ABSTRACT

Title of Dissertation:	SPATIOTEMPORAL ANALYSIS OF VEHICLE MOBILITY PATTERNS USING MACHINE LEARNING APPROACHES
	Guimin Zhu, Doctor of Philosophy, 2023
Dissertation directed by:	Professor Kathleen Stewart Department of Geographical Sciences

Vehicle mobility is important to a diverse range of disciplines, e.g., geography, transportation, and public health. Machine Learning algorithms have been applied in geospatial analysis related to vehicle mobility and travel pattern research, which provided researchers with more flexibility and capabilities for complex mobility pattern analyses. This dissertation aims to explore how different Machine Learning models (e.g., regression and clustering) can be applied to enhance the interpretability of vehicle mobility patterns by conducting explanatory analyses on factors that may impact different mobility patterns (i.e., trip volume changes and travel times) over space and time (e.g., different stages of the COVID-19 Pandemic at regional and nationwide scales). In this dissertation, three studies were undertaken to investigate the spatiotemporal trends of vehicle trip changes and travel behaviors, using passively-collected mobile device data. The first study examined mobility patterns over different time periods during the summer 2020 when COVID-19 cases were spiking in Florida

(locations with large numbers of vulnerable individuals) and analyzed a set of underlying drivers for mobility and how these factors changed over time using Machine Learning approaches. The second study investigated changing mobility patterns across the U.S. during 2021 when COVID-19 vaccinations were becoming available to understand whether changing vaccination rates led to a change in the rate of trips using Machine Learning clustering methods. The third study investigated reasons impacting travel times for two origin-destination pairs using a Machine Learning approach to better understand how different factors can affect travel times over different trip purposes and different trip lengths in Maryland. The contributions of this dissertation are that it provided new insights into how different types of mobility patterns evolved over space and time, especially during a major public health crisis, and the results are useful for policy and planning implications for local and regional officials, e.g., mobility restriction measurements, decision support for economic recovery, and public health strategies. The integration of diverse data sources (e.g., passively-collected mobility data and other mobility data from different public and private sources) and the utilization of multiple Machine Learning models enhanced the interpretability of vehicle mobility patterns.

SPATIOTEMPORAL ANALYSIS OF VEHICLE MOBILITY PATTERNS USING MACHINE LEARNING APPROACHES

By

Guimin Zhu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Kathleen Stewart, Chair
Professor Taylor Oshan
Professor Yiqun Xie
Professor Deb Niemeier
Professor Vanessa Frias-Martinez

Dedication

To my parents,

Huaqian Zhu (朱华钱) and Yonglan Cai (蔡永兰),

and my beloved families

for their endless love, support, and treasured memories throughout the journey.

Acknowledgements

I would like to express my gratitude to my advisor, Professor Kathleen Stewart, for her advice, guidance, and help throughout my PhD journey. Professor Stewart provided feedback on my research, guidance on my Teaching Assistant work, and support in both my study and life in the past few years. My PhD journey could not be accomplished without Professor Stewart's consistent support. Working with Professor Stewart and serving as Professor Stewart's TA were the best things that ever happened to my PhD career.

I would like to also express my sincere appreciation to my dissertation committee members, Professor Taylor Oshan, Professor Yiqun Xie, Professor Deb Niemeier, and Professor Vanessa Frias-Martinez for their generous suggestions and insights on my research and dissertation. I am deeply thankful to Professor Deb Niemeier for being my dean's representative and co-authoring my first paper.

My warmest thanks are also reserved for my colleagues, Dr. Junchuan Fan, Dr. Yanjia Cao, Dr. Hai Lan, Dr. Yao Li, Dr. Zhiyue Xia, Dr. Jeff Sauer, Dr. Zheng Liu, and Peiqi Zhang, for helping solve research-related technical issues. Thank you to my friends, Dr. Xin Xu, Dr. Chu-Chun Chang, Yunting Song, Yuehui Qian, Weiye Chen, Yingrui Zhao, Haley Mullen, Xin Dong, Yuhao Wang, Jiaming Lu, and Dr. Aolin Jia in the Department of Geographical Sciences.

A special thank goes to my board game friends, Xingyue Huang, Sheng-Min Huang, Yue Huang, Xiarong Xu, Yu Nie, Qiyu Zhang, Gege Wang, Yudong Li, Chujun Lu, and Zhang Tian, as they supported me through this lonely PhD journey, especially

the COVID-19 Pandemic days. Sincere thanks to Haotian Wang for being an enjoyable and supporting friend. A special thanks to Guanhong Wang.

This dissertation is partly supported by the National Science Foundation under Grant No. BCS-2027412.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACS: American Community Survey

ANOVA: Analysis of Variance

BLS: Bureau of Labor Statistics

CATT Lab: Center for Advanced Transportation Technology Laboratory

CCP: Cost-Complexity Pruning

CDC: Centers for Disease Control and Prevention

COVID-19: SARS-CoV-2 coronavirus disease

DTW: Dynamic Time Warping

FHWA: Federal Highway Administration

GHCN: Global Historical Climatology Network

GPS: Global Positioning System

LAUS: Local Area Unemployment Statistics

MAE: Mean Absolute Error

ML: Machine Learning

MTI: Maryland Transportation Institute

NAICS: North American Industry Classification System

NCHS: National Center for Health Statistics

NOAA: National Oceanic and Atmospheric Administration

O-D: Origin-Destination

OSM: OpenStreetMap

POI: Point of Interest

r: Pearson's Correlation

R^2: Coefficient of Determination

R-OC: Rockville – Ocean City

RFE: Recursive Feature Elimination

RMS: Root Mean Square Error

T-SS: Towson – Silver Spring

Tukey HSD: Tukey Honestly Significant Difference

WHO: World Health Organization

Chapter 1: Introduction

1.1 Background and motivation

Human mobility is important to a wide range of disciplines, including geography, transportation, sociology, and public health, among others. From the perspective of geographers, vehicle mobility studies play an important role in understanding how people's day-to-day activities are structured in space and carried out across numerous sub-areas, including health geography, transportation geography, geospatial intelligence, etc. (Castles, 2018; Wolfe et al., 2020). Studies on mobility have been pushed forward to a new era due to huge amounts of location-based data generated for different purposes. Identifying and understanding mobility patterns has been an important area of research in the field of Geographic Information Science for urban planning, traffic forecasting, and mitigating the spread of infectious diseases among other application areas (Belik et al., 2011; M. C. González et al., 2008; Siła-Nowicka et al., 2016; Xia et al., 2018). Topics on mobility include, for example, space-time relationships, e.g., recovery of human mobility after the global pandemic and other natural hazards (Elliott, 2015; Griffiths et al., 2021; Huang et al., 2020; Q. Wang & Taylor, 2014); durations of mobility in different contexts, e.g., travel time estimation and reliability (Jenelius & Koutsopoulos, 2013; Sanaullah et al., 2016; Tang et al., 2016; Xu et al., 2019); and trajectory analysis, e.g., trip purpose computation and trajectory characteristics (Kwan, 2000, 2004; J. G. Lee et al., 2008; Zamir et al., 2014); among other topics.

Geospatial modeling techniques, from traditional modeling methods to Machine Learning (ML) methods to recent deep learning methods (e.g., GeoAI), have been applied to vehicle mobility and travel pattern research (Fan et al., 2019; Mollalo, Vahedi, et al., 2020; Xu et al., 2019), for example, to analyze the spatiotemporal trends of mobility changes and understand how mobility patterns evolved over space and time. Before the rise of ML methods for different geospatial modeling tasks, analysis relied on prerequisite domain knowledge of the study subjects. ML algorithms provided researchers with insights on different impact factors including revealing factors that were otherwise undetected but have an impact on the social phenomenon being studied. For example, researchers investigated the risk of getting infected by SARS-CoV-2 coronavirus disease (COVID-19) by three transportation modes, public transit, walking, and driving, and surprising, walking was not as safe as the general public perceived even with an increased rate of facemask wearing (R. Zhu et al., 2021). In addition, the enormous volumes of data, generated from Global Positioning System (GPS), social media check-ins, and from apps, offer more and more possibilities for analyzing underlying patterns of mobility, but leaves traditional modeling methods more limited for analysis tasks. ML methods, applied to different regions and time periods, provide flexibility and stronger capabilities for complex pattern recognition and predictive analytics needed for mobility analyses especially those involving passively-collected mobile device data (Luca et al., 2021; Song et al., 2016).

The term *human mobility* contains a diverse range of transportation modes, including vehicle, pedestrian, cyclist, public transit, air, etc. (Barbosa et al., 2018). In this dissertation, I focused on vehicle movement, particularly trip volume changes and

travel time analyses. This dissertation investigates how ML models can be applied to interpret vehicle mobility patterns by examining factors that may impact mobility at different spatial and temporal scales (e.g., regional and nationwide study at census tract and county level) and under different contexts (e.g., different stages of the COVID-19 Pandemic) (Figure 1.1). More specifically, in this dissertation, three studies related to vehicle mobility were undertaken using multiple ML approaches, e.g., regression and clustering, as well as statistical tests, e.g., Analysis of Variance (ANOVA) and Tukey's Honestly Significant Difference test (Tukey HSD).

