.0%
2-5		9.3%	9.9%	10.8%	0.0%	30.0%
5-15		9.9%	7.5%	12.8%	1.4%	31.6%
15-20		2.6%	0.0%	0.6%	0.6%	3.9%
20-25		1.5%	0.0%	0.0%	0.0%	1.5%
25-30		1.3%	0.0%	0.0%	0.0%	1.3%
30-45		1.3%	0.0%	0.0%	0.0%	1.3%
45-60		0.6%	0.0%	0.0%	0.0%	0.6%
60-75		0.2%	0.0%	0.0%	0.0%	0.2%
> 75		1.0%	0.4%	0.1%	0.1%	1.5%
Sum		54.3%	19.2%	24.3%	2.1%	100.0%

6.2.2 Separating Tanker Trucks from Other Trucks

One of the tasks in this project was to separate the GPS data belonging to tanker trucks from the GPS data belonging to other trucks. It has to be mentioned that tanker trucks that carry

fuel commodities from PEV to fuel recipients in 12-county area, load the commodities at designated terminals in PEV. Figure 6.3 illustrates the location of these terminals in PEV. Initially a GIS shapefile containing 13 terminal points was provided to the research team. After removing and adding a couple of terminals by the project team, the analysis was continued with 14 terminals in PEV.

To separate tanker trucks from other trucks, a polygon was drawn around each terminal so that if a tanker truck had stopped at that terminal to load fuel commodities, its GPS data would have been captured in the polygon. Subsequently, these polygons were saved as a GIS shapefile and sent to ATRI. ATRI then provided the research team with the GPS data that fell inside the polygons. In total there were 14 terminals for which the polygons were drawn. Figure 6.4 shows the polygon around terminal 1. The polygon is in red and the terminal is in yellow circle. The polygons were usually extended beyond limits of the actual terminals to capture any GPS point with a small distance from the actual terminal due to GPS spatial errors. The rest of the polygons around other terminals can be found in Appendix A.

6.3 Fuel Recipient Data

Gas stations are the main delivery destinations of tanker trucks that load fuel commodities at PEV. In addition to gas stations, there are other fuel recipients such as government agencies, agricultural establishments, and industrial establishments that receive fuel from tanker trucks in the 12-county region. In order to investigate what proportion of the gas stations or other fuel recipients receive fuel from PEV two sets of fuel recipient data were provided to the research team. One set of data came from Department of Revenue (DOR) surveys and the second dataset came from HERE which is a map service and location data

provider. The two datasets were investigated and compared and their advantages and disadvantages are discussed in the remaining of this section.



Figure 6.3 Location of the terminals (yellow points) in PEV and their associated number (in red)



Figure 6.4 Terminal 1 (in yellow circle) and the red polygon used for GPS data extraction

6.3.1 DOR Data

This data set included 2315 facilities that consisted of gas stations and other fuel recipients such as agricultural, industrial, and government facilities. Gas stations account for 69% of the facilities in DOR data and the rest of the facilities (i.e., agricultural, industrial, and

government facilities) are considered as “other fuel recipients”. This dataset was incomplete in terms of covering all the active gas stations in the 12-county region. This was revealed through a comparison between gas stations available in Google Earth and gas stations available in DOR data.

DOR data was the only source of information on any fuel recipients other than gas stations. As mentioned before these fuel recipients were agricultural, industrial, or government facilities that received fuel commodities from PEV. Such fuel recipients are referred to as “other fuel recipients” from this point forward in this thesis. Moreover, in this project it was important to know what percentages of trucks serve other fuel recipients and therefore, they were separated from gas stations in future analysis.

6.3.2 HERE Data

HERE data contained 1841 gas stations in the 12-county region. There were no other facility type other than gas stations in HERE data. HERE data was also incomplete in terms of encompassing all the active gas station in the 12-county region. This was revealed through comparing active gas stations in HERE with active gas stations in Google Earth. Therefore, both datasets, namely DOR and HERE, were incomplete. Moreover, both datasets had some level of gas station data overlap when compared to each other.

6.3.3 Comparison between DOR and HERE Data

The two datasets were not complete as mentioned earlier. Moreover, they showed some degree of overlap in terms of geocoded gas stations when both layers of DOR and HERE were compared. Figure 6.5 demonstrates two overlapping points, one from HERE and the other from DOR, that geocode one gas station. The variable “x” is the spatial difference between these two overlapping points. For most of overlapping points in HERE and DOR data, x was found highly

varied and in turn, it was not possible to establish a limit on x to distinguish between overlapping and non-overlapping points. This led to using a combination of both DOR and HERE data (without removing overlapping points) to identify trip origins (destinations) location descriptions. In order to accurately identify trip origins (destinations) location descriptions, each trip origin (destination) was observed in Google Earth using clusters technique. This technique will be explained in details in the coming chapters.

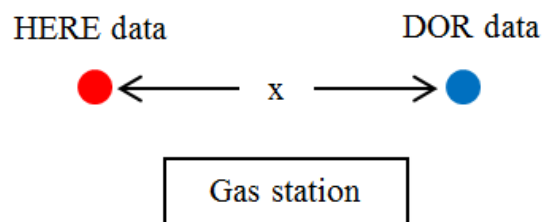


Figure 6.5 Overlapping points from HERE (red) and DOR (blue)

CHAPTER 7 : DERIVING TRIP CHAINS FOR TANKER TRUCKS

7.1 Introduction

This chapter goes over the procedure used for converting raw GPS data into truck trips and deriving trip chains. American Transportation Research Institute (ATRI) provided the research team with GPS data belonging to tanker trucks travelling within the 12-county region. First, this data was converted to truck trips using a trip conversion algorithm developed by Thakur et al. (2). Then trip characteristics such as trip length, trip time, and trip speed distributions were measured. Later, certain criteria were introduced to build the chain of trips made by each truck. By following such trip chains one can learn the trip patterns of tanker trucks that carry fuel commodities in the 12-county region.

7.2 Algorithm Description

The overall procedure to convert raw GPS data to truck trips is listed below:

- 1) Clean, read and sort raw GPS data in a chronological order for each truck ID. At the end of this step all the GPS data belonging to each truck ID is grouped together in the chronological order.
- 2) Identify stops (i.e., trip ends) based on spatial movement, time gap, and speed between consecutive GPS points.
 - a) Derive a preliminary set of trips based on a minimum stop dwell-time buffer value. Use 5 minutes of dwell-time.
- 3) Conduct additional quality check and eliminate trips that do not satisfy quality criteria

The original version of the above-mentioned algorithm is developed and explained by Thakur et al. (2). However for this project, the following changes were made to the original version of the algorithm.

First, a minimum dwell-time of 5 minutes was used in order to capture all the possible stops that tanker trucks make. This is because the dwell-time for fuel delivery can be quite varied based on type of truck or fuel recipient. As a result, in order to avoid missing any fuel delivery stops the minimum dwell time was set to 5 minutes so that all the possible valid stops could be captured.

Second, all the trips that were shorter than 1 mile were captured because short fuel delivery trips could happen considering the notable number of gas stations located within 1 mile of PEV. Moreover, a tanker truck can make multiple fuel delivery stops at multiple gas stations that are located within 1 mile of each other.

Third, no consecutive trips were joined based on destination or origin facility type at this step. This means that if a truck ended its first trip at a rest stop, and started its next trip from the same rest stop, the two trips were not joined (in the original version of the algorithm such trips would be joined). The reason for not joining such trips was to capture all the possible fuel delivery stops. It was observed that there are quite a few rest stops that have gas stations. Therefore, most of the stops made there by tanker trucks were for fuel delivery purposes rather than recreational purposes.

This algorithm was applied to 242,218 raw GPS points which resulted in 14,598 trips. Table 7.1 summarizes the results from converting 242,218 GPS points into 14,598 trips.

Table 7.1 General trip statistics for 14,598 extracted trips

		D1 data	D2 data	All data
September 2014	Number of GPS records	8,621	86,606	95,227
	Number of trips extracted	92	6,435	6,527
	Number of unique truck IDs	11	44	55
	Average trip length (miles)	47	28	28
	Average trip time (minutes)	49	35	35
	Average trip speed (mph)	41	35	35
March 2015	Number of GPS records	35,182	112,009	147,191
	Number of trips extracted	344	7,727	8,071
	Number of unique truck IDs	34	52	86
	Average trip length (miles)	41	33	33
	Average trip time (minutes)	51	38	39
	Average trip speed (mph)	43	38	38
All two months	Number of GPS records	43,803	198,615	242,418
	Number of trips extracted	436	14,162	14,598
	Number of unique truck IDs	45	96	122
	Average trip length (miles)	42	31	31
	Average trip time (minutes)	51	37	37
	Average trip speed (mph)	43	37	37

7.2.1 Validation Checks

100 trips were randomly selected to check their trip ends' locations. The purpose of this validation check was to test if origins and destinations of extracted trips were in valid locations. The definition of a valid location is a gas station, PEV terminals, distribution center, other fuel recipients, etc. If a trip end (i.e., origin or destination) fell on the roadway then that trip end was invalid and flagged as "roadway". Table 7.2 illustrates the location description of trip ends belonging to 100 trips selected for validation checks.

Figures 7.1, 7.2, and 7.3 illustrate trip length, trip time, and average trip speed distributions for 100 validation trips, respectively. Same distributions for the whole 14,598 extracted trips are provided in Section 7.3. Comparing length, time and average speed

distributions between all 14,598 trips and 100 trips selected for validation, selected trips seem to be a good representative of the population. Moreover, the validation check results summarized in Table 7.2 show that most of trip ends fall on valid locations. Only 4% of trip origins and 11% of trip destinations were found on the roadway. Even though such percentages are still at a satisfactory level, there is a major reason why still a few trip ends were observed on the roadway. This reason is explained in detail in Section C.3 of Appendix C.

Table 7.2 Summary of trip end location description of 100 trips selected for validation checks

	Origin	Destination
PEV Terminal	37	34
Gas Station	47	44
Distribution Center	9	8
Other	3	3
Roadway	4	11
Sum	100	100

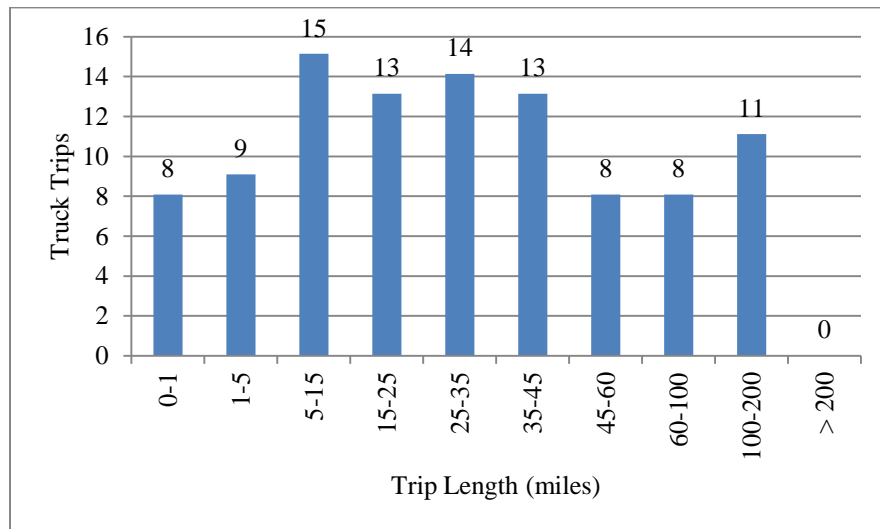


Figure 7.1 Trip length distribution for 100 validation trips

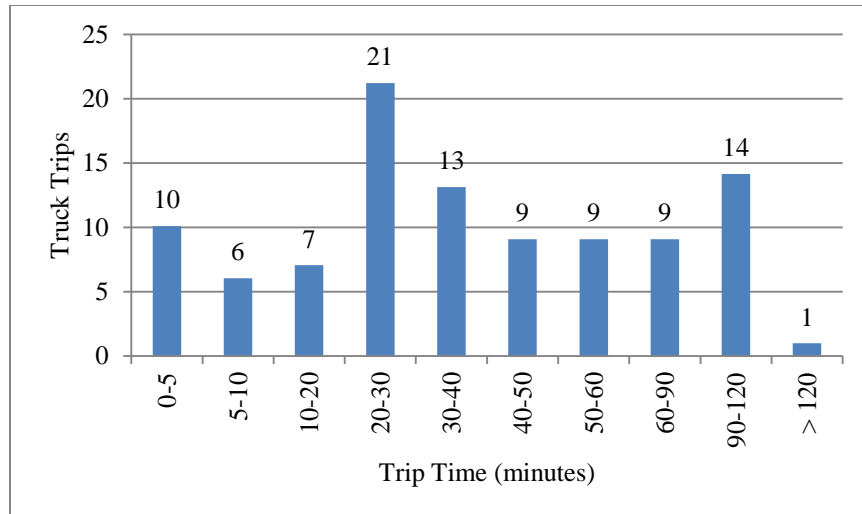


Figure 7.2 Trip time distribution for 100 validation trips

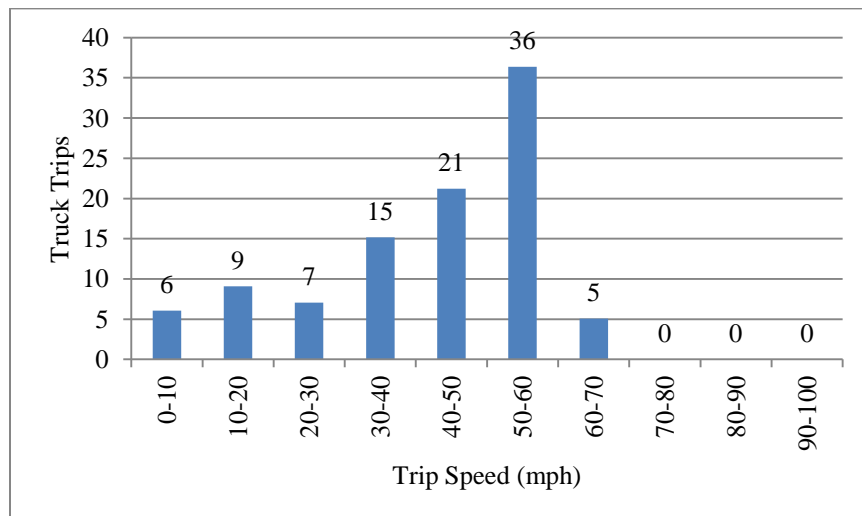


Figure 7.3 Trip speed distribution for 100 validation trips

7.3 Characteristics of Truck Trips Derived from ATRI Data

This section provides some analyses of truck trip data derived from ATRI GPS data. These analyses include trip length distribution, trip time distribution, average trip speed distribution, and time of day profiles. Figure 7.4 shows the 12-county region along with the study boundary (blue polygon). The red dots in the figure show a sample of GPS points obtained from ATRI. The figure illustrates that most of the truck data is concentrated on the east coast of the

region while rest of the data is stretched to the west. This geographical distribution of GPS points impacts the distribution of trip characteristics, namely, trip length and trip time. Figure 7.5 shows the trip length distribution of all 14,598 trips derived from the data. As can be observed, the distribution is divided into two portions. The first portion is trips less than 35 miles which mostly cover the eastern area of 12-county region. The second portion is trips more than 35 miles that cover the western area of 12-county region. The reason for this division is that gas stations or other fuel recipients are predominantly located on the east coast rather than west coast of 12-county region, and in the middle (gator alley) there are not many gas stations or fuel recipients that can attract fuel delivery trips. Figure 7.6 illustrates the trip time distribution of all 14,598 trips derived from two months of data. Similarly, the trip time distribution is also divide into two portions, one portion belongs to trips covering the eastern area of 12-county region whose trip time is less than 50 minutes, and the other portion belongs to trips covering the western area of 12-county region whose trip time is more than 50 minutes. It must be noted that trip time represents the time interval during which the truck was moving and it also includes any traffic stops less than 5 minutes. Finally, Figure 7.7 shows average trip speed distribution for all 14,598 trips. Expectedly, there are few trips with average speed of 70 mph or above. Each of the following distributions for each separate month and each separate data type (i.e., D1 or D2) is included in the appendix B.

