Development and Analysis of Truck Route Choice Data for the Tampa Bay Region using GPS Data

BDV25-730-3

Draft Final Report

Prepared for:

Florida Department of Transportation District 7

Mr. Ken Spitz, Project Manager

Mr. Brian Hunter, Project Manager

Prepared by:

UNIVERSITY OF SOUTH FLORIDA
COLLEGE OF ENGINEERING

University of South Florida

Mr. Divyakant Tahlyan

Ms. Trang D. Luong, EIT

Dr. Abdul R. Pinjari

Dr. Seckin Ozkul

March 2017

DISCLAIMER

METRIC CONVERSION

SYMBOL	WHEN YOU KNOW	MULTIPLY BY	TO FIND	SYMBOL
LENGTH				
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km

TECHNICAL REPORT DOCUMENTATION

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle *Development and Analysis of Truck Route Choice Data for the Tampa Bay Region using GPS Data*		5. Report Date *March 2017* 6. Performing Organization Code *USF*
7. Author(s) *Divyakant Tahlyan, Trang Luong, Abdul R. Pinjari, Seckin Ozkul*		8. Performing Organization Report No.
9. Performing Organization Name and Address *College of Engineering* *Department of Civil and Environmental Engineering* *University of South Florida* *4202 E. Fowler Avenue, ENB118, Tampa FL 33620*		10. Work Unit No. (TRAIS) 11. Contract or Grant No. *BDV25-730-3*
12. Sponsoring Agency Name and Address *Florida Department of Transportation District 7* *Intermodal Systems Development* *11201 North McKinley Drive, MS 7-500* *Tampa, FL 33612-6456*		13. Type of Report and Period Covered *Draft Final Report* *04/05/2016 to 03/22/2017* 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

This project investigated the use of truck-GPS data to develop route choice data for Tampa Bay region. The project used the American Transportation Research Institute's (ATRI's) truck-GPS data to first convert them into truck trips and then derive chosen routes for the trips. The procedure was applied to the first 15 days of truck-GPS data from each month of October 2015, December 2015, April 2016, and June 2016 in 6 Central Florida counties – Hillsborough, Pinellas, Polk, Pasco, Hernando, and Citrus – comprising more than 96 million GPS records. The raw GPS data were used to derive a database of more than 230,000 truck trips and the corresponding chosen routes. A part of the data was used to analyze the travel behavior of truck trips departing from and arriving at Port Tampa Bay. The data products from this project will be used by the research team to develop a truck route choice model for the Tampa Bay region, which in turn, can be used to improve the truck modeling components of the Tampa Bay Regional Planning Model.

17. Key Word *Truck-GPS data, map-matching, route choice, Port Tampa Bay*	18. Distribution Statement		
19. Security Classify. (of this report) *Unclassified*	20. Security Classify. (of this page) *Unclassified*	21. No. of Pages *51*	22. Price

ACKNOWLEDGMENTS

The authors express their sincere appreciation to the Florida Department of Transportation (FDOT) District 7 managers on this project, Mr. Kenneth Spitz and Mr. Brian Hunter, for their support and feedback during this research. We also thank Ms. Elaine Martino of FDOT District 7 for not only providing the geo-spatial data needed for this project but also answering our inquiries about the data. The authors thank Ms. Menna Yassin for her involvement during the conceptualization stage of this project. The American Transportation Research Institute (ATRI) provided the truck-GPS data used in this project. Our thanks are due to Mr. Dan Murray and Mr. Jeffrey Short from ATRI, who played a key role in the provision of the truck-GPS data needed for this project. We also extend our thanks to Mr. Richard Tillery and Dr. Zahra Pourabdollahi from RS&H for their inputs in the project. Finally, we thank Mr. Eren Yuksel, a graduate student at the University of South Florida, for helping us with validation of the derived routes.

EXECUTIVE SUMMARY

Project Objectives

This project aims at providing important data useful for understanding and modeling truck route choice behavior in the Tampa Bay region, which in turn, can be used to improve the truck modeling components of the Tampa Bay Regional Planning Model (TBRPM).

The goal of this task work order was to use truck-GPS data from the American Transportation Research Institute (ATRI) to derive data on the routes that trucks use to complete their travel between different origins and destinations in the Tampa Bay region. To this end, the following specific tasks were performed in the project:

1. Secure truck-GPS data from ATRI.
2. Convert raw GPS data into database of truck trips.
3. Derive chosen routes for truck trips obtained from tasks above.
4. Analyze routes used by trucks between selected origin and destination locations in the region.

Each of these tasks is briefly discussed in the following sections.

Task 1. Secure Truck-GPS Data from ATRI

For this project, the research team used ATRI's truck-GPS data for the first 15 days of each month of October 2015, December 2015, April 2016, and June 2016 in the six Central Florida counties of Hillsborough, Pinellas, Polk, Pasco, Hernando, and Citrus and 15 miles beyond their geographic boundaries. This resulted in 96,438,457 raw GPS records belonging to 110,475 truck IDs.

To protect confidentiality of the data, ATRI required the University of South Florida (USF) to sign a non-disclosure agreement (NDA). According to the NDA, the raw GPS data shared by ATRI with USF was to be used only for the purpose of analysis by USF researchers and was not to be shared with anyone outside the research team except the Florida Department of Transportation (FDOT) consultant team of Reynolds, Smith and Hills (RS&H). In addition, the NDA allowed for the aggregate results and data products (such as a database of trips and a database of travel routes) from the research to be submitted to FDOT as long as the locations of the trip origins, destinations, and intermediate locations were not revealed in a high spatial resolution.

Task 2. Convert Raw GPS Data into Database of Truck Trips

The procedure used to convert raw truck-GPS data into trips utilized the same algorithm used for a previous FDOT study by Pinjari et al. (2016), which was originally derived in another FDOT study by Pinjari et al. (2014). The procedure was applied to all four months (October and December 2015 and April and June 2016) of refined truck-GPS data, comprising 90,184,869 GPS records. This resulted in 1,076,371 truck trips.

This initial set of trips extracted from ATRI data was further refined by removing trips that started and ended in the same TAZ, trips that were less than two miles, and trips with a ratio between direct origin-destination (OD) distance and trip length less than 0.7. After implementing the above criteria, a refined and validated set of 676,902 trips was retained for subsequent use.

Task 3. Derive Chosen Routes from Refined, Validated Truck Trips

The procedure for deriving routes using GPS data consisted of three broad steps: 1) map-matching data preparation, 2) map-matching, and 3) route generation. A modified version of the procedure developed by Kamali et al. (2016) was used to derive the routes, and routes were generated for 243,438 trips. The routes derived for each trip included all the network links traversed by the truck between the origin and destination of the trip. Validation of 300 randomly-selected trips indicated a high level of accuracy of the derived routes. An algorithm also was developed to remove all routes with loops of more than one mile. The final product of the project was provided in the form of a GIS shape file of the generated routes for 233,329 trips along with trip-level information such as trip length, trip time, and origin/destination TAZs. Another product of the project was in the form of a comma-separated file (.csv) containing unique OBJECT IDs of each link traversed along the generated routes for each trip. Appendix A describes the GIS shapefile of routes and Appendix B describes the .csv file.

Task 4. Analyze Routes Used by Trucks between Selected Origin and Destination Locations in the Region

After discussions with FDOT District 7, analysis of routes used by trucks departing from and arriving at Port Tampa Bay was conducted. Of 233,329 final trips, there were 3,755 trips departing from the Port and 3,774 trips arriving at the Port. A descriptive analysis of these trips was conducted to understand the travel behavior of trucks traveling into and out of Port Tampa Bay.

Uses of Data Developed in this Project

The data developed in this project offer significant opportunities to understand truck route choice behavior in the Tampa bay region. Future work on analyzing the determinants of truck route choice patterns can help forecast aggregate-level network performance for medium- to long-term decisions such as designation of truck routes, addition of new truck corridors, and by-pass routes. Measuring and monitoring the travel paths (or routes) trucks take and understanding why they do so can help design short-term truck routing policies aimed at congestion mitigation, improving reliability, and maintaining a state of good repair. A parallel project being led by this project's Principal Investigator is using the route choice data derived from this project to develop truck route choice models for the Tampa Bay region. Therefore, the data products derived in this project are already bearing fruit. Such truck route choice models potentially can be used to improve the truck modeling components of the TBRPM.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Freight is gaining increasing importance in transportation planning and decision-making in the Tampa Bay region, with freight truck volumes projected to increase by as much as 65% by 2040. As the freight movement in the region increases, the region's economic productivity will increasingly rely on a transportation system that can accommodate goods movement efficiently, reliably, and safely. Recognizing this, the Tampa Bay region is at the forefront of recognizing and addressing freight movement issues for improving freight mobility in the region.

An essential step toward enhancing highway freight mobility is to gain a thorough understanding of freight truck behavior in the transportation networks traveled by trucks, including the demand for travel between origins and destinations, interaction with other modes of travel, and routes of travel. An understudied dimension of all these aspects is truck travel route choice. Measuring and monitoring the travel paths (or routes) trucks take and understanding why they do so can help design short-term truck routing policies aimed at congestion mitigation, improving reliability, and maintaining a state of good repair. In addition, it is essential to understand and forecast truck travel route choice and aggregate level network performance for medium- to long-term decisions such as designation of truck routes, addition of new truck corridors, and by-pass routes.

A primary challenge for all such investigations, however, is the unavailability of observed data on truck routes. Traditional travel surveys do not allow for the observation and measurement of truck travel routes. In the absence of such data, transportation planners and analysts must make assumptions on truck route choice behavior that may not hold true in the context of freight movement. Another challenge that also stems from the unavailability of data is the dearth of truck route choice models that can be used to analyze and foresee truck travel routes under alternative scenarios of traffic performance (congestion, reliability, etc.) and routing policies.

Truck-GPS data from the American Transportation Research Institute (ATRI) provide an opportunity to measure and understand truck route choice at an unprecedented scale. To do so, however, the raw GPS data need to be converted to a more useful form that can be used to identify the trucks' trip origins and destinations as well as the specific routes used by the trucks between their travel origins and destinations. Therefore, this project aims at using ATRI's truck-GPS data in the Tampa Bay region to derive data on the routes trucks use to complete their travel between different origins and destinations in the region. Doing so will provide important data that are useful for understanding and modeling truck route choice behavior in the Tampa Bay region, which, in turn, can be used to improve the truck modeling components of the Tampa Bay Regional Planning Model (TBRPM).

The specific objectives of this task work order (TWO) are two-fold, as outlined below:

1. Develop data for analyzing truck route choice patterns between different origins and destinations in the Tampa Bay region using truck-GPS data. To do so, the project will rely on the truck-GPS data from ATRI.

2. Analyze the routes commonly used by trucks between selected origin and destination locations of interest to FDOT.

To achieve the above objectives, the following tasks were conducted:

1. Secure the truck-GPS data from ATRI.
2. Convert raw GPS data to database of truck trips.
3. Derive chosen routes for truck trips derived in above task.
4. Analyze routes used by trucks between selected origin and destination locations in region.

This report describes these tasks in detail and presents the results from each. Specifically, the report is structured as follows. Chapter 2 provides a description of the data used for this project, including ATRI's truck-GPS data, information on traffic analysis zones, and the Navteq road network. Chapter 3 outlines the procedure used for converting ATRI's raw GPS data into truck trips and presents descriptive characteristics of the trips derived from the raw GPS data. Chapter 4 outlines the procedure used to derive chosen routes for the truck trips derived in the project and presents descriptive characteristics of the derived routes as well as results from the validation of a sample of the routes derived in the project. Chapter 5 presents an analysis of the routes commonly used by trucks between selected origin and destination locations in the Tampa Bay region. Finally, Chapter 6 concludes the report and identifies avenues for future research.

It is worth noting that a parallel project sponsored by the United States Department of Transportation (USDOT) being led by this project's Principal Investigator is using the data derived from this project to develop truck route choice models for the Tampa Bay region. Therefore, the data products derived in this project are already bearing fruit. Such truck route choice models potentially can be used to improve the truck modeling components of the TBRPM.

CHAPTER 2: OVERVIEW OF DATA

2.1 Introduction

This chapter provides a brief overview of the data used for this project. The data consists of ATRI's truck-GPS data with the following geographic and temporal coverage: GPS data for 6 Central Florida counties and 15 miles beyond their geographic boundaries—Hillsborough, Pinellas, Polk, Pasco, Hernando, and Citrus—for a total of 8 weeks. Specifically, this study used data for two continuous weeks each from the four months of October 2015, December 2015, April 2016, and June 2016.

In addition to the truck-GPS data, the following other data were used in the work:

- Shape file of traffic analysis zones (TAZs) in the six-county region.
- Shape file of a detailed highway network in the study area.

2.2 ATRI's Truck-GPS Data

ATRI obtains truck position data from partnerships with trucking companies and data vendors throughout the US and North America from a large sample of trucks that use onboard wireless communication systems such as GPS. For this project, the research team used ATRI's truck-GPS data for the first 15 days of each month of October 2015, December 2015, April 2016, and June 2016. ATRI's truck-GPS data came in two formats—format S (static) and format R (rotating.) The term "static" implies that the unique truck ID assigned to a particular truck remains static or does not change. The term "rotating" implies that the unique truck ID assigned to a truck changes every month. Each record in ATRI's truck-GPS data included the following:

- **Geographical Information** (x [longitude], y [latitude]) – the geographic location of the corresponding truck at the point at which the truck location was recorded.
- **Unique Truck IDs** (truckid) – an anonymized random digit unique identifier for the truck. In the dataset S, the truck ID remains the same throughout the dataset for a given truck. In dataset R, the truck ID rotates every month.
- **Time Stamp** (readdate) – the data and time at which the corresponding information was recorded in the format YYYY-MM-DD HH:MM:SS
- **Spot Speed** (speed) – information on the instantaneous speed (in mph) of the corresponding truck at the point at which the information was recorded.
- **Geographic Direction** (heading) – information on the geographic direction in which the corresponding truck is heading at the point at which the information was recorded.

Information on the specific format of the data for format S and format R is as follows:

- Data in the S format.
  - Sample record of dataset S:

x	y	truckid	readdate	speed	heading
-82.31849	27.98362	117717	2015-03-01 00:00:00	0	ST

- In the dataset of format S, most truck IDs are 6 numeric digits (e.g., 101695) but there are truck IDs that are 3, 4, or 5 numeric digits and 9- or 11-character alpha-numeric (e.g., a2420559639) in nature.
- Heading values are divided into 8 geographic categories—E, N, NE, NW, S, SE, SW, W—plus ST ("stop") and NA ("not available").
- Longitude and latitude values in the dataset are precise up to 5 decimal places.

• Data in the R format:

- Sample record of dataset R:

x	y	truckid	readdate	speed	heading
-82.35787	27.97245	014bd2a1f0590a26009c24b3ff14	2015-03-01 00:00:00	0	0.0

- In the dataset of format R, all truck IDs are 30-character alpha-numeric (e.g., 0003b7d7d5414b8f88d22da27f96b4) in nature.
- Heading values vary from 0 to 360 degrees at intervals of 22.5 degrees. The heading is measured from north in the clockwise direction.
- Similar to dataset S, the longitude and latitude values in dataset R are precise up to 5 decimal places.

To protect confidentiality of the data, ATRI required the University of South Florida (USF) to sign a non-disclosure agreement (NDA). According to the NDA, the raw GPS data shared by ATRI with USF were to be used only for the purpose of analysis by USF researchers. The raw data must not be shared with anyone outside the research team except FDOT's consultant of Reynolds, Smith and Hills (RS&H) in Tampa. In addition, the NDA allows for the aggregate results and data products (such as a database of trips and a database of travel routes) from the research to be submitted to FDOT as long as the locations of the trip origins, destinations, and intermediate locations are not revealed in a high spatial resolution.

2.3 Refinement of ATRI's Truck-GPS Data

Before further processing of the GPS data to derive truck trips and the corresponding routes, the data were examined and refined to remove GPS records that exhibit anomalies. Specifically, the following types of anomalous records were removed from the raw GPS data:

- GPS records where only one GPS record was present corresponding to a unique truck ID.
- Multiple GPS records with same truck ID and the same time stamp (only one such GPS record was retained while removing other repetitions of the same data point).
- GPS records with spot speed of more than 100 mph and data with average speed between consecutive GPS records for a truck ID greater than 100 mph. Such high-speed GPS records were removed because of the high likelihood of such data occurrence due to GPS error.