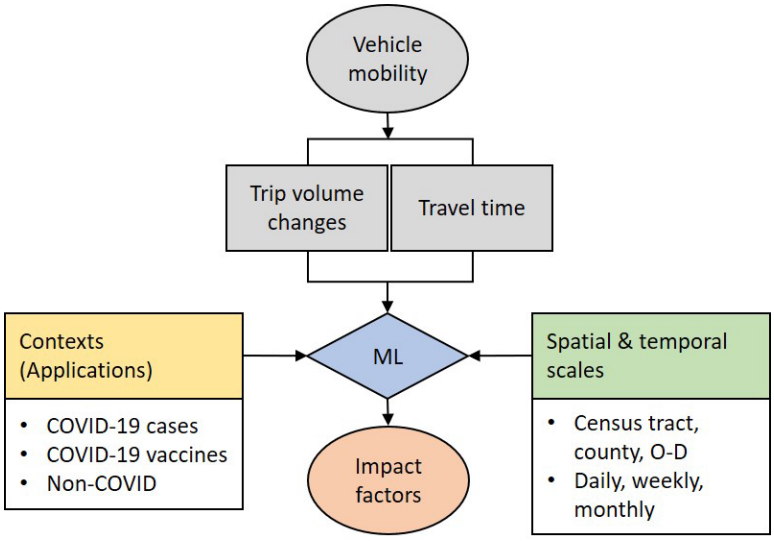

Figure 1.1 Conceptual flowchart for dissertation topics.

The first research study (Chapter 2) in this dissertation considered patterns of driving during the early pandemic when the COVID-19 case numbers were high and rising even more in certain parts of the U.S., and analyzed a set of factors to understand how these different factors impacted driver's trips in a highly populated, region of the southeastern U.S. The first confirmed COVID-19 case in the U.S. was reported on January 20, 2020, and shortly after, World Health Organization (WHO) declared

COVID-19 a global pandemic on March 11, 2020, with more than 118,000 cases in 114 countries and over 4,000 deaths (Centers for Disease Control and Prevention, 2023). This study was undertaken when the pandemic was ongoing and there was not much known at the time about the relationship between the mobility and the changing levels of COVID-19 illness in different parts of the U.S. California was the first state to issue a statewide mandatory stay-at-home order that restricted mobility to reduce the transmission of the highly infectious COVID-19 on March 19, 2020. By May 31, 2020, 42 states and territories issued mobility-restriction orders that required all residents to remain at home except for essential activities (Centers for Disease Control and Prevention, 2023). In the early pandemic, questions about how mobility patterns changed after the stay-at-home orders were issued, how the movement of people's daily lives and travel was impacted by the pandemic, and the role of mobility in sustaining the level of infection and transmission were key topics of study (Gao, Rao, Kang, Liang, Kruse, et al., 2020; Kraemer et al., 2020; Nouvellet et al., 2021). For my research in Study 1, three counties in Florida (Miami-Dade County, Broward County, and Palm Beach County) were selected as a study area to analyze the spatiotemporal trends of dynamic mobility patterns in this tri-county region. These three counties were hard hit by the COVID-19 pandemic, contributing to about 57% of the total Florida positive COVID-19 cases and 55% of the total deaths in summer 2020 (Florida Department of Health, 2021a). With a large vulnerable population including Black and Hispanic populations as well as a significant population over the age of 65 in these three counties, we investigated the changing relationships between mobility patterns and increasing COVID-19 cases in the three counties in this early stage of the COVID-19 pandemic

(U.S. Census Bureau, 2022). The contribution of this research is that the dynamic relationships between mobility patterns and COVID-19 infections were examined at a fine spatial granularity in a hotspot region in the U.S. when there were not many understandings on the mobility patterns during the early-Pandemic period. The Random Forest results provided county-level insights that are useful for both public health officials and those stakeholders at the intersection of transportation planning and health management for decision support on managing expectations regarding expected travel by local populations during a major health crisis.

Chapter 2 used a Random Forest model to examine explanatory factors including sociodemographic, travel-related, built environment, and health factors, and revealed how these factors contributed to mobility patterns among the three study counties and over different time periods at a time when the relationship between mobility and a high level of COVID-19 was not well understood. Generally, Random Forest models are a good choice for regression and classification tasks, as Random Forest requires little processing of the data, can handle both numerical and categorical values, and are typically robust to outliers and unbalanced data compared to other ML algorithms, e.g., artificial neural network and support vector machine (Nguyen et al., 2021; Rodriguez-Galiano et al., 2015). The contributions of explanatory factors for trip patterns across the three counties were separately assessed using Random Forest models. The results revealed differences among the three study counties and over two time periods.

The second study (Chapter 3) in this dissertation investigated the pattern of drivers' trips across the U.S. during a time when COVID-19 vaccines were being

introduced and undertook a spatiotemporal ML analysis to detect trends in both mobility and vaccination rates, and analyze the degree to which vaccinations rates may have impacted the level of mobility in different parts of the U.S. While the first study focused on the mobility patterns during the early COVID-19 pandemic, the second study was conducted on data from the mid-Pandemic in 2021, after COVID-19 vaccines were made available in late December 2020 (U.S. Food and Drug Administration, 2020). The administration of COVID-19 vaccines helped mitigate the spread of COVID-19, as while there were about 225,000 daily new COVID-19 cases on January 1, 2021, this level of infection decreased to a much lower level of about 9,000 cases by the end of May 2021, approximately a 96% decrease, after vaccines were introduced. As the general public got vaccinated and the U.S. Government relaxed mobility-related policies (e.g., activity restrictions and social distancing measures), people were expected to feel safer once vaccinated and be more willing to leave their homes and travel (Fiori & Lacoviello, 2021). This research addressed a gap in knowledge regarding an understanding of how ongoing COVID-19 vaccination rates were associated with mobility across the U.S., i.e., the spatial and temporal distributions of counties with different mobility-vaccination profiles. The clustering analysis results identified the county differences, e.g., counties in large population centers and metropolitan areas and counties in the Mountain and Southern states.

For Chapter 3, the dynamic relationships between mobility and vaccination rates were examined using a K-means time-series clustering approach. The time-series clustering method is used to investigate compound dynamic relationships over time for multi-variates, instead of only one variable at a time (Giordano et al., 2021; Siebert et

al., 2021). The different spatial and temporal distributions of detected clusters represented counties with different mobility-vaccination profiles. A set of demographic and socioeconomic factors (e.g., race and ethnicity, education, and work-related factors) was examined to investigate how these factors may serve as drivers for mobility behaviors and how they were related to the different mobility-vaccination profiles in the U.S., using the ANOVA test and the Tukey HSD test. Finally, to understand how different trip purposes (e.g., trips to retail/recreation locations and workplace locations) may also be associated with the different mobility-vaccination profiles was examined for a case study involving urban (dense population) and rural (less dense population) counties in Maryland and Alabama respectively. The findings of the research for Chapter 3 demonstrated the spatial and temporal distributions of counties with different mobility-vaccination profiles across the U.S., and factors related to education, economic, and race/ethnicity significantly contributed to these differences, among others. The results could guide policymakers, businesses, and individuals, in transportation demand analysis, economic recovery, and public health policies as people adapted to a post-pandemic world.

The third study (Chapter 4) in this dissertation sought to understand how driving times can vary for the same origin-destination (O-D) and understand the impact of different factors on driving time using a ML model. The model was tested for two different trip lengths and different trip purposes – urban commuting and rural recreational – to understand how the importance of these factors changes for the different contexts of travel. Currently research mostly focuses on how advanced models could provide more accurate travel time prediction for a given O-D, while the gap that

this study fills is to answer the questions why drivers are taking different amount of time between the same O-D, what factors are the key factors contributing to the travel times, and whether the key factors change over different O-Ds.

While Chapter 2 and Chapter 3 investigated mobility patterns in the context of the COVID-19 pandemic (confirmed cases and vaccines), the third study switched the focus to travel behaviors during a non-COVID time using trip data for 2018-2019. The mobility patterns examined in the first and second studies were represented by trip volumes (e.g., numbers of trips per person and relative numbers of trips), while in the third study, the data was also trips, but the focus for analysis was on travel time for these trips, i.e., the driving time to a destination. By examining and understanding how different factors impact travel time in the two contexts (urban commuting and rural recreation), transportation planners and policymakers can develop more effective strategies to enhance driver awareness, reduce traffic congestion and improve overall transportation efficiency (Carrion & Levinson, 2012; Z. Wang, Fu, et al., 2018). Existing research, including commercial navigation applications, mainly focus on providing more accurate travel time prediction using advanced ML algorithms, e.g., Artificial Neural Network and Graph Convolutional Neural Network (Jin et al., 2021; Xu et al., 2019) using the large volume of historical travel speed data and real-time traffic information (Epstein, 2013; Ireland, 2011). Current research falls short in interpreting how different factors contribute to different travel times even for the same trip and how factors vary for different trip purposes and trip lengths.

The set of factors that were analyzed to understand their impacts on travel time in Chapter 4 were categorized into three groups, namely driver behavior, built

environment and road network, and external factors (e.g., weather, traffic incidents, and holidays). Other studies focused on the prediction of travel time, have analyzed departure date and time, weekday/weekend, public holidays, speed limit, functional class, intersections, season, temperature, and precipitation (Jenelius & Koutsopoulos, 2013; Tang et al., 2016; Y. Wang et al., 2014). In my study, driver route choice and driver travel speed behaviors, two new factors not considered in previous travel time-related research, were clustered with the other factors analyzed using a time-series clustering algorithm. All factors were trained using a Random Forest model on two selected O-D pairs representing different trip purposes and trip lengths. The results were constructive and could provide insights for travel time prediction researchers into how they could better weight these explanatory variables.