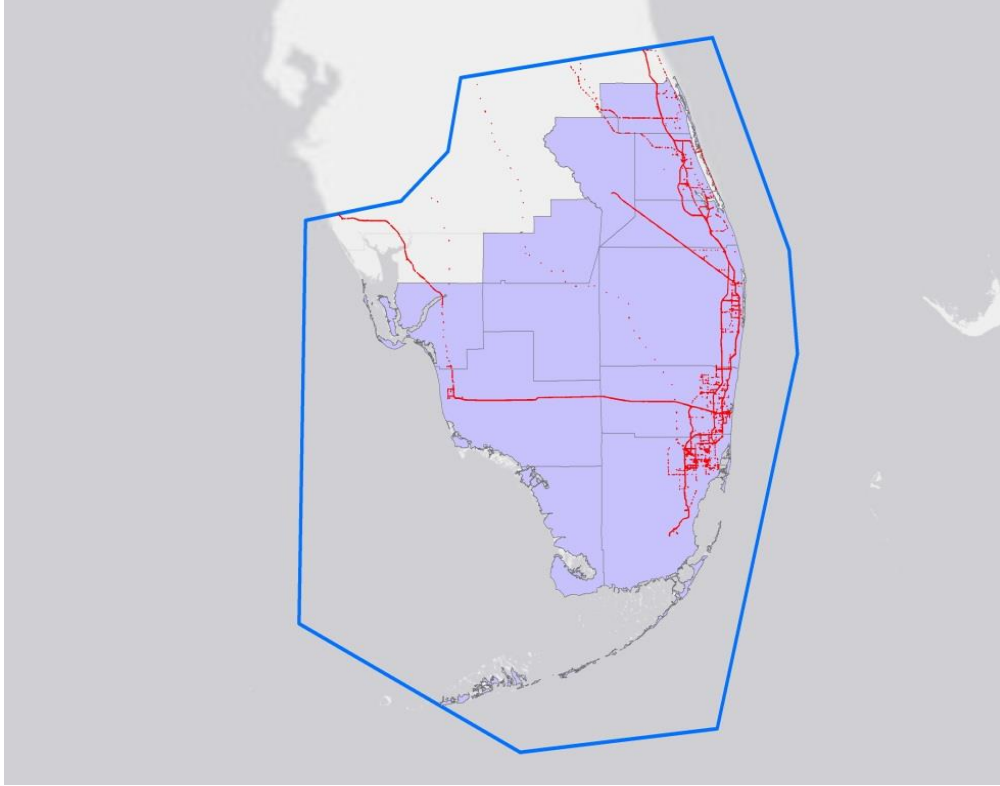


Figure 7.4 12-county region in south Florida, sample of ATRI GPS points (red dots), and study boundary (large blue polygon)

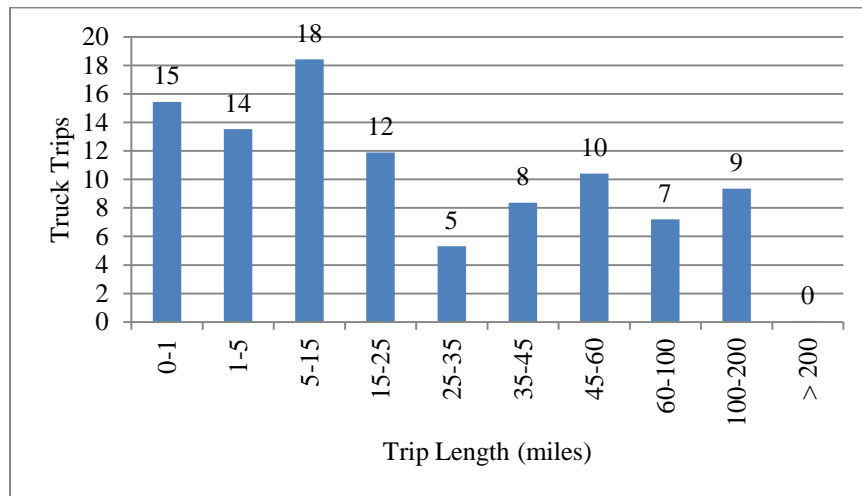


Figure 7.5 Trip length distribution of all 14,598 trips

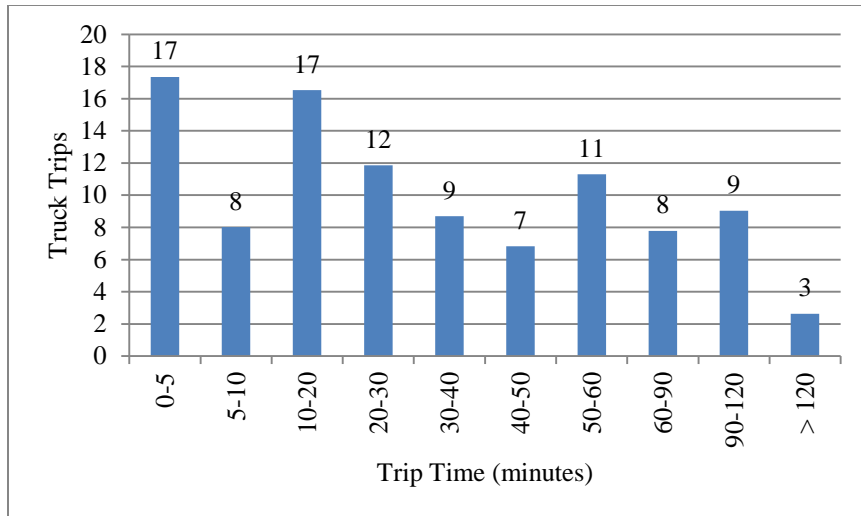


Figure 7.6 Trip time distribution of all 14,598 trips

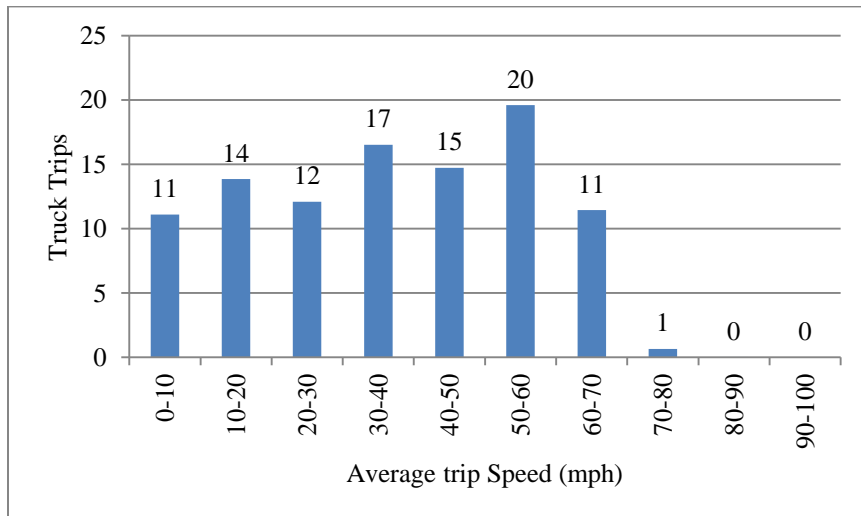


Figure 7.7 Average trip speed distribution of all 14,598 trips

7.4 Identifying Trip Ends' Location Descriptions

One of the main tasks in this project was to distinguish between fuel delivery trips and other trips. Moreover, gas stations or other fuel recipients that did not exist in HERE or DOR data but receive fuel commodities from tanker trucks had to be identified. In addition, there were some anomalies in both HERE and DOR data regarding the correct coordinates of gas stations or other fuel recipients that had to be resolved and corrected based on the observed trips in ATRI

data. To address all the issues mentioned above, the location description of extracted trip ends had to be identified.

The main idea for identifying trip ends' location descriptions was based on observing the land use of trip ends in Google Earth. Since going through all the 14,598 trip ends was practically impossible the idea of grouping close trip ends was proposed. This idea helped reduce the load of trip end land uses that had to be observed in Google Earth. The algorithm below illustrates how these grouped trip ends (hereafter clusters) were made. First, two terms have to be explained. Place ID: an ID number for a unique location visited by trip ends. A gas station is an example of a unique location. Unique ID: a combination of rounded trip end longitude (hereafter X) and latitude (hereafter Y) to three decimal places. X and Y belong to trip ends. An example of unique ID: 27.122_-82.454.

The following describes the algorithm:

- 1) Round the GPS Y and X to three decimal places to capture points within 320ft of each other. After rounding, there are unique values of rounded Y coordinates (e.g.: 27.122) and X coordinates (e.g.: -82.454).
- 2) Combine the rounded Y and X coordinates to create unique IDs
- 3) Extract the first original GPS coordinates for each unique ID (i.e., non-rounded Y and X) and call them "unique ID representatives".
- 4) First Run:
 - a) Sort the dataset by original Y coordinates and then original X coordinates
 - b) Calculate the distance between consecutive unique ID representatives
 - c) Calculate the difference in GPS coordinates between unique ID representatives

- d) Combine consecutive unique ID representatives that are less than 1000 ft apart to identify place IDs
 - e) Calculate the total distance between the first and the last unique ID representatives of a place ID. If the total distance is more than 1000 ft, the place ID is then subdivided into two or more place IDs to make the total distance less than 1000 ft.
 - f) Select all new place IDs obtained after the first run.
- 5) Second Run:
- a) Sort the first set of new place IDs by original X coordinates and then original Y coordinates to recapture the points that may satisfy the spacing conditions but were too far apart due to the nature of the data sorting order. Perform the same steps from 4.b to 4.e.
 - b) Select all of the new place IDs obtained after the second run
- 6) Repeat the same procedure for the third and fourth
- 7) End the sorting and iterating process to identify place IDs because the number of unique place IDs has reached its minimum
- 8) Label each place ID with “New cluster #”. With # ranging from 1 to the total number of clusters

Identifying the location description of clusters was done using a combination of Google Earth, HERE, and DOR data. As discussed earlier, HERE and DOR data do not include the entire active gas stations or other fuel recipients in the 12-county region in Florida. Moreover, both sources of data have shown some anomalies in terms of geocoded gas stations or other fuel recipients. For example, a geocoded point representing a gas station in HERE or DOR data is not

exactly falling on its corresponding gas station in Google Earth. These anomalies were corrected while identifying the trip ends' location descriptions. To this end, all three layers of clusters, HERE, and DOR data were imported in Google Earth. Then, clusters' locations were observed. If HERE or DOR data points were falling on the exact location of clusters, then the clusters were just marked with the location description of HERE or DOR points. If HERE or DOR points were not falling on the exact location of clusters but were within 500 ft of the clusters then the coordinates of HERE or DOR points were updated to the exact coordinates of clusters. Otherwise, the coordinates of clusters were recorded as new gas stations or fuel recipients if clusters were in gas stations or fuel recipients.

Table 7.3 illustrates the land use description distribution of origins and destination for all 14,598 trips after implementing the algorithm described above. As can be observed around 40% and 35% of trip origins and destinations are observed in gas stations and PEV, respectively. This result is expected because these trips belong to tanker trucks that mainly deliver fuel to gas stations. Next large percentage in the table belongs to distribution centers (13 % for both origins and destinations). The fact that the share of distribution centers are very close for both origin and destinations suggests that some trucks mainly travel between distribution centers and in turn, are not delivering fuel. It is possible to have a few trucks in the data that do not deliver fuel. Such trucks were eliminated in further steps (next sections explains how this elimination was done). Lastly, some trip ends were observed on the roadway. This is not an issue of the trip conversion algorithm but rather an issue of the study boundary. Section C.3 in Appendix C explains how the study boundary causes some on-the-road trip ends.

Table 7.3 Land use description distribution for the trip origins and destinations of the total 14,598 trips

Case	Location	Number of trip origins	Number of trip destinations
1	Gas station	5,957 (40.8 %) trip origins	5,552 (38 %) trip destinations
2	PEV	5,245 (35.9 %) trip origins	5,005 (34.3 %) trip destinations
3	Other fuel recipients	747 (5.1 %) trip origins	368 (2.5 %) trip destinations
4	Distribution center	1,906 (13.1 %) trip origins	1,910 (13.1 %) trip destinations
5	On the road	463 (3.2 %) trip origins	1,510 (10.3 %) trip destinations
6	Rest stop	280 (1.9 %) trip origins	253 (1.7 %) trip destinations
	Total	14,598 (100%) trip origins	14,598 (100%) trip destinations

7.5 Cleaning the Trip Dataset

Extracted trips from ATRI data needed to be cleaned to be suitable for further analysis. To this ends, trucks that did not predominantly visit PEV had to be removed from the trip file. These kinds of trucks most probably are not tanker trucks and therefore are not of interest in this project. Moreover, if the first (last) trip of a truck has its origin (destination) on the roadway, that trip should be removed. This had to be done in order to have the first (last) trip of a truck start (end) at a valid location. In addition, trips with origin, or destination, or both on the roadway had to be addressed. These trips were either joined to their next (previous) trip or had their origin (destination) replaced by the previous (next) destination (origin). Lastly, truck IDs that belonged to the same trucks in D1 data had to be identified. Since truck IDs rotate every 24 hours in D1 data, it might include trucks with two different truck IDs in two consecutive days. By finding those trucks and changing the truck ID in the second day back to the truck ID in the first day, that truck could be followed for two consecutive days. All the four data cleaning tasks that were described above are listed below:

- 1) Remove 27 truck IDs that did not mainly deliver fuel commodities
- 2) Remove 36 trips at the start or end of data stream whose ends are on the roadway
- 3) Join or remove trips with origin or destination on the roadway based on certain criteria
- 4) Join 3pairs of truck IDs from D1 data based on certain criteria

Appendix C will illustrate each trip cleaning task in more details. Table 7.4 illustrates the land use description distribution of origins and destinations after taking the above-mentioned trip cleaning steps. As can be observed, shares of “distribution center” and “on the road” have dropped after cleaning the trip dataset and a total number of 12,649 trips were remained for further analysis.

Table 7.4 Land use description distribution for the trip origins and destinations for 12,649 trips after cleaning the trip file

Case	Location	Number of trip origins	Number of trip destinations
1	Gas station	5,683 (44.9 %) trip origins	5,370 (42.5 %) trip destinations
2	PEV	5,163 (40.8 %) trip origins	4,968 (39.3 %) trip destinations
3	Other fuel recipients	539 (4.3 %) trip origins	278 (2.2 %) trip destinations
4	Distribution center	694 (5.5 %) trip origins	665 (5.3 %) trip destinations
5	On the road	310 (2.5 %) trip origins	1,129 (8.9 %) trip destinations
6	Rest stop	260 (2.1 %) trip origins	239 (1.9 %) trip destinations
	Total	12,649 (100 %) trip origins	12,649 (100 %) trip destinations

7.6 Deriving Trip Chains

This section illustrates the process of deriving trip chains. A trip chain is a series of trips made by a truck in chronological order. A truck makes a chain of trips per day and therefore, building trip chains helps understand the behavior of tanker trucks. It also makes it possible to follow a chain of trips made by a truck for trip measurement analysis purposes.