2.4 Descriptive Information on ATRI's Truck-GPS Data

Table 2.1 presents basic descriptive characteristics of all four months of data combined, before and after the above-mentioned refinement steps. As can be observed from the table, although the number of GPS records in dataset R is about 6 times the number of records in dataset S, the number of trucks in dataset R is only about 1.5 times the number of trucks in dataset S. This is largely because the dataset R has a higher frequency GPS data than that in dataset S.

Table 2.1: Number of Records and Truck IDs Before and After Refinement

		Dataset S	Dataset R	Total
Before cleaning data	# Truck IDs	41,306	69,169	110,475
	# GPS Points	14,104,992	82,333,465	96,438,457
After cleaning data	# Truck IDs	41,171	69,064	110,235
	# GPS Points	13,669,502	76,515,367	90,184,869

Figure 2.1 provides the distribution of the duration of data available for different truck IDs in datasets S and R. It can be observed that dataset S can be used to trace trucks for longer durations than dataset R. Figure 2.2 provides the distribution of instantaneous speed across GPS records. It is interesting to note that the percentage of GPS records with zero instantaneous speed is greater in dataset R than in dataset S.

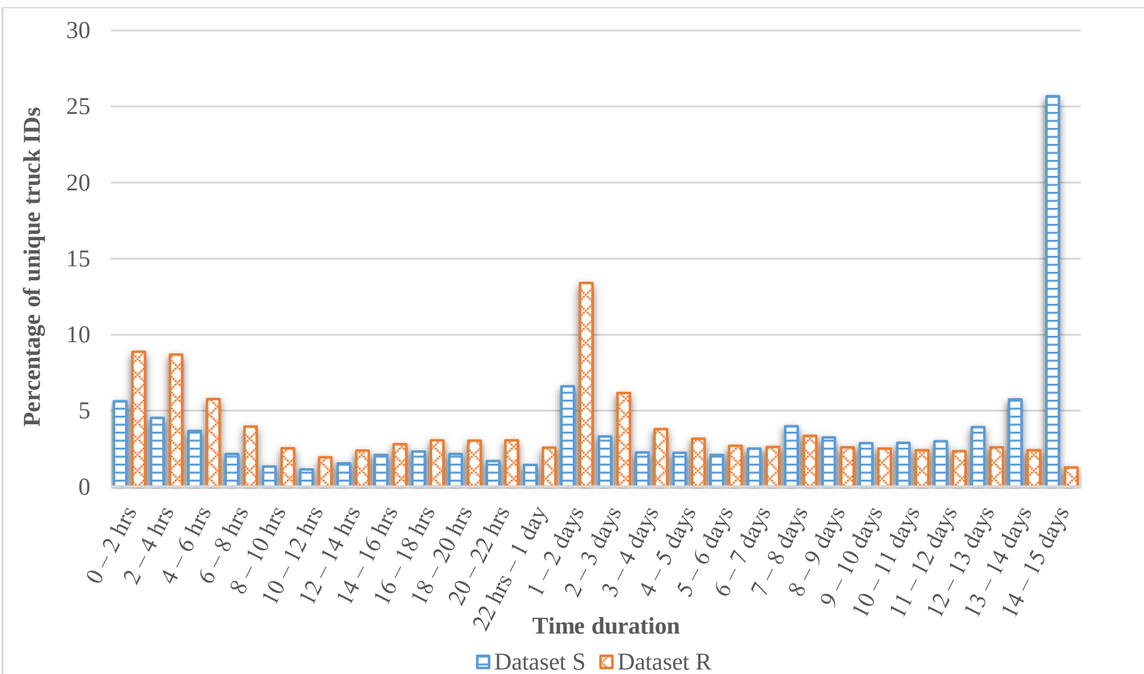

Figure 2.1: Distribution of duration of data available for each truck ID

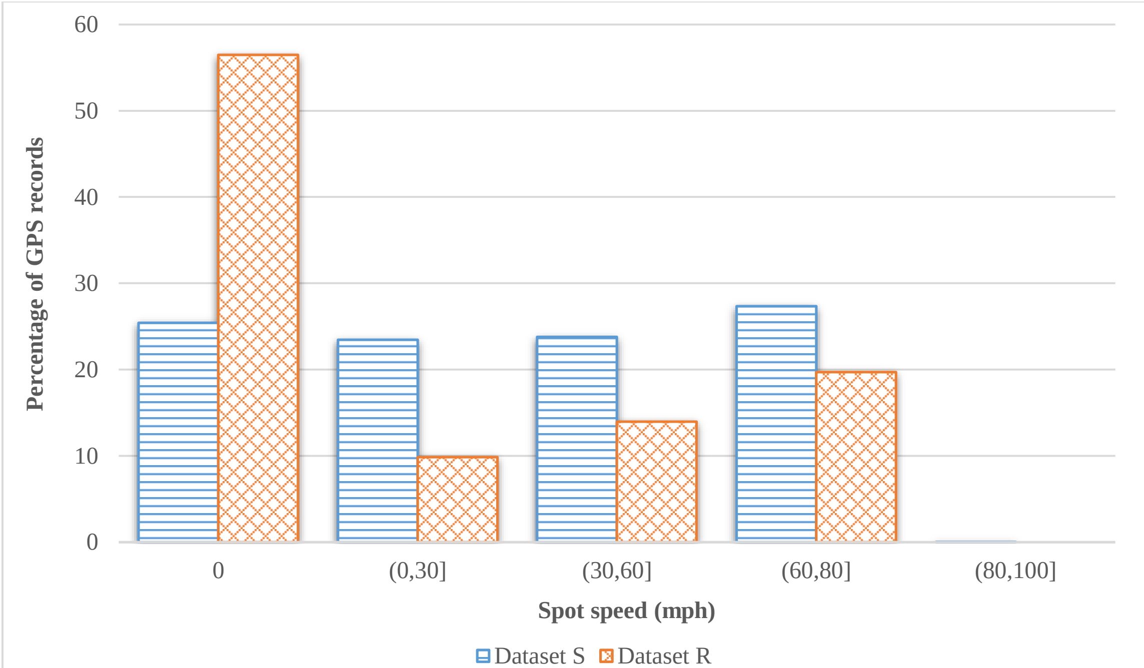

Figure 2.2: Distribution of instantaneous speeds (in mph) of GPS records

An important attribute of GPS data is the time gap between consecutive GPS points (referred to as "ping rate" hereafter). The higher the ping rate (or lower the time gap), the more frequent the GPS data. Another key attribute of GPS data is the geodetic distance between the consecutive GPS points (referred to as "spatial gap" hereafter). The geodetic distance (or spatial gap) between two consecutive points was calculated using the Haversine formula as follows:

$$hav(\frac{d}{r}) = hav(\varphi_2 - \varphi_1) + \cos(\varphi_2)\cos(\varphi_2)hav(\lambda_2 - \lambda_1)$$

where:

- $hav(\theta) = \sin^2(\frac{\theta}{2}) = \frac{1-\cos(\theta)}{2}$
- d is the distance between the two GPS points
- r is the radius of the earth
- φ_1, φ_2 are latitudes of GPS points 1 and point 2, respectively, in radians
- λ_1, λ_2 are longitude of point 1 and point 2, respectively, in radians.

Tables 2.2 and 2.3 provide cross-tabulations between the spatial gap (in miles) and the time gap between consecutive GPS records for datasets S and R, respectively. It can be observed from these tables that about 67% of the data in dataset S and 96% of the data in dataset R have a ping rate (or time gap) less than 5 minutes and a spatial gap less than 5 miles. Such high frequency data are useful for accurately deriving information on truck travel routes. Although the GPS data in dataset R are of higher frequency than those in dataset S, both datasets were used to derive truck travel routes because both have sufficiently high frequent GPS data.

Table 2.2: Cross-Tabulation between Spatial Gap (in miles) and Ping Rate for Dataset S

Spatial Gap / Ping Rate (min)	(0,1]	(1,5]	(5,15]	(15,20]	(20,25]	(25,30]	> 30	Total
< 1 min	34.18	3.31	0	0	0	0	0	37.49
(1-2] min	8.98	1.75	0	0	0	0	0	10.73
(2-5] min	8.59	10.53	4.29	0	0	0	0	23.41
(5-14] min	6.29	1.86	1.86	0.01	0	0	0	10.02
(14-15] min	1.24	0.32	3.13	1.42	0	0	0	6.11
(15-20] min	1.77	0.03	0.05	0.04	0.01	0	0	1.9
(20-25] min	0.98	0.01	0.01	0.01	0.03	0.07	0	1.11
(25-30] min	2.44	0.01	0.02	0.01	0.02	0.05	0.06	2.61
(30-45] min	1.63	0.02	0.01	0.01	0.01	0.01	0.03	1.72
(45-120] min	1.97	0.08	0.05	0.01	0.01	0.01	0.04	2.17
(2 hrs – 1] day	1.74	0.27	0.19	0.03	0.02	0.01	0.1	2.36
> 1 day	0.13	0.09	0.06	0.01	0.01	0	0.07	0.37
Total	69.94	18.28	9.67	1.55	0.11	0.15	0.3	100%

Table 2.3: Cross-Tabulation between Spatial Gap (in Miles) and Ping Rate for Dataset R

Spatial Gap / Ping Rate (min)	(0,1]	(1,5]	(5,15]	(15,20]	(20,25]	(25,30]	> 30	Total
< 1 min	74.42	13.88	0	0	0	0	0	88.3
(1-2] min	4.13	2.05	0	0	0	0	0	6.18
(2-5] min	1.43	0.22	0	0	0	0	0	1.65
(5-14] min	1.37	0.12	0.08	0	0	0	0	1.57
(14-15] min	0.98	0	0.01	0	0	0	0	0.99
(15-20] min	0.19	0	0.01	0.01	0	0	0	0.21
(20-25] min	0.04	0	0	0.01	0.04	0	0	0.09
(25-30] min	0.05	0	0	0	0	0	0	0.05
(30-45] min	0.06	0	0	0	0	0	0	0.06
(45-120] min	0.6	0.01	0.01	0	0	0	0	0.62
(2 hrs – 1] day	0.1	0.01	0.01	0	0	0	0.01	0.13
> 1 day	0.06	0	0	0	0.01	0	0.01	0.08
Total	83.43	16.29	0.12	0.02	0.05	0	0.02	100%

Figure 2.3 presents the distribution of average trip speed between consecutive GPS records. It must be noted that for dataset R, the percentage of GPS records with average speed between consecutive GPS records for a truck ID less than 1 mph is more than 50% and is also almost double that in dataset S.

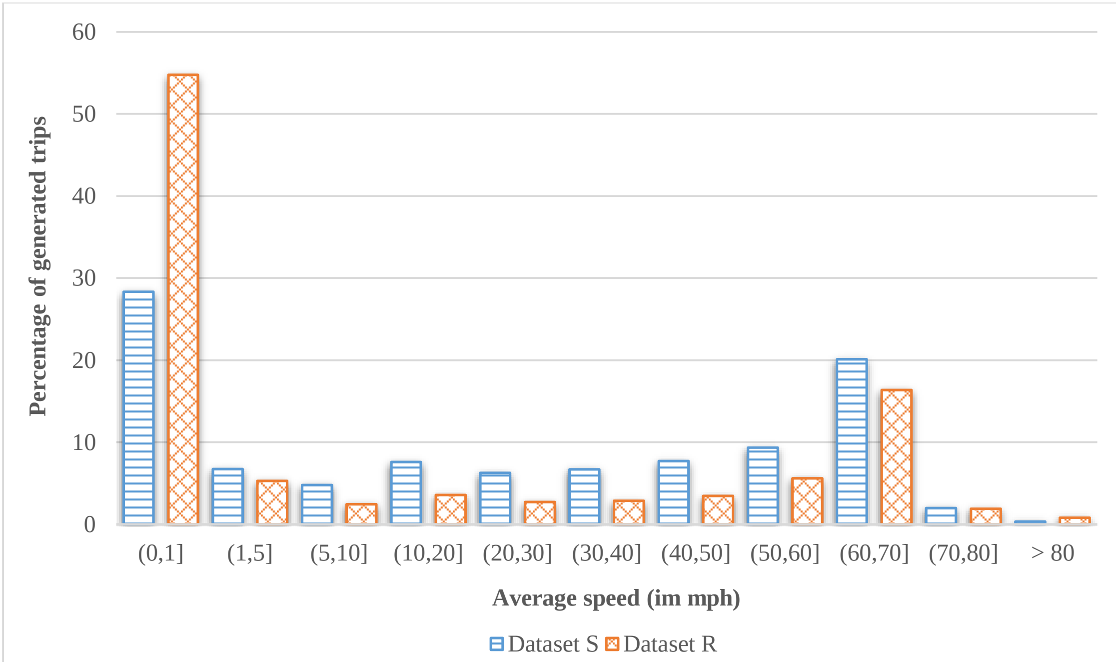

Figure 2.3: Distribution of average speed between consecutive GPS records of a truck ID

2.5 Descriptive Information on TAZs in Six-County Region

Figure 2.4 shows the six-county study region of focus in this study, which includes Citrus, Hernando, Hillsborough, Pasco, Pinellas, and Polk counties. In the Florida Statewide Model (FLSWM), this six-county region is divided into 891 traffic analysis zones (TAZs)—Citrus–28, Hernando–64, Hillsborough–304, Pasco–78, Pinellas–278, and Polk–139. Once the raw GPS data were converted into truck trips, the origin and destination locations of the trips were identified at a spatial resolution of these FLSWM TAZs. Doing so helped to identify the origin and destination TAZs at an aggregate level without revealing the exact location of the trip ends. This helped to comply with the non-disclosure agreement USF signed with ATRI. Since the GPS data were available for 15 miles beyond the six-county region, a 15-mile buffer around the six-county region was also considered part of the study region, so the routes taken by the trips with GPS points falling in the 15-mile buffer might be beyond the 15-mile geographic boundary as well.

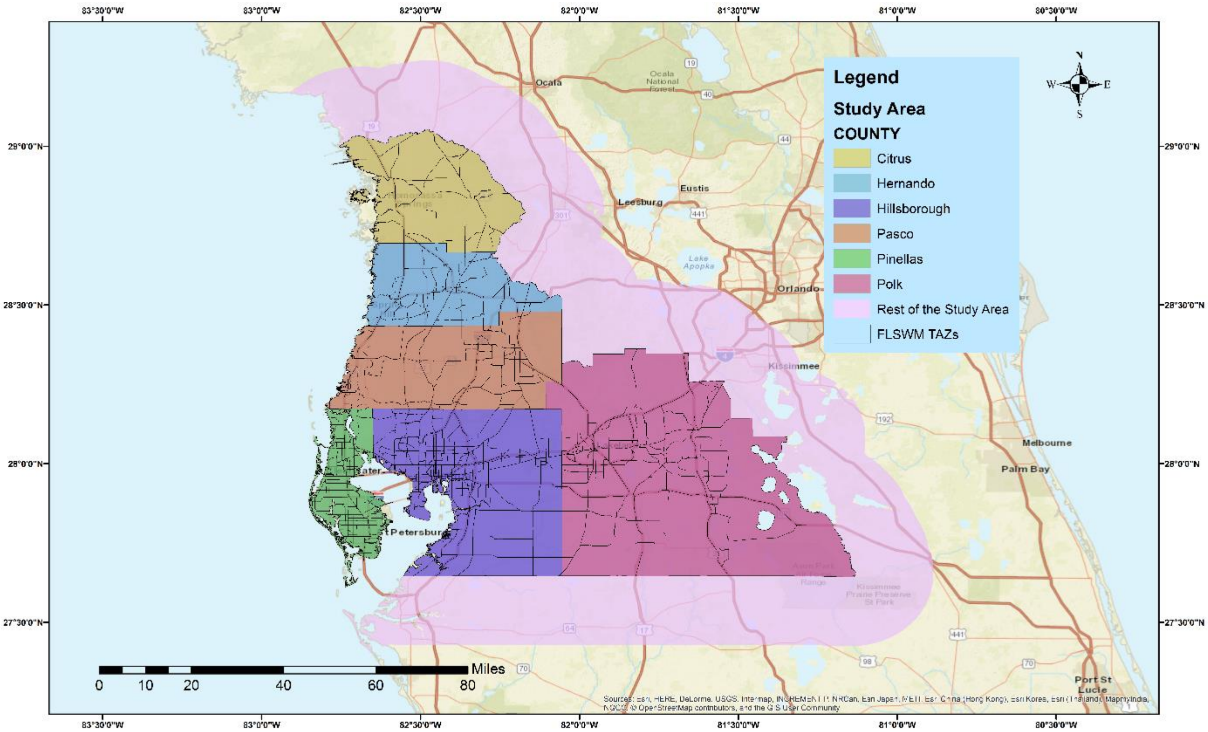

Figure 2.4: Six-county study region and 15-mile buffer

2.6 Descriptive Information on Detailed Road Network in Six-County Region

Another type of data used in this project in addition to ATRI's truck-GPS data and FLSWM's TAZs is the Navteq road network data. The Navteq data were used to derive routes after map-matching the GPS records of the corresponding trips. For the six-county region, Figure 2.5 shows 490,690 network links covering 43,612 miles of road network.