1.2 Dissertation structure and research objectives

Overall, this dissertation explores how different ML models can be applied to interpret vehicle mobility patterns (i.e., trip volume changes and travel time) by examining factors that may impact mobility over space and time in different contexts.

This dissertation is composed of five chapters. Chapter 1 presents an introduction to the dissertation and describes the structure. Chapter 2 discusses an investigation of mobility patterns over different time periods during the early Pandemic when COVID-19 cases were spiking in Florida and analyzed the set of drivers for mobility in locations with large numbers of vulnerable individuals and how the factors changed over time using ML approaches. Chapters 3 examines changing mobility patterns across the U.S. during the mid-Pandemic when COVID-19 vaccinations were becoming available to understand whether changing vaccination rates led to a change

in the rate of trips using ML clustering methods. Chapter 4 investigates reasons impacting travel times for two O-D pairs using a ML approach to better understand how different factors can affect travel times over different trip purposes and different trip lengths in Maryland. Chapter 5 concludes the dissertation by summarizing the major findings of this dissertation and discussing future research directions relating to applications of advanced ML and even Deep Learning algorithms on mobility-related topics.

The research objectives for Chapter 2 are:

(1) Derive mobility patterns at census tract level during a peak period of the early Pandemic (May-July 2020) for three Florida counties (Miami-Dade, Broward, and Palm Beach counties) from passively-collected mobile device data to understand how trips were distributed and how they changed over the 3-month period.

(2) Select and examine a set of more than 30 factors, including sociodemographic, travel-related, built environment and COVID-19 factors, and detect the spatial and temporal distributions of these factors in order to understand the difference characteristics among the three counties.

(3) Build and evaluate the performance of Random Forest models to reveal the changing importance ranking for factors in order to determine what factors were key factors contributing to the mobility patterns among the three counties and over different time periods.

The research objectives of Chapter 3 are:

(1) Determine the patterns of county-level mobility across the U.S. in the context of increasing COVID-19 vaccination rates by week during January-May 2021 by developing a mobility index and explore how a spatiotemporal clustering approach can examine the compound associations between mobility and vaccination rates in different locations of the U.S.

(2) Examine demographic and socioeconomic factors as well as numbers of COVID-19 cases and deaths to investigate how these drivers were related to the different mobility-vaccination clusters.

(3) Examine county-level mobility by analyzing trip purposes for selected urban and rural counties to better understand how trip purposes such as retail/recreation trips and work trips among others may be associated with locations having different clusters.

The research objectives of Chapter 4 are:

(1) Identify and collect factors including driver behaviors, built environment and road network characteristics, and external factors (e.g., traffic incidents, weather, and holidays) that may impact driving time in an urban area.

(2) Compute driver route choice and travel speed behaviors, two previously understudied factors, from GPS waypoints using time-series clustering algorithms.

(3) Examine different trip lengths and trip purposes for trips representing

urban commuting driving and trips to a recreational or vacation destination to understand how different factors impact driving time and how the importance of factors changes with trip length/purpose using a ML model.

Chapter 2: Understanding the drivers of mobility during the COVID-19 pandemic in Florida, USA using a Machine Learning approach

2.1 Abstract

As of March 2021, the State of Florida, USA had accounted for approximately 6.67% of total COVID-19 cases in the US. The main objective of this study is to analyze mobility patterns during a three month period in summer 2020, when COVID-19 case numbers were very high for three Florida counties, Miami-Dade, Broward, and Palm Beach counties. To investigate patterns, as well as drivers, related to changes in mobility across the tri-county region, a random forest regression model was built using sociodemographic, travel, and built environment factors, as well as COVID-19 positive case data. Mobility patterns declined in each county when new COVID-19 infections began to rise, beginning in mid-June 2020. While the mean number of bar and restaurant visits was lower overall due to closures, analysis showed that these visits remained a top factor that impacted mobility for all three counties, even with a rise in cases. Our modeling results suggest that there were mobility pattern differences between counties with respect to factors relating, for example, to race and ethnicity (different population groups factored differently in each county), as well as social distancing or travel-related factors (e.g., staying at home behaviors) over the two time periods prior to and after the spike of COVID-19 cases.

2.2 Introduction

Since January 2020, when the first confirmed case of the COVID-19 was reported in the United States, the pandemic has ravaged the United States, with the number of confirmed cases and deaths at over 30.2 million and 551,000 respectively, as of March 2021 (Centers for Disease Control and Prevention, 2023). Questions about how to best slow or stop the spread of this highly infectious disease, including what are the key factors that have enabled the spread of the virus and what can be done to impede its deadly progress, remain under study. The movement of people as they go about their daily lives or travel over larger spatial extents (e.g., travel by air) has been a key focus of study, throwing a spotlight on the role of mobility in sustaining the level of infection and transmission (Kraemer et al., 2020; Nouvellet et al., 2021). Tracking the movement of individuals as they undertake daily activities using the expanding location-based services via applications that apply passive tracking technologies (Gao, Rao, Kang, Liang, Kruse, et al., 2020; Kishore et al., 2020; Xiong et al., 2020), allows us to dig deeper into the role of mobility in infectious disease modeling.

In this study, we investigate mobility patterns, i.e., mean inflow trip patterns, during a peak period of the pandemic, May, June, and July 2020 for three Florida counties, Miami-Dade, Broward, and Palm Beach counties. We use a random forest regression model to determine how a set of more than thirty different factors, including sociodemographic (e.g., median household income, age, race, and ethnicity), travel (e.g., mean travel time to work, percent of the population working from home), and built environment factors (e.g., road network density, street intersection density) as well as the changing number of COVID-19 positive cases, relates to changing levels of

mobility across the tri-county region. Our study is at the detailed granularity of census tracts, highlighting how human behaviors relating to mobility across tracts and between counties varied over time and space and providing insights for planning as well as possible consequences for pandemic outcomes.

Florida's unique attractions (highly regarded oceanside beaches, hotels and resorts, and year-round warm weather) make the state a draw for tourists and travelers year-round, giving Florida a unique status of possibly being a driver for virus transmission beyond its borders (Mangrum & Niekamp, 2020). Local population groups of diverse race and ethnicity succumbed to high levels of infection, which combined with the high number of elderly residents, contributed to over 2 million confirmed cases and 33,000 deaths as of March 2021 (Florida Department of Health, 2021b).

ML algorithms (Breiman, 2001) and random forest models in particular (Liaw & Wiener, 2002) are widely used in geospatial modeling by providing determinant-specific spatial contexts. These models have been especially useful for identifying explanatory variables and assessing the importance of these variables with respect to dependent variables such as transport mode choice decision prediction, transportation mode recognition, travel demand system prediction, and explanation of drivers for forest change (Ghasri et al., 2017; Jahangiri & Rakha, 2015; Rasouli & Timmermans, 2012; Santos et al., 2019). A random forest regression model is a meta estimator that fits a number of decision trees to various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (Chen et al., 1999; Hao & Ho, 2019). Generally, random forest models are a good choice for regression and

classification tasks based on their advantages, e.g., little pre-processing (rescaling or transforming) of the data is required, the modeling can be parallelizable, are compatible with high dimensional data, and are typically robust to outliers and unbalanced data (Rodriguez-Galiano et al., 2015). Comparisons of random forest models with other ML algorithms (e.g., linear regression, decision tree, artificial neural network, and support vector machine) for geospatial modeling find that the random forest model performance, in terms of both computation time and prediction accuracy, is generally positive (Hagenauer et al., 2019; Nguyen et al., 2021).

We used a random forest model for examining explanatory factors (i.e., sociodemographic, travel-related, built environment, and health factors) and their relative importance for revealing drivers underlying patterns of mobility based on inflow trips in the context of rising COVID-19 cases in three key counties in Florida.

2.3 Related work

Studies published since the pandemic began to show the effect that COVID-19 has had on employment, education, and the economy. Franch-Pardo et al. conducted a systematic review of scientific articles on geospatial and spatial-statistical analysis of COVID-19 using perspectives drawn from spatiotemporal analysis, health and social geography, environmental variables, data mining, and web-based mapping (Franch-Pardo et al., 2020). New mobility platforms using mobile device data from SafeGraph, Google Mobility Reports, and Descartes Labs (Gao, Rao, Kang, Liang, & Kruse, 2020; Gao, Rao, Kang, Liang, Kruse, et al., 2020; Kang et al., 2020; Warren & Skillman, 2020) have shown the dynamic nature of mobility data at different granularities, e.g., county, metropolitan area, and state. The University of Maryland's COVID-19 Impact

Analysis Platform reports daily updated mobility-related data products (e.g., social distancing index and trip distances) (L. Zhang et al., 2020). Facebook, in partnership with academic institutions, created a global COVID-19 symptom survey that invites users to report on COVID-19 related symptoms, social distancing behaviors, and vaccine acceptance on a daily basis (Kreuter et al., 2020).

Mobility restrictions have been posited to be effective for constraining disease transmission within and between communities (Espinoza et al., 2020), and mobility data that has been collected from mobile devices and location-based applications can be measured against a baseline from pre-pandemic times to provide insights for policymakers and epidemiologists interested in monitoring social distancing and the spread of COVID-19 (Chang et al., 2020; Kishore et al., 2020). Investigations of mobility trends indicate that stay-at-home orders were largely effective (M. Lee et al., 2020).