Total number of trips derived from ATRI data was 14,598. Subsequently, a trip clearing process was implemented to make the dataset ready for building trip chains. This process resulted in keeping 12,649 trips in the dataset.

For the final 12,649 trips there were 12,538 consecutive trip pairs corresponding to 111 unique truck IDs. A trip pair in this context refers to two consecutive trips made by a truck. For example, if truck A makes three trips and truck B makes four trips, then there will be two trip pairs for truck A and three trip pairs for truck B. Figure 7.11 illustrates the relationship between spatial gap and temporal gap for 12,538 trip pairs. Spatial gap is the spatial distance between the destination of the first trip and the origin of the second trip in a trip pair. Similarly, temporal gap is the temporal difference between the destination of the first trip and the origin of the second trip in a trip pair. As can be observed in Figure 7.11, the majority of trip pairs have a spatial gap of less than 1 mile. Moreover, 8% of trip pairs have a temporal gap of more than 120 minutes (two hours).

There are 10,794 trip pairs (86% of the total 12,538 trip pairs) in which the land use description of the first trip's destination and the second trip's origin are the same. These 10,794 trip pairs are called "matching trip pairs". 98% of matching trip pairs (10,624 trip pairs) have the spatial gap of less than 1 mile. These statistics mean that there was strong connectivity between consecutive trips. Moreover, these statistics were used to define the criteria for building trip chains.

As long as trip chain criteria are concerned, it is worth understanding the spatial gap distribution for non-matching trip pairs (i.e. trip pairs in which the location description for the first trip's destination and the second trip's origin does not match). Figure 7.8 shows the spatial gap distribution for 1,744 non-matching trip pairs. The figure shows that a significant portion

(38%) of non-matching trip pairs fall within 1 mile of spatial gap. This means that choosing 1 mile spatial gap as one of the trip chain building criteria would also capture a significant portion of non-matching trips.

7.6.1 Procedure for Deriving Trip Chains

To build trip chains the following criteria were used.

- 1) Spatial gap < 1 mile
- 2) Temporal gap < 4 hours

The spatial gap less than 1 mile was proposed based on the discussion in the previous section. To put it in a nutshell, the spatial gap for the majority of trip pairs was less than 1 mile and in turn, 1 mile of spatial gap was chosen as the first criteria for building trip chains.

Moreover, based on the discussions between the research team and FDOT 4 officials and consultants, it was decided to add a temporal gap of 4 hours as the second criteria. This means that if the temporal gap between the one trip's origin and the next trip's destination was more than 4 hours then the chain of trips was broken.

Based on the above-mentioned criteria, 1,320 trip chains were built that included 11,918 trips. Figure 7.9 shows the distribution of number of trips existing in trip chains. As can be observed the majority of trip chains include five or less number of trips. Figure 7.10 shows a zoomed-in distribution of trip chains with five or less number of trips. It is noteworthy that there are a significant number of trip chains with high number of trips (Figure 7.9). Specifically 2% of trip chains include more than 50 trips. This is because trucks could be tracked for several days in a row and therefore, they have built trip chains with high number of trips.

Figure 7.12 and 7.13 show the length distribution and time distribution, respectively, for all 1,320 trip chains. As can be observed, there were a few trip chains with more than 1000 miles

length and consequently more than 1440 minutes (one day) duration. This is because some trucks have been tracked for more than one day and as a result, their trip chains are long in terms of length and time.

This project particularly aims at deriving routes of tanker trucks that carry fuel commodities from PEV to fuel recipients (including gas stations and other fuel recipients) in the 12-county region. To this end, it was important to identify those trip chains that visit PEV at least once. There were 807 trip chains out of 1,320 total trip chains that visited PEV at least once. Figure 7.16 illustrates the distribution of trip chains among number of trips for 807 trip chains that visit PEV at least once. Figure 7.17 shows a similar distribution to Figure 7.16 for trip chains with 5 trips or less. Comparing Figure 7.16 with 7.9 reveals that trip chains visiting PEV generally have higher number of trips. This is encouraging because trip chains that visit PEV are of interest in the context of this project. Additionally, such trips are less likely to be broken due to the study boundary and therefore, they contain higher number of trips. Figure 7.18 and 7.19 illustrate distributions of trip chain length and trip chain time, respectively, for 807 trip chains visiting PEV. Comparing these figures with those of 1,320 trip chains shows that 807 trip chains are relatively longer both in terms of trip length and trip time. This is expected because trip chains that visit PEV contain higher number of trips compared to all 1,320 trip chains. Table 7.5 illustrates the location description distribution of origins and destinations of 807 trip chains. Shares of PEV and gas station in this table are relatively higher than other location descriptions. This is expected because these trip chains visit PEV at least once and therefore, they contain fuel delivery trips. Table 7.6 shows the location description distribution of other 513 trip chains that did not visit PEV. In this table, the percentages of “On the road” are quite high for both origins and destinations. On the other hand, shares of PEV and gas stations are relatively low compared

to Table 7.5. This is because these trip chains were originally part of a bigger chain but were cut because their corresponding trucks had crossed the study boundary.

Figure 7.14 and 7.15 show profiles of starting time and ending time of 1,320 trip chains, respectively. As expected, there is a spike at 8:00 AM in Figure 7.14 which corresponds to the AM peak. There is also a spike in Figure 7.15 around 17:00 PM which corresponds to the PM peak. In both figures there are spikes close to midnight. This is because the stream of truck data usually starts or ends around midnight (12:00 AM). These figures show that tanker truck trip chains start or end around usual AM and PM peak hours, respectively.

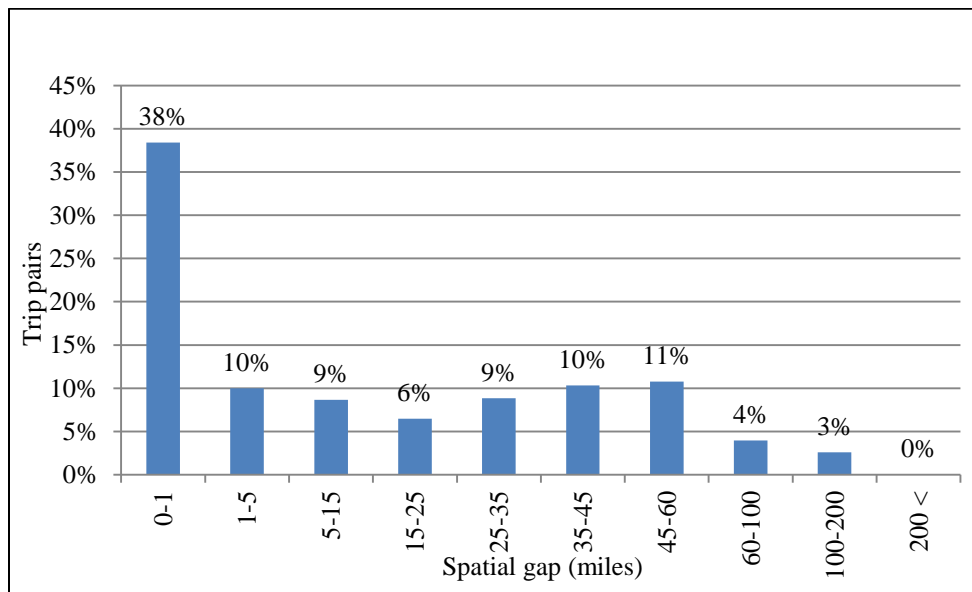


Figure 7.8 Spatial gap distribution for non-matching trip pairs (N = 1,744 trip pairs)

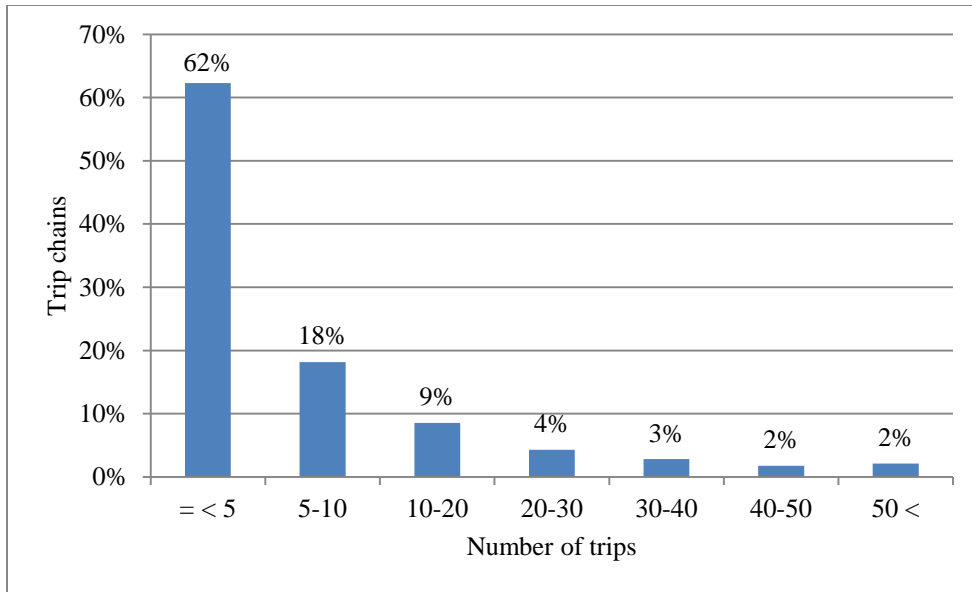


Figure 7.9 Distribution of trip chains among number of trips (N = 1,320 trip chains)

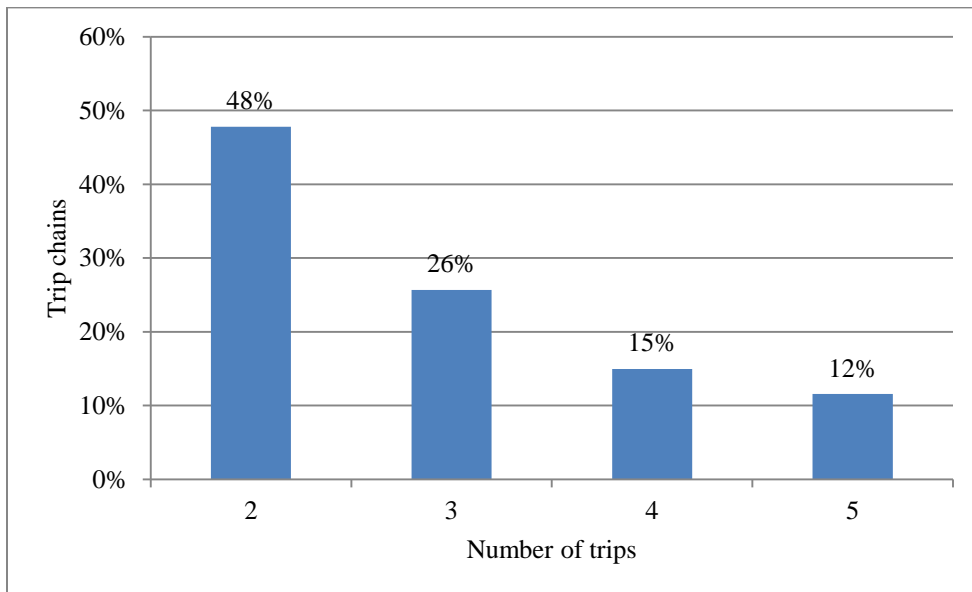


Figure 7.10 Distribution of trip chains among number of trips (number of trips =< 5) (n=822 trip chains)

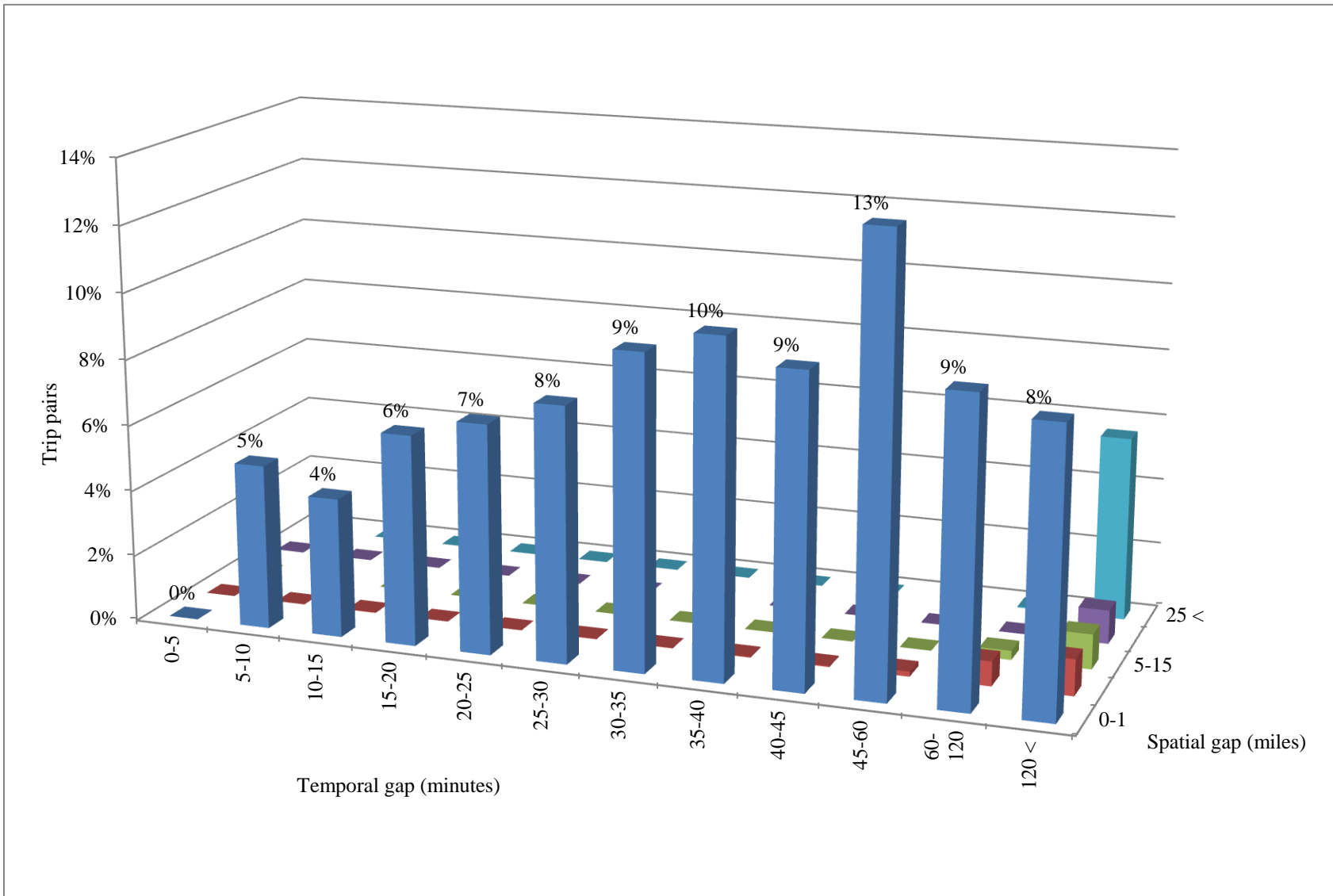


Figure 7.11 3-D histogram of spatial gap vs. temporal gap (N = 12,538 trip pairs)

