Exploring Targeted Long-Distance Travel Survey Sampling Frame Approaches to Support Travel Survey Collection

by

Mitchell P. Fisher II

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 6, 2022

Keywords: Long-Distance, Travel, Survey, Sampling

Approved by

Jeffrey J. LaMondia, Chair, Associate Professor, Department of Civil and Environmental Engineering
Lisa Aultman-Hall, Professor and Chair, Systems Design Engineering, University of Waterloo
Laurence Rilett, Director, Auburn University Transportation Research Institute
Huaguo Zhou, Professor, Department of Civil and Environmental Engineering

Abstract

Long-distance travel continues to grow in economic, environmental, and infrastructural importance. To inform policy, infrastructural funding, and travel demand forecasting, detailed travel survey findings are needed at the national-level. However, the costs of running a full-scale national long-distance travel survey, both fiscally and temporally, has limited recent attempts. As such, survey data users have had to use older national survey data and then optimize their findings based on more recent, smaller-scale state travel surveys. These smaller-scale surveys have sample framing limitations, but while these limitations may be allowable given the immediate needs and scope associated with its original purpose, their utility for extrapolation and aggregation for broader use compared to the scope of national-level annual panel surveys is still fairly unknown.

This dissertation aims to identify how targeted sampling frame approaches can be used by national long-distance travel surveys. Having a full-sized, population-proportioned sample is always a tenant of good survey design, but the associated costs, especially for more niche subjects, can lead to an unwillingness to fully commit resources to survey deployment. This can lead to patchwork solutions to reduce costs such as with asking respondents for their long-distance travel habits over a very small timeframe or using regional survey findings to update national travel models. While these patchwork solutions offer potentially valid solutions, the actual impacts on data validity are practically unknown. This dissertation aims to fully explore these aspects, by not only exploring how targeted sampling frames might affect long-distance travel survey data accuracy, but also the best approaches for creating a targeted sample frame for the purposes of capturing nation-wide US long-distance travel.

Results suggest that the long-distance travel survey sampling frame can be targeted to reduce both fiscal and temporal costs considering seasonal variability trends, using targeted sociodemographic sampling, or using targeted geo-economic sampling. At the same time, these reductions can still provide statistically viable samples for sociodemographic groupings, travel volumes, mode splits, and purpose splits comparative to full-scale national surveys like the 1995 American Travel Survey, 2001 National Household Travel Survey, and 2013 Longitudinal Survey of Overnight Travel.

Acknowledgments

I am blessed to have an amazing and supportive family and network of friends who were willing to put up with the countless long nights, stressful moments, and venting conversations while finishing my degree. Without their support I don't think I could've made it through this with my sanity intact. To my parents, Chip and Teri, you made me who I am today and instilled hard work, dedication, integrity, and pride in everything I do. To my siblings, Tucker, Emme, and Sophia, I couldn't ask for a better group to grow up with. To Emily, you joined me on the tail-end of this journey, but your support kept me focused and motivated to seeing this finally through.

To my committee, I thank you all for the support, feedback, and encouragement throughout this process. Dr. Jeff LaMondia, your guidance and mentorship over these past six years has been nothing short of amazing. You encourage me to approach research the best way I can while being innovative and loving what I do. Dr. Lisa Aultman-Hall, thank you for all you have done, and it has been a blast working together over the last few years. Dr. Laurence Rilett, you helped give me a new perspective on how this work fits into the grander scheme of transportation, as well as invaluable feedback and suggestions to ensure this is the best it can be. Dr. Huaguo Zhou, it has been an absolute pleasure to have you as a teacher and supporter through my years at Auburn. Thank you all for everything!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADT	Average Daily Traffic
ATS	American Travel Survey
BLS	Bureau of Labor Statistics
BTS	Bureau of Transportation Statistics
CDR	Call Data Record
DOT	Department of Transportation
FHWA	Federal Highway Administration
GDP	Gross Domestic Product
GIS	Geographic Information System
GPS	Global Positioning System
HBTE	Home-Based Trip Ends
LBS	Location-Based Services
LRTP	Long Range Transportation Plan
LSOT	Longitudinal Survey of Overnight Travel
MPO	Metropolitan Planning Organization
NCHRP	National Cooperative Highway Research Program
NHTS	National Household Travel Survey
OD	Origin-Destination
ORNL	Oak Ridge National Laboratory
TRB	Transportation Research Board
VMT	Vehicle Miles Traveled

Chapter 1: Introduction

Long distance travel, which characterizes non-routine and out of town travel not typically captured in daily travel surveys, has grown exponentially over the past decade. This dramatic increase in total miles driven, flown and on rails is fueled by the continued growth of megaregions [1], increased dispersity of social networks [2], continued variability in work and travel patterns [3], and greater affordability of air travel [4]. This travel is an overall estimated 30 percent of passenger miles in recent decades [5] accounting for an estimated 1.9 billion leisure and 464.4 million work related trips in 2019 [6]. In fact, air passenger-milage has increased 44 percent from 2000 to 2018, while highway passenger-milage increasing roughly 27 percent within that same timeframe [7]. If previous trends persist, these travel volumes will continue to rise [8, 9].

As such, the economic, environmental, and infrastructural impacts of long-distance trips are becoming a forefront issue for transportation planners and decision-makers across the country. In 2019, the estimated economic output from long-distance travel was $2.6 trillion supporting around 15.8 million US jobs [6]. An increase of $500 billion in economic output and 700,000 US jobs since 2015 [10]. This impact has also been demonstrated to be independent of U.S. GDP growth [11]; however, it is still susceptible to recessions [12-14] and extreme times of turmoil (i.e. the 9/11 terrorist attacks [15-17] or COVID-19 pandemic [18-20]), albeit to a lesser degree than other economic sectors [13, 21]. Individuals are still traveling, and with increased travel comes an elevated concern of long-distance travel's environmental impacts. These range from the obvious, such as increased travel-related pollution [22-25], to the more specific—as seen with tourism impacts on sensitive wildlife areas [26, 27] or the concerns of indirect effects on seasonal tourist locations [28-30]. Given the scale of its impacts, it would be considered imperative that the factors driving its demand be well documented and studied, but this is regrettably not the case.

While the impacts of long-distance travel are well studied, collection of annual long-distance travel data to support short- and longer-term modeling, forecasting, and planning has fallen behind. Daily US travel behavior is currently captured by the National Household Travel Survey (NHTS), which can capture a respondent's long-distance trip if it occurs on their travel day. However, given the seasonality and sparseness of long-distance travel, a panel-type survey has been necessary to truly capture long-distance travel behavior. This is even referenced in the 2001 NHTS User Guide, the last true national-level long-distance travel survey, which warns practitioners of the fallibility of scaling its short-term panel trip volumes (28-day travel period) to annual long-distance volumes [31]. Additionally, administrating these surveys can be quite cost prohibitive (for example, the 1995 American Travel Survey (ATS) roughly cost $466 per household in 2020 USD [32]). As such, researchers and planners have had to rely on either the "outdated" [33] 1995 ATS or smaller scoped annual travel surveys for their modeling needs; even weighing more recent trends and volumes back to ATS findings. While both options maybe valid, there is a need for newer representative travel survey data.

The importance for a new national long-distance survey expands beyond national modeling implications. Statewide modeling efforts rely on quality travel data to forecast trip volumes. As such, states rely on either national surveys (such as the NHTS series) or state/MPO run travel surveys which both mainly capture urban travel patterns. While urban travel constitutes a large share of the total number of trips, it is the rural and long-distance travelers which proportionately effect VMT the most. NCHRP Report 735 [34] noted that while three-quarters of vehicle trips in the 2009 NHTS where less than 10 miles in length, they accounted for less than one-third of all VMT. In comparison, trips of over 100-miles accounted for less than one percent of all vehicle trips, but 15.5 percent of all household-based vehicle miles [34]. This impact on VMT from such a small share of trips makes understanding the nature and trends of long-distance travel a priority for travel demand forecasting. With changing travel patterns, demographic shifts, and economic conditions potentially changing long-distance travel volumes and behavior, having solid long-distance travel data could help states, MPOs, and municipalities better forecast travel demand to prevent unforeseen congestion and infrastructure maintenance costs.

There have been efforts to alleviate these data issues by utilizing passive data sources, but they still need to be supported by traditional travel surveys. These sources—including smartphone-based applications, GPS, Location-Based Services (LBS), Bluetooth, and Call Data Record (CDR)—can provide accurate, real-time origin-destination counts, and have been shown to be able to approximate trip modes and purposes [35-42]. However, they cannot provide detailed sociodemographic or trip information such as income, age, number of household vehicles, or travel party size—all important variables needed in travel demand forecasting. Traditional surveys provide this information by adding context to these passively collected origin-destination counts, making it vital that regular long-distance travel survey data be available to ensure continued accuracy of merged passive-survey datasets. Though this merger approach only becomes a feasible and cost-effective method for travel demand and behavior analysis if the burden for collecting the supporting travel surveys is reduced.

Therefore, this dissertation aims to identify methods of targeting the sampling frame of national long-distance travel surveys as they relate to data capture quality. Having a full-sized, population-proportioned sample is always a tenant of good survey design, but the associated costs, especially for more niche subjects, can lead to an unwillingness to fully commit resources to survey deployment. This can lead to patchwork solutions to reduce costs such as with asking respondents for their long-distance travel habits over a very small timeframe or using regional survey findings to update national travel models. While these patchwork solutions offer potentially valid solutions, the actual impacts on data validity are practically unknown. This dissertation aims to fully explore this fact, by not only exploring how limited sampling frames might affect long-distance travel survey data accuracy, but also the best approaches for creating a limited sample frame for the purposes of capturing US long-distance travel.

1.1 Important Terminology

Before further exploring this topic, it is best to set clear definitions for some terms used in this dissertation—notably defining long-distance trips, representativeness, sampling frame, accuracy, and effectiveness.

- Long-Distance Trip
    - Defined for this dissertation as any passenger trip that is at least 50-miles one-way *and* has an overnight component. This was chosen to both a) prevent the capture of daily commute and errand trips (trip types that could arguably be defined as "routine travel" for those individuals); and b) create a baseline definition that works for the three surveys utilized in this dissertation. These restrictions allow for a sole focus on those non-routine trips that traditionally cost more to adequately capture.
- Representativeness
    - The concept of a sample adequately replicating a population according to whatever characteristics or quality is under study [43]. For this dissertation,

representativeness is defined as how well a sampling frame captures the demographic distributions and travel behavior of the US population as compared to the "ground truth" set forth by national long-distance travel surveys.

- Sampling Frame
    - A list, or the source material, from which a sample will be drawn from. It should be representative of the population [44]. For this dissertation, the sampling frame(s) considers long-distance temporal, trip, and sociodemographic characteristic distribution approaches.
- Accuracy
    - Defined by the Organization for Economic Co-operation and Development as "the degree to which [information] correctly estimate[s] or describe[s] the [phenomena it] was designed to measure" [45]. For this dissertation, accuracy is contextualized by how well the sampling frames proposed in this dissertation reflect the "ground truth" set forth by national, full-scaled national long-distance travel surveys.
- Effectiveness
    - The extent to which the activity's stated objectives have been met [45]. For this dissertation, this is a measure of if the stated hypotheses are met in both a statistical and practical sense.

1.2 Research Question and Hypotheses

With long-distance increasing in importance, the need to effectively identify new trends and changing behavior is highly relevant to the field—particularly within the modeling community. Traditionally this is accomplished by developing models and forecasts based on travel survey data, but with the heavy fiscal and temporal costs associated with a nationally scaled, annual longitudinal survey, there has not been a recent national survey to reference. While there have been national and state surveys that have measured long-distance travel, their scope has been either regionally-focused or limited as “snapshot”-style travel capture, which, while cost effective, has yet to be fully explored in terms of data capture breadth. Can these smaller-scale surveys adequately capture annual long-distance travel trends? As such this dissertation aims to answer the following question: can a national long-distance travel survey sampling frame be targeted without a significant loss in data quality compared to a full-scaled surveys?

To appropriately answer this question, this dissertation will consider how separately targeting three major framing approaches—temporally, demographically, and geographically—captures the depth of long-distance travel demonstrated by three nationally-scaled long-distance travel surveys. By utilizing these nationally scaled, and mostly annual panel, surveys as test beds, the effects of targeted sample framing can be measured against national, proportional samples as opposed to comparing data from the more common smaller-scaled long-distance travel surveys. To fully clarify, these approaches are arranged as the following hypotheses:

- I hypothesize that there are unique groups of travel days that have statistically similar trip rate patterns for various populations within each travel day group but are also statistically different between travel day groups.
- I hypothesize that there are unique groups travelers that have statistically similar trip rate, mode choice, and trip purpose patterns within each population group but are also statistically different between population groups.

- I hypothesize that there are unique socio-geographic groups that have statistically similar trip rate patterns within each socio-geographic group but are also statistically different between socio-geographic groups.

Each hypothesis will be evaluated by this dissertation for its validity within its respected tested travel survey. While validation against an additional dataset would be ideal, the already lacking availability of annual long-distance travel survey datasets and, more importantly, the lack of available detailed trip records prevents true validation testing. As such, this dissertation aims to serve as an exploration into what could possibly work to target long-distance travel survey sampling frames under future iterations.

1.3 Dissertation Objectives

The main goal of this dissertation is to determine how targeting the sampling frame of long-distance travel surveys affects travel behavior capture. This is accomplished by testing the effectiveness of reducing the sampling frame temporally, socio-demographically, and socio-geographically. Results of this dissertation will aid in the discussion of how various sampling framing options may help reduce the overall costs of effective long-distance travel surveying.

First, determining how targeting the data capture time from an annual scope to a smaller timeframe effects travel behavior will be explored. Long-distance travel is irregular by nature and capturing a household's travel over a year offers the best insight into their long-distance travel behavior. However, yearlong panel surveys are both costly and run the risk of high dropout rates among participants. As such, many travel surveys now capture long-distance travel behavior by recording several weeks-worth of travel; potentially reducing the depth of available data. This chapter aims to determine a) if there are differences in the day-to-day long-distance travel patterns across unique sociodemographic groups across a year, and b) how long-distance travel days can be clustered such that the long-distance travel volumes (reflecting the range of sociodemographic groups) are similar for days within each group *and* the travel volumes are different between groups. These tasks are accomplished using the 2013 Longitudinal Survey of Overnight Travel (LSOT) dataset given its inclusion of specific tour beginning and end dates. Travel activity for this objective is defined as the volume of home-based trip ends (HBTEs) occurring on a single day, i.e., all long-distance tours either beginning *or* ending on that day of the year. Each day is characterized by HBTEs for six sociodemographic groups relating to their respective mode and purpose splits. Findings from this task should not only highlight the need for considering long-distance trip seasonal variability in setting the travel survey capture timeframe, but also help lay a path forward for exploring how to better capture long-distance travel on a non-annual survey scale.

The second objective focuses on general survey representativeness, consistency, and travel equity. This is accomplished through defining sociodemographic groupings and minimum sampling numbers needed to repeat the travel volumes seen in the 1995 ATS, 2001 NHTS, and 2013 LSOT. This task will a) collapse vital socio-demographic information—simplifying and streamlining comparison methodology between surveys; b) identify key long-distance traveling socio-demographic groupings; and c) suggest group sampling rates to best reflect larger survey trends and account for non-travelers. As more focus is placed into integrating passive travel data in modeling/surveys, the ability to aggregate detailed socio-demographic information for privacy concerns as well as reduce survey overhead is a forefront motivation. This is completed by creating a universal socio-demographic categorization and outlier detection methodology, ANOVA analysis for grouping validations, Thiel indices to identify group equality, and traditional sample size calculations. Not only does this task further reduce survey fiscal overhead, but these results are also carried through and utilized throughout the remaining dissertation objectives.

The third objective analyzes long-distance travel surveys for traveler trend consistencies relative to socio-geographic classifications. This is done for trip volumes, mode choice distributions, and trip purpose distributions. Socio-geographic groupings will be based on Oak Ridge National Laboratory's work on scaling the 2001 NHTS [46]. This involves classifying all US counties based on five characteristic influencers of long-distance travel: income, age, locale, airport access, and Amtrak access. Analysis for this objective will include both ANOVA and two-sample testing. The goal here is to identify if any groupings stay relatively consistent between geographies in their long-distance travel behavior.

1.4 Dissertation Organization

This dissertation is organized in seven additional chapters after this introduction. First, is a detailed literature review of the importance of long-distance travel survey data, how the data is used, trends in long-distance travel behavior, and known challenges associated with capturing long-distance travel. The next chapter explores the need for consistent long-distance travel survey collection efforts. This includes a review of notable US long-distance travel surveys over the past three decades and a comparison breakdown of travel trends between the three major surveys—the 1995 ATS, 2001 NHTS, and 2013 LSOT—used throughout this dissertation. Next, chapter four explores how long-distance travel surveys can be targeted via seasonal and annual trends. Using the 2013 LSOT dataset, travel activity for each day of the year is analyzed and classified to understand annual travel variability. The next chapter, chapter five, focuses on targeting long-distance travel survey sampling using sociodemographic groupings. This is followed by chapter six which explores the geo-demographic patterns with long-distance travel among US counties with a final goal of creating a targeted sampling approach based on these findings. Finally, chapters seven and eight summarize the findings of this dissertation; offering sampling frame considerations for future long-distance surveying efforts as well as identify general future research topics.

Chapter 2: Literature Review

Transportation planners, modelers, and decisionmakers throughout the United States have documented their need to accurately forecast long-distance travel to support regional/state project prioritization and decision-making. However, they have also documented the challenges of capturing quality long-distance travel survey data to support these forecasts is currently limited. One solution is the use of targeted sampling approaches, which can allow for smaller-scale surveys providing accurate and quality travel behavior data at reduced costs. This literature review explores the needs for forecasting long-distance travel, , known factors influencing long-distance travel behavior, and the challenges inherent to travel surveying.

This literature review is organized into six distinct sections. In the first section, a review of the rising trends in US long-distance travel and the need to accurately forecast it will be covered. This includes its overall growth, economic impacts, and environmental impacts to fully illustrate both the positives and negatives this type of travel brings. The next section discusses the role and scope long-distance travel demand modeling has in the United States at the national, state, and regional levels. Next, the factors influencing long-distance travel decision-making are presented. This section covers the sociodemographic aspects of travel, destination and purpose, and mode choice considerations in travel decision-making. Next, an overview of select long-distance travel demand models is given. Here, the role of good travel survey data in modeling is presented focusing on what type of data is needed. The next section covers the challenges inherent with capturing long-distance travel behavior with surveys as well as how passive data techniques can help alleviate these challenges. The chapter concludes with a summary of findings, particularly highlighting what factors are essential to streamline long-distance survey sampling practices.

2.1 The Need for Accurate Forecasting of Long-Distance Travel

Demand for long-distance travel has grown continuously over the past few decades. The increased affordability of air travel [4], growth of megaregions [1], and evolving characteristics of work and travel patterns [3] has resulted in a steady rise in long-distance travel volumes. In fact, Bureau of Transportation Statistics (BTS) estimates show increases of 44 and 27 percent for air and highway passenger-milage between 2000 and 2018 [7]. This has translated to an estimated 1.9 billion leisure and 464.4 million work related trips in 2019 alone [6]. While it can still be impacted by economic recessions [12-14] and instances of sudden major change, such as with COVID-19 [18-20], travel volumes continue to steadily increase. This continued rise in long-distance travel volumes has not gone unnoticed by decisionmakers or advocates, as the impact of this travel has widespread positive and negative effects. These impacts—economic, land-use and congestion, and environmental—call for the forecasting of long-distance travel volumes at the national, state, and regional levels to best prepare and inform shareholders to changes in trends.

Perhaps the most attractive impacts of rising long-distance travel volumes are the economic benefits. The tourism industry is a major economic sector in the US supporting countless jobs, local/regional economies, and partner industries. In 2019, it was estimated long-distance travel attributed to $2.6 trillion in economic output, supporting around 15.8 million US jobs [6]. While the COVID-19 pandemic has dampened these impacts over the past few years, experts still believe the tourism industry will bounce back, albeit with potential changes in mode and destination choices [47-50]. As such, areas traditionally seen as gateway communities to larger attractions (commonly seen outside landmark areas such as national parks, amusement parks, etc.) could see changes, either positive or negative, to their local economies if trends shift [51-54].

Some of these changes include land-use impacts and roadway congestion. As communities evolve, what was once farmland and wilderness may slowly transform into new residential housing, commercial ventures, or recreational destinations. These land-use changes can have drastic effects on surrounding communities and demand for long-distance travel. For example, in the 1960s, the construction of Walt Disney World from swamplands to one of the world's forefront vacation destinations had a massive impact on the neighboring city of Orlando, Florida. Over the next few decades, this small agricultural city experienced hypergrowth transforming into not only a major tourist destination, but also one of the nation's fastest-growing urban regions [55]. While this is an extreme example, the interconnections of land-use and long-distance travel volumes still affect transportation infrastructure one in the same, particularly when observing rural development. NCHRP web-only document 211's study of rural communities and land-use changes found that, in general, areas that saw new household and job growth but already had a high amount of existing development, a mixture of industries, and good access to regional commercial developments saw little to no increase in daily VMT. However, a large increase in daily VMT for the surrounding region was observed when job and household growth was observed together in a relatively underdeveloped and isolated area [56]. Basically, rapid, unmitigated growth in underdeveloped rural areas, while economically advantageous and lucrative, can greatly increase regional VMT and long-distance travel volumes. These rises in travel volumes can in turn stimulate congested conditions [33, 34, 56], increase transportation entropy of regional roadway networks [57], and accelerate roadway rehabilitation timelines [33, 34, 56]; affecting regional economics both directly and indirectly.

While economic impacts can be seen as a major driving force for forecasting long-distance travel volumes, the environmental impacts, and associated environmental influencing factors, attributed to long-distance travel have become a recent focus. With the increase in long-distance travel, emissions associated with travel modes have also risen. In fact, air passenger-milage has increased 44 percent from 2000 to 2018, while highway passenger-milage increasing roughly 27 percent within that same timeframe [7]. This has been attributed as nearly 29 percent of US greenhouse gas emissions in 2019 alone, with passenger cars (40.5 percent) and commercial aircraft (7.2 percent) making up nearly half of emissions [58]. These impacts have not gone unnoticed by the public. As concern for climate change increases among US adults [59] and the fallout of the COVID-19 pandemic has accelerated activism efforts [60], awareness of the impacts long-distance travel has on the environment may be at an all-time high. In Sweden, an entire movement, called flygskam ("flight shaming"), has gained traction which encourages the public shaming of individuals who partake in air travel as a method of eco-activism [61]. This movement was even credited with a nine percent decrease in domestic Swedish air travel in 2019 [62]. While this can be seen as a net positive for reducing air travel volumes, the question then falls onto if these effected trips just change modes and indirectly impact the environment in another way. For example, it is theorized that the widespread adoption of automated vehicles will not only pull travelers away from the personal vehicle mode share, but also passenger air travel, particularly for shorter distance trips [63, 64]. While this could potentially reduce short-haul flight frequencies (a known significant contributor of emissions [65]), the associated increased VMT on highways could increase congestion and accelerate pavement rehabilitation needs—potentially offsetting the emissions saved from reduced flying. As such, forecasting how these scenarios may impact mode shifts needs to be explored to ensure decisionmakers are informed on changes to the transportation network.

Indirect environmental impacts caused by long-distance travel could also contribute to changes in travel patterns. For example, Sivaraman's study of US vacation travel found a correlation between weather conditions and coastline length with a household's destination choice [66]. Considering the seasonality dependence of some types of vacation travel and how individuals are dissuaded to complete long-distance travel due to extreme weather [67], impacts of climate change could result in changing destination patterns or associated volumes. Winter recreational areas in particular are concerned about these possible impacts due to climate change. With unpredictable winter snow patterns occurring more commonly, areas that rely on a predictable season could experience unsustainable economic lulls [28-30]. These changes have also been a concern for other recreational vectors such as hiking, boating, fishing, and hunting as changes in temperatures, precipitation, and water levels influence visitor volumes [68].

Overall, concerns for the impacts of climate change on weather patterns and the destination environment could have lasting effects on long-distance travel behavior and associated economic benefits. Unlike localized economic, political, or environmental changes that may change how one may travel around their city; the impacts of long-distance travel are vast and can affect individuals on the other side of the globe. Arguably, long-distance travel should be viewed under the same lens as the global economy—a policy change in Europe regarding air travel emissions may impact tourism in the US. While this may be an extreme example, the point still holds when looking at interstate commerce. For example, some well-known spring break locations in recent years, notably Panama City Beach, Florida, have implemented ordinances to combat the large unruly crowds. In response, spring breakers who may have visited this location in the past may now decide to visit another beach community. This cause and effect can extend past the local level impacting not only the original area but now also surrounding communities or maybe even other states. Situations like this illustrate the importance of understanding long-distance travel behavior. Its impacts extend beyond the traditional boundaries set by city limits, state lines, or country borders, and as such, effectively forecasting volumes should be a decisionmaker's goal.

Forecasting long-distance travel is our best technique for anticipating and proactively planning for economic, land-use, congestion, and environmental concerns. However, these travel demand forecasts can only happen if quality travel behavior data is available. This is where the role of a regular long-distance travel survey becomes important, but due to the costs associated with the traditional sampling approach, there has not been a dedicated national long-distance travel survey in several years. By exploring targeted sampling techniques for long-distance travel surveys, this dissertation offers an approach to provide more regularly captured travel behavior trends, allowing for better forecasting efforts.

2.2 The Role and Scope of Modeling Long-Distance Travel

Travel demand modeling plays an important role at all levels of planning—from the localized levels of Metropolitan Planning Organizations (MPOs), cities, and counties; to the broader-scales seen with states, regions, and nationally. These models help decisionmakers understand a variety of topics from predicting traffic volumes to understanding transportation policy impacts, as well as inform the long range transportation plans (LRTP) of MPOs and states. Interestingly, in the earliest days of federally mandated U.S. travel planning, long-distance and intercity travel was the *main* focus [69]. Intraurban travel was dominated by public/private transit options while the automobile was seen more as a means for connecting rural areas to urban areas or complementing interurban travel. As such, highway infrastructure focused on ensuring adequate regional connectivity considering population density, manufacturing, and agricultural production. Demand and planning emphasis was mainly placed on not *who* was taking these trips, but *where* and how many. Eventually, with the continued urbanization of the mid-20th century, the focus (and funding) of modeling efforts centered around short-distance urban travel, and the creation of MPOs cemented these smaller-scale modeling concerns. The result of which created vast networks of disaggregated modeling focused on how many trips were made within a planning boundary, but with little regard for the those travelers who crossed this boundary or their origins/destinations. This has also been seen with state-wide modeling efforts. While there is some flow of information between modeling hierarchies, each level of planning is chiefly concerned with its immediate realm. As models address larger areas, they tend to focus on less detail.

In the context of long-distance travel, there is further disconnect between levels with planning bodies either having differing definitions of long-distance travel or being unconcerned with this travel completely [33, 34]. While the arguable consensus definition of long-distance travel among researchers is at least 50-miles one-way, the majority of MPO practitioners (up to 83 percent [70]) define any travel outside their planning boundary to be long-distance. At the state level, this definition changes to be more in line with the traditional 50-mile benchmark, but the habit that travel outside the planning area—in this case state boundary—regardless of distance is labeled long-distance, permeates the idea that the current local modeling focus neglects long-distance travel. This was backed up by Cordero's thesis exploring the state of practice of long-distance and intercity modeling in the US. He found up to 41 percent of MPOs did not consider long-distance travel (including intercity and/or outer planning boundary) to be important in their LRTPs, and there was also a disconnect between the importance of rural and urban long-distance travel with only up to 24 percent of MPOs considering both types to be important [70]. These inconsistencies suggest that planning organizations may see their outer-bound trips as not their primary concern, relying on the next modeling hierarchy to capture these trips.

Smaller scale modeling efforts may also lack detailed long-distance travel forecasting due to data limitations. A topic that comes up in NCHRP Report 735 regarding transferable parameters for statewide long-distance and rural travel modeling is the lack of available data [34]. Rural and long-distance trips are both, by definition, harder to capture due to both irregularity and the effected population. Coupled with geographic limitations, national surveys (like the NHTS) may not provide detailed enough travel information for a county, region, or state to effectively forecast their travel. This has led to transferability studies (such as in NCHRP Report 735) into how this data can still be used in finer geographic areas. However, other studies have found transferability to be more haphazard at best, recommending local survey data is still needed to ensure modeling quality [46].

However, maybe it should not be up to smaller planning areas to fully model long-distance travel. As long-distance travel consists of distances and travel modes that can easily span a vast geographic area, smaller planning organizations may not be equipped with the data, expertise, or funding to model travel coming or passing through their planning area. Instead, long-distance modeling should be a concern for larger planning organizations and travel corridors. Aultman-Hall's whitepaper on "Incorporating Long-Distance Travel into Transportation Planning in the United States" makes this case by explaining how the scope, impact, and domain of long-distance travel demand modeling extends well beyond the abilities of smaller municipalities and organizations, leaving it up to the USDOT as the only agency with the means to take a leadership role in this endeavor [5]. This has been echoed by other long-distance travel proponents calling for a nationally-scoped long-distance travel demand model such as in TRB Special Report 320 [33], NCHRP Report 735 [34], and the Outwater *et. al. Foundational Knowledge to Support a Long-Distance Travel Demand Model Framework* exploratory research report [71].

Since these calls, there has been headway made in this area, such as the NextGen NHTS series, and a clear understanding of the role long-distance travel demand modeling plays on the national scale. On the national level, modeling long-distance travel volumes serves mainly to inform infrastructure investment decisions, evaluate the impacts of transportation policies (such as modal alternatives [namely high-speed rail [72] corridors], fuel and fare price changing impacts, and economic changes) on mobility and the economy, and to understand the impacts of private-sector transportation decisions. Its secondary goal would be to support statewide modeling needs [71]. These tenets clearly tie not only the capabilities of long-distance travel demand modeling in informing an everchanging world, but also provides insights to finer levels of planning—the who, what, where, when, and why of those external trips—supporting local efforts in preparing against changing external factors.

2.3 Factors Influencing Long-Distance Travel Decision-Making

Modelers need detailed travel survey data to ensure valid national travel forecasting. However, the cost of capturing this data at the national-scale has prevented full-scale efforts and has encouraged research into targeting sample framing. To accomplish this, understanding the nuances and patterns of long-distance travel behavior is the key to creating effective targeted sample framing techniques. While the mode share, destination choices, and trip purposes of long-distance may evolve, research suggests the underlying temporal and sociodemographic factors influencing trends are less likely to change. The past few decades have seen multiple works studying the factors influencing long-distance travel, and by using their findings, targeted sampling approaches can be created.

First, the question of *who* takes long-distance trips is answered. Long-distance travel is not an equitable affair. Different sociodemographic groups complete long-distance trips at different rates, modes, and purposes. While a general effort will be made to isolate influential sociodemographic factors, it is important to note that all influential factors affecting long-distance travel making decisions are intertwined. This will be evident throughout this review. Next, the questions of *where*, *when*, and *why* individuals travel is explored. These questions are answered in the context of destinations, seasonality, and purpose—all of which can influence the other. From this section, it should become increasingly clear that capturing long-distance travel on anything less than an annual scale is incredibly difficult. Finally, the question of *how* travelers complete long-distance travel is answered. Here, not only is mode discussed, but also habitual influences. At the conclusion of each subsection this dissertation will define how it will attempt to account for the challenges associated with each question regarding study design.

2.3.1 Who is Traveling? Known Sociodemographic Trends

Long-distance travel has made up an overall estimated 30 percent of passenger miles in recent decades [5]. However, not everyone participates in this type of travel equally. While everyday local trips show a sense of pattern and equal need (e.g., the weekly trip to the grocery store nearly all households make regardless of mode), long-distance travel frequency, mode choice, and purpose have been found to be greatly influenced by sociodemographic factors. Prior research has identified four major influencers: income, age, gender, and household makeup.

Perhaps the greatest influential factor of long-distance travel involves employment and income [73-87]. Higher income individuals tend to complete more long-distance trips [75, 82, 87, 88] as well as less likely to choose personal vehicles [73, 75, 86, 88]. In Cho's work, he related to the concept of value of time as being heavily influential in how income plays a role in long-distance travel volume, as well as mode choice [75]. Using the 2009 NHTS, Cho demonstrated how high-income individuals are more sensitive to travel time over travel cost. In contrast, lower income individuals are more likely to use either public transit or personal vehicles over air as they are more sensitive to costs rather than time [75, 76, 81-83]; or in the case of households making less than $20,000 annually, see a decreased likelihood of taking leisure trips altogether [83]. These relations have been demonstrated internationally [5]. However, as long-distance travel continues to become affordable [89], there is potential for lower income households to increase their travel volumes.

Additionally, age has been found to be an influential factor, however, there is greater discourse among results. Aultman-Hall et al. found age to not be a correlating factor in determining long-distance travel frequency, however this same study *did* find an increase in age did increase the likelihood of completing a greater number of tours in a year [3]. This has also been seen in the 2001 NHTS results, particularly with leisure travel [90]. However, the same study, as have others [91], have found that older individuals do eventually succumb to mobility issues and frequency begins to drop off, specifically those above the age of sixty-five. Regarding mode choice, some works suggest that individuals over the age of 60 years old are less likely to travel by air [73-76, 79, 80], but there is disagreement [77, 81, 82, 92]. It should be noted that when it comes to the study of age's impact on long-distance travel, there is no set definition of measurement. Some studies treat age as a continuous variable, while others bin age into various-sized ranges. These ranges have been observed to be based on generations, decades, or even common age groupings (i.e., children, young adults, middle age, etc.). As such, the reasons for some long-distance relationships and behaviors cannot be fully measured [5]. Other influences, such as income, may further confound age's true role in long-distance travel behavior.

Another known lesser influencer is traveler gender. Most studies have found men as more likely to complete long-distance trips compared to women, which could relate to traditional work habits [77, 82, 91-93]. Mallett's study on women's long-distance travel habits found nearly 80 percent of the difference in trip disparity could be attributed to business trips, despite the at the time major growth in female business travel [93]. It was also proposed women's lower incomes, employment rates, and likelihood of being the primary caregiver of the household attributed to this stark difference in travel between the sexes.

Household lifecycle is another major influencer of long-distance travel volume; particularly, the presence of children under eighteen in the household as well as household size. LaMondia et al. found that larger families or families with small children traveled shorter distances than other groups [94]. Additionally, others have demonstrated that that an increase in household size resulted in a decrease in long-distance travel frequency [82]. However, others, such as the results of the 2015 Michigan Travel Counts III household travel survey, suggest the opposite; with long-distance likelihood increasing with family size (particularly among couples) [95]. The influence on mode choice is also rather prominent with families with children more price-sensitive to air travel costs resulting in a favoritism to traveling by car [81-83, 88, 94, 96, 97]. However, there are other studies demonstrating children do not have a significant effect on mode choice [75].

2.3.2 The Where, When, and Why: Destinations, Seasonality, and Purpose

Long-distance travel is influenced by seasonal career responsibilities, economics, family activities, and social events. [76, 78, 81-83, 88, 98-100]. While annual data collection does capture seasonal highs and lows of travel, as discussed, trends and response rates vary for socioeconomic groups [76, 81, 88, 98]. Coupled with recall biases and burden, annual data may not simply be missing trips but skewed towards specific seasons or socioeconomic groups. For example, states such as Texas, Wisconsin, and Montana demonstrate significant differences in long-distance travel volumes across seasons, while states such as California, Florida, and Michigan maintained high demand concentrations throughout the year [101]. Additionally, the US Travel Association states that holidays (e.g. Christmas, Easter, Thanksgiving, and Independence Day) result in higher travel frequencies along with the months of June, July, and August [99]. Studies that use smaller temporal sample sizes than yearlong would then be at risk of not capturing major spikes in travel behavior. This could result in sample bias towards higher income households who traditionally complete more long-distance trips than lower income households [76, 81, 88]. Additionally, air travel has been found to be income-elastic, especially during different seasons, [76, 78] with lower income households preferring personal vehicles or public transportation over air [76, 81, 100, 102]. Another impact on sample integrity is "immobile" respondents, or respondents who do not complete a trip during the survey timeframe. Focused towards smaller survey timeframes, Madre found that researchers should expect 8 to 12 percent of possible respondents to be immobile on any one-day or weekday survey with a higher rate associated with weekend surveys [100]. These factors each effect long-distance survey results and should be in the forefront of every researcher's mind when designing a long-distance survey component.

While the 1995 ATS provided 12 unique trip purposes, recent studies, such as Bradley *et al.*'s national long-distance tour-base activity model developed for the FHWA, reduced the number of purpose categories to five: business, commuting, leisure, visiting friends and relatives (VFR), and personal business [103]. Other studies have learned the same and have adjusted the number of available trip/tour purposes accordingly (such as in the 2013 LSOT [98]). This has been done as the understanding of long-distance travel has improved and major trip purposes have been charted. While multi-purpose trips occur, generally, long-distance travel consists of mainly general leisure/vacation, VFR, and business trips, in that order. For example, in the 2015 Michigan Travel Counts III household travel survey, 45.2 percent of recorded long-distance trips were categorized as leisure or leisure adjacent [95]. This was followed by VFR at 31.9 percent (for a total of 77.1 percent non-work trip purpose split) and business trips at 12.8 percent (10.1 percent classified as "other"). In the case of the 2013 LSOT, the split between work and non-work trips was a 32.2/67.7 split [3].

Another important connection in measuring long distance travel is that trip purpose and seasonality are interconnected. For example, Figure 1 breaks the 1995 ATS dataset down by business and leisure trip volumes over the standard annual quarter system. Each of the counted households completed at least one long-distance trip and trip counts are limited to only those completed in that particular quarter.

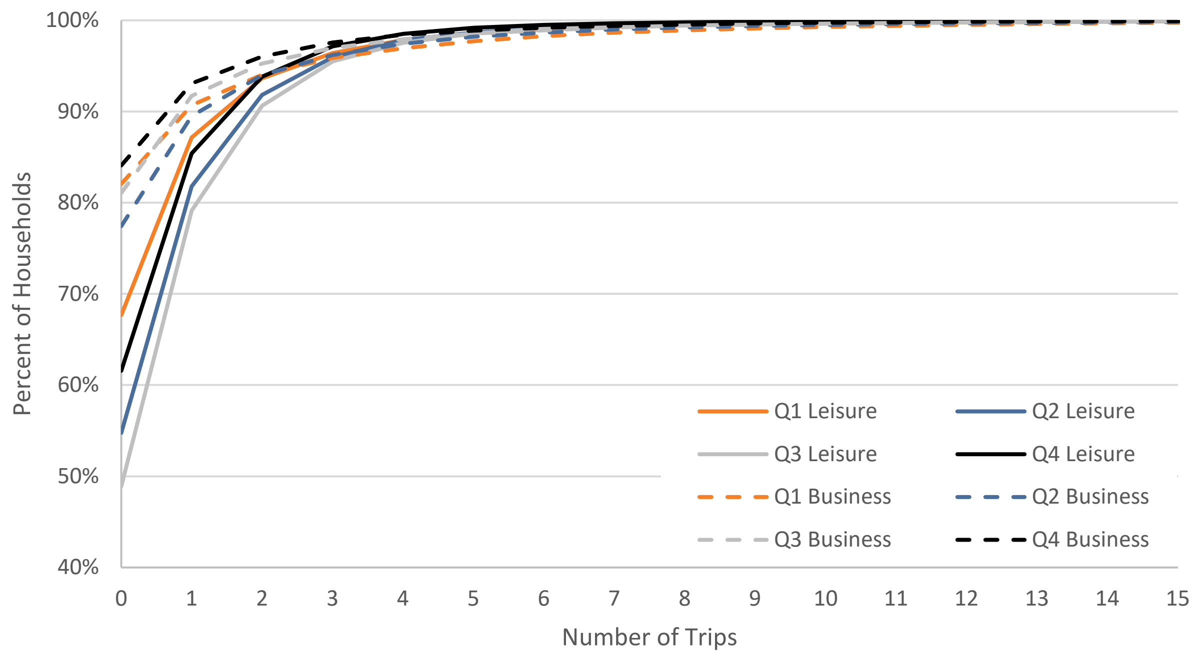


Figure 1: Cumulative Distribution of 1995 ATS Quarterly Trips by Trip Purpose

While visually the cumulative households approach 100 percent roughly around the same number of trips, what should be noted is where the percentage of zero trips begins. It not only varies by trip purpose (with business trips a less likely occurrence), but also by quarter. Thus, illustrating the desire, and need, for annual-scale long-distance travel data. However, with this scale comes increased respondent burden and monetary costs. Therefore, identifying a shorter timeframe that accurately reflects the overwhelming majority of long-distance travel in a year would prove vital to surveying efforts.

2.3.3 How Do People Travel? Mode Choice and Habitual Influences

While the factors influencing demand for long-distance travel have been well covered, another important aspect—in the context of traditional modeling and infrastructure concerns—is *how* individuals complete long-distance travel. Knowing what modes and the paths and habits that effect the choice individuals make can help determine a multitude of things; from which roadways and airports to improve, to how destinations should manage incoming tourist traffic.

This aspect of long-distance travel arguably has the greatest overall impact on economic and policy considerations, and as such cannot be overstated.

Long-distance travel in the U.S. can best be characterized by car and air travel. While other modes do exist—such as rail, bus, and ship—these two modes account for the overwhelming majority of long-distance travel, and as such, account for the majority of research. Mode choice was lightly touched upon with demographic factors, but in the case of air travel, a myriad of factors influence behavior. Air travel has been described as income-elastic [78, 82], but airport access distance, flight choice, and potential layover times have been found to greatly influence air travel, even greater than fare prices [77, 104]. Additionally, work travel has been found to make travelers less sensitive to travel costs since their employer is usually paying. This has resulted in work trips to be more likely to be traveled via air [75, 77, 83-86]. However, this has been disputed depending on the type of work [73]. Another issue of note that could have future impacts on air travel is the continued growth of eco-activism. The concept of "flygskam", Swedish for flight shaming, has gained traction in Europe as a way to publicly shame air travel in the name of eco-activism [105]. This in turn has driven study into the potential impacts eco-activism could have on air travel volumes with some seeing reductions [105-107], and others seeing no major threat [108].

Personal vehicles on the other hand have been found to be more frequently used for leisure trips since they allow travelers greater flexibility at their destination and generally cost less [75, 80, 83]. Trip distance and duration have also been found to be influential, particularly on mode choice. While it is has been noted that the majority of long-distance trips are less than 500 miles one-way [33], the exact distance where travelers prefer a certain mode is still debated. In fact, the measurement of distance is debated in these findings with some researchers measuring distance temporally and other measuring distance in the physical sense [86, 88, 109]. Findings have also been contradictory regarding trip duration with some stating the flexibility of personal vehicles being key for long duration trips [88] and others arguing the speed of air travel allows greater time at the destination [83, 86].

Regarding mode choice, researchers agree that long-distance travel mode choices are highly dependent on an individuals' perceived value of time and the efficiency with which they can reach their destination via each mode. Additionally, researchers agree that expectations for costs and travel times can vary greatly across different demographic groups and for different types of tours. Algers [74] first identified sensitivities to socioeconomic, destination, modal, and access characteristics, and Outwater *et al [96]*, LaMondia *et al* [94], Van Nostrand *et al* [110], and others have continued to study how these factors can significantly influence long-distance mode choices. However, differences exist in terms of how significant these values of time are and how to best express them. While some researchers [92, 96] conclude shorter trips have more value placed on time and cost than longer trips (which are impacted by other more inelastic factors), others believe the opposite [75, 109, 111]. These values of time are complicated by the inclusion of access and egress mode combinations, which is especially important for individuals accessing airports [77, 78, 88, 92, 104, 112]. Regardless, the monetary and temporal properties of a mode have been shown to be significant predictors in mode choice. LaMondia *et al.* found that personal vehicle travel time and air cost travel costs (including access and egress costs) could almost solely predict an individual's mode choice behavior [113]. However, travel diary design does limit the ability to account in detail the actual costs of travel, especially with regards to alternative modes. As such, while incredibly important factors in decision making, the required guest-work needed on the backend of data collection opens the door to potential falsehoods in travel values. More work is needed to best figure a way for researchers to gather this information as accurately as possible in travel surveys without greatly increasing survey burden. With sociodemographic, temporal, and fiscal factors rightfully influencing behavior, another major source of mode choice decision making lies in the characteristics of the tour itself.

The third, and perhaps most important, background topic covers the interplay between tour distance, duration, and destination. While researchers agree that distance influences mode choices, the exact threshold for which different modes are preferred is still debated, with some modeling travel times and some distances [86, 88, 109]. Researchers also have contradictory findings on the preferred mode for longer duration tours, with some pointing out that personal vehicles allow more flexibility [88] and others recognizing that faster air modes allow more time at the destination [83, 86]. Of course, these factors are interrelated, and their relative importance needs to be explored.

While other long-distance modes such as rail and bus should not be ignored in long-distance travel studies, previous research has shown these modes to be highly regional-dependent, particularly in the U.S., limiting mainstream use [114]. Another major issue plaguing both rail and bus are the associated access and egress requirements needed to reach major terminals. However, in recent years, there has been a building novelty associated with Amtrak long-distance travel, especially regarding sleeper car usage. Outside of offering regional commuter rail, particularly in the Northeast corridor, Amtrak also offers some of the only interregional passenger rail service in the U.S. In a recent interview with *Slate*, William J. Flynn, current CEO of Amtrak, discussed how he sees Amtrak evolving under his watch as well as how COVID-19 has affected ridership. While COVID has resulted in a capacity percentage in the low 20s for normal operations, sleeper trains have fared better operating at roughly 35 percent of normal capacity [115]. This coupled with the recent rise of newer private passenger rail services and the promises of high-speed rail akin to European and Asian markets could result in an increase in demand for long-distance rail; however, this is yet to be seen.

Compounding these topics is the recognition that individuals can experience an inherent enjoyment of travel in and of itself [116]. Mohktarian at al. emphasize that the act of travel can provide a positive utility distinct from the destination, which is especially true for leisure activities. Families and other closely related travel parties can use travel time as a bonding experience and treat it as a group shared activity. Researchers that focus on tourism often find that individuals demonstrate a predisposition to choose one mode over another, even if an unchosen mode is more efficient, based on their past experiences and comfort level. This is referred to as "habitual/self-selection" choice patterns [117-119]. Verplanken [120] studied the self-selection of individuals that demonstrated past behaviors that significantly minimized monetary cost and found they also chose personal vehicles due to costs. Aarts found that mode choices were more significantly tied to travel goals [121], although observed mode choices were not statistically repetitious [117].

Considering these findings, this dissertation focuses mainly on the trends associated with personal vehicle and air long-distance travel. This is to offset two major concerns: lack of valid trip data and modal regional dependency. Travel volumes for rail, bus, and other forms of long-distance transportation consist of such a small portion of survey-captured long-distance travel that sampling minimums can become an issue. However, the impact of these lesser modes should not be ignored, as they capture a unique set of travelers. As such, their inclusion in future survey framework is addressed at the end of this dissertation.

2.3.4: A Summary of Notable Influential Factors for Long-Distance Travel Volumes

Long-distance travel represents the non-routine tripmaking of individuals. Like daily travel, travelers are influenced by a multitude of factors—such as mode access, time of day, sociodemographic tendencies, or purposes—when deciding their trips. However, unlike daily travel, these factors are arguably more pronounced as the pure nature of long-distance travel relies on travelers deviating from their normal daily behavior. This results in two major findings: first, the amount of resources needed to fully capture an individual's long-distance travel behavior greatly exceeds the costs of capturing daily travel; and second, the factors influencing long-distance travel behavior are more pronounced than that of daily travel. For the purposes of sampling, these findings can be further summarized into two distinct fields: basic trip characteristics and major sociodemographic influences.

Understanding the when, how, and why long-distance trips occur can help strengthen targeted sampling approaches. Based on literature findings, three major trends can be concluded:

- Long-distance travel is seasonally dependent. Therefore, from a collection standpoint, sampling should consider the temporal scope of collection.
- Long-distance travel mode choice is influenced by trip distance, purpose, costs, destination, and duration. While these trips are accomplished by a multitude of mode types, U.S. long-distance travel mode share is dominated by personal vehicle and air. Therefore, sampling approaches should focus on the factors influencing these mode choice volumes.
- Long-distance travel purposes are influenced not only by sociodemographic factors, but also influence tour destination, mode choice, duration, and other factors. Generally speaking, long-distance travel purpose can be characterized into two major types: leisure

and work travel. Therefore, focused sampling efforts should consider the travel characteristics of these travel purposes when creating targeted sampling approaches.

While knowing the general trip characteristic trends of long-distance travel can help pinpoint the when and why of travel volumes, there is still the question of sample collection equity. Not everyone participates in long-distance travel equally, and understanding the sociodemographic factors and trends associated with long-distance travel volumes ensures targeted sampling approaches adequately capture marginalized travel groups. While literature review findings highlight multiple sociodemographic factors as being influential in long-distance travel volumes and behavior, three factors in particular standout in ensuring travel survey collection equity.