Numerous researchers have examined the relationship between human mobility and COVID-19 infection rates. For example, analysis using mobile device location data from across the U.S. and a Simultaneous Equations Model found a positive relationship between inflow trips for each U.S. county and COVID-19 infections, which may be useful for gauging the relationship between mobility and COVID-19 transmission risks (Xiong et al., 2020). Gao et al. examined the association between the rate of human mobility changes of mobile phone users (i.e., change rates of median travel distance and median home dwelling time), and the rate of confirmed COVID-19 cases in 50 U.S. states and the District of Columbia, finding that social distancing mandates were associated with the slowing of COVID-19 spread, especially when stay-at-home orders

were to be lifted and states were planning for reopening their economies (Gao, Rao, Kang, Liang, Kruse, et al., 2020). Other dimensions were also studied, including socioeconomic factors, such as population, household income (Huang et al., 2021), age, race, and ethnicity. A multinational study investigated the relationship between the severity of COVID-19, mobility changes, and lockdown measures, found that lockdown measures were significant with respect to encouraging people to maintain social distancing, while the severity of socioeconomic and institutional factors (e.g., median age, percentage of the population employed in services, and percentage of health expenditure) may have limited effects to sustain social distancing (Rahman et al., 2020). It has also been demonstrated that COVID-19 case positivity during spring break in New York City was independently associated with mobility, and largely driven by residents' socioeconomic status, including proportion of population living in households with more than three inhabitants and proportion of the 18- to 64-year-old population that is uninsured (Lamb et al., 2021). Behavioral changes, measured by multiple mobility metrics for March to May 2020, also seem to matter, with senior communities reacting faster and longer in response to the stay-at-home orders compared to younger communities (Kabiri et al., 2020). Research by Lou et al. involved a comparative analysis of responses between lower-income and upper-income groups and assessed their relative exposure to COVID-19 risks at the county level (Lou et al., 2020). Analysis results showed that higher incomes were related to an improvement in social distancing behavior (Q. Sun et al., 2020). This research informed our study such that levels of income and poverty were included in the random forest model as explanatory variables.

A variety of regression models and algorithms have been used to predict or explain the occurrence of COVID-19. Mollalo et al. modeled over 50 environmental, socioeconomic, topographic, and demographic candidate explanatory variables as well as age-adjusted mortality rates of several disease factors at the county level across the U.S. using Geographically Weighted Regression and ML algorithms such as Artificial Neural Network. The interest was in identifying significant explanatory variables (e.g., median household income, income inequality, and age-adjusted mortality rates of ischemic heart disease) and hotspots of COVID-19 incidence (Mollalo, Rivera, et al., 2020; Mollalo, Vahedi, et al., 2020).

2.4 Materials and methods

2.4.1 Data and study area

The study area for this research comprises three counties in Florida, Miami-Dade, Broward, and Palm Beach, located in the southeastern tip of Florida. One of the unique characteristics of Florida is the large population of retirees (over 65 years), approximately 18% of the state's total population. The southeastern part of Florida has also a diverse population with respect to race and ethnicity, for example, Hispanics comprise 68% of Miami-Dade and 30% of Broward counties respectively, Blacks represent approximately 29% of Broward County, and White Non-Hispanics represent 55% of Palm Beach County (Table 2.1) based on the 2019 American Community Survey (ACS) (*2019 American Community Survey Single-Year Estimates*, 2019).

Table 2.1 Demographics of Miami-Dade, Broward, and Palm Beach counties.

	Miami-Dade		**Broward**		**Palm Beach**	
# of census tracts	519		362		338	
Total population	2,699,428		1,926,205		1,465,027	
Race and ethnicity						
Black	469,202	17.38%	551,097	28.61%	273,384	18.66%
White	2,028,500	75.15%	1,170,083	60.75%	1,077,422	73.54%
Non-Hispanic	850,503	31.51%	1,351,916	70.19%	1,137,087	77.62%
Black Non-His	426,336	15.79%	530,990	27.57%	266,676	18.20%
White Non-His	356,026	13.19%	698,805	36.28%	799,422	54.57%
Hispanic	1,848,925	68.49%	574,289	29.81%	327,940	22.38%
Black His	42,866	1.59%	20,107	1.04%	6,708	0.46%
White His	1,672,474	61.96%	471,278	24.47%	278,000	18.98%
Gender						
Male	1,311,459	48.58%	938,043	48.70%	710,241	48.48%
Female	1,387,969	51.42%	988,162	51.30%	754,786	51.52%
Median household income ($)	52,669		57,433		62,571	
Age group						
0-19	615,919	22.82%	451,353	23.43%	313,436	21.39%
20-39	736,246	27.27%	501,570	26.04%	338,567	23.11%
40-59	765,800	28.37%	539,530	28.01%	373,605	25.50%
60-79	459,748	17.03%	349,128	18.13%	331,428	22.62%
80 and above	121,715	4.51%	84,624	4.39%	107,991	7.37%

We used mobility data provided by the Maryland Transportation Institute (MTI) at the University of Maryland. These data included origin-destination trips data computed from mobile device locations that capture travel patterns at the granularity of census tracts for four time periods per day (6am - 10am, 10am - 2pm, 2pm - 6pm, and 6pm - 6am) (Xiong et al., 2020). The origin and destination trips data were aggregated into inflow (the number of trips per person flowing into a specific census tract from all other places) and outflow (the number of trips per person flowing out of

a specific census tract to all other tracts). As there was very little difference in the patterns of inflow and outflow trips per person per census tract, i.e., when there is a trip flowing into a specific census tract there is usually a trip going out, the number of inflow trips per person per tract was used to analyze mobility in this study (Figure 2.1). Inflow trips per person per unit have also been used in other studies for analyzing mobility (M. Lee et al., 2020; Xiong et al., 2020).

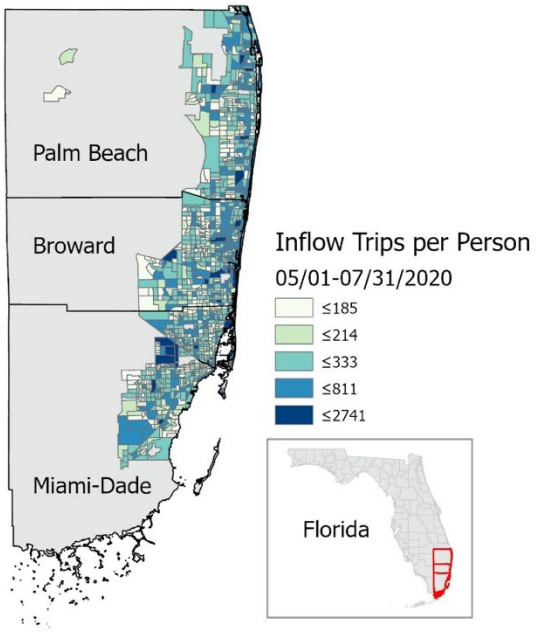

Figure 2.1 Inflow trips per person per census tract 05/01-07/31/2020 in Miami-Dade, Broward, and Palm Beach counties.

As of March 2021, these three counties had the highest COVID-19 severity in the state of Florida, contributing a total of about 38% of the total positive cases and about 33% of total deaths (Florida Department of Health, 2021b), while these three counties comprise over 28% of the total population of Florida. Miami-Dade County was the first county to implement a stay-at-home order among all Florida counties (March 2020) and was the last to lift the order and enter a reopening phase (May 2020).

During this March-May 2020 stay-at-home order period, the cumulative COVID-19 cases reached a total of over 31,000 in the three counties; the number of cases in Florida during the same period reached over 55,000 (Centers for Disease Control and Prevention, 2023). After the stay-at-home order was lifted, COVID-19 cases remained low for the month of May, and then in mid-June, cases began to increase. We examined data for May, June, and July 2020 (a total of 92 days).

County-level data were available from March 2, 2020, when the first COVID-19 case was reported in Florida; ZIP code level COVID-19 case number data was made available from the Florida Department of Health public dashboard from May 18, 2020 (Florida Department of Health, 2021a).

The first two weeks of May were extrapolated based on the overall COVID-19 trend at county level. To be consistent with the other study variables, the ZIP code level data were converted to census tracts using the HUD USPS ZIP Code Crosswalk provided by the U.S. Department of Housing and Urban Development's Office of Policy Development and Research (United States Department of Housing and Urban Development, 2021). The relationship between the daily median inflow trips per person per census tract and daily new COVID-19 cases shows an increase in the number of cases in all three counties after the middle of June 2020 (Figure 2.2). We divided the 3-month period into two time segments, i.e., May 1 to June 15, 2020, and June 16 to July 31, 2020 (both 46 days), and ran random forest models separately for these two periods in order to investigate any changes in the factors that might underlie mobility during these times.