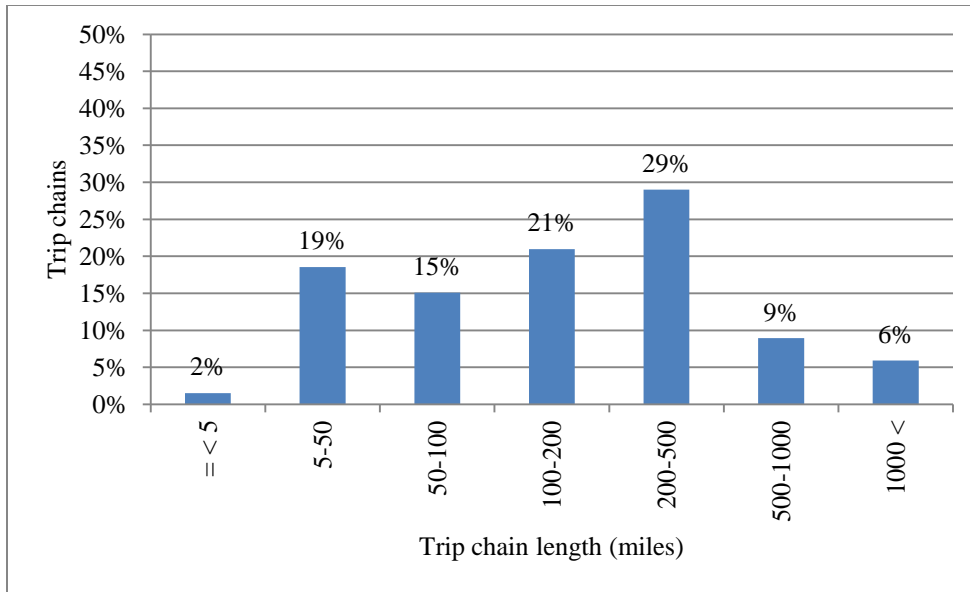


Figure 7.12 Trip chain length distribution (N=1,320 trip chains)

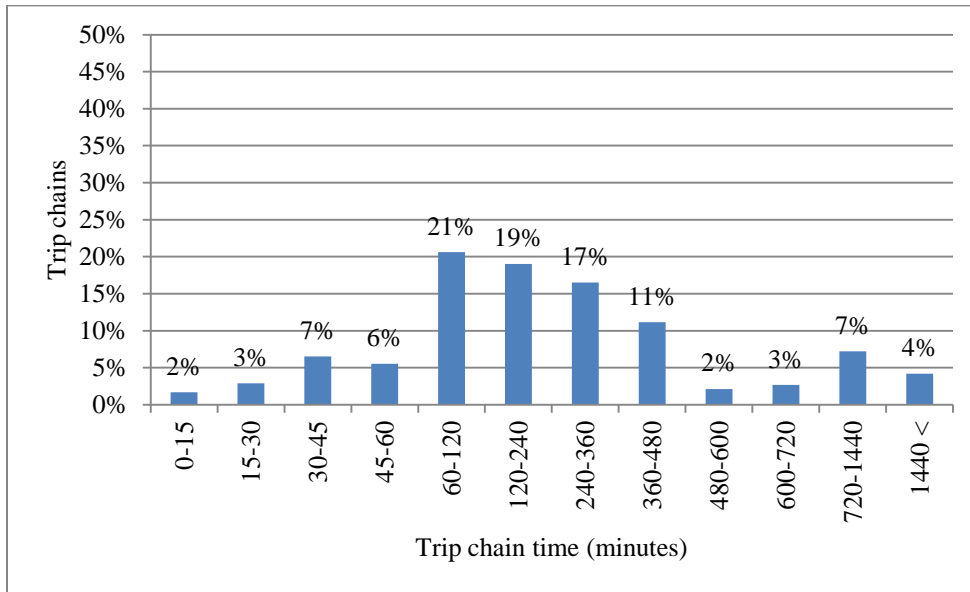


Figure 7.13 Trip chain time distribution (N = 1,320 trip chains)

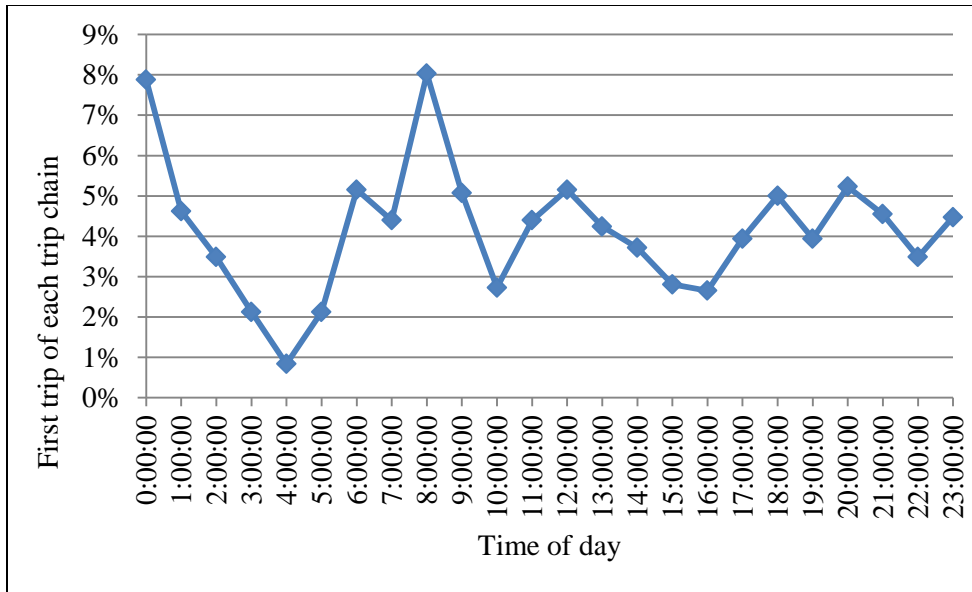


Figure 7.14 Profile of starting time of the trip chains (N = 1,320 trip chains)

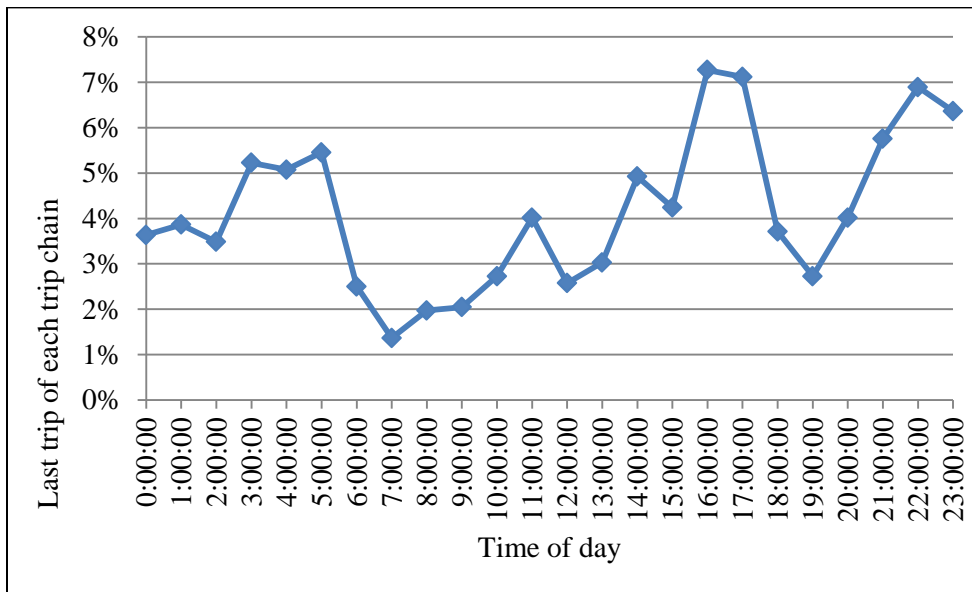


Figure 7.15 Profile of ending time of the trip chains (N = 1,320 trip chains)

Table 7.5 Location distribution for starting and ending points of 807 trip chains that have visited PEV at least once

Case	Location	Number of trip chain origins	Number of trip chain destinations
1	Gas station	249 (30.9 %) trip chain origins	153 (19.0 %) trip chain destinations
2	PEV	388 (48.1 %) trip chain origins	369 (45.7 %) trip chain destinations
3	Other fuel recipients	66 (8.2 %) trip chain origins	66 (8.2 %) trip chain destinations
4	Distribution center	55 (6.8 %) trip chain origins	52 (6.4 %) trip chain destinations
5	On the road	32 (4.0 %) trip chain origins	159 (19.7 %) trip chain destinations
6	Rest stop	17 (2.1 %) trip chain origins	8 (1.0 %) trip chain destinations
	Total	807 (100 %) trip chain origins	807 (100 %) trip chain destinations

Table 7.6 Location distribution for starting and ending points of 513 trip chains that do not visit PEV

Case	Location	Number of trip chain origins	Number of trip chain destinations
1	Gas station	181 (35.3 %) trip chain origins	46 (9.0 %) trip chain destinations
3	Other fuel recipients	28 (5.5 %) trip chain origins	14 (2.7 %) trip chain destinations
4	Distribution center	47 (9.2 %) trip chain origins	27 (5.3 %) trip chain destinations
5	On the road	174 (33.9 %) trip chain origins	388 (75.6 %) trip chain destinations
6	Rest stop	83 (16.2 %) trip chain origins	38 (7.4 %) trip chain destinations
	Total	513 (100 %) trip chain origins	513 (100 %) trip chain destinations

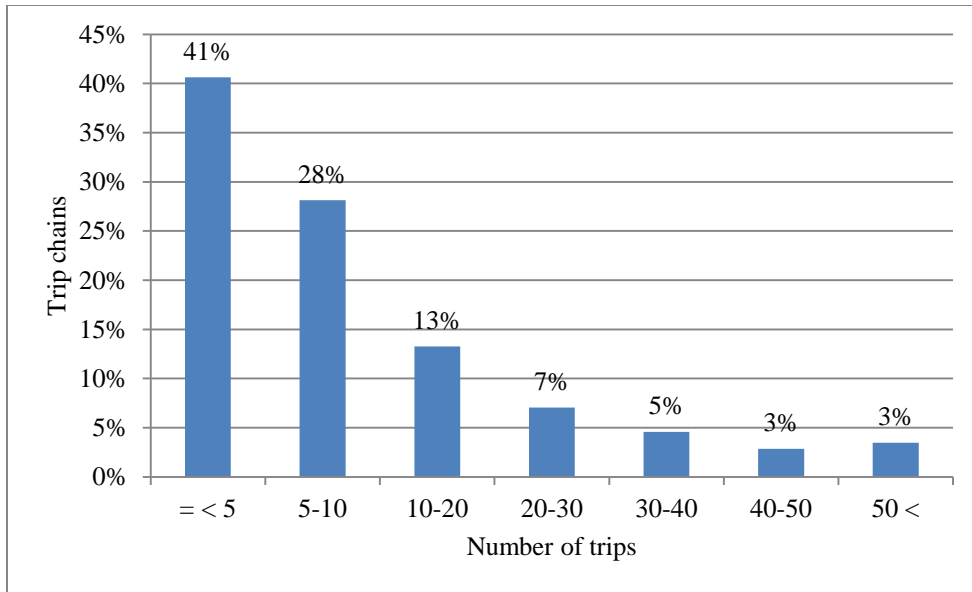


Figure 7.16 Distribution of number of trips per trip chain for trip chains that visit PEV at least once (N = 807 trip chains)

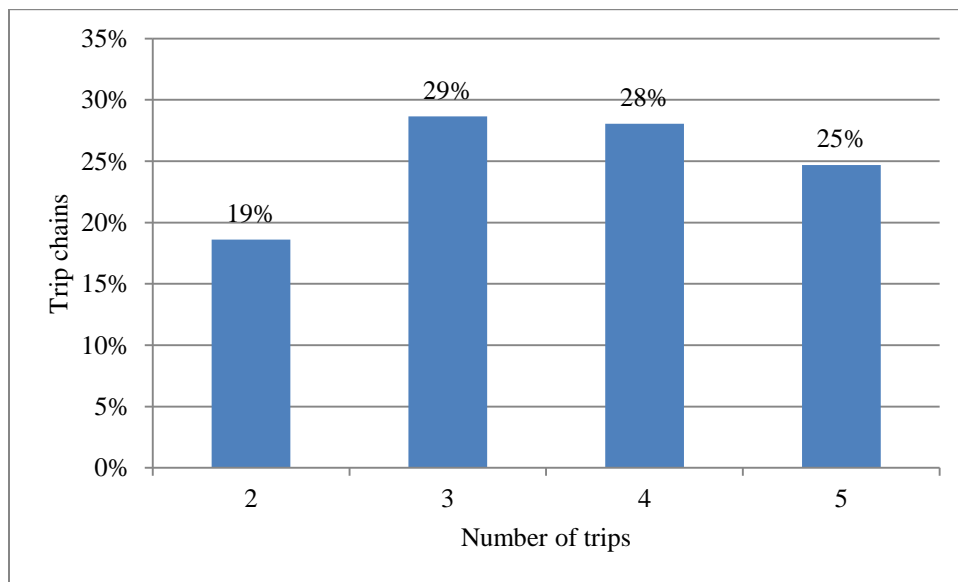


Figure 7.17 Distribution of number of trips per trip chain for trip chains that visit PEV at least once (N = 328 trip chains)

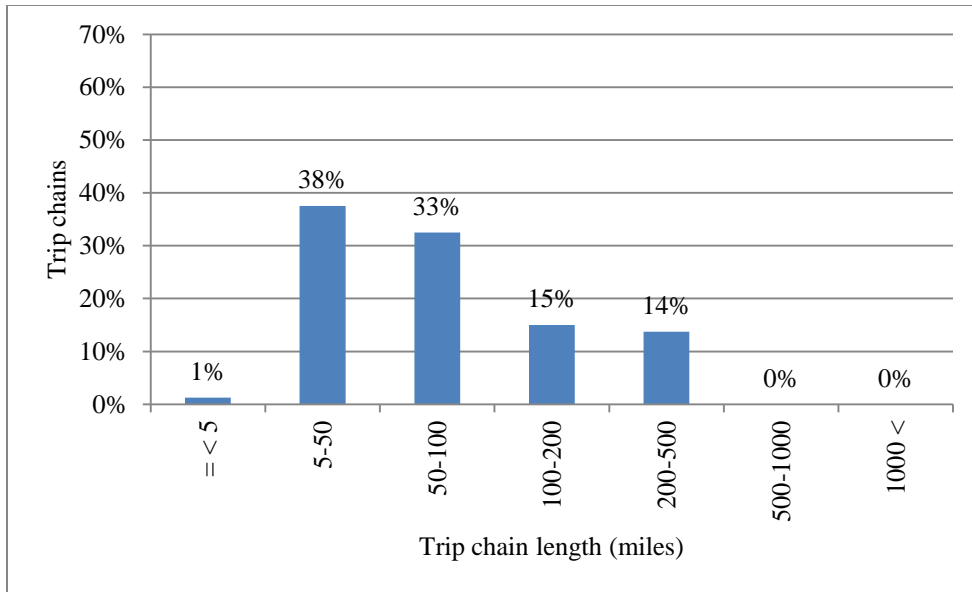


Figure 7.18 Trip chain length distribution for trip chains that visit PEV at least once (N=807 trip chains)

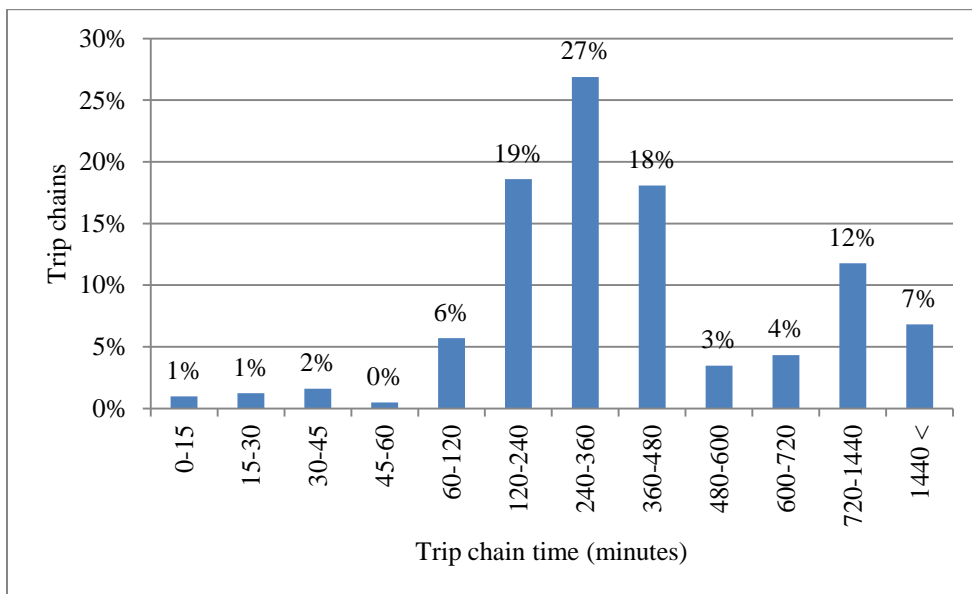


Figure 7.19 Trip chain time distribution for trip chains that visit PEV at least once (N = 807 trip chains)

CHAPTER 8 : DERIVING TRIP ROUTES

8.1 Introduction

The ultimate objective of this project was to derive routes for tanker trucks that load fuel commodities at PEV. 14,598 tanker truck trips were extracted from GPS data, cleaned, and analyzed in previous steps. Subsequently, 1,320 trip chains were built for which routes had to be generated using the techniques presented in the first part of this thesis. This part consisted of two steps, namely, map-matching and route generation. These two steps are explained in the coming sections.