- Long-distance travel volumes are influenced by household income, with higher income associated with increased travel volumes. Therefore, sampling should consider household income as a primary selection factor in collection equity.
- Long-distance travel behavior is influenced by age, particularly in association with travel mode and purpose. As such, older travelers have been associated with greater leisure and vehicle volumes. Therefore, sampling should consider traveler age as a primary selection factor in collection equity.
- Long-distance travel behavior is influenced by a household's lifecycle, particularly when children are present. These households have been linked with decreased long-distance travel volumes, decreased travel distances, and a favoritism for vehicle travel over air. Therefore, sampling should consider a household's presence of children as a primary selection factor in collection equity.

Overall, literature review findings suggest targeted long-distance sampling approaches should chiefly consider these six significant factors. While there are a multitude of other factors influencing long-distance travel behavior, these six factors stand out as being the most consistent findings among previous studies and best help ensure sampling equity, data applicability, and approach legitimacy.

2.4 The Importance of Survey Data in Modeling Long-Distance Travel

Understanding how surveys are utilized by long-distance travel demand modeling is vital. It can help ensure this dissertation's findings contribute to the larger field of travel demand modeling meaningfully to their needs. This section serves as a demonstration of how long-distance travel is modeled with a particular focus on data needs. Particularly, it explores the importance of travel survey data, particularly national-level data, for modeling needs. It is important to recognize that changes in travel surveys have the potential to change the data used in developing travel demand models. As such, the learnings about the travel behavior patterns generated from this dissertation can be later used in future research to understand how changes in sample sizes impact travel demand modeling results.

Long-distance travel demand models can be mainly classified into one of three categories: trip-based, tour-based, or activity-based. Trip-based models, or more commonly known as four-step models, are the traditional modeling structure most frequently used in transportation modeling. It consists of four major steps (trip generation, trip distribution, mode choice, and rout assignment) and characterizes travel based simply on the number of one-way trips between origin and destination pairings. Therefore, each trip an individual takes throughout a given day (ex. home to work, work to store, store to home) is treated as a separate event with no relation to the prior trip. Tour-based models take this a step further and group trips into a logical flow. In this case, the three trips listed prior would be calculated as a single tour. However, both the trip- and tour-based methods fail to account for trip details (such as with vehicle capacity) or the time of day a trip would logically occur. These models operate normally at a macro level, considering an entire area's production/attraction of travel. This is solved in an activity-based model which takes a more micro-simulation approach. These models, which build off tour-based models, treat "travel as a demand derived from the desire for activities" [122] and model travel based more on realistic behaviors and time constraints at the household- or even individual-level. As such, the computation and data needs required for the models is greater than the four-step approach—needing detailed travel survey information and computing higher fidelity travel behavior—which limits its usage to areas with the proper resources and support to fully implement.

The data needs for each of these modeling approaches also varies in complexity, but generally they have similar needs. Chiefly, socioeconomic and geographic data including population and household data, employment, and land-use are some of the most important sources needed [123]. This can be data from census sources, labor statistics, GIS databases and other sources. The data gathered here informs the model of what/where is available to travel to and who can travel. Next, is infrastructure network data such as roads, transit networks, and airports among other mode paths. This data is vital to proper model usage as it helps inform which routes an individual may take to complete their trip/tour/activity. It's importance "cannot be overstated" [71]. The last major data source for modeling is the travel behavior survey. This information informs the model of who, where, why, and how travel is completed for the study area. It is vital for the mode choice step in modeling as well as providing activity-based models when one might travel. These three data types are the keystones to effective travel demand modeling, but are not fully fleshed out for use or application within the long-distance domain. Domestically, long-distance modeling, particularly at the state and MPO levels, usually are based in the four-step modeling approach. This is the case for the Florida [124, 125], Kentucky [126], Tennessee [127], and Wisconsin [71] statewide models. Major exceptions to this are California's HSR activity-based model [128] and Ohio's dedicated long-distance travel model which is tour-based [111]. Nationally, an advanced tour-based approach is used [71]. European-based models focus more on activity-based approaches. National models such as the Swedish [129] and United Kingdom [130] models rely on nest logit-based activity models, while the Norwegian [131] and Portuguese [132], while simpler, still follow activity-based modeling approaches. The reason for these differences varies from case to case, but from the US state perspective, these models are approached using the traditional four-step process to ease integration and data sharing with other planning entities statewide [71]. Summarizing these models gives these general themes to take away from the whole field:

- No Single Definition of Long-Distance Travel
    - While this dissertation defines long-distance trips as 50-miles one-way and overnight, this definition is not universal. The reviewed models showed a variety of long-distance travel definitions: the majority were distance-based [111, 124, 126, 127, 129-132], but some were geographic-based (as interregional travel) [128] or even mode-based [132]. Even when focusing on a distance definition, there was discourse with some defining a trip as 40-miles to others considering over 100-miles. Overall, distance definitions, as well as other stated definitions, seemed arbitrary in nature and based more on the model's purpose.

- Modeling Applications Vary in Purpose and Scope
  - Most reviewed long-distance models were built with a very specific purpose in mind. While some were made to be integrated into daily travel models and thus serving a multitude of applications [124, 126, 127, 131], others focused solely on specific travel corridors or mode feasibility studies, namely HSR [111, 128-130, 132]. Additionally, the ability to use the model for forecasting travel behavior changes due to socio-economic or demographic changes was also not consistent. However, most considered this to some degree.
- Needed Input Data Shares a Theme
  - Each model required multiple datasets including geospatial, mode usage counts, census-type, and travel survey data; to properly function. This data was site-specific and in the cases of using nationally-scaled travel survey data sources, *was normally supplemented with a local-scale travel survey*.

With all the aforementioned models differing in structure, travel definition, and use, they all share one similarity: *the need for long distance travel survey data*. Particularly, a majority of these models depend on nationally scaled travel surveys, with the US-based models favoring the NHTS and ATS—surveys that are potentially outdated. The US national long-distance travel demand model further explains this issue in detail. Here, researchers had to compile multiple national-level long-distance travel surveys, such as the 1995 ATS and 2001 NHTS with more recent state-level long-distance travel surveys (the 2012 California Household Travel Survey (CHTS), 2003 Ohio Household Travel Survey, and 2010 Colorado Front Range Travel Survey) to essentially update travel behavior to more modern times. For example, the 1995 ATS was used to estimate time budget constraints, tour generation, tour scheduling, tour duration, and travel party size, but was not used to estimate mode choice and destination. These last two steps were estimated using the more recent travel surveys since it was understood that travel behavior patterns may have changed over time (particularly with air travel becoming increasingly affordable) forcing the modelers to use the most recent data available, even if it wasn't at the preferred national scale. In fact, recommendations from this report stated that the greatest modeling data limitation was "the lack of survey data on actual long-distance tours and trips" [71].

While there are current efforts to mitigate this lack of national long-distance travel survey data, as seen with the NextGen NHTS, there is still a need to understand exactly how we can efficiently capture long-distance travel behavior without sacrificing its value. This becomes a major concern when considering the evolution of sample sizes of nationally-focused surveys. With the NHTS series, there has been a steady decrease in overall sample size with 63,000 households sampled in the 2001 NHTS to the proposed 7,500 households sampled in the NextGen NHTS. These smaller sampling sizes limit the allowable geographic fidelity the survey data can be used at. For example, the 2017 NHTS is recommended to be used at no less than the Census Region level due to sampling size limitations at the lower levels of geography. With the NextGen NHTS series, its smaller sample size of 7,500 households would suggest a recommended level of geography no less than the national level. These limitations are mitigated for some agencies purchasing survey add-ons that run in conjunction to the national survey, but the fact remains that these national surveys mainly reflect *national* travel behaviors. As such, understanding the intricacies of long-distance travel behavior and potential approaches to efficiently capture it can not only inform national-level surveying efforts, but guide regional and state surveying efforts.

2.5 Challenges with Capturing Long-Distance Data and Improvement Efforts

Households' long distance travel patterns vary significantly across the year. When models of this travel are estimated using data from shorter timeframes, conclusions about travel choices can be conflicting, especially as they relate to household income [75, 94, 96], travel party size [88, 94], and traveler age [74, 77, 78, 92, 133]. However, there are three interrelated challenges that influence the ability to collect this desired annual data: respondent burden, recall bias, and seasonal/socioeconomic variability.

- Respondent Burden
  - Collecting data from a single participant over the course of an entire year can lead to perceived burden and dropout. While survey design and scope fundamentally affect burden [134], research shows that the length of a single survey and duration of time over which one is asked to complete surveys significantly impacts willingness to complete the survey [98, 135, 136]. In particular, researchers conducting the LSOT found a retention rate of only 51.5 percent of participants at the end of the yearlong study period [98].
- Recall Bias
  - Annual data may be collected at many different rates, from daily to monthly to at the end of the year. "Memory decay, failure to understand or to follow survey instructions, unwillingness to report full details of travel, and simple carelessness have all contributed to the incomplete collection of travel data in self-reporting surveys" [137]. As the time between surveys and the retrospective time increases, recall bias, in the form of missing details or even complete trips increases as well [101, 138]. It has also been found that respondents—when

asked for a numerical answer such as how many trips did you take this year? — would infer their response based on partial information and memory recall [138]. This could result in either underrepresentation or overrepresentation of trips, potentially skewing the validity of responses.

- Seasonal Variability
  - Since a large percentage of long-distance travel can be described as leisure travel, travel volumes have been found to be seasonally dependent. Holidays such as Christmas, Thanksgiving, and Independence Day; and the summer months have been shown to have much higher frequencies of travel [99]. Additionally, some destinations—including Texas, Wisconsin, and Montana—demonstrate significant differences in long-distance travel volumes across seasons [101]. Surveys that use smaller temporal sample sizes may be at risk of not capturing representative data due to missed volume spikes or the capture of only periods of high travel.
- Socioeconomic Variability
  - Higher income households have been found to complete more long-distance trips and are more likely to choose air travel than lower income households [76, 81, 88]. Since an estimated 8 to 12 percent of respondents may not travel on any given day [100], smaller scoped surveys risk not fully capturing long-distance travel behavior, particularly for harder to reach households and rare travelers.

While recall bias and respondent burden are inherent to most surveys by nature, seasonal and socioeconomic variability can be better controlled for through the sampling process. These challenges need to be addressed during the survey design phase, and when it cannot be avoided, listed as a potential caveat for use of the survey data.

2.5.1 Reducing Burden and Bias Through Passive Data

Recently, technological innovations and the ubiquity smartphones has offered a new approach that can help alleviate recall bias and respondent burden: passive data collection. Coupled with the monetary and temporal costs of conducting annual surveys, the ability to passively collect data with accuracy and in conjunction with occasional active surveys for updating purposes is enticing. However, with any passive data collection, the data source needs to be of the utmost quality and representative of the population. The Pew Research Center, a nonpartisan "fact tank" states cellphone ownership tends to favor younger age groups and higher household incomes—with smartphone ownership having greater variability [139].

If the data focus is solely reliant on Location Based Services (LBS) data and ignores the less accurate Call Detail Record (CDR) data recorded from cellular tower pings, the ownership rate of smartphones illustrates the theoretical maximum data penetration possibilities of an LBS-based passive data approach. In fact, some companies have already recognized this data source and have used it to model travel pattern metrics; albeit at a much smaller penetration rate. For example, Streetlight Data, Inc. states that their services process analytics on about 65 million devices—roughly 23 percent of the adult US and Canadian population [140]. In addition, regional and daily sample sizes have been found to vary greatly from one percent to up to 35 percent; which further limit the effectiveness of analytics in some locations. Streetlight's suggestion to clients is thus to "encourage clients to evaluate the total [daily trip] sample across the entire study period instead of focusing on per-day penetration rates" [140]. Therefore, the use of LBS-only data needs to further be investigated to determine if there are any major demographic groups under or overrepresented by this source.

Another pioneer in the application of smartphone-based data collection is RSG's rMove™ application. This is a targeted smartphone survey app that not only passively records GPS, accelerometer, and compass data of a user's trip, but also actively collects trip and demographic information from user input [141]. The software can also identify habitual travel and allows for users to modify passive trip information upon review, limiting mistakes. As for data integrity, RSG has performed several case studies coupling traditional active survey methods with their passive collection technology on the same samples [141, 142] with relatively positive results—concluding that rMove™ should be coupled with traditional travel diaries to offer an additional data source, rather than a complete switch. Although strides in data cleaning and data validation techniques could prove otherwise long-term. As for long distance travel applications, a 2017-2018 panel survey performed in Franklin County, Ohio, confirmed proof of concept but identified rMove™ panel members as younger and of higher income than the associated household travel survey [143].

There have been multiple studies investigating the use of passive data collection in inferring tour purpose [35-39], mode choice [36-42], and some sociodemographic information [38, 140, 144-149]. For tour purpose, most studies have focused on overlaying locational user data with land use and point-of-interest data with varying levels of success. Mode choice has also seen strides in recent years with the addition of newer technologies and improved connection capabilities. However, dense urban cores still prove challenging to infer since modal options tend to be more numerous (walking, biking, driving, and transit); particularly with underground transit [41]. In general, mode choice is either inferred via travel speeds, proximity, or traditional

4-step model assignment methods [39]. Sociodemographic inference has proven to be the most interesting with most approaches looking solely at home-based census block distributions [38, 148, 150]. However, cellphone usage has been proven to be capable of predicting age [145], income levels [144, 149], gender [144, 145], and household type [145]. The most interesting case in this field has been Blumenstock's use of CDR data to predict income level in Rwanda [149]. Using a tree-based ensemble model his team was able to better predict changes income distribution than census survey projections, however, it should be noted the penetration of cellphone usage in Rwanda presented a clear bias towards higher income individuals.

Passive travel data collection over a long period of time (e.g., through cellular data to track and individuals' travel habits) can generate different types of burden for participants and researchers. Participants may experience some survey burden if they are asked to verify data or respond to questions based on passive movement counts [151], or their burden could be limited to social network activities they already participate in [152]. Consequently, the majority of the survey burden is transferred to researchers who must compile and clean the raw data. Jazen's study found that due to privacy laws as well as the nature of passive data, trip purpose could not be ascertained so researchers had to approximate based on location [151]. While this could still provide accurate results, actual trip purpose was still unknown. However, trip numbers, distances, durations, and modes proved to be more accurate than traditional surveys but did require more data cleaning/coding from the research team. Cho's team completed a similar study but included social network check-in data along with mobile phone location data, but trip purposes were still based on assumptions [152]. Advances in smartphone survey apps have attempted to merge passive movement data with short questionnaires [153, 154], but concerns about the number of questions, frequency, detail of data collection, and length of time in which respondents keep the app installed are still being studied.

Overall, research has shown there is a solid possibility of using cellphone data for inferring sociodemographic information from passively collected trip information, but travel survey data is still needed to provide a) validation data and b) inform of changes in travel behavior trends. While theoretical smartphone penetration rates account for over 80 percent of the US population, there is still bias towards younger age and higher income individuals. In addition, market research has shown that actual penetration rates are significantly lower, resulting in the questioning in data validation. Previous research has also shown smartphone data can be used to reliably infer tour purpose, mode, and some sociodemographic information; the latter being more reliant on actual phone use or paired census block data. The undertaking of this dissertation thus has the potential to offer a new approach to the field of passive data collection by exploring the further validation and accurate prediction of sociodemographic information solely from observed travel patterns.

2.6 Conclusions

Long-distance travel continues to grow in importance, scope, and impact globally. It has been attributed to economic growth as well as environmental impact and maintains itself as a major contributor to transportation network burden. However, even given its significance, there is still a lag in data collection efforts. The reasons for this among proponents have been explained as limited resources to a need for national leadership, but all agree that a lack of updated long-distance travel survey data has contributed the most.

Reviewing modeling efforts worldwide highlights a multitude of scopes, methods, and purposes for creating long-distance models, but all share a common data need: *relevant long-*

distance travel surveys. For the United States, modelers have had to rely on either outdated long-distance travel surveys such as the 1995 ATS and 2001 NHTS, or smaller scoped, local surveys including the 2013 CHTS. While these smaller scoped surveys help fill in the gaps, they still do not provide enough detail to a) fully capture the entire scope and scale of annual long-distance travel behavior and b) provide a reliably geographically-transferable source of data.

Since administering a national annual long-distance travel survey has proven to be very financially and temporally expensive, a need to explore targeted sampling approaches is called for. To best understand how to tackle this task, a detailed review of known influencing factors in long-distance travel tripmaking and behavior is required. While a multitude of factors influence long-distance travel volumes, six factors stood out as being the most consistent findings among previous studies and best help ensure sampling equity, data applicability, and approach legitimacy. Trip characteristics such as mode choice trends, trip purpose, and the seasonality of long-distance travel; and sociodemographic characteristics such as income, age, and the household presence of children were found to be the most influential for potential long-distance travel survey targeted sampling approaches. As such, this dissertation will consider these factors when creating its targeted sampling approaches.

Chapter 3: The Case for Consistent Survey Collection Efforts

So far, this dissertation has covered the need for long-distance data as well as the major trends associated with said data; however, what has yet to be explored are the surveys themselves. While there has not been a concerted national long-distance travel survey in some years, there have still been several notable surveys capturing long-distance travel over the past three decades. These surveys vary in scope and scale from smaller-timeframe regional surveys to annual-level national surveys. Each approach offers an incredible wealth of data on travel behavior as well as serve as a "who's who" list for best practices in survey design. No survey is perfect but understanding the limitations and purpose of a survey will help guide researchers into picking the right survey for their research.

Beyond the usual caveats with a survey, another limitation outside its design is its longevity. Long-distance travel behavior is constantly growing in the US, and it is not fully known how sociodemographic groups are (or are not) evolving their travel trends. With this comes the question, are the findings from a twenty-year-old survey still relevant? This chapter aims to explore that question by examining three national surveys for changes in long-distance travel trends: the 1995 American Travel Survey (ATS), 2001 National Household Travel Survey (NHTS), and 2013 Longitudinal Survey of Overnight Travel (LSOT). The findings from this analysis will clearly make the argument for continuously capturing long-distance travel through a detailed travel survey, or at the very least, demonstrate how these surveys should be seen as a product of their own, and not a glimpse into changing long-distance travel trends.

This chapter consists of six major sections exploring long-distance travel surveys. The first section reviews the notable US long-distance travel surveys over the past three decades. This involves brief summaries of each survey, their scopes, any known caveats, and their sample sizes. Results from this section should assist researchers in choosing an appropriate survey dataset for their research, as well as set the scene for the latter part of this chapter. Next, the second through fourth sections of this chapter explores how three national-level long-distance travel surveys compare to one another regarding trends. Finally, the last two sections of this chapter summarize these findings and what we can do moving forward in the field. The findings from this section help make the case for consistently capturing travel data with a regularly applied national survey.

3.1 Notable US-Based Long-Distance Travel Surveys

Travel surveys can cover a wide range in scope, scale, and timeframe. This is very well observed with long-distance U.S. travel surveys conducted since the 1995 ATS. Here, not only have surveys fluctuated in capture time (from annual levels to daily capture), but geographic scale and how they define long-distance travel. Over a twenty-year timeframe, eight regional- and national-level travel surveys, presented in Table 1, were gathered and compared to highlight similarities and shortcomings. No survey is perfect, with each displaying some form of limitation due to design. This subsection is not meant to detract from the validity of any of the surveys presented. It is solely to provide other practitioners a comprehensive guide in comparing common long-distance travel survey data sources to best suite their research needs. Below is a quick summary of each dataset:

- **1995 American Travel Survey (ATS):**
    - Conducted by the Bureau of Transportation Statistics (BTS), this national-scale panel survey sampled nearly 80,000 households over the course of a year. Long-distance travel was defined as 100-miles roundtrip. It is still used by researchers and modelers even over 25 years later.

- **2001 National Household Travel Survey (NHTS):**
    - Also conducted by BTS, this national-scale panel survey sampled over 63,000 households on their travel habits over a 28-day period. It is considered the last true national survey capturing long-distance travel. Long-distance was defined as 50 miles one-way. However, due to the September 11th terrorist attacks and following economic recession, quarter four travel validity is questionable.
- **2004 Commute Atlanta:**
    - Conducted by the Georgia Department of Transportation (GDOT) and Georgia Tech, this Atlanta-area travel survey used vehicle-mounted GPS devices to mainly monitor local commute and daily travel trends, as well as cost sensitivity to variable tolls/taxes in its later years. About 275 households were surveyed over multiple years, however the long-distance portion of this study draws from the first year of data (base condition) and is solely trips identified as leaving the Atlanta-metro area and exceeding 50-miles one-way. Limited data is available on actual trip party characteristics.
- **2004 Michigan Travel Counts I:**
    - Conducted by the Michigan Department of Transportation (MDOT), this 14,315-household travel survey asked participants to log 48 hours' worth of travel on their given travel day. For long-distance travel, participants were asked to recall trips exceeding 100 miles one-way over the past 3-month period and 12-month period.

- **2009 Michigan Travel Counts II:**
    - Also conducted by MDOT, over 2,000 households that had participated in the previous Travel Counts survey were asked to provide a 24-hour travel diary on a given travel day. Like the previous survey iteration, recall periods of 3-months and 12-months were used to capture long-distance travel defined as exceeding 100 miles one-way.
- **2013 California Household Travel Survey (CHTS):**
    - Conducted by the California Department of Transportation (Caltrans), over 14,000 Californian households were asked to complete a 24-hour travel survey for a given travel day. Long-distance trips, defined as at least 50 miles one-way, were captured using a recall period of 8-weeks prior. However, recalled trips had very limited information recorded outside of final destinations, main mode choice, and number of days.
- **2013 Longitudinal Survey of Overnight Travel (LSOT):**
    - Conducted by Auburn University and University of Vermont researchers, over 1,000 individuals located nation-wide were asked to participate in a year-long panel survey recording their long-distance travel. This survey defined long-distance travel as any tour with an overnight component, a first.
- **2015 Michigan Travel Counts III:**
    - Conducted by MDOT, the most recent iteration of the Michigan Travel Counts survey series implemented major changes in travel capture. Over 16,000 households were asked to complete a 24-hour long travel diary, with a smaller proportion of households provided with a GPS device to track their movements in

detail for a 3-day timeframe. Long-distance trips not captured during studied travel days were compiled through a 3-month and 12-month recall of trips exceeding 100 miles one-way.

Perhaps the greatest differences between these surveys are their definitions of long-distance travel and what constitutes a trip/tour. While most of these surveys consider long-distance travel to be at least 50 miles one-way, some, such as the Michigan Travel Counts series and LSOT consider long-distance to be defined as 100 miles one-way or overnight, respectfully. However, unlike the Michigan Travel Counts series, the LSOT can further limit its long-distance definition to include distance from the final dataset, albeit long-distance day trips are lost. The other major difference is how each survey treats the capture of a long-distance trip/tour. While surveys such as the Michigan Travel Counts series and CHTS only capture the farthest origin-destination pair, others such as the LSOT and Commute Atlanta surveys capture the long-distance tour in its entirety, from origin back to origin. The ATS and NHTS treat this slightly differently. While every leg of a tour is captured, given the minimum distance requirement is met, the entire tour cannot be ascertained since each leg is treated as a separate long-distance trip. As such, consistency between surveys for comparison reasons must be careful to identify these potential issues.

Table 1: Comparing Different Long-Distance Survey Scopes [98, 155-159]

Survey	Conducted By	Structure (Lowest Level)	Trip/Tour Definition	Long-Distance Travel Definition	Geographic Scale	Timeframe	Survey Type	Data Relationship	Known Biases/Issues	Sample Size
1995 ATS	BTS	Person Trips	Any stop from one address to the next is a separate trip	100 Miles Round Trip	United States	1995 – 1996	Year-Long Panel	Defined Trip and Household IDs	Respondent Recall	80,000
2001 NHTS	BTS	Person Trips	Same as ATS but stops only to change a mode of transportation excluded	50 Miles One-Way	United States	2001 – 2002	28-Day Time Period	Can collapse by Household ID, but No Trip ID to link members	Shorter Time Period; 9/11; "Telescoping"	63,163
2004 Commute Atlanta	GDOT and Georgia Tech	Vehicle Trips	The entire tour is cataloged	50 Linear Miles One-Way and Outside the Atlanta Metro Area	Atlanta Metro Area (13 Counties)	Baseline Data Jan 2004 – Dec 2004	Year-Long Panel	Unknown	Unknown Trip Purpose/Party Info; Limited to Trips Completed by Vehicle	273
2004 Michigan Travel Counts I	MDOT	Person Trips	Furthest destination from origin considered	100 Miles One-Way	Michigan	Feb 2004 – Jan 2005	48-Hour Travel Diary; 3-Month and 12-Month Recall	Unknown	Shorter Time Period; Respondent Recall	14,315
2009 Michigan Travel Counts II	MDOT	Person Trips	Furthest destination from origin considered	100 Miles One-Way	Michigan	2009	24-Hour Travel Diary; 3-Month and 12-Month Recall	Unknown	Shorter Time Period; Respondent Recall	2,395*
2013 California Household Travel Survey	Caltrans	Person Trips	Most recent trip entirely cataloged; recalled trips limited to O-D	50 Miles One-Way	California	Dec 2011 – Feb 2013	24-Hour Travel Diary; 8-Week Recall	Unknown	Shorter Time Period; Respondent Recall; Limited Information for Recall Trips	42,431
2013 LSOT	Auburn and Vermont Researchers	Person Trips	The entire tour is cataloged (origin - origin; legs)	Overnight Tour Element	Focused mainly on Alabama, Vermont, and California	2013 – 2014	Year-Long Panel	No Specific Household ID; Individual-centric	High Income Individuals	1,024
2015 Michigan Travel Counts III	MDOT	Person Trips	Furthest destination from origin considered	100 Miles One-Way	Michigan	Jan 2015 – Dec 2015	24-Hour Travel Diary; Limited 3-Day GPS Collection; 3-Month and 12-Month Recall	Unknown	Shorter Time Period; Respondent Recall	16,276

*Household Must Have Participated in the 2004 Michigan Travel Counts Survey

Another major difference and source of concern is the scale of the survey periods. While the ATS, LSOT, and Commute Atlanta surveys were panel surveys occurring over an entire year, the other surveys relied on anywhere from a single day's worth of travel capture to a year's worth of travel recall. As long-distance travel is dependent heavily on seasonal and irregular travel patterns [76, 78, 81-83, 88, 98, 99, 101], this creates an issue where panel participants' habits would not be fully captured. This is even acknowledged in the NHTS user guide regarding scaling/weighing trip counts on an annual level [160]. As for recall, it is well known survey respondents have a hard time accurately providing details, or even accurate trip counts, the further from the dates of travel it gets [101, 138]. This issue is discussed in detail later in this chapter.

Regarding survey limitations, most of these surveys suffer from potential recall issues or shorten studied time periods. A few of these, such as the Commute Atlanta, Michigan Travel Counts, and CHTS, also suffer from a lack of captured data. For example, the Commute Atlanta survey accurately captured travel movements via GPS but could not capture the details of a trip such as the travel party size, age, or trip purpose. This shows the importance of merging traditional travel diary data with GPS data; providing researchers accurate trip information but also providing much needed travel context. As for the Michigan Travel Counts and CHTS surveys, their travel recall questions, in an effort to reduce respondent burden, limited the information asked of the respondent to the bare minimum (such as final destination, purpose, and duration). Finer details of these trips are therefore lost. Other travel surveys suffer from more unique limitations. For example, the 2001 NHTS has several notable caveats with a) the occurrence of 9/11 potentially effecting travel, b) the inability to capture annual long-distance trends, and c) the "telescoping" effect of partially complete trips being brought into or outside of the survey period. This once again limits the versatility of the capture trip data, and as such, researchers must be diligent to ensure data integrity. Finally, the LSOT has two notable limitations: first, a tighter sampling window resulted in a bias of higher income households being sampled; and second, most respondents resided in the states of Alabama, California, and Vermont. Knowing these general limitations would serve researchers well in properly applying these datasets.

3.2 The Case for Consistent Long-Distance Travel Collection

While the major long-distance travel surveys have provided valuable data for a multitude of studies, there is still the issue with working with potentially out-of-date data. We know long-distance travel is constantly increasing in volume [8, 9], and with greater affordability in air travel, potential changes in work travel due to COVID-19, and increased dispersity of social networks, the relevancy of reliable travel data is at an all-time high. What is not known is if, or how, long-distance travel behavior is changing year-to-year due to cohort and generational effects. Are mode splits the same today as they were ten years ago? Has the number of long-distance work trips decreased due to the adoption of remote meeting platforms? There is anecdotal evidence that change is occurring, but to the author's knowledge, there has yet to be a large-scale effort to explore these claims.

This section explores the three national long-distance travel surveys: the 1995 ATS, the 2001 NHTS, and the 2013 LSOT. First, to compare these vastly different surveys, a universal sociodemographic classification system is created to not only standardize terms across surveys, but also to vastly simplify what information is needed for meaningful analyses. Next, the three surveys are compared against one another for trends in mode choice, trip volumes, and trip distances to identify what claims about evolving long-distance travel trends can be made, and validate the argument for continuously capturing long-distance travel behavior through survey instruments.

3.2.1 Comparison Methodology

A universal data problem is the inconsistencies with variable definitions. While two datasets can have the same variable, such as household income, how it is logged in the dataset can vary greatly. Each of the aforementioned datasets used in the analysis categorized household income differently—some categorized in $5,000 increments, and others categorized in varying ranges. Coupled with average inflation and varying definitions of poverty, the datasets cannot be immediately normalized and consist of a very wide range of subgroups. To best compare the datasets to themselves and each other, a universal household categorical system was created utilizing three known factors affecting long distance and intercity travel volumes: household income, respondent age, and the presence of children in the household. Each category was broken down as such:

- Household Income (income range varies, defined in the following paragraph):
    - Low
    - Medium
    - High
- Age Group of Respondent or Oldest Member in Household:
    - Under 25 years old
    - 25 to 44 years old
    - 45 to 64 years old
    - 65 years or older

- Household Children:
    - Yes, at least one household individual is under 18 years of age
    - No individuals under the age of 18 are present in the household

While more demographics would be preferred, these groups were selected to ensure that a) each dataset was able to represent most if not all the groups and b) there were adequate sample sizes to make conclusions about travel variation for each group. The reasoning and logic for these choices are fully explained in chapter 5 of this dissertation. Each household was then assigned into one of 24 combined categories using the income level, age group, and presence of children (ex. low income, age 25 to 44, and no children; would be one category).

As income levels varied by year, the decision was made to create a universal low, medium, and high-income level band system using the specific year's median income and poverty threshold for an average four-person household. Income data was pulled directly from US Census estimates for the specific year [161-163]. Equation 1 defines the definition of each income level.

$$Income\ Level = \begin{cases} \boldsymbol{low} < \tilde{x}_{inc} - \frac{\tilde{x}_{inc} - PT}{2} \\ \tilde{x}_{inc} - \frac{\tilde{x}_{inc} - PT}{2} \leq \boldsymbol{medium} \leq \tilde{x}_{inc} + \frac{\tilde{x}_{inc} - PT}{2} \\ \boldsymbol{high} > \tilde{x}_{inc} + \frac{\tilde{x}_{inc} - PT}{2} \end{cases} \tag{1}$$

Where $\tilde{x}_{inc}$ is the median household income and PT is the poverty threshold.

The income level threshold values were further rounded to the nearest $5,000 increment to better match the datasets. The exception to this was for the 2013 LSOT dataset where the threshold values were rounded to the nearest household income level categorical cutoff value. Table 2 shows the income level values and the resultant thresholds used for this analysis.

Table 2: US Census Household Income Estimates and Income Level Band Calculations

		1995	2001	2013
Base Income Statistics	*Median*	$34,076	$42,228	$51,939
	Poverty Threshold (Average 4 Person Household)	$15,569	$18,104	$14,053
Income Level Band Calculations	*Band Size*	$9,254	$12,062	$14,053
	Low	$24,823	$30,166	$37,887
	High	$43,330	$54,290	$65,992
	Low (Rounded to Nearest Dataset Category)	$25,000	$30,000	$50,000
	High (Rounded to Nearest Dataset Category)	$45,000	$55,000	$75,000

Chapter 5 results recommended consolidating the Over 65 age groupings as well as remove trips completed by those in the Under 25 age groupings (due to sampling restrictions). Table 3 presents the sample size breakdown of this study's travelers for each survey. Note these counts are based on travelers that completed at least one overnight trip as well as had known demographic/trip variables (i.e., no "N/A" or "SYS MISS" variables).

Table 3: Demographic Grouping Proportions of Travelers

Income Category	Age Category	Presence of Children in Household	1995 ATS		2001 NHTS		2013 LSOT	
			Count	*% of Sample*	*Count*	*% of Sample*	*Count*	*% of Sample*
Low	*25 to 44*	*No*	1,612	3.7%	214	2.2%	101	10.8%
		Yes	1,638	3.8%	266	2.8%	19	2.0%
	45 to 64	*No*	2,225	5.1%	421	4.4%	29	3.1%
		Yes	387	0.9%	109	1.1%	6	0.6%
	65 or Older	*All*	4,035	9.3%	489	5.1%	8	0.9%
Medium	*25 to 44*	*No*	3,210	7.4%	376	3.9%	70	7.5%
		Yes	5,026	11.6%	625	6.5%	38	4.1%
	45 to 64	*No*	4,845	11.2%	740	7.7%	64	6.9%
		Yes	1,395	3.2%	233	2.4%	22	2.4%
	65 or Older	*All*	2,959	6.8%	529	5.5%	11	1.2%
High	*25 to 44*	*No*	2,270	5.2%	772	8.0%	110	11.8%
		Yes	4,785	11.0%	1,694	17.6%	117	12.5%
	45 to 64	*No*	5,205	12.0%	1,728	18.0%	206	22.1%
		Yes	2,117	4.9%	999	10.4%	99	10.6%
	65 or Older	*All*	1,678	3.9%	427	4.4%	34	3.6%
	n		43,387	100.0%	9,622	100.0%	934	100.0%

While the ATS and NHTS had much higher sampling counts among all demographic groups compared to the LSOT, the overall proportions were generally the same. All three datasets tended to skew towards the high-income groups, as expected by the literature. Even before statistical testing, it is obvious that there is quite some variety between survey proportion samplings. It would be expected this could play a role in chi-square results indicating statistical differences between the surveys.

Comparisons between dataset distributions was completed using Pearson's chi-square test for independence. While mainly used in the medical field for identifying if a treatment and another factor are independent of each other [164], it can also be applied here to determine if there is an independent relationship between the surveys and another, standardized categorical variable. Therefore, if the associated chi-square value is found to be statistically significant, the null hypothesis—that the outcome distributions are independent of (similar in distribution to) one another—can be rejected showing that there is some dependence among the categorical groupings; or simply put, there is a difference in the survey distributions.

To perform the Pearson's chi-square test of independence, the observed values must be compared to an associated expected value. In some cases, this would be left to the researcher to determine or possibly from a control group. However, in this study's case, the expected value must be calculated based on an expected proportion of the column (survey [LSOT, ATS, NHTS]) and row (distance, mode choice, or demographic grouping) totals of the observed values. Equation 2 shows this calculation.

$$Expected\ Value\ = \frac{Row\ Total * Column\ Total}{Table\ Total} \tag{2}$$

The expected value is calculated for each observed value. For example, in the comparison of two variables which each consisted of two categories, there would be four observed variables in a crosstab table. Therefore, four expected values would be calculated. This can be scaled as needed. In this study, there are three columns (surveys), and a variable set of rows representing different trip breakdowns (i.e., mode choice, one-distance, etc.). A sample crosstab of surveys and three distance categories would therefore be a 3x3 crosstab which each cell being an observed value (actual data) *and* a calculated expected value to compare against.

The actual chi-square value is calculated using Equation 3.

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{3}$$

Calculated values were compared to an associated critical value as defined by the degrees of freedom of the crosstab table as well as the alpha value. For this study, alpha values of 0.1 and 0.05 (providing confidence levels of 90 and 95 percent, respectively) were used.

The Pearson's chi-square test of independence assumes a simple random sample of a given population, independent observations, adequate total sample size, and expected values of five or greater.

3.3 Comparing Survey Travel Trends

The objective of this subsection is to determine if there are consistencies in long-distance travel behavior trends and representation among defined demographic groups between different long-distance surveys. Using three national surveys, travelers are binned into 20 demographic groupings based on income, age, and household makeup. To control for the increasing volumes in long-distance commuting patterns [87], long-distance trips are limited to meeting a minimum distance of 50-miles one-way as well as an overall trip duration of greater than one day. Trips were then analyzed for distribution trends using Pearson's chi-square test of independence among a) respondent and trip volume distributions, b) one-way distance, and c) mode choice. To the authors' knowledge, this is the first instance of testing these surveys for trend consistencies. Results from this chapter will help understand potentially two factors: if overnight long-distance travel trends are changing over time within demographic groupings, highlighting a potential need for modeling changes; or that there are various inconsistencies between national surveys due to sampling/design and a call for a future national-level travel survey should be warranted.

3.3.1 Demographic Breakdown of Travelers and Trips

The surveys were first tested for consistencies in sampled demographic groups. While weighing is not applied to these trips/demographics, the idea here was to determine if there were any major differences in demographic distributions of those who took at least one overnight long-distance trip. Therefore, it could be argued that if long-distance travel trends were not changing among demographic groups, the distribution of the groups would not change between surveys, regardless of sampling differences. Table 4 in the previous section shows the raw counts of travelers in each demographic group and those distributions were tested for independence. It was noted that there was an observable variety in survey proportion samplings with both the NHTS and LSOT datasets showing a greater proportion of high-income individuals (58.4 and 68.6 percent, respectively) and the ATS showing a slightly greater proportion of medium income individuals (40.2 percent) compared to high income (37.0 percent). Comparing age groups showed further variety in trends. Both annual panel surveys, the ATS and LSOT, had smaller proportion shares as the age category increased. The NHTS, however, had a slightly higher proportion of travelers in the "45 to 64 years old" category (44.0 percent) compared to the "25 to 44 years old" category (41.0 percent). All three surveys had the smallest proportion of travelers in the "65 years or older" category. As for the presence of children, only comparing "non-65 or older" category travelers, since the child category was aggregated for this age group, found all three surveys had a higher proportion of travelers within non-children households.

Chi-square testing between demographic distributions found them to be statistically different of one another (*p-value* < 0.0001). Comparing surveys in pairs found all combinations to also be statistically different with associated p-values less than 0.0001. Based on these results, each survey represents one of two things: there is a variance in sampling methods, or long-distance travel trends between demographic groups are changing between survey years. Determining which is the cause is outside the scope of this study and would most likely require a multi-year longitudinal survey to determine.

Further comparing respondent distributions to trip distributions does highlight a pattern across all three surveys: high income individuals generally complete a greater share of overnight long-distance trips than their associated respondent distribution. As previous research shows high income as an influencing factor in long-distance trip making, it would be expected that a greater proportion of both made trips and respondent representation would be made in this category. Table 4 compares the grouping proportions of both trips and travelers for the sample.

Table 4: Demographic Grouping Proportions of Trips and Travelers (Columns Sum to 100 percent)

Income Category	Age Category	Presence of Children in Household	1995 ATS		2001 NHTS		2013 LSOT	
			Trips	*Travelers*	*Trips*	*Travelers*	*Trips*	*Travelers*
Low	*25 to 44*	*No*	3.2%	3.7%	2.2%	2.2%	9.1%	10.8%
		Yes	2.5%	3.8%	2.7%	2.8%	0.9%	2.0%
	45 to 64	*No*	3.8%	5.1%	4.2%	4.4%	2.6%	3.1%
		Yes	0.7%	0.9%	1.1%	1.1%	0.6%	0.6%
	65 or Older	*All*	5.2%	9.3%	4.6%	5.1%	0.6%	0.9%
Medium	*25 to 44*	*No*	7.4%	7.4%	4.0%	3.9%	7.4%	7.5%
		Yes	10.0%	11.6%	6.4%	6.5%	2.7%	4.1%
	45 to 64	*No*	10.7%	11.2%	7.5%	7.7%	5.9%	6.9%
		Yes	3.2%	3.2%	2.4%	2.4%	1.9%	2.4%
	65 or Older	*All*	5.3%	6.8%	5.0%	5.5%	1.2%	1.2%
High	*25 to 44*	*No*	7.0%	5.2%	8.6%	8.0%	13.3%	11.8%
		Yes	13.4%	11.0%	17.3%	17.6%	13.1%	12.5%
	45 to 64	*No*	17.2%	12.0%	19.1%	18.0%	25.8%	22.1%
		Yes	6.9%	4.9%	10.7%	10.4%	11.3%	10.6%
	65 or Older	*All*	3.6%	3.9%	4.3%	4.4%	3.6%	3.6%
	n		232,914	43,387	12,976	9,622	7,080	934

Trip distributions were also tested for consistencies across surveys and were also found to be statistically different from each other on both the global-level (*p-value* < 0.0001) and the pairwise-level (all *p-values* < 0.0001). As the Pearson's test takes into consideration both the row and column totals for determining the expected distribution, the fact the NHTS is not an annual survey and is partially capturing long-distance behavior is a slight non-factor. In fact, the greatest dissonance between observed and expected trip counts was with the LSOT dataset where the overrepresentation of travelers/trips made by the "low-income, 25 to 44 years old, no children" demographic group greatly influenced comparison results negatively. However, as pairwise comparisons between surveys found similar results, it can best be concluded that each survey's sample pool showed different trip making volume behaviors. As with respondent demographic distribution conclusions, it cannot be determined if this is a function of a change in travel behavior over time, or a function of survey sampling variance.

3.4 Trip Characteristic Consistencies

While the sociodemographic samples of each survey might vary, it does not necessarily mean trip behavior changes between surveys. Comparing raw respondent and trip sampling distributions showed that each survey captured the population differently. However, focusing on the actual trip characteristics helps determine if there are any consistencies in demographic group behavior across the three surveys.

In the following subsections, the number of trips each sociodemographic group completed were compared across surveys based on stated purpose, one-way trip distance, and mode choice. For example, a trip could be binned as a work trip, under 200 miles, traveled by a personal vehicle.

This section is structured as follows: first, trip one-way distance trends are compared between surveys, demographic groups, and trip purposes. This is followed by a comparison of mode choice trends with each one-way distance category further broken down by primary mode choice.

3.4.1 Distance Consistencies

National passenger-mile volumes have steadily increased over the years, so determining if that impacts long-distance travel distance trends was a crucial facet. Here, it would be viewed as either the passenger-mile volumes effected total trip volumes or effected actual trip distance. As the previous section did find universal differences in total trip volumes, the expectation would be similar for trip distances. With trips analyzed further as either work or leisure trips, which has been found to have differing behavioral factors [75, 77, 80, 81, 83], a more nuanced comparison could be completed.

Trips were divided into three one-way distance categories: up to 200 miles, 201 to 500 miles, and over 500 miles. Each trip record was then sorted into one of these distance categories, trip purpose (work or leisure), and their associated demographic group. A chi-square value was then calculated for each demographic group comparing the consistencies of trip distance distribution of the three surveys. Figures 2a through 2c show the results for the work and leisure distribution comparisons by income group. Bolded X^2 values signify significant differences between the survey distributions.