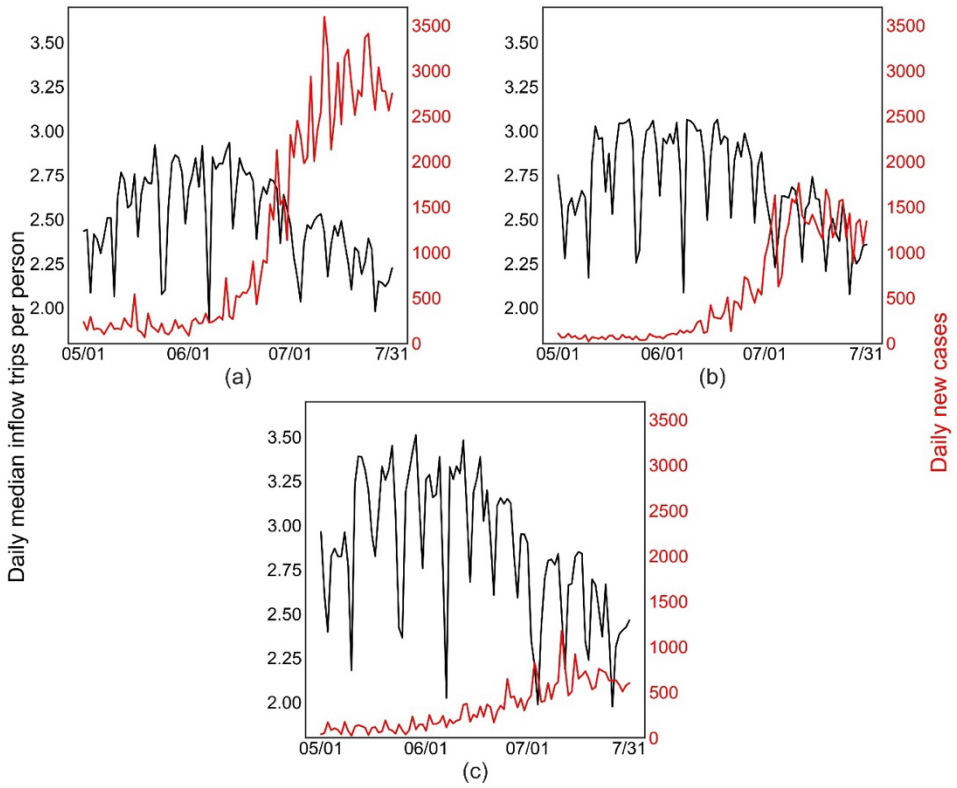

Figure 2.2 Median daily inflow trips per person and daily new COVID-19 cases (05/01-07/31/2020) for (a) Miami-Dade County, (b) Broward County, and (c) Palm Beach County.

We collected additional explanatory variables across three different categories: sociodemographic, travel, and built environment. Sociodemographic factors refer to sociological and demographic population characteristics collected from 2019 ACS, including income, employment, education, race and ethnicity (Figure 2.3), gender, age, and work-related measures. These variables were collected and processed at census tract level. Population demographic details have already been listed in Table 2.1. In this study, Black Non-Hispanic populations refer to Black, and White Non-Hispanic populations refer to White. Based on previous studies finding that different income groups respond differently to the COVID-19 outbreak in terms of practicing social distancing (Lou et al., 2020; Q. Sun et al., 2020), a factor representing essential workers was included in the model using 2019 ACS data and calculated based on a ratio of

service and production occupations, transportation, and material moving occupations to all occupations.

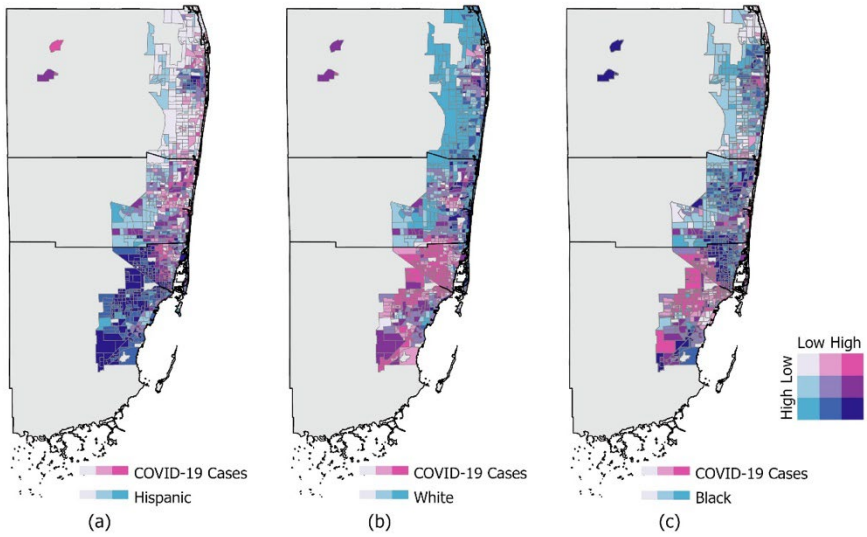

Figure 2.3 Bivariate mappings of COVID-19 cases per 10k people and (a) percent of Hispanic population, (b) percent of White population, and (c) percent of Black population.

Travel-related factors included vehicle mobility behavioral changes impacted by stay-at-home orders, work travel movements, travel distance to beaches, etc. The principal beaches in each county (i.e., Miami Beach, Fort Lauderdale Beach, and Palm Beach), attract both tourists and local people and we assumed these points of interest play an important role in daily mobility patterns during the COVID-19 pandemic. For this reason, the Euclidean distance from census tracts to their corresponding nearest beaches was calculated as one of the travel-related factors. To capture how people's behaviors changed under social distancing requirements, SafeGraph's Social Distancing Metrics dataset consisting of three different variables: percent of time dwelling at home, percent of devices completely at home, and percent of both full time and part time work behaviors (defined as devices spending over 3 hours at a location other than their home from 8am to 6pm) at census block group level were used in this

study (SafeGraph, 2021). The data were generated using GPS locations from anonymous mobile devices to census tract level for consistency. In addition, SafeGraph also provided POI daily visit pattern data at census block group level. Among all the POIs, bars (NAICS code = 722410) and restaurants (NAICS code = 722511) are typically correlated with higher exposure to COVID-19, and limits on bar and restaurant operations have been considered one of the most effective social distancing implementations (Wellenius et al., 2020). The numbers of bar- and restaurant-related POIs for the three counties during May-July 2020 vary by county (Table 2.2). The numbers of bars open in all three counties were likely lower than normal due to COVID-19 business closures. We processed and aggregated the mean daily bar and restaurant visits by census tract for processing in the random forest model.

Table 2.2 Numbers of bars and restaurants in Miami-Dade, Broward, and Palm Beach counties during May-July 2020 (from SafeGraph).

POI	Miami-Dade	Broward	Palm Beach
Bars	68	48	45
Restaurants	5609	3725	2605

Built environment factors were obtained from the Smart Location Database, which is a nationwide geographic data resource for measuring location efficiency maintained by the United States Environmental Protection Agency (United States Environmental Protection Agency, 2021). Among the more than ninety attributes summarizing characteristics, e.g., neighborhood design, transit service, and employment, a set of four spatial and built environmental variables that are most relevant to this study were selected: gross employment density, road network density, street intersection density, and distance to the nearest transit stop. The dataset was

available at the census block group level, which was processed to census tract level for the random forest model. Details of the explanatory and dependent variables used in this analysis and data sources for the variables are provided (Table 2.3).

Table 2.3 Explanatory variables and the dependent variable used in this study.

Category	Variables	Sources
	Explanatory variables	
Sociodemographic	Median household income Unemployment rate Average household size Percent of population with low, medium, and high wages Percent of population with high school degree Percent of population with bachelor's degree or above Percent of the Black population Percent of the White population Percent of the Hispanic population Sex ratio (number of males per 100 females) Age groups 0-19, 20-39, 40-59, 60-79, 80+ Percent of the population working from home Percent of population defined as essential workers	2019 ACS
Travel-related	Mean time travel to work Distance to beach Percent of time dwelling at home Percent of devices completely at home Percent of full time and part time work behaviors Mean bar/restaurant visits	2019 ACS and SafeGraph
Built environment	Gross employment density Total road network density Street intersection density Distance from centroids to the nearest transit stop	Smart Location Database
COVID-19	Cumulative COVID-19 positive cases (05/01-07/31/2020) per 10k people	Florida DOH
	Dependent variable	
Mobility	Inflow trips per person per census tract (05/01-07/31/2020) at census tract level	MTI

2.4.2 Random Forest model

We used Python as the processing language and Scikit-learn as the Python ML package. Before splitting the dataset into training and testing sets, extreme observations

were filtered out in order for these values not to influence the regression model. This included census tracts with a total population less than 500 and population density less than 0.0001 as these were considered to be not representative (e.g., tracts containing the Miami International Airport and the Everglades National Park). Also, outliers in the daily trips per person (i.e., the dependent variable), exceeding the 90% percentile, were removed to avoid the influence of extreme and unusual values skewing the models. The remaining data contained 1065 observations at census tract level which were randomly divided into two subsets. A training set comprising 80% of the data was used to develop the random forest model with 5-fold cross-validation (we also tested with 10-fold cross-validation), and the testing set comprising 20% of the data to assess model performance. To analyze the effect of the training and testing set split ratios, other split ratios, including 60%-40%, 70%-30%, and 75%-25%, were also tested to understand the impact on model performance. Four evaluation measures were used to assess the model performance: (1) Pearson correlation coefficient (r) between the observed values and predicted values, (2) the coefficient of determination (R^2), (3) root mean square error ($RMSE$), and (4) mean absolute error (MAE). $RMSE$ and MAE are defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

While parameter tuning is often applied to avoid overfitting, this step also seeks the optimal combination of given parameters for the best model performance. Four parameters were tuned including the number of trees (*n_estimators*), maximum depth of trees (*max_depth*), the number of features considered when looking for the best split

(*max_features*), and the minimum number of samples required to be at a leaf node (*min_samples_leaf*). Then each combination of parameters was trained with 5-fold cross-validation while the optimal parameters were selected, and the best model performance was returned.