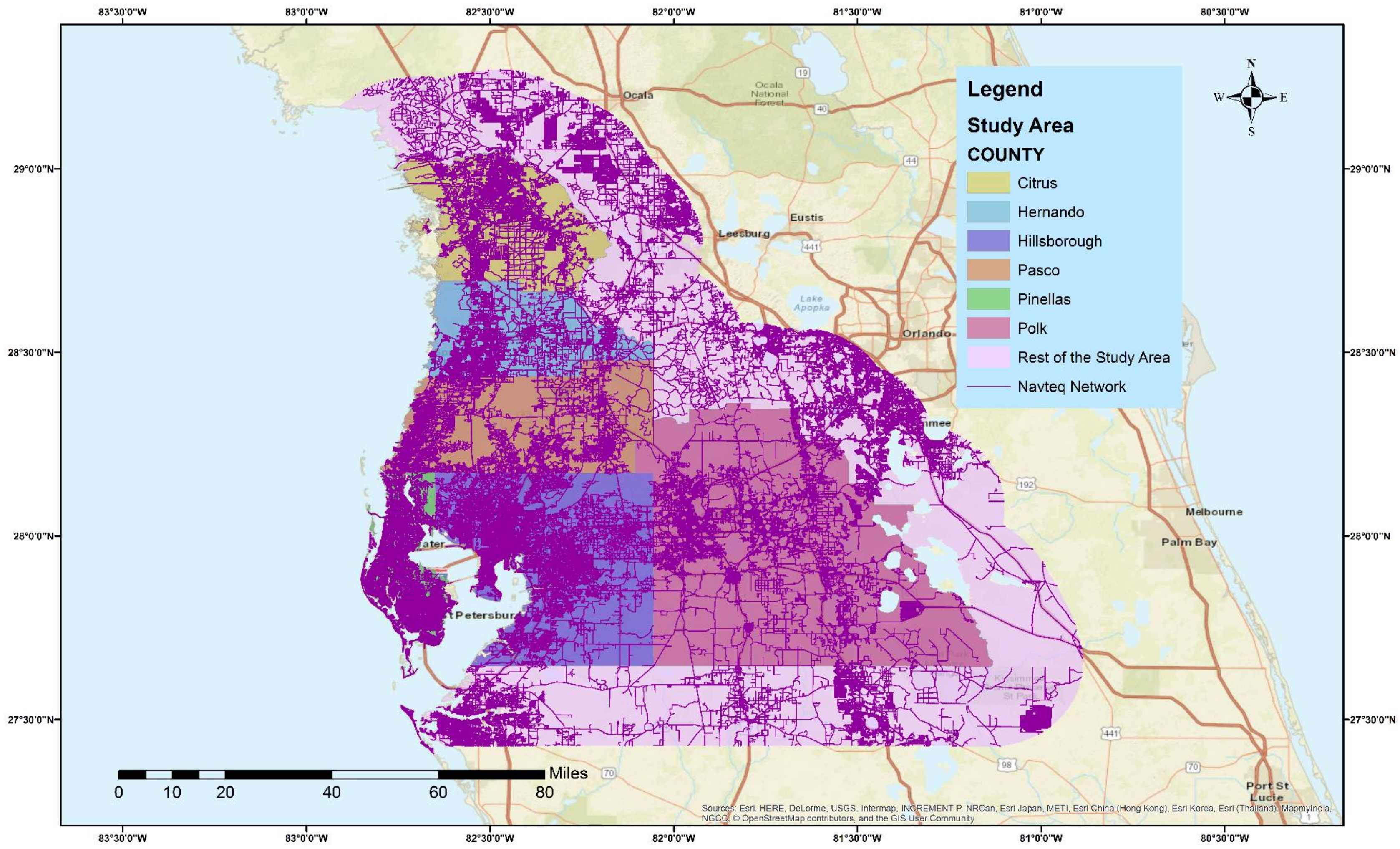

Figure 2.5: Navteq road network used in project

CHAPTER 3: DERIVING TRUCK TRIPS

3.1 Introduction

This chapter describes the procedure used for converting ATRI's raw truck-GPS data into truck trips and also includes an overview of the total trips derived, criteria for cleaning out some trips, and an overview of the final database of the refined trips. In addition, the chapter presents descriptive characteristics of the final refined trips and a summary of the validation of the trips database. It is worth noting that the procedure used to convert raw truck-GPS data into trips was the same procedure used for a previous FDOT study by Pinjari et al. (2016), which was originally derived in another FDOT study by Pinjari et al. (2014). Therefore, much of the procedure described in Section 3.2 is drawn from the FDOT report by Pinjari et al (2016). The procedure is described here only to make this report stand alone, without having to refer readers to the previous FDOT report.

3.2 Algorithm Description

The overall procedure to convert ATRI's truck-GPS data into a database of truck trips can be described in the following three broad steps, each of which is detailed in this section.

1. Clean, read, and sort raw GPS data in chronological order for each truck ID. At the end of this step, all GPS data belonging to each truck ID are grouped together in chronological order.
2. Identify an initial set of truck trip stops (i.e., trip ends) based on spatial movement, time gap, and speed between consecutive GPS points. In this step, a truck is considered to have stopped at a destination if it stops (i.e., if the average travel speed between two consecutive GPS points is less than 5 mph) for at least 5 minutes. A truck stop of less than 5-minute duration is considered to be a traffic stop (i.e., not a valid destination) and, therefore, is considered part of the travel between origin and destination.
3. Conduct quality checks and refine or eliminate trips that do not satisfy quality criteria.

3.2.1 Data Cleaning and Sorting

The raw GPS data were first screened for basic quality checks such as the presence of spatial and temporal information and the presence of at least 6 hours of data for each truck ID. All truck IDs in which the duration of data available was less than 6 hours were removed from the dataset, as the authors believed that the amount of data available for these trucks is too small to derive useful information. The data were then sorted in chronological order for each truck ID, beginning with the GPS record with the earliest date and time stamp.

3.2.2 Identification of Truck Stops (i.e., Truck Trip-Ends) to Generate Truck Trips

This step comprises a major part of the procedure to convert raw GPS data into truck trips. The details of the algorithmic procedure in this step are presented in Figure 3.1, which was modified from a previous FDOT project by Pinjari et al. (2016) from the algorithm developed by Pinjari et al. (2014). Following is a list of the terms used in the algorithm along with their definitions:

- *Travel distance* (td): Spatial (geodetic) distance between two consecutive GPS records.
- *Travel time* (trt): Time gap between two consecutive GPS records.
- *Average travel speed* (trs): Average travel speed between consecutive GPS records (td/trt).
- *Trip length* (tl): Total distance traveled by truck from origin of trip to current GPS point. This becomes equal to trip distance, when destination is reached.
- *Trip time* (tpt): Total time taken to travel from origin of trip to current GPS point.
- *Trip speed* (tps): Average speed of trip between origin and current GPS point.
- *Origin dwell-time* (odwt): Total time duration of stop at origin; i.e., when truck is not moving (wait time for truck before starting its trip)
- *Destination dwell-time* (ddwt): Total duration truck stops at destination of a trip.
- *Stop dwell-time* (sdwt): Duration of intermediate stop (e.g., traffic stop).
- *Total stop dwell-time* (tsdwt): Total duration at all intermediate stops during trip.

The first three terms—td, trt, and trs—are measures of movement between consecutive GPS data points. The next three terms—tl, tpt, and tps—are measures of total travel between the trip origin and the current GPS data point. When the truck destination is reached, these measures are for the entire trip beginning from its origin to the destination. The last four terms are dwell times (i.e., stop durations) at different stages during the trip—odwt is the dwell-time at the origin of a trip, ddwt is the dwell-time at the destination of the trip, sdwt is the dwell-time at an intermediate stop (e.g., traffic stop) that is not the destination of the trip, and tsdwt is the sum of dwell-time at all intermediate stops during the trip.

For each truck ID, the algorithm begins with reading its first GPS record and initializing all the terms. Then, the algorithm reads the next record and computes average travel speed between the two records to verify if the truck is moving or if it is at rest. The subsequent parts of the algorithm are described below.

3.2.2.1 Determination of truck stops and moving instances

An important component of the algorithm involved determining whether a truck was at a stop (i.e., rest) or in motion. As can be observed from the flowchart in Figure 3.1, the primary condition used to determine whether a truck was at a stop (which could be an origin, a destination, or an intermediate stop) or it was moving was based on the average travel speed between consecutive GPS data points. A cut-off speed of 5 mph was used, which means if the average travel speed between consecutive GPS records was less than 5 mph (i.e., if trs < 5 mph), the truck was assumed to have stopped.

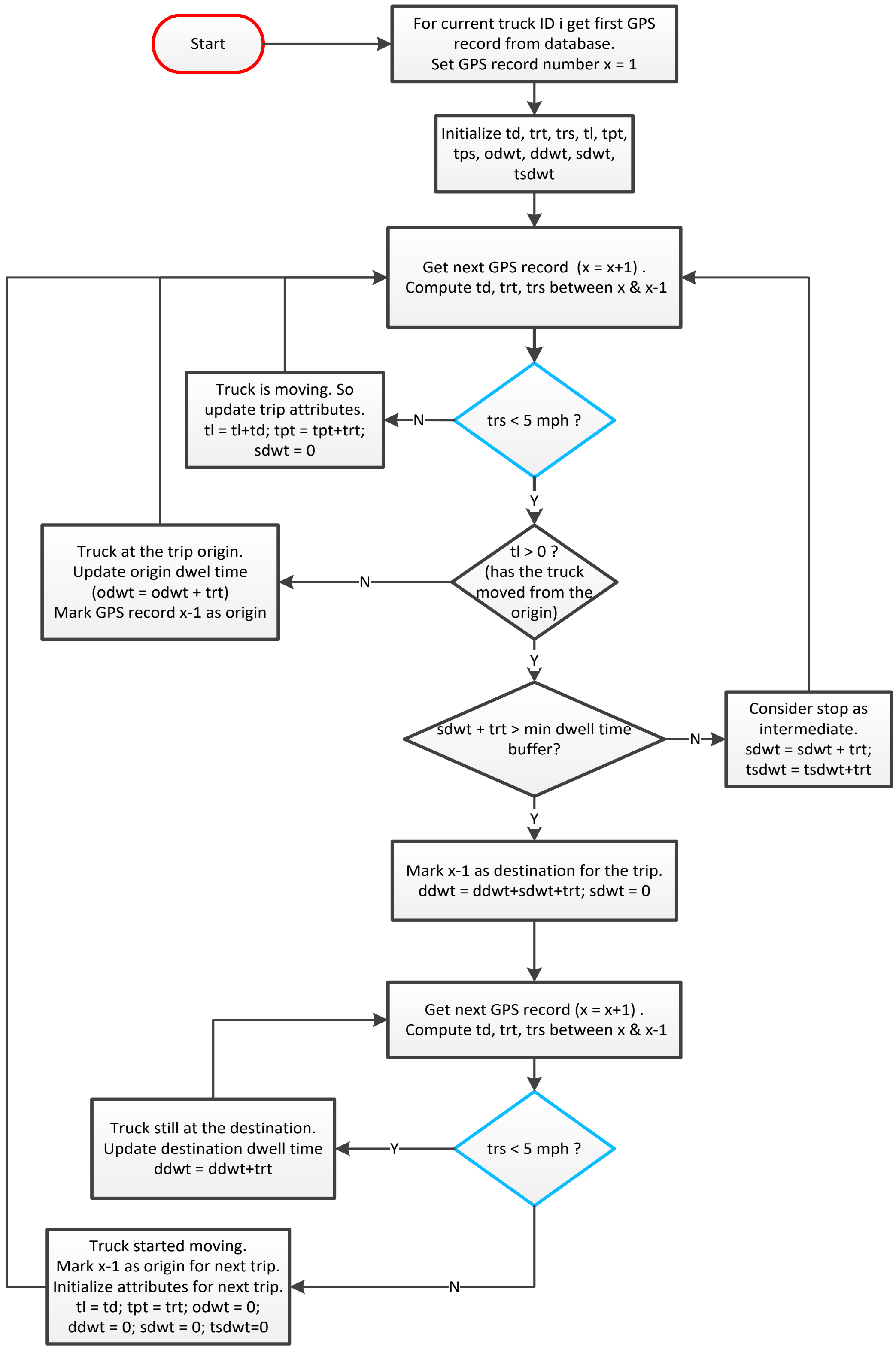

Figure 3.1: Algorithm for identifying truck trip ends from raw GPS data (Pinjari et al. 2016)

3.2.2.3 Specific characteristics of algorithm designed for short-haul trips

The original algorithm developed by Pinjari et al. (2014) was geared toward deriving long-haul freight trips. However, Pinjari et al. (2016) modified the original algorithm for short-haul trips.

First, a minimum dwell-time buffer of 5 minutes was used to capture all the possible stops that trucks make. As short-haul trucks generally make delivery stops that are small in time duration, by choosing a dwell-time buffer greater than 5 minutes, many valid stops could be missed. The Pinjari et al. (2014) algorithm used longer durations for minimum dwell-time buffer because of their focus on long-haul trips.

Second, Pinjari et al. (2014) eliminated truck stops in rest areas, regardless of the duration of those stops. However, in Pinjari et al (2016), truck stops in rest areas were also retained.

3.2.2.4 Trip outputs from algorithms implemented in study

The above-discussed procedure was applied to all four months (October and December 2015 and April and June 2016) of refined truck-GPS data comprising 90,184,869 GPS records. This resulted in 1,076,371 truck trips. Table 3.1 shows the summary of the trips derived from the data. Summaries are provided for both data formats (i.e., formats S and R) separately.

Table 3.1: Descriptive Statistics of Trips Extracted from Raw Data

	Format S	Format R	Both format S and R
Number of trips extracted	673,334	403,037	1,076,371
Number of unique truck IDs	32,180	46,306	78,486

For each of the over 1 million trips derived, the trip origin and destination locations, trip start and end times, trip travel time, and trip destination dwell time were output from the above-described algorithm. In addition, for each trip, an additional set of software codes was written to extract all the GPS data points (from the raw data files) between the origin and destination locations. This raw data was used later for deriving the truck travel route between the origin and destination locations.

3.3 Validation of Truck Trips Derived from GPS Data

To validate the truck trips derived from the ATRI data, a random sample of 200 trips was selected from the derived trips. These trips and their corresponding GPS records were checked thoroughly for any anomalies. First, the origins and destination of all 200 trips were manually verified by overlaying them on Google Earth. Second, all other trips for the trucks corresponding to the selected sample were verified to be sure that the destination of the first trip should be same as the origin of the next trip. In addition to the trips, it was also verified that the extracted GPS records corresponding to the derived trips should not have missing records, especially near the geographic boundaries of the study region where certain trips might start and end inside the study are but still take a route that is outside the study region. In addition, certain trips with extremely low or high travel times and travel speeds were verified to check for any anomalies.

It was observed that 47% origins and 41.94% destinations of the trips derived from the GPS data were within 50 feet of the roadway network. To confirm that these origins and destinations made sense and were not on a roadway in the middle of nowhere, their proximity to a freight activity generator was checked. It was found that out of 47% of origins that are within 50 feet of the road network, 60.19% were within 1,000 feet of a freight activity generator and 75.86% were within 2,000 feet of a freight activity generator. Similarly, out of 41.94% of destinations that were within 50 feet of the road network, 58.78% were within 1,000 feet of a freight activity generator and 75.59% were within 2,000 feet of a freight activity generator. Proximity to freight activity generators for a large proportion of data gave us greater confidence in the trip ends.