Figure 2a: Global One-Way Distance Distribution Comparisons for Low Income Groups
(= 90% Confidence Level, ** = 95% Confidence Level)*

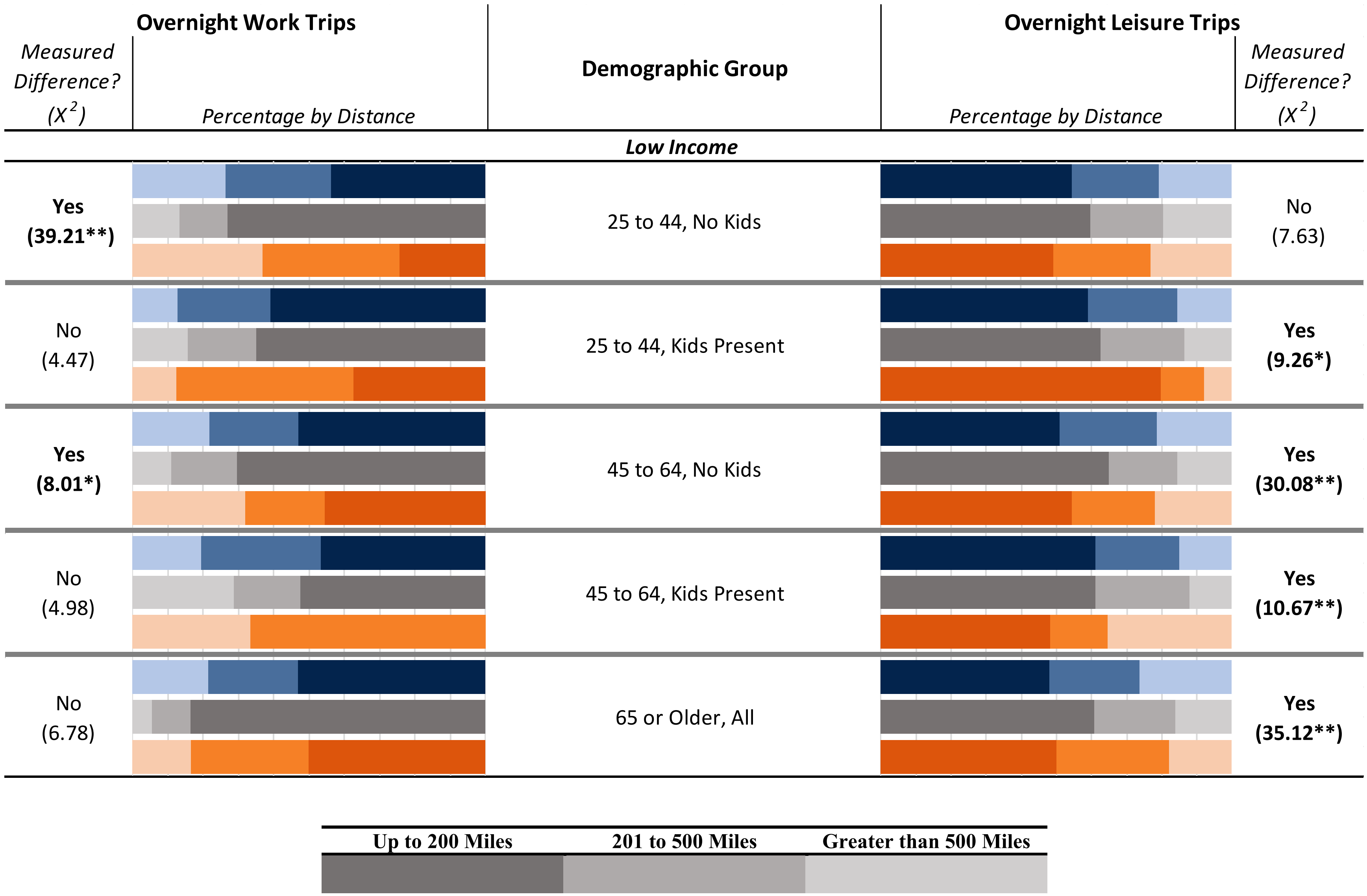


Figure 2b: Global One-Way Distance Distribution Comparisons for Medium Income Groups
(= 90% Confidence Level, ** = 95% Confidence Level)*

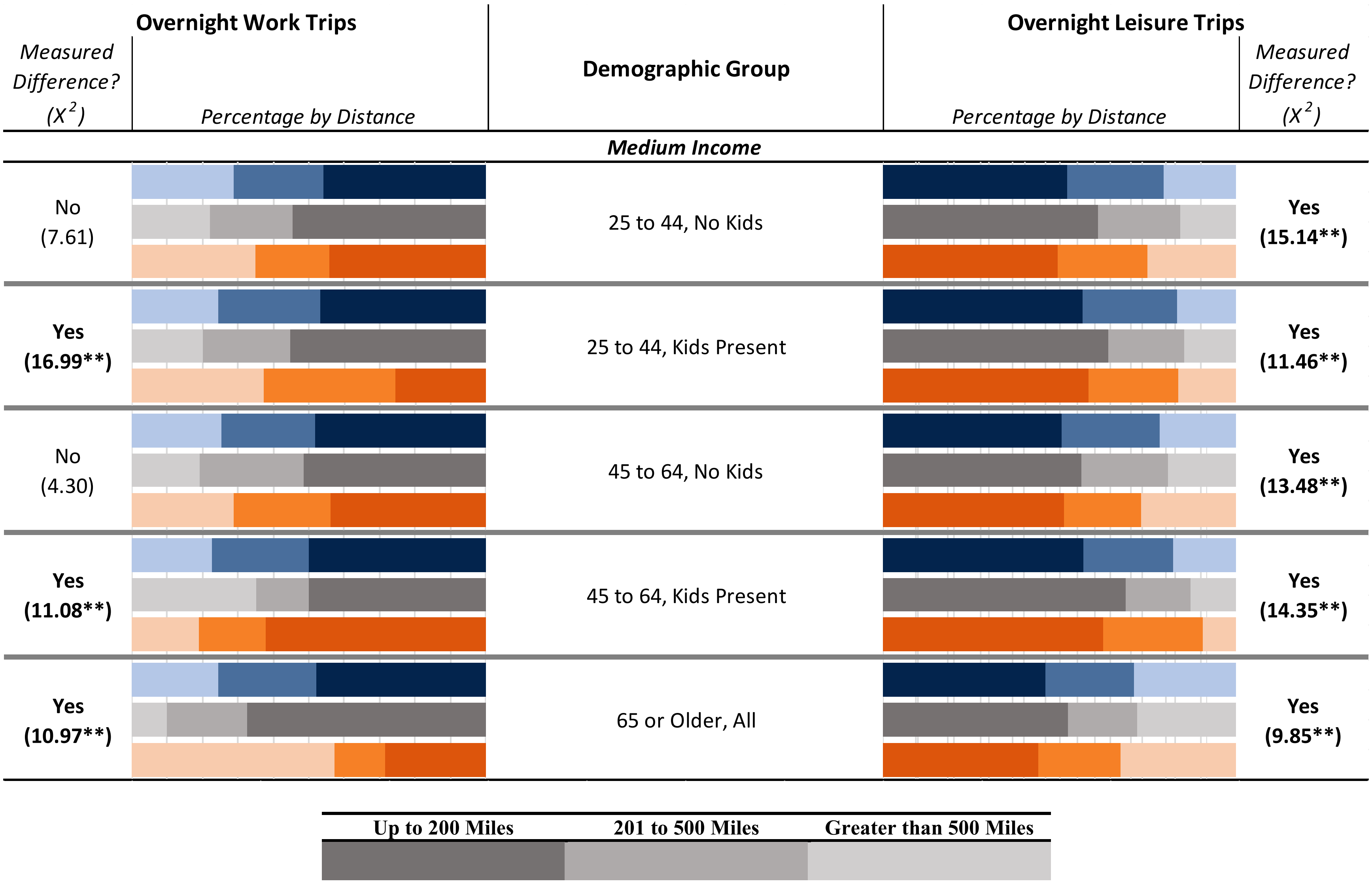
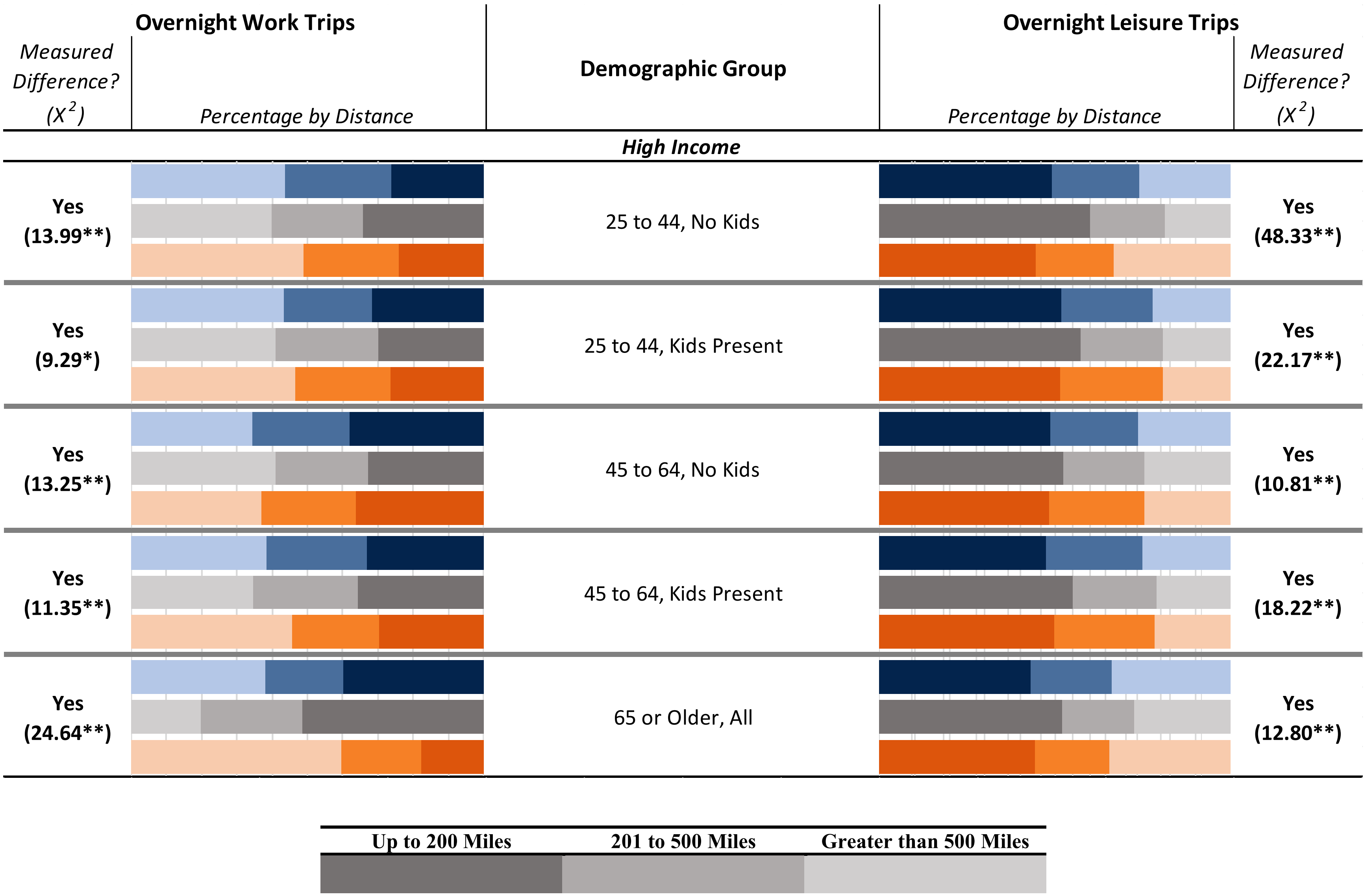


Figure 2c: Global One-Way Distance Distribution Comparisons for High Income Groups
(= 90% Confidence Level, ** = 95% Confidence Level)*



General trends across all three figures suggest there is a greater consistency within work trips than leisure trips. However, this is a very limited difference. Overall, there seems to be a definitive difference between surveys in terms of distance traveled. In fact, there seems to be very little pattern at all between distributions, suggesting the distributions are closer in relation to the survey sample than any outside travel factor. The only notable trend is with the NHTS distributions favoring shorter trip distances—a result of the non-annual nature of the survey missing the rarer, extreme-distance trips.

Between income groups, those labeled as high-income saw complete dissonance between survey distributions for all sub-demographic categories. This was also present within the medium-income groups, especially within leisure trips. Low-income groups saw the most discourse within leisure travel too, with work travel seeing discourse in only two of the five demographic groups. The work travel differences can most likely be explained by the types of jobs generally associated with each income group: work travel is more likely to occur within higher income jobs which potentially would allow greater travel variety. However, leisure travel differences are not as easy to explain outside of either survey sample travel bias or another external factor such as economic conditions.

Although general trends can be gathered from the previous figures, further analyzing how each survey compared to the other allows identifying if a single survey is distributed differently, or if all three surveys are in discontent. Tables 5, 6, and 7 present a more detailed breakdown of survey comparisons. Standard p-values were utilized in this model with bolded numbers highlighting values reaching at least a 90 percent confidence level. It should be noted that these chi-square values were calculated using specific expected values computed from the distribution of the *paired* surveys resulting in some differences between pairwise results and overall comparison results illustrated in the previous figures.

These pairwise comparisons provide rather interesting results compared to the overall chi-square values of earlier. Where there was discontent between survey distributions, the NHTS was generally the survey at fault. An example would be the results for high income groups where the NHTS was inconsistent with at least one other survey for every age/child group. This is once again a factor of the NHTS not being an annual panel survey and failing to capture the nuances of long-distance travel. However, both the ATS and LSOT also showed differences in distributions throughout the groupings. Only three groupings were consistent between all three surveys: "low-income, 25 to 44, children present"; "low-income, 45 to 64, children present"; and "medium-income, 45 to 64, no children"; all relating to work travel. These groupings were also found to be in convention among the global comparisons, however, the "medium-income, 25 to 44, no children" work travel grouping—found to be statistically similar in the global comparisons—found discontent between the NHTS and LSOT surveys in the pairwise comparisons. Exploring the possible cause of this points once again to the NHTS's bias towards shorter distance trips as well as the influence the LSOT's small sample size has on the actual chi-square calculation. As sample size becomes smaller, the potential error between the observed and expected values used in the chi-square calculation increases exponentially. Other notable results showed only five groupings had complete dissonance among all three surveys. Looking at visuals in Figures 2a through 2c clearly shows the vast differences in travel distance patterns between the three surveys for all five of the groupings. The most likely explanation for this would be from each survey's individual sample behavior rather than an overall temporal travel trend.

Table 5: One-Way Distance Distribution Comparisons between Dataset Pairs for Low Income Households
(90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value) NHTS	Overnight Work Trips Measured Difference? (X^2 p-value) ATS		Low Income: Presence of Children in Household	Low Income: Age Category	Low Income: Presence of Children in Household		Overnight Leisure Trips Measured Difference? (X^2 p-value) ATS	Overnight Leisure Trips Measured Difference? (X^2 p-value) NHTS
-	**Yes (0.000)**	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	No (0.329)	-
Yes (0.000)	**Yes (0.000)**	*LSOT*	*No*	*24 to 44*	*No*	*LSOT*	**Yes (0.083)**	**Yes (0.042)**
-	No (0.366)	*NHTS*	*Yes*	*24 to 44*	*Yes*	*NHTS*	No (0.598)	-
No (0.139)	No (0.306)	*LSOT*	*Yes*	*24 to 44*	*Yes*	*LSOT*	**Yes (0.015)**	**Yes (0.075)**
-	**Yes (0.035)**	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.000)**	-
Yes (0.061)	No (0.528)	*LSOT*	*No*	*45 to 64*	*No*	*LSOT*	No (0.487)	**Yes (0.064)**
-	No (0.314)	*NHTS*	*Yes*	*45 to 64*	*Yes*	*NHTS*	No (0.671)	-
No (0.137)	No (0.268)	*LSOT*	*Yes*	*45 to 64*	*Yes*	*LSOT*	**Yes (0.009)**	**Yes (0.012)**
-	**Yes (0.038)**	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes (0.000)**	-
No (0.264)	No (0.898)	*LSOT*	*All*	*65 or Older*	*All*	*LSOT*	No (0.535)	No (0.480)

Table 6: One-Way Distance Distribution Comparisons between Dataset Pairs for Medium Income Households (90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			Medium Income				Overnight Leisure Trips Measured Difference? (X^2 p-value)	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	No (0.109)	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	**Yes (0.005)**	-
Yes (0.058)	No (0.215)	*LSOT*				*LSOT*	No (0.117)	**Yes (0.003)**
-	**Yes (0.067)**	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.003)**	-
Yes (0.000)	**Yes (0.003)**	*LSOT*				*LSOT*	No (0.926)	No (0.553)
-	No (0.212)	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.026)**	-
No (0.160)	No (0.567)	*LSOT*				*LSOT*	**Yes (0.046)**	**Yes (0.040)**
-	**Yes (0.012)**	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.004)**	-
No (0.212)	No (0.327)	*LSOT*				*LSOT*	No (0.161)	No (0.180)
-	**Yes (0.033)**	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes (0.009)**	-
Yes (0.009)	No (0.135)	*LSOT*				*LSOT*	No (0.775)	No (0.399)

Table 7: One-Way Distance Distribution Comparisons between Dataset Pairs for High Income Households
(90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			High Income				Overnight Leisure Trips Measured Difference? (X^2 p-value)	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	**Yes (0.004)**	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	**Yes (0.000)**	-
Yes (0.015)	No (0.262)	*LSOT*				*LSOT*	**Yes (0.000)**	**Yes (0.000)**
-	**Yes (0.077)**	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.000)**	-
No (0.218)	No (0.118)	*LSOT*				*LSOT*	No (0.126)	**Yes (0.021)**
-	**Yes (0.003)**	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.021)**	-
No (0.263)	No (0.408)	*LSOT*				*LSOT*	No (0.227)	**Yes (0.053)**
-	No (0.253)	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.000)**	-
Yes (0.005)	**Yes (0.018)**	*LSOT*				*LSOT*	No (0.295)	No (0.143)
-	**Yes (0.010)**	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes (0.002)**	-
Yes (0.000)	**Yes (0.001)**	*LSOT*				*LSOT*	No (0.861)	No (0.169)

In conclusion, distance comparisons found several notable trends and results:

- Global chi-square comparisons between all three surveys showed major discourse in travel distance trends for most demographic groups.
- Work trip distance results showed more consistencies within groups than leisure trip distances but results still suggest there are vast differences inherent to each survey's sample.
- NHTS 2001 distributions were the most inconsistent when compared individually to the ATS and LSOT distributions.

Thus, utilizing the NHTS 2001 for long-distance travel in direct comparison to the annual-level panel surveys is not recommended. It would best be served to further analyze the trends between surveys over a smaller temporal frame such as a single month, season, week, or quarter. It is still indeterminable if the major inconsistencies between surveys are a product of changes in travel behavior or survey discourse. However, these findings do provide further evidence each survey is more a product of their own than a consistent, continuous source of long-distance travel data.

3.4.2 Mode Choice Consistencies

As with passenger-mile volume increases possibly affecting travel distance trends, the potential for these increases affecting mode choice was also theorized. Particularly, air travel passenger-mile volumes have steadily increased each year so an expected trend of increased air travel as a long-distance mode choice was assumed. By testing for choice consistencies between surveys, each demographic group and distance category could be reviewed for mode choice behavior consistencies. If a demographic group was consistent across surveys, it suggests there are no major differences in trends from expected volumes. This does not necessarily mean there isn't a change in mode choice trends, but that somewhere, the travelers are behaving unexpectedly (i.e., a much larger share of trips was completed by vehicle than air when it would be expected the vehicle share would be much less). It should also be noted that by breaking trip volumes to such specific bins, the possibility of a very small sample size was highly likely. With how the chi-square value is calculated, smaller sample sizes result in smaller expected values, which in turn produce a more volatile chi-square cell value. The difference between an observed 5 trips and expected 8 trips is greater than the difference between an observed 105 trips and expected 108 trips.

Like the distance comparisons, each trip was binned into the appropriate mode choice, distance category, trip purpose, and demographic group. As vehicle and air mode choices dominate long-distance travel, the analysis focused solely on those trips; trips that had another main mode (cruise, train, bus, etc.) were removed from this analysis. Global comparisons for trips up to 200 miles (Figure 3), trips between 201 and 500 miles (Figure 4), and trips over 500 miles (Figure 5) are broken down by demographic group and purpose. Any charts labeled as "N/A" had less than 20 observations, but this did not affect chi-square calculations. What will become obvious is that with some smaller sample sizes, the chi-square values and related pie charts might sometimes appear to be in complete disagreement. A visual difference does not necessarily result in a significant chi-square value. This relates back to how the chi-square value compares raw observation counts and not the percentages displayed in the pie charts. These charts largely function as a visual aid rather than the statistical truth.

Figure 3: Global Mode Choice Comparisons for Trips up to 200 Miles
(= 90% Confidence Level, ** = 95% Confidence Level)*

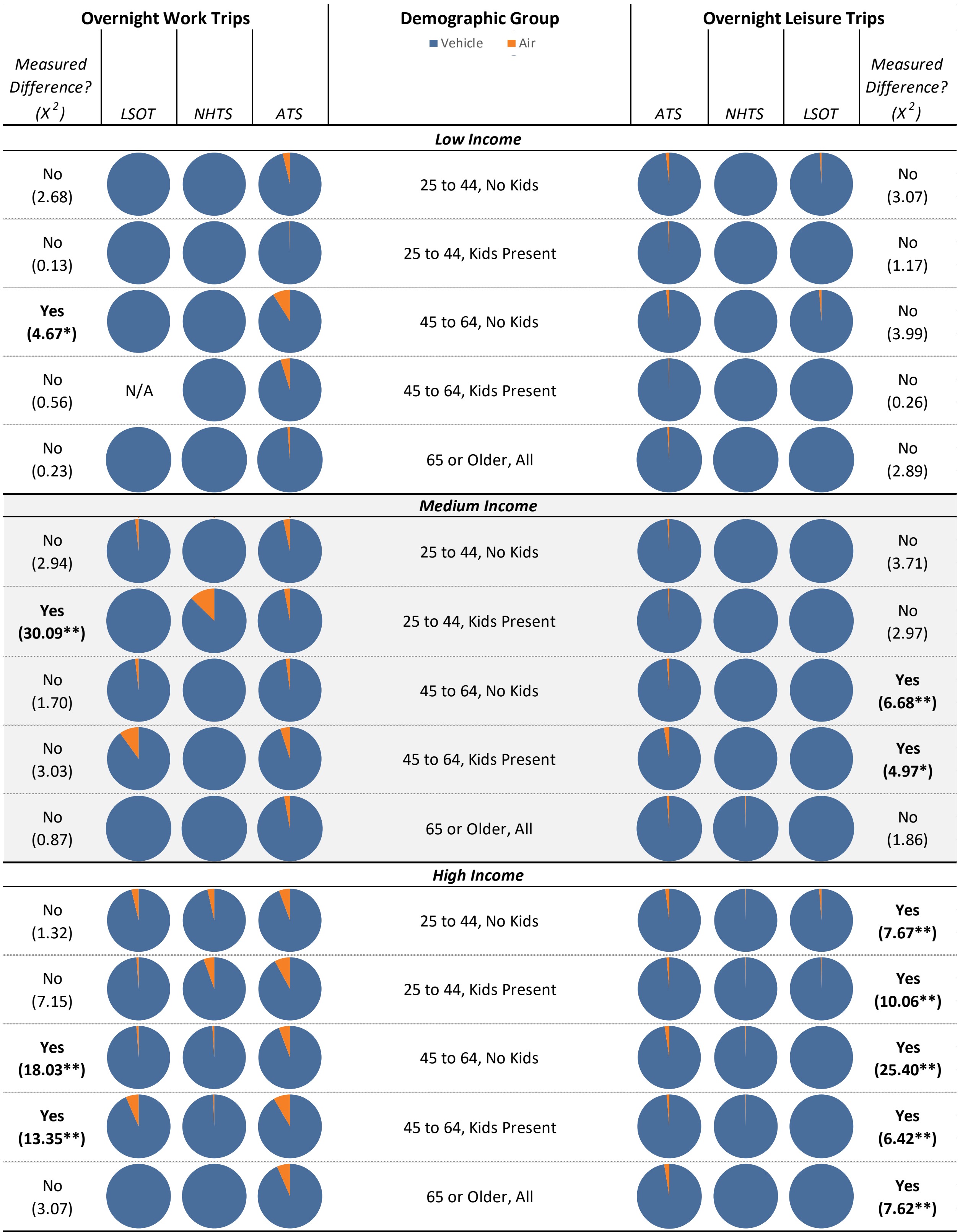


Figure 4: Global Mode Choice Comparisons for Trips Between 201 and 500 Miles (* = 90% Confidence Level, ** = 95% Confidence Level)

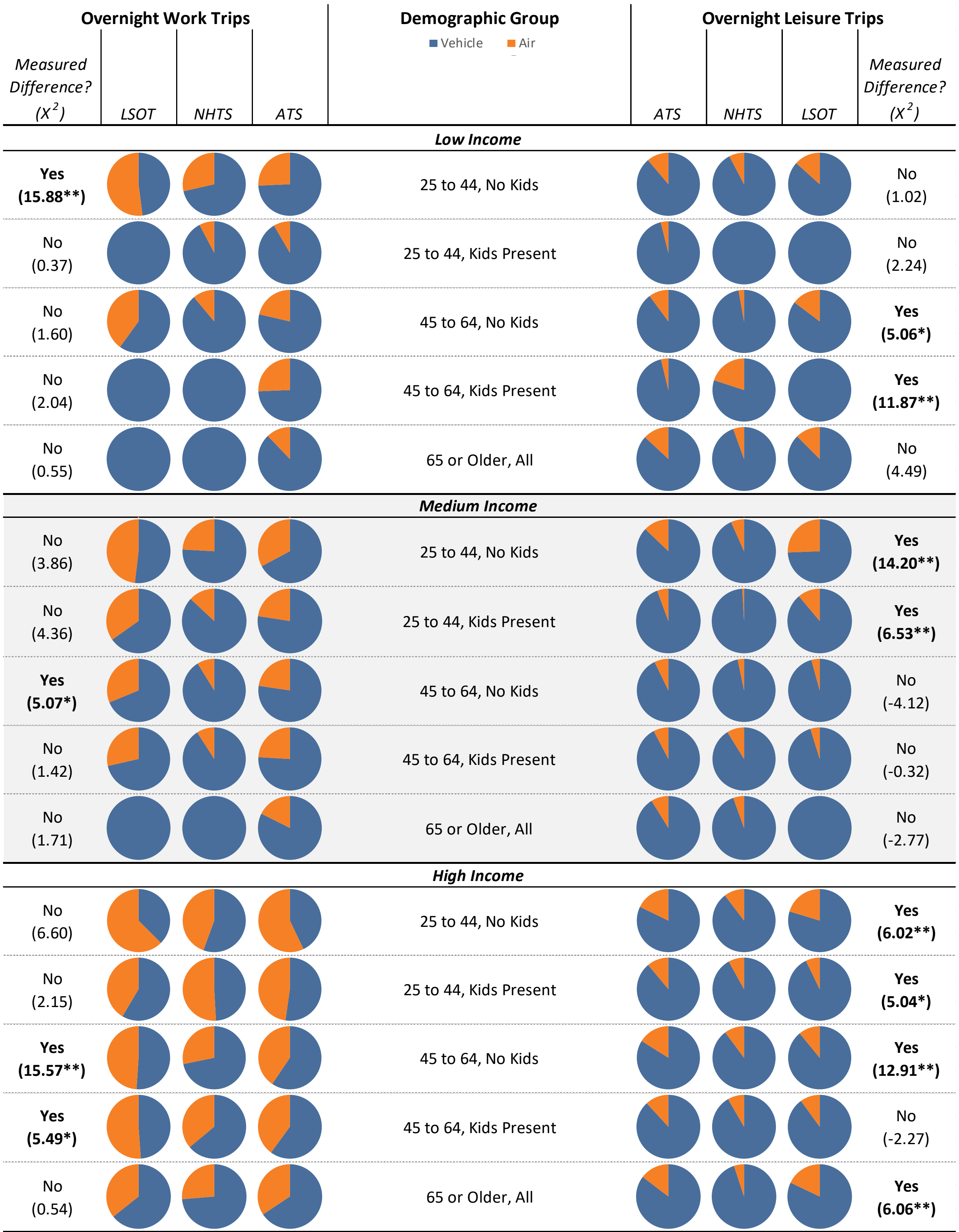
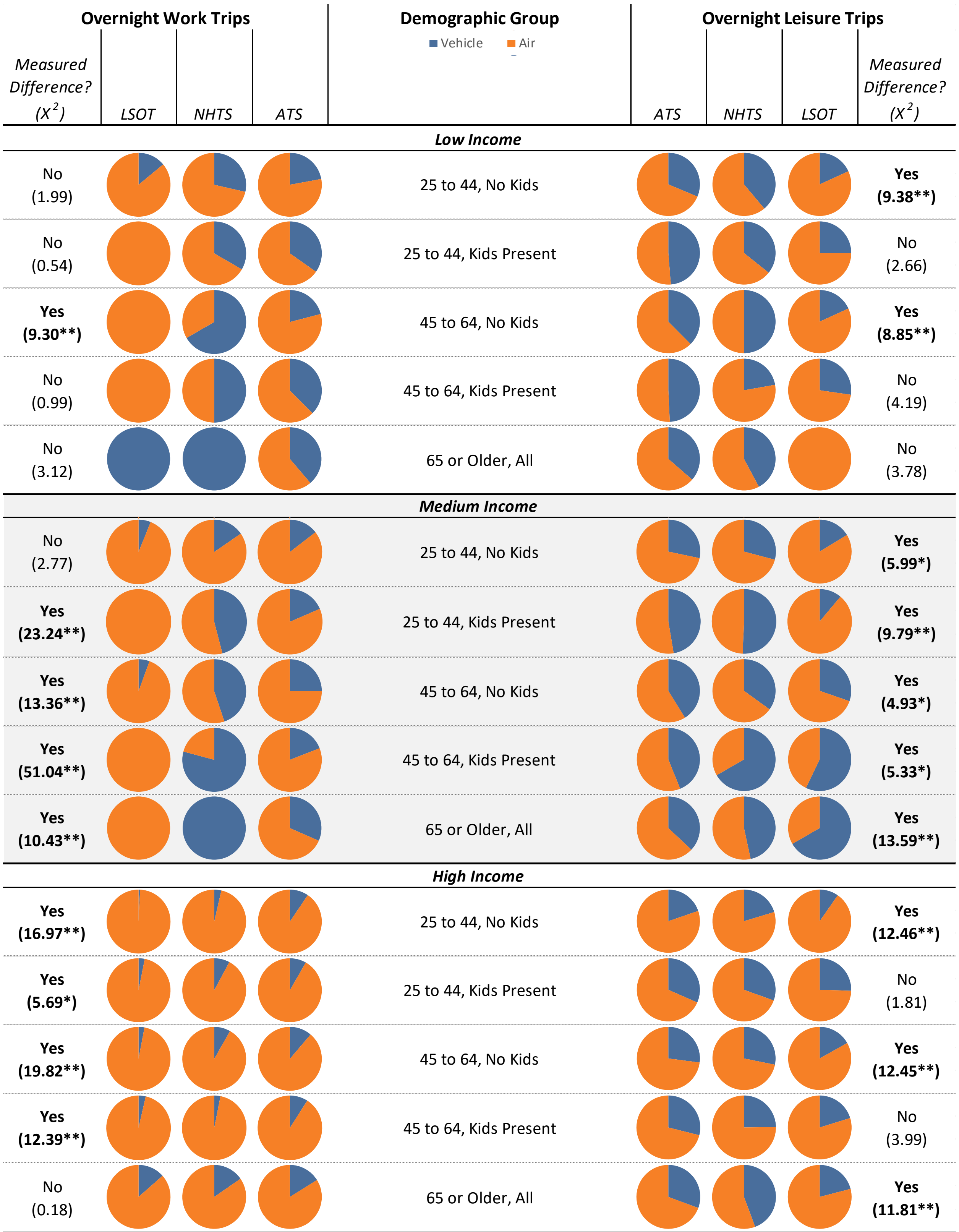


Figure 5: Global Mode Choice Comparisons for Trips Over 500 Miles
(= 90% Confidence Level, ** = 95% Confidence Level)*



Overall results suggest a greater consistency in mode choice trends between surveys among all demographic groups and distance categories. These distance categories also visually represent the concept of value of time where trips up to 200 miles one-way were predominately completed via personal vehicle, and trips greater than 500 miles one-way were completed via air. Trips between 201 and 500 miles one-way showed an interesting split in mode choice, particularly with work trips. This would suggest that the "inflection point" of travel time versus travel cost for air is somewhere in this category, or within the early tail of trips greater than 500 miles. In other words, the value of time for most individuals would be within this distance range. Exactly where this point is could be a notable future study.

Work and leisure travel results suggest air travel is more valid for work trips than leisure trips, particularly for trips under 500 miles one-way. Previous research supports long-distance air travel as having greater viability with work travel citing shorter trip durations and usually being financed by the employer rather than the employee [75, 77, 83-86]. Income levels also reflect work trip mode choice with generally the higher the income group, the greater the air mode share. Leisure travel results show a similar pattern among income levels, but there is more discontent among survey results in this purpose category. Overall, mode choice regarding trip purpose follows a more individualistic approach for leisure travel than work travel.

Respondent income levels showed different trends regarding survey consistencies on mode choice over all distance categories. For both low- and medium-income travelers, trips under 500 miles showed general mode choice consistency within demographic groups, slightly favoring low-income groupings. This trend changes once distance exceeds 500 miles one-way as low-income demographics mainly stay consistent and medium-income groups show all but one group being in dissonance. High-income groups are for the most part inconsistent within groups across all distance categories. This could reflect the cost of air travel in comparison to personal vehicle, with lower-income groups having less financial freedom to choose air travel for shorter distance trips [76, 78, 81-83]. High-income travelers' general inconsistencies in leisure travel mode choice could also reflect this idea as these individuals would have a greater freedom in choosing air travel especially in shorter distance situations.

Breaking down by pairwise survey comparisons shows similar individual results as the global comparisons of the previous figures. Most dissonance was observed in the higher income and distance categories. Tables 8 through 16 present these detailed breakdowns of survey comparisons. Standard p-values were utilized in this model with bolded numbers highlighting values reaching at least a 90 percent confidence level. Tables are broken between income level and distance categories. Comparison p-values represent each mode choice distribution comparison between surveys. Values listed as "NA" are where a mode choice category recalled zero trips.

While LSOT and NHTS chi-squares could not be ascertained for much of the "up to 200 miles" comparisons, it appears most survey inconsistencies in this distance category occurred between the ATS and NHTS. As the ATS and LSOT comparison values were slightly more consistent, this could suggest the NHTS's design or sample population prevented an accurate sample. However, as LSOT and NHTS comparison values could not be computed for much of this distance category, it is difficult to pinpoint this as a function of panel time-period disagreement. Similarly, for the "201 to 500 miles" and "greater than 500 miles" distance categories, no clear pattern could be ascertained on which survey(s) were causing the most discourse. While the LSOT did have several instances (particularly within the "greater than 500 miles" category), the same could be said for the ATS and NHTS surveys.

Table 8: Trips Up to 200 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for Low Income Households
(90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			**Low Income**				**Overnight Leisure Trips Measured Difference? (X^2 p-value)**	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	No (0.230)	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	No (0.149)	-
No NA	No (0.264)	*LSOT*				*LSOT*	No (0.310)	No (0.315)
-	No (0.727)	*NHTS*	*Yes*		*Yes*	*NHTS*	No (0.341)	-
No NA	No (0.927)	*LSOT*				*LSOT*	No (0.610)	No NA
-	**Yes** **(0.054)**	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes** **(0.047)**	-
No NA	No (0.323)	*LSOT*				*LSOT*	No (0.786)	**Yes** **(0.085)**
-	No (0.455)	*NHTS*	*Yes*		*Yes*	*NHTS*	No (0.647)	-
No NA	No NA	*LSOT*				*LSOT*	No (0.818)	No NA
-	No (0.665)	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes** **(0.097)**	-
No NA	No (0.846)	*LSOT*				*LSOT*	No (0.707)	No NA

Table 9: Trips Up to 200 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for Medium Income Households (90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value) NHTS	Overnight Work Trips Measured Difference? (X^2 p-value) ATS		Medium Income: Presence of Children in Household	Medium Income: Age Category	Medium Income: Presence of Children in Household		Overnight Leisure Trips Measured Difference? (X^2 p-value) ATS	Overnight Leisure Trips Measured Difference? (X^2 p-value) NHTS
-	No (0.115)	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	No (0.154)	-
No (0.263)	No (0.495)	*LSOT*	*No*	*24 to 44*	*No*	*LSOT*	No (0.195)	No NA
-	**Yes (0.000)**	*NHTS*	*Yes*	*24 to 44*	*Yes*	*NHTS*	No (0.114)	-
No (0.132)	No (0.485)	*LSOT*	*Yes*	*24 to 44*	*Yes*	*LSOT*	No (0.493)	No NA
-	No (0.195)	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.025)**	-
No (0.238)	No (0.862)	*LSOT*	*No*	*45 to 64*	*No*	*LSOT*	No (0.194)	No NA
-	No (0.164)	*NHTS*	*Yes*	*45 to 64*	*Yes*	*NHTS*	**Yes (0.055)**	-
Yes (0.050)	No (0.310)	*LSOT*	*Yes*	*45 to 64*	*Yes*	*LSOT*	No (0.254)	No NA
-	No (0.369)	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	No (0.223)	-
No NA	No (0.807)	*LSOT*	*All*	*65 or Older*	*All*	*LSOT*	No (0.538)	No (0.731)

Table 10: Trips Up to 200 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for High Income Households (90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			High Income				Overnight Leisure Trips Measured Difference? (X^2 p-value)	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	No (0.326)	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	**Yes (0.009)**	-
No (0.949)	No (0.534)	*LSOT*				*LSOT*	No (0.300)	No (0.136)
-	No (0.223)	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.004)**	-
Yes (0.084)	**Yes (0.016)**	*LSOT*				*LSOT*	No (0.148)	No (0.722)
-	**Yes (0.004)**	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.000)**	-
No (0.878)	**Yes (0.002)**	*LSOT*				*LSOT*	**Yes (0.001)**	No (0.187)
-	**Yes (0.000)**	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.050)**	-
Yes (0.005)	No (0.491)	*LSOT*				*LSOT*	No (0.104)	No (0.498)
-	No (0.129)	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes (0.018)**	-
No NA	No (0.380)	*LSOT*				*LSOT*	No (0.153)	No NA

Table 11: Trips 201 to 500 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for Low Income Households (90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			Low Income				Overnight Leisure Trips Measured Difference? (X^2 p-value)	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	No (0.862)	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	No (0.496)	-
No (0.246)	**Yes (0.000)**	*LSOT*	*No*	*24 to 44*	*No*	*LSOT*	No (0.480)	No (0.345)
-	No (0.929)	*NHTS*	*Yes*	*24 to 44*	*Yes*	*NHTS*	No (0.158)	-
No (0.567)	No (0.545)	*LSOT*	*Yes*	*24 to 44*	*Yes*	*LSOT*	No (0.621)	No NA
-	No (0.456)	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.041)**	-
No (0.207)	No (0.314)	*LSOT*	*No*	*45 to 64*	*No*	*LSOT*	No (0.383)	**Yes (0.021)**
-	No (0.243)	*NHTS*	*Yes*	*45 to 64*	*Yes*	*NHTS*	**Yes (0.001)**	-
No NA	No (0.408)	*LSOT*	*Yes*	*45 to 64*	*Yes*	*LSOT*	No (0.658)	No (0.273)
-	No (0.599)	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes (0.034)**	-
No NA	No (0.599)	*LSOT*	*All*	*65 or Older*	*All*	*LSOT*	No (0.962)	No (0.426)

Table 12: Trips 201 to 500 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for Medium Income Households (90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			Medium Income				Overnight Leisure Trips Measured Difference? (X^2 p-value)	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	No (0.325)	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	No (0.113)	-
Yes (0.061)	**Yes (0.094)**	*LSOT*				*LSOT*	**Yes (0.001)**	**Yes (0.001)**
-	No (0.125)	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.026)**	-
Yes (0.034)	No (0.165)	*LSOT*				*LSOT*	No (0.221)	**Yes (0.004)**
-	**Yes (0.055)**	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.060)**	-
Yes (0.022)	No (0.252)	*LSOT*				*LSOT*	No (0.436)	No (0.741)
-	No (0.248)	*NHTS*	*Yes*		*Yes*	*NHTS*	No (0.804)	-
No (0.280)	No (0.782)	*LSOT*				*LSOT*	No (0.619)	No (0.573)
-	No (0.222)	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	No (0.267)	-
No NA	No (0.644)	*LSOT*				*LSOT*	No (0.213)	No (0.337)

Table 13: Trips 201 to 500 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for High Income Households
(90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			**High Income**				**Overnight Leisure Trips Measured Difference? (X^2 p-value)**	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	**Yes (0.018)**	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	**Yes (0.021)**	-
Yes (0.022)	No (0.361)	*LSOT*				*LSOT*	No (0.472)	**Yes (0.021)**
-	No (0.423)	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.086)**	-
No (0.143)	No (0.232)	*LSOT*				*LSOT*	No (0.131)	No (0.746)
-	**Yes (0.002)**	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.003)**	-
Yes (0.000)	**Yes (0.020)**	*LSOT*				*LSOT*	**Yes (0.028)**	No (0.754)
-	No (0.368)	*NHTS*	*Yes*		*Yes*	*NHTS*	No (0.157)	-
Yes (0.025)	**Yes (0.035)**	*LSOT*				*LSOT*	No (0.558)	No (0.630)
-	No (0.470)	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes (0.018)**	-
No (0.561)	No (0.915)	*LSOT*				*LSOT*	No (0.566)	**Yes (0.024)**

Table 14: Trips Greater than 500 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for Low Income Households (90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)				**Low Income**			**Overnight Leisure Trips Measured Difference? (X^2 p-value)**	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	No (0.682)	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	No (0.337)	-
No (0.322)	No (0.182)	*LSOT*	*No*	*24 to 44*	*No*	*LSOT*	**Yes (0.004)**	**Yes (0.011)**
-	No (0.921)	*NHTS*	*Yes*	*24 to 44*	*Yes*	*NHTS*	No (0.178)	-
No (0.488)	No (0.466)	*LSOT*	*Yes*	*24 to 44*	*Yes*	*LSOT*	No (0.344)	No (0.673)
-	**Yes (0.007)**	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	**Yes (0.065)**	-
Yes (0.009)	No (0.174)	*LSOT*	*No*	*45 to 64*	*No*	*LSOT*	**Yes (0.023)**	**Yes (0.003)**
-	No (0.545)	*NHTS*	*Yes*	*45 to 64*	*Yes*	*NHTS*	No (0.115)	-
No (0.350)	No (0.440)	*LSOT*	*Yes*	*45 to 64*	*Yes*	*LSOT*	No (0.160)	No (0.795)
-	No (0.210)	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	No (0.342)	-
No NA	No (0.210)	*LSOT*	*All*	*65 or Older*	*All*	*LSOT*	**Yes (0.091)**	**Yes (0.063)**

Table 15: Trips Greater than 500 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for Medium Income Households (90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			Medium Income				Overnight Leisure Trips Measured Difference? (X^2 p-value)	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	No (0.904)	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	No (0.898)	-
No (0.189)	**Yes (0.097)**	*LSOT*				*LSOT*	**Yes (0.015)**	**Yes (0.070)**
-	**Yes (0.000)**	*NHTS*	*Yes*		*Yes*	*NHTS*	No (0.568)	-
Yes (0.000)	**Yes (0.023)**	*LSOT*				*LSOT*	**Yes (0.002)**	**Yes (0.002)**
-	**Yes (0.015)**	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	No (0.169)	-
Yes (0.000)	**Yes (0.007)**	*LSOT*				*LSOT*	**Yes (0.075)**	No (0.517)
-	**Yes (0.000)**	*NHTS*	*Yes*		*Yes*	*NHTS*	**Yes (0.027)**	-
Yes (0.000)	No (0.200)	*LSOT*				*LSOT*	No (0.482)	No (0.643)
-	**Yes (0.004)**	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes (0.026)**	-
Yes (0.005)	No (0.176)	*LSOT*				*LSOT*	**Yes (0.003)**	**Yes (0.071)**

Table 16: Trips Greater than 500 Miles Distance Mode Choice Distribution Comparisons between Dataset Pairs for High Income Households (90% Confidence Level Bolded)

Overnight Work Trips Measured Difference? (X^2 p-value)			High Income				Overnight Leisure Trips Measured Difference? (X^2 p-value)	
NHTS	*ATS*		**Presence of Children in Household**	**Age Category**	**Presence of Children in Household**		*ATS*	*NHTS*
-	**Yes (0.019)**	*NHTS*	*No*	*24 to 44*	*No*	*NHTS*	No (0.869)	-
No (0.110)	**Yes (0.001)**	*LSOT*				*LSOT*	**Yes (0.000)**	**Yes (0.006)**
-	No (0.797)	*NHTS*	*Yes*		*Yes*	*NHTS*	No (0.704)	-
Yes (0.046)	**Yes (0.017)**	*LSOT*				*LSOT*	No (0.192)	No (0.344)
-	No (0.144)	*NHTS*	*No*	*45 to 64*	*No*	*NHTS*	No (0.638)	-
Yes (0.008)	**Yes (0.000)**	*LSOT*				*LSOT*	**Yes (0.001)**	**Yes (0.001)**
-	**Yes (0.009)**	*NHTS*	*Yes*		*Yes*	*NHTS*	No (0.274)	-
No (0.814)	**Yes (0.015)**	*LSOT*				*LSOT*	**Yes (0.085)**	No (0.418)
-	No (0.942)	*NHTS*	*All*	*65 or Older*	*All*	*NHTS*	**Yes (0.003)**	-
No (0.867)	No (0.672)	*LSOT*				*LSOT*	No (0.103)	**Yes (0.002)**

Overall mode choice comparison results highlight similar issues as the distance, trip volume, and respondent distribution comparisons did: each survey performs very differently from the other, albeit at to a lesser degree. However, several general conclusions can still be made from the mode share results:

- Mode choice consistencies for work travel were slightly better than for leisure travel, reflecting the individualistic nature of leisure travel mode choice and work travel's usually shorter trip durations.
- Trips under 500 miles one-way showed higher consistencies than trips over 500 miles. However, overall mode shares for "up to 200 miles" and "greater than 500 miles" pointed towards monolithic shares to a singular mode: vehicle and air, respectively.
- High-income groups had showed more inconsistencies between mode choice than either low- or medium-income groups. This potentially reflects the financial freedom of the high-income groups in being able to easily choose either mode regardless of trip distance. Low-income groups tended to be consistent across all distance categories suggesting favorability towards vehicle travel until distance becomes too great. This was also seen to a lesser extent with the medium-income groups with more discontent shown in the "201 to 500 miles" distance range.
- Pairwise survey consistency testing could not determine if a single survey was the main source of discourse. While NHTS comparisons were prominent, both ATS and LSOT values showed inconsistencies throughout. Further suggesting each survey as more a product of its own and potentially hinting at changes in long-distance travel behavior over time.

While it is notable that mode choice comparisons showed the greatest consistencies among this research, it still highlighted the apparent randomness in travel behavior further found in distance, respondent, and trip volume results. Thus, it could be concluded that each one of these three surveys is a product of its own and cannot be a viewed a continuous study of long-distance travel over the years.

3.5 Conclusions

As long-distance travel continues to grow in both economic and transportation network impact, the need for accurate travel trend information in application for modeling remains a major research sector. Defining long-distance travel as trips at least 50-miles one-way and containing an overnight component, three national surveys over a nearly 20-year period, influential demographic groups were tested for their consistencies in their sampling, trip volume, distance, and mode choice trends to determine if any changes in long-distance travel trends were present. Results showed global comparisons of respondent, trip, and distance distributions were for the most part in complete dissonance suggesting each survey either a unique product of its sample's individual behavior, or the result of an external change in long-distance travel behavior due to untested temporal, societal, or economic factors. Major conclusions can be summarized as follows:

- **Respondent and Trip Volume Consistencies**
    - Respondent distributions between each survey was found to be vastly different at both the global- and pairwise-levels, suggesting either travel behavior changed, or each survey captured the travel population differently.

- This was also the case for trip volume consistencies with both global and pairwise comparisons highlighting vast differences in trip distributions among demographic groupings.

- **Trip Distance Consistencies**
    - Global chi-square comparisons between all three surveys showed major discourse in travel distance trends for most demographic groups.
    - Work trip distance results showed more consistencies within groups than leisure trip distances but results still suggest there are vast differences inherent to each survey's sample.
    - NHTS 2001 distributions were the most inconsistent when compared individually to the ATS and LSOT distributions.
- **Mode Choice Consistencies**
    - Mode choice consistencies for work travel were slightly better than for leisure travel, reflecting the individualistic nature of leisure travel mode choice and work travel's usually shorter trip durations.
    - Trips under 500 miles one-way showed higher consistencies than trips over 500 miles. However, overall mode shares for "up to 200 miles" and "greater than 500 miles" pointed towards monolithic shares to a singular mode: vehicle and air, respectively.
    - Income levels showed differing levels of distribution discourse with inconsistencies becoming more extreme as the income level increased.

- Pairwise survey consistency testing could not determine if a single survey was the main source of discourse. While NHTS comparisons were prominent, both ATS and LSOT values showed inconsistencies throughout.

Overall, results suggest that these three long-distance travel surveys vary greatly in terms of travelers, trip capture, and travel behavior. Thus, it could be concluded that each one of these three surveys is a product of its own and cannot be a viewed a continuous study of long-distance travel over the years.

3.5.1 Limitations

There are several limitations to this research. Considering how chi-square tests general distribution trends, it only highlights where something is significantly unexpected. Distributions between surveys can still differ (suggesting trend changes), while also producing an insignificant chi-square value. Additionally, these tests ignore macroeconomic influences that could affect long-distance travel such as fuel price, economic stability, and airline travel prices. Further testing and potentially adding more long-distance surveys for comparison could either improve or disprove this claim. Additionally, consistencies between surveys were mainly grouped via income level. Comparisons across age groups and the presence of children could highlight unseen trends. Another alternative approach to binning demographics would be to focus on "geo-economic clustering" as used in the Hu's analysis of the NHTS [46]. This would group trips at the census-tract level as either extreme poverty, mega-urban, suburban, urban, or rural origins. Potentially the best approach to determining if this study highlights survey inconsistencies or changes in long-distance travel behavior is by reinstating a national-level annual long-distance travel panel survey, similar in scope and design to the ATS for direct comparison.

3.6 Moving Forward with More Efficient Sampling

This chapter highlighted the various techniques and approaches some US long-distance travel surveys have taken over the past few decades and how they compare to one another. While each survey fulfilled its stated purpose, the shear variety of definitions, capture periods, and geographic scopes reenforce the need for a consistent national long-distance travel survey approach. Countless researchers have utilized these surveys for behavioral research, travel demand modeling, and other novel projects, but as the years go by, these same surveys are used with the assumption that they are still factually accurate of US long-distance travel trends.

In the latter half of this chapter, three nationally-scaled surveys—the 1995 ATS, 2001 NHTS, and 2013 LSOT—were compared to see if similar long-distance travel trends between all three surveys were present. Results suggested a few major lessons: first, each of these three surveys was unique in how they captured, defined, and sampled long-distance travel. Both the ATS and LSOT captured a respondent's entire years' worth of long-distance travel behavior, while the NHTS only captured a respondent's long-distance travel behavior over a 28-day period. Additionally, the definition of long-distance travel, while controlled for in this study, differed between surveys with the LSOT capturing any trip that had an overnight component, and the ATS and NHTS defined by trip distance. This resulted in either a partial capture of long-distance travel behavior (as evident by the NHTS's stark difference in trends compared to the ATS and LSOT) or a disconnect in how a respondent would be prompted for long-distance travel information. Second, each of these surveys utilized a unique sample frame. As such, there was no way to determine if the changes in traveler trends and behaviors was a result of an observed evolution of long-distance travel behavior, or just the behaviors adherent to their respective sample. These results lead to the conclusion that each of these three surveys should be viewed as standalone products, and not a continuous study of long-distance travel.

So, what does this mean? Data users need the most up to date travel behavior information possible, but the possibility of another survey on a scale akin to the American Travel Survey is most likely not going to happen any time soon. While this can be seen is detrimental to the field, there is also opportunity. The surveys of the past offer a wealth of knowledge that, while demonstrably different in execution and scope, can still be used to make the surveys of the future more cost and temporally efficient. Do we need population-proportioned sample for our surveys, or can we target our sample around specific sociodemographic groups? Is an entire year of travel behavior needed to effectively describe a respondent’s long-distance travel, or can we target a few representative days? Do we need a representative sample of the entire US, or can a few geographic areas be just as representative? Some of these questions have been demonstrated in the past (such as with the 2001 NHTS asking for only 28 days’ worth of travel, or the current US national long-distance travel model being calibrated with the trends of more recent state long-distance travel surveys), but there has been little to no research done on the validity of these approaches. The following chapters in this dissertation aim to, if not answer these questions, at least start the conversation on how to answer these questions. By exploring the possibilities of targeted sampling in this dissertation, future national long-distance travel surveys may be completed more efficiently and regularly to ensure data users are supported with the most up-to-date data.

Chapter 4: Targeting the Survey Timeframe

Perhaps the greatest question posed by researchers is what temporal scope is needed to adequately capture long-distance travel. While the transportation industry recognizes long-distance travel volumes vary throughout a year, it is still unclear a) how much long-distance travel volume variability exists in travel days throughout a year as well as b) how to collect travel survey data to capture this tripmaking variability. Some have argued that a year-long panel survey is needed such as with the ATS or the LSOT surveys, while others, including many more state-level and more recent national-level travel surveys, rely on smaller capture periods: often seen as 3-weeks. While any approach offers insights into long-distance travel trends, the pure penetration into individual/household travel trends may vary greatly. Long-distance travel is, by definition, a non-routine event. Only capturing a fragment of an individual's annual travel behavior can fail to accurately tell their full story. This would suggest that the best way to capture someone's long-distance travel behavior would be through a longitudinal-type survey. However, the shear fiscal and temporal costs associated with this type of survey makes it unattractive to the average decisionmaker. As such, these shorter "snapshot-type" survey approaches are more common, and while they can show long-distance travel trends at a macro population level, their usefulness for minute analyses on the individual/household levels is practically impossible.