Overfitting occurs when the model is overly trained, resulting in a good fit for a limited set of data, but performs unsatisfactorily when it comes to the unseen out-of-bag testing samples. To prevent overfitting, several techniques were applied in this study, including recursive feature elimination (RFE), which is a feature selection algorithm, parameter tuning, oversampling (Branco et al., 2017; Lemaitre et al., 2015), and adding cost-complexity pruning (CCP) for regularization.

After the optimal model was trained and tested, the contributions of explanatory variables for mobility patterns (i.e., inflow trips) in each county were assessed by visualizing a ranked list of feature importance. In this study, we used the Gini importance to evaluate the feature importance (Breiman et al., 1984). Gini importance is computed as the (normalized) total reduction of a criterion, i.e., the function to measure the quality of a split of randomized decision trees (i.e., the random forest) brought about by a specific feature. We use mean squared error (MSE) as the criterion, and the function was computed by the Sci-kit learn package. The three counties were trained first as one model, and then a model for each county was trained separately for the two time periods so that any differences with respect to feature importance could be compared, and county patterns and trends could be identified.

2.5 Results

2.5.1 Mobility patterns and related sociodemographic factors in the three counties

Our primary interest was in investigating how mobility patterns changed across the three counties during a time in the pandemic when cases were rising, and what were the driving factors underlying these changes. At the county level, the pattern of COVID-19 daily new cases with daily median inflow trips per person (Figure 2.2) showed an increase in the number of cases beginning in mid-June 2020 and continuing into July. In contrast, mobility changes from the first time period to the second declined by -6.07%, -6.29%, and -10.62% for Miami-Dade, Broward, and Palm Beach counties, respectively (Table 2.1). Prior to mid-June 2020, Palm Beach and Broward counties experienced higher inflow trips per person than Miami-Dade County, and Palm Beach County experienced the largest decrease in mobility overall from the first time period to the second compared to the other two counties. Palm Beach County maintained the highest inflow trips per person and the lowest COVID-19 case numbers in the second time period.

Table 2.4 Total inflow trips for 05/01- 06/15/2020 and 06/16-07/31/2020 for Miami-Dade, Broward, and Palm Beach counties.

County	05/01-06/15	06/16-07/31	Change (%)
Miami-Dade	388,724,381	365,125,529	-6.07
Broward	280,165,073	2,62,556,430	-6.29
Palm Beach	219,750,854	196,404,838	-10.62

Pearson correlation coefficients were computed to determine the relationships between inflow trips per person and sociodemographic variables including median

household income and age with significance levels of $p < 0.05$, $p < 0.01$, and $p < 0.001$ (Table 2.5). For the first time period, for Miami-Dade and Palm Beach counties, the correlation between mobility and median household income was weakly positive, while for Broward County it was weakly negative. For the second time period when COVID-19 cases were spiking, Miami-Dade dipped to a weakly negative correlation with median household income, while Palm Beach (with fewer COVID-19 new cases) remained weakly positive (relationship for Broward County didn't change). Examining the relationships between mobility and age groups, showed that younger aged groups tended to be negatively correlated with mobility both before and after the peak in cases, while for older age groups (over 60 years) there was a weak positive correlation in Miami-Dade and Broward counties and a weak negative correlation in Palm Beach County. For the second period where COVID-19 was higher, these relationships continued to hold suggesting that in Palm Beach County there was more concern about the increase in COVID-19 among older-aged individuals.

Table 2.5 Pearson correlation analyses between inflow trips per person and median household income and age groups for Miami-Dade, Broward, and Palm Beach counties for 05/01-06/15/2020 and 06/16-07/31/2020.

	Miami-Dade		**Broward**		**Palm Beach**	
	5/1-6/15	6/16-7/31	5/1-6/15	6/16-7/31	5/1-6/15	6/16-07/31
Income	0.0957*	-0.0268	-0.0301	-0.1097*	0.1570**	0.0247
Age group						
0-19	-0.0701	-0.0717	-0.1742***	-0.1593**	0.0379	0.0517
20-39	0.0576	0.1256**	-0.0069	0.0303	0.1434**	0.1732**
40-59	-0.0965*	-0.1566***	0.1398**	0.1146*	0.1296*	0.1262*
60-79	0.0344	0.0148	0.0652	0.0386	-0.0993	-0.1244*
80 or above	0.0973*	0.0771	0.0081	0.0068	-0.1338*	-0.1404**

Note: * p<0.05, ** p<0.01, *** p<0.001

2.5.2 Mobility patterns and travel-related behaviors

The stay-at-home orders for these three counties were issued at similar times, Miami-Dade County on March 26, and Broward County and Palm Beach County on March 27. Palm Beach County lifted its stay-at-home order on May 11, while Miami-Dade and Broward counties were part of the reopening phase on May 18. Two variables that related to how individuals responded to restrictions in travel, median percent of time dwelling at home (Figure 2.4a) and percent of population staying completely at home (Figure 2.4b) were analyzed at county level. The figures suggest that after the stay-at-home orders were lifted, the percent of time people spent dwelling at home decreased and remained relatively low through mid-June when COVID-19 cases began to spike in this part of Florida and continued to be relatively low compared to the stay-at-home period through the end of July (Figure 2.4a). Miami-Dade County had the highest overall percent of the population who stayed at home throughout the three-month period (Figure 2.4b) while Palm Beach County had the lowest percent.

Patterns associated with either full-time and/or part-time work behaviors were captured through tracking mobile devices that spent more than 3 hours per day away from home (Figure 2.4c). While all three counties had similar patterns with respect to the percent of devices that spent more than 3 hours per day away from home, steadily increasing from early May to mid-June followed by a decrease from mid-June to the end of July, Miami-Dade County had the highest proportion of devices with such pattern suggesting either full-time and/or part-time work behaviors, while Palm Beach County had the lowest, suggesting different rates of work-related behaviors in the three counties.

While there was an overall lower level of mean bar and restaurant visits for the three counties due to COVID-19-related closures, our analysis showed that there was a steady increase in bar and restaurant visits until mid-June when these types of outings showed a sudden decrease followed by a subsequent increase again in early July (Figure 2.4d).

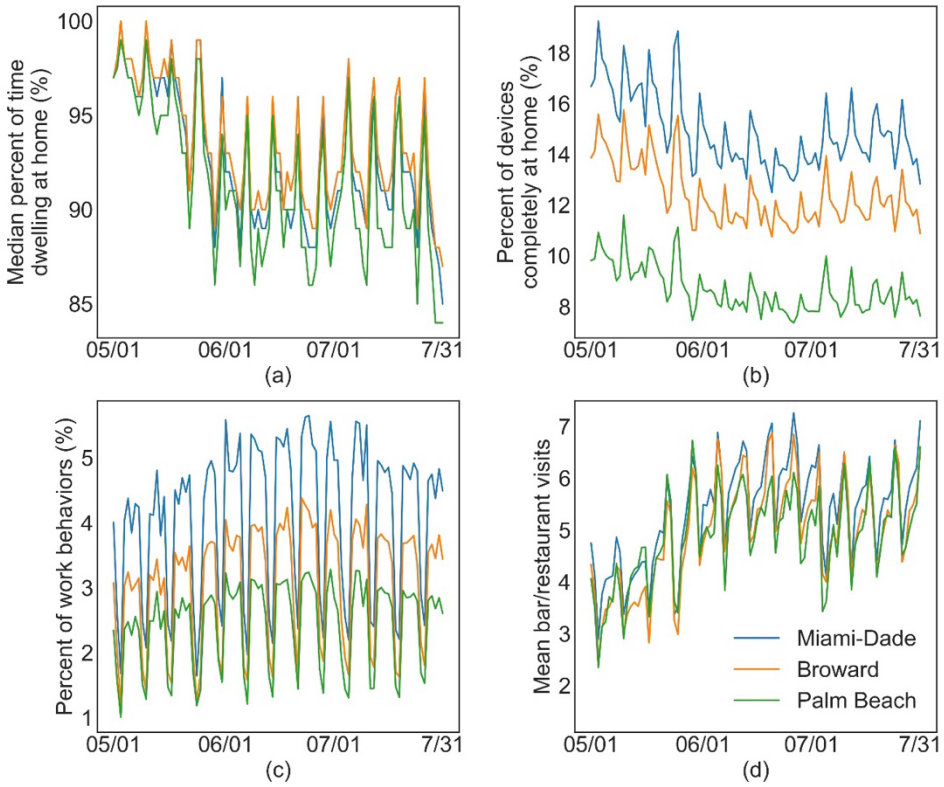

Figure 2.4 Mobility-related behaviors during 05/01-07/31/2020, including (a) median percent of time dwelling at home, (b) percent of devices completely at home, (c) percent of both full time and part timework behaviors, and (d) mean bar and restaurant visits.

2.5.3 Random Forest models

Model Performance

Thirty explanatory variables (Table 2.3) were trained separately for each of the two time periods as features for the random forest regression models. The performance

of all random forest models was assessed using the measures of r, R^2, $RMSE$, and MAE (Table 2.6). We found some interesting variations between the models for each of the counties. With respect to values of r, i.e., the correlation between the observed values and predicted values that reflect how well the predictive model performed, the Palm Beach model returned the highest r values (0.6781 and 0.6766 respectively), followed by Broward and Miami-Dade. This suggests perhaps that the set of analyzed variables performed slightly better for Palm Beach when it came to being able to predict mobility patterns than for the other two counties.