3.4 Refinement of Truck Trips Database Derived from Raw Data

The initial set of trips extracted from ATRI data was refined in different ways before proceeding further. First, trips that start and end in same TAZ were removed, because they were likely to be of very short length and were less likely to be of interest for route choice analysis. Highway networks typically do not provide more than a single route option for such short trips. For the same reason, trips between different TAZs of less than 2 miles in length were also removed. Second, trips with a ratio between direct OD distance (or airline OD distance) and trip length less than 0.7 were removed. This is because such trips were likely to have valid intermediate stops or destinations (between the origin and destination) that were not captured by the algorithm. The routes for such trips were likely to be circuitous (due to intermediate stops) and did not necessarily represent routes without intermediate stops. Table 3.2 provides the number (percentage of initial trips in parentheses) of trips retained after each stage of refinement.

Table 3.2: Number of Trips Retained after Each Refinement Stage

	Dataset S	Dataset R
Initially derived trips	673,334	403,037
After removing trips that start and end in same TAZ	554,427 (82.34%)	275,454 (68.34%)
After removing trips of less than 2 miles in length	499,349 (74.16%)	263,667 (65.42%)
After removing trips with ratio of direct OD distance and trip length less than 0.7	448,213 (66.57%)	228,689 (56.74%)

3.5 Descriptive Summary of Derived Trips

This section provides the descriptive characteristics of the validated and/or refined truck trips from both dataset S and dataset R, as described earlier. In total, 676,902 trips were derived collectively from both dataset S and dataset R. Of these 676,902 trips, 228,689 were derived from dataset R and 448,213 were derived from dataset S. Figure 3.1 shows the trip length distribution of the derived trips with very few trips being more than 150 miles in distance. Figure 3.2 shows the distribution of trip time across the derived trips and highlights that most trips were less than 2 hours in duration.

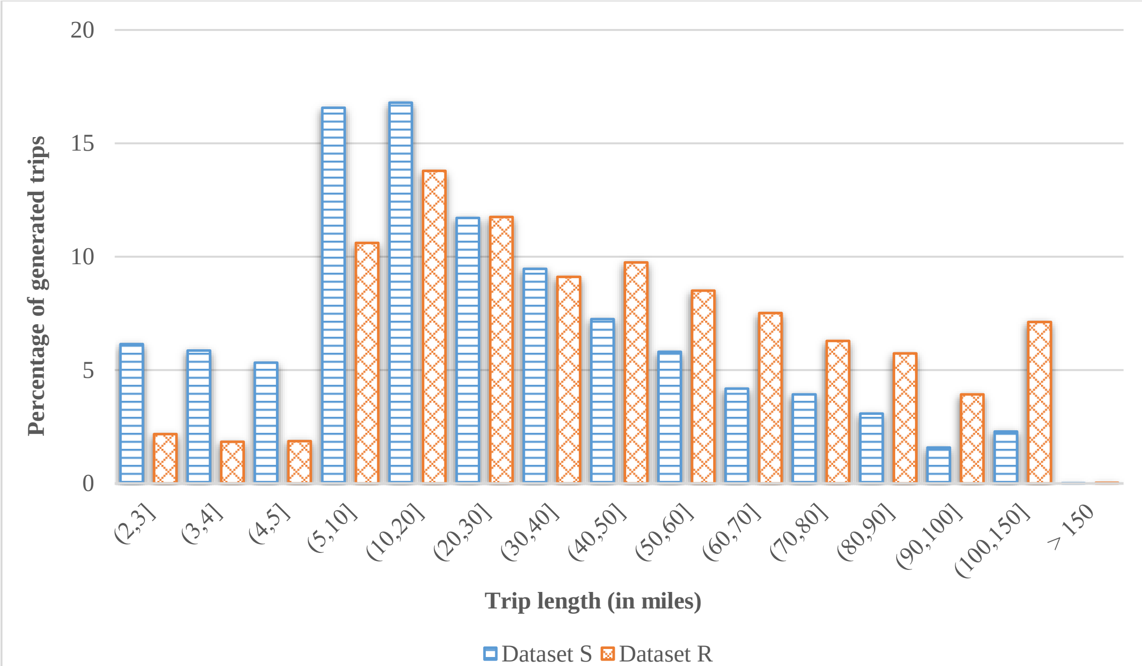

Figure 3.1: Distribution of estimated length of derived trips

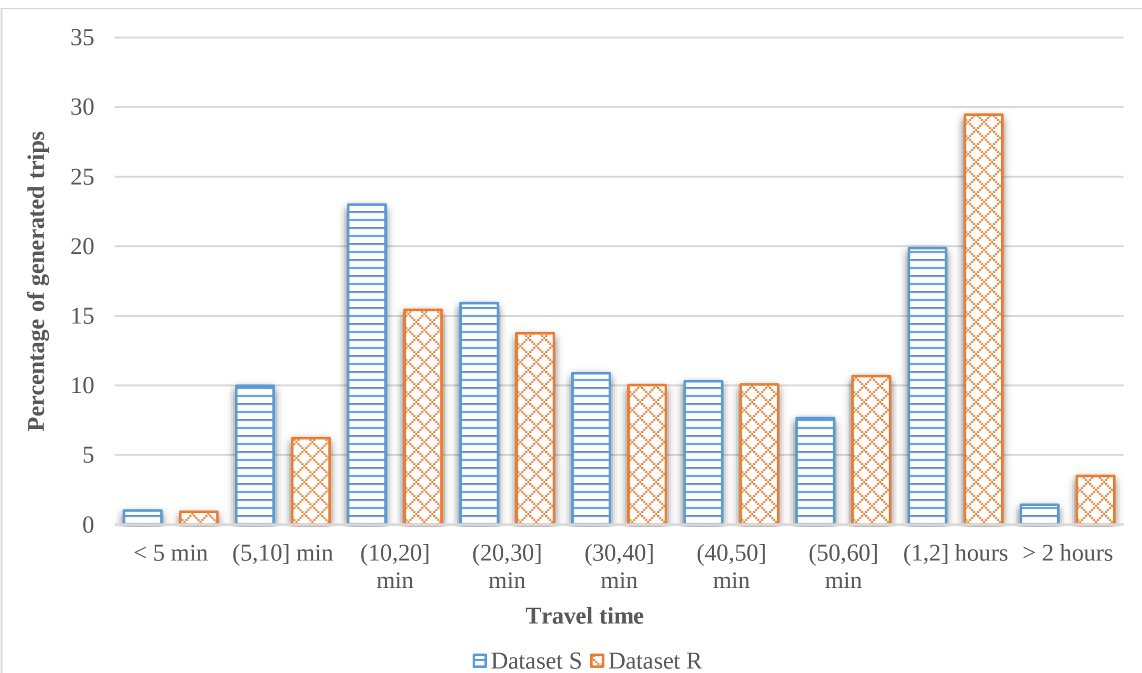

Figure 3.2: Distribution of travel time of derived trips

Figure 3.3 presents the distribution of average trip speed across derived trips, and Figure 3.4 shows the distribution of direct OD distance and trip length across derived trips. It must be noted that in Figure 3.4, there are no trips with the ratio of direct OD distance and trip length of less than 0.7, as such were already removed from the data.

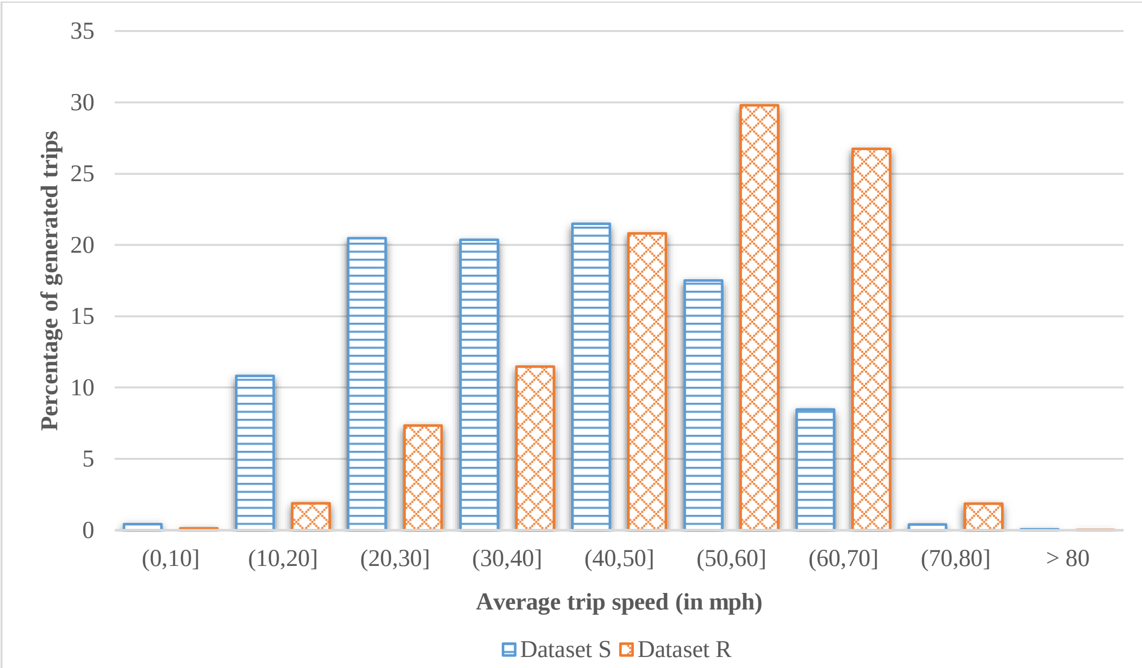

Figure 3.3: Distribution of average speed of derived trips

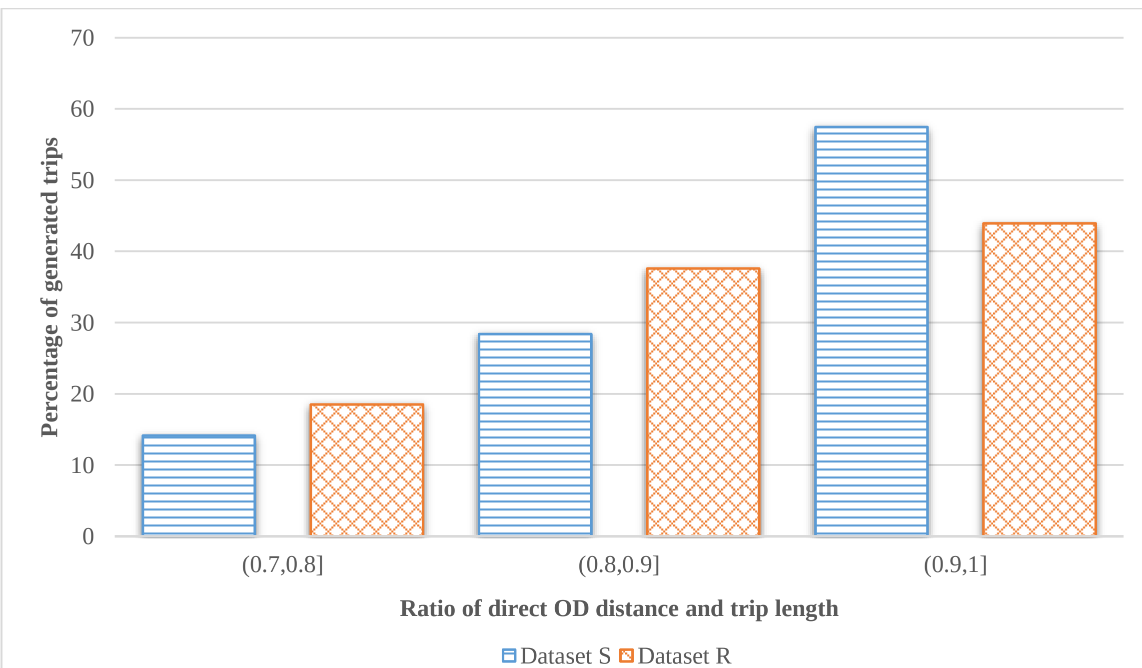

Figure 3.4: Distribution of ratio of direct OD distance and trip length across derived trips

Tables 3.3 and 3.4 provide the cross-tabulation between the number of GPS records corresponding to a trip and the trip length for datasets S and R, respectively. The percentage of trips with less than five GPS records is 9.63 and 6.35 in datasets S and R, respectively.

Table 3.3: Cross-Tabulation between Trip Length and Number of GPS Records for Dataset S

Trip Length (miles) / No. of GPS Records	(2,5]	(5,15]	(15,20]	(20,25]	(25,30]	(30,50]	(50,100]	Total
(0,4]	3.94	2.83	0.49	0.41	0.32	0.89	0.75	9.63
(4,10]	10.91	14.93	3.49	2.42	2.3	4.84	1.74	40.63
(10,20]	2.61	7.31	2.13	2.18	2.24	7.92	8.57	32.96
(20,30]	0.16	1.44	0.53	0.53	0.47	1.7	4.33	9.16
(30,40]	0.06	0.37	0.27	0.36	0.35	0.5	1.19	3.1
(40,50]	0.03	0.09	0.07	0.09	0.17	0.51	0.37	1.33
(50,100]	0.03	0.14	0.04	0.05	0.08	0.7	1.73	2.77
(100,150]	0.01	0.02	0.01	0	0	0.02	0.32	0.38
> 150	0	0.01	0	0	0.01	0.01	0.02	0.05
Total	17.75	27.14	7.03	6.04	5.94	17.09	19.02	100%

Table 3.4: Cross-Tabulation between Trip Length and Number of GPS Records for Dataset R

Trip Length (miles) / No. of GPS Records	(2,5]	(5,15]	(15,20]	(20,25]	(25,30]	(30,50]	(50,100]	Total
(0,4]	0.63	0.96	0.28	0.34	0.32	1.31	2.51	6.35
(4,10]	1.54	1.29	0.07	0.03	0.03	0.08	0.03	3.07
(10,20]	1.65	4.1	0.75	0.2	0.08	0.44	0.07	7.29
(20,30]	0.88	4.13	2.09	1.36	0.47	0.76	0.27	9.96
(30,40]	0.44	2.49	1.31	1.4	1.26	1.84	1.44	10.18
(40,50]	0.28	1.56	0.84	1.05	1.04	2.77	2.02	9.56
(50,100]	0.59	3	1.53	1.84	1.93	10.06	14.46	33.41
(100,150]	0.16	0.72	0.31	0.4	0.37	2	10.26	14.22
> 150	0.15	0.58	0.26	0.29	0.25	1.05	3.36	5.94
Total	6.32	18.83	7.44	6.91	5.75	20.31	34.42	100%

CHAPTER 4: DERIVING TRUCK ROUTES

4.1 Introduction

This chapter discusses the procedure used to derive the travel paths (or routes) of truck trips derived in Chapter 3. The procedure for deriving routes using GPS data consisted of three broad steps: 1) map-matching data set preparation, 2) map-matching process, and 3) route generation. These three steps are explained in detail in this chapter, along with validation, refinement results, and a descriptive summary of the derived routes. Appendix A provides the description of the GIS shape file for all the generated routes, and Appendix B provides the description of the *link-list* comma separated file (.csv), which can be used, along with the Navteq network, to determine the link level attributes of any given route.

4.2 Map-Matching Dataset Preparation

After conversion of raw GPS data into 676,902 truck trips, the data were further refined to obtain a data set that could be easily map-matched to an underlying network. This section discusses the procedure used to refine the GPS data to obtain the dataset that can be easily map-matched. Much of this discussion is derived from Kamali et al. (2016) and was developed as part of a previous FDOT project by Pinjari et al. (2016), albeit with some improvements to better suit the data used in the current project. Steps involved in the map-matching dataset preparation are discussed in the following subsections.