Researchers and practitioners have understood this tradeoff for years with a multitude of solutions proposed to ascertain long-distance travel habits. One common approach is by asking a respondent to recall the number of trips, and details of said trips, they have taken over a period of time. This approach, while demonstrably good at assessing those who do *not* travel often, bodes poorly for those who do travel. Dowds, in his study comparing actual respondent long-distance travel volumes to respondent self-estimated travel volumes, found that estimates of overnight tours, air travel, and international tours were consistent, at best, for 70 percent of respondents [165]. While it was concluded that the aggregated trip rates were rather accurate, the concern was still within the details—particularly the lack of verifiable information regarding trip durations, distances, modes, party sizes , and other relevant information needed for travel behavior analyses. Other researchers have tackled this issue with novel approaches such as capturing long-distance travel behavior based on the most recent respondent trip [166], but these attempts have been skeptical at best. What is generally agreed upon is that while annual data is arguably the best form of capture, its associated costs are great enough to warrant alternative capture efforts, with the understanding that the resulting data is limited at best.

Therefore, this chapter aims to measure the variability in long-distance travel volumes per travel day throughout a year considering seasons and sociodemographics. Specifically, this chapter determines a) if there are differences in the day-by-day long-distance travel patterns across six unique sociodemographic groups across a year and b) how long-distance travel days can be clustered such that the long-distance travel volumes (reflecting the range of sociodemographic groups) are similar for days within each group *and* the travel volumes are different between groups.

These tasks are accomplished using the 2013 LSOT dataset given its inclusion of specific tour beginning and end dates. Travel activity in this paper is defined as the volume of home-based trip ends (HBTEs) occurring on a single day, *i.e.*, all long-distance tours either beginning *or* ending on that day of the year. Each day is characterized by HBTEs for six sociodemographic groups relating to their respective daily total HBTE volume, percent of daily HBTE leisure purpose, percent of daily HBTE work purpose, percent of daily HBTE completed by personal vehicle, and percent of daily HBTE completed by air travel. Findings highlight the need for considering long-distance trip seasonality in setting the travel survey capture timeframe, and lay a path forward for exploring how to better capture long-distance travel, thus supporting one of this dissertation's core objectives and hypotheses.

This chapter is organized into three sections. First, an overview of the LSOT's trends and associated sociodemographic travel behavior, along with general methodology, is presented. This will not only illustrate the variability in long-distance travel throughout a year, but also how it varies greatly among demographic groupings. Next, clustering days based on travel activity results are presented. Finally, notable impacts, survey considerations, and other recommendations are consolidated in the conclusion for this chapter.

4.1 Methodology and Data

To accomplish this task, an annual-level longitudinal survey of long-distance travel was needed. While the 1995 ATS provides a respectable, national-level source of data meeting this criterion, the public file of the ATS only lists a trip's starting and ending quarter as the most detailed temporal scale. Effort was made to secure the more detailed dataset containing a trip's starting and ending dates, but after consulting with several sources, the complete ATS dataset's whereabouts are unknown, and the public file is currently the only available source of the ATS. Additionally, the need for an annual-level panel survey ruled out the use of the NHTS and other long-distance travel surveys that captured only a respondent's travel for a day or other snapshot approaches (such as a 28-day reflection or asking for just the total number of long-distance trips recently taken). This research aimed to measure the variability in long-distance travel volumes per travel day throughout a year considering seasons and sociodemographics, and as such, needed a full-detailed, annual-level, panel survey showcasing an individual's *entire* long-distance travel behavior for a year. Thus, the 2013 LSOT was chosen for this study.

In 2013, Auburn University and University of Vermont researchers conducted the Longitudinal Survey of Overnight Travel (LSOT). This survey aimed to create one of the first national surveys of long-distance travel for the US since the 1995 ATS. For this survey, long-distance travel was defined as any trip a respondent took that included an overnight segment. While this survey is slightly biased towards higher income households, it offers the most recent collection of long-distance travel in the US at the annual level and provided the necessary temporal detail needed to explore targeting the sampling timeframe. After cleaning and limiting the respondents to only those who completed the survey for the entire year, the final dataset consisted of 567 respondents who completed 5,323 long-distance trips.

This section is presented in three subsections. First, how respondents were classified into sociodemographic groups is explained. The methodology here reflects the methodology used in chapters three and five of this dissertation. The next subsection gives a review on the analysis methods used in this chapter. This includes how long-distance travel behavior was defined per day and the k-mean clustering approach. Finally, a review of the dataset used in this chapter, the 2013 LSOT is given, including basic observed trends for each sociodemographic group.

4.1.1 Sociodemographic Groupings

Respondents were placed into one of six sociodemographic groups based on the household income and the presence of a child, under the age of 18, in the household, using the methodology supported in chapter five of this dissertation. While age is a highly influential factor in long-distance tripmaking, the LSOT with its smaller sample size drove this analysis to consolidate sociodemographic groups to ensure an adequate sample for each group. In total, the following sociodemographic groups and their annual long-distance travel rates were formed into Table 17.

Table 17: Summary of Sociodemographic Groups and Associated Long-Distance Travel Rates

Household Type		N (Households)	Mean Number of Long-Distance Tours per Respondent in a Year…				
			…in Total	…for Work	…for Leisure	…by Personal Vehicle	…by Air
Low Income	*No Children in the Household*	64	8.25	1.58	6.08	5.58	2.41
	Children in the Household	14	5.64	0.50	4.36	5.21	0.29
Medium Income	*No Children in the Household*	88	8.63	2.28	5.56	6.06	2.39
	Children in the Household	33	6.91	2.03	3.85	5.36	1.39
High Income	*No Children in the Household*	231	10.20	3.58	6.05	6.14	3.88
	Children in the Household	137	10.01	3.94	5.47	6.35	3.60
Total		567	9.39	3.08	5.67	6.05	3.18

Of note is the final sample sizes of each dataset. While there were 934 participants in the panel dataset as used in the other chapters of this dissertation, only 567 of those participants completed the entire survey for the year. This sample limitation was done to control for early response bias. While beginning this chapter's analysis, findings showed that there was an abnormal influx of travel occurring over the first few months of 2013 (January through mid-March in particular) which did not reflect literature review suggestions such as higher summer and end-of-year holiday travel. Investigation showed that this spike was due to survey drop-off: the final dataset kept all valid survey participants regardless of if they completed the entire yearlong survey, thus earlier year travel would have more instances than later year travel as survey burden resulted in respondent dropout as the year progressed. Since this chapter's goal is to investigate how best to limit the temporal scope of long-distance travel surveys, the focus is less on having as many instances of long-distance travel as possible and more on ensuring a respondent's entire annual long-distance travel behavior is present. Spot checking was also done to ensure these full panel participants truly reported their entire year of travel behavior.

4.1.2 Analysis Methods

Long-distance travel consists of a multitude of unique factors, influencers, and trends that define the variety of peoples' travel behavior. However, capturing all the nuances of long-distance travel is both difficult and time consuming. If a survey's main goal is to best inform data users on updated trends and behaviors, then the survey should be designed to best maximize its immediate usage. Yes, capturing the nuances of long-distance travel behavior for more niche travel purposes and modes supports research efforts and informs on emerging trends, but for this purpose, a full-scale annual survey is best suited to capture these subtilities. This chapter aims to measure the variability in long-distance travel volumes per travel day throughout a year. These findings will help highlight the need for considering long-distance trip seasonality in setting a travel survey's capture timeframe, as well as lay a path forward for exploring how to better capture long-distance travel. . As such, this analysis characterizes daily long-distance travel with five variables related to major travel days, trip purpose, and main mode usage as listed below. Note these variables were defined for each of the six defined sociodemographic groups creating a total of 30 long-distance travel behavior variables.

- Home-Based Trip Ends (HBTEs)
    - Defined as when any long-distance trip segment either begins or ends at the respondent's home on that travel day. Thus, for a given day, this is the summation of recorded long-distance trip segments with either the origin or destination listed as the respondent's home address.

- Percent of HBTEs Listed as Work Travel
    - Defined as the percentage of HBTEs for a calendar day where the associated long-distance trip whose listed purpose is work.
- Percent of HBTEs Listed as Leisure Travel
    - Defined as the percentage of HBTEs for a calendar day where the associated long-distance trip whose listed purpose is leisure.
- Percent of HBTEs Listing Main Mode as Vehicle
    - Defined as the percentage of HBTEs for a calendar day where the associated long-distance trip whose listed main mode is vehicle.
- Percent of HBTEs Listing Main Mode as Air
    - Defined as the percentage of HBTEs for a calendar day where the associated long-distance trip whose listed main mode is air.

Summarizing how HBTE rates vary by sociodemographic groups broadly illustrates the differences inherent in mean tripmaking volumes. These rates, shown in Table 18, demonstrate how varied long-distance travel is not just by sociodemographic group, but also trip purpose and main travel mode.

Table 18: Summary of Mean HBTEs for Days by Sociodemographic Group

Household Type		N (days)	Mean Number of HBTEs per Day Over a Year Timeframe…				
			…in Total	…for Work	…for Leisure	…by Personal Vehicle	…by Air
Low Income	*No Children in the Household*	365	2.89	0.55	2.13	1.96	0.84
	Children in the Household	365	0.43	0.04	0.33	0.40	0.02
Medium Income	*No Children in the Household*	365	4.16	1.10	2.68	2.92	1.15
	Children in the Household	365	1.25	0.37	0.70	0.97	0.25
High Income	*No Children in the Household*	365	12.92	4.54	7.65	7.78	4.91
	Children in the Household	365	7.52	2.96	4.11	4.77	2.70
	Total	365	4.86	1.59	2.93	3.13	1.65

While literature review findings support the focus on work/leisure purpose and vehicle/air mode splits as the majority of long-distance travel behavior, how best to identify travel activity for a given day was more difficult. Two approaches were considered: first a day's long-distance travel would consider all respondents currently leaving or currently on a long-distance trip (ex., if the capture day was a Saturday and two people were on a Friday to Sunday trip, one was leaving home on a long-distance trip, and three were returning home from a long-distance trip; then the total travel activity for that Saturday would be five long-distance trips). However, this method presented two problems: one, if a respondent is already on a long-distance trip, then the survey might only capture the information related to their destination, thus showing the respondent on an "island" (away from home without any idea of how, when, and where they arrived) for that travel day if no follow up is given. This is a common problem with long-distance trips captured in daily travel formats. The second problem with this approach, is it distorts the daily long-distance travel rates by falsely capturing high duration long-distance trips as daily travel burden.

The second approach for identifying travel activity for a given day considered only when a respondent was leaving or returning to their home. This method solved the issues listed in the other method while also relating long-distance travel back to daily travel survey capture. By identifying what days overnight long-distance travel begins or ends throughout the year, findings from this chapter could be compared to findings seen in daily travel surveys, such as the NHTS, for confirmation of this approach. Granted, the NHTS in its current format (pre-NextGen implementation) does not offer full long-distance travel details like a full scale long-distance longitudinal survey, but the beginnings and endings of long-distance trips could still be captured for comparison. However, most importantly, this approach helps identify the days where long-distance travel behavior is most likely to be captured by different sociodemographic groups; thus, illustrating how shorter recall surveys (like past 28-days or past month recall) fail to capture some sociodemographic groups' long-distance travel behavior.

While identifying variability in long-distance travel volumes per travel day throughout a year considering seasons and sociodemographics is the goal for this chapter, ensuring capture equity complicated this process. Based on previous research, certain days are more likely to experience heightened long-distance volumes such as major US holidays or weekends. The question then becomes determining if patterns, or clusters, for long-distance travel behavior exist among sociodemographic groups. To accomplish this, the K-means algorithm was used to determine the optimal number of clusters needed to characterize travel days. This step is defined as hard clustering, where each day could only be classified into a single cluster. K-means clustering was applied to the dataset iteratively using the 30 long-distance travel behavior variables (six sociodemographic groups each with five long-distance travel behavior variables) to graph a distortion score curve. This process allowed for the identification of the optimum number of clusters that both has the lowest distortion score (the average of the Euclidean squared distance from the centroid of the clusters) and lowest computational fit time. Figure 3 demonstrates the distortion score elbow (where the sum of squared distances begins to lessen with each additional value of *k* as well as minimize computation time) occurring around 8 clusters. This cluster size was then used for the remainder of this chapter.

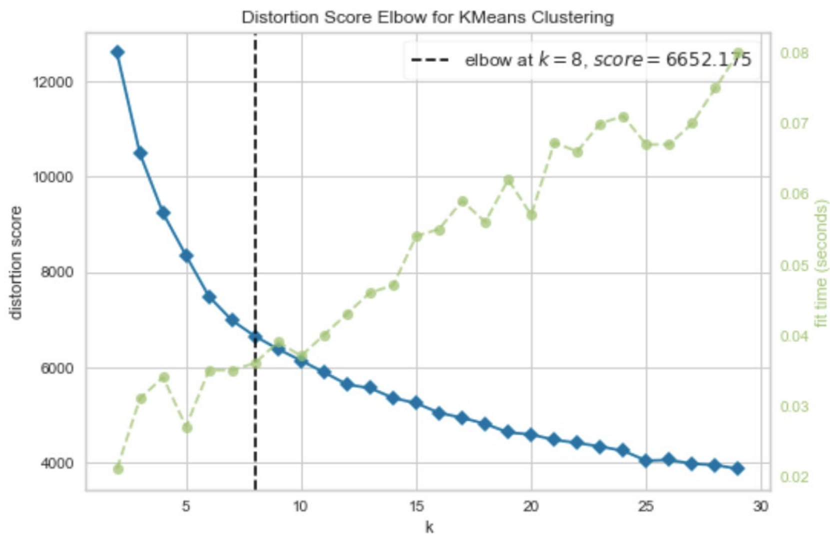


Figure 3: Distortion Score Curve (Blue Line Plots Distortion Score and Green Line Plots Algorithm Fitting Time)

4.1.3 Differences in Day-by-Day Long-Distance Travel Patterns Across Six Sociodemographic Groups

In order to appreciate the differences in day-by-day long-distance travel patterns across the six sociodemographic groups, HBTEs for each day and sociodemographic group were mapped to the 2013 calendar. This was done to a) illustrate the differences in travel behavior between sociodemographic groups, and b) highlight notable high long-distance travel periods such as peak days, weekends, and major holidays. Tables 20 through 26 break down HBTE volumes as follows:

- Table 20: Total HBTE Activity for All Sociodemographic Groups
- Table 21: HBTE Activity for Low Income, No Kid Households
- Table 22: HBTE Activity for Low Income, Kids Households
- Table 23: HBTE Activity for Medium Income, No Kids Households
- Table 24: HBTE Activity for Medium Income, Kids Households
- Table 25: HBTE Activity for High Income, No Kids Households
- Table 26: HBTE Activity for High Income, Kids Households

Table 19 acts as the legend for these tables with HBTE activity for a day binned into one of eight categories from no trip ends occurring that day, to more than 30 trip ends occurring that day. For additional context, major US federal holidays, such as extended weekends and major commercial/religious holidays are highlighted throughout the calendar year with specific dates provided below.

Table 19: Legend for Calendar Home-Based Trip End Activity and Highlighted Holidays

No Trip Ends	1 to 5 Trip Ends	6 to 10 Trip Ends	11 to 15 Trip Ends	16 to 20 Trip Ends	21 to 25 Trip Ends	25 to 30 Trip Ends	Greater than 30 Trip Ends

Holiday	Date (2013)
Martin Luther King Jr. Day	January 21
President's Day	February 18
Saint Patrick's Day	March 17
Easter	March 31
Mother's Day	May 12
Memorial Day	May 27
Father's Day	June 16
Independence Day	July 4
Labor Day	September 2
Columbus Day	October 14
Veteran's Day	November 11
Thanksgiving	November 28
Christmas Day	December 25
New Year's Eve	December 31

Table 20: 2013 Home-Based Trip End Activity by Day

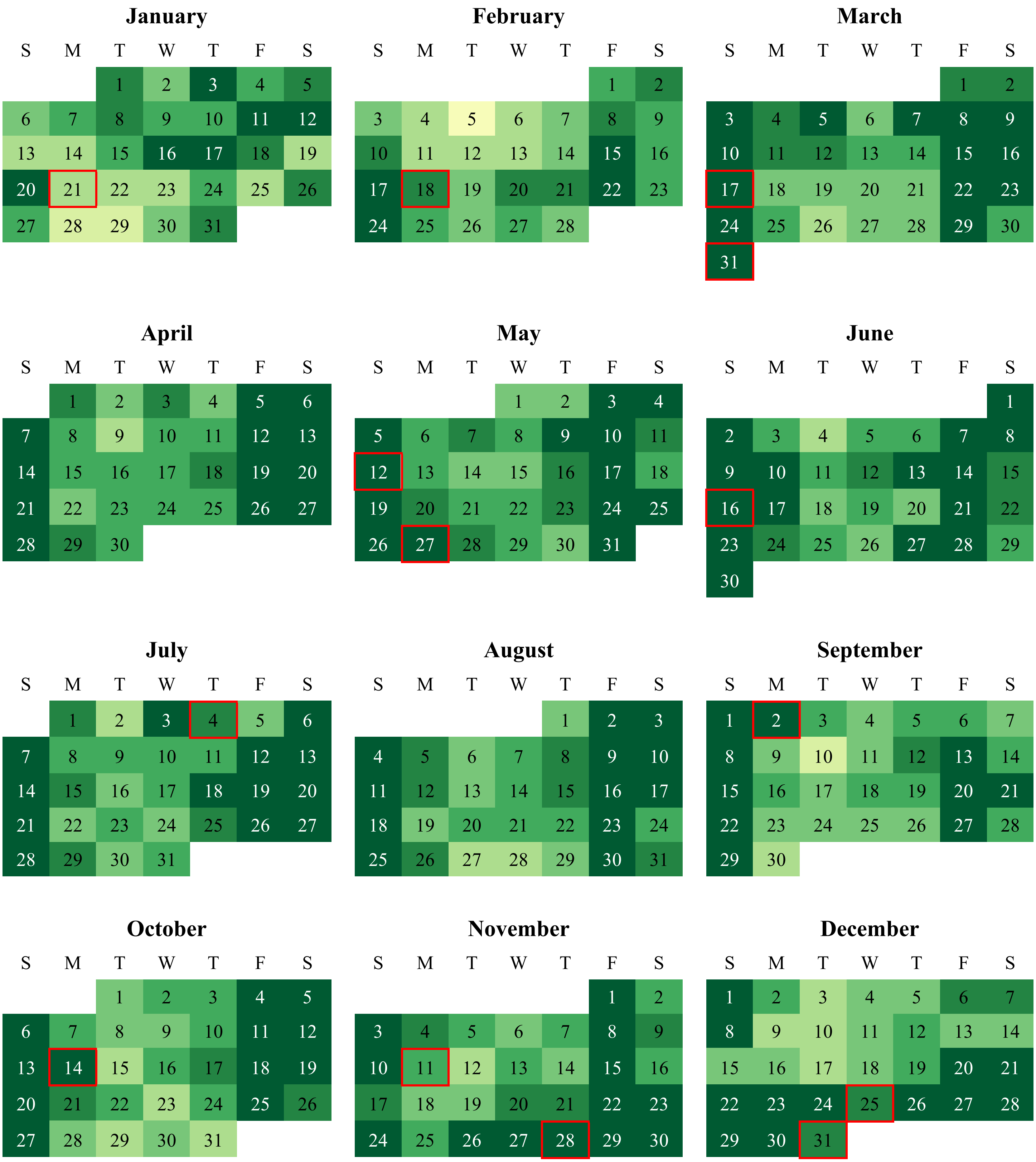


Table 21: Home-Based Trip End Activity for Low Income, No Kids Households

January

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

February

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

September

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

October

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

November

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

December

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Table 22: Home-Based Trip End Activity for Low Income, Kids Households

January

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

February

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

September

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

October

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

November

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

December

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Table 23: Home-Based Trip End Activity for Medium Income, No Kids Households

January

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

February

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

September

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

October

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

November

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

December

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Table 24: Home-Based Trip End Activity for Medium Income, Kids Households

January

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

February

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

September

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

October

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

November

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

December

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Table 25: Home-Based Trip End Activity for High Income, No Kids Households

January

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

February

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

September

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

October

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

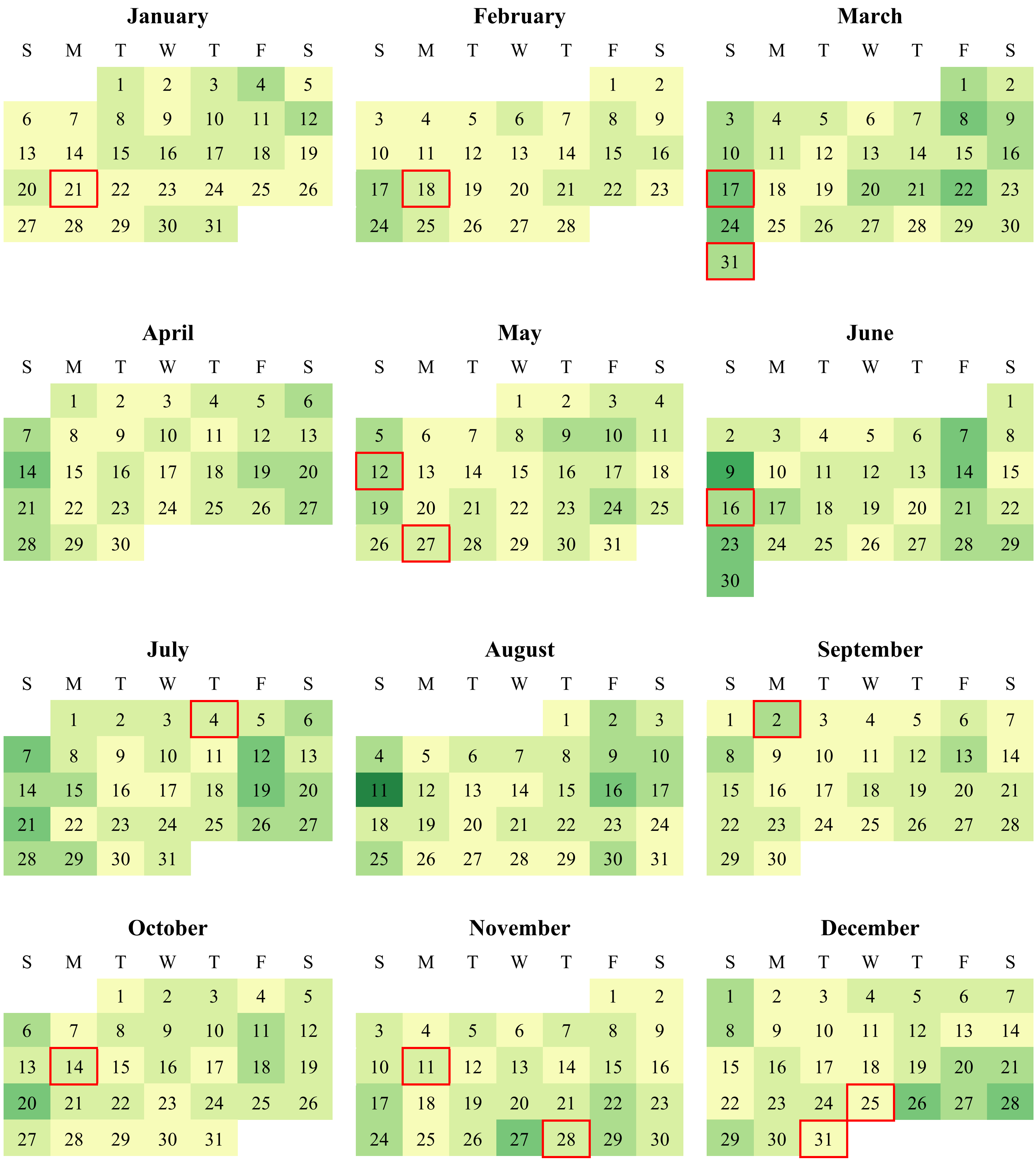
November

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

December

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Table 26: Home-Based Trip End Activity for High Income, Kids Households



Mapping HBTEs for all annual travel and each sociodemographic group highlights several major patterns. First, long-distance travel activity volumes orbit mainly around weekends and major US holidays. The months of January, February, and September (barring Labor Day Weekend) showcased the least amount of travel overall, while the summer months (May through August) and the latter halves of November and December (Thanksgiving, Christmas, and News Years) had the densest travel activity. Regarding weekend travel, the typical weekend showed an increase in HBTEs from Friday through Sunday, with pure HBTE volumes showing that Friday and Sunday exhibited higher HBTE volumes than Saturday. However, when a US federal holiday created an extended weekend, such as with Memorial Day or Labor Day, the HBTE volumes for those weekends shifted to a heavy Friday and Monday travel volume presence. These results not only reenforce literature review findings and common planning knowledge, but also show how individuals mainly perform long-distance travel over weekends, and more importantly, extended holiday weekends. The other holiday travel activity of note would be July 4th and Christmas: holidays that, unlike Thanksgiving, are defined by a date rather than a specific day of a month (Thanksgiving is always the fourth Thursday of November, while July 4th and Christmas shift the day of the week year-to-year). Here, Long-distance travel volumes appear to reflect how respondents receive holiday leave or utilize paid time off. For July 4th, there is high HBTE activity occurring on July 3rd (a Wednesday) and high (but not as great) HBTE activity occurring on July 4th. While it is known July 4th is major travel holiday, what is interesting about this is how HBTE activity is observed immediately after Thursday, July 4th. Calendar results would suggest that travel on July 5th, a Friday, were rather low, and that the following Saturday and Sunday saw high HBTE activity. This suggests that holidays that fall during the middle of the week may cause respondents to combine these normally excused work holidays with the closest weekend to extend their leisure activities. Therefore, creating situations, like transit and parking lot arrival terminology, of batch arrival or random departure for the preceding/receding days to a mid-week holiday. This would still need to be confirmed from other travel surveys where the holiday is on another weekday. For Christmas, it appears that HBTE activity is rather steady throughout the entire week, barring Christmas Day, illustrating a more random arrival/departure type structure. This may reflect the more liberal corporate holiday leave schedules instituted by different businesses or respondents' willingness to try and maximize holiday travel around this time.

Breaking HBTE activities up based on sociodemographic groups created Tables 20 through 25. These results highlighted similar findings regarding weekend and holiday travel as previously discussed, but also showcased the influence household income and the presence of children in the household has on long-distance travel activity. It is well known both these variables are highly influential in long-distance travel behavior, but these tables visually illustrate how higher income individuals partake in more long-distance travel and that the presence of children, regardless of income level, reduces the volume of long-distance travel. Households with children appear to produce higher HBTE activity around major travel holidays and the occasional weekend, while households without children also showcase similar trends, but with more weekday travel appearing, Additionally, travel volumes for households with children tend to be more active in the summer months suggesting a higher propensity to travel when the child is not in school. For weekday travel volumes, respondents without children in their household had higher weekday HBTE activity volumes than those with children. This trend also increased with income level. This could be a reflection of the nature of the respondents' job requiring travel, and that a respondent without children, or higher income, may be more willing to partake in a job that requires frequent out-of-town travel to complete.

Overall, mapping long-distance travel activity as HBTE volumes per day for a year offers an interesting and informative take on understanding long-distance travel. Results illustrate how long-distance travel fluctuates not only by the day of the week, but also seasonally and based on where major travel holidays occur during a week. Additionally, long-distance travel behavior varies greatly based on a respondent's sociodemographic characteristics. All of this supports proponents who call for a yearlong panel long-distance travel survey to best capture travel behavior equitably. For certain these findings highlight the haphazardness of asking a respondent for their long-distance travel activity for a random time period. However, there is still evidence from these findings that suggest through smart targeted sampling, the number of days needed to capture an individual's annual long-distance travel behavior can be reduced. This includes how most respondents travel around major holidays, weekends, and most importantly, not normally travel on weekdays. By identifying the underlying patterns associated with long-distance travel activity—be it overall travel, mode split, or trip purpose—future surveyors may be able to capture an entire year's worth of long-distance travel behavior, equitably, without sacrificing the data fidelity associated with yearlong survey formats.

4.2 Clustering Travel Days Based on Long-Distance Travel Volumes for Sociodemographic Groups, Modes, and Purposes

The second task seeks to determine how long-distance travel days can be clustered such that a) the long-distance travel volumes (reflecting the range of sociodemographic groups) are similar for days *within each group,* and b) the travel volumes are different *between groups*. In this task, it is important to combine the travel patterns across different sociodemographic groups.

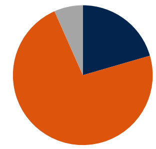
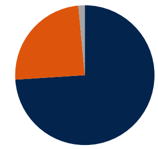
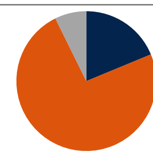
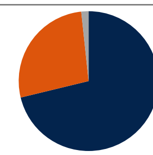
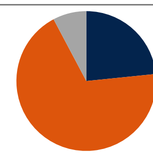
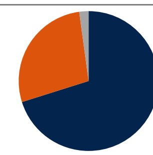
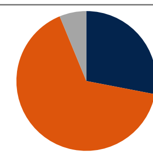
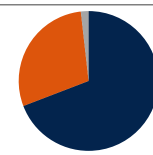
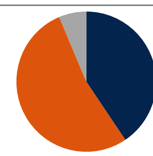
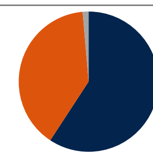
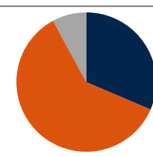
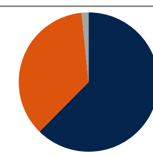
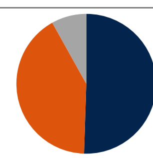
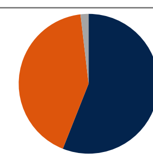
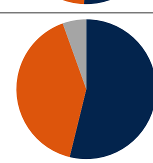
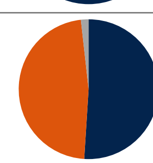
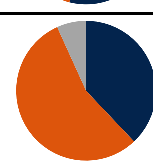
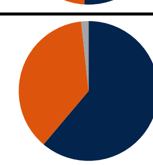
Travel days for the 2013 LSOT were categorized into eight clusters based on the number of HBTEs, trip purpose percentage splits, and trip mode percentage splits for each of the six sociodemographic groups resulting in a total of 30 independent variables. Cluster results were then ordered by descending average HBTEs creating four major Travel Activity Groups (TAGs): extreme travel activity, high travel activity, medium travel activity, and low travel activity; with the higher activity clusters showcasing major travel holidays and weekends. These TAGs were first determined based on the general breaks in the *overall* mean number of HBTEs, but within each cluster there were distinct variations in the mean percentages of trips completed by different modes and for different purposes. Table 27 presents summarized trends of each cluster for each activity categories. The purpose and mode split pie charts were created by calculating the mean purpose and mode splits for all the days contained within each cluster. TAGs can be summarized as:

- **The Extreme Travel Activity Group (43 Days)**
    - This TAG consisted of heavy leisure and personal vehicle percentages for the majority of sociodemographic groups. Specifically, this TAG had the highest personal vehicle percentage for 4 of the 6 sociodemographic groups with the exceptions being "Low Income, No Kids"; and "Medium, Kids". Additionally, 5 of 6 sociodemographic groups had their greatest leisure percentage in this TAG with the exception of "Medium Income, Kids". Overall sociodemographic trends also highlighted consistently low work travel capture in this TAG. High income groups and the "Low Income, No Kids" group exhibited their greatest mean daily HBTE activity within this TAG.

- **The High Travel Activity Group (79 Days)**
    - This TAG consisted of heavy leisure and personal vehicle percentages for all sociodemographic groups but with more focused capture on low and medium income travel. Captured work travel was low for all sociodemographic groups but was greater than captured in the Extreme Travel Activity Group. Additionally, the "Low Income, No Kids" and both medium income groups exhibited their greatest mean daily HBTE activity within this TAG.
- **The Medium Travel Activity Group (160 Days)**
    - Generally, work and air travel capture was greater for this TAG compared to the Extreme and High Travel Activity Groups. This TAG exhibited the highest work and air travel capture for both low income groups, as well as the highest work travel capture for the "Medium Income, No Kids" group. Mean daily HBTE activity was low for all sociodemographic groups, with this TAG showcasing the lowest mean daily HBTE activity for both low income groups; the "Medium Income, No Kids" group; and the "High Income, Kids" group.
- **The Low Travel Activity Group (83 Days)**
    - This TAG exhibited the overall lowest mean daily HBTE activity with both the "Medium Income, Kids" and "High Income, No Kids" groups having their lowest mean daily HBTE activity captured here. This TAG also had the most zero travel activity days (daily sociodemographic group HBTE equals zero) of any TAG/cluster. Additionally, this TAG consisted of the greatest work travel capture percentage for both high income groups and the "Low Income, No Kids" group.

The greatest air travel capture percentage for all medium and high income groups was also represented in this TAG.

Table 27: Average Cluster Characteristics Organized by Home-Based Trip Ends

Cluster Description		Days in Cluster	Average Home-Based Trip Ends	Average Purpose Split (■ Work ■ Leisure ■ Other)	Average Mode Split (■ Vehicle ■ Air ■ Other)
Extreme Activity	E1	15	51.73		
	E2	28	50.21		
High Activity	H1	39	40.77		
	H2	40	38.43		
Medium Activity	M1	68	25.78		
	M2	47	25.55		
	M3	45	23.00		
Low Activity	L1	83	16.24		
Annual-Level		365	29.17		

Tables 28 and 29 provide further insight into the distribution of these clustered days throughout the year by showcasing the number of days for in every cluster each month by count and visually, respectively. Generally, as the mean number of daily HBTEs decreased per cluster group, the percentage of trips identified as being for leisure or completed by vehicle also decreased. Additionally, mapping cluster membership to the 2013 calendar showed a pattern reflecting this decrease in leisure and vehicle travel as higher travel activity clusters reflected more holiday and weekend travel, and the lower travel activity clusters captured more weekday, work, and air travel days.

Table 28: Distribution of Cluster Days by Month

Month	**Days in Cluster**								**Total Days in Month**
	E1	**E2**	**H1**	**H2**	**M1**	**M2**	**M3**	**L1**	
Jan	0	0	2	2	5	7	4	11	**31**
Feb	0	0	1	3	3	5	3	13	**28**
Mar	3	2	4	4	9	1	3	5	**31**
Apr	1	1	6	3	5	4	6	4	**30**
May	0	4	0	6	8	5	4	4	**31**
Jun	3	2	3	4	8	4	4	2	**30**
Jul	3	2	6	2	9	1	3	5	**31**
Aug	2	3	5	2	3	4	2	10	**31**
Sep	0	3	3	3	3	4	8	6	**30**
Oct	1	3	1	5	8	2	3	8	**31**
Nov	1	4	4	4	5	5	2	5	**30**
Dec	1	4	4	2	2	5	3	10	**31**
Total	**15**	**28**	**39**	**40**	**68**	**47**	**45**	**83**	**365**

Table 29: 2013 Calendar of Daily Cluster Membership (Major Holidays Highlighted in Black)

January

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

February

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

September

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

October

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

November

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

December

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Breaking down by cluster activity group, days falling under the extreme levels of travel activity, such as E1 and E2, represented mainly weekend and holiday travel. In fact, 87 percent of days classified as E1 fell on Sundays or Fridays with the only one day in this group falling on a weekday: the Wednesday before Thanksgiving. The days in this cluster also are predominantly in the summer months (June, July, and August) with 53 percent of cluster days showing such. For days in cluster E2, similar patterns are shown with the majority of days occurring on Fridays and Sundays and the only two weekday cluster members being July 3rd (a Wednesday) and the day before New Year's Eve, December 30th (a Monday). Instead of most of these days falling on summer months, the majority in this cluster occur in the Fall months (September, October, November). Overall, these extreme travel activity clusters reflect the highest travel periods for weekends and day-before holiday travel for holidays that usually do not warrant extended time off (such as common for the week of Christmas).

Clusters labeled in the high travel activity group—H1 and H2—contrasted slightly from the extreme activity group by representing the mainly other weekend days throughout the Spring, Summer, and Fall months. Notably, both Memorial Day (Monday, May 27th) and Labor Day (Monday, September 2nd) were classified in this group. For the H1 cluster, these days were mainly in the summer months and covered mainly Saturdays and Sundays. Only two weekdays were included in the cluster, the aforementioned Labor Day Monday, and the Thursday after Christmas Day (December 26th). In fact, the Thursday and Friday following Christmas Day both were categorized in the H1 cluster. The greatest concentration of H1 cluster members were the Fridays and Saturdays of the months of April, July, and August. April activity may reflect "spring fever" travel as weather begins to steadily improve, while the July and August weekends signifying the height of summer vacation travel. For days in cluster H2, the majority of members were Fridays concentrated in the spring months, however, this cluster was the first cluster type to show a higher concentration of Monday through Thursday members. These weekday cluster members tended to be the immediate days prior to a major holiday or other extreme travel day such as the Tuesday before Thanksgiving and the Monday and Tuesday before Christmas Day (including Christmas Eve). This suggests that this cluster captured the holiday travel associated more closely with holidays, such as Thanksgiving and Christmas, which have a higher likelihood of extended time off, allowing travel for those holidays to occur over a greater number of days rather than all at once. For this cluster, there was also one concentration of days occurring the week March 4th through 8th. Since the LSOT dataset has a slight bias towards university employees and students, this might be a reflection of spring break travel.

The next travel group were those clusters deemed of medium travel activity, clusters M1, M2, and M3. This group exhibited a higher percentage of members being weekdays (Monday through Thursday), with M3 consisting of 89 percent weekday members. Like the high travel activity groups, the medium activity groups concentrated around weekends of lesser travel volumes such as weekends in January and February, or days immediately surrounding higher travel activity groups such as summer weekdays or preceding/following major holidays. If one thinks of travel activity as a bell curve with E1 and E2 being the peak, the medium travel activity group can be seen as making up the beginnings of the curve tails. All three medium activity group clusters had the majority of their members days listed in the non-summer months. For M1, this was concentrated in Spring, M2 was concentrated in Winter, and M3 was concentrated across both Spring and Fall months. It should be noted that M3 is the first cluster to see the major purpose for its travel as work travel, albeit just scratching past the majority. Additionally, the holidays of MLK Day, July 4th, Veteran's Day, Thanksgiving Day, Christmas Day, and New

Year's Eve all classified into one of the medium travel activity clusters. This could reflect the lower travel associated with the actual holidays rather than the immediate days leading up to and following these major travel holidays. As individuals are more likely to already be at their holiday destination *on* the holiday, then the amount of travel actually occurring on the holiday would most likely be lower.

Finally, the low travel activity group consisted of one cluster type, L1, and its members made up the majority of low traveling weekdays at 88 percent, but also captured the most work travel. While some weekends did include L1 members, these weekends were either low travel colder months (as seen in January and February), or the weekends preceding major holiday travel (such as in mid-November and December) where individuals may be limiting long-distance travel in anticipation of a planned long-distance trip for the major winter holiday season. All-in-all, this cluster group captured the lulls in long-distance leisure travel throughout the year mainly reflecting around work travel occurring during the weekdays.

4.3 Conclusions

This chapter explored trends in the annual variability of long-distance travel trends. Long-distance travel often represents non-routine travel, such as holidays, vacations, and the occasional work conference. As such, much discussion has focused on the need to collect travel survey data for a year to fully capture the patterns of different sociodemographic groups who would be missed if only surveyed for a shorter period of time.

However, due to high costs associated with such a survey effort, it is beneficial to review trends in daily long-distance travel across these sociodemographic groups. Therefore, this research measured the variability in long-distance travel volumes per travel day throughout a year considering seasons and sociodemographics. Specifically, this chapter determined a) if there are differences in the day-by-day long-distance travel patterns across six unique sociodemographic groups across a year and b) how long-distance travel days can be clustered such that the long-distance travel volumes (reflecting the range of sociodemographic groups) are similar for days within each group *and* the travel volumes are different between groups.

These tasks were accomplished through K-means clustering days based on 30 variables considering five types of long distance HBTEs by six different sociodemographic groups. Eight clusters were identified that characterize days with extreme long-distance travel demand to days with low demand. Even though they are not part of the clustering process, the results highlighted how much these volumes are tied to seasonal travel needs and holidays.

Recalling the objectives and hypotheses of this dissertation, the findings of this chapter support the idea that long-distance travel behavior might be able to be adequately captured using a subset of representative days, even accounting for differences in travel by different sociodemographic groups. This means surveys could be focused to collect travel diaries for specific days, and travel trends for certain days could be transferrable to other similarly clustered days.

Of course, there are many opportunities for further work to explore these relationships deeper. For example, holidays and seasonal impacts vary greatly across the country, so geographic variations could be considered. Additionally, more data on different sociodemographic groups could determine if there are more unique trends not captured with the current dataset. Finally, bootstrap simulations selecting different survey days from the LSOT to determine accuracy of travel forecasts for different clusters could be completed to understand implications for minimum sample sizes. It is also suggested that this chapter's methods be applied to the upcoming NextGen NHTS for similarities in results.

Chapter 5: Targeting Respondent Demographic Groupings

The second major objective of this dissertation considers how to best target sociodemographic groupings in long-distance travel survey sampling. The traditional approach to survey sampling, particularly with travel survey sampling, is to employ a population proportion approach. This method considers the sociodemographic distributions for a defined area, and then randomly samples based on these distributions to ensure a representative sample of the inhabitants of an area. The sample is then weighted back to reflect the actual population distributions. While this approach provides a valid sampling frame, it is costly and requires a vast amount of resources to appropriately implement, putting outside the reach of the average researcher or planning organization.

An alternative to this approach is to employ a method collectively known as probability-based sampling. This sampling method considers patterns inherent to a field and adjusting the sampling frame to best capture the desired behavior efficiently and equitably. In the case of long-distance travel, it is known certain groups generate higher volumes of travel (such as higher income households and households with less children) than other groups. By understanding which groups are more *and* less likely to be generating long-distance travel, the sampling frame can be adjusted to maximize capture of rarer travel groups (such as lower income households) while reducing capture of high travel likelihood groups to ensure sample equity for a smaller-scale travel survey sample.

Therefore, this chapter aims to identify how sociodemographic groups can be best targeted to minimize surveying burdens yet remain equitable. This is accomplished by a) identifying common sociodemographic groupings based on previous literature findings and observed trends in previous long-distance travel surveys, b) validating these groupings in terms of statistical independence, and c) determining the minimum required sample sizes needed to effectively replicate previous long-distance travel surveys. Findings and methods from this chapter are not only used throughout this dissertation (as seen in chapters three, four, and six), but can also aim researchers, planners, and decisionmakers towards how best to capture long-distance travel effectively and equitably when faced with minimum funding or other hurdles preventing a full-scale population proportioned sampling approach.

This chapter is organized as follows: first, a brief review of the three long-distance travel surveys used for analysis (the 1995 ATS, 2001 NHTS, and 2013 LSOT) is presented. This is followed by the applied methodologies: the universal household definitions, outlier identification, ANOVA, Theil index, and minimum sample sizes. Results are presented next for each of the previously applied methodologies. These results not only validate the universal household definition methodology employed throughout this definition, but also identify how long-distance travel surveys can be greatly reduced in sample size without sacrificing data fidelity. Finally, conclusions are summarized explaining key findings of the analyses as well as where future work should focus on.

5.1 Long-Distance Travel Survey Data Review

For this chapter, three unique long-distance travel surveys were used to best capture and compare trends across different definitions, collection methods, and timelines. The three datasets span nearly a 20-year period from 1995 to 2013 which captures the changes in long-distance travel throughout the years. Two datasets, the 1995 ATS and the 2001 NHTS, were conducted by the United States Bureau of Statistics (BTS) reflecting a national focus. The third dataset, the 2013 LSOT, was conducted by Auburn University and University of Vermont researchers and has a tighter sample window (a noted bias of a higher household income sample) with the majority of respondents based out of Alabama, California, and Vermont. Table 30 compares the datasets.

Table 30: Dataset Characteristics

	1995 ATS	2001 NHTS	2013 LSOT
Conducted By	BTS	BTS	Auburn and Vermont Researchers
Definition of LONG-DISTANCE Travel	100 Miles One-Way	50 Miles One-Way	Overnight Tour Element
Geographic Scale	United States	United States	Focused on Alabama, Vermont, and California
Timeframe	1995 – 1996	2001 – 2002	2013 – 2014
Survey Type	Year-Long Panel	28-Day Time Period	Year-Long Panel
Household/Individual/Trip Relationship	Defined Trip and Household IDs	Can collapse by Household ID, but No Trip ID to link members	No Specific Household ID; Individual-centric
Known Biases	Respondent Recall	Shorter Time Period; 9/11	High Income Individuals
Total Number of Households/Persons	80,000	63,163	1,024
Number of Usable Households/Persons	48,527	12,241	1,013
Number of Usable Tours	337,520	45,165	8,367

The most notable differences in the datasets are the definition of long-distance trips and the collection time period. The definition of a long-distance trip has been debated for years and these three survey definitions reflect this debate. As mentioned in the literature review and in chapter four, long-distance travel sees fluctuations throughout the year. It is quite seasonal. As such, panel surveys are recommended to best capture a typical household's long-distance travel trends. This is the case for both the ATS and LSOT, but the 2001 NHTS adopted a 4-week travel period to provide "…information on a larger sample of long-distance trips than the [Nationwide Personal Transportation Survey] NPTS and better recall of trips than [the] ATS…" *[8]*. The 2001 NHTS did acknowledge the limitations of the 4-week travel period versus a comparable panel survey.

Table 29 also lists each dataset's number of usable households/persons and number of usable tours. These values were created from the original public files by removing all cases where the household income, respondent age, or trip distance was either unknown or under 50-miles in linearly measured distance one-way. This was done as part of the normal data cleaning procedure, and in the case of the 50-mile minimum, removed outlier trips (mainly from the LSOT).

5.2 Dataset Comparison Methodology

This section breaks down the applied household category system as well as the three statistical methods applied to each dataset: ANOVA, Theil Index, and sample size calculations. Each subsection goes over the methods, logic, and interpretation for the given aspect. The methodology used in this chapter is applied throughout this dissertation namely in chapters three, four, and six.

5.2.1 Household Definitions

A universal problem big data faces is the inconsistencies with variable definitions. While two datasets can have the same variable, such as household income, how it is logged in the dataset can vary greatly. Each of the aforementioned datasets used in the analysis categorized household income differently—some categorized in \$5,000 increments, and others categorized in varying ranges. Coupled with average inflation and varying definitions of poverty, the datasets cannot be immediately normalized and consist of a very wide range of subgroups. To best compare the datasets to themselves and each other, a universal household categorical system was created utilizing three known factors affecting long-distance and intercity travel volumes: household income, respondent age, and the presence of children in the household. Each category was broken down as such:

- Household Income (income range varies, defined in the following paragraph):
    - Low
    - Medium
    - High
- Age Group of Respondent or Oldest Member in Household:
    - Under 25 years old
    - 25 to 44 years old
    - 45 to 64 years old
    - 65 years or older
- Household Children:
    - Yes, at least one household individual is under 18 years of age
    - No individuals under the age of 18 are present in the household

While more demographics would be preferred, these groups were selected to ensure that a) each dataset was able to represent most if not all the groups and b) there were adequate sample sizes to make conclusions about travel variation for each group. Each household was then assigned into one of 24 combined categories using the income level, age group, and presence of children (ex. low income, age 25 to 44, and no children; would be one category).

As income levels varied by year, the decision was made to create a universal low, medium, and high-income level band system using the specific year's median income and poverty threshold for an average four-person household. Income data was pulled directly from US Census estimates for the specific year [161-163]. The Equation 1 below defines the definition of each income level.

$$Income\ Level = \begin{cases} \boldsymbol{low} < \tilde{x}_{inc} - \frac{\tilde{x}_{inc} - PT}{2} \\ \tilde{x}_{inc} - \frac{\tilde{x}_{inc} - PT}{2} \leq \boldsymbol{medium} \leq \tilde{x}_{inc} + \frac{\tilde{x}_{inc} - PT}{2} \\ \boldsymbol{high} > \tilde{x}_{inc} + \frac{\tilde{x}_{inc} - PT}{2} \end{cases} \tag{1}$$

Where $\tilde{x}_{inc}$ is the median household income and PT is the poverty threshold.

The income level threshold values were further rounded to the nearest $5,000 increment to better match the datasets. The exception to this was for the 2013 LSOT dataset where the threshold values were rounded to the nearest household income level categorical cutoff value. Table 31 shows the income level values and the resultant thresholds used for this analysis.

Table 31: US Census Annual Household Income Estimates and Income Level Band Calculations

		1995	2001	2013
Base Income Statistics	*Median*	$34,076	$42,228	$51,939
	Poverty Threshold (Average 4 Person Household)	$15,569	$18,104	$14,053
Income Level Band Calculations	*Band Size*	$9,254	$12,062	$14,053
	Low	$24,823	$30,166	$37,887
	High	$43,330	$54,290	$65,992
	Low (Rounded to Nearest Dataset Category)	$25,000	$30,000	$50,000
	High (Rounded to Nearest Dataset Category)	$45,000	$55,000	$75,000

This method was also repeated using thresholds found by adding/subtracting two bands from the median (similar to a 2 standard deviation approach). This process found a much more uneven distribution of households, especially with the low-income category, and was thus abandoned in favor of the previously discussed single band approach to improve group sample sizes.