8.2 Map-matching Algorithm

The first step toward deriving truck routes using GPS data is map-matching. The methodology developed in Chapter 4 was implemented here with a few changes. First, an extensive data preparation like Section 4.3 was not necessary here. This is because such data preparation eliminates many small trips in trip chains. Eliminating such trips would break the continuity of trip chains. However, a minor data processing was done in order to improve the quality of derived routes. This data processing removed all the GPS points within 30 ft of origin and destination of each trip (a similar procedure to Stage 5 of Section 4.3). Doing so reduced the chance of wrong route estimations near origins or destinations while kept most of the GPS points in the dataset.

Second, in the map-matching algorithm used in the project a buffer zone of 500 ft was used as opposed to the previously-used 1000 ft in Step 2 of Section 4.4 . This is because tanker

trucks mostly make trips in urban areas as opposed to rural. Therefore, they are mostly observed in dense places on the network where the distance between links is usually small. As a result, a smaller buffer zone has to be used in Step 2 of map-matching to decrease the chance of snapping GPS points to wrong links in dense areas.

11,907 trips were successfully map-matched using the above-mentioned algorithm. Only 11 trips from the original 11,918 trips were missed due to elimination of some GPS points in the algorithm. The next step in deriving routes was generating the routes using map-matched GPS points.

8.3 Route Generation

Due to the infrequent nature of the GPS data in this project, the map-matching algorithm explained above does not capture all links in a trip. Consequently, missing links need to be found so that a route can be generated for each trip. To this end, ArcMap 10.3 Network Analyst extension was employed to generate the routes for map-matched GPS points. Network Analyst utilizes a modified version of Dijkstra's algorithm (2) to find shortest paths between two points. For each trip, the shortest path between consecutive GPS points was found based on minimizing travel time.

The final output of route generation was a GIS shapefile in which each feature is a route link that contains trip information. Figure 8.1 shows an overall view of the generated routes for 11,907 trips that belong to 1,320 trip chains in the 12-county region.

8.4 Validation

Generated routes were validated in terms of feasibility and consistency. Routes are consistent if:

- 1) The direction of the travel is consistent throughout the entire route

- 2) There are no loops throughout the entire route
- 3) There are no missed links in routes

And feasible if and only if:

- 4) There are no impossible maneuvers throughout the entire route (impossible maneuvers such as jumping off a bridge)

50 trips were selected and then followed on Google Earth for validation checks.

Consistency and feasibility checks are done simultaneously when the route is followed on Google Earth. To check the consistency, each generated route was compared to the route from Google Earth to determine if the generated route shows the same direction through the entire trip. For feasibility check, each trip is observed at interchanges or overpasses or ramp junctions to see if the generated route shows any impossible maneuvers at such locations. To check the connectivity, each route was checked for any missing links while being followed on Google Earth. Table 8.1 lists all the 50 trips that were selected for feasibility, consistency, and connectivity checks with their corresponding trip time and trip length information. The last column from left in this table illustrates the status of each trip with regard to feasibility and consistency checks. As can be observed, all the 48 out of 50 routes are marked as “Ok” which means they are all feasible, consistent, and connected. Results from validation checks show that only two trips out of 50 had a loop in their derived routes which is satisfactory. Figure 8.2 illustrates the trip length distribution of trips selected for validation checks. As can be observed, the distribution encompasses almost all type of trip lengths in order to be a good representative of the population.

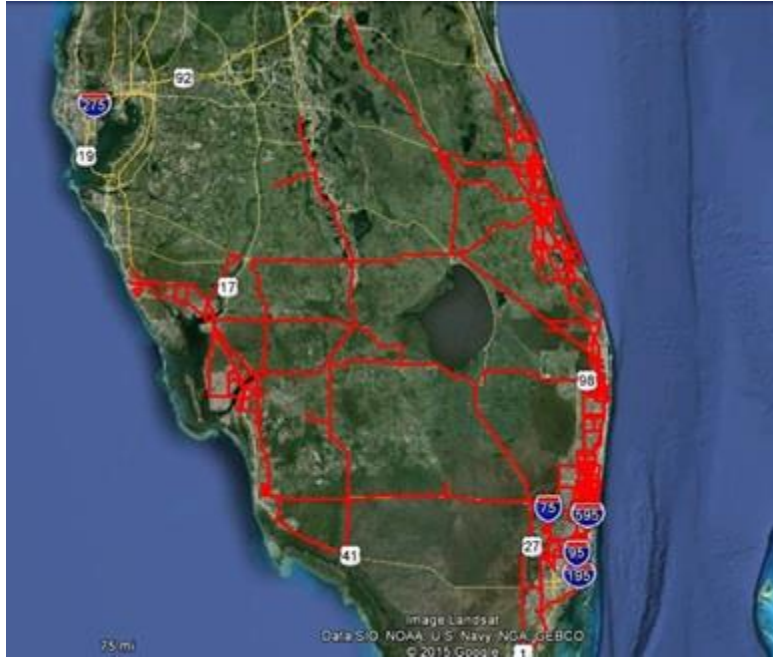


Figure 8.1 Final 11,907 derived routes

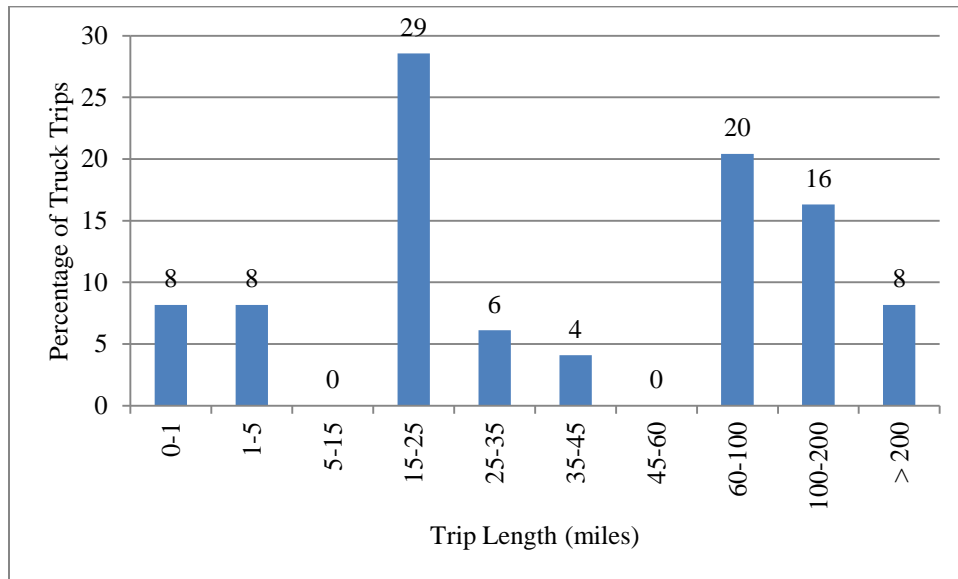


Figure 8.2 Trip length distribution of 50 trips selected for validation checks

Table 8.1 General trip information for 50 routes selected for feasibility, consistency, and connectivity checks

Case Number	Route number	Trip Time(minutes)	Trip Length(miles)	Status
1	2	97.3	94	Ok
2	3	63.9	62	Ok
3	4	1.2	4	Ok

Table 8.1 (Continued)

4	6	0.2	1	Ok
5	7	3.0	15	Ok
6	9	0.1	1	Ok
7	10	0.8	5	Ok
8	11	79.4	86	Ok
9	12	21.5	27	Ok
10	13	102.8	120	Ok
11	14	22.2	43	Ok
12	15	24.8	56	Ok
13	16	3.0	10	Ok
14	17	79.4	89	Ok
15	18	60.8	70	Ok
16	19	138.7	154	Ok
17	20	21.7	39	Ok
18	21	21.1	42	Ok
19	23	78.9	89	Ok
20	24	21.4	26	Ok
21	26	96.1	115	Ok
22	27	99.9	110	Ok
23	28	40.4	44	Ok
24	29	136.6	138	Ok
25	30	0.4	2	Ok
26	31	135.6	126	Ok
27	32	28.8	32	Ok
28	33	100.0	93	Ok
29	34	65.7	63	Ok
30	35	23.3	37	Ok
31	36	22.4	35	Ok
32	37	33.7	59	Ok
33	38	2.3	7	Ok
34	39	24.3	41	Not Ok
35	46	36.9	70	Ok
36	47	33.3	46	Ok
37	48	23.0	44	Ok
38	49	20.6	29	Ok
39	50	15.9	28	Ok
40	1993	212.3	256	Not Ok
41	2248	224.7	200	Ok
42	5836	176.7	155	Ok
43	6082	202.2	180	Ok
44	6091	181.9	184	Ok
45	6280	205.6	191	Ok
46	6372	177.1	156	Ok
47	10009	59	43	Ok
48	11678	183.5	283	Ok
49	12010	12	8.2	Ok
50	12088	35	24	Ok

Figure 8.3 illustrates an example of a route that has been followed on Google Earth for feasibility, consistency, and connectivity checks.

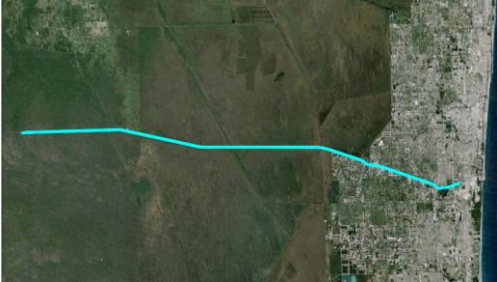

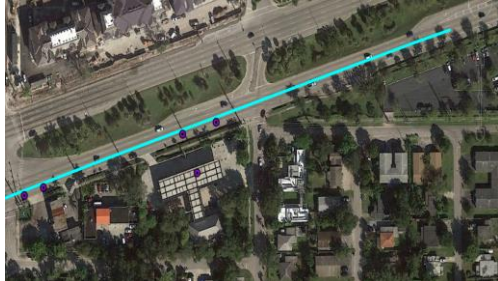
Trip number	3
The whole route	
Origin	
Overpass	
Destination	

Figure 8.3 An example of following a route for feasibility and consistency check

CHAPTER 9 : CONCLUSION OF THE SECOND PART

9.1 Conclusion

The second part of this thesis showed the successful implementation of the route generation procedures developed in the first part for an FDOT sponsored project. The goal of the project was to use ATRI GPS data to derive routes for tanker trucks' serving Port Everglades and 12 counties in southern Florida.

9.2 Gathering ATRI's Truck GPS Data and Separating Petroleum Tanker Trucks

The project resulted in combining objective 1 and objective 2 in one step process. Tanker truck trips originate from certain fuel terminals at PEV. Tanker trucks load fuel commodities at these terminals and then carry the fuel to their destinations. A polygon was drawn around each terminal in PEV and sent to ATRI so that ATRI would extract only those trucks' GPS points that fell inside the polygons. ATRI provided the research team with GPS data from two months of September 2014 and March 2015. The polygon technique mostly eliminated the chance of other trucks' GPS data being in the final dataset. Further analysis of tanker truck trips revealed that there were still some trucks in the data that did not predominantly make fuel delivery trips. Such trucks and their respective trips were removed before trip routes were derived.

9.3 Derive Trip Chains of Trucks Originating at PEV

1,320 trip chains were derived that corresponded to 11,918 trips and 95 unique truck IDs. 807 trip chains out of 1,320 visited PEV at least once. Such trip chains mostly started (ended) in either PEV or gas stations. 513 trip chains out of 1,320 did not visit PEV and they were mostly

incomplete because their respective truck had exited the 12-county study area (12-county study area consisted of: Miami-Dade, Broward, Palm Beach, Monroe, Martin, St. Lucie, Indian River, Okeechobee, Glades, Hendry, Lee, and Collier Counties). The effort to derive trip chains included the steps below.

9.3.1 Modifying Trip Extraction Algorithm and Extracting Truck Trips

The raw GPS data was converted into a database of truck trips. The algorithms developed previously by Thakur et al. (2) were utilized in this step. However, the algorithms were developed primarily for the purpose of deriving long-haul trips. In this project these algorithms were modified so that urban trips could be extracted from raw GPS data. Running the algorithm resulted in 14,598 truck trips from two months of data.

9.3.2 Identifying Characteristics of Truck Trips

One of the outcomes of this project was calculating trip measurements such as trip length, trip time, and trip speed for extracted trips. This information can be used for further tanker truck travel modelling and analysis. Moreover, one of the important outcomes of this project was identifying land use description of origins and destinations of all 14,598 trips. This has been done through developing an algorithm for grouping trip ends in “clusters” and then identifying each cluster’s land use using Google Earth.

9.3.3 Rectifying and Enriching Existing Data on Fuel Recipients

The project consultants provided with two sources of data on fuel recipients in 12-county region, namely Department of Revenue (DOR) data and HERE data. Both datasets were incomplete in terms of active gas station coverage in the 12-county region. In addition, both datasets showed some anomalies with regard to geocoded gas stations or other fuel recipients. This project resulted in correcting wrongly geocoded gas stations or other fuel recipients in both

datasets and also adding around 100 new gas stations to the dataset through analyzing tanker truck trips' destinations. The rectified and enriched final fuel recipient dataset that included DOR, HERE, and new gas station data, was produced in form of a GIS shapefile.