4.2.1 Selection of Trips Suitable for Accurate Derivation of Routes

Spatial proximity between consecutive coordinates in the GPS data plays an important role in the accuracy of routes generated from the data. The larger the spatial gap between consecutive coordinates, the less accurate a derived route would be. Since spatial separation in the data is dependent on the temporal frequency of the data, it is likely that the higher-frequency data sets provide more accurate measurement of routes as well. In this step, trips that follow any one of the following criteria were selected from the data set of 676,902 derived truck trips:

- Maximum time gap (or ping rate) between any consecutive coordinates < = 10 minutes and corresponding spatial gap < = 10 miles
- Maximum time gap (or ping rate) between any consecutive coordinates > 10 minutes and corresponding spatial gap < = 5 miles

As shown in the highlighted areas of Figures 4.1 and 4.2, 68.63% of the derived truck trips in dataset S and 80.31% of derived truck trips in dataset R were retained and used for further processing. In total, 432,188 trips were retained after applying the selection criteria mentioned in this section.

Table 4.1: Cross-Tabulation of Maximum Time Gap (Ping Rate) and Corresponding Spatial Gap (in miles) for Trips in Dataset S

Max Ping Rate \ Space Gap	(0,1]	(1,5]	(5,10]	(10,15]	(15,20]	(20,25]	(25,30]	> 30	Total
< 1 min	1.65	3.37	0	0	0	0	0	0	5.02
(1-2] min	1.65	2.29	0	0	0	0	0	0	3.94
(2-5] min	0.42	23.35	12.60	0	0	0	0	0	36.37
(5-10] min	0.02	6.34	2.11	0.22	0	0	0	0	8.69
(10-15] min	0	4.45	8.61	13.73	10.44	0.01	0	0	37.25
(15-20] min	0	0.16	0.26	0.20	0.17	0.05	0	0	0.85
(20-25] min	0	0.06	0.09	0.10	0.16	0.43	0.78	0	1.62
(25-30] min	0	0.06	0.08	0.09	0.17	0.30	0.42	0.50	1.62
(30-45] min	0	0.10	0.12	0.09	0.13	0.18	0.15	0.33	1.10
(45-120] min	0	0.16	0.29	0.23	0.25	0.23	0.20	1.19	2.55
> 2 hours	0	0.06	0.11	0.11	0.10	0.07	0.06	0.48	0.99
Total	3.75	40.39	24.28	14.78	11.43	1.27	1.62	2.49	100%

Table 4.2: Cross-Tabulation of Maximum Time Gap (Ping Rate) and Corresponding Spatial Gap (in miles) for Trips in Dataset R

Max Ping Rate \ Space Gap	(0,1]	(1,5]	(5,10]	(10,15]	(15,20]	(20,25]	(25,30]	> 30	Total
< 1 min	7.01	9.92	0	0	0	0	0	0	16.92
(1-2] min	1.27	30.79	0	0	0	0	0	0	32.06
(2-5] min	0.21	13.03	0.18	0	0	0	0	0	13.42
(5-10] min	0.08	7.28	3.81	0.06	0	0	0	0	11.22
(10-15] min	0.01	2.13	3.16	2.66	0.10	0	0	0	8.06
(15-20] min	0	0.19	0.21	0.56	0.27	0.26	0	0	1.49
(20-25] min	0	0.07	0.07	0.23	1.04	7.10	0.04	0	8.55
(25-30] min	0	0.04	0.05	0.04	0.49	0.44	0.14	0.02	1.22
(30-45] min	0	0.05	0.07	0.08	0.13	0.69	0.13	0.18	1.33
(45-120] min	0	0.12	0.20	0.23	0.24	0.36	0.25	2.68	4.08
> 2 hours	0	0.15	0.15	0.13	0.12	0.09	0.10	0.90	1.64
Total	8.57	63.78	7.88	3.99	2.40	8.94	0.66	3.78	100%

4.2.2 Preparation of GPS Data for Map-Matching

For all the trips retained after the selection criterion mentioned in Section 4.2.1, the corresponding raw GPS data had to be prepared for map-matching. To do so, the following three steps were employed on the raw GPS data of each trip:

1. GPS coordinates within a 1,000-ft radius of the origin or destination of each trip were removed. For trips that started or ended in urban areas with a high density of highway network links, it was not easy to accurately map-match the raw GPS data because of the absence of many minor roadway network links in the Navteq network used in the study. Mismatching these GPS points would lead to loops (circuitous maneuvers) in the generated routes. As this step also removes the origin and destination GPS points, these points were later added back to the set of GPS points corresponding to a trip.
2. The remaining GPS coordinates were space-sampled to be at least 1,000 ft apart. This was done to eliminate GPS points that were too close to each other and did not help enhance the accuracy of matching the points to the road network. The space-sampling approach is different from the time-sampling approach mentioned in Kamali et al. (2016), but space sampling helps to keep consistency across the spatial distribution of the consecutive GPS points. In the time-sampling approach, GPS points can still be very close to each other after a period of time (e.g., 1 or 2 minutes when a truck stops at a traffic light, etc.)
3. In this step, trips with less than 5 GPS points were removed from the database, as the number of GPS points in those trips was considered below the number needed to accurately derive the travel route.

After these steps, 288,468 truck trips were retained from the initial 688,638 truck trips. Table 4.3 provides a summary of the number of truck trips and GPS points removed after each step for both dataset S and dataset R. The percentage values in parentheses show the percentage of total trips or GPS points lost after each step employed above.

Table 4.3: Summary of Remaining Trips and GPS Points after Each Map-Matching Data Processing Stage

Steps	Data S		Data R	
	Number of Trips	***Number of GPS Points***	***Number of Trips***	***Number of GPS Points***
Initial data	442,593	6,116,191	226,045	11,477,043
1) Maximum time gap and spatial gap	259,679	3,995,810	172,509	9,004,085
	Lost: 182,914 (41.3%)	Lost: 2,120,381 (34.7%)	Lost: 53,536 (23.7%)	Lost: 2,472,958 (21.5%)
2) Remove GPS points within 1,000ft radius of endings	259,679	2,803,210	172,509	7,768,555
	Lost: 0	Lost: 1,192,600 (19.5%)	Lost: 0	Lost: 1,235,530 (10.8%)
3) Space sampling of every 1,000 ft	259,679	2,601,599	172,509	6,913,898
	Lost: 0	Lost: 201,611 (3.3%)	Lost: 0	Lost: 854,657 (7.4%)
4) Remove trips with less than 5 GPS points	127,272	2,345,443	161,196	6,892,297
	Lost: 132,407 (29.9%)	Lost: 256,156 (4.2%)	Lost: 11,313 (5%)	Lost: 21,601 (0.2%)

4.3 Map-Matching Process

After selecting the truck trips that are suitable for generating routes and preparing the data for map-matching, the retained GPS points were map-matched to an underlying network. Map-matching is a technique that uses a combination of GPS location data and roadway network data to identify the correct link that has been traversed by the vehicle on the network. The map-matching procedure presented here is a modified version of the one presented in Kamali et al. (2016), which was originally proposed by Yang et al. (2005). The methodology is briefly described below.

- *Step 1* – All the GPS points that were within 500-ft buffer around the highway interchanges were removed. Points close to highway interchanges are difficult to map-match and can lead to major detours from the actual route. This was done by drawing a 500-ft buffer around the highway interchanges; these interchanges were identified using the attributes present in the Navteq network that indicated if a network link belongs to a highway interchange. The GPS points falling within the 500-ft buffer were identified by intersecting the GPS data layer with the 500-ft buffer layer around the highway interchanges.
- *Step 2* – The distance of each GPS point from the first nearest link in the network was calculated. This distance was denoted as D1, and all GPS points where D1 > 500 ft were removed. Subsequently, distance of each GPS point from the second nearest link in the network was calculated and denoted as D2. This was done using the "Generate Near Distance Table" tool in the ArcGIS environment.
- *Step 3a* – The angle between the geographic north and a perpendicular line drawn from each GPS point to the first- and second-nearest links in the network was calculated. If the

location of the link was east of the location of the GPS point, the angle was measured in the clockwise direction; if the location of the link was west of the GPS point, the angle was measured in the anti-clockwise direction. These angles were denoted as A1 and A2. If 170° < (A1 + A2) < 190°, the GPS point was supposedly between two parallel roads and was difficult to be map-matched accurately. All such GPS points were removed from the dataset.

- *Step 3b* – Of the points removed in Step 3a, those that fell within the 65-ft buffer of just one roadway intersection were retained and were map-matched to the intersections instead of the links.
- *Step 4* – All trips with fewer than five GPS points were removed, as these trips did not have enough GPS points for accurate route determination.

Table 4.4 provides a summary of the GPS points lost at each stage of the map-matching process, for both dataset S and dataset R. In the end, GPS data corresponding to 243,438 trips were retained and used to generate the routes.

Table 4.4: Summary of Retained Trips and GPS Points after Each Trip-Selection Criteria

Steps		Data S	Data R
Input data from Section 4.2	Trips	127,272	161,196
	GPS points	2,345,443	6,892,297
1) Remove GPS points within 500ft of highway interchanges		1,808,143	5,422,817
		Lost: 537,300 records (22.9%)	Lost: 1,469,480 records (21.3%)
2) Remove GPS points farther than 500ft of any road link		1,793,786	5,386,991
		Lost: 14,357 records (0.6%)	Lost: 35,826 records (0.5%)
3a) Remove GPS points between two parallel links		1,430,319	4,099,804
		Lost: 363,467 records (15.5%)	Lost: 1,287,187 records (18.7%)
3b) For GPS points between two parallel links, keep points within 65ft of only one intersection		11,446	23,615
		Gain: 11,446 records (0.5%)	Gain: 23,615 records (0.3%)
Output data after map-matching process	Trips with at least 5 GPS points	92,868	150,570
	Corresponding GPS points	1,526,193	4,394,770

4.4 Route Generation

Due to the infrequent nature of the GPS data in this project, the GPS points retained for the map-matching process may not be able to capture all links traversed by a truck between its origin and destination. Thus, for each trip, the complete travel route was generated by deriving the shortest time path between each trip's origin and destination and through the intermediate, map-matched roadway links. Using ArcGIS's Network Analyst extension, all truck routes were generated based on the map-matched GPS data and Navteq road network. Network Analyst uses a modified version of Dijkstra's shortest path algorithm to find shortest “weight” path between a set of points. Free flow travel time was used as the weight here. A python script with ArcPy library was used to automate the process of iteratively accessing Network Analyst to extract the route and the underlying network links for an OD pair.

The python script gave two outputs. The first output was a GIS shape file of all generated routes, where each geographic feature in the shape file corresponded to a trip. The attribute table of the shape file contains all the necessary trip information like origin and destination TAZ, start and end time of the trip, trip length, trip travel time, start and end county etc. Another output was a comma separate file (.csv) containing a list of the links traversed by trucks on each trip. The link list was extracted to retain the attributes of each link traveled in a route for any subsequent analysis such as route choice modeling and route diversity measurement.

In total, routes were generated for 243,438 trips generated from datasets S and R. The route-generation process does not guarantee that all generated routes are accurate, as it tries only to find the shortest “weight” path traversing the map-matched points. If a GPS point was wrongly map-matched, it was possible that the generated route for that particular trip might have irrational detours. Therefore, a validation exercise was conducted to verify the accuracy of the generated routes.

4.5 Route Validation

The routes generated using the procedure described above were validated in terms of their consistency and feasibility. The routes were considered to be consistent and feasible if:

- The direction of the travel was consistent throughout the entire route.
- There were no major loops or detours throughout the entire route.
- There were no impossible maneuvers throughout the entire route (such as switching to a lane underneath a bridge).

A sample of 300 randomly-selected trips from the generated routes was validated in terms of feasibility and consistency. During the validation procedure, the generated routes were plotted on top of an aerial image to ensure that the direction of travel was consistent throughout. The generated routes also were plotted against the raw GPS data to check if they followed the raw records. These visual checks helped to detect routes with problems such as loops, missing links, detours, or a different path due to eliminated GPS points after the map-matching stage.

Of 300 routes, only 9 had the following issues:

- 1 route with a 5-mile detour due to a GPS mapped to another link at an intersection.
- 2 routes with loops on highways due to one GPS point mapped to the other side of the opposite travel direction.
- 6 routes with different segment ranges of 5–10 miles due to eliminated 2–5 GPS points between that segment.

4.6 Route Refinement

The validation process showed that although the majority of routes were feasible and consistent, up to 10% of routes still had some loops due to some wrongly map-matched GPS points. One important output mentioned earlier was a list of all links traversed to reach the destination in the order in which they were traversed. Each link was denoted by a unique OBJECT ID. A route with no loops will contain no repeated links, and a route that traverses the same road link more than once will have repeated OBJECT IDs appearing in the output file. Thus, a systematic way to identify routes with major loops is to measure the length of the repeated links. All routes with a total length of repeated links more than 1 mile are routes with major loops or inconsistent direction of travel. After removal of 10,109 such routes (~4.2% of all generated routes), 233,329 routes were retained. Figure 4.1 shows the final set of routes derived using the map-matched GPS data.

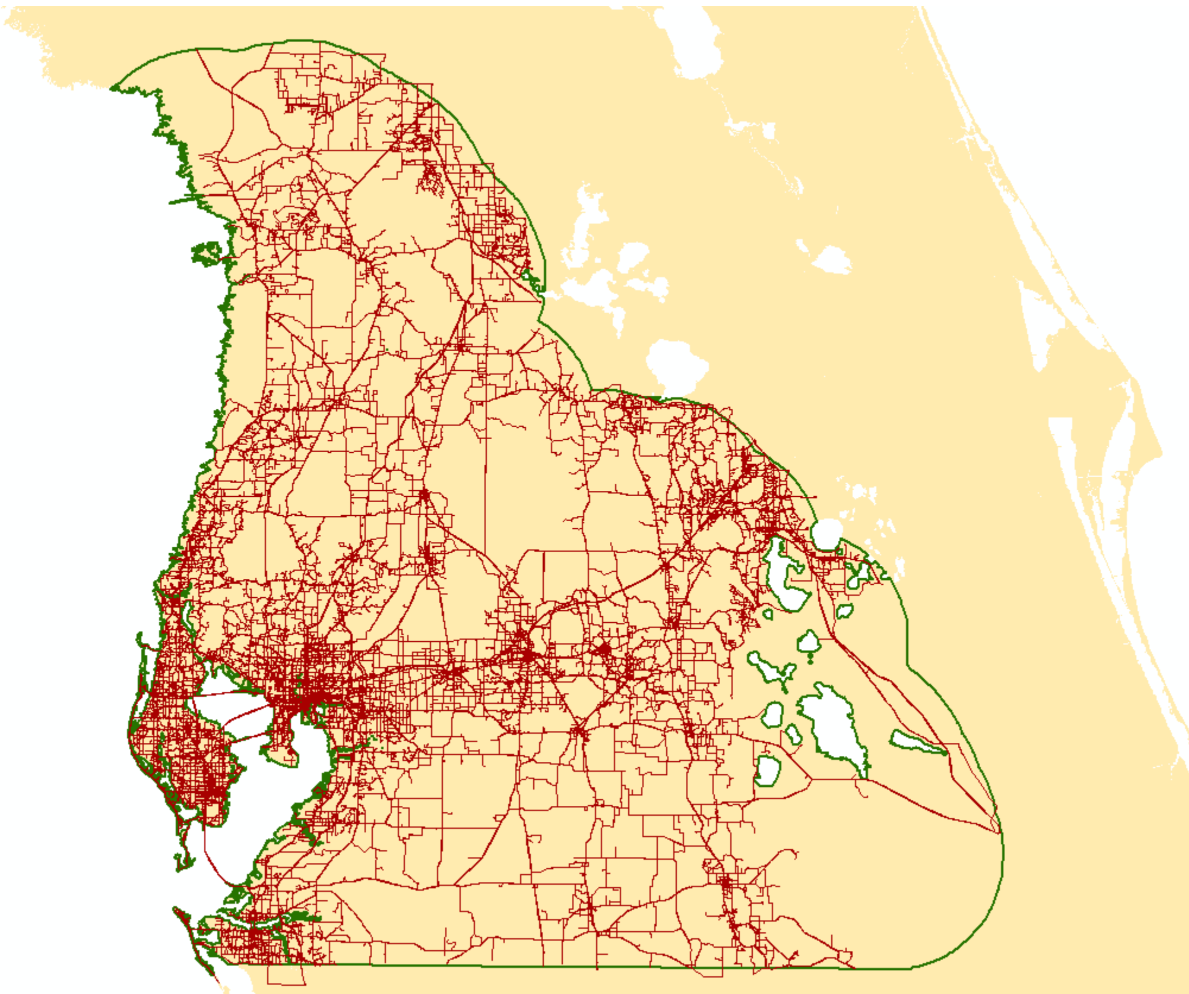

Figure 4.1: Final Tampa Bay routes with no major loops

4.7 Descriptive Summary of Derived Routes

Of 233,329 routes, there were 90,816 routes from dataset S and 142,513 routes from dataset R. Figure 4.2 shows the route length distribution of the derived routes. Route length is an accurate measure of the travel path calculated using the actual route traversed on the roadway network.