The coefficient of determination (R^2) that measures the percentage of the response variable variation that is explained by the random forest model, was also found to be highest for Palm Beach County, while the R^2 values for both Miami-Dade and Broward counties for the second time period (when cases were rising) were higher than that of the first time period. As we were not able to collect and include all the variables that could be impactful for mobility, for example, changes in employment due to the pandemic and COVID-19 mortality and hospitalization data, it is not completely surprising that the models showed room for improvement. In terms of prediction errors, Broward County had the highest $RMSE$ and MAE, although the values were similarly strong across all models. In general, the model performance for the second time period was better than that of the first time period with higher r values and lower error values.

Table 2.6 Random Forest model performance for 05/01-06/15/2020 and 06/16-07/31/2020 for Miami-Dade, Broward, and Palm Beach counties.

	Miami-Dade		Broward		Palm Beach	
	5/1-6/15	6/16-7/31	5/1-6/15	6/16-7/31	5/1-6/15	6/16-7/31
r	0.5104	0.6068	0.5496	0.6712	0.6781	0.6766
R^2	0.2555	0.3549	0.2964	0.3666	0.4358	0.4415
$RMSE$	34.03	33.31	44.22	42.67	37.27	37.80
MAE	27.21	26.64	36.61	35.48	31.14	28.89

Feature contributions for the period prior to the rise in the COVID-19 cases

Feature importance scores for the three counties were analyzed to obtain an understanding of how the different factors ranked in importance according to the random forest model with respect to the number of inflow trips per person. During the first time period (05/01-06/15/2020) when mobility was relatively high, COVID-19 cases were still relatively low, the number of new COVID-19 cases was ranked 7th in importance in Broward, 8th in Miami-Dade, while for Palm Beach County, this variable was not among the top 15 factors ranked by importance scores. While COVID-19 cases were not so high, the importance scores for both the built environment factors and travel-related factors ranked higher overall than sociodemographic factors (Figure 2.5). Gross employment density was ranked very highly for all three counties (1st for Broward and Palm Beach, and 2nd for Miami-Dade). Other built environment factors, e.g., street intersection density and road network density, were also present in the top 15 factors for all three counties. With respect to travel factors for the first period, these were highly ranked in all three counties, with mean bar and restaurant visits ranked 1st for Miami-Dade, 2nd for Palm Beach, and 5th for Broward. Time spent completely at home, full-time and part-time work behaviors (based on devices being away from home for more than 3 hours), median percent of time dwelling at home, and other social

distancing factors were also in the top 15 factors for all three counties suggesting that the population was also sensitive to the ongoing COVID-19 situation in their region.

With regard to sociodemographic factors during the first time period for Miami-Dade County, the percent of White and Hispanic population was ranked 3rd and 4th respectively for Miami-Dade County. White and Hispanic populations contribute, respectively, approximately 13% and 68% of the total population for Miami-Dade (Figure 2.5a). In Broward County, the percent of both Black and White populations were also in the top 15 rankings, albeit not as highly ranked (positions 9 and 12 respectively), and the percent of Hispanic population was 13th in the rankings (Figure 2.5b). For Palm Beach County, the results were different with important sociodemographic factors relating to income (median household income ranked 5th), employment (general unemployment levels ranked 8th), and education (bachelor's degree and high school degree ranked 13th and 15th respectively) rather than race and ethnicity (not one of the top 15 factors) (Figure 2.5c). These inter-county differences in the model results relating to sociodemographic factors are interesting to note and underscore the kinds of population differences that exist between the counties.

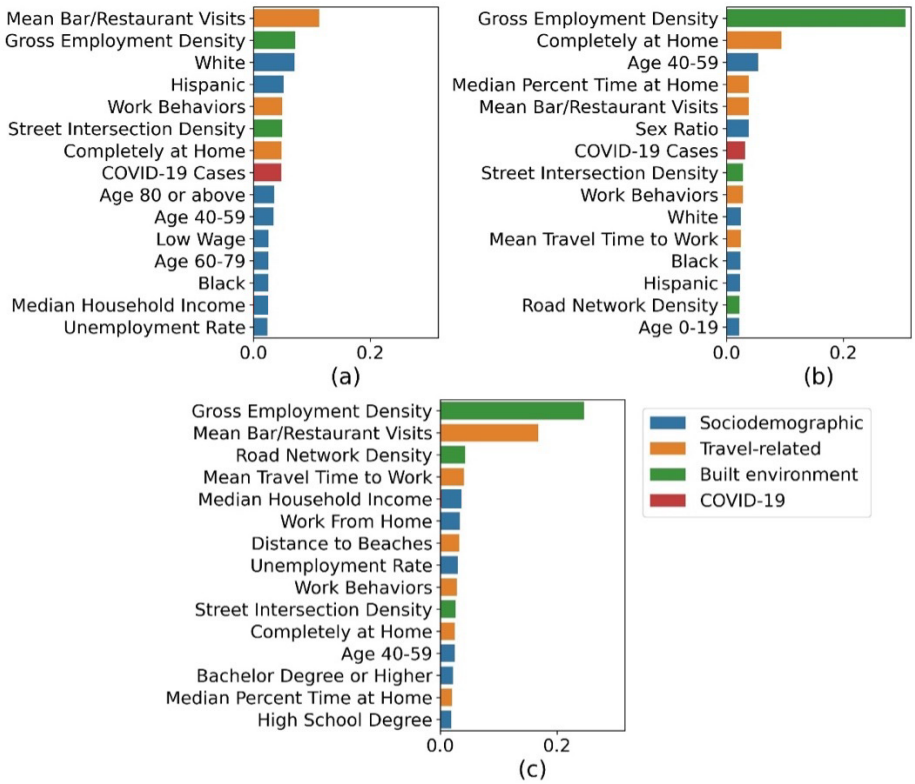

Figure 2.5 The relative importance of the top 15 variables to the number of inflow trips per person (05/01-06/15/2020) using random forest models for (a) Miami-Dade County, (b) Broward County, and (c) Palm Beach County.

Feature contributions for the period following the rise in COVID-19 cases

As the number of new COVID-19 cases began to spike in mid-June 2020, the second period captured some changes in the ranking of variables based on importance scores. Factors that ranked highest in importance during this period continued to be those related to travel and built environment (Figure 2.6). Both gross employment density (1st for all three counties) and the mean number of bar and restaurant visits (2nd for all three counties) continued to be top factors for all the models. In Palm Beach County, the importance scores for these two factors were much higher than for the other counties (Figure 2.6c). Built environment factors, e.g., street intersection density and road network density, were still present in the rankings. Job- and work-related factors, i.e., mean travel time to work and full-time and part-time work behaviors were most

important in Palm Beach County (ranked 3rd and 4th respectively), while for Miami-Dade County, full-time and part-time work behaviors were ranked 6th and for Broward County, they ranked 10th. Mean travel time to work ranked 3rd in Palm Beach, 12th in Miami-Dade, and 15th in Broward County, underscoring how work-related factors seemed to continue as strong drivers in Palm Beach County even with cases rising. Travel distance to beaches was ranked 5th for Broward and 8th for Palm Beach, while this factor was not in the top 15 for Miami-Dade County.

With respect to sociodemographic factors for the second time period, the percent of Hispanic population was a factor in all three county models but was much more of a factor for Miami-Dade County where it ranked 3rd while it was 12th in Broward and 13th in Palm Beach. Black population was 8th in importance in Miami-Dade and 14th in Broward County (not present in the Palm Beach rankings). The age group 40-59 years was another common factor but with different importance, as it ranked 4th for Miami-Dade, 7th for Broward, and 14th for Palm Beach, although the percent population corresponding to ages 40-59 were similar across the three counties (approximately 28%, 28%, and 26% respectively). The factor of age 80 or above ranked at 10 in Miami-Dade and 15 in Palm Beach County. Conversely, the youngest age group (0-19 years) appeared only in Broward County and at rank 13.

The most noticeable change between the two time periods was that the factor representing the number of new COVID-19 cases was much higher ranked for the second time period, being 5th, 3rd, and 9th for Miami-Dade, Broward, and Palm Beach counties respectively. The random forest model was able to discern that the increase in COVID-19 was increasingly important for mobility, even in Palm Beach County where

for the first period of time, COVID-19 cases were not in the top 15 factors explaining inflow mobility.