9.4 Deriving Trip Routes

This project resulted in deriving 11,907 trip routes out of 11,918 trips belonging to 1,320 trip chains. An effective yet simple algorithm developed in the first part of this thesis was used to derive these routes. Derived routes were later converted into route links and were provided in form of a GIS file. This GIS file which was delivered to FDOT district 4 as the final product included some trip level information such as trip length, trip time, origin or destination land use description, etc.

9.5 Opportunities for Future Research

The work done in this project can be extended in a few directions. First, extracted trips from this project can be used for further travel behavior analyses such as origin-destination matrix estimation (ODME). Such analyses provide valuable insights into fuel commodity flows throughout the region. Moreover, the results from the project can be used to build tanker truck route choice sets. Route choices sets then can be further explored for tanker truck route choice modelling. Route choice modelling can lead to important interpretations regarding which routes are usually used by tanker trucks or what incentives impact tanker trucks to choose a particular route. The output of such analysis will be very useful for freight policy makers and stakeholder.

Another opportunity that results from this study provide is the chance to introduce route variability measures specifically for tanker trucks. These measurements can be utilized to improve tanker truck choice set generation models. Rich choice set generation models then open the path for better modelling of tanker trucks' route choice.

The fact that the data used in this project is specific to tanker trucks opens this opportunity to compare the results with similar studies that include all types of trucks. The focus of such comparisons will be if the travel behavior of tanker trucks is significantly different from other types of trucks. The interpretations and insights from such efforts can be used by researchers and freight policy makers.

REFERENCES

- 1) Strocko, E., Sprung, M., Nguyen, L., Rick, C., & Sedor, J. (2014). Freight Facts and Figures 2013 (No. FHWA-HOP-14-004).
- 2) Thakur, A., Zanjani, A. B., Pinjari, A. R., Short, J., Mysore, V., & Tabatabaee, S. F. (2015). Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. In Transportation Research Board 94th Annual Meeting (No. 15-6031).
- 3) Quddus, M. A., Ochieng, W. Y., & Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5), 312-328.
- 4) Kim, J. S. (1996). Node based map-matching algorithm for car navigation system. In *International Symposium on Automotive Technology & Automation (29th: 1996: Florence, Italy)*. Global deployment of advanced transportation telematics/ITS.
- 5) Ochieng, W. Y., Quddus, M., & Noland, R. B. (2003). Map-matching in complex urban road networks. *Revista Brasileira de Cartografia*, 2(55).
- 6) Yang, J. S., Kang, S. P., & Chon, K. S. (2005). The map-matching algorithm of GPS data with relatively long polling time intervals. *Journal of the Eastern Asia Society for Transportation Studies*, 6, 2561-2573.
- 7) Dhakar, N. S. (2012). *Route Choice Modeling Using GPS Data* (Doctoral dissertation, University of Florida).
- 8) Xu, H., Liu, H., Tan, C. W., & Bao, Y. (2010). Development and application of an enhanced Kalman filter and global positioning system error-correction approach for improved map-matching. *Journal of Intelligent Transportation Systems*, 14(1), 27-36.
- 9) Chen, B. Y., Yuan, H., Li, Q., Lam, W. H., Shaw, S. L., & Yan, K. (2014). Map-matching algorithm for large-scale low-frequency floating car data. *International Journal of Geographical Information Science*, 28(1), 22-38.
- 10) Hess, S., Quddus, M., Rieser-Schüssler, N., & Daly, A. (2015). Developing advanced route choice models for heavy goods vehicles using GPS data. *Transportation Research Part E: Logistics and Transportation Review*, 77, 29-44.

- 11) Jagadeesh, G. R., Srikanthan, T., & Zhang, X. D. (2004). A map-matching method for GPS based real-time vehicle location. *Journal of Navigation*, 57(03), 429-440.
- 12) DoD, U. S. (2001). Global positioning system standard positioning service performance standard. Assistant secretary of defense for command, control, communications, and intelligence.
- 13) Nadine Schuessler, I. V. T., & Axhausen, K. W. (2009). Accounting for Route Overlap in Urban and Sub-urban Route Choice Decisions Derived from GPS Observations.

APPENDIX A: POLYGONS AROUND PORT EVERGLADES FUEL TERMINALS

This appendix provides all the polygons drawn around fuel terminals at Port Everglades. The polygons are shown in red and the actual fuel terminals are shown in yellow circles. These red polygons were used to separate tanker truck GPS data from other types of trucks' GPS data. Basically, ATRI provided the research team with only those GPS points that fell inside the red polygons. Since mostly tanker trucks stop at the fuel terminals, by capturing the GPS points falling in the red polygons, one can capture tanker truck GPS data.



Figure A.1 Terminal 1 (in yellow circle) and the red polygon used for GPS data extraction



Figure A.2 Terminal 2 (in yellow circle) and the red polygon used for GPS data extraction



Figure A.3 Terminal 5 (in yellow circle) and the red polygon used for GPS data extraction

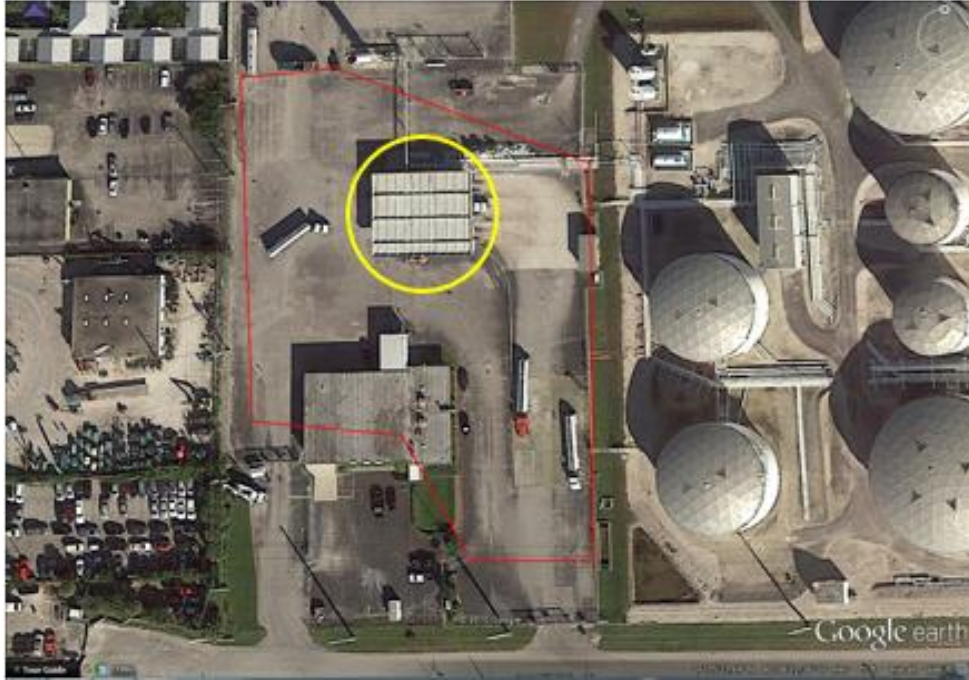


Figure A.4 Terminal 7 (in yellow circle) and the red polygon used for GPS data extraction



Figure A.5 Terminal 11 (in yellow circle) and the red polygon used for GPS data extraction



Figure A.6 Terminal 12 (in yellow circle) and the red polygon used for GPS data extraction

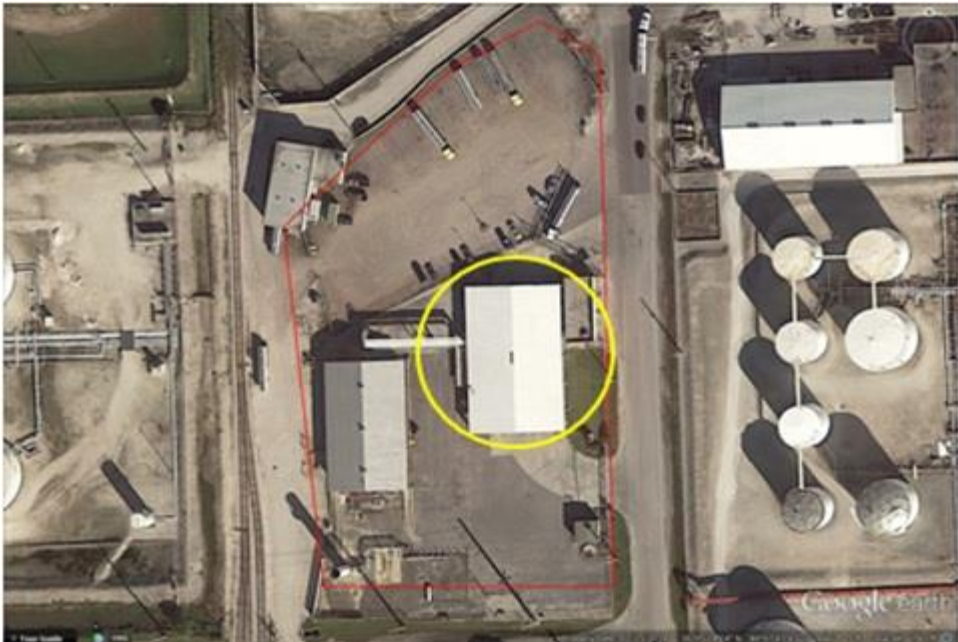


Figure A.7 Terminal 13 (in yellow circle) and the red polygon used for GPS data extraction



Figure A.8 Terminal 14 (in yellow circle) and the red polygon used for GPS data extraction

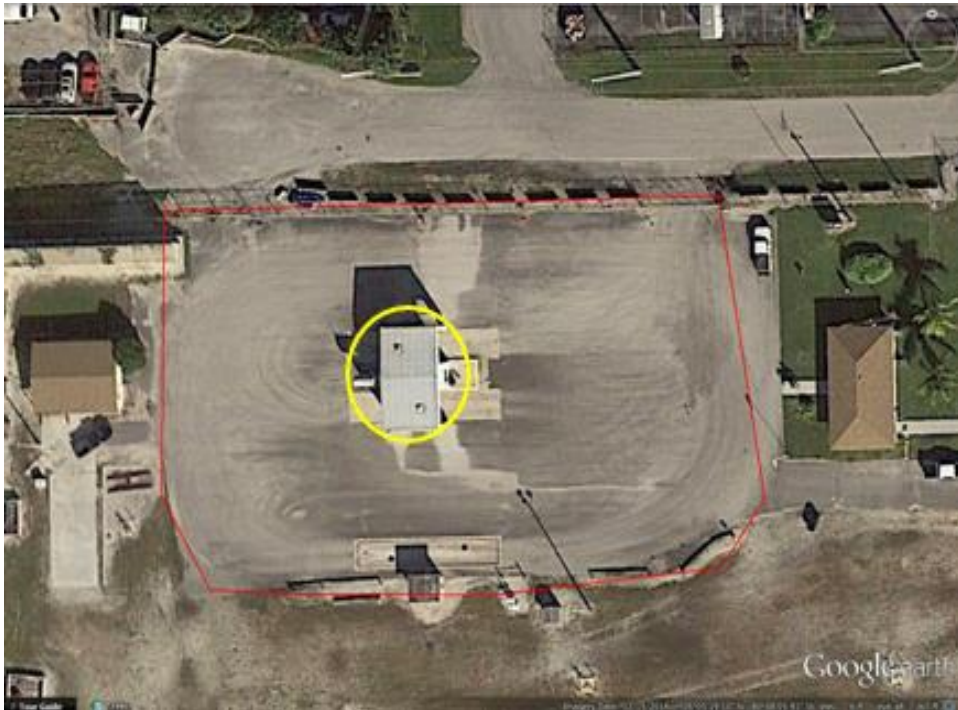


Figure A.9 Terminal 15 (in yellow circle) and the red polygon used for GPS data extraction

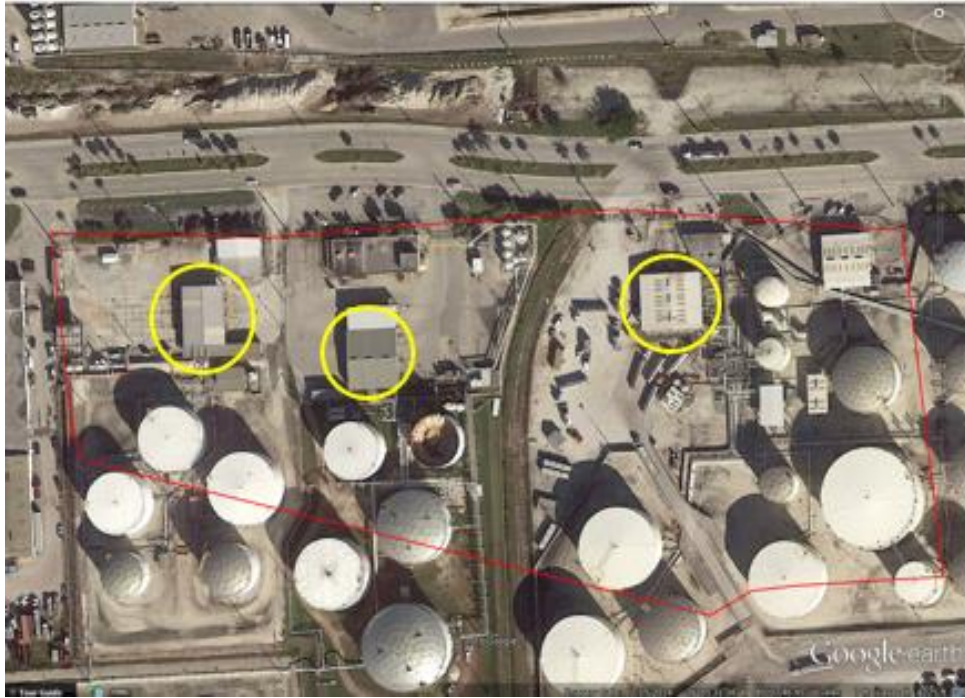


Figure A.10 Terminals 3, 4, and 10 (in yellow circle) and the red polygon used for GPS data extraction



Figure A.11 Terminals 6 and 9 (in yellow circle) and the red polygon used for GPS data extraction

APPENDIX B: TRUCK TRIP CHARACTERISTICS

This appendix provides distributions of trip length, trip time, and trip average speed for extracted trips from two months of data. The above-mentioned distributions are provided for each data type (i.e., D1 data and D2 data) separately. These distributions can be used for tanker-truck modelling purposes in the 12-county region in Florida.