5.2.2 Outlier Identification

With all datasets, extreme outliers threaten the integrity of statistical analysis, and as such, must be identified and removed. The approach used in this case was both standard and unique. Rather than remove an entire case if a single outlier was found, each relevant analysis variable (trip rate, trip duration, roundtrip distance, etc.) was subjected to its own detection procedure with outlier cases removed only from that variable subset. Outliers were defined as any value 1.5 times the interquartile range away from the lower or upper quartiles using Equation 5:

$$Outlier = \begin{cases} Value < \ Q_{25} - 1.5 * IQR \\ Value > \ Q_{75} + 1.5 * IQR \end{cases} \tag{5}$$

Where Q_{25} is the 25 percent quartile value, Q_{75} is the 75 percent quartile value, and IQR is the interquartile range value.

An example of this application would be if an individual completed one, 10,000-mile roundtrip long-distance trip during the entire survey period. Results show that one trip falls within the normal number of long-distance trips completed for this survey, but the 10,000-mile roundtrip is flagged as an outlier case. Therefore, this individual's long-distance trip would be included in the analysis of long-distance travel volumes but would *not* be included when analyzing long-distance travel distance. This process maximizes the available long-distance travel data for one analysis while also reducing statistical noise for another.

5.2.3 ANOVA

To determine if the aforementioned household classifications characteristics sufficiently captured differences in long-distance travel, the analysis of variance (ANOVA) approach was used. This approach allows researchers to study the effect that one or more categorical variables has on a continuous outcome using each groups' variance. If the result of the test is statistically significant (tested at a 95 percent confidence level), then the chosen categories are sufficient at explaining data variation. This method does make three assumptions: all sample values are independent from each other, there is a linear relationship between variables, and the variance of within each group is similar.

5.2.4 Theil Index

The Theil index is traditionally used to measure economic equality [167], but it can also be an effective way to measure equality among any number of things. The Theil index is utilized in two different scenarios: the Theil T index is more sensitive towards the high-end (rich) of a distribution, while the Theil L index is more sensitive towards the low-end (poor). Both measure the entropic "distance" a population is from the equalitarian state [167]. For this analysis, the Theil T index, shown below as Equation 6:

$$T_T = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\mu} ln\left(\frac{x_i}{\mu}\right) \tag{6}$$

Where T_T is the Theil T index, N is the population, μ is the mean of characteristic x, and x_i is the characteristic value of individual i.

The Theil T index can have a value from 0, perfect equality, with the value steadily rising to represent greater inequality.

5.2.5 Sample Size

Sample size was calculated for each dataset and its household categories to determine the minimum number of households needed to confidently match the observed means and standard deviations. While not particularly interesting on a whole-dataset level, when applied to the household group level, focus areas begin to emerge. For this application, as well as seen in chapter four, Equation 4 was used:

$$n = \left(\frac{z * \sigma}{e}\right)^2 \tag{4}$$

Where n is sample size, z is z-score, σ is the standard deviation, and e is the margin of error.

Sample sizes were calculated using post-cleaned data (all outliers were removed). Final sample sizes were corrected for expected outliers by applying a percentage increase of observed outlier cases.

5.3 Results

The following section presents results for dataset trends, group comparisons, sample size calculations, and summarized actions. Subsections 5.3.1 and 5.3.2 cover household group breakdowns, outlier calculations, and notable trends found between trip volumes, median roundtrip distances, median trip durations, and percentage of leisure trips. Subsections 5.3.3 and 5.3.4 talk in detail the results of the ANOVA and Theil T Index calculations. Finally, subsection 5.3.5 discusses the minimum sample size calculations for each survey and sociodemographic group.

5.3.1 Dataset Distributions

Once datasets were cleaned, each dataset was divided into the 24 household groups discussed in the methodology section (household income, age group, and presence of children). Initial results showed several groups lacking a minimum of 50 instances, thus deeming these groups too small to be statistically sufficient for this analysis. In response, the children/no children subgroups for "65 or Older" respondents were collapsed into single groups, leaving 21 unique sociodemographic household groups. Table 32 displays the raw household group breakdown for each dataset. Groups with less than 50 instances are bolded.

Table 32: Group Sample Sizes (Prior to Outlier Removal)

			ATS	LSOT	NHTS
LOW INCOME	**Under 25**	*No Children*	740	-	157
		Children	229	-	159
	25 to 44	*No Children*	1,760	101	268
		Children	1,866	**19**	356
	45 to 64	*No Children*	2,483	**29**	497
		Children	448	**6**	168
	65 or Older	*All*	4,569	**8**	715
MEDIUM INCOME	**Under 25**	*No Children*	622	-	56
		Children	145	-	174
	25 to 44	*No Children*	3,372	70	416
		Children	5,399	**38**	731
	45 to 64	*No Children*	5,169	64	804
		Children	1,542	**22**	318
	65 or Older	*All*	3,248	**11**	595
HIGH INCOME	**Under 25**	*No Children*	216	**1**	**42**
		Children	**49**	-	193
	25 to 44	*No Children*	2,295	110	661
		Children	4,955	117	1,483
	45 to 64	*No Children*	5,339	206	1,601
		Children	2,258	99	1,083
	65 or Older	*All*	1,780	**34**	457
TOTALS			48,484	935	10,934

While the ATS and NHTS have better representation across all groups, all three datasets favored higher income groups. As these households tend to complete long-distance trips more frequently, this was expected. The LSOT had the smallest representation for households in the

“Under 25” age groups, however, proportionally, this was in line with the other datasets which saw no more than two percent representation in any “Under 25” group.

Using these sociodemographic groups, each dataset was subjected to an intensive review for pattern recognition regarding trip rates, maximum/minimum/median roundtrip distance, maximum/minimum/median trip duration, and business/leisure trip splits. This process was completed with sub-datasets composed of vehicle-only trips and air-only trips. Using probable and cumulative distribution function charts, as well as boxplots, each of the previous variables were manually inspected across all three datasets for pattern recognition for a total of 243 unique plots. After careful consideration, the most notable trends were found among the overall trip groups (all trips regardless of mode) for trip rates, median roundtrip distance, median trip duration, and percentage of leisure trips. Figures 4 and 5 show the trip rate cumulative distribution plots and boxplots.

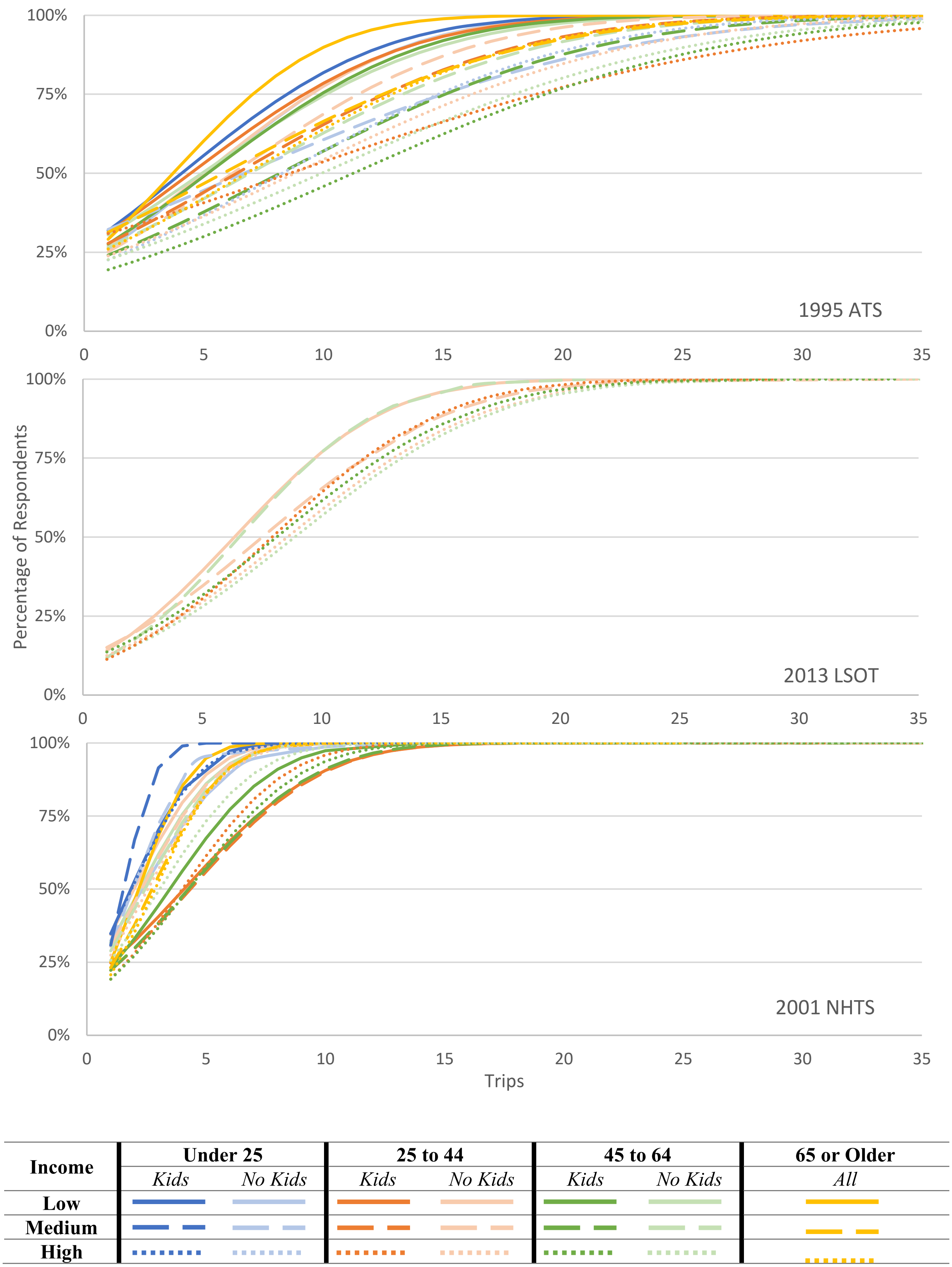
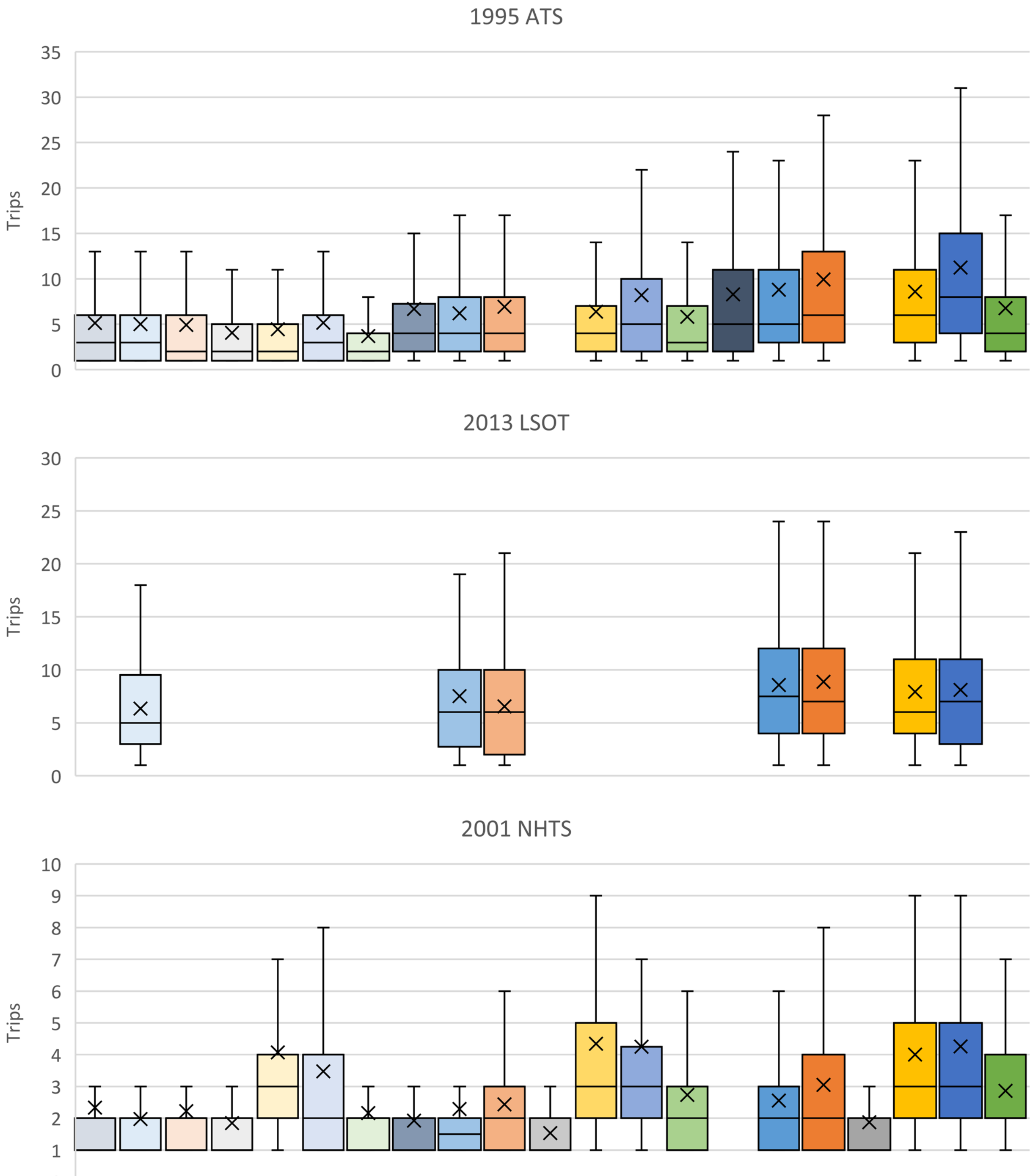


Figure 4: Trip Rate Cumulative Distribution Plots

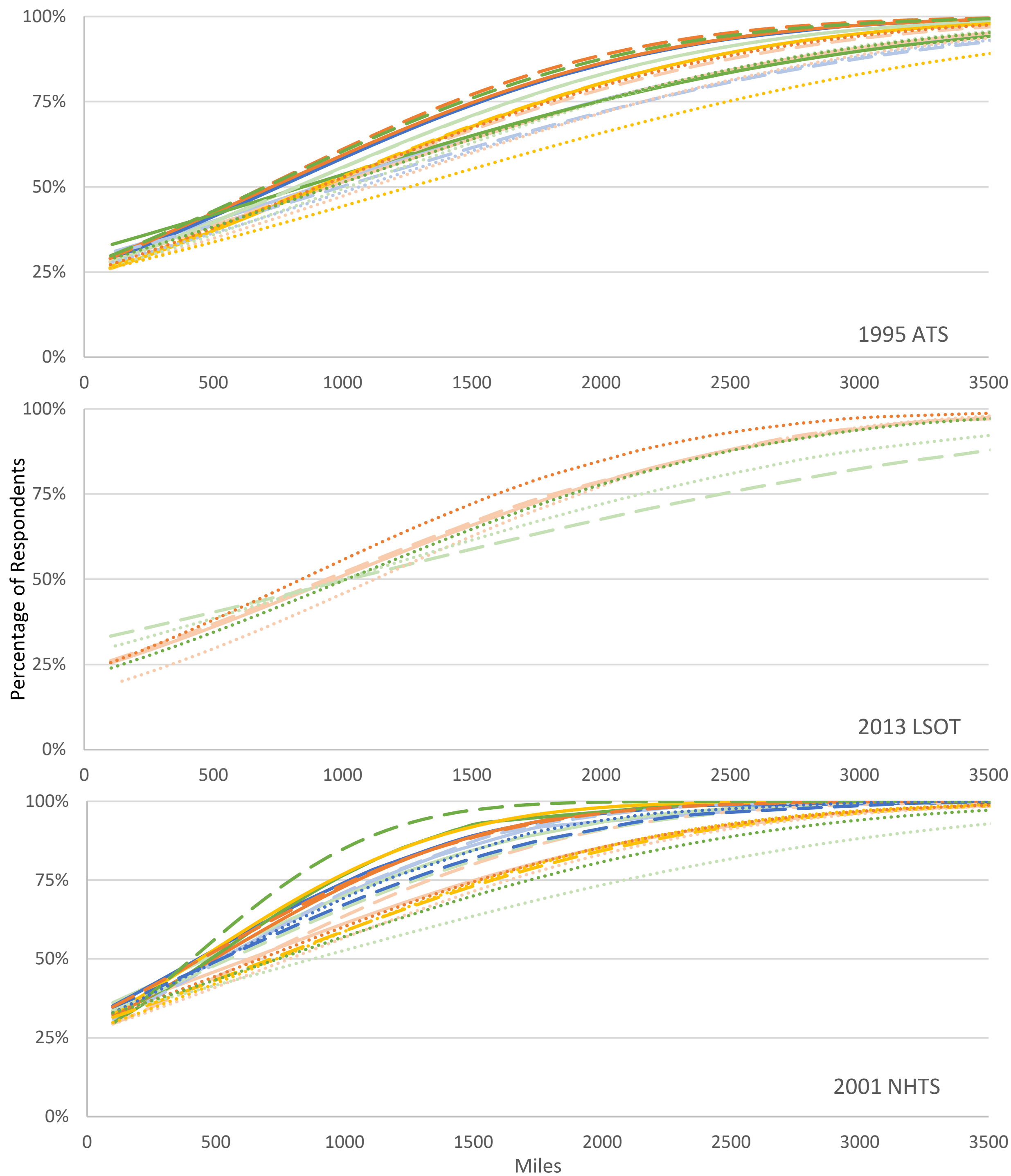


Income	Under 25		25 to 44		45 to 64		65 or Older
	Kids	*No Kids*	*Kids*	*No Kids*	*Kids*	*No Kids*	*All*
Low							
Medium							
High							

Figure 5: Trip Rate Boxplots (X = Mean)

Trip rates charts visually showed the difference annual panel data and the NHTS's 4-week collection period had on total long-distance travel rates. However, the general pattern of higher income households taking more trips was evident across all three datasets. Low-income households traveled less overall while medium and high income households mingled among the high-rate groups. While no single household group clearly stood out, high income, with children, age group 45 to 64; consistently appeared towards the higher range in trip making. Boxplot analysis showed this group as having one of the greatest ranges in trip making—particularly among the ATS dataset.

Figures 6 and 7 show median roundtrip distance cumulative distribution plots and associated boxplots.



Income	Under 25		25 to 44		45 to 64		65 or Older
	Kids	*No Kids*	*Kids*	*No Kids*	*Kids*	*No Kids*	*All*
Low							
Medium							
High							

Figure 6: Median Roundtrip Distance Cumulative Distribution Plots

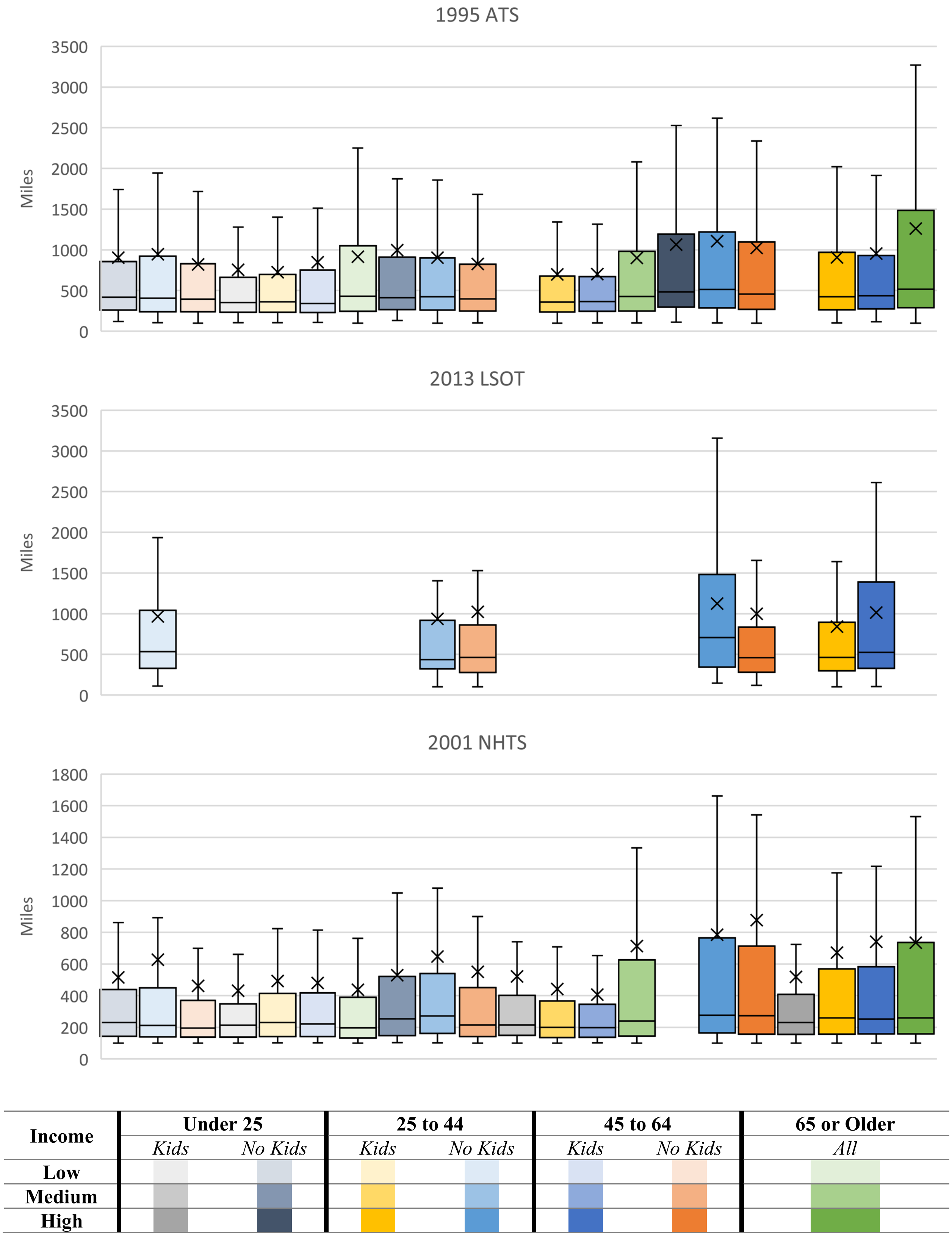


Figure 7: Median Roundtrip Distance Boxplots (X = Mean)

Median roundtrip distance had overall similar trends among all datasets, especially among the two panel surveys. No clear groupings were discernable which would support the idea that most long-distance trip making is universally similar in roundtrip distance. Boxplot analysis did show the high-income groups to have greater range than the other income groups. The presence of children also seemed to affect roundtrip distance as households without children had a wider range for distance than households with children.

Figures 8 and 9 display the median trip duration plots.

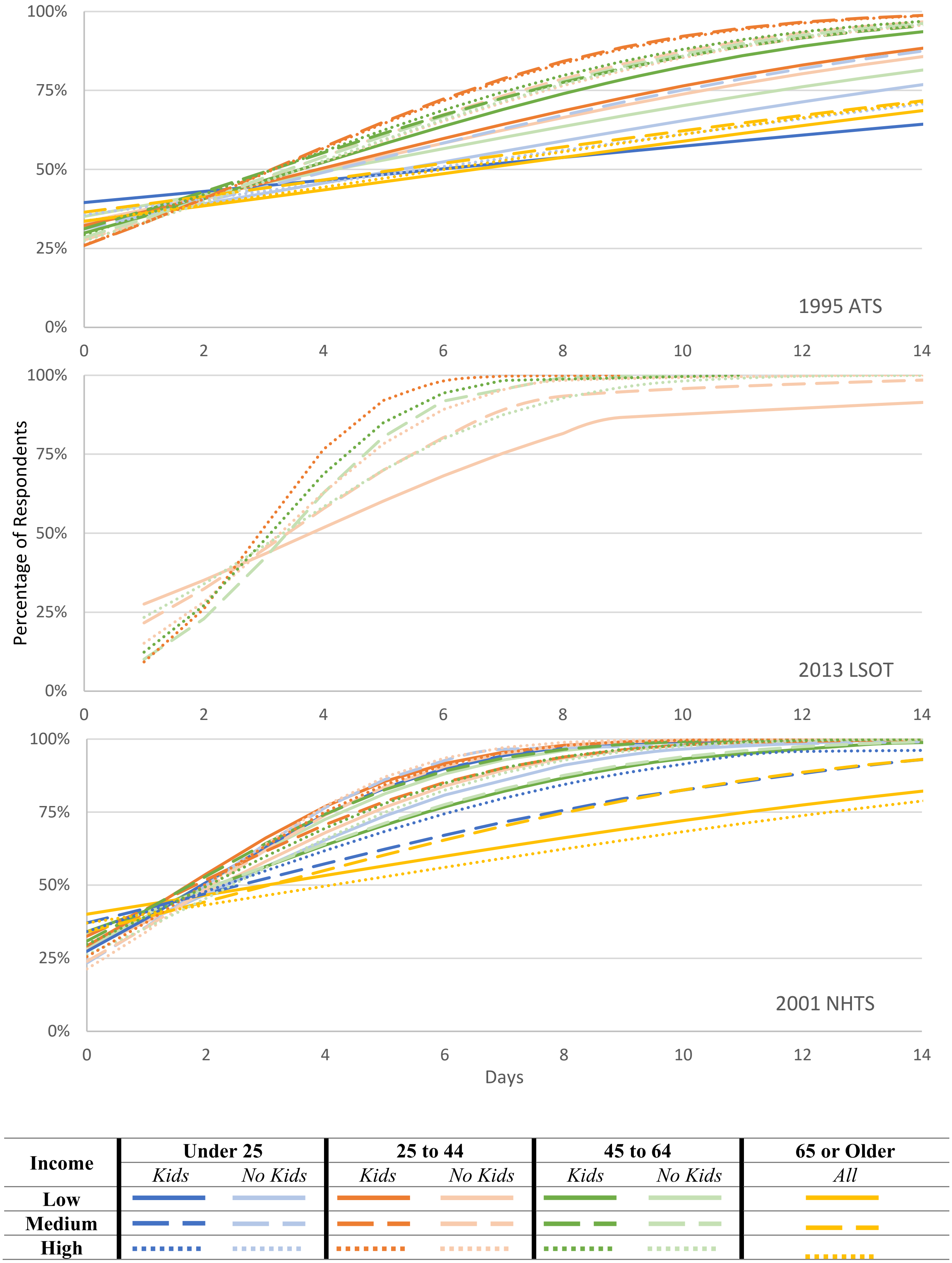


Figure 8: Median Trip Duration Cumulative Distribution Plots

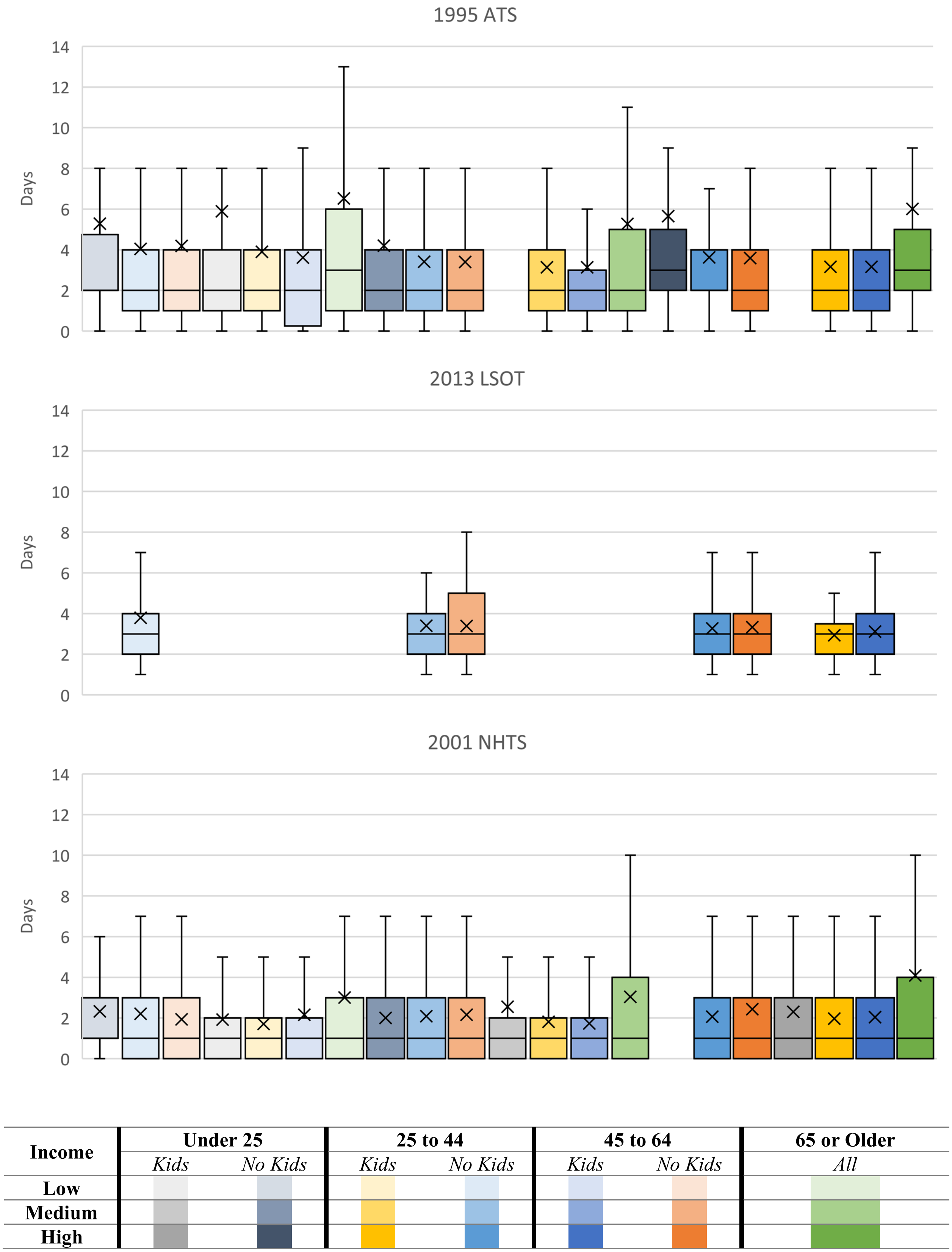


Figure 9: Median Trip Duration Boxplots (X = Mean)

Median trip duration presented rather interesting results with "65 or Older" age group households having flatter distributions than other age groups suggesting this age group partook in longer duration trips overall. Boxplots confirmed this pattern. Comparison of LSOT data reflected the scope of this survey. As this survey only recorded trips involving an overnight component, day trips (evident in the ATS and NHTS distributions) were not present. Although the NHTS and ATS had different collection periods, both datasets still trended surprisingly closely for durations under two days (with both datasets seeing 40 percent of each group hitting around that duration). As the NHTS was only a 4-week time collection period, trip duration over eight days became increasingly rare compared to the flatter curves highlighted with the ATS. Interestingly, the two panel surveys, the ATS and LSOT, did not have similar curve patterns; with the LSOT having tighter distributions on the lower duration end.

Figures 10 and 11 present the percentage of leisure trips cumulative distribution plots and associated boxplots.

Figure 10: Percentage of Leisure Trips Cumulative Distribution Plots

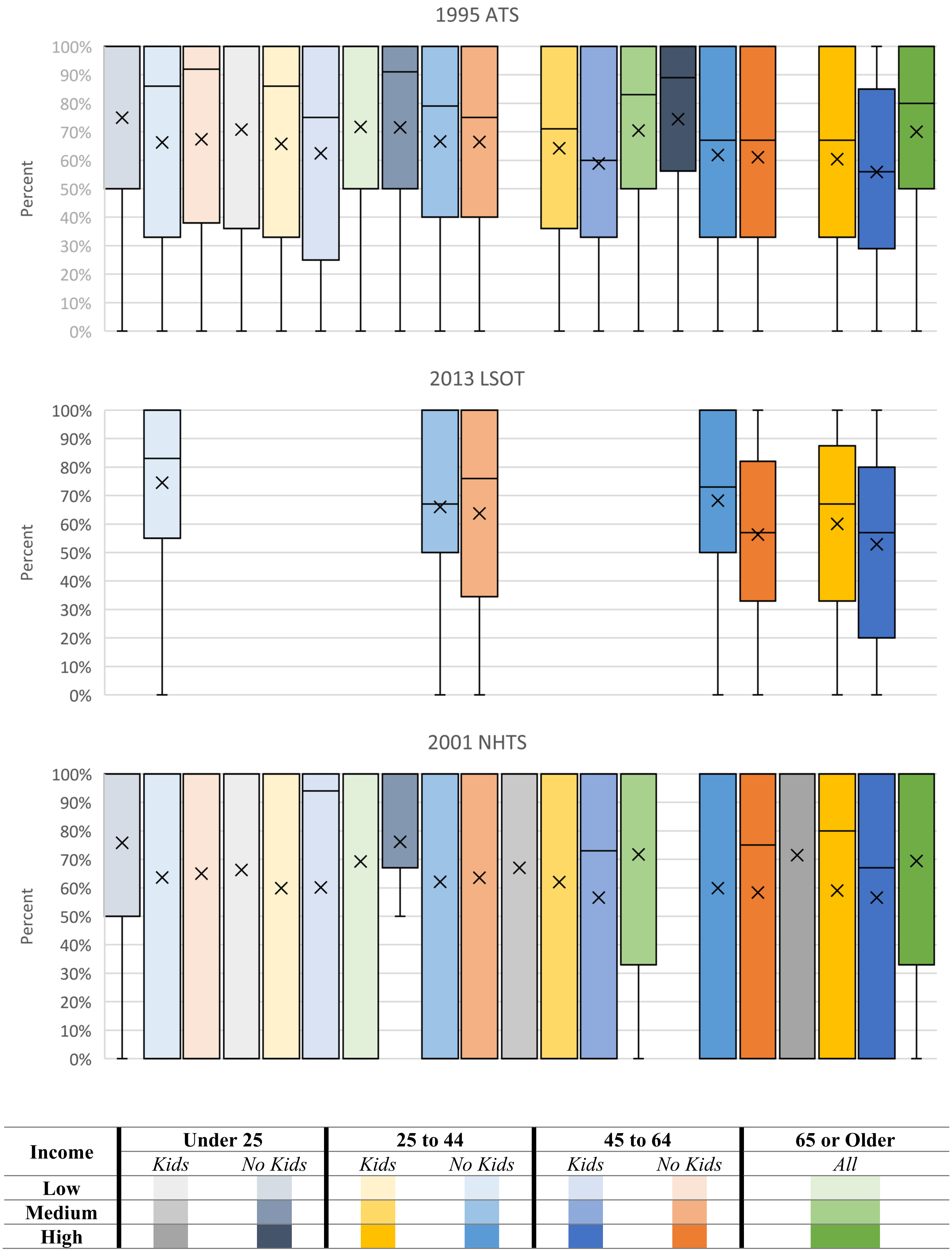


Figure 11: Percentage of Leisure Trips Boxplots (X = Mean)

As the percentage of leisure trips is nearly the direct inverse of the percentage of business trips (the percent of “other” trips was minute), the decision was made to highlight leisure trip distributions. This variable saw the most consistent patterns among all explored variables. Again; high income, with children, age group 45 to 64; stood out by showing a higher concentration of households taking a lower percentage of leisure trips than the other groups. For all three surveys, the group with the highest proportion of leisure trips was from the low-income group—particularly from the no children and younger age subgroups. The other sociodemographic groups were interspersed between these two notable groups. Once again, the effects of annual versus smaller time period data was evident with the NHTS boxplots hosting full-spectrum ranges. As noted with trip rates, the smaller collection period does not adequately capture typical long-distance travel patterns. This resulted in a number of households logging only one or two long-distance trips for the collection period which would naturally skew the trip purpose percentage breakdowns.

5.3.2 Outlier Calculations

To better apply the data for the ANOVA, Theil Index, and sample size calculations, each of the aforementioned variables and surveys were analyzed and voided of outliers. Table 33 highlights the raw variable summary statistics and their post-reduction statistics.

Table 33: Summary Statistics Pre and Post Outlier Detection

	Number of Trips			Median Roundtrip Distance			Median Trip Duration		
	ATS	*LSOT*	*NHTS*	*ATS*	*LSOT*	*NHTS*	*` ATS*	*LSOT*	*NHTS*
Original Cases	48,484	935	10,934	48,484	935	10,934	48,484	935	10,934
Min	1	1	1	100	102	100	0	1	0
Max	847	38	41	20,450	16,441	22,487	341	41	199
Median	4	6	2	414	480	236	2	3	1
Mean	6.93	7.57	3.13	895.68	969.16	640.64	3.95	3.30	2.33
Std. Dev.	10.216	5.953	3.234	1,358.082	1,446.421	1,249.014	9.234	2.993	5.591
Q_{25}	2	3	1	252	314	148	1	2	0
Q_{75}	8	10	4	920	965	516	4	4	3
IQR	6	7	3	668	656	368	3	2	3
Reduced Cases	43,921	899	10,400	43,061	823	9,394	44,717	874	10,336
Reduction	9.41%	3.85%	4.88%	11.19%	11.98%	14.08%	7.77%	6.52%	5.47%
Min	1	1	1	100	102	100	0	1	0
Max	16	20	8	1,921	1,946	1,064	8	7	7
Median	3	6	2	368	424	205	2	3	1
Mean	4.66	6.84	2.57	523.04	563.41	281.18	2.42	2.77	1.46
Std. Dev.	3.817	4.694	1.715	401.570	401.739	202.938	1.941	1.214	1.743

While four major variables are used in this analysis, all but the percentage of leisure trips variable were subjected to outlier detection. Since the leisure percentages were a) defined by a finite range, and b) displayed an already wide range of values in both extremes, the need for outlier detection was unnecessary. The most notable result of this process regarded the median roundtrip distance outlier reduction process. Unlike with the number of trips and median trip duration, median trip distance had the most room for variability among travelers which resulted in up to 14 percent of cases being removed as outliers on the upper end. The average percentage of outliers among the subgroups was roughly 12.5 percent for both the ATS and NHTS datasets.

This average jumped to 16.0 percent for the LSOT dataset due to the 65 or older age groups, which saw the greatest percentage of outliers. This was from the trip duration (for low and medium income) and roundtrip distance (for high income) sub-datasets. It should also be noted the extremeness of some of the raw maximum values. These represent the absolute fringe traveler groups who travel with such frequency and range, their inclusion greatly skews the data. The reduction percentages played a role in the sample size calculations presented later to account for these extreme travelers.

5.3.3 ANOVA Results

To determine if the chosen sociodemographic group characteristics best described long-distance travel trends, two ANOVAs were carried out for one-way, two-way, and three-way variable interactions. Since the “65 or Older” age groups were collapsed on the child/no child level as described in subsection 5.3.1, an ANOVA omitting the entire age group had to be run. A second ANOVA was compiled omitting the child/no child level all together; allowing for testing of all age groups and income levels. Table 34 presents the results of the “65 or Older” omitted ANOVA and the “Child-Omitted” ANOVA.

Table 34: ANOVA Results

		Child-Defined Groups (No Households over the Age of 65)			All Groups Based on Age and Income		
		ATS	**LSOT**	**NHTS**	**ATS**	**LSOT**	**NHTS**
		F-Stat	*F-Stat*	*F-Stat*	*F-Stat*	*F-Stat*	*F-Stat*
NUMBER OF TRIPS	**Age**	**38.052*****	**2.476***	**61.731*****	**55.686*****	1.183	**67.245*****
	Children	**4.233****	**3.251***	**98.924*****	~	~	~
	Income Level	**127.552*****	**7.395*****	**15.815*****	**331.773*****	**3.491****	**37.335*****
	Age*Children	**28.547*****	1.658	**39.339*****	~	~	~
	Age*Income Level	**5.451*****	0.810	**5.001*****	**10.099*****	0.122	**8.143*****
	Children*Income Level	1.460	0.234	0.848	~	~	~
	Age*Children* Income Level	1.158	1.609	0.234	~	~	~
MEDIAN ROUNDTRIP DISTANCE	**Age**	**5.105****	0.266	**2.348***	**13.054*****	1.452	0.991
	Children	**16.033*****	**5.668****	**5.052****	~	~	~
	Income Level	**11.666*****	**6.588*****	**6.652*****	**21.704*****	0.611	**12.544*****
	Age*Children	**2.331***	**2.831***	1.367	~	~	~
	Age*Income Level	1.462	0.456	0.198	**1.976***	0.731	0.804
	Children*Income Level	0.398	1.357	1.693	~	~	~
	Age*Children* Income Level	1.146	**2.990***	0.424	~	~	~
MEDIAN TRIP DURATION	**Age**	**7.389*****	0.632	**4.443****	**53.202*****	**2.520***	**2.326***
	Children	**28.817*****	**11.698*****	**15.280*****	~	~	~
	Income Level	**6.472****	0.702	**6.935*****	**19.037*****	0.078	**10.199*****
	Age*Children	**3.662****	1.892	0.716	~	~	~
	Age*Income Level	0.949	0.488	0.018	**3.440****	0.141	0.760
	Children*Income Level	2.286	**2.744***	0.257	~	~	~
	Age*Children* Income Level	0.848	2.207	0.527	~	~	~
PERCENTAGE LEISURE TRIPS	**Age**	**42.739*****	1.021	**12.932*****	**105.514*****	1.979	**25.627*****
	Children	**10.891*****	0.643	**5.665****	~	~	~
	Income Level	**3.916****	**8.893*****	1.640	**11.375*****	**5.277****	1.535
	Age*Children	**11.230*****	0.003	0.869	~	~	~
	Age*Income Level	**2.190***	1.737	0.271	**4.932*****	1.692	0.948
	Children*Income Level	0.370	0.345	2.028	~	~	~
	Age*Children* Income Level	0.641	0.257	0.830	~	~	~

***90% Confidence Level, **95% Confidence Level, ***99% Confidence Level**

On an individual variable level, each statistically differed to some degree amongst its categories in a majority of the cases. This trend was mainly observed with the ATS and NHTS surveys. The LSOT saw conflict within the median trip duration and leisure trip percentage groups. What was most interesting was the lack of two-way and thee-way interactions being statistically significant. To see if this was related to the dropping of the 65 or older age group, the second ANOVA considering only age and income variables was completed. While similar results were present for the individual grouping variables, the main interest was in the two-way interaction term. For the ATS survey, this interaction term was statistically significant in determining the defined age and income groups were statistically different from one another. However, outside the significance of this term for the number of trips in the NHTS survey, the two-way interaction was not found to be statistically significant on any meaningful confidence level.

A series of Tukey post hoc tests was administered to the household grouping variables (three-way interactions) to identify where the disconnect was occurring. Results found that group heterogeneity was prominent for both the ATS and NHTS surveys, while the LSOT survey showed practically no grouping uniqueness. Focusing on the ATS number of trips results, the defined household groups differed mainly along income levels as opposed to age or presence of children. The NHTS survey showed a more evenly diverse grouping among the income levels, age groups, and presence of children which would suggest the defined household groupings to be appropriate for this shorter-term survey period. One noticeable modification to the NHTS groupings would be the collapse of all age group 1 (under 25 years of age) household groups into a single group regardless of presence of children or income level. With the ATS results, this was not the case as the high income, no children, age group 1 grouping was statistically different from all other similar age groups to at least the 90 percent confidence level.

Median roundtrip distance and duration results showed only the ATS having any notable statistical patterns. Understandably, the NHTS—having a shorter survey capture period—

showed little grouping differences on a three-way interaction level. The only statistically significant differencing measure was high income households. In comparison, the ATS results not only reflected the high-income housing difference, but also the presence of children and age group (in particular the 65 or older age group) impacted the median duration and travel distance. As such, a yearlong panel survey would best serve for finding minute group differences for trip durations and distances. However, the LSOT, also a panel survey, had very little three-way interaction distinctions. On an individual variable level, only median trip duration showed statistical differences at the 95 percent confidence level. This was only with the 65 years or older age group and the presence of children (reflecting ATS results). The cause of this discrepancy could be the more focused and high-income biased LSOT survey sample.

For percentage of leisure trips, only the ATS results showed any significant differences on the three-way interaction level. The trends were very similar to the ATS and NHTS trip number results with distinct differences between income groups but not within income groups. Like the NHTS, the entire age group 1 (under 25 years of age) could be collapsed into a single grouping. On the individual variable level, the presence of children was statistically different for all three survey datasets at the 95 percent confidence level. Comparing income levels, both panel survey results (ATS and LSOT) showed significant differences among all three income levels. The NHTS results showed only significant differences among the high-income level respondents (similarly to the distance and duration results). This once again shows the finer detail panel surveys can give in comparison to shorter duration survey periods in respect to subgroupings of travelers.

Overall, ANOVA results and related post hoc tests suggested the three sociodemographic variables and their categories adequately characterized long-distance trip making. However, these groupings could be further collapsed to account for correlations, such as with the under 25 age groups or overall income levels. Of the three sociodemographic variables, income level was the best individual variable at describing long-distance travel behavior. Additionally, ATS results showed panel data better characterized the sociodemographic subgroupings compared to the NHTS's shorter survey timeframes.

5.3.4 Theil Index Results

While normally used as a measure of income equality, the Thiel T Index was applied to determine the equality within each sociodemographic group. This can in turn better illustrate how members within the group compare amongst their peers. By calculating a unitless Thiel T Index value for each group and dataset, group equality can be compared more directly between sociodemographic groups *and* surveys than the ANOVA or visual analysis options. Table 35 displays these findings.

Table 35: Theil T Index Values

			Number of Trips			Median Roundtrip Distance			Median Trip Duration			Percentage of Leisure Trips		
			ATS	*LSOT*	*NHTS*	*ATS*	*LSOT*	*NHTS*	*ATS*	*LSOT*	*NHTS*	*ATS*	*LSOT*	*NHTS*
LOW INCOME	**Under 25**	***No Children***	0.32	-	0.18	0.22	-	0.21	0.29	-	0.41	0.17	-	0.21
		Children	0.36	-	0.20	0.23	-	0.18	0.41	-	0.86	0.24	-	0.40
	25 to 44	***No Children***	0.35	0.25	0.18	0.26	0.23	0.22	0.38	0.10	0.76	0.27	0.11	0.41
		Children	0.33	0.25	0.17	0.25	0.07	0.19	0.44	0.12	0.88	0.27	0.18	0.45
	45 to 64	***No Children***	0.34	0.22	0.18	0.25	0.15	0.20	0.42	0.09	0.87	0.26	0.11	0.38
		Children	0.36	0.09	0.19	0.25	0.08	0.22	0.54	0.09	0.78	0.31	0.03	0.41
	65 or Older	***All***	0.34	0.42	0.19	0.26	0.05	0.20	0.43	0.04	0.90	0.21	0.09	0.33
MEDIUM INCOME	**Under 25**	**No Children**	0.32	-	0.15	0.23	-	0.19	0.31	-	0.62	0.19	-	0.25
		Children	0.33	-	0.17	0.25	-	0.20	0.36	-	0.75	0.21	-	0.38
	25 to 44	**No Children**	0.30	0.25	0.21	0.23	0.19	0.20	0.31	0.10	0.62	0.21	0.15	0.42
		Children	0.30	0.34	0.17	0.24	0.15	0.21	0.37	0.12	0.90	0.22	0.19	0.40
	45 to 64	**No Children**	0.30	0.24	0.18	0.24	0.24	0.22	0.39	0.11	0.81	0.22	0.20	0.38
		Children	0.27	0.19	0.16	0.23	0.19	0.20	0.41	0.08	0.89	0.23	0.19	0.44
	65 or Older	**All**	0.30	0.16	0.17	0.25	0.17	0.22	0.40	0.06	0.82	0.18	0.05	0.28
HIGH INCOME	**Under 25**	**No Children**	0.29	0.00	0.20	0.24	0.00	0.25	0.26	0.00	0.49	0.14	0.00	0.32
		Children	0.45	-	0.19	0.28	-	0.18	0.31	-	0.75	0.19	-	0.30
	25 to 44	**No Children**	0.26	0.21	0.20	0.24	0.22	0.22	0.25	0.08	0.54	0.21	0.12	0.42
		Children	0.25	0.20	0.17	0.25	0.20	0.22	0.30	0.08	0.66	0.22	0.19	0.42
	45 to 64	**No Children**	0.25	0.21	0.19	0.24	0.22	0.22	0.31	0.09	0.65	0.21	0.22	0.44
		Children	0.21	0.23	0.15	0.23	0.23	0.20	0.32	0.08	0.69	0.21	0.28	0.39
	65 or Older	**All**	0.27	0.16	0.16	0.24	0.17	0.23	0.28	0.07	0.67	0.17	0.11	0.30

Trip number results reinforced the patterns seen earlier with the ATS survey having the highest Theil index for most groupings and the NHTS having the lowest. As the NHTS was not an annual panel survey, participants had fewer opportunities to travel so the total number of trips was relatively low. The ATS, however, reflected its broad survey population with having moderate Theil indices. The LSOT showed similar results to the ATS. The two highest Theil indices observed for the number of trips was for the ATS's high income, under 25 years of age, with children group (0.45); and the LSOT's low income, 65 years or older group (0.42). Both groups had low sample sizes (49 and 8, respectively) which might have contributed to the higher numbers.