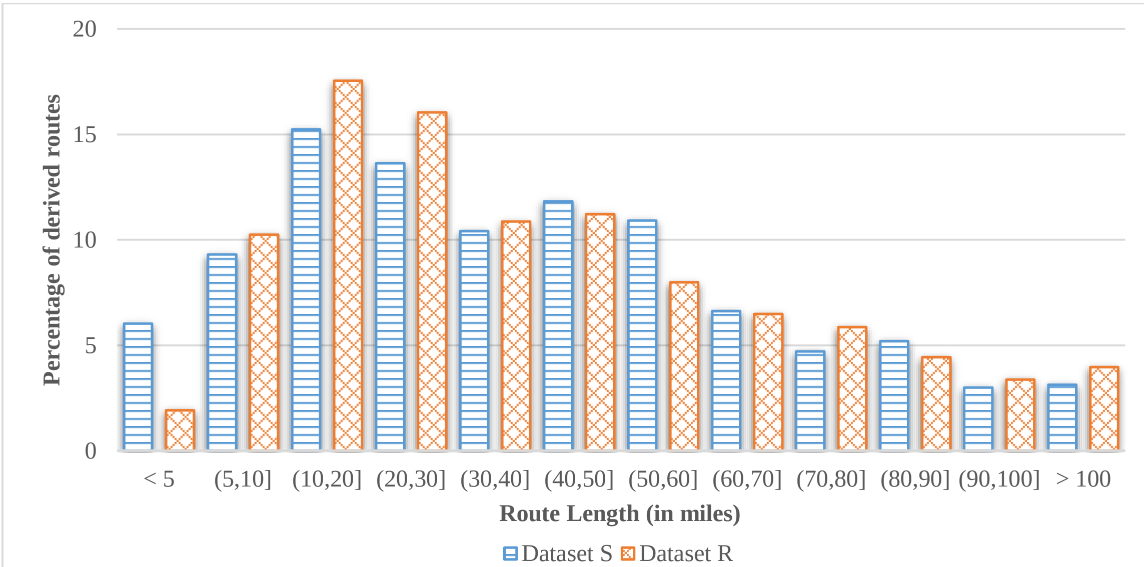

Figure 4.2: Distribution of route length of final trips

Figure 4.3 shows the free-flow travel time distribution for the final trips. Free flow travel time is the time a truck would take to travel along a route during free flow conditions. Travel time for a trip is typically more than or equal to the free flow travel time.

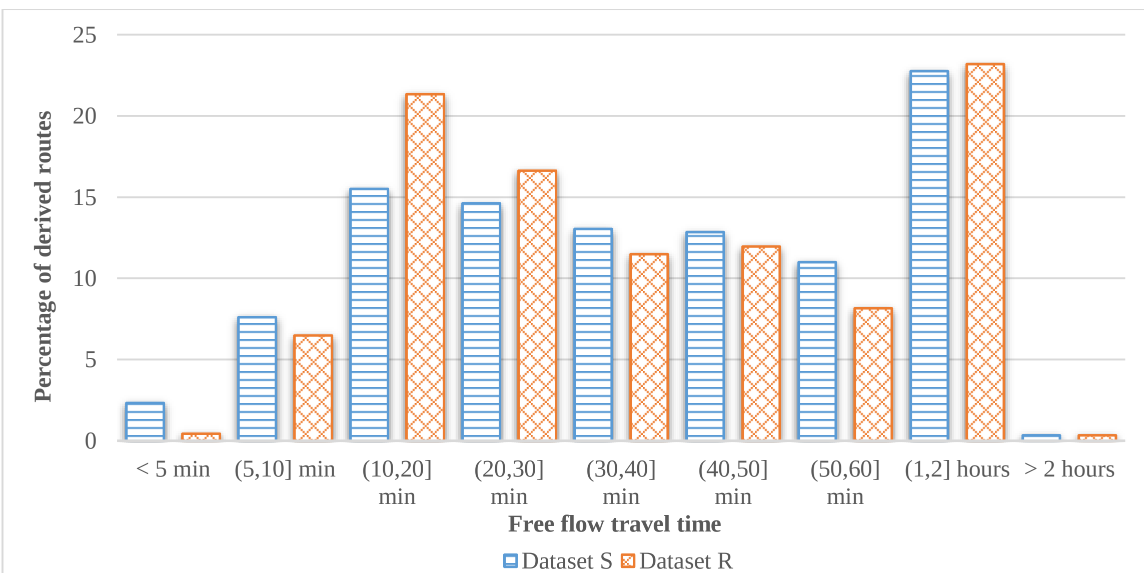

Figure 4.3: Distribution of free flow travel time of final trips

Figure 4.4 shows the distribution of the number of map-matched GPS points for each route. Routes generated for the trips with a higher number of map-matched GPS points are more likely to be close to the actual path taken by the truck. In that sense, routes generated for the trips belonging to dataset R were likely to be more accurate than trips belonging to the dataset S.

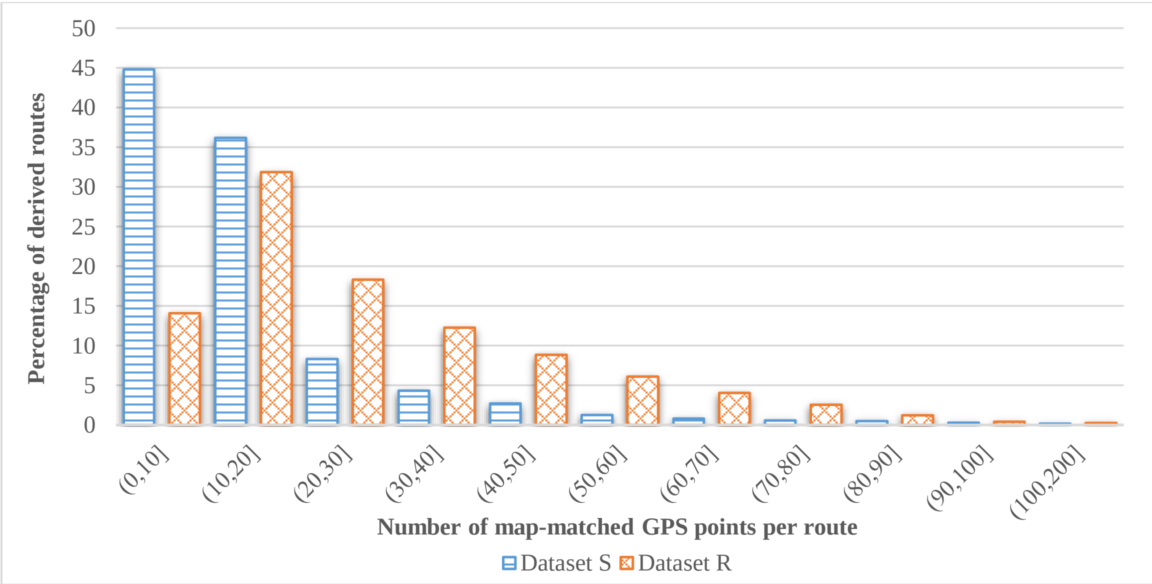

Figure 4.4: Distribution of number of map-matched GPS points per route

Figure 4.5 shows the distribution of the number of links for each route. Trips with longer route length generally have higher number of links.

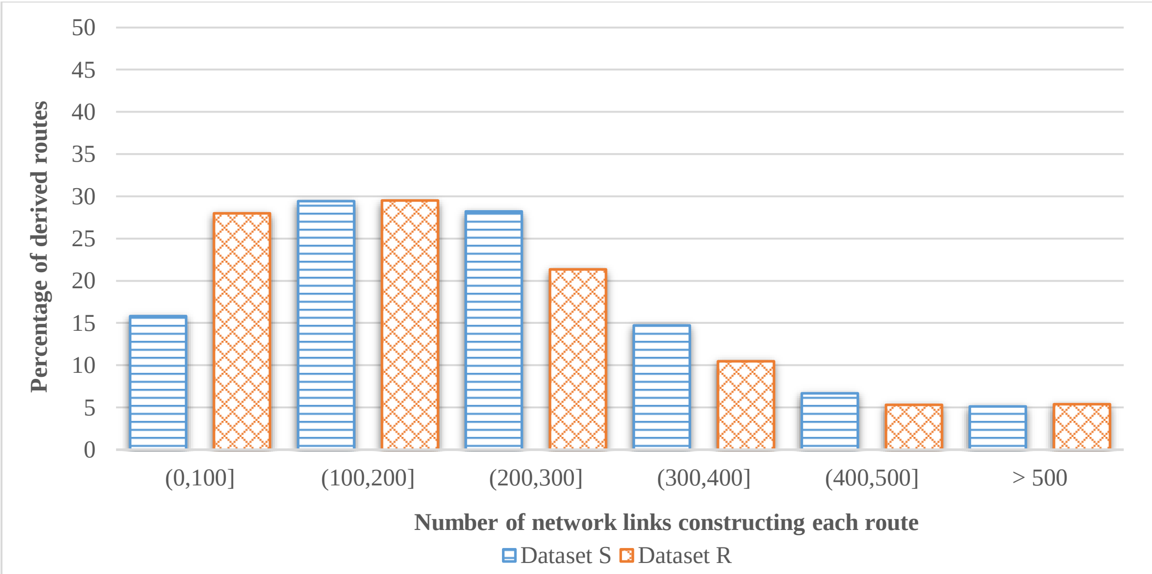

Figure 4.5: Distribution of number of network links per route

CHAPTER 5: ANALYSIS OF TRUCK TRIPS DEPARTING FROM AND ARRIVING AT PORT TAMPA BAY

5.1 Introduction

One task of the project was to analyze the routes taken by trucks between specific origin destination locations (or OD pair) of interest to FDOT District 7. After discussion with FDOT District 7, it was decided to analyze the trips made by trucks originating from Port Tampa Bay. This chapter presents the descriptive summary of these trips. In particular, the distribution of trip length, trip duration, and departure time of day of the trips made by trucks originating from and ending at Port Tampa Bay are discussed, as is the spatial distribution across different counties and regions of trips made by trucks originating from and ending at Port Tampa Bay and the network-level distribution of these trips.

5.2 Descriptive Summary of Selected Truck Routes

This section provides a descriptive summary of trips made by trucks originating from or ending at Port Tampa Bay. This analysis was conducted only for trips that were map-matched to derived routes. Of 233,329 final truck trips, there were 3,755 trips with origins inside Port Tampa Bay and 3,774 trips with destinations within Port Tampa Bay.

Figure 5.1 shows the route length distribution of trips originating from and ending at Port Tampa Bay.

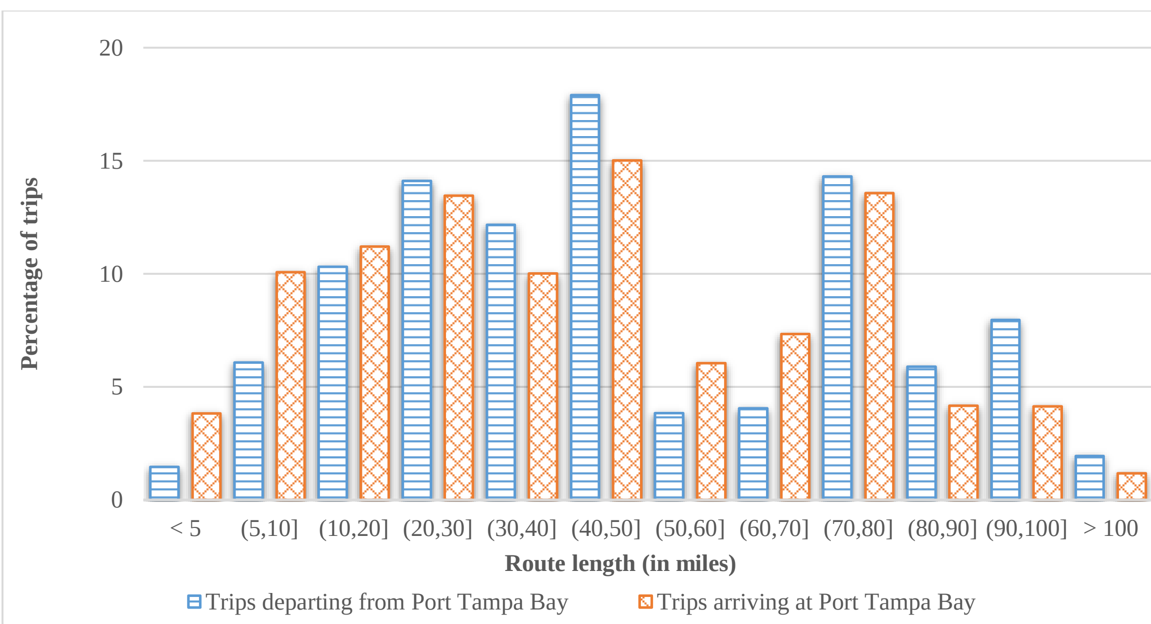

Figure 5.1: Distribution of route length of truck trips originating from and ending at Port Tampa Bay

Figure 5.2 shows the distribution of the trip duration of trips originating form and ending at Port Tampa Bay. Figure 5.3 shows the distribution of the average speed of the trips originating from and ending at the Port Tampa Bay.

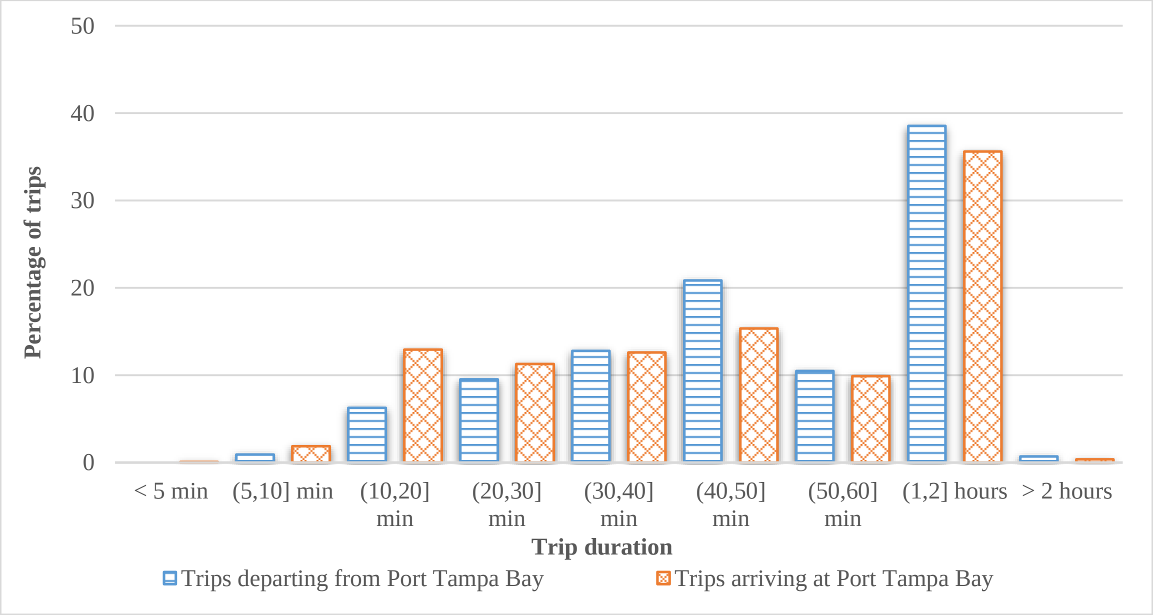

Figure 5.2: Distribution of travel time of truck trips originating from and ending at Port Tampa Bay