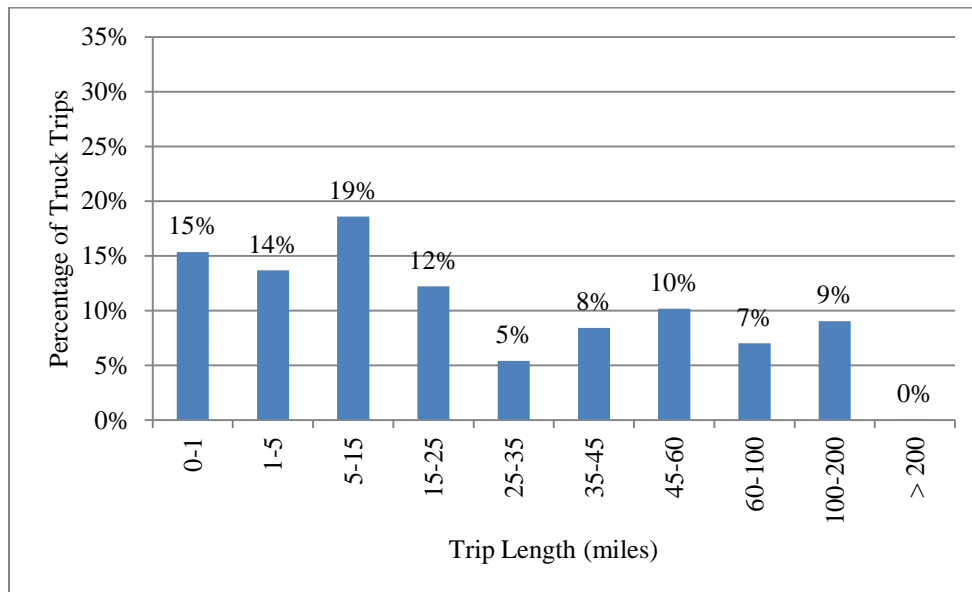


Figure B.1 Trip length distribution for all trips extracted from D2 data (N = 14,162 trips)

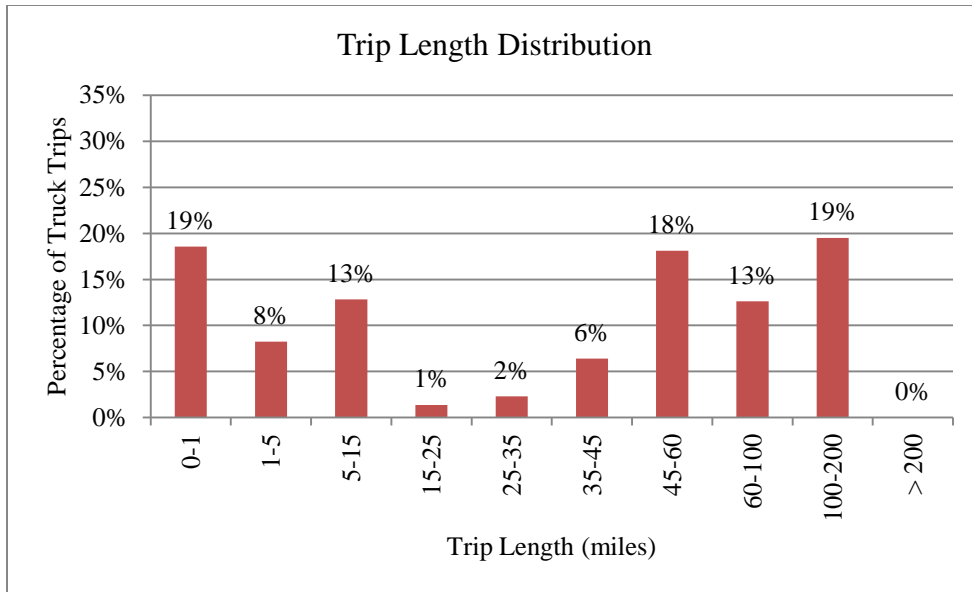


Figure B.2 Trip length distribution for all trips extracted from D1 data (N = 436 trips)

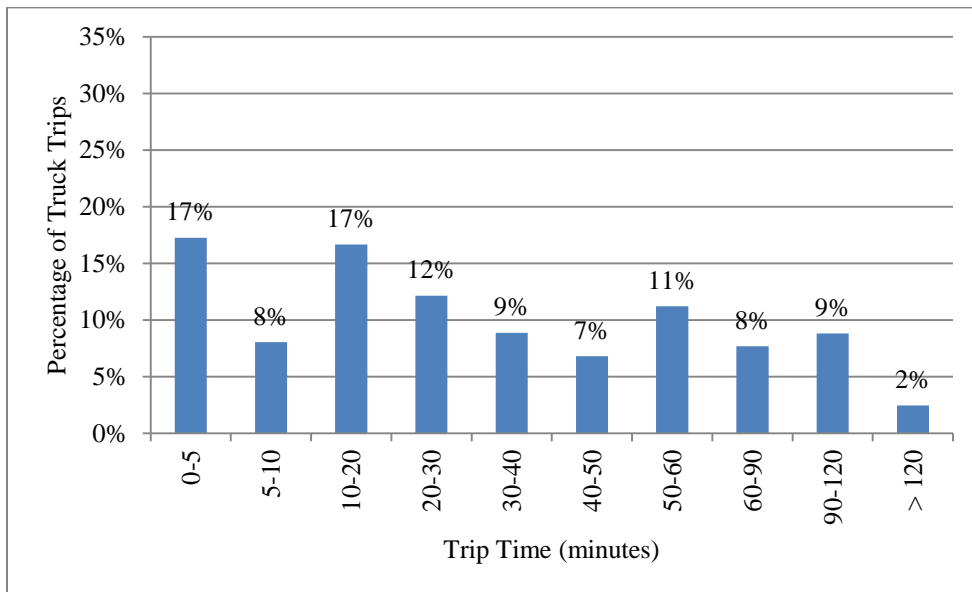


Figure B.3 Trip time distribution for all trips extracted from D2 data (N = 14,162 trips)

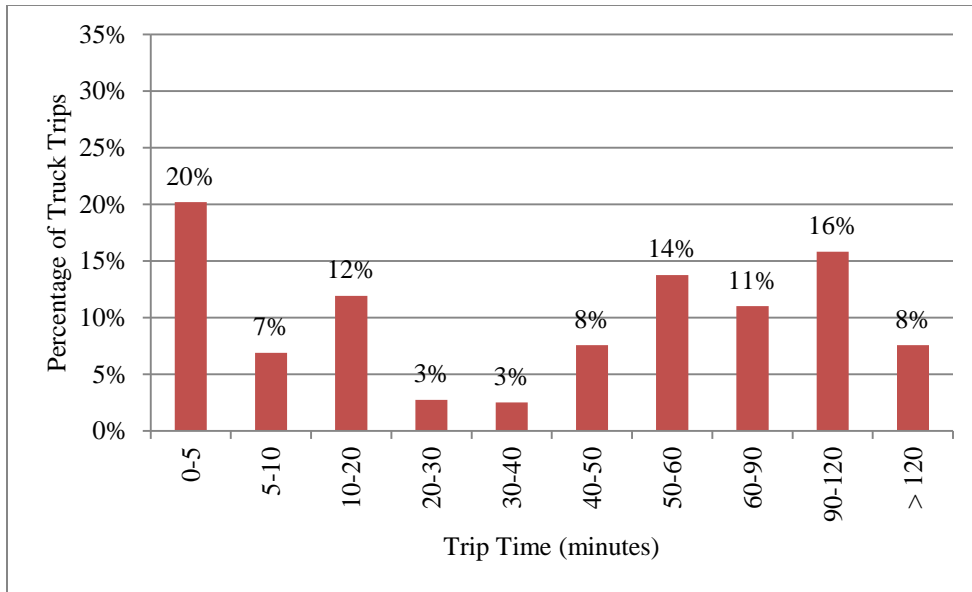


Figure B.4 Trip time distribution for all trips extracted from D1 data (N = 436 trips)

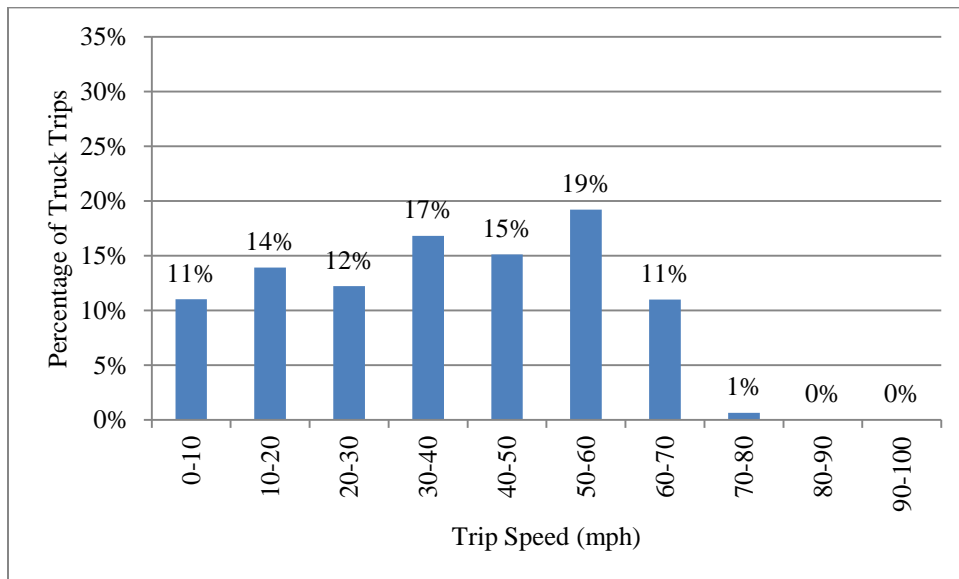


Figure B.5 Trip average speed distribution for all trips extracted from D2 data (N = 14,162 trips)

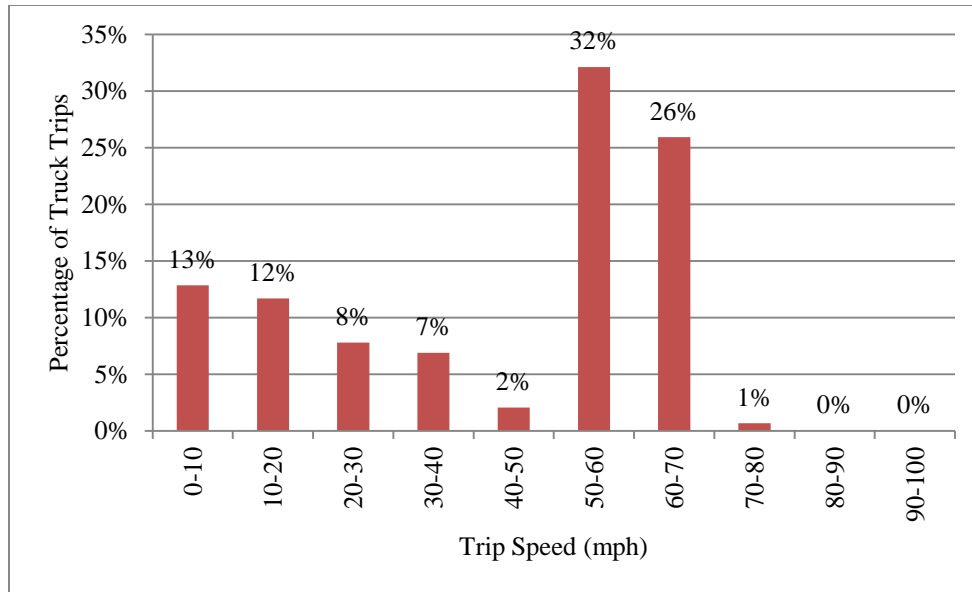


Figure B.6 Trip average speed distribution for all trips extracted from D1 data (N = 436 trips).

APPENDIX C: TRIP FILE CLEANING TASKS

This appendix presents trip cleaning tasks in details. There were 14,598 total trips extracted from two months of ATRI data, namely September 2014 and March 2015. Some of these trips were non-fuel delivery trips, or had trip ends on the roadway. This appendix explains the four different tasks taken to address issues such as non-fuel delivery trips or on-the-road trip ends.

C.1 Removing Non-Fuel Delivery Trips

In this project the main goal is to derive the routes for fuel-delivery trips going from (to) PEV. Since fuel-delivery trips are made by tanker trucks, tanker trucks' GPS data had to be separated from other trucks' GPS data. Currently ATRI do not offer GPS data classified based on the type of truck. Therefore, a solution had to be proposed to only capture tanker trucks' GPS data from the pool of ATRI data.

The proposed solution was drawing polygons around fuel terminals in PEV and selecting only those GPS data that fell inside the polygons. Although this method proved to be effective, there were still some GPS data obtained from ATRI that belonged to other types of trucks rather than tanker trucks. Such GPS data corresponded to non-fuel-delivery trips made by non-tanker. Therefore, non-tanker trucks and their respective trips had to be removed.

27 truck IDs were identified that seemed no to belong to tanker trucks. These truck IDs had either of the following features:

- 1) Mainly served distribution centers as opposed to gas stations and other fuel recipients

2) Visited PEV relatively fewer times than gas stations or distribution centers

These trucks either mainly served distribution centers or visited gas stations much more frequently than PEV. The latter happened probably because these trucks needed to buy fuel while the former suggests that these trucks were not tanker trucks. Either case, such truck IDs had to be removed from the data.

Figure C.1 shows the trip length distribution of trips belonging to truck IDs removed from the data. This figure shows that the majority (60%) of these trips are short trips (less than 5 miles). Therefore, it is unlikely that these trips belong to tanker trucks that deliver fuel commodities from PEV to fuel recipients. Moreover, a significant portion (26 %) of trips in Figure C.1 is less than 1 mile. This means that these trips were most probably happening at the same location and therefore, were better to be removed.

C.2 Solving Trip Ends on the Road

Table C.1 shows that some portions of trip ends fell on the roadway. This issue happened mainly because a study boundary was imposed on the GPS data and therefore, when trucks exited the study area their stream of GPS data was cut off until they entered the study area again. If the origin (destination) of the first (last) trip of a truck ID was on the roadway, that trip was considered incomplete. Such incomplete trips were removed because the first (last) trip of a truck ID had to start (end) at a valid location.

36 trips were identified that were the starting (ending) trip of a truck ID and their origins (destinations) were on the roadway. These trips were removed immediately after the procedure in Section C.1. Table C.3 shows the location description distribution of remaining trips after removing non-tanker truck IDs (and their respective trips) and the above-mentioned 36 trips. As

can be observed, the percentages of “Distribution center” have dropped from 13.1 % in Table C.1 to around 5% in Table C.2.

C.3 Joining/Replacing Trip Ends on the Roadway

There were 1,273 trip destinations and 378 trip origins on the roadway after removing non-tanker trucks and 36 trips (after steps taken in Section C.1 and Section C.2). There were three different cases for trip ends on the roadway. Case (1): current trip destination on the roadway and the next trip origin in a valid location, case (2): current trip origin on the roadway and the previous trip destination in a valid location, and case (3): current trip destination and next trip origin both on the roadway. These three cases were addressed as below:

- 1) For case (1): replace 110 on-the-road destinations by the next valid origin within 1 mile distance
- 2) For case (2): replace 34 on-the-road origins by the previous valid destination within 1 mile distance
- 3) For case (3): Join 34 trip pairs in which the current trip destination and next trip origin both are on the roadway. There were 309 trip pairs with case (3) condition, 34 out of 309 trip pairs were within one mile and 45 minutes of each other.

Having addressed three different cases where trip ends fell on the roadway, it is worth mentioning why this phenomenon happened. The list below explains the reasons why this many trip ends were observed on the roadway:

- 1) **The study boundary:** This is the primary reason for observing trip ends on the roadway. The stream of GPS data for trucks traveling outside the study area was cut by the study boundary. Therefore, there were some incomplete trips whose destinations were on the roadway and close to the study boundary. Figure C.2 shows all the destinations falling on

the roadway and the 12-county study boundary (red polygon) implemented by ATRI. These destinations were also observed on the north bound of the roadway. That means the stream of data was cut when the trucks were going out of the boundary. Similarly, Figure C.3 shows all the origins falling on the roadway. Red polygon is the study boundary implemented by ATRI. Close observations using Google Earth showed that almost all of on-the-road origins on I-75 and I-95 were on the south bound of the roadway. That means GPS data stream was resumed as soon as trucks had entered the study area.

- 2) Slow movements of trucks: If a truck moves slower than 5 mph for more than 5 minutes then it is considered that the truck has stopped. Some trucks met this criterion in traffic stops and therefore, their trip origins or destinations were captured on the road. This case was usually observed far from the study boundary and mostly in urban areas. The criteria for joining on-the-road trips were designed to resolve on-the-road trip ends that happen for slow movement of trucks.