Median roundtrip distance results again had the ATS with the highest Theil indices, but the lowest alternated between the NHTS and LSOT datasets. However, all three surveys were relatively equal in terms of Theil numbers with most being within the 0.17 and 0.25 range. While these values were calculated sans outliers, the tight groupings suggest that all three surveys captured expected median roundtrip distance rather well, regarding group equality. This was not the case for median trip duration, as the NHTS results had the highest Theil values out of the entire analysis. This is once again most likely a reflection of the short surveying periods used in the NHTS which would not adequately capture a respondent's full year of long-distance travel, thus potentially missing the rarer, longer duration trips taken throughout the year. As the ATS results were relatively lower, it would support this idea: while trip duration equality naturally varies with annual panel data, it captures a more realistic depiction of typical travel than a shorter duration survey platform. The LSOT results for this section had almost the lowest Theil values of the entire analysis, however as the definition of a long-distance trip with this survey was defined as "overnight" the results are considered biased (missing all day trips).

A very similar story is told with the “percentage of leisure trips” Theil results. Since the NHTS had such a large percentage of respondents with low trip totals, the variance of the percent of leisure trips was quite high. This in turn would lead to the recommendation of ignoring those results due to data skewing. However, what was interesting was the rather similar results within the ATS survey. Each subgroup had very similar Thiel index numbers regardless of age, income, or presence of children. This was similar with the LSOT results with the notable exception of the high income 45 to 65 age group, which had the highest Thiel numbers in the LSOT set. As higher income households tend to partake in more long-distance business trips, this is most likely its reflection in the numbers; but as the LSOT was organized at the respondent, not household-level, this could reflect possibly multiple household respondents having different travel patterns (i.e., two respondents living together but one travels more for business, yet both report the same sociodemographic information used in this analysis).

5.3.5 Sample Size Results

Utilizing the cleaned group means and standard deviations, the estimated sample size for each sub-group was calculated to a 95 percent confidence level (z-score = 1.96). The margins of error allowed for each variable were one trip, 100 miles, one day, and 10 percent of trips; for number of trips, median roundtrip distance, median trip duration, and percentage of leisure trips, respectively. Table 36 shows the calculated sample size needed for each variable, survey, and household group. The sample sizes shown here only reflect traveling, compliant (not an outlier), households/persons.

Table 36: Minimum Estimated Sample Sizes (Maximum Percentage of Missing Outliers)

			ATS	LSOT	NHTS
LOW INCOME	**Under 25**	***No Children***	53 (10.5%)	-	57 (10.2%)
		Children	60 (11.8%)	-	85 (6.3%)
	25 to 44	***No Children***	63 (11.8%)	73 (11.9%)	84 (13.1%)
		Children	61 (8.5%)	51 (15.8%)	83 (10.1%)
	45 to 64	***No Children***	59 (10.1%)	79 (10.3%)	80 (8.7%)
		Children	63 (8.0%)	33 (16.7%)	77 (10.1%)
	65 or Older	***All***	70 (16.9%)	134 (37.5%)	75 (9.1%)
MEDIUM INCOME	**Under 25**	**No Children**	57 (12.4%)	-	66 (10.7%)
		Children	59 (10.3%)	-	82 (8.6%)
	25 to 44	**No Children**	60 (11.2%)	87 (11.4%)	80 (15.1%)
		Children	53 (7.5%)	76 (7.9%)	79 (10.5%)
	45 to 64	**No Children**	59 (9.9%)	73 (10.9%)	78 (11.6%)
		Children	59 (12.8%)	68 (9.1%)	77 (11.0%)
	65 or Older	**All**	64 (11.9%)	91 (36.4%)	69 (17.3%)
HIGH INCOME	**Under 25**	**No Children**	73 (13.4%)	-	74 (11.9%)
		Children	70 (8.2%)	-	74 (8.8%)
	25 to 44	**No Children**	75 (15.1%)	83 (17.3%)	79 (19.4%)
		Children	68 (12.4%)	77 (11.1%)	76 (14.8%)
	45 to 64	**No Children**	67 (17.1%)	103 (10.7%)	78 (19.8%)
		Children	69 (21.4%)	91 (13.1%)	69 (16.5%)
	65 or Older	**All**	72 (19.8%)	71 (20.6%)	69 (17.9%)
TOTALS			1,334 (11.2%)	1,190 (12.0%)	1,591 (14.1%)

Overall, the minimum sample size needed for each variable was quite small. The tightest groupings were for median trip duration (reflecting distribution findings from the previous section). The limited sampling period (4-weeks) of the 2001 NHTS was also evident as respondents were more likely to attempt fewer long-distance trips which in turn effected not only the range of expected number of trips, but also the percentage leisure trips. As fewer trips are made, the percentages tended to have a greater range, resulting in the needed larger sample size. This effect did not carry over to the median roundtrip distance estimates which would suggest the long-distance trips mainly captured by the 2001 NHTS were shorter in distance overall (reflective of the need for annual panel data to capture rarer, grander long-distance trips).

Regarding outliers, the average percentage of outliers was roughly 12.5 percent for both the ATS and NHTS surveys. This average jumped to 16.0 percent for the LSOT survey due to the 65 or older age groups, which saw the greatest percentage of outliers. This was from the trip duration (for low and medium income) and roundtrip distance (for high income) sub-datasets.

The maximum sample size for each survey's group was isolated and then increased by the percentage of outliers found in the raw dataset. This creates a new, minimum sample size for each household group needed to account for expected outliers and confidently recreate the group trends discussed throughout this paper. Table 37 presents these minimum sample sizes.

Table 37: Minimum Sample Size Needed for Each Dataset (Corrected for Outliers)

			ATS	LSOT	NHTS
LOW INCOME	**Under 25**	***No Children***	59	-	57
		Children	60	-	85
	25 to 44	***No Children***	70	81	84
		Children	61	51	83
	45 to 64	***No Children***	64	79	80
		Children	63	33	77
	65 or Older	***All***	79	134	75
MEDIUM INCOME	**Under 25**	**No Children**	64	-	66
		Children	64	-	82
	25 to 44	**No Children**	67	91	80
		Children	57	76	79
	45 to 64	**No Children**	65	78	78
		Children	67	68	77
	65 or Older	**All**	72	108	69
HIGH INCOME	**Under 25**	**No Children**	83	-	74
		Children	70	-	74
	25 to 44	**No Children**	86	97	79
		Children	75	81	76
	45 to 64	**No Children**	78	108	78
		Children	84	96	69
	65 or Older	**All**	86	86	69
TOTALS			**1,473**	**1,264**	**1,591**

At the 95 percent confidence level and previously stated margins of error, the minimum sample sizes for each survey are surprisingly similar. While the ATS and NHTS sample sizes are significantly less than the original surveys' scopes, the LSOT's is pretty much the same sample size (1,013) which suggests that survey's scope to be appropriate in capturing general long-distance travel trends.

It should be noted that an effort to determine the non-traveling surveyed households for each dataset was made with little progress. Long-distance travel is an irregular event to most respondents and as such, a sizeable portion of those surveyed most likely recorded zero long-distance trips during the survey period. In fact, using the NHTS's and ATS's stated total sample sizes (63,000 and 80,000, respectively) identified 83 percent and 39 percent of surveyed households as non-traveling or invalid responses. Due to dataset design, these non-travelers could not be confidently identified (with the exception of the NHTS) on the socioeconomic group-level and as such were ignored for these sample size calculations. Further investigation into non-traveling households would be useful for future survey sample research.

5.4 Conclusions

This chapter explored how long-distance survey sampling could be improved by targeting sociodemographic groupings. The traditional approach to travel survey sampling is population proportion sampling which considers an equal chance of capturing travel behavior based on the study area's sociodemographic distributions. However, by utilizing known patterns and behaviors inherent to long-distance travel within sociodemographic groups, the sampling effort can be targeted to reduce the overall sampling frame without losing data fidelity and maintaining capture equity, while also reducing administrative costs.

To accomplish this, three US-based surveys—the 1995 ATS, 2001 NHTS, and 2013 LSOT—were analyzed for patterns both between and within different sociodemographic groups. These sources were selected as they represent a range of time periods, slightly varied definitions of long-distance travel, and varied geographical locations. Long-distance travel analysis was completed by characterizing travel behavior considering four travel metrics a) the volume of long-distance trip making, b) median roundtrip distance, c) median trip duration, and d) the percentage of trips labeled as leisure purpose.

Compiling the results from this analysis showed the distillation of complex sociodemographic characteristics into groups defined by household income, respondent age, and the presence of children in the household to be promising for describing general long-distance travel trends. In particular, defining household income into either low, medium, or high levels using relative poverty threshold measures not only normalized the differences in each dataset's categorical income definitions, but also adequately described travel trends between groups distinctly in most cases. ANOVA and post hoc testing highlighted these insights but also shed further light on both the NHTS and LSOT surveys' flaws:

- The NHTS's short survey timeframe (28 days) reflected its inability to capture the nuances in long-distance travel particularly with trip duration and roundtrip distance. While the high-income group distinguished itself in this case, further detail and group identity among the other income groups as well as the age groups was lacking.
- The LSOT's smaller sample size and bias towards higher income individuals not only limited group calculations due to inadequate candidate pools, but also contrasted the results of the other panel survey, the ATS. While the ATS clearly highlighted group heterogeneity (especially among trip rates), the LSOT failed to at the same level.

However, the individual statistical difference results among each sociodemographic grouping variable were similar, which would suggest a retooling of the chosen grouping variable interactions. One such retooling identified was the collapse of age group 1 (under 25 years old) into a single group regardless of child presence or income level. Collapsing into a single group would not only reduce the overall number of subgroups thus alleviating small sample size issues, but also still adequately distinguish this group's long-distance travel habits.

Theil index results showed equality within groups to be rather high. The noted exception was with trip duration and leisure trip percentage within the NHTS, reflecting the survey's shorter survey period. Roundtrip distance results were rather comparable amongst all three datasets which would suggest long-distance travelers act similarly, at least with median roundtrip distance, when controlled for outlier trips.

Finally, sampling techniques had two major findings: first, the needed valid sample size, even controlled for outliers, was multitudes smaller than the actual survey sample size (in particular the national-level surveys). Second, non-traveling households need to be identified and accounted for in sample size calculations. Regarding the second finding, only the NHTS offered any accurate national counts of these non-traveling households on the defined group level. While the LSOT did have some non-traveling households participate throughout the survey's duration, this was identified as five records, hardly a sample size from which to make conclusions. Surface-level analysis suggested up to 83 percent of those surveyed in the 2001 NHTS did not partake in any long-distance travel over the survey period. It would then be recommended the defined sample sizes be doubled or even tripled to account for these non-travelers for non-panel, annual surveys. However, surface-level ATS numbers would suggest only 39 percent of those surveyed were non-traveling households, which is a decent reduction, but individual group capture rates were still unidentifiable given the full ATS survey dataset is unavailable for analysis. It was therefore recommended further analysis on identifying non-traveling rather than traveling households would be beneficial in determining sample size calculations, particularly in the application of non-panel surveys.

While more work is needed to better identify non-traveling households, planners and researchers can still utilize the general results of this dissertation to better approach long-distance travel surveying efforts. Overall, this chapter supported the second objective and hypothesis identified for this dissertation. The long-distance sociodemographic sampling process can be streamlined in a way that a) reduces the overall number of surveys needed for a statistically valid sample, and b) employs a universal household categorical system that limits the necessary data fidelity reducing survey burden while also supporting respondent privacy. This system also allows for easier comparisons between differing survey datasets by harmonizing the sociodemographic definition universe; with this system employed throughout this dissertation.

Chapter 6: Targeting the Geographic Capture Area

The last sampling frame targeting approach this dissertation considers is the idea of exploring similarities in long-distance travel behavior among geographic areas. Long-distance travel is difficult to capture using traditional household travel survey methods, given both the lower incidence of travel as well as the lack of a sampling frame or universe of long-distance travelers. Building on focused household travel survey sampling techniques that have been studied extensively related to *demographics*, this approach also considers how *regional geography* could be used to target long-distance traveling households [46, 168, 169]. This chapter highlights an approach based on how households in similar geographic areas (defined by regional access, air travel opportunities, etc.) may demonstrate similar long-distance tripmaking compared to peers in dissimilar areas. In practice, this would mean two households with similar sociodemographics would make *similar* long-distance travel behaviors if they both resided in areas with similar geographic access to long-distance destinations and modes. However, these two households would make *different* long-distance travel behaviors if they resided in areas with different geographic access characteristics. While much anecdotal evidence exists of this behavior across the US, it is not currently being used in practice for data collection or analyses.

Several studies have indirectly highlighted the potential of using national travel survey data to understand the regionality of long-distance travel behavior. For example, the Oak Ridge National Lab's (ORNL) study, while mainly focused on short-distance travel, found scaling and integrating the national-level NHTS and other data sources for application to smaller geographic areas (such as Census tract or transportation analysis zones) *could* be possible, but those smaller geographies might be better served by local household surveys [46]. This has also been hinted in Tennessee's integration of the FHWA's national long-distance passenger model to its statewide model [169]. Here, researchers had to calibrate the national model using state data to better forecast long-distance travel demand. Points were made on how the national model did an "impressive job of reproducing the number of trips observed in each district" but struggled to account for regional issues such as overpredicting in the northeast districts of the state (where travel could be reduced due to both physical and psychological barriers attributed to mountain travel). Other studies have also acknowledged directly surrounding locale influences local travel [170-175], but it is not farfetched to consider these findings could support how surrounding locale (such as mode accessibility or regional habits/tendencies) influences long-distance travel behavior choices.

In the previous chapters, representative data on long-distance travel could be achieved by first identifying groups of respondents or travel days with similar travel patterns, and then targeting these groups with smaller sample sizes. Therefore, the objectives of this study are to a) develop a socio-geographic cluster classification system for stratifying US counties and b) measure differences in long-distance travel behavior across these classifications. Specifically, this was accomplished via three major tasks: first, machine learning was used to develop a cluster classification system for US counties based on median household income, median age, percent urban (based on percentage of total population residing in urban areas), access to passenger rail, and access to sizeable airports. Second, the 2001 NHTS long-distance person-trip volumes were weighted and summarized for each of the calculated geographic clusters. Mean household total long-distance trips were measured by trip rates, by mode (air, vehicle, other), and by purpose (leisure, work, other). Third, ANOVA testing was used to determine if there were statistical differences in long-distance trip-making between classification clusters and/or census regions. This research directly builds on previous studies by ORNL considering the impact geographic characteristics have on daily travel [46]. Results are then used to develop a socio-geographic sampling technique that can more efficiently support long-distance travel survey data collection.

This chapter is organized as follows: first, the socio-geographic data and methods for clustering counties are outlined. Next, the steps used to weight and aggregate the NHTS data for these clusters are described. Third, ANOVA analyses are conducted to determine the statistical differences in tripmaking between clusters and regional levels. Finally, results are summarized and applications/approaches for targeted long-distance travel survey sampling techniques are presented.

6.1 Methodology and Data

The first step for understanding regional trends in tripmaking (and thus informing alternative sampling approaches) was to identify trends in community types, for which trip volumes can then be compared across. Geography was defined as counties for this work to balance representative sample sizes with likely differences in land uses.

6.1.1 Characterizing Counties

Working off the ORNL report's classification scheme and literature review, all 3,141 US counties were characterized by five variables related to long-distance travel behavior: urban percentage, median household income, median household age, largest airport hub access, and Amtrak access. All county demographic data was collected for the year 2000 to match with the travel survey data presented in the next section. Historic county GIS files, population, income, age, and county population density centroids were collected from the IPUMS NHGIS website [176]. Urban percentage was calculated as the percentage of the county's population residing in an urban area. County median household income was categorized into High, Medium, and Low, using the method utilized throughout this dissertation.

Airport and Amtrak location GIS data for 2000 was collected from the BTS NTAD [177] and USDHS HIFLD [178] GIS websites, respectively. ArcGIS was used to determine (a) the largest airport hub type hub (according to 2001 FAA estimates [179] of Large, Medium or Small) within a 2hr/130-mile straight-line distance around each county's population density centroid and (b) whether an Amtrak station was within a 0.5hr/33-mile straight-line distance around each county's population density centroid. Both variables were stored as categorical variables (e.g. largest airport hub type and yes/no Amtrak is present). Different access time/distances thresholds were selected based on observed long-distance travel mode choice trends [77, 78, 88, 92, 104] and professional judgement. Figure 12 and Table 38 present geographical and statistical summaries of the clustering variables, respectively.

Table 38: County Summary Statistics at the National and Regional Levels

Variable	National	Northeast	Midwest	South	West
Counties (n)	3,141	217	1,055	1,424	445
Urban Percentage					
Mean	40.12%	55.60%	36.12%	38.82%	46.23%
St. Dev.	30.99%	29.87%	29.61%	30.37%	33.51%
Median Age					
Mean	37.35	37.99	38.19	36.8	36.77
St. Dev.	4.013	2.329	3.885	3.74	5.235
Median Income					
% High	4.11%	13.36%	2.75%	3.32%	5.62%
% Medium	71.54%	83.87%	83.03%	58.92%	78.65%
% Low	24.36%	2.76%	14.22%	37.85%	15.73%
Largest Airport Hub Access					
% Large	40.78%	65.90%	40.28%	40.59%	30.34%
% Medium	23.30%	15.67%	19.81%	30.06%	13.71%
% Small	20.15%	17.51%	12.13%	26.97%	18.65%
% None	15.76%	0.92%	27.77%	2.39%	37.30%
Amtrak Access					
% Yes	37.06%	64.06%	35.45%	35.32%	33.26%
% No	62.94%	35.94%	64.55%	64.68%	66.74%

Figure 12a: Median Income

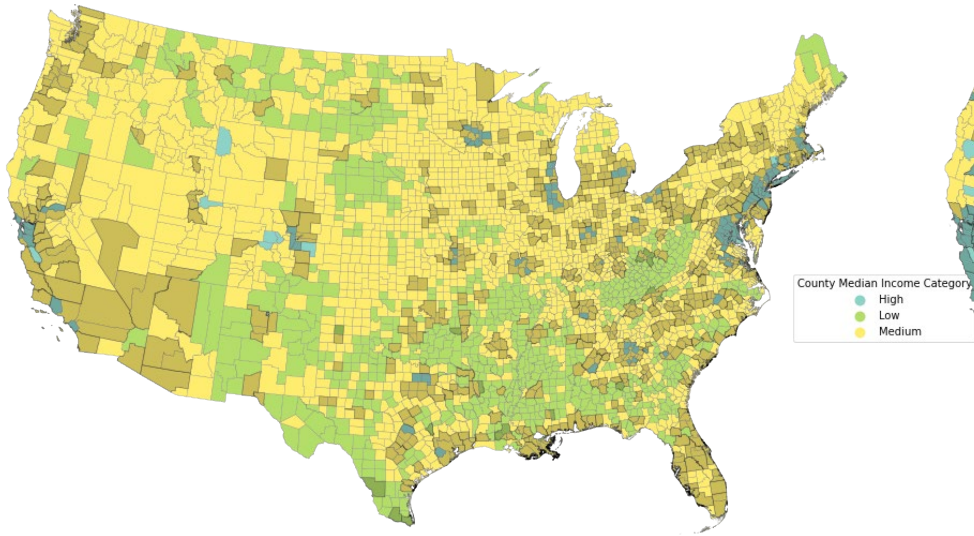


Figure 12b: Amtrak Station Accessibility

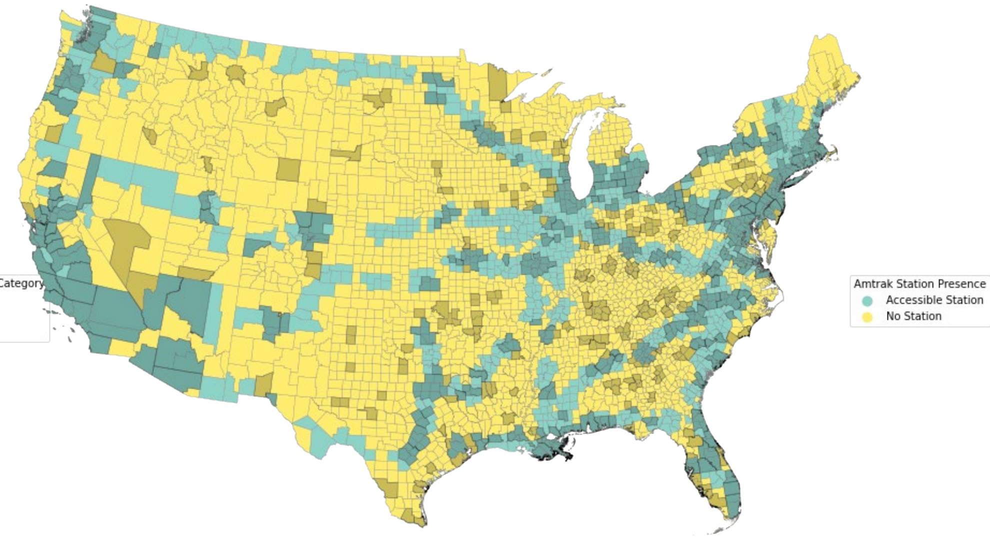


Figure 12c: Largest Accessible Airport

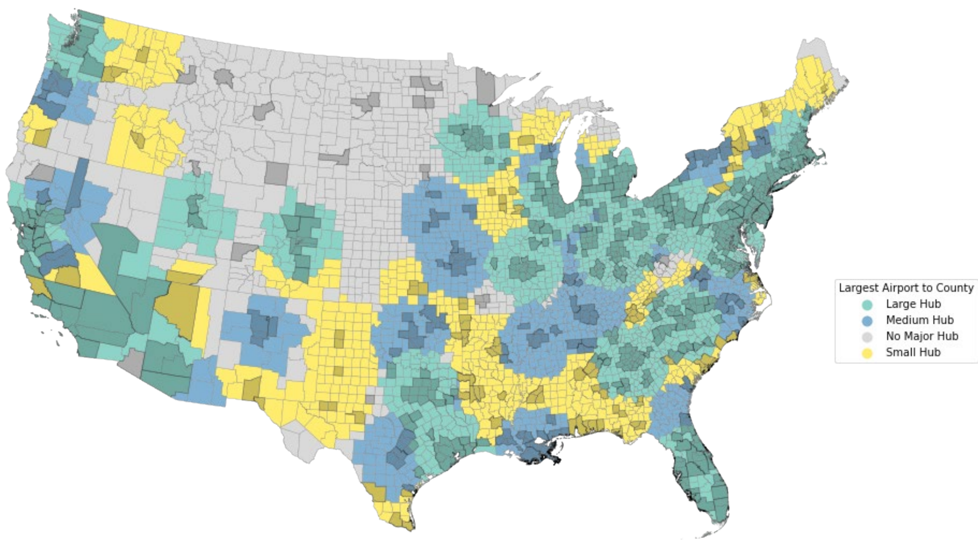


Figure 12d: Urban Population Percentage

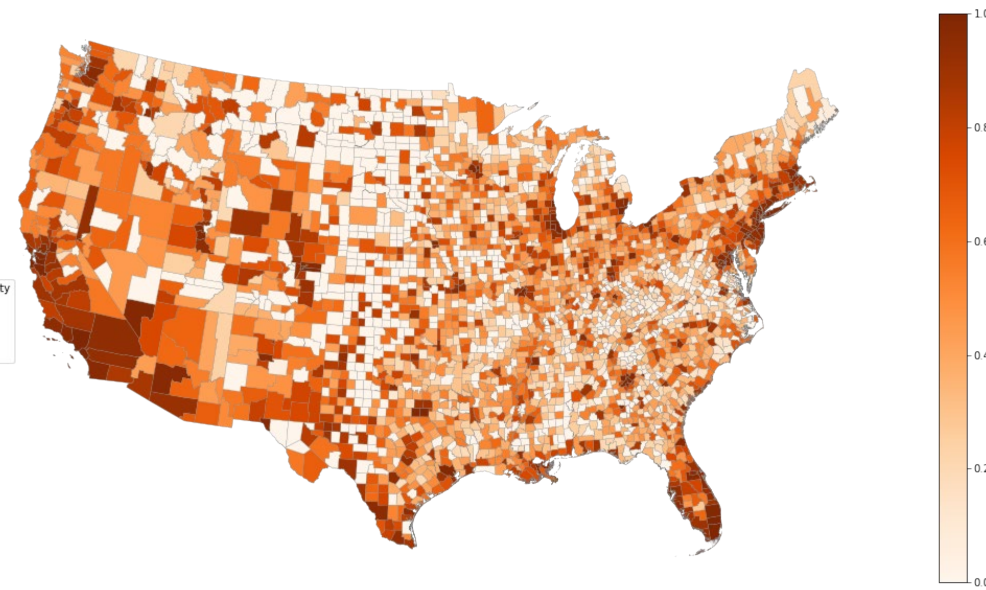


Figure 12: Categorical Variable Mapping of US Counties (MSAs Identified as Darkened Silhouettes)

6.1.2 K-means and Fuzzy C-Means Clustering

Naturally-occurring groups of the US counties were then identified using K-means and Fuzzy C-means clustering in Python, machine learning techniques widely used in geospatial data analyses due to their flexibility in spatial classification [180-183].

First, the K-means algorithm was used to determine the optimal number of clusters needed to characterize counties. This step is defined as hard clustering, where each county could only be classified into a single cluster. K-means clustering was applied to the dataset iteratively to graph a distortion score curve. This process allowed for the identification of the optimum number of clusters that both has the lowest distortion score (the average of the Euclidean squared distance from the centroid of the clusters) and lowest computational fit time. Figure 13 demonstrates the distortion score elbow (where the sum of squared distances begins to lessen with each additional value of *k* as well as minimize computation time) occurring around 11 clusters.

Second, the Fuzzy C-means algorithm was used to determine counties' cluster classification. This step is defined as soft clustering, where each county can be classified as a probability of falling into different analysis) and the fuzziness parameter (which defines the rigidity of cluster membership and was kept at the default package value). The Python *fuzzy-c-means* package [184] was used to classify counties based on the five variables of interest. Final cluster membership for each county was defined by looking at the cluster with which the county had the highest association probability.

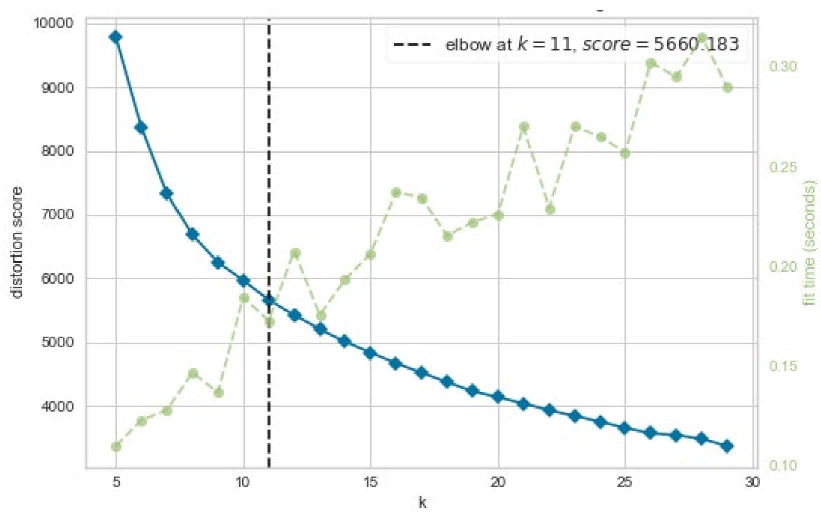


Figure 13: Distortion Score Curve (Blue Line Plots Distortion Score and Green Line Plots Algorithm Fitting Time)

6.2 Clustering Results

The eleven county-types identified through the K-means and Fuzzy C-means clustering process are shown in Table 39. They are listed in order of average urban percentage to aid in discussion. Additionally, Figure 14 showcases how this classification scheme is dispersed geographically across the United States.

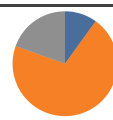
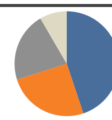
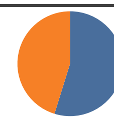
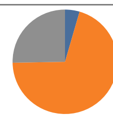
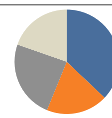
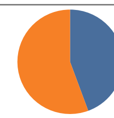
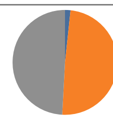
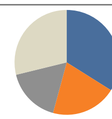
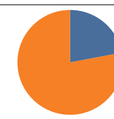
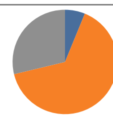
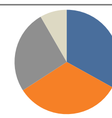
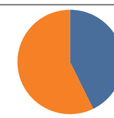
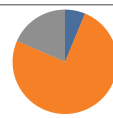
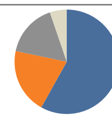
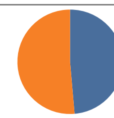
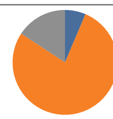
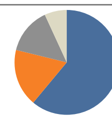
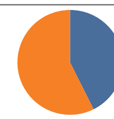
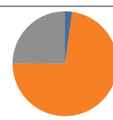
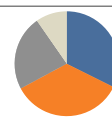
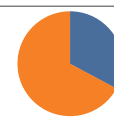
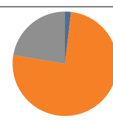
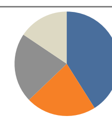
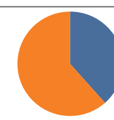
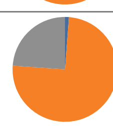
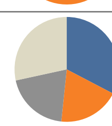
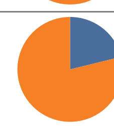
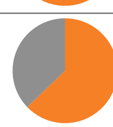
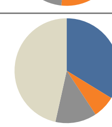
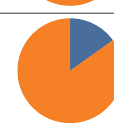
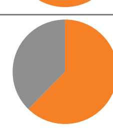
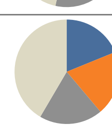
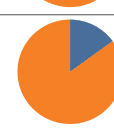
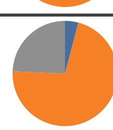
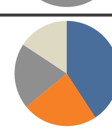
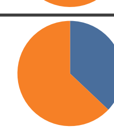
Clusters U-1 through U-3 consisted of the most urban-dense counties. Clusters U-1 and U-2 comprised the majority of counties within MSAs and had relatively similar socio-geographic characteristics. However, counties in Cluster U-1 had more access to airports and Amtrak stations. Cluster U-3 captured younger and less wealthy urban counties, which highly correlated with counties with military bases or colleges/universities.

Clusters M-1 through M-4 characterized counties with a more mixed density development. Notably, all four clusters had high percentages of counties with medium-level incomes. Counties within clusters M-2 and M-3 stood out as having excellent access to large airport hubs, whereas counties within clusters M-1 and M-4 make up by having increased access to medium airport hubs. Inspection of these counties in Table 39 shows they are often concentrated in the East North Central and upper East South Central census divisions.

Clusters R-1 through R-4 represented the rural counties across the nation. Counties within clusters R-1 and R-2 have slightly higher median incomes, but very few counties in any of these clusters have median incomes that fall into the high category. Cluster R-1 has the best long-distance mode accessibility among rural-dominate clusters, with 84-percent of counties having some airport access. Airport access quickly declines among the remaining clusters, with R-3 and R-4 counties being the most remote of the entire classification. R-1 and R-2 occur mostly within the Piedmont Region (particularly rural Pennsylvania and West Virginia) as well as concentrated in the West North Central census division, whereas R-3 and R-4 describe rural communities across the entire US.

Overall, the clustering results highlight the vastly different communities that exist across the United States in terms of urban development, income levels, and access to long-distance travel modes. It is hypothesize these different communities will pursue different amounts of long-distance trips due to the relative ease (or difficulty) in completing them. Additionally, the identified clusters visually match with previously observed regional trends, which further supports their use in the next analysis.

Table 39: Characteristic Summaries of Clusters

Cluster	Character	Count	**Urban Percentage** Average (Std. Dev.)	**Median Age** Average (Std. Dev.)	**Median Income** ■ High ■ Medium ■ Low	**Largest Airport Hub Access** ■ Large ■ Medium ■ Small ■ Other/None	**Amtrak Access** ■ Yes ■ No
U-1	Urban Higher Income Average Access	261	64.02% (28.481%)	32.97 (0.670)			
U-2	Urban Mixed Income Average Access	178	60.15% (28.332%)	30.11 (1.021)			
U-3	Urban Lower Income Lower Access	59	59.05% (31.656%)	25.31 (1.944)			
M-1	Mixed Density Higher Income Average Access	372	50.41% (29.506%)	34.79 (0.450)			
M-2	Mixed Density Higher Income Higher Access	401	47.03% (28.159%)	36.04 (0.346)			
M-3	Mixed Density Higher Income Higher Access	406	43.20% (27.595%)	37.18 (0.344)			
M-4	Mixed Density Mixed Income Higher Access	402	34.80% (25.994%)	38.21 (0.364)			
R-1	Rural Mixed Income Average Access	385	30.99% (26.949%)	39.45 (0.417)			
R-2	Rural Mixed Income Lower Access	331	25.57% (27.136%)	40.95 (0.508)			
R-3	Rural Lower Income Lower Access	86	18.43% (29.682%)	47.15 (2.346)			
R-4	Rural Lower Income Lower Access	260	15.37% (25.681%)	43.06 (0.775)			
All		3141	40.12% (30.989%)	37.35 (4.013)			

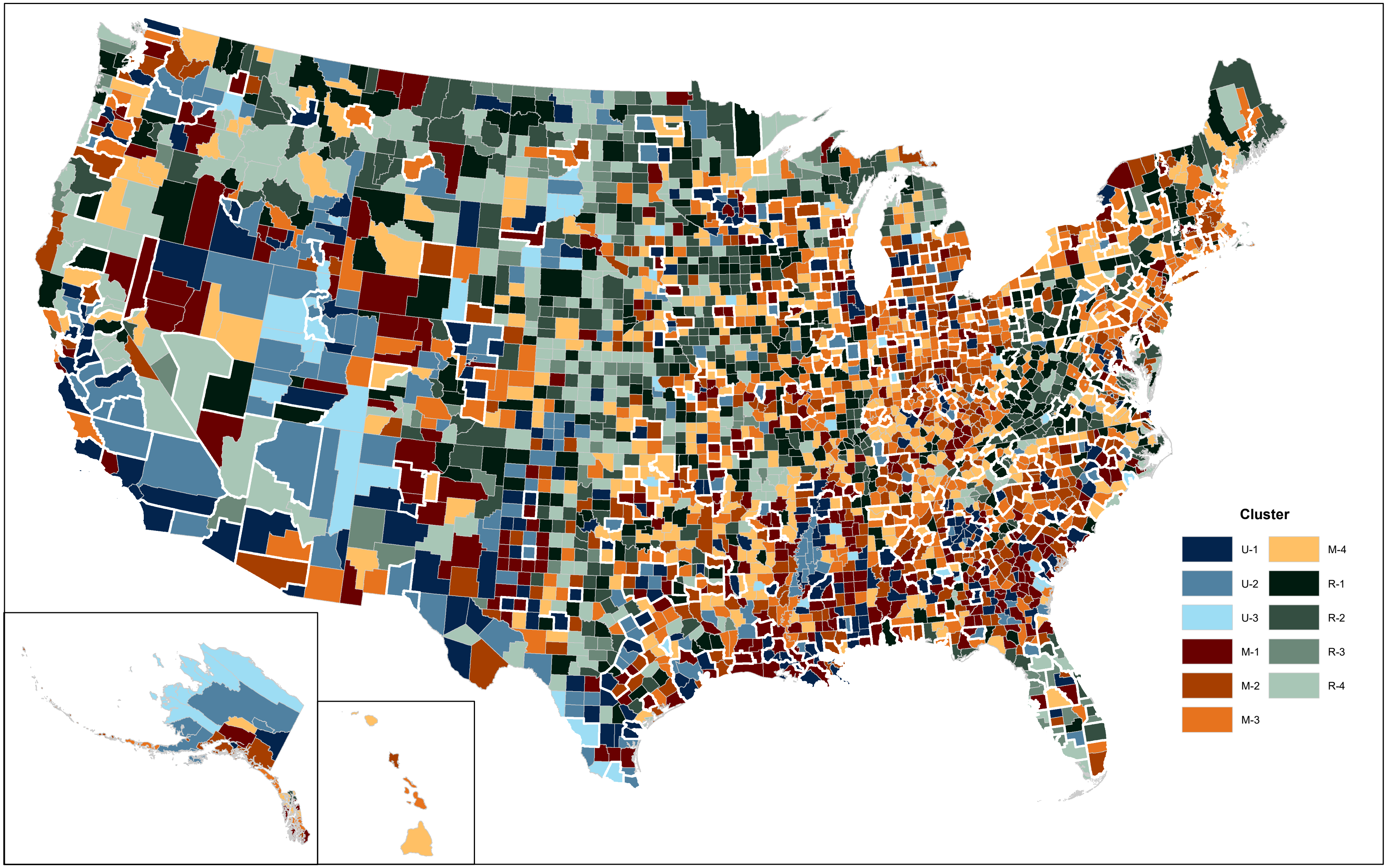


Figure 14: County Clustering Identification (MSAs Outlined in White)

6.3 Long-Distance Trip Volumes by Socio-Geographic Cluster & Census Region Results

The second step of the analysis was to weight and summarize 2001 NHTS long-distance trip volumes for households in each of the different classification clusters. Specifically, we focused on household long-distance trip volumes of total trips, total trips by mode (air, vehicle, other), and total trips by purpose (leisure, work, other). As such, all references to trips and travel in this section are in relation to long-distance travel only (i.e., no daily travel). It should be noted the NHTS data used for this chapter was from the restricted, non-public files which were requested and approved by FHWA personnel.

6.3.1 Cleaning & Weighting the 2001 NHTS Long-Distance Travel Supplement

The 2001 NHTS was the latest national survey to capture long-distance travel. It was conducted under the sponsorship of the US DOT's Bureau of Transportation Statistics (BTS), Federal Highway Administration (FHWA), and National Highway Traffic Safety Administration (NHTSA). A total of 69,817 households were surveyed about their long-distance travel during an assigned 28-day period. The 2001 NHTS defined long-distance travel as any trip that was 50-miles one-way. Trips were recorded at the person-level with a unifying household ID to aggregate tripmaking trends across all household members. Access to restricted use files was granted for this research effort, which provided county of residence for participating households.

This dissertation defined long-distance travel as any trip that was both over 50-miles one-way and having an overnight component, resulting in a final cleaned sample size of 17,310 50-mile or more, overnight long-distance person-trips for this chapter. The final project dataset reflected the travel of 63,162 households (after removal of households that were missing the critical demographic variables) across 2,567 counties (81.7 percent of all U.S. counties). Of these, 7,373 households reporting long-distance travel were located in 1,664 counties, or 53.0 percent of all U.S. counties and 64.8 percent of NHTS surveyed counties.

It is recognized that a) long-distance travel is heavily influenced by seasonality [76, 78, 81-83, 88, 98, 99, 101] and b) the dataset does include travel disrupted by 9/11, but the fact households are represented across the entire year helps distribute any variability associated with any single group. Given the NHTS sampled respondents based on geographic requirements throughout the year, a basic understanding of long-distance travel behavior at an *aggregated level* can still be inferred from the data.

After cleaning the data to remove households with missing data, characteristics for the remaining 63,162 households were then weighted in relation to the 2000 Census county-level data. First, a county was assigned to each respondent household based on residence. Second, weights to adjust for probability of selecting a household from a specific county, *x*, were calculated using Equation 7:

$$Weight_x = \frac{Census\ Population_x / Census\ Population_{US}}{Survey\ Sample_x / Survey\ Sample_{US}} \tag{7}$$

Third, weights were calculated to adjust for fielding bias based on age groups and income. Specifically, for each cluster, *z*, the NHTS survey responses and the 2000 census were summated relating to four different demographic groups, *d*, using combinations of age (< 65 vs ≥ 65) and income (low vs. med/high). Weights were calculated using Equation 8:

$$Weight_{zd} = \frac{Census\ Population_{zd} / Census\ Population_z}{Survey\ Sample_{zd} / Survey\ Sample_z} \tag{3}$$

Fourth, the two weights were applied to every household in the sample, based on their county, *x*; cluster, *z*; and demographic group, *d*. Finally, the records were normalized to keep the same total number of long-distance trips in our final dataset as we observed in the original. Table 40 summarizes the total long-distance trips completed within each cluster with weights and normalization. The weighted trip volumes and means were used in the rest of this chapter.

Table 40: Long-distance Trips by Cluster After Weighting

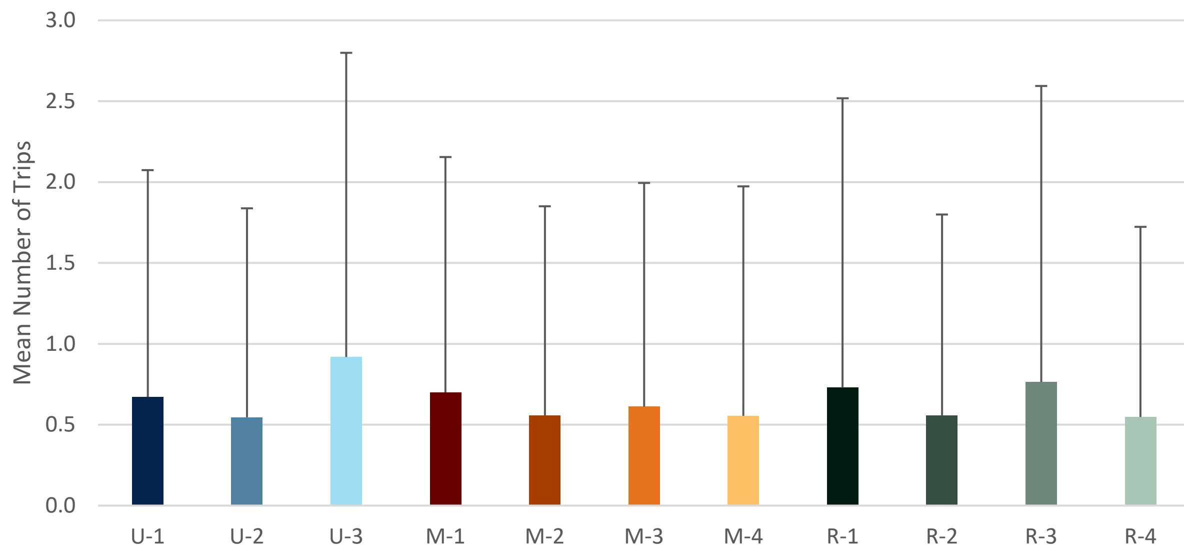
Cluster	Observed	Weighted	Normalized and Weighted
U-1	4,131	10,417	4,973
U-2	1,314	2,640	1,260
U-3	312	591	282
M-1	2,610	5,771	2,755
M-2	2,891	5,706	2,724
M-3	2,413	5,025	2,399
M-4	1,316	2,062	984
R-1	1,071	1,811	865
R-2	635	1,105	527
R-3	206	472	225
R-4	411	660	315
All Clusters	17,310	36,261	17,310

6.3.2 Visualizing and Summarizing Long-Distance Travel Trends

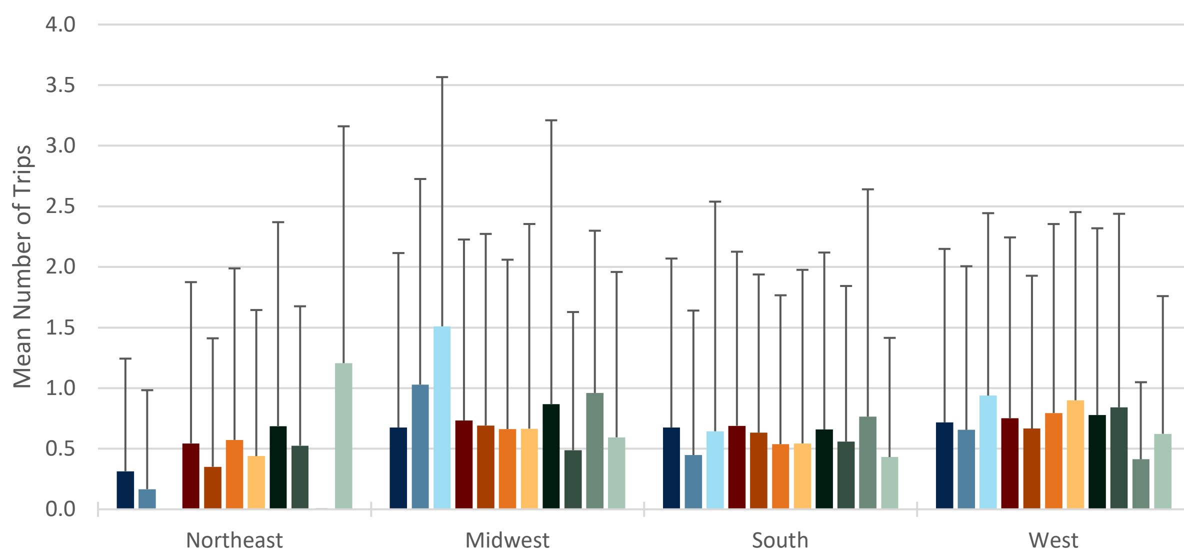
The weighted and normalized long-distance trip rates for households in each county and census region are presented in Figures 15 (all long-distance trips), 16 (by trip mode), and 17 (by trip purpose). It is important to recognize while differences between mean long-distance trip rates per household may seem small, they magnify when full county populations are considered. These figures demonstrate interesting trends:

- High long-distance trip rates are associated with all cluster types; geography does not discriminate for observed long-distance travel behavior.

- The Northeast has, on average, the lowest long-distance trip rates per household in each cluster compared to those in other census regions.
- Households in the same cluster often have different trip rates (overall, by mode, and by purpose) in different census regions.
- Rural community clusters, even with their inherently dispersed geography, have similar (and sometimes higher) long-distance trip rates than other clusters, across all census regions.
- There is no discernible relationship between personal vehicle trip rates and air travel trip rates across clusters and census regions.
- There is high trip rate variability for households in every cluster in each census region.