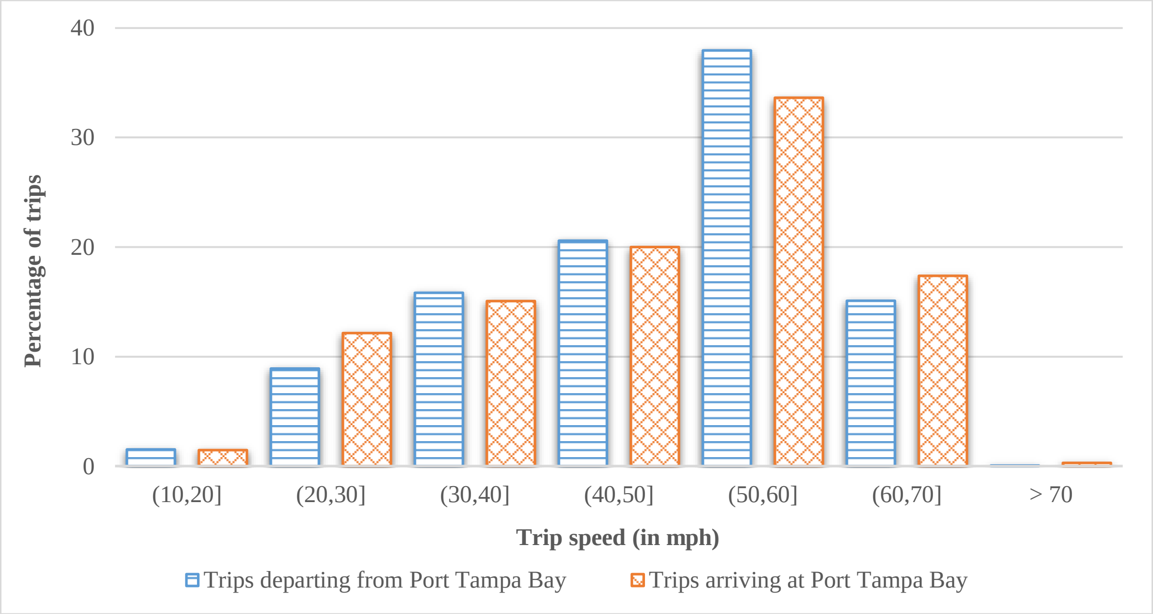

Figure 5.3: Distribution of average speed of truck trips originating from and ending at the Port Tampa Bay

Figure 5.4 shows the distribution of the departure and arrival times of trips originating from and ending at Port Tampa Bay. The arrival and departure times of trucks are distributed across the day; the peak departure time was between 1:00–2:00 PM and the peak arrival time was 12:00–1:00 PM.

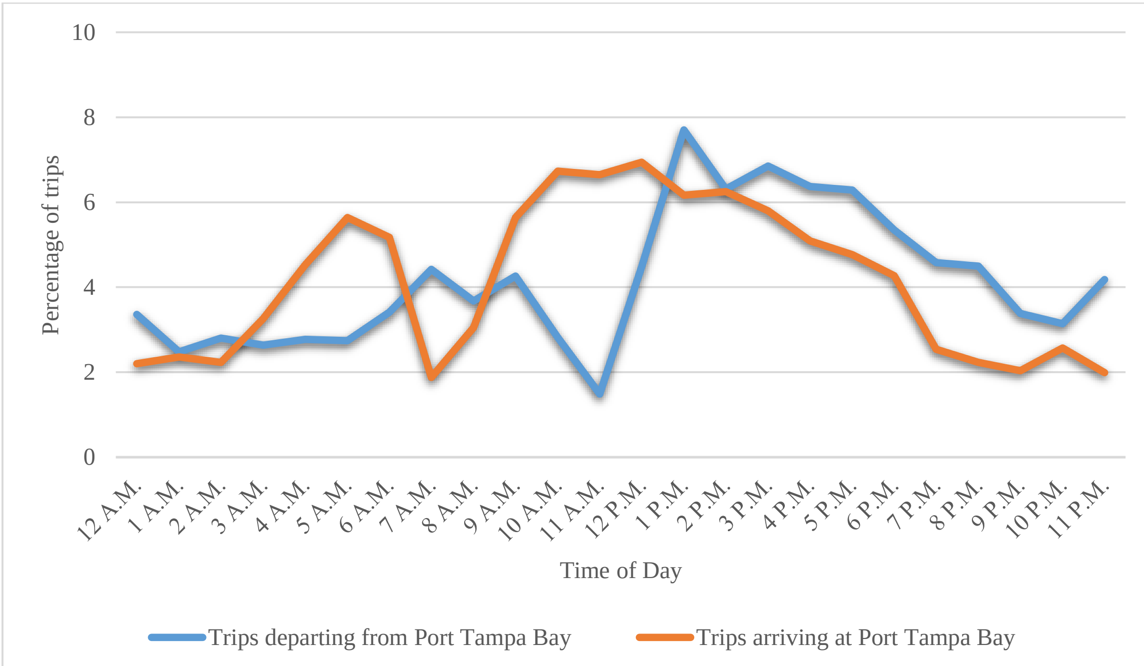

Figure 5.4: Distribution of trip departure/arrival times at Port Tampa Bay

Figures 5.5 and 5.6 illustrate the spatial distribution across different counties in the study area of the trips made by trucks originating from and ending at Port Tampa Bay, respectively. As shown in Figure 5.5, 0.66% of trips made by trucks that departed from Port Tampa Bay ended in Citrus County, 1.82% ended in Hernando County, 10.08% ended in Pasco County, 7.92% ended in Pinellas County, 21.91% ended in Hillsborough County, 11.06% ended in Polk County, and 46.56% ended in the rest of the study area. Figure 5.6 shows that 0.34% of trips made by the trucks that arrived at Port Tampa Bay started from Citrus County, 2.12% started in Hernando County, 6.08% started in Pasco County, 7.83% started in Pinellas County, 29.41% started in Hillsborough County, 16.21% started in Polk County, and 38.03% started in the rest of the study area.

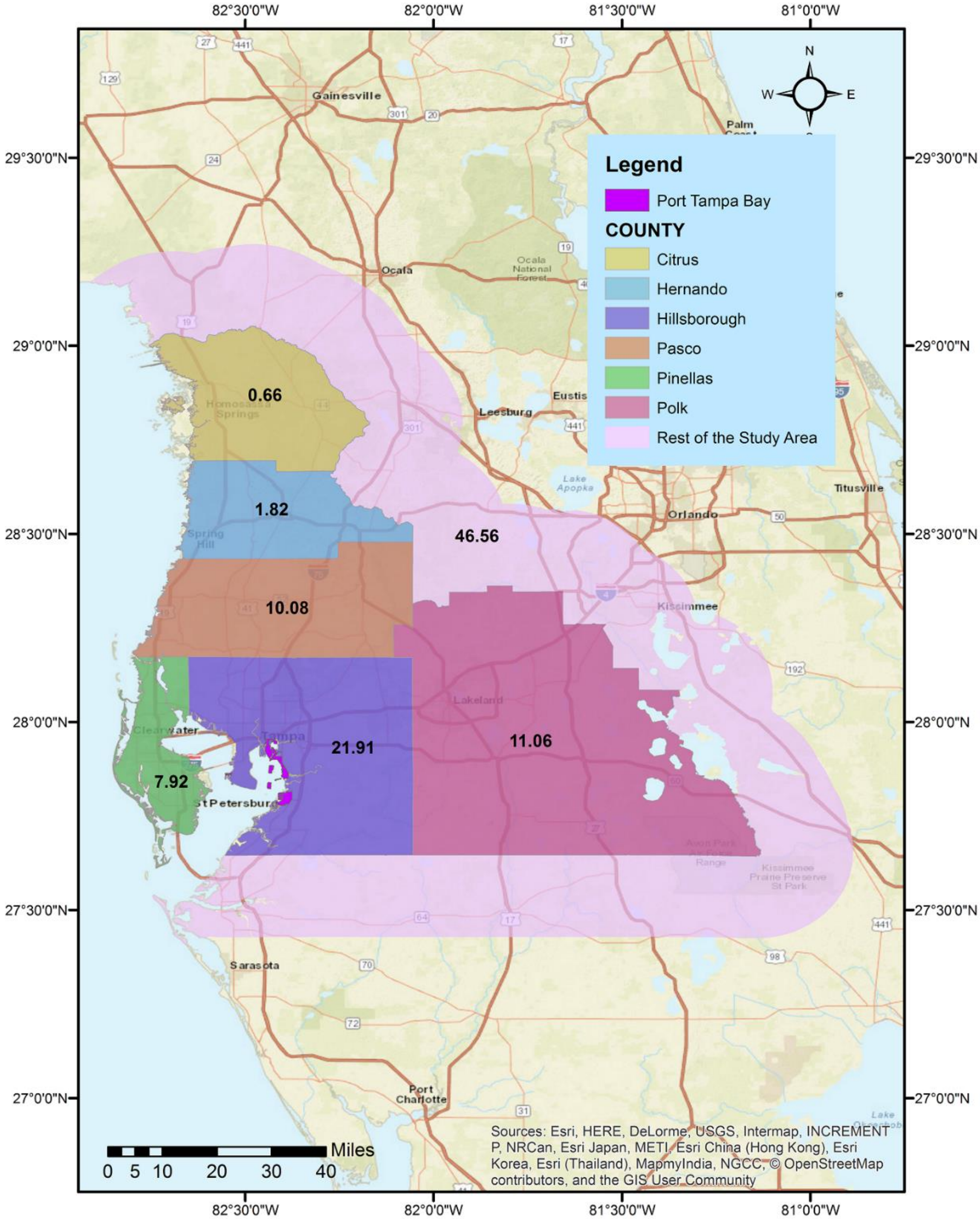

Figure 5.5: County-wise distribution of destination of truck trips originating from Port Tampa Bay

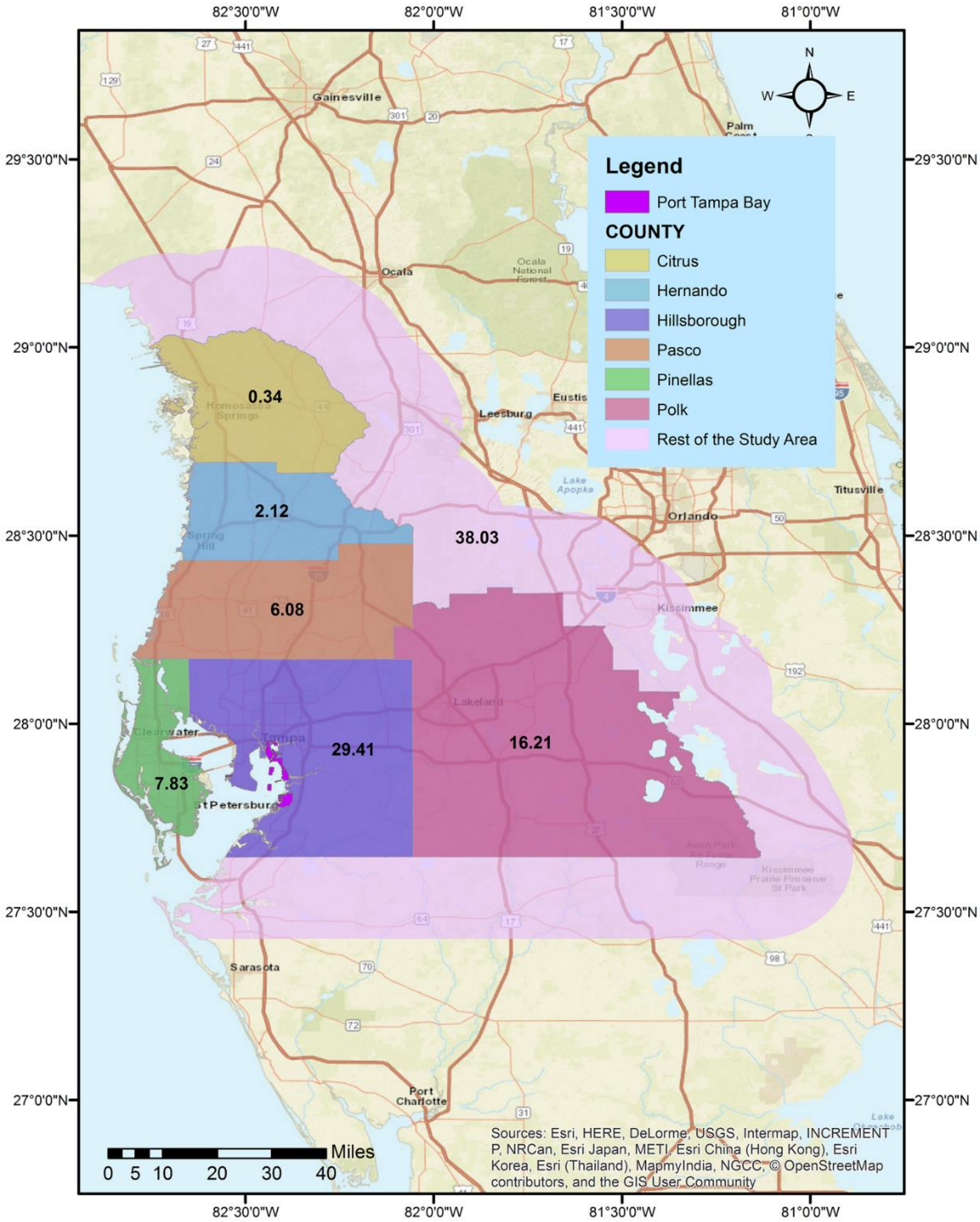

Figure 5.6: County-wise distribution of origins of truck trips ending at Port Tampa Bay

Figure 5.7 shows the spatial distribution across different radial zones of the destinations of trips originating from Port Tampa Bay.

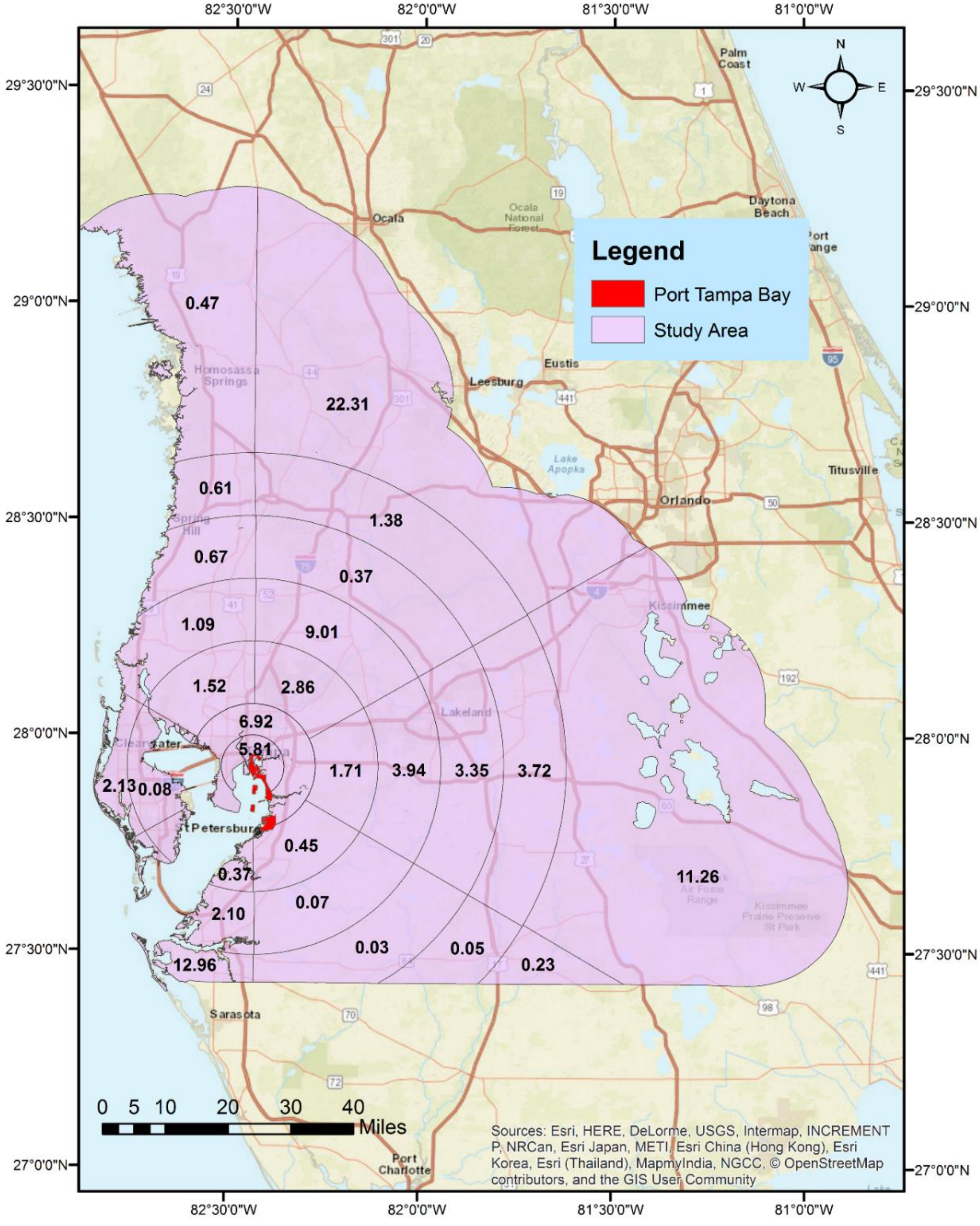

Figure 5.7: Spatial distribution of destination locations of trips originating from Port Tampa Bay

Figure 5.8 provides the distribution of routes taken by the trucks departing from Port Tampa Bay. Specifically, the percentage values shown in Figure 5.8 are the proportions of the total number of trips passing through various key network locations around Port Tampa Bay. As evidenced by Figure 5.8, a large proportion of trips made by trucks originating from Port Tampa Bay take I-75 northbound and I-4 eastbound. This high usage of I-75 and I-4 by trucks is also consistent with the high percentage of destination locations of the trips made by trucks originating from Port Tampa Bay falling in Hillsborough and Polk counties.