C.4 Joining Truck IDs from D1 Data

In D1 data truck IDs rotate every 24 hours. Therefore, there might be some tanker trucks that had appeared in D1 data with two different truck IDs in two consecutive days. Since it was important in this project to follow the chain of tanker trucks as much as possible, a task was created to identify tanker trucks that appeared with different truck IDs in two or more consecutive days in D1 data.

To this end, the spatial gap and temporal gap between truck IDs in D1 data had to be observed. The spatial gap (temporal gap) is the distance (time difference) between the last GPS point of current truck ID and the first GPS point of the next truck ID. Figure C.4 and C.5 show

the distribution of spatial gap and temporal gap, respectively, for 44 truck IDs in D1 data (September 2014 and March 2015 combined). Consequently, if the spatial gap was less than one mile and the temporal gap was less than 60 minutes, the two truck IDs were considered to belong to the same tanker truck. Three pairs of truck IDs were found that satisfied these criteria and therefore, the second truck ID in each pair was changed to the first truck ID.

Table C.5 shows the land use description distribution of remaining trips after trip ends on the road were joined or replaced, and three pairs of truck IDs in D1 data were joined. These 12,649 trips were the final output of trip cleaning tasks.

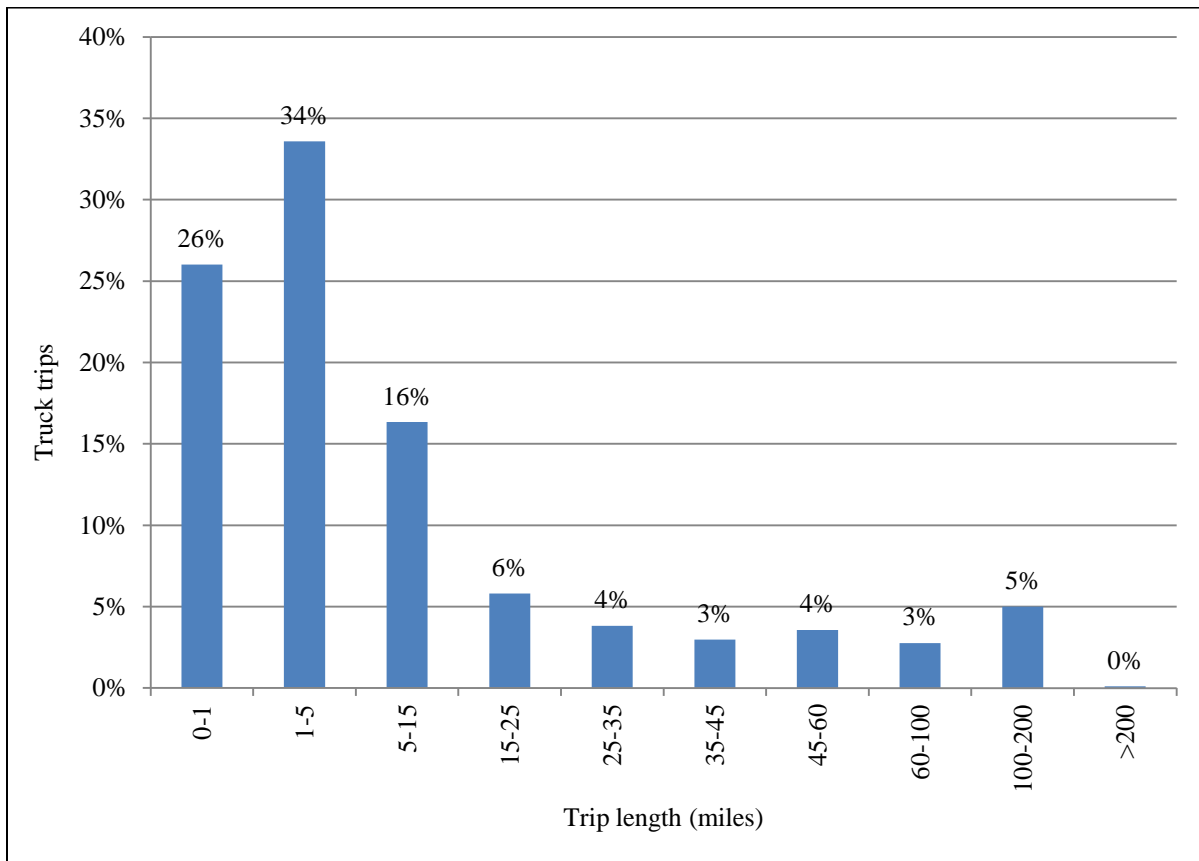


Figure C.1 Trip length distribution of 27 truck IDs removed from D1 and D2 data (September 2014 and March 2015 combined) (N = 1,879 trips)

Table C.1 Land use description distribution for the trip origins and destinations of the total 14,598 trips

Case	Location	Number of trip origins	Number of trip destinations
1	Gas station	5,667 (44.7 %) trip origins	5,325 (42 %) trip destinations
2	PEV	5,163 (40.7 %) trip origins	4,931 (38.9 %) trip destinations
3	Other fuel recipients	539 (4.3 %) trip origins	277 (2.2 %) trip destinations
4	Distribution center	681 (5.4 %) trip origins	655 (5.2 %) trip destinations
5	On the road	378 (3 %) trip origins	1,273 (10 %) trip destinations
6	Rest stop	255 (2 %) trip origins	222 (1.8 %) trip destinations
	Total	12,683 (100 %) trip origins	12,683 (100 %) trip destinations

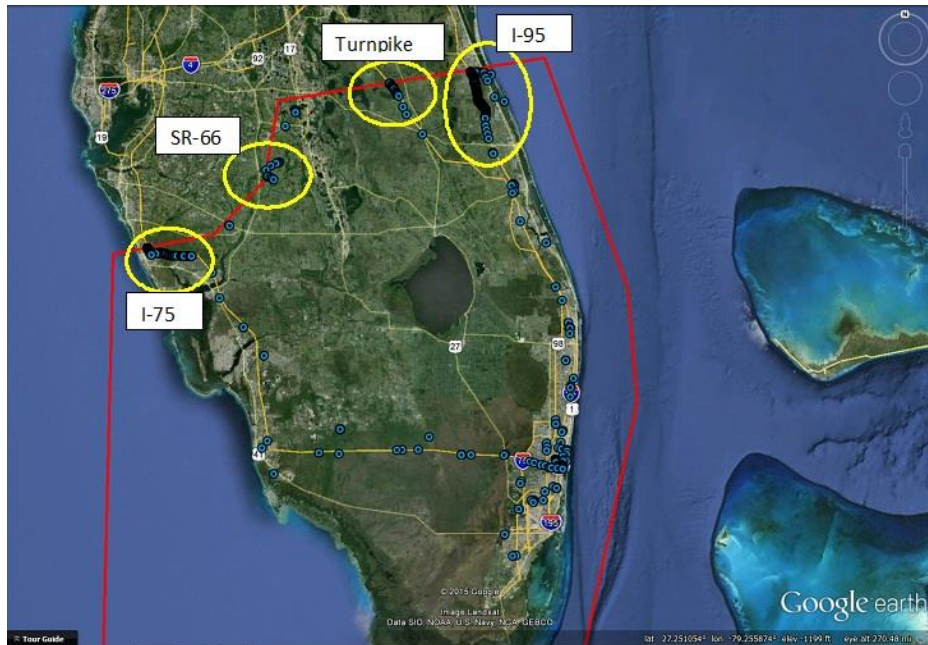


Figure C.2 All 1,273 on-the-road trip destinations shown in blue dots

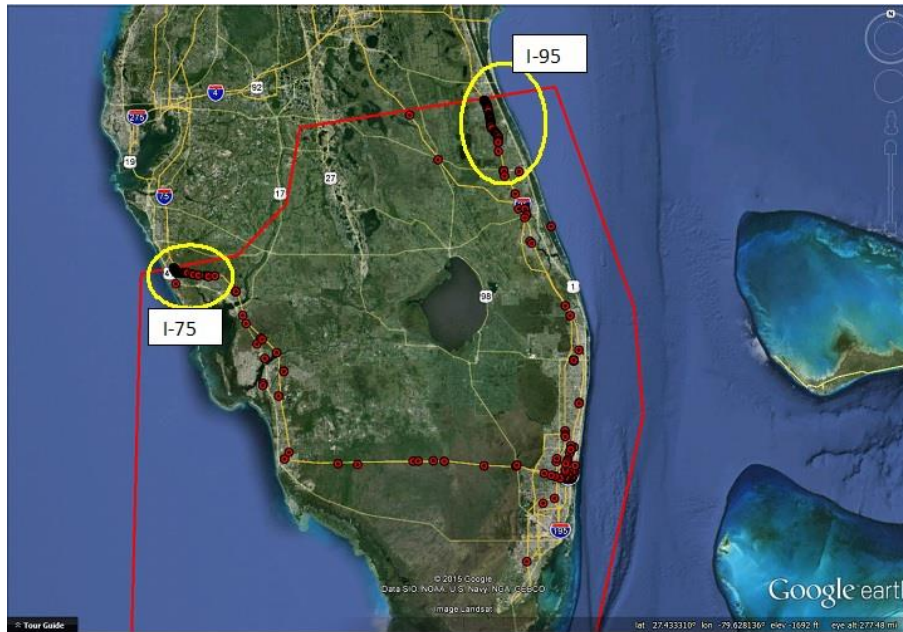


Figure C.3 All of the 378 on-the-road trip origins

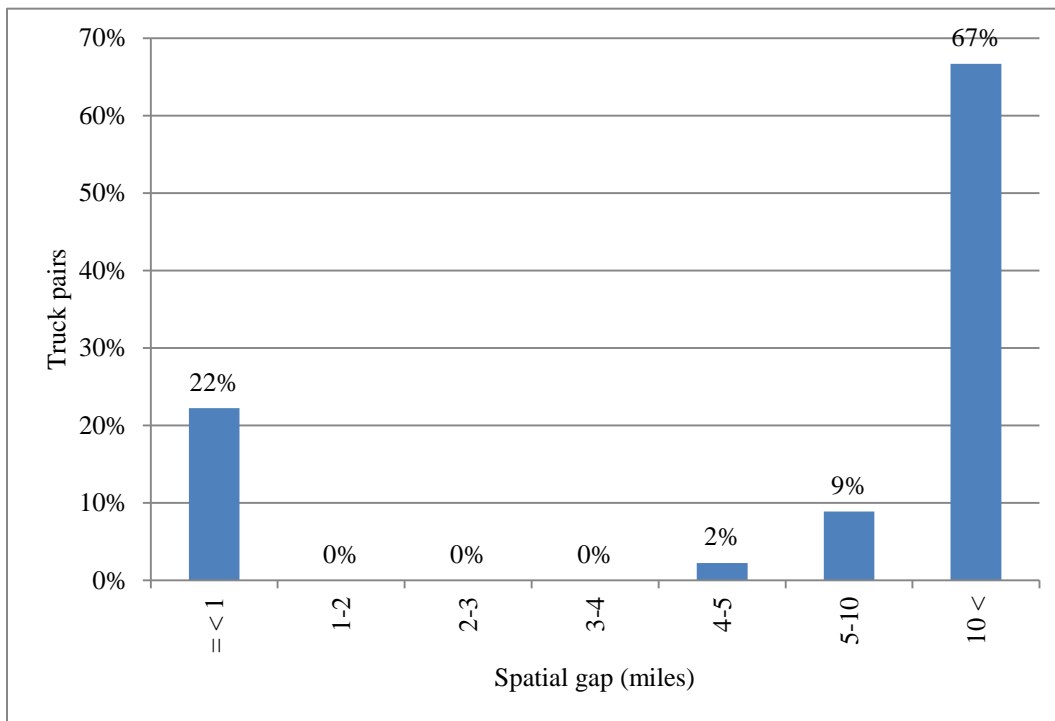


Figure C.4 Distribution of spatial gap between consecutive truck IDs (N = 44 truck ID pairs)

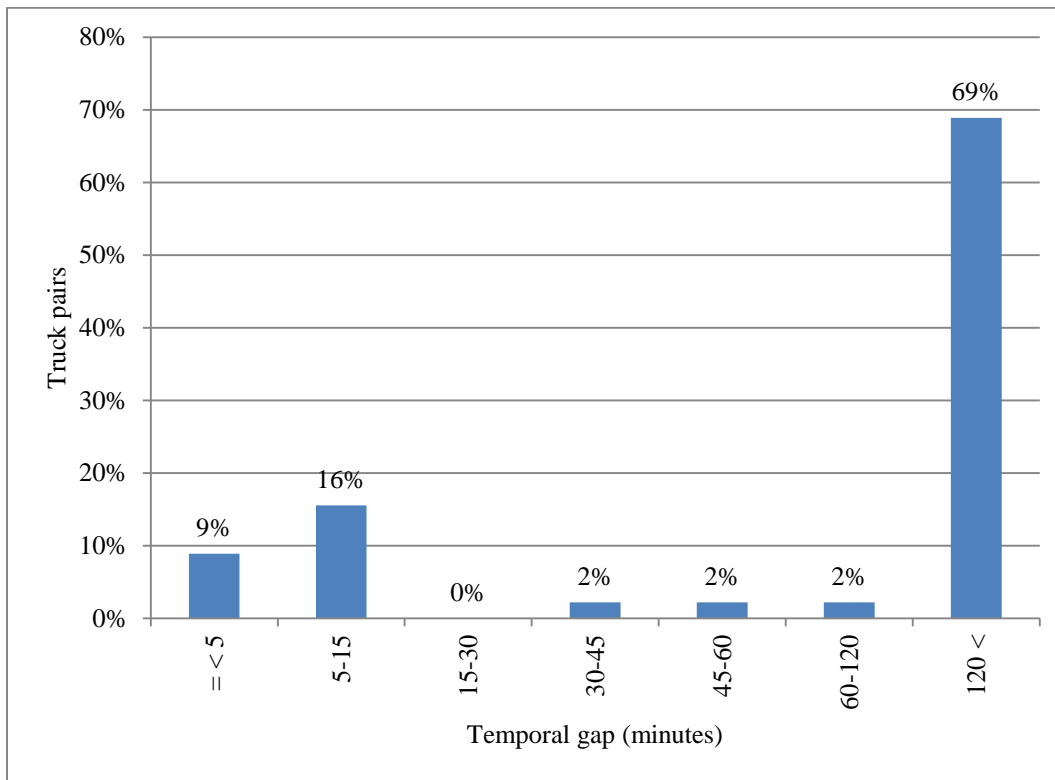


Figure C.5 Distribution of temporal gap between consecutive truck IDs (N = 44 truck ID pairs)

Table C.2 Land use description distribution for the trip origins and destinations of 12,649 trips

Case	Location	Number of trip origins	Number of trip destinations
1	Gas station	5,683 (44.9 %) trip origins	5,370 (42.5 %) trip destinations
2	PEV	5,163 (40.8 %) trip origins	4,968 (39.3 %) trip destinations
3	Other fuel recipients	539 (4.3 %) trip origins	278 (2.2 %) trip destinations
4	Distribution center	694 (5.5 %) trip origins	665 (5.3 %) trip destinations
5	On the road	310 (2.5 %) trip origins	1,129 (8.9 %) trip destinations
6	Rest stop	260 (2.1 %) trip origins	239 (1.9 %) trip destinations
	Total	12,649 (100 %) trip origins	12,649 (100 %) trip destinations