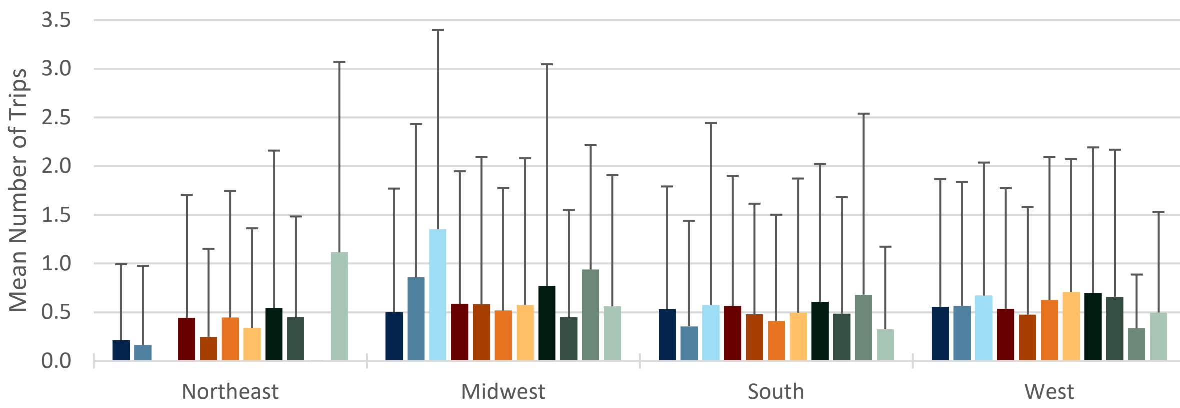


(a) Total Long-Distance Mean Trip Rates by Cluster

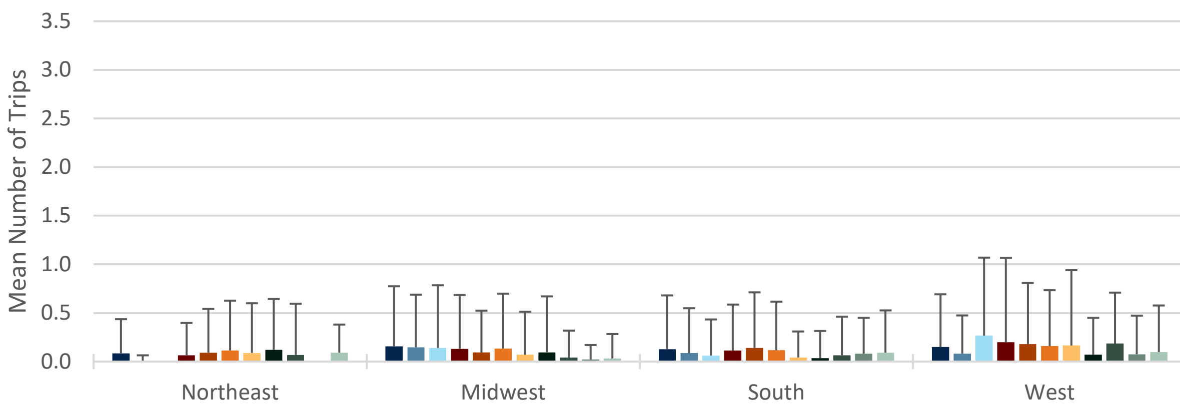


(b) Total Long-Distance Mean Trip Rates by Cluster and Census Region

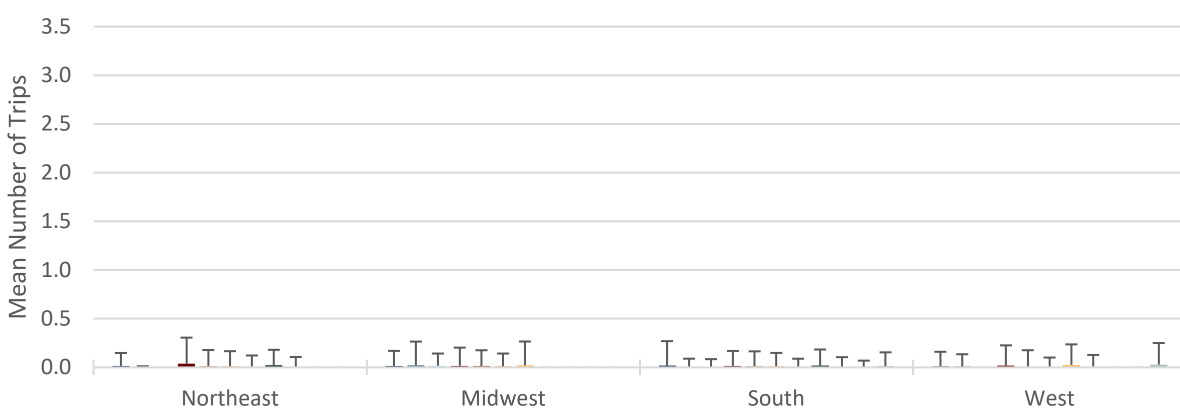
Figure 15: Distributions of Overall Mean Long-Distance Trips



(a) Mean Long-Distance Trips Completed by Personal Vehicle by Cluster and Census Region

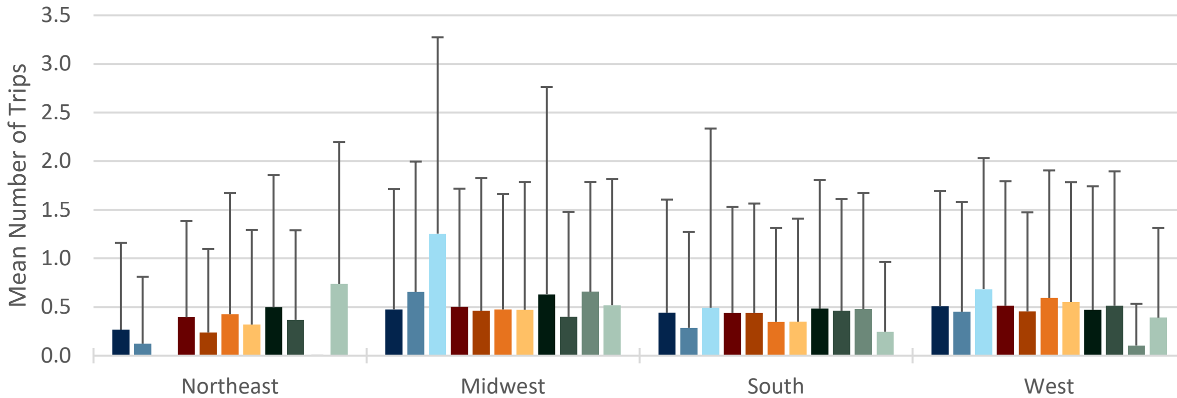


(b) Mean Long-Distance Trips Completed by Air Travel by Cluster and Census Region

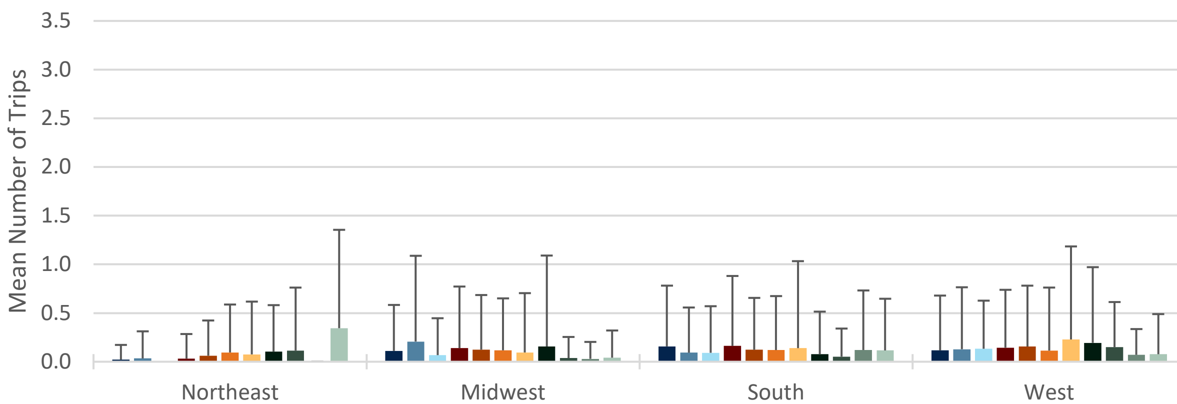


(c) Mean Long-Distance Trips Completed by Other Modes

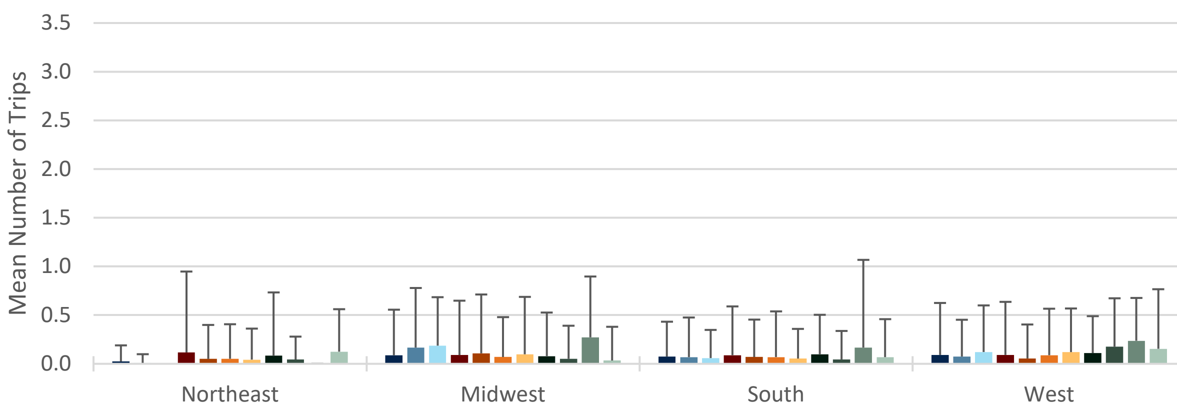
Figure 16: Distributions of Mean Long-Distance Trips by Mode



(a) Mean Long-Distance Leisure Trips by Cluster and Census Region



(b) Mean Long-Distance Work Trips by Cluster and Census region



(c) Mean Long-Distance Other Purpose Trips by Cluster and Census Region

Figure 17: Distributions of Mean Long-Distance Trips by Purpose

6.4 Differences in Trip Volumes Between Clusters & Census Regions Results

In the third step of the analysis, ANOVA and two-sample hypothesis testing was used to determine if there were statistical differences in tripmaking between classification clusters and/or census regions. An ANOVA test compares the means and variances within and between groups to determine if there are statistical differences. In this chapter, two sets of seven ANOVA tests were conducted. The seven dependent variables were the mean number of long-distance trips of the seven different types: all, personal vehicle, air travel, other modes, leisure purpose, work purpose, and other purpose trips. Two ANOVAs were calculated for a) one-way ANOVAs to test if at least one cluster exhibited statistically different tripmaking volume and b) two-way ANOVAs to test if tripmaking volume was statistically different between clusters, regions and/or both. The ANOVA testing was followed up with two-sample hypothesis tests to determine specifically which clusters and/or census regions shared statistically similar tripmaking volumes.

6.4.1 Differences Between Clusters

Results of the one-way ANOVA of trip volumes for the seven different long-distance types across the eleven clusters can be seen in Table 41. They indicated at least one cluster made a statistically different number of trips of each type (at a 99.9 percent confidence level). The results from the two-sample hypothesis tests to determine exactly which pairs of clusters produced different mean household trip volumes of each type can be seen in Figure 18. These results are organized by trip type, and cluster pairs that have statistically different volumes (at a 90 percent confidence level) are shaded blue. A tally of the differences is compiled in the Cluster Pair Uniqueness Factor figure.

Overall, there are significant differences in trip volumes between many of the clusters, however these differences are not consistent across all the types of trips. This emphasizes how important these cluster identities are, and how the nuance of the different socio-geographic areas influences the generation of different types of long-distance tripmaking. Trips made using other modes, for work, and for other purposes were less variable across all clusters, likely a reflection of the smaller sample sizes.

The urban clusters differ from each other most in leisure trip volumes but show little variation for work trip volumes. U-1 and U-2 differ on air trips, which is not surprising given the differences in airport access between those clusters. U-3 differs from U-1 and U-2 on personal vehicle trip volumes, most likely owing to this cluster's lower income compared to the other two. This also could reflect previous research demonstrating the differences in travel behavior university students have compared to the general population [185].

The mixed density clusters saw the most variability with M-1 especially differing from the rest. This especially emphasizes the accuracy of the socio-geographic clustering; even though M-1's characteristics were not notably different from the other mixed density clusters, it was unique enough to capture notably different long-distance travel behaviors. M-4 differed from the rest on air travel (again likely due to less airport access). The mixed density clusters also were demonstrably different in their tripmaking for leisure, personal vehicle, and air travel compared to the urban and rural cluster groupings, highlighting how different geographies are inherently related to different levels of long-distance travel.

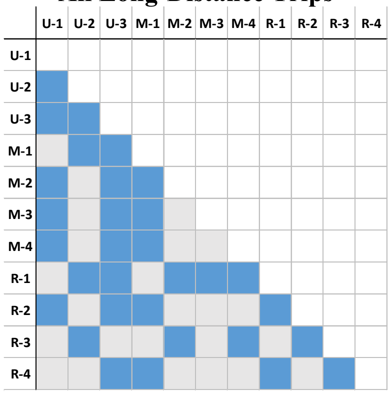
The rural clusters had differences within the group as well, but to a lesser extent. R-1 differed the most from the rest of the set (especially related to leisure and air travel); it also had the highest median income and airport access from the rural group. R-1 and R-3 took statistically the least personal vehicle trips than any other cluster, and R-2 took statistically the fewest work trips than any other cluster. Interestingly, R-3, with the oldest median age of any cluster in the nation, had statistically more “other” trip purposes than any other cluster. This finding reinforces the critical literature focused on providing long-distance access to health care for those living in rural areas.

Table 41: Mean Trip Volumes for Clusters and ANOVA Results

Cluster	**n**	Mean Number of						
		All	*Personal Vehicle*	*Air Travel*	*Other Mode*	*Leisure Purpose*	*Work Purpose*	*Other Purpose*
		Household Long-Distance Trips						
U-1	15,514	0.67	0.52	0.14	0.02	0.47	0.12	0.08
U-2	4,835	0.55	0.45	0.09	0.01	0.36	0.11	0.08
U-3	643	0.92	0.78	0.13	0.01	0.72	0.10	0.10
M-1	8,244	0.70	0.55	0.13	0.02	0.47	0.14	0.09
M-2	10,234	0.56	0.42	0.12	0.01	0.38	0.11	0.07
M-3	8,220	0.61	0.47	0.12	0.01	0.44	0.11	0.06
M-4	3,717	0.55	0.46	0.08	0.01	0.38	0.11	0.06
R-1	2,485	0.73	0.63	0.09	0.01	0.52	0.12	0.09
R-2	1,988	0.56	0.48	0.07	0.01	0.42	0.08	0.05
R-3	617	0.76	0.68	0.08	0.00	0.48	0.11	0.17
R-4	1,204	0.55	0.45	0.09	0.02	0.34	0.12	0.09
ANOVA	*F*	14.60	13.71	10.74	2.06	13.02	2.87	5.95
	Sig.	0.00	0.00	0.00	0.02	0.00	0.00	0.00

Post-Hoc Cluster Comparisons: National Level

All Long-Distance Trips

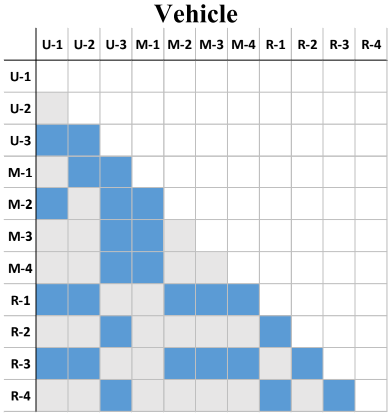


Cluster Pair Uniqueness Factor

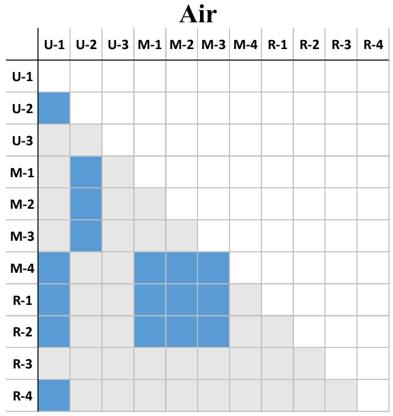
	U-1	U-2	U-3	M-1	M-2	M-3	M-4	R-1	R-2	R-3	R-4
U-1											
U-2	3										
U-3	3	3									
M-1	0	5	3								
M-2	3	1	3	5							
M-3	1	2	3	4	1						
M-4	3	0	3	5	1	1					
R-1	2	3	2	1	4	4	3				
R-2	3	0	3	4	1	1	0	2			
R-3	2	3	1	1	3	2	3	1	3		
R-4	2	0	3	2	0	0	0	3	0	3	

By Primary Mode Choice…

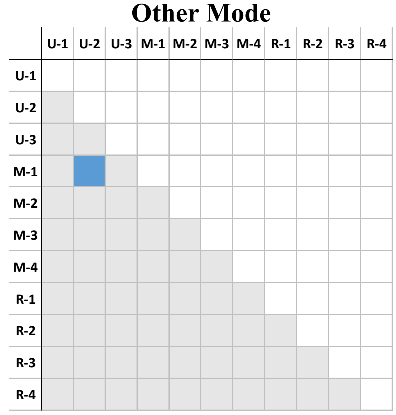
Vehicle



Air

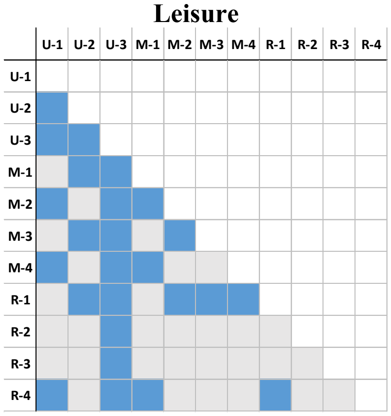


Other Mode

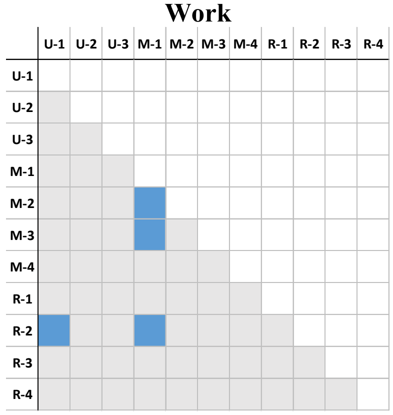


By Trip Purpose…

Leisure



Work



Other Purpose

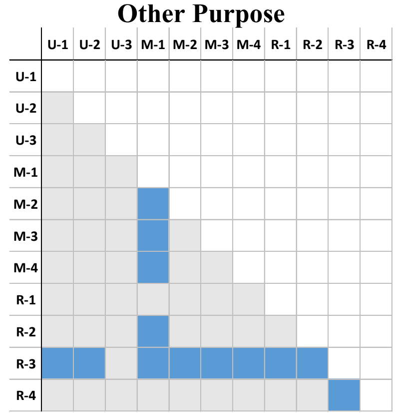


Figure 18: Post-Hoc Pairwise Cluster Comparisons (Blue Indicates Statistical Difference [90% CL])

6.4.2 Differences Between Clusters and Census Regions

To further consider if these results varied not just by cluster but also by census region, a further set of analyses was conducted. Results of the two-way ANOVA of trip volumes for the seven different long-distance types across the eleven clusters and four census regions can be seen in Table 42. They indicate a) some clusters, b) some census regions, and c) some combinations of both statistically capture variability in long-distance tripmaking (at a 99.9 percent confidence level) for all trip types. The one exception was regions do not statistically impact volumes of trips made by other modes (likely due to sample size issues). Two-sample hypothesis tests were then conducted, but this time measuring whether the mean trip volume is the same between every pair of clusters within each census region (Figures 19-22). These results are organized by census region and trip type. Cluster pairs that have statistically different volumes (at a 90 percent confidence level) are shaded blue. A tally of the differences is compiled in the Cluster Pair Uniqueness Factor figures.

Overall, we see many of the same patterns in households' long-distance trip volumes between clusters reappear in these visuals. However, comparing clusters by census region highlights a) how some trends exist regardless of region and b) other trends exist only in specific regions. For example, the volumes of trips completed by different clusters differ most in the Northeast census region. In the Northeast, outside of air travel, nearly every cluster demonstrates statistically different trip volume rates across all long-distance trip types. Air travel is consistently accessible in this region, allowing for more consistent air trips across all socio-geographic groups. These results reinforce the need for any long-distance survey sampling approach to consider all types of county clusters to capture the breadth of activity in any one region.

Alternatively, the West census region demonstrates the least variability in long-distance trip volume rates across clusters. In this region, households in urban and mixed density geographies were observed to pursue statistically different numbers of long-distance trips. Outside of that pair, however, trip rates are relatively consistent for households (even within the same category). Again, air travel rates differ the most across urban and rural communities in the West. This is likely related to the access to airports and other public transportation options. These results reinforce the need for survey clusters to consider density as an organizing characteristic.

While there are slight discrepancies, the differences between clusters in the South and Midwest census regions were surprisingly similar. This indicated these two regions share additional geographical and societal factors influencing long-distance travel beyond those considered in the socio-geographical clusters. However, one major difference is in air travel, where the Midwest has fewer differences between all the clusters (emphasizing the urban counties pursue significantly more long-distance trips than mixed density and rural counties) and the South had more variability across all cluster types. It is important to recognize even though the *patterns are similar*, the mean household long-distance trip *rates are different* between the two census regions. Therefore, these results indicate it is important to collect representative responses from each cluster in these two regions.

Table 42: Mean Trip Volumes for Clusters within Census Regions and ANOVA Results

Region	Cluster	n	*All*	*Personal Vehicle*	*Air Travel*	*Other Mode*	*Leisure Purpose*	*Work Purpose*	*Other Purpose*
			Mean Number of Household Long-Distance Trips						
Northeast	U-1	698	0.35	0.24	0.09	0.02	0.30	0.02	0.03
	U-2	389	0.17	0.17	0.00	0.00	0.13	0.03	0.01
	U-3	0	-	-	-	-	-	-	-
	M-1	721	0.65	0.53	0.08	0.05	0.47	0.04	0.14
	M-2	2792	0.42	0.29	0.11	0.02	0.29	0.07	0.06
	M-3	2804	0.62	0.48	0.12	0.01	0.46	0.10	0.06
	M-4	1341	0.55	0.42	0.11	0.01	0.41	0.09	0.05
	R-1	918	0.72	0.58	0.12	0.02	0.52	0.12	0.09
	R-2	679	0.54	0.46	0.07	0.01	0.38	0.12	0.05
	R-3	0	-	-	-	-	-	-	-
	R-4	95	1.20	1.12	0.08	0.00	0.73	0.35	0.13
Midwest	U-1	3683	0.72	0.54	0.16	0.02	0.50	0.12	0.10
	U-2	447	1.02	0.85	0.15	0.02	0.65	0.21	0.16
	U-3	143	1.54	1.37	0.15	0.01	1.27	0.07	0.20
	M-1	2045	0.82	0.66	0.14	0.02	0.56	0.16	0.10
	M-2	1793	0.74	0.62	0.10	0.02	0.49	0.13	0.12
	M-3	2324	0.69	0.54	0.14	0.01	0.50	0.12	0.07
	M-4	674	0.83	0.72	0.08	0.03	0.59	0.12	0.11
	R-1	507	0.99	0.89	0.11	0.00	0.72	0.18	0.09
	R-2	195	0.57	0.52	0.05	0.00	0.47	0.05	0.06
	R-3	27	1.19	1.15	0.04	0.00	0.81	0.04	0.33
	R-4	57	0.86	0.82	0.04	0.00	0.82	0.04	0.00
South	U-1	4462	0.68	0.54	0.13	0.02	0.45	0.16	0.07
	U-2	2619	0.46	0.37	0.09	0.00	0.29	0.10	0.07
	U-3	318	0.74	0.66	0.07	0.01	0.57	0.10	0.06
	M-1	3155	0.73	0.60	0.12	0.01	0.47	0.17	0.09
	M-2	3168	0.68	0.51	0.15	0.02	0.47	0.14	0.08
	M-3	1747	0.61	0.46	0.14	0.02	0.40	0.13	0.08
	M-4	843	0.56	0.50	0.05	0.01	0.36	0.14	0.06
	R-1	658	0.66	0.61	0.03	0.02	0.49	0.07	0.10
	R-2	888	0.57	0.50	0.07	0.01	0.47	0.06	0.05
	R-3	575	0.77	0.69	0.08	0.00	0.48	0.12	0.16
	R-4	728	0.45	0.34	0.10	0.02	0.25	0.13	0.07

Table 42: Mean Trip Volumes for Clusters within Census Regions and ANOVA Results (continued)

Region	Cluster	n	*All*	*Personal Vehicle*	*Air Travel*	*Other Mode*	*Leisure Purpose*	*Work Purpose*	*Other Purpose*
			Mean Number of Household Long-Distance Trips						
West	**U-1**	6127	0.71	0.55	0.15	0.01	0.51	0.12	0.09
	U-2	1417	0.66	0.57	0.08	0.01	0.46	0.13	0.07
	U-3	166	0.94	0.69	0.25	0.00	0.67	0.15	0.11
	M-1	1771	0.75	0.54	0.20	0.02	0.51	0.15	0.09
	M-2	1341	0.74	0.53	0.20	0.01	0.50	0.17	0.06
	M-3	776	0.81	0.64	0.16	0.01	0.60	0.12	0.09
	M-4	267	1.04	0.81	0.20	0.03	0.64	0.27	0.13
	R-1	246	0.75	0.67	0.07	0.01	0.45	0.19	0.11
	R-2	130	0.85	0.66	0.19	0.00	0.53	0.16	0.16
	R-3	15	0.40	0.27	0.13	0.00	0.13	0.07	0.20
	R-4	251	0.62	0.49	0.10	0.03	0.39	0.08	0.15
ANOVA	**Region**	*F*	24.93	23.68	7.43	0.17	22.70	3.28	6.14
		Sig.	0.00	0.00	0.00	0.92	0.00	0.02	0.00
	Cluster	*F*	9.47	11.99	5.59	2.74	8.99	2.27	3.43
		Sig.	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	Region* Cluster	*F*	6.92	7.09	3.98	1.84	5.12	4.71	2.75
		Sig.	0.00	0.00	0.00	0.00	0.00	0.00	0.00

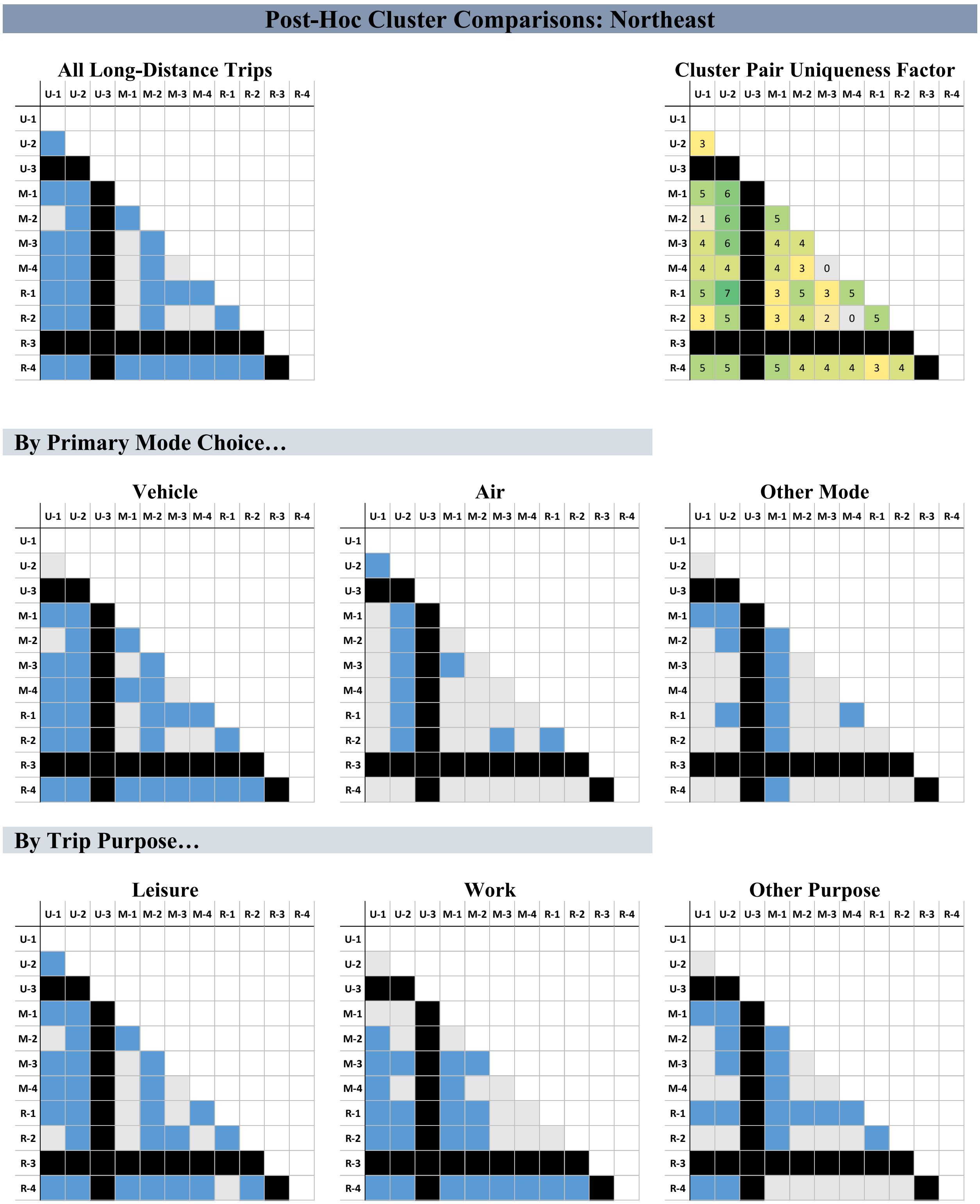


Figure 19: Post-Hoc Pairwise Cluster Comparisons for the Northeast Census Region (Black Indicates No Cluster Members; Blue Indicates Statistical Difference [90% CL])

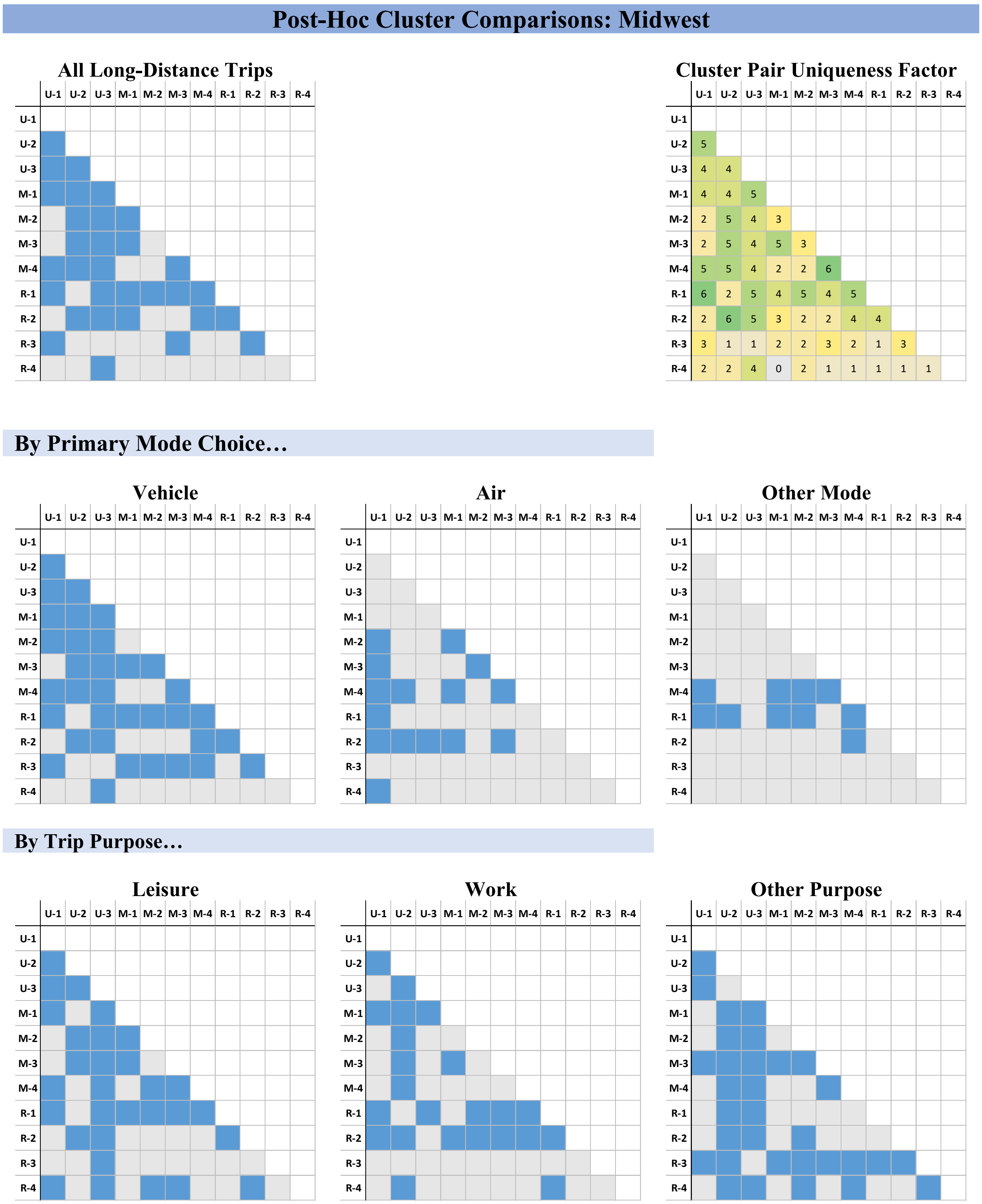
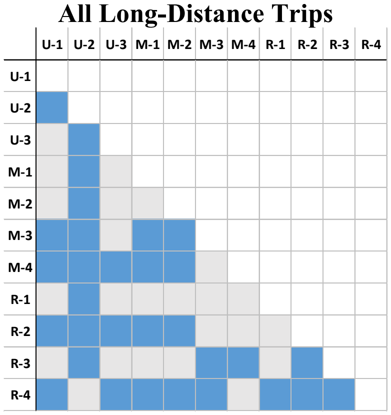


Figure 20: Post-Hoc Pairwise Cluster Comparisons for the Midwest Census Region (Blue Indicates Statistical Difference [90% CL])

Post-Hoc Cluster Comparisons: South

All Long-Distance Trips

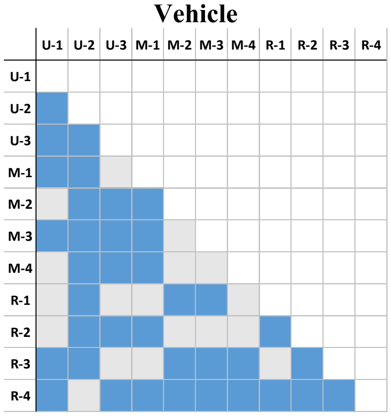


Cluster Pair Uniqueness Factor

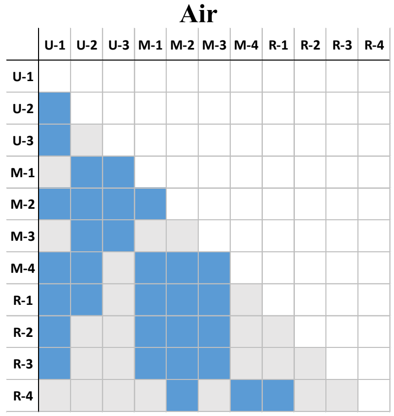
	U-1	U-2	U-3	M-1	M-2	M-3	M-4	R-1	R-2	R-3	R-4
U-1											
U-2	6										
U-3	4	3									
M-1	1	7	2								
M-2	2	6	2	3							
M-3	2	6	3	4	2						
M-4	4	4	3	5	3	1					
R-1	2	4	0	2	3	4	3				
R-2	5	4	2	5	4	3	2	2			
R-3	4	4	1	3	3	4	4	1	4		
R-4	3	0	3	4	4	3	3	5	4	4	

By Primary Mode Choice…

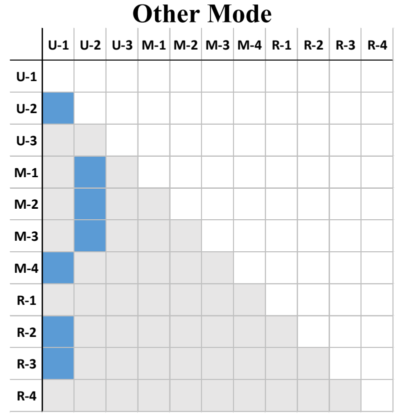
Vehicle



Air

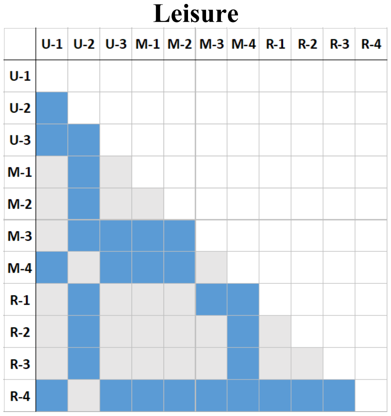


Other Mode

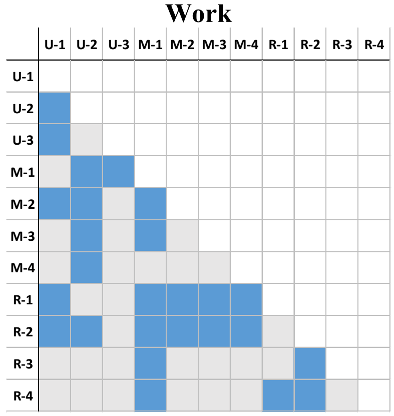


By Trip Purpose…

Leisure



Work



Other Purpose

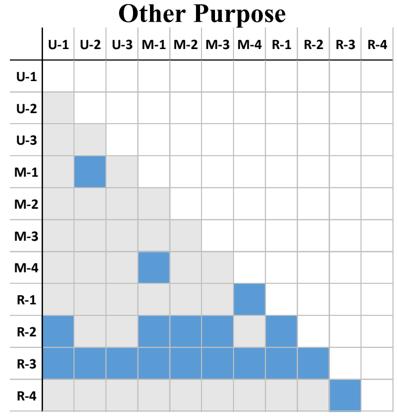
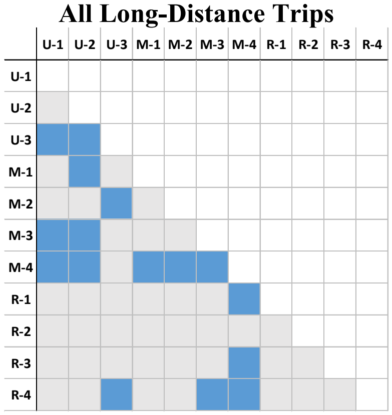


Figure 21: Post-Hoc Pairwise Cluster Comparisons for the South Census Region (Blue Indicates Statistical Difference [90% CL])

Post-Hoc Cluster Comparisons: West

All Long-Distance Trips

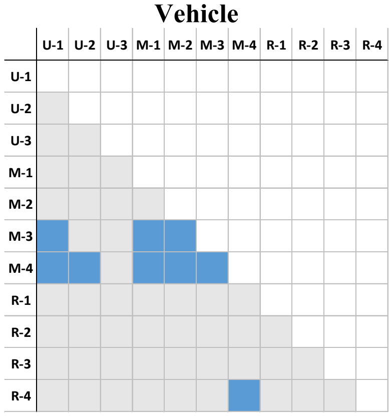


Cluster Pair Uniqueness Factor

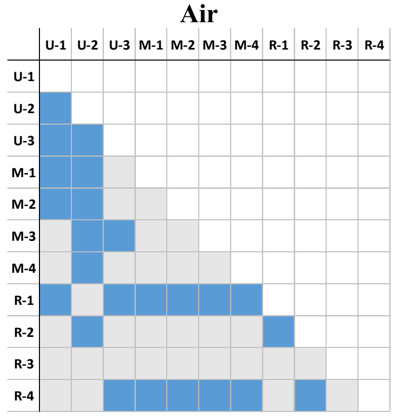
	U-1	U-2	U-3	M-1	M-2	M-3	M-4	R-1	R-2	R-3	R-4
U-1											
U-2	1										
U-3	3	3									
M-1	2	2	1								
M-2	3	2	2	1							
M-3	3	3	1	2	3						
M-4	5	7	2	3	6	4					
R-1	2	0	2	1	1	2	3				
R-2	1	2	0	0	1	0	2	1			
R-3	0	0	1	0	0	0	1	0	0		
R-4	2	2	4	3	4	5	5	1	2	0	

By Primary Mode Choice…

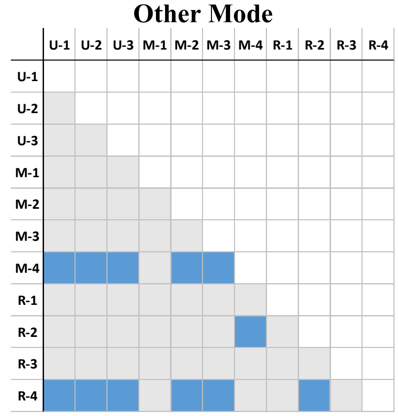
Vehicle



Air

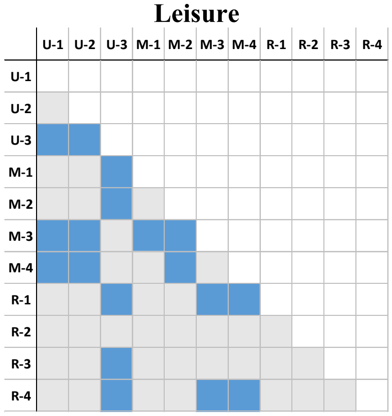


Other Mode

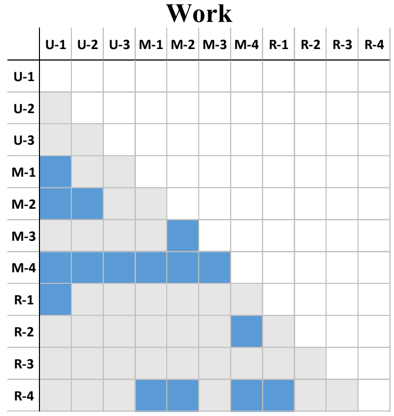


By Trip Purpose…

Leisure



Work



Other Purpose

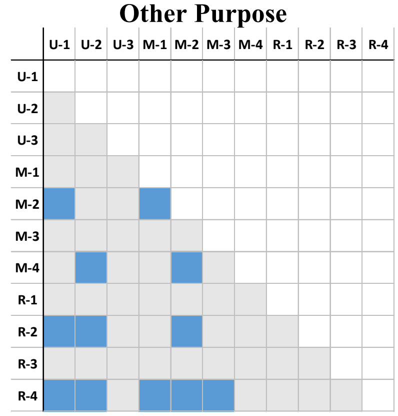


Figure 22: Post-Hoc Pairwise Cluster Comparisons for the West Census Region (Blue Indicates Statistical Difference [90% CL])

6.5 Conclusions

As long-distance travel volumes continue to increase, transportation decisionmakers are seeking data to help manage infrastructure, congestion, regional economies, and environmental impacts associated with these trips. Household travel surveys offer the detail necessary for developing policy-responsive forecasting models and other analyses but tend to focus on daily travel. Efforts to survey long-distance travelers are difficult due to the lack of a clearly defined universe or sampling frame given the lower incidence of long-distance trip-making. As such, this chapter a) developed a socio-geographic cluster classification system for stratifying US counties and b) measured differences in long-distance travel behavior across these county classifications using the 2001 NHTS. This resulted in three notable conclusions:

- U.S. counties can be successfully classified into socio-geographic clusters for the purposes of measuring long-distance tripmaking behavior. This work demonstrated 11 clusters formed around five characteristics was successful at differentiating communities based on urban development, age, income, and long-distance mode accessibility.
- Households within each county cluster pursued a) statistically similar long-distance trip rates with other households in their cluster and b) statistically different trip rates from households in other clusters at a 90 percent confidence level. This work demonstrated leisure trips, those completed by personal vehicle, and those completed by air had the most variability between clusters, and urban density had a notably important influence on this variability.
- Differences in household long-distance trip rates also extended to the census region level, meaning census regions as well as socio-geographic cluster assignment are significant

predictors of long-distance trip rates. For example, this work demonstrated clustered counties in the Northeast census region had the most variation in trip rates across the clusters, whereas the West census region had the least variability between clusters (especially in rural areas).

These results have several implications for developing more efficient and cost-effective sampling approaches for long-distance travel surveys. Most significantly, rather than collect a long-distance travel survey from a sampling approach that is drawn proportionate to the population alone, the inclusion of county-based *regional socio-geographic clusters* provide a more efficient sampling framework. This chapter highlighted 44 different regional socio-geographic clusters (11 clusters across 4 regions) which can be easily identified using readily available GIS layers and census data to create valid sampling strata from which random samples can be drawn. Additionally, there are many regional socio-geographic clusters that shared similar trip rates, indicating the number of unique clusters needed to represent in the sample framework could be reduced even further. Another important finding identified mode-availability as an essential element to consider when developing sampling frames, particularly access to different sized airports and rail service, and should cover the range of options available in the individual socio-geographic clusters (as was implemented in this study). For example, some households had access to more than one airport, so those choosing to fly to their destinations have options based on value of time and cost sensitivities, both of which influence long-distance trip making characteristics. Combined, the factors tested in this clustering approach show promise for a more focused, cost-effective means for collecting representative long-distance travel surveys.

This chapter's findings offered the third and final support to the objectives and hypotheses of this dissertation. Not only do the results support the hypotheses that similar socio-geographic areas exhibit similar long-distance travel behavior, but also showcase a new method to target long-distance travel survey sampling. Of course, there are many opportunities for further expanding this work, including creating a more robust air access measure, considering how trip rates change for specific times of the year, how response rates for different populations may be used to tailor data collection, what the minimum sample sizes might be within these regional socio-geographic clusters, how these results may change across different national survey years, and which clusters have the option of being further aggregated. However, data limitations inherent with the 2001 NHTS such as the limited sample size, data age, and travel behavior capture scope begs for this analysis to be revisited with a more recent, and larger, sample. This does not disparage the findings of this chapter, as the methods and findings here can serve as a guide for future efforts by researchers and planners to validate this approach.

Chapter 7: Sampling Framework Considerations and Conclusions

Long-distance travel continues to grow in economic, environmental, and infrastructural importance; but the lack of availability for recent national-scale long-distance travel data has led to calls from a new national long-distance travel survey. Having up-to-date travel survey data is vital to researchers, planners, and modelers alike. Modelers use this data to accurately forecast travel volumes, a necessity for DOTs and municipalities; and researchers use this data to further improve the field by exploring relationships, identifying trends, and discovering new approaches to collecting travel behavior. Both these uses of survey data have major implications on not just these communities, but also on transportation infrastructure decisions including funding, project prioritization, and policy impacts.

However, the costs of running a full-scale long-distance travel survey, both fiscally and temporally, has limited recent attempts. While there have been smaller scale, long-distance travel surveys over the past few years, the last annual-panel, national-level survey capturing long-distance travel was the 1995 ATS—nearly 30 years ago as of the writing of this dissertation. As such, survey data users have had to use older national survey data and then optimize their findings based on more recent, smaller-scale state travel surveys. These smaller-scale surveys have sampling limitations regarding either geographic scope, capture timeframe scope, or demographic groupings. While these limitations may be allowable given the immediate needs and scope associated with its original purpose, how they fair comparatively to the scope of national-level annual panel surveys is still fairly unknown.

This dissertation aimed to identify how targeted sampling frame approaches can be used by long-distance travel surveys. Having a full-sized, population-proportioned sample is always a tenant of good survey design, but the associated costs, especially for more niche subjects, can lead to an unwillingness to fully commit resources to survey deployment. This can lead to patchwork solutions to reduce costs such as with asking respondents for their long-distance travel habits over a very small timeframe or using regional survey findings to update national travel models. While these patchwork solutions offer potentially valid solutions, the actual impacts on data validity are practically unknown. This dissertation aims to fully explore this fact, by not only exploring how targeted sampling frames might affect long-distance travel survey data accuracy, but also potential approaches for creating a targeted sample frame for the purposes of capturing US long-distance travel.

To answer this question, this dissertation considered how targeting three major framing approaches—temporally, demographically, and geographically—captured the depth of long-distance travel demonstrated by three nationally-scaled long-distance travel surveys. These approaches were arranged as the following hypotheses:

- I hypothesize that annual long-distance travel exhibits *temporal* trends considering seasons and sociodemographic groups, therefore, the number of days needed in a long-distance travel survey can be reduced.
- I hypothesize that long-distance travel behavior within a sociodemographic group does not vary, therefore, the needed number of households/individuals in a sampling frame can be reduced.
- I hypothesize that similar socio-geographic groups have similar long-distance travel behavior, therefore, the geographic frame of a long-distance survey can be reduced

Each hypothesis was tested by this dissertation for their validity within its respected tested travel survey. While validation against an additional dataset would have been ideal, the already lacking availability of annual long-distance travel survey datasets and, more importantly, the lack of available detailed trip records prevents true validation testing. As such, this dissertation aimed to serve as an exploration into what could possibly work to target *national* long-distance travel survey sampling frames under future iterations.

This chapter summates the findings of this dissertation as well as provides final thoughts. This is presented as four subsections. First, the findings of targeted temporal sampling are presented. This is followed by the findings of targeted sociodemographic sampling. Next, the findings of targeted geo-economic sampling and its associated framework suggestions are given. Finally, final thoughts and suggestions for targeted long-distance travel survey sampling frameworks are presented.

7.1 Targeting the Sampling Timeframe

This chapter explores trends in the annual variability of long-distance travel trends. Long-distance travel often represents non-routine travel, such as holidays, vacations, and the occasional work conference. As such, much discussion has focused on the need to collect travel survey data for a year to fully capture the patterns of different sociodemographic groups who would be missed if only surveyed for a shorter period of time.

However, due to high costs associated with such a survey effort, it is beneficial to review trends in daily long-distance travel across these sociodemographic groups. Therefore, this research measured the variability in long-distance travel volumes per travel day throughout a year considering seasons and sociodemographics. Specifically, this chapter determined a) if there are differences in the day-by-day long-distance travel patterns across six unique sociodemographic groups across a year and b) how long-distance travel days can be clustered such that the long-distance travel volumes (reflecting the range of sociodemographic groups) are similar for days within each group *and* the travel volumes are different between groups.

These tasks were accomplished through K-means clustering days based on 30 variables considering five types of long distance HBTEs by six different sociodemographic groups. Eight clusters were identified that characterize days with extreme long-distance travel demand to days with low demand. Even though they are not part of the clustering process, the results highlighted how much these volumes are tied to seasonal travel needs and holidays.

Recalling the objectives and hypotheses of this dissertation, the findings of this chapter support the idea that long-distance travel behavior might be able to be adequately captured using a subset of representative days, even accounting for differences in travel by different sociodemographic groups. This means surveys could be focused to collect travel diaries for specific days, and travel trends for certain days could be transferrable to other similarly clustered days.

Of course, there are many opportunities for further work to explore these relationships deeper. For example, holidays and seasonal impacts vary greatly across the country, so geographic variations could be considered. Additionally, more data on different sociodemographic groups could determine if there are more unique trends not captured with the current dataset. Finally, bootstrap simulations selecting different survey days from the LSOT to determine accuracy of travel forecasts for different clusters could be completed to understand implications for minimum sample sizes. It is also suggested that this chapter's methods be applied to the upcoming NextGen NHTS for similarities in results.