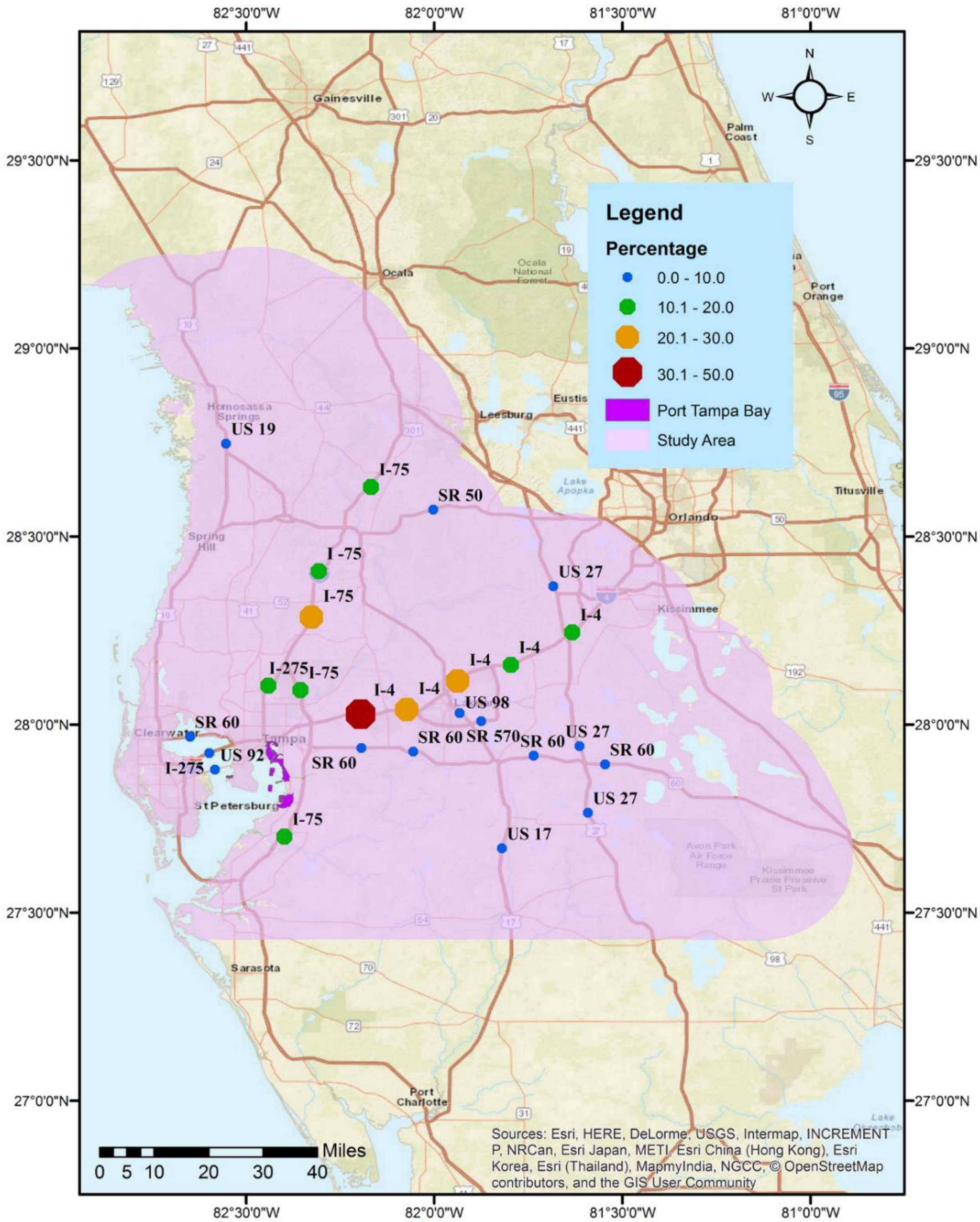

Figure 5.8: Distribution of routes along major corridors for trips originating from Port Tampa Bay

CHAPTER 6: SUMMARY AND CONCLUSIONS

6.1 Summary

This project aims at providing important data useful for understanding and modeling truck route choice behavior in the Tampa Bay region, which, in turn, can be used to improve the truck modeling components of the Tampa Bay Regional Planning Model (TBRPM).

The goal of this task work order was to use truck-GPS data from ATRI to derive data on the routes trucks use to complete their travel between different origins and destinations in the Tampa Bay region. To this end, the following specific tasks were performed in the project:

1. Secure truck-GPS data from ATRI.
2. Convert raw GPS data into database of truck trips.
3. Derive chosen routes for truck trips obtained from tasks above.
4. Analyze routes used by trucks between selected origin and destination locations in the region.

Each of these steps is discussed briefly in the following sections.

Task 1. Secure Truck-GPS Data from ATRI

The research team used ATRI's truck-GPS data for the first 15 days of each month of October 2015, December 2015, April 2016, and June 2016 for the Central Florida counties of Hillsborough, Pinellas, Polk, Pasco, Hernando, and Citrus and 15 miles beyond their geographic boundaries. This resulted in 96,438,457 raw GPS records belonging to 110,475 truck IDs.

Task 2. Convert Raw GPS Data into Database of Truck Trips

The procedure used to convert raw truck-GPS data into trips used the same algorithm used for a previous FDOT study by Pinjari et al. (2016), which was originally derived in another FDOT study by Pinjari et al. (2014). The procedure was applied to all four months (October and December 2015 and April and June 2016) of refined truck-GPS data, comprising 90,184,869 GPS records. This resulted in 1,076,371 truck trips.

This initial set of trips extracted from ATRI data was further refined by removing trips that started and ended in the same TAZ, trips that were less than 2 miles, and trips with a ratio between direct OD distance and trip length less than 0.7. After implementing these criteria, a refined and validated set of 676,902 trips was retained for subsequent use.

Task 3. Derive Chosen Routes from Refined, Validated Truck Trips

The procedure for deriving routes using GPS data consisted of three broad steps, namely 1) map-matching data preparation, 2) map-matching, and 3) route generation. A modified version of the procedure developed by Kamali et al. (2016) was used to derive the routes, which were generated for 243,438 trips. The routes derived for each trip includes all network links traversed by trucks between the origins and destinations of trip. Validation of 300 randomly-selected trips indicated a high level of accuracy of the derived routes. An algorithm was also developed to remove all routes with loops of more than one mile to avoid routes with unrealistic loops or detours. The final product of the project was provided in the form of a GIS shape file of the generated routes for 233,329 trips along with trip-level information such as trip length, trip time, and origin/destination TAZs. Another product of the project was in form of a comma-separated file (.csv) containing unique OBJECT IDs of each link traversed along the generated routes for each trip. Appendix A describes the GIS shapefile of routes and Appendix B describes the .csv file.

Task 4. Analyze Routes Used by Trucks between Selected Origin and Destination Locations In the Region

After discussions with FDOT District 7, analysis of routes used by trucks departing from and arriving at Port Tampa Bay was conducted. Of 233,329 final trips, there were 3,755 trips departing from Port Tampa Bay and 3,774 trips arriving at Port Tampa Bay. A descriptive analysis of these trips such as route length distribution, trip duration distribution, average travel speed distribution, time of day distribution, destination location spatial distribution, and network distribution was conducted to understand the travel behavior of trucks traveling into and out of Port Tampa Bay.

6.2 Opportunities for Future Research

The data developed in this project offer significant opportunities to understand truck route choice behavior in urban regions. Although the current project focused on developing these data, future work on analyzing the determinants of truck route choice patterns can help forecast aggregate-level network performance for medium- to-long-term decisions such as designation of truck routes, addition of new truck corridors, and by-pass routes. Measuring and monitoring the travel paths (or routes) trucks take and understanding why they do so can help design short-term truck routing policies aimed at congestion mitigation, improving reliability, and maintaining a state of good repair.

It is worth noting that a parallel project sponsored by USDOT being led by this project's Principal Investigator is using the route choice data derived from this project to develop truck route choice models for the Tampa Bay region. Therefore, the data products derived in this project are already bearing fruit. Such truck route choice models potentially can be used to improve the truck modeling components of the TBRPM.

REFERENCES

Kamali, M., Ermagun, A., Viswanathan, K., and Pinjari, A. R. (2016). Deriving Truck Route Choice from Large GPS Data Streams. Transportation Research Record, 2563: 62-70.

Pinjari, A. R., Zanjani, A. B., Thakur, A., Irmania, A. N., Kamali, M., Short, J., and Park, L. (2014). Using Truck Fleet Data in Combination with Other Data Sources for Freight Modeling and Planning. Florida Department of Transportation Research Report BDK84-977-20). http://www.fdot.gov/research/Completed_Proj/Summary_PL/FDOT-BDK84-977-20-rpt.pdf.

Pinjari, A. R., Kamali, M., Luong, T., and Ozkul, S. (2016). GPS Data for Truck-Route Choice Analysis of Port Everglades Petroleum Commodity Flow. Florida Department of Transportation Research Report BDV25-977-17. http://www.fdot.gov/research/Completed_Proj/Summary_PL/FDOT-BDV25-977-17-rpt.pdf.

Yang, J. S., Kang, S. P., and 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.

APPENDIX A: DESCRIPTION OF GIS SHAPEFILE

The GIS shape file of the derived truck routes consists of 233,329 features, which is equal to the number of final derived routes. Each feature in the shape file represents the route corresponding to a trip. Table A1 presents all attributes present in the shape file's attribute table. The definition of each of these attributes is provided below.

Table A-1: Attributes of Route Shape File

No.	Attribute Name in Shape File	Description
1	T_Num	Trip number
2	Tr_ID	Truck ID number
3	Data_Type	Month_R or Month_S
4	T_Start_T	Trip start date and time
5	Org_TAZ	Origin TAZ
6	Org_Cnty	Origin county
7	T_End_T	Trip end date and time
8	Dest_TAZ	Destination TAZ
9	Dest_Cnty	Destination county
10	R_Lngth	Route length (mi)
11	T_Time	Trip time (min)
12	FF_TT	Free flow travel time (min)
13	Or_Dw_Tm	Origin dwell time (min)
14	Dest_Dw_Tm	Destination dwell time (min)
15	Stop_Dw_Tm	Stop dwell time (min)
16	GPS_num	Number of map-matched GPS points for each trip
17	O_GPS_num	Number of original GPS points for each trip
18	Links_num	Number of links in the link list for each trip

Trip number (T_Num): Unique ID number for final trips. To visualize a trip belonging to a trip number, select trip based on this particular attribute.

Truck ID number (Tr_ID): Unique ID number assigned to each individual truck. To visualize routes of all trips belonging to a specific truck ID, select all trip numbers belonging to that truck ID and view in a GIS layer.

Type of data (Data_Type): Combination of recorded month and type of dataset. Four months of data are available in this project (October and December 2015 and April and June 2016) for two types of data format – R (rotating) and S (static).

Trip start date and time (T_Start_T): Start date and time of the trip in following format: YYYY-MM-DD HH:MM:SS (e.g., 2015-12-15 00:00:00).

Origin TAZ (Org_TAZ): Traffic Analysis Zone (TAZ) number in which the trip started; same as that followed in FLSWM.

Origin county (Org_Cnty): Name of county in which trip started.

Trip end date and time (T_End _T): Start date and time of trip in following format: YYYY-MM-DD HH:MM:SS (e.g., 2015-12-15 00:10:01).

Destination TAZ (Dest_TAZ): TAZ number in which trip ended; same as that followed in FLSWM.

Destination county (Dest_Cnty): Name of county in which trip ended.

Route length (R_Lngth): Trip length in miles (obtained from actual length of traveled route measured from Navteq network for each trip).

Trip time (T_Time): Trip duration in minutes (obtained from algorithm to convert raw GPS data into trips), including traffic stops of duration less than 5 minutes.

Free flow travel time (FF_TT): Trip travel time in minutes under free-flow conditions.

Origin dwell time (Or_Dw_Tm): Dwell time at origin of trip, in minutes.

Destination dwell time (Dest_Dw_Tm): Dwell time at destination of trip, in minutes.

Stop dwell time (Stop_Dw_Tm): Total stop time during a trip in minutes, including traffic stops of duration less than 5 minutes. It should be noted here that the difference between the trip end time and trip start time is equal to sum of trip time, origin dwell time and stop dwell time. (T_End_T – T_Start_T = T_Time + Or_Dw_Time + Stop_Dw_Tm).

Number of map-matched GPS records (GPS_num): Number of map-matched GPS points for each trip.

Number of original GPS records in each trip (O_GPS_num): Number of original GPS points for each trips.

Number of links in each trip (Links_num): Number of links in a given route; unique OBJECT IDs of links for each route provided in *link-list* (.csv) file.

APPENDIX B: DESCRIPTION OF LINK-LIST FILE

The *link-list* in a comma separated file (.csv) format contains a list of the links traversed by trucks on each trip. The *link-list* was extracted to retain the attributes of the links traveled by a truck on a given trip. These links are of prime importance for developing a route choice model. The link-list file has four attributes namely: Trip Number (T_Num), Truck ID number (Tr_ID), Link ID number (OBJECTID) and Link Sequence Identifier (link_seq). A sample of the *link-list* is provided in the Table B-1. A description of the four attributes is as discussed below:

Table B-1: Sample of link-list File

T_Num	Tr_ID	OBJECTID	link_seq
1	ABCD	423230	1
1	ABCD	423305	2
1	ABCD	423306	3
1	ABCD	428503	4
1	ABCD	1382530	5
2	XYZ	423230	1
2	XYZ	423305	2
2	XYZ	423306	3
2	XYZ	1334690	4
2	XYZ	1334691	5
2	XYZ	1382528	6

- **Trip number (T_Num):** Unique ID number for final set of trips.
- **Truck ID number (Tr_ID):** Anonymized unique ID number of the truck making the trip.
- **Link ID number (OBJECTID):** Unique ID number assigned to each link in Navteq network. All the information related to a link (length, travel time, roadway type etc.) can be retrieved from Navteq network using attribute this attribute.
- **Link sequence (seq):** Sequence of the links for a given trip. For a given trip, this number starts from 1 at the origin link and increases by one unit for each successive link in that trip.

As seen in Table B-1, all the records in the link-list with same T_Num attribute belongs to a trip. The OBJECTID attribute of all those records that have same T_Num attribute gives the link IDs of the links that were traversed in that trip to travel between origin and destination locations. The link_seq attribute represents the order in which these links were traversed during the travel. To visualize the links of a trip in the GIS environment, one needs to select links in the Navteq network with OBJECTIDs same as the one given in the link-list for that trip. This can be easily done using the ‘selection by attribute tool’ present in the ArcGIS environment.