7.2 Targeting Sociodemographic Groups

This chapter explored how long-distance survey sampling could be improved by targeting sociodemographic groupings. The traditional approach to travel survey sampling is population proportion sampling which considers an equal chance of capturing travel behavior based on the study area's sociodemographic distributions. However, by utilizing known patterns and behaviors inherent to long-distance travel within sociodemographic groups, the sampling effort can be targeted to reduce the overall sampling frame without losing data fidelity and maintaining capture equity, while also reducing administrative costs.

To accomplish this, three US-based surveys—the 1995 ATS, 2001 NHTS, and 2013 LSOT—were analyzed for patterns both between and within different sociodemographic groups. These sources were selected as they represent a range of time periods, slightly varied definitions of long-distance travel, and varied geographical locations. Long-distance travel analysis was completed by characterizing travel behavior considering four travel metrics a) the volume of long-distance trip making, b) median roundtrip distance, c) median trip duration, and d) the percentage of trips labeled as leisure purpose.

Compiling the results from this analysis showed the distillation of complex sociodemographic characteristics into groups defined by household income, respondent age, and the presence of children in the household to be promising for describing general long-distance travel trends. In particular, defining household income into either low, medium, or high levels using relative poverty threshold measures not only normalized the differences in each dataset's categorical income definitions, but also adequately described travel trends between groups distinctly in most cases. ANOVA and post hoc testing highlighted these insights but also shed further light on both the NHTS and LSOT surveys' flaws:

- The NHTS's short survey timeframe (28 days) reflected its inability to capture the nuances in long-distance travel particularly with trip duration and roundtrip distance. While the high-income group distinguished itself in this case, further detail and group identity among the other income groups as well as the age groups was lacking.

- The LSOT's smaller sample size and bias towards higher income individuals not only limited group calculations due to inadequate candidate pools, but also contrasted the results of the other panel survey, the ATS. While the ATS clearly highlighted group heterogeneity (especially among trip rates), the LSOT failed to at the same level.

However, the individual statistical difference results among each sociodemographic grouping variable were similar, which would suggest a retooling of the chosen grouping variable interactions. One such retooling identified was the collapse of age group 1 (under 25 years old) into a single group regardless of child presence or income level. Collapsing into a single group would not only reduce the overall number of subgroups thus alleviating small sample size issues, but also still adequately distinguish this group's long-distance travel habits.

Theil index results showed equality within groups to be rather high. The noted exception was with trip duration and leisure trip percentage within the NHTS, reflecting the survey's shorter survey period. Roundtrip distance results were rather comparable amongst all three datasets which would suggest long-distance travelers act similarly, at least with median roundtrip distance, when controlled for outlier trips.

Finally, sampling techniques had two major findings: first, the needed valid sample size, even controlled for outliers, was multitudes smaller than the actual survey sample size (in particular the national-level surveys). Second, non-traveling households need to be identified and accounted for in sample size calculations. Regarding the second finding, only the NHTS offered any accurate national counts of these non-traveling households on the defined group level. While the LSOT did have some non-traveling households participate throughout the survey's duration, this was identified as five records, hardly a sample size from which to make conclusions. Surface-level analysis suggested up to 83 percent of those surveyed in the 2001 NHTS did not partake in any long-distance travel over the survey period. It would then be recommended the defined sample sizes be doubled or even tripled to account for these non-travelers for non-panel, annual surveys. However, surface-level ATS numbers would suggest only 39 percent of those surveyed were non-traveling households, which is a decent reduction, but individual group capture rates were still unidentifiable given the full ATS survey dataset is unavailable for analysis. It was therefore recommended further analysis on identifying non-traveling rather than traveling households would be beneficial in determining sample size calculations, particularly in the application of non-panel surveys.

While more work is needed to better identify non-traveling households, planners and researchers can still utilize the general results of this dissertation to better approach long-distance travel surveying efforts. Overall, this chapter supported the second objective and hypothesis identified for this dissertation. The long-distance sociodemographic sampling process can be streamlined in a way that a) reduces the overall number of surveys needed for a statistically valid sample, up to 97 percent; and b) employs a universal household categorical system that limits the necessary data fidelity reducing survey burden while also supporting respondent privacy. This system also allows for easier comparisons between differing survey datasets by harmonizing the sociodemographic definition universe; with this system employed throughout this dissertation.

7.3 Targeting the Geographic Capture Area

The last sampling frame targeting approach this dissertation considered was the idea of concentrating the geographic scale of a survey. Long-distance travel is difficult to capture using traditional household travel survey methods, given both the lower incidence of travel as well as the lack of a sampling frame or universe of long-distance travelers. Building on focused household travel survey sampling techniques that have been studied extensively related to *demographics*, this approach also considers how *regional geography* could be used to target long-distance traveling households [46, 168, 169]. As such, this chapter a) developed a socio-geographic cluster classification system for stratifying US counties and b) measured differences in long-distance travel behavior across these county classifications using the 2001 NHTS. This resulted in three notable conclusions:

- U.S. counties can be successfully classified into socio-geographic clusters for the purposes of measuring long-distance tripmaking behavior. This work demonstrated 11 clusters formed around five characteristics was successful at differentiating communities based on urban development, age, income, and long-distance mode accessibility.
- Households within each county cluster pursued a) statistically similar long-distance trip rates with other households in their cluster and b) statistically different trip rates from households in other clusters. This work demonstrated leisure trips, those completed by personal vehicle, and those completed by air had the most variability between clusters, and urban density had a notably important influence on this variability.
- Differences in household long-distance trip rates also extended to the census region level, meaning census regions as well as socio-geographic cluster assignment are significant predictors of long-distance trip rates. For example, this work demonstrated clustered counties in the Northeast census region had the most variation in trip rates across the clusters, whereas the West census region had the least variability between clusters (especially in rural areas).

These results have several implications for developing more efficient and cost-effective sampling approaches for long-distance travel surveys. Most significantly, rather than collect a long-distance travel survey from a sampling approach that is drawn proportionate to the population alone, the inclusion of county-based *regional socio-geographic clusters* could provide a more efficient sampling framework. This chapter highlighted 44 different regional socio-geographic clusters (11 clusters across 4 regions) which can be easily identified using readily available GIS layers and census data to create valid sampling strata from which random samples can be drawn. Additionally, there are many regional socio-geographic clusters that shared similar trip rates, indicating the number of unique clusters needed to represent in the sample framework could be reduced even further. Another important finding identified mode-availability as an essential element to consider when developing sampling frames, particularly access to different sized airports and rail service, and should cover the range of options available in the individual socio-geographic clusters (as was implemented in this study). For example, some households had access to more than one airport, so those choosing to fly to their destinations have options based on value of time and cost sensitivities, both of which influence long-distance trip making characteristics. Combined, the factors tested in this clustering approach show promise for a more focused, cost-effective means for collecting representative long-distance travel surveys.

This chapter's findings offered the third and final support to the objectives and hypotheses of this dissertation. Not only do the results support the hypotheses that similar socio-geographic areas exhibit similar long-distance travel behavior, but also showcase a new method to target long-distance travel survey sampling. Of course, there are many opportunities for further expanding this work, including considering how trip rates change for specific times of the year, how response rates for different populations may be used to tailor data collection, what the minimum sample sizes might be within these regional socio-geographic clusters, how these results may change across different national survey years, and which clusters have the option of being further aggregated. However, data limitations inherent with the 2001 NHTS such as the limited sample size, data age, and travel behavior capture scope begs for this analysis to be revisited with a more recent, and larger, sample. This does not disparage the findings of this chapter, as the methods and findings here can serve as a guide for future efforts by researchers and planners to validate this approach.

7.4 Final Conclusions

This dissertation aimed to identify how targeted sampling frame approaches could be used by long-distance travel surveys. Results from this dissertation showed how daily travel variation patterns, sociodemographic patterns, and geo-economic patterns could be used to better sample for long-distance travel, reducing fiscal and temporal costs. At the same time, these reductions can still provide statistically viable samples for sociodemographic groupings, travel volumes, mode splits, and purpose splits comparative to full-scale national surveys like the 1995 ATS, 2001 NHTS, and 2013 LSOT.

Practitioners can use the findings and lessons from this dissertation in their future long-distance travel survey design regardless of scale. Particularly regarding the reductions in the needed sampling pool at the national level. Chapter five illustrated how the needed sample for national-level surveys could be reduced up to 97 percent from tens of thousands, to just a few thousand. While this mainly focuses for national-level surveys, the lessons could still be applied to state and region-level surveys to consider that their needed samples may be much less than they think. Additionally, this survey’s findings can help guide current and future works as a general baseline for QA/QC purposes. As not many long-distance travel surveys exist, there is a lack of control data to confirm findings/trends are as expected. This dissertation's findings could serve as a general trend guide (such as with general sociodemographic trends and annual travel patterns) to ensure survey findings behave correctly.

While this dissertation shows promise in creating targeted long-distance travel survey sampling frames, it is not without its caveats. Chiefly, the unavailability of other, more recent national long-distance travel survey data prevents full validity testing of these findings. Thus, it is suggested that this dissertation's findings be revisited for retesting once a newer national long-distance travel behavior dataset, such as the NextGen NHTS, is available. Additional caveats for this dissertation's chapters include limited sample sizes, dataset ages, and travel behavior fidelity which, while not disparaging of any findings, did prevent some additional analysis from being completed.

Chapter 8: Future Work and Final Thoughts

This dissertation presented a concerted effort illustrating how the national long-distance travel survey sampling frame can be targeted to reduce overall costs. However, there are many other opportunities in exploring long-distance travel moving forward. Beyond what can be expanded upon in the concepts discussed in this dissertation, there is also a need to discuss the general nature of the long-distance travel behavior field as a whole. This chapter covers several potential future work ideas and general thoughts on where we go moving forward.

8.1 Further Validation Efforts

The first topic would be on further validating this dissertation's works. All three approaches offer promise but given the impacts of the COVID-19 pandemic, everchanging travel trends, and data limitations, newer sources of travel data would further warrant merit and update these findings. For the timeframe approach, only one year of travel data was available at the detail level required for analysis. Having another dataset would be helpful in validating findings as well as further exploring how the weekly position of some major "floating" travel holidays (such as July 4th, Christmas, and New Year's) influences travel volumes before and after the holiday. Additionally, exploring this approach by using traffic volume counts (a much more readily available data source) could be another way to determine the effects these "floating" holidays have on vehicle travel volumes. For the sociodemographic sampling approach, having an additional dataset to test the groupings and minimum sampling approaches would be beneficial to warrant further validation.

Finally, for the geographic approach, monitoring a) how US counties have changed in composition since 2001, and b) how the associated travel behavior differences may have changed would be highly beneficial. Therefore, it is suggested that these methods be tested against a new, preferably nationally-scaled survey for validation. In the context of the NextGen NHTS being completed this year which does ask for some long-distance travel history, at least the geographic approach can be compared to more recent travel data. For the other two approaches, smaller validation efforts can be made but with the noted caveat of comparing aggregated travel volumes/behavior rather than a year-long description of travel behavior by respondents.

8.2 Combining Approaches into a Single Method

The second topic is how this dissertation's three targeting approaches could be combined in the future. There are natural interrelationships between geographies, times, and sociodemographics. By taking these three approaches used in this dissertation and exploring the correlations and patterns inherent between the approaches, a more efficient sampling frame for national long-distance travel surveys can be found. This approach also needs to consider the validation concerns and limitations mentioned throughout this dissertation.

8.3 Redefining Long-Distance Travel

The third topic is the very definition of long-distance travel. Recalling Chapter two and three's discussions of long-distance modeling and surveying, there many definitions of what constituted a long-distance trip: travel distance, duration, mode choice, or travel time. With so many differing opinions of what should be considered a long-distance trip, there have understandably been efforts to standardize this definition. Discussions at the 2022 International Steering Committee for Travel Survey Conferences (ISCTSC) conference with worldwide colleagues suggested that defining long-distance travel could be approached in two major ways. First, is the idea that long-distance travel, as defined currently by distance/duration/etc., can be further subdivided into regular and irregular travel. Second, is redefining the definition of long-distance travel to consider a) a more respondent-unique approach, and b) an approach better suited to future passive data gathering efforts.

8.3.1 Regular versus Irregular Long-Distance Tripmaking

Long-distance travel can be viewed in two different classes: regular and irregular travel. A daily commuter trip where the respondent crosses the 50-mile one-way travel threshold widely used for defining long-distance travel is considered long-distance, but this type of trip can be captured in a standard daily travel survey format. However, a respondent going on a vacation trip either once a year or for the first time, would be less likely to have this travel captured without the aid of a panel-type survey, thus being an irregular trip.

Therefore, it is suggested that surveys capturing long-distance travel add an additional question asking for trip regularity status. This designation would help further define what type of travel is captured adding needed detail to daily travel surveying formats as well as better defined long-distance travel volumes for modeling purposes.

8.3.2 Universal Definition of Long-Distance Travel

The consensus of what constitutes long-distance travel are those trips that cannot normally be captured in a daily travel survey—truly the non-routine trips one would take. This is regardless of distance, duration, mode choice, or crossed planning boundaries. Given the current designation format, maybe another definition would be better for the field. For example, European colleagues found that defining long-distance travel by a certain distance tended to not be as helpful for them since urban density is much higher than that of the United States. Additionally, the usage of a distance metric is arbitrary since a) metric and imperial units are not equal, and b) a respondent is less likely to recall the last time they went 50-miles, accurately, when asked.

Therefore, it is suggested that a consensus on what should be defined as long-distance travel be made moving forward. This definition should ideally be internationally consistent and should attempt to avoid using any defined travel distance. While having a definitive distance is convenient for researchers, planners, and modelers, it is not the best aspect for capturing non-routine travel and is not the easiest way to ask survey respondents for travel recall.

8.4 The Future of Active Long-Distance Travel Surveying

Finally, with the continuous improvements to technology and our understandings of travel behavior, future efforts in active travel surveying should take advantage of patterns, available data, and emerging technologies to reduce sampling efforts and respondent burden. One such way is with the merging of passive data with active data collection. While this data fusion is already done to an extent, the concept of accurately inferring mode choice, purpose, household characteristics, and trip party characteristics is still in its infancy. Some of these inference choices such as mode choice and trip purpose can be more easily inferred than household and trip party characteristics given origin-destination and travel time/distance/route information, but the latter are still dependent on survey data for likelihoods. As such, it is most likely that a hybridized, smaller-scale, constant method of travel capture, maybe with a GPS-supported smartphone survey, such as RSG's rMove, will be the future of travel surveys. This is already seen partially with the NextGen NHTS series which will be occurring biennially starting in 2022, with a smaller sample size traditional 24-hour recall travel diary active survey component of 7,500 households, and with an origin-destination component [186]. This approach will provide data users with a constant stream of nationally-focused travel behavior data, potentially allowing for quicker updates to travel mode trends, purposes, destinations, and origins.

References

1. Nelson, A. and R. Lang, *Megapolitan america*. 2018: Routledge.
2. Larsen, J., K.W. Axhausen, and J. Urry, *Geographies of social networks: meetings, travel and communications.* Mobilities, 2006. **1**(2): p. 261-283.
3. Aultman-Hall, L., et al., *The implications of long-distance tour attributes for national travel data collection in the United States.* Transportation, 2018. **45**(3): p. 875-903.
4. Statistics, U.S.B.o.L., *Consumer Price Index for All Urban Consumers: Airline Fares in U.S. City Average [CUSR0000SETG01]*. 2021: Federal Reserve Bank of St. Louis.
5. Aultman-Hall, L., *Incorporating Long-Distance Travel into Transportation Planning in the United States.* 2018.
6. Association, U.S.T., *U.S. Travel Answer Sheet*. 2020, U.S. Travel Association: www.ustravel.org.
7. Statistics, B.o.T., *National Transportation Statistics: Table 1-40: U.S. Passenger Miles (Millions)*. 2020, U.S. Department of Transportation: www.bts.gov.
8. Hu, P.S. and T.R. Reuscher, *Summary of travel trends: 2001 national household travel survey.* 2004.
9. Schäfer, A.W., *Long-term trends in domestic US passenger travel: the past 110 years and the next 90.* Transportation, 2017. **44**(2): p. 293-310.
10. Association, U.S.T., *U.S. Travel and Tourism Overview (2015)*, U.S.T. Association, Editor. 2016.
11. Aratuo, D.N. and X.L. Etienne, *Industry level analysis of tourism-economic growth in the United States.* Tourism Management, 2019. **70**: p. 333-340.
12. Alegre, J. and L. Pou, *US household tourism expenditure and the Great Recession: An analysis with the Consumer Expenditure Survey.* Tourism economics, 2016. **22**(3): p. 608-620.
13. Smeral, E., *Impacts of the world recession and economic crisis on tourism: Forecasts and potential risks.* Journal of Travel Research, 2010. **49**(1): p. 31-38.
14. Ritchie, J.R.B., C.M. Amaya Molinar, and D.C. Frechtling, *Impacts of the world recession and economic crisis on tourism: North America.* Journal of travel research, 2010. **49**(1): p. 5-15.
15. Ito, H. and D. Lee, *Assessing the impact of the September 11 terrorist attacks on US airline demand.* Journal of Economics and Business, 2005. **57**(1): p. 75-95.
16. Blunk, S.S., D.E. Clark, and J.M. McGibany, *Evaluating the long-run impacts of the 9/11 terrorist attacks on US domestic airline travel.* Applied economics, 2006. **38**(4): p. 363-370.
17. Bonham, C., C. Edmonds, and J. Mak, *The impact of 9/11 and other terrible global events on tourism in the United States and Hawaii.* Journal of Travel Research, 2006. **45**(1): p. 99-110.
18. Gössling, S., D. Scott, and C.M. Hall, *Pandemics, tourism and global change: a rapid assessment of COVID-19.* Journal of Sustainable Tourism, 2020. **29**(1): p. 1-20.
19. Folinas, S. and T. Metaxas, *Tourism: The great patient of coronavirus COVID-2019.* 2020.
20. Lyulyov, O., et al., *The Link Between Economic Growth and Tourism: Covid-19 Impact.* Proceedings of the 36th International Business Information Management Association (IBIMA), 2020: p. 4-5.

21. Wilkerson, C., *Travel and tourism: An overlooked industry in the US and Tenth District.* Economic Review-Federal Reserve Bank of Kansas City, 2003. **88**(3): p. 45-72.
22. Becken, S., *Analysing international tourist flows to estimate energy use associated with air travel.* Journal of sustainable tourism, 2002. **10**(2): p. 114-131.
23. Gössling, S. and P. Peeters, *Assessing tourism's global environmental impact 1900–2050.* Journal of Sustainable Tourism, 2015. **23**(5): p. 639-659.
24. Dubois, G. and J.P. Ceron, *Tourism/leisure greenhouse gas emissions forecasts for 2050: Factors for change in France.* Journal of Sustainable Tourism, 2006. **14**(2): p. 172-191.
25. Becken, S., D.G. Simmons, and C. Frampton, *Energy use associated with different travel choices.* Tourism Management, 2003. **24**(3): p. 267-277.
26. Obua, J., *Environmental impact of ecotourism in Kibale national park, Uganda.* Journal of Sustainable Tourism, 1997. **5**(3): p. 213-223.
27. Eagles, P.F.J. and S.F. McCool, *Tourism in national parks and protected areas: Planning and management*. 2002: Cabi.
28. Burakowski, E. and M. Magnusson, *Climate impacts on the winter tourism economy in the United States.* 2012.
29. Koenig, U. and B. Abegg, *Impacts of climate change on winter tourism in the Swiss Alps.* Journal of sustainable tourism, 1997. **5**(1): p. 46-58.
30. Scott, D., J. Dawson, and B. Jones, *Climate change vulnerability of the US Northeast winter recreation–tourism sector.* Mitigation and adaptation strategies for global change, 2008. **13**(5): p. 577-596.
31. *2001 National Household Travel Survey User's Guide*, B.o.T. Statistics, Editor. 2004.
32. National Research Council . Committee to Review the Bureau of Transportation Statistics' Survey, P., *Measuring Personal Travel and Goods Movement: A Review of the Bureau of Transportation Statistics' Surveys.* 2004.
33. National Academies of Sciences, E.a.M., *Interregional Travel: A New Perspective for Policy Making*. 2016: National Academies Press.
34. National Academies of Science, E., and Medicine, *Long-Distance and Rural Travel Transferable Parameters for Statewide Travel Forecasting Models*. 2012.
35. Krause, C.M. and L. Zhang, *Short-term travel behavior prediction with GPS, land use, and point of interest data.* Transportation Research Part B: Methodological, 2019. **123**: p. 349-361.
36. Lu, Y. and L. Zhang, *Imputing trip purposes for long-distance travel.* Transportation, 2015. **42**(4): p. 581-595.
37. Chen, C., et al., *The promises of big data and small data for travel behavior (aka human mobility) analysis.* Transportation research part C: emerging technologies, 2016. **68**: p. 285-299.
38. Çolak, S., et al., *Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities.* Transportation Research Record, 2015. **2526**(1): p. 126-135.
39. Toole, J.L., et al., *The path most traveled: Travel demand estimation using big data resources.* Transportation Research Part C: Emerging Technologies, 2015. **58**: p. 162-177.
40. Xiao, G., Z. Juan, and C. Zhang, *Travel mode detection based on GPS track data and Bayesian networks.* Computers, Environment and Urban Systems, 2015. **54**: p. 14-22.

41. Yang, D., et al., *Travel Mode Detection using Smartphone GPS Data: a Comparison between Random Forest and Wide-and-Deep Learning*. 2019.
42. Shin, D., et al., *Urban sensing: Using smartphones for transportation mode classification.* Computers, Environment and Urban Systems, 2015. **53**: p. 76-86.
43. Lin, E. *Representativeness Analysis: How Our Data Reflects the Real Labor Market Dynamics*. 2017; Available from: https://www.dssgfellowship.org/2017/10/06/representativeness_analysis/.
44. Taherdoost, H., *Sampling methods in research methodology; how to choose a sampling technique for research.* How to Choose a Sampling Technique for Research (April 10, 2016), 2016.
45. *Glossary of Statistical Terms*. 2007, Organization for Economic Co-Operation and Development. p. 863.
46. Hu, P.S., T. Reuscher, and R.L. Schmoyer, *Transferring 2001 National Household Travel Survey*, F.H. Administration, Editor. 2007 Department of Energy.
47. Assaf, A.G., F. Kock, and M. Tsionas, *Tourism during and after COVID-19: An Expert-Informed Agenda for Future Research.* Journal of Travel Research, 2022. **61**(2): p. 454-457.
48. Rice, W.L., et al., *Changes in recreational behaviors of outdoor enthusiasts during the COVID-19 pandemic: analysis across urban and rural communities.* Journal of Urban Ecology, 2020. **6**(1): p. juaa020.
49. Landry, C.E., et al., *How has the COVID-19 pandemic affected outdoor recreation in the US? A revealed preference approach.* Applied Economic Perspectives and Policy, 2021. **43**(1): p. 443-457.
50. Mackenzie, S.H. and J. Goodnow, *Adventure in the age of COVID-19: Embracing microadventures and locavism in a post-pandemic world.* Leisure Sciences, 2020. **43**(1-2): p. 62-69.
51. Prey, J., D.W. Marcouiller, and D. Kim, *Economic Impacts of the Wisconsin State Park System: Connections to Gateway Communities*, W.D.o.N. Resources, Editor. 2013: Madison, Wisconsin. p. 50.
52. Chan, Y. and T.O. Carroll, *Estimating recreational travel and economic values of state parks.* Journal of urban planning and development, 1985. **111**(1): p. 65-79.
53. Stynes, D. and J. Daniel, *Economic significance of recreational uses of national parks and other public lands.* 2005.
54. Stevens, T.H., T.A. More, and M. Markowski-Lindsay, *Declining national park visitation: An economic analysis.* Journal of leisure research, 2014. **46**(2): p. 153-164.
55. Foglesong, R., *Walt Disney World and Orlando*, in *The Tourist City*. 1999, Yale University Press. p. 89-106.
56. National Academies of Sciences, E., and Medicine, *Close to Home: A Handbook for Transportation-Efficient Growth in Small Communities and Rural Areas*. 2015: Washington, DC.
57. Zhou, H., R. Bouyekhf, and A. El Moudni, *Concept of Transportation Entropy and Its Application in Traffic Signal Control.* IFAC Proceedings Volumes, 2013. **46**(25): p. 37-42.
58. EPA, *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990 - 2019*, U.S.E.P. Agency, Editor. 2021.
59. Brenan, M. *Water Pollution Remains Top Environmental Concern in U.S.* 2021.

60. O'Connor, P. and G. Assaker, *COVID-19's effects on future pro-environmental traveler behavior: an empirical examination using norm activation, economic sacrifices, and risk perception theories.* Journal of Sustainable Tourism, 2021: p. 1-19.
61. Bhowmik, A.K. *Flight shaming: how to spread the campaign that made Swedes give up flying for good*. The Conversation, 2020.
62. *Sweden sees rare fall in air passengers, as flight-shaming takes off*. BBC News, 2020.
63. Perrine, K.A., K.M. Kockelman, and Y. Huang, *Anticipating long-distance travel shifts due to self-driving vehicles.* Journal of Transport Geography, 2020. **82**: p. 102547.
64. LaMondia, J.J., et al., *Shifts in long-distance travel mode due to automated vehicles: Statewide mode-shift simulation experiment and travel survey analysis.* Transportation Research Record, 2016. **2566**(1): p. 1-11.
65. Pukhova, A., et al., *Agent-Based Simulation of Long-Distance Travel: Strategies to Reduce CO2 Emissions from Passenger Aviation.* Urban Planning, 2021. **6**(2): p. 271-284.
66. Sivaraman, V., *Analysis of long-distance vacation travel demand in the United States: a multiple discrete–continuous choice fra.* Transportation, 2013. **40**: p. 151-171.
67. Zanni, A.M., et al., *Improving scenario methods in infrastructure planning: A case study of long distance travel and mobility in the UK under extreme weather uncertainty and a changing climate.* Technological Forecasting and Social Change, 2017. **115**: p. 180-197.
68. Hand, M.S., et al., *Effects of climate change on outdoor recreation [Chapter 10].* In: Halofsky, Jessica E.; Peterson, David L.; Ho, Joanne J.; Little, Natalie, J.; Joyce, Linda A., eds. Climate change vulnerability and adaptation in the Intermountain Region [Part 2]. Gen. Tech. Rep. RMRS-GTR-375. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 316-338., 2018. **375**: p. 316-338.
69. Weiner, E., *Urban transportation planning in the United States: an historical overview*. 1997: US Department of Transportation.
70. Cordero, F., *State of the Practice of Long Distance and Intercity Travel Modeling in US Metropolitan Planning Organizations and State Departments of Transportation.* 2019.
71. Outwater, M., et al., *Foundational Knowledge to Support a Long-Distance Passenger Travel Demand Modeling Framework*, F.H. Administration, Editor. 2015, US Department of Transportation.
72. Lu, Y., et al., *National Travel Demand Model for the United States: Applications to Statewide Modeling, Fuel Price, and HSR*. 2018.
73. *America on the Go... Findings from the National Household Travel Survey*, B.o.T. Statistics, Editor. 2006, U.S. Department of Transportation: bts.gov.
74. Algers, S., *Integrated structure of long-distance travel behavior models in Sweden.* Transportation Research Record, 1993. **1413**: p. 141-141.
75. Cho, H.D., *The factors that affect long-distance travel mode choice decisions and their implications for transportation policy*. 2013: University of Florida.
76. Limtanakool, N., M. Dijst, and T. Schwanen, *The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium-and longer-distance trips.* Journal of transport geography, 2006. **14**(5): p. 327-341.
77. Blackstone, E.A., A.J. Buck, and S. Hakim, *Determinants of airport choice in a multi-airport region.* Atlantic Economic Journal, 2006. **34**(3): p. 313-326.
78. Hess, S., *Advanced discrete choice models with applications to transport demand*. 2005, University of London.

79. Kim, S. and G.F. Ulfarsson, *Travel mode choice of the elderly: effects of personal, household, neighborhood, and trip characteristics.* Transportation Research Record, 2004. **1894**(1): p. 117-126.
80. LaMondia, J., C.R. Bhat, and D.A. Hensher, *An annual time use model for domestic vacation travel.* Journal of Choice Modelling, 2008. **1**(1): p. 70-97.
81. Mallett, W.J., *Long-distance travel by low-income households.* TRB Transportation Research Circular E-C026—Personal travel: The long and short of it, 2001: p. 169-177.
82. Dargay, J.M. and S. Clark, *The determinants of long distance travel in Great Britain.* Transportation Research Part A: Policy and Practice, 2012. **46**(3): p. 576-587.
83. Rasmidatta, I., *Mode Choice Models for Long Distance Travel in USA.* 2007.
84. Carlsson, F., *Private vs. Business and Rail vs. Air Passengers: willingness to pay for transport attributes.* rapport nr.: Working Papers in Economics, 1999(1999).
85. Monzon, A. and A. Rodríguez-Dapena, *Choice of mode of transport for long-distance trips: Solving the problem of sparse data.* Transportation Research Part A: Policy and Practice, 2006. **40**(7): p. 587-601.
86. Wilson, F.R., S. Damodaran, and J.D. Innes, *Disaggregate mode choice models for intercity passenger travel in Canada.* Canadian Journal of Civil Engineering, 1990. **17**(2): p. 184-191.
87. Liss, S., et al., *Our Nation's Travel: Current Issues-2001 National Household Travel Survey (NHTS)*. 2005, United States. Federal Highway Administration.
88. Thrane, C., *Examining tourists' long-distance transportation mode choices using a Multinomial Logit regression model.* Tourism Management Perspectives, 2015. **15**: p. 115-121.
89. Shaw, S. and C. Thomas, *Discussion note: Social and cultural dimensions of air travel demand: Hyper-mobility in the UK?* Journal of Sustainable Tourism, 2006. **14**(2): p. 209-215.
90. Collia, D.V., J. Sharp, and L. Giesbrecht, *The 2001 national household travel survey: A look into the travel patterns of older Americans.* Journal of safety research, 2003. **34**(4): p. 461-470.
91. Limtanakool, N., M. Dijst, and T. Schwanen, *On the participation in medium-and long-distance travel: A decomposition analysis for the UK and the Netherlands.* Tijdschrift voor economische en sociale geografie, 2006. **97**(4): p. 389-404.
92. De Lapparent, M., A. Frei, and K.W. Axhausen, *Choice of mode for long distance travel: current SP-based models from three European countries*. 2009: Citeseer.
93. Mallett, W.J., *Long-distance travel by women: results from the 1995 American travel survey.* transportation research record, 1999. **1693**(1): p. 71-78.
94. LaMondia, J., T. Snell, and C.R. Bhat, *Tourism Travel within the European Union: The Impact of Personal Preferences and Perceptions on Vacation Destination and Travel Mode Choices.* 2008.
95. McGuckin, N., J. Casas, and M. Wilaby, *MI Travel Counts III: Travel Characteristics - Technical Report*. 2016, Michigan Department of Transportation: Lansing, Michigan. p. 76.
96. Outwater, M.L., et al., *Attitudinal market segmentation approach to mode choice and ridership forecasting: Structural equation modeling.* Transportation Research Record, 2003. **1854**(1): p. 32-42.

97. Davison, L. and T. Ryley, *The relationship between air travel behaviour and the key life stages of having children and entering retirement.* Journal of Transport Geography, 2013. **26**: p. 78-86.
98. Aultman-Hall, L., et al., *Design and response quality in a one-year longitudinal survey of overnight and long-distance travel.* Transportation Research Procedia, 2015. **11**: p. 136-153.
99. Selke, L., *Top 10 Worst Travel Days*. 2017 USA Today.
100. Madre, J.-L., K.W. Axhausen, and W. Brög, *Immobility in travel diary surveys.* Transportation, 2007. **34**(1): p. 107-128.
101. Uysal, M., D.R. Fesenmaier, and J.T. O'Leary, *Geographic and seasonal variation in the concentration of travel in the United States.* Journal of travel research, 1994. **32**(3): p. 61-64.
102. Hwang, H.-L. and J. Rollow, *Data processing procedures and methodology for estimating trip distances for the 1995 American Travel Survey (ATS)*. 2000, Oak Ridge National Laboratory. Center for Transportation Analysis.
103. Bradley, M., M.L. Outwater, and N. Ferdous, *Implementation of a Practical Model System to Predict Long-Distance Travel for the Entire US Population.* Transportation Research Record, 2016. **2563**(1): p. 9-17.
104. Basar, G. and C. Bhat, *A Parameterized Consideration Set Model for Airport Choice: An Application to the San Francisco Bay Area 6. Performing Organization Code.* 2003.
105. Mkono, M., *Eco-anxiety and the flight shaming movement: Implications for tourism.* Journal of Tourism Futures, 2020.
106. Jochimsen, L.M., *'I can do more.': investigating frequent flyers' responses to flight shaming and encouragement interventions for reducing air travel and their effects on perceived self-efficacy.* Master Thesis Series in Environmental Studies and Sustainability Science, 2020.
107. Gössling, S., et al., *Can we fly less? Evaluating the 'necessity'of air travel.* Journal of Air Transport Management, 2019. **81**: p. 101722.
108. Flaherty, G.T. and A. Holmes, *Will flight shaming influence the future of air travel?* Journal of travel medicine, 2020. **27**(2): p. taz088.
109. Koppelman, F.S. and V. Sethi, *Incorporating variance and covariance heterogeneity in the generalized nested logit model: an application to modeling long distance travel choice behavior.* Transportation Research Part B: Methodological, 2005. **39**(9): p. 825-853.
110. Van Nostrand, C., V. Sivaraman, and A.R. Pinjari, *Analysis of long-distance vacation travel demand in the United States: A multiple discrete–continuous choice framework.* Transportation, 2013. **40**(1): p. 151-171.
111. Erhardt, G.D., et al., *Ohio long-distance travel model.* Transportation Research Record, 2007. **2003**(1): p. 130-138.
112. Pinjari, A.R. and C. Bhat, *Nonlinearity of response to level-of-service variables in travel mode choice models.* Transportation research record, 2006. **1977**(1): p. 67-74.
113. LaMondia, J., et al., *Modeling Overnight Out-of-State Long Distance Mode Choice*. 2018.
114. Chester, M.V. and M.S. Ryerson, *Grand challenges for high-speed rail environmental assessment in the United States.* Transportation Research Part A: Policy and Practice, 2014. **61**: p. 15-26.

115. Grabar, H. *The CEO of Amtrak Thinks Americans Are Ready for Trains Again*. Slate, 2021.
116. Ory, D.T. and P.L. Mokhtarian, *When is getting there half the fun? Modeling the liking for travel.* Transportation Research Part A: Policy and Practice, 2005. **39**(2-3): p. 97-123.
117. Pinjari, A.R., et al., *Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions.* Transportation, 2011. **38**(6): p. 933.
118. Gärling, T. and K.W. Axhausen, *Introduction: Habitual travel choice.* Transportation, 2003. **30**(1): p. 1-11.
119. Frank, L., et al., *Urban form, travel time, and cost relationships with tour complexity and mode choice.* Transportation, 2008. **35**(1): p. 37-54.
120. Verplanken, B., et al., *Context change and travel mode choice: Combining the habit discontinuity and self-activation hypotheses.* Journal of Environmental Psychology, 2008. **28**(2): p. 121-127.
121. Aarts, H. and A.P. Dijksterhuis, *The automatic activation of goal-directed behaviour: The case of travel habit.* Journal of environmental psychology, 2000. **20**(1): p. 75-82.
122. Rossi, T.F. and Y. Shiftan, *Tour based travel demand modeling in the US.* IFAC Proceedings Volumes, 1997. **30**(8): p. 381-386.
123. National Academies of Sciences, E., and Medicine, *Travel Demand Forecasting: Parameters and Techniques*. 2012: Washington, DC.
124. Wu, Y., *The Integrated Florida Statewide Model*. 2008, FDOT Systems Planning Office.
125. *Use of the Florida Statewide Model*, F.D.o. Transportation, Editor. 2018.
126. Han, J., A. Bakhshi, and K. Kaltenbach, *KYSTMv17: The Latest Statewide Model*. 2017, The Corradino Group.
127. Stammer Jr, R.E. and G. Pratt, *Statewide Modeling Practices and Prototype Statewide Model Development for Tennessee*. 2002.
128. Outwater, M., et al., *California statewide model for high-speed rail.* Journal of Choice Modelling, 2010. **3**(1): p. 58-83.
129. Beser, M. and S. Algers, *SAMPERS—The new Swedish national travel demand forecasting tool*, in *National Transport Models*. 2002, Springer. p. 101-118.
130. Rohr, C., et al., *Modeling long-distance travel in great britain.* Transportation research record, 2013. **2344**(1): p. 144-151.
131. Steinsland, C. and L. Fridstrom, *Travel demand models on the edge: Exploring the NTM5 model's limits of extrapolation*, N.C.f.T. Research, Editor. 2012, Institute of Transport Economics.
132. de Bok, M., et al. *Estimation of a mode choice model for long distance travel in Portugal*.
133. Statistics, B.o.T., *America on the Go... Findings from the National Household Travel Survey*, F.H. Administration, Editor. 2006, U.S. Department of Transportation: bts.gov.
134. Axhausen, K.W., *Defining the scope of a long-distance travel survey*, in *Capturing long-distance travel*. 2003, Research Studies Press. p. 8-25.
135. Axhausen, K.W. and C. Weis, *Predicting response rate: A natural experiment.* Survey Practice, 2010. **3**(2).
136. Axhausen, K.W., B. Schmid, and C. Weis, *Predicting response rates updated.* Arbeitsberichte Verkehrs-und Raumplanung, 2015. **1063**.
137. Tourangeau, R., *Remembering what happened: Memory errors and survey reports*, in *The science of self-report*. 1999, Psychology Press. p. 41-60.

138. Paulin, G., *Travel expenditures, 2005-2011: spending slows during recent recession.* 2012.
139. Center, P.R., *Mobile Fact Sheet*. 2019, The Pew Charitable Trusts: Pew Research Center.
140. *Streetlight InSight Metrics: Our Methodology and Data Sources*. 2018, Streetlight Data, Inc. San Francisco, CA.
141. Flake, L., et al., *Use of smartphone panels for viable and cost-effective gps data collection for small and medium planning agencies.* Transportation Research Record, 2017. **2643**(1): p. 160-165.
142. Bradley, M., et al. *The Transition from Diary-Based to Smartphone-Based Travel Survey Data: Implications for Travel Demand Modeling*. in *5th Transportation Research Board Innovations in Travel Modeling Conference, Denver, CO*. 2016.
143. Ritter, C., et al., *Long-Distance Smartphone-Based Travel Surveys in Ohio*. 2018, RSG Inc.
144. Junco, R., D. Merson, and D.W. Salter, *The effect of gender, ethnicity, and income on college students' use of communication technologies.* Cyberpsychology, Behavior, and Social Networking, 2010. **13**(6): p. 619-627.
145. Nadeem, S. and M.C. Weigle, *Demographic prediction of mobile user from phone usage.* Age, 2012. **1**: p. 16-21.
146. Eagle, N., Y.-A. de Montjoye, and L.M. Bettencourt. *Community computing: Comparisons between rural and urban societies using mobile phone data*. in *2009 international conference on computational science and engineering*. 2009. IEEE.
147. Soto, V., et al. *Prediction of socioeconomic levels using cell phone records*. in *International Conference on User Modeling, Adaptation, and Personalization*. 2011. Springer.
148. Becker, R., et al., *Human mobility characterization from cellular network data.* Communications of the ACM, 2013. **56**(1): p. 74-82.
149. Blumenstock, J., G. Cadamuro, and R. On, *Predicting poverty and wealth from mobile phone metadata.* Science, 2015. **350**(6264): p. 1073-1076.
150. Calabrese, F., L. Ferrari, and V.D. Blondel, *Urban sensing using mobile phone network data: a survey of research.* Acm computing surveys (csur), 2014. **47**(2): p. 1-20.
151. Janzen, M., et al. *Estimating long-distance travel demand with mobile phone billing data*. in *16th Swiss Transport Research Conference (STRC 2016)*. 2016. Swiss Transport Research Conference (STRC).
152. Cho, E., S.A. Myers, and J. Leskovec. *Friendship and mobility: user movement in location-based social networks*. in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011.
153. Wolf, J., M. Oliveira, and M. Thompson, *The impact of trip underreporting on VMT and travel time estimates: preliminary findings from the California statewide household travel survey GPS study.* Transportation Research Record, 2003. **1854**: p. 189-198.
154. Smith, R.S. and J.E. Wood, *Memory--its reliability in the recall of long distance business travel*. 1977.
155. Li, H., et al. *Descriptive Analysis of Long-Distance Travel by Personal Vehicles Using 2004 Commute Atlanta Data*. in *Transportation Research Board 86th Annual Meeting*. Washington, DC.
156. *MI Travel Counts: Final Report*. 2005, Michigan Department of Transportation. p. 154.

157. *MI Travel Counts II: Final Methodological Report*. 2010, Michigan Department of Transportation. p. 29.
158. Wilaby, M. and J. Casas, *MI Travel Counts III: Household Travel Survey - Final Methodology Report*. 2016, Westat. p. 60.
159. *2010-2012 California Household Travel Survey Final Report: Version 1.0*. 2013, California Department of Transportation. p. 147.
160. Statistics, B.o.T., *2001 National Household Travel Survey User's Guide*, F.H. Administration, Editor. 2004.
161. *Poverty Thresholds by Size of Family and Number of Children: 1995*, U.C. Bureau, Editor. 1995: Web.
162. *Poverty Thresholds by Size of Family and Number of Children: 2001*, U.C. Bureau, Editor. 2001: Web.
163. *Poverty Thresholds by Size of Family and Number of Children: 2013*, U.C. Bureau, Editor. 2013: Web.
164. Hazra, A. and N. Gogtay, *Biostatistics series module 3: comparing groups: numerical variables.* Indian journal of dermatology, 2016. **61**(3): p. 251.
165. Dowds, J., L. Aultman-Hall, and J.J. LaMondia, *Comparing alternative methods of collecting self-assessed overnight long-distance travel frequencies.* Travel Behaviour and Society, 2020. **19**: p. 124-136.
166. Richardson, A.J. and R.K. Seethaler, *Estimating Long-Distance Travel Behavior from the Most Recent Trip.* TRB Transportation Research Circular E-C026—Personal Travel: The Long and Short of It, 2001.
167. *Theil Index*. Income Equality Metrics 2021; Available from: https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/theil-index.html.
168. Festin, S.M., *Summary of national and regional travel trends: 1970-1995*. 1996, United States. Federal Highway Administration.
169. Bernardin Jr, V.L., et al., *Integration of National Long-Distance Passenger Travel Demand Model with Tennessee Statewide Model and Calibration to Big Data.* Transportation Research Record, 2017. **2653**(1): p. 75-81.
170. Shiftan, Y., *The use of activity-based modeling to analyze the effect of land-use policies on travel behavior.* The Annals of Regional Science, 2008. **42**(1): p. 79-97.
171. Buliung, R.N. and P.S. Kanaroglou, *Urban form and household activity-travel behavior.* Growth and Change, 2006. **37**(2): p. 172-199.
172. Manaugh, K. and A.M. El-Geneidy, *What makes travel'local' Defining and understanding local travel behavior.* Journal of Transport and Land Use, 2012. **5**(3): p. 15-27.
173. Krizek, K.J., *Operationalizing neighborhood accessibility for land use-travel behavior research and regional modeling.* Journal of Planning Education and Research, 2003. **22**(3): p. 270-287.
174. Nasri, A. and L. Zhang, *Assessing the impact of metropolitan-level, county-level, and local-level built environment on travel behavior: Evidence from 19 US urban areas.* Journal of Urban Planning and Development, 2015. **141**(3): p. 04014031.
175. Kotval-K, Z. and I. Vojnovic, *The socio-economics of travel behavior and environmental burdens: A Detroit, Michigan regional context.* Transportation Research Part D: Transport and Environment, 2015. **41**: p. 477-491.

176. Manson, S., et al., *IPUMS National Historical Geographic Information System*. 2020, IPUMS: Minneapolis, MN.
177. *Airports*, A.S.f.R.a.T.B.o.T.S.N.T.A. Database, Editor. 2021, US Department of Transportation: Bureau of Transportation Statistics: National Transportation Atlas Database.
178. *Amtrak Stations*, F.R. Administration, Editor. 2020, U.S. Department of Homeland Security: Homeland Infrastructure Foundation-Level Data (HIFLD).
179. *National Plan of Integrated Airport Systems (NPIAS): 2001 - 2005*, U.S.D.o.T.F.A. Administration, Editor. 2002, U.S. Department of Transportation: faa.gov. p. 251.
180. Cuong, B.C., P.L. Lanzi, and N.T. Thong, *A novel intuitionistic fuzzy clustering method for geo-demographic analysis.* Expert Systems with applications, 2012. **39**(10): p. 9848-9859.
181. Grekousis, G. and H. Thomas, *Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods.* Applied Geography, 2012. **34**: p. 125-136.
182. Son, L.H., *Enhancing clustering quality of geo-demographic analysis using context fuzzy clustering type-2 and particle swarm optimization.* Applied Soft Computing, 2014. **22**: p. 566-584.
183. Cuong, B.C. and H.V. Long, *Spatial interaction–modification model and applications to geo-demographic analysis.* Knowledge-Based Systems, 2013. **49**: p. 152-170.
184. Dias, M.L.D., *fuzzy-c-means: An implementation of Fuzzy C-means clustering algorithm*. 2019, Zenodo: pypi.org.
185. Khattak, A., et al., *Travel by university students in Virginia: Is this travel different from travel by the general population?* Transportation Research Record, 2011. **2255**(1): p. 137-145.
186. *NextGen NHTS Newsletter: Summer 2020*, U.D.o. Transportation, Editor. 2020, Federal Highway Administration. p. 5.

Appendix A: Standard Deviations of Day Clusters

			HBTE	Purpose Percentage Split		Mode Percentage Split	
				Work	*Leisure*	*Vehicle*	*Air*
Low Income	*No Kids*	*E1*	2.498	0.167	0.199	0.249	0.269
		E2	2.527	0.129	0.191	0.217	0.229
		H1	2.707	0.246	0.285	0.271	0.277
		H2	1.655	0.186	0.238	0.256	0.258
		M1	1.479	0.275	0.324	0.344	0.334
		M2	1.541	0.282	0.393	0.370	0.314
		M3	1.044	0.400	0.473	0.437	0.395
		L1	1.310	0.387	0.431	0.428	0.401
		Annual-Level	2.306	0.305	0.393	0.376	0.335
	Kids	*E1*	0.799	0.000	0.500	0.516	0.000
		E2	0.897	0.197	0.459	0.501	0.063
		H1	0.999	0.167	0.480	0.487	0.135
		H2	0.632	0.158	0.496	0.500	0.158
		M1	0.553	0.179	0.396	0.418	0.170
		M2	0.712	0.073	0.428	0.452	0.000
		M3	0.318	0.149	0.149	0.288	0.000
		L1	0.581	0.195	0.332	0.388	0.055
		Annual-Level	0.710	0.163	0.421	0.453	0.105
Medium Income	*No Kids*	*E1*	1.922	0.208	0.293	0.179	0.189
		E2	1.771	0.193	0.195	0.135	0.123
		H1	2.476	0.234	0.264	0.228	0.194
		H2	2.362	0.228	0.237	0.188	0.189
		M1	2.008	0.277	0.304	0.305	0.270
		M2	1.440	0.243	0.285	0.263	0.244
		M3	1.340	0.416	0.431	0.406	0.381
		L1	1.539	0.331	0.369	0.398	0.380
		Annual-Level	2.448	0.293	0.324	0.322	0.294
	Kids	*E1*	1.373	0.188	0.386	0.366	0.208
		E2	1.362	0.270	0.435	0.437	0.286
		H1	1.503	0.303	0.399	0.310	0.248
		H2	1.374	0.354	0.412	0.455	0.229
		M1	1.074	0.417	0.435	0.452	0.297
		M2	1.039	0.235	0.475	0.465	0.235
		M3	1.036	0.388	0.341	0.446	0.306
		L1	0.852	0.417	0.350	0.420	0.364
		Annual-Level	1.247	0.362	0.434	0.462	0.292
High Income	*No Kids*	*E1*	2.764	0.145	0.152	0.086	0.077
		E2	2.808	0.121	0.126	0.119	0.113
		H1	2.299	0.188	0.176	0.177	0.170
		H2	2.190	0.150	0.162	0.115	0.114
		M1	1.848	0.233	0.244	0.196	0.189
		M2	2.058	0.212	0.214	0.179	0.184
		M3	1.753	0.188	0.188	0.168	0.157
		L1	1.994	0.227	0.209	0.210	0.191
		Annual-Level	5.421	0.239	0.236	0.187	0.178
	Kids	*E1*	3.283	0.116	0.125	0.132	0.132
		E2	2.529	0.133	0.147	0.137	0.138
		H1	2.258	0.198	0.194	0.151	0.153
		H2	1.884	0.169	0.179	0.178	0.185
		M1	1.748	0.235	0.238	0.205	0.204
		M2	1.412	0.361	0.372	0.332	0.332
		M3	1.691	0.288	0.280	0.294	0.289
		L1	1.794	0.320	0.320	0.287	0.286
		Annual-Level	4.240	0.302	0.300	0.261	0.261