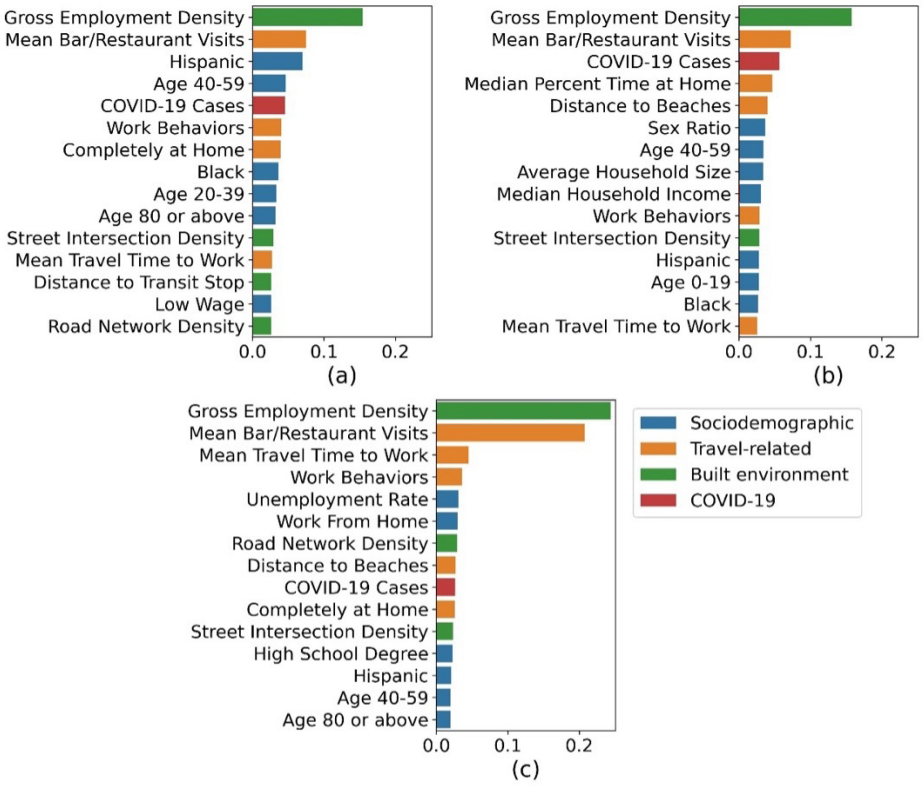

Figure 2.6 The relative importance of the top 15 variables for the inflow trips per person (06/16-07/31/2020) using random forest models for (a) Miami-Dade County, (b) Broward County, and (c) Palm Beach County.

We also analyzed a random forest model trained using all three months together. The Palm Beach model returned the highest r value (0.6672), followed by Broward and Miami-Dade (0.5774 and 0.4946 respectively), which is similar to the order of model performance for the two separate time periods. The results showed that the rankings of important features were similar to the period from mid-June to late July (i.e., the second time period) with mean bar and restaurant visits, gross employment density, and the percent of Hispanic population being the top three factors for Miami-

Dade. These three factors were within our expectations since Miami-Dade County is different from the other two counties in terms of race and ethnicity. Gross employment density, mean bar and restaurant visits, and median percent of time dwelling at home were the top three factors for the Broward model. Similarly, mean bar and restaurant visits, gross employment density, and mean travel time to work were the top three factors for the Palm Beach model. The time spent dwelling at home for Broward County and the mean travel time to work factor for Palm Beach County both relate to social distancing and suggest local county populations were sensitive to the changing COVID-19 situation and how that affected work travel decisions. In this model, COVID-19 new cases were ranked 4th for Broward, 5th for Miami-Dade, and 12th for Palm Beach, reflecting the situation that with the lowest number of COVID-19 new cases, mobility in Palm Beach County was not as influenced by COVID-19 cases, while Miami-Dade and Broward counties experienced higher numbers of COVID-19 new cases and mobility appeared to be sensitive to this situation. The increasing importance of COVID-19 cases as a driver for changing mobility patterns is evident in our models, demonstrating that the pandemic was indeed impacting mobility.

2.6 Discussion

For this research, we used random forest models to understand mobility patterns during the COVID-19 pandemic in three Florida counties, including Miami-Dade, Broward, and Palm Beach counties, and examined a set of sociodemographic, travel, and built environment explanatory factors and their relative importance for explaining patterns of mobility in the context of rising COVID-19 cases. Much of recent research investigating mobility under COVID-19 is at county-level or state-level across the U.S.

(Gao, Rao, Kang, Liang, Kruse, et al., 2020; Mollalo, Rivera, et al., 2020; Mollalo, Vahedi, et al., 2020; Xiong et al., 2020) or at nation-level (Nouvellet et al., 2021; Rahman et al., 2020). However, this research was undertaken at census-tract granularity to discover finer-grained patterns of mobility as well as the drivers for mobility based on the number of inflow trips for each county.

Using a random forest model, we were able to compare the contributions of the explanatory variables over the three counties and over the two time periods. A changing relationship between important features was identified. Previous research suggested an association with COVID-19 cases, and reductions in mobility were correlated with the slowing of COVID-19 spread (Badr et al., 2020; Gao, Rao, Kang, Liang, Kruse, et al., 2020; Xiong et al., 2020). The results of our random forest model analysis indicated that new COVID-19 cases did have an overall impact on mobility for the three counties we analyzed. In Palm Beach County, for example, this factor was much less important until when COVID-19 case numbers started to rise, when this factor shifted to become increasingly important for mobility. Other studies showed that socioeconomic and institutional factors (e.g., median age, percentage of the population employed in services, and percentage of health expenditure) may have limited effects for sustaining social distancing and reduced mobility (Rahman et al., 2020), and studies have also indicated a noticeable correlation between mobility and socioeconomic factors (Gao, Rao, Kang, Liang, Kruse, et al., 2020; Kabiri et al., 2020; Lou et al., 2020). Our random forest models revealed that sociodemographic factors (e.g., race, ethnicity, and age groups) did affect the number of inflow trips (e.g., the percent of the Hispanic population in Miami-Dade County, the age group of 40-59 in Broward County, and

income and employment factors in Palm Beach County) and that based on this result, this group of factors should be considered by decisionmakers and healthcare providers when considering strategies to reach different population groups during a spike in infections.

Due to not being able to collect and include all the variables that could be impactful for mobility, the model performance and overfitting issues could perhaps be improved by including more dimensions of data, e.g., COVID-19 mortality and hospitalization data that are strongly related to healthcare resource availability (Baud et al., 2020; Ji et al., 2020) and changes in employment due to the pandemic. In addition, estimates for essential workers were made using sub-categories of occupation data in the 2019 ACS while 2020 estimates might differ, which might also affect the random forest model results. In terms of mobility data accessibility, the O-D matrix data used in this study was obtained from MTI under a restricted data agreement, and the SafeGraph data was obtained under a restricted data agreement for academic research only. The SafeGraph data currently is no longer available, even for the academic research purposes.

2.7 Conclusions

As the COVID-19 pandemic impacted the daily lives of individuals, this research found that based on tracking inflow trips at census tract level for three counties in Florida, mobility was indeed impacted by COVID-19, especially when compared to mobility during the pre-COVID period (i.e., in 2019). And that during a summertime spike in COVID-19 cases, there were further impacts on the number of trips being made in each county. The set of key explanatory factors revealed by the random forest model

were travel-related factors (e.g., social distancing and work travel-related variables) and built environment factors (e.g., gross employment density and street and road network density), while sociodemographic factors (race and ethnicity, age, household income) were also present. These three counties represent an urban region in the United States that has had a very high number of COVID-19 cases and that has high Black and Hispanic populations that have been particularly vulnerable to COVID-19 infections, as well as a significant population of individuals over the age of 65, also vulnerable to this infectious disease. These different factors that affect the number of trips made across this tri-county region (e.g., social distancing, work travel-related variables, and gross employment density) may be helpful for local officials and public health experts as they review steps and strategies, such as stay-at-home orders and business restrictions or closures. It is also important to note that counties have their unique local characteristics (sociodemographic, economic, points of interest) and our analysis showed how these different characteristics resulted in different sets of factor rankings for each county. While this study focused on counties in Florida, the methodology is generalizable to other locations across the U.S. and other regions. Future research could focus on the model performance improvement and overfitting elimination by including more variables that may be impactful on mobility, e.g., changes in employment during the pandemic, mortality and testing data if available, and trips to additional POIs. Further research on modified random forest approaches, e.g., geographically weighted random forest could offer new opportunities for improved spatial data handling.

Chapter 3: Space-time relationships between COVID-19 vaccinations and human mobility patterns in the United States

3.1 Abstract

As COVID-19 vaccines were administered in early 2021, they helped to mitigate the spread of the COVID-19 virus and signaled an important shift in the pandemic. To better understand how ongoing COVID-19 vaccinations were related to human mobility across the U.S, we identified different mobility-vaccination profiles between January and May 2021 by county in the U.S., using K-means multivariate time-series clustering. The impacts of demographic, socioeconomic, and COVID-19-related variables on different profiles were examined. Results showed 5 different clusters of mobility-vaccination profiles were found for the U.S. One cluster represented counties in metropolitan areas and tourist destinations (e.g., Los Angeles and New York) that had estimated 25% higher mobility and 75% higher vaccination rates than rural counties in the Mountain and South U.S. Census regions (e.g., counties in Arkansas and Mississippi), where people were mobile despite not getting vaccinated. Higher education and household income were found to impact counties' mobility-vaccination profiles. Examination of trip purposes for selected counties returned higher trips to retail/recreation and workplaces for rural counties with relatively lower mobility-vaccination profiles. The results can serve as input for regional and local health officials regarding population responses to a pandemic relevant to economic recovery and future disease prevention.

3.2 Introduction

The SARS-CoV-2 coronavirus disease (COVID-19) has impacted populations around the world since March 2020, when the WHO officially declared COVID-19 as a global pandemic (WHO, 2023). With a total of over 767 million COVID-19 positive cases and over 6.9 million deaths worldwide as of July 2023 (WHO, 2023), the spread of the COVID-19 virus also contributed to a paradigm shift in people's day-to-day activities, making it challenging for a period of time for individuals to carry out their activities in the same way as before the pandemic. In the United States, an important shift in the pandemic was when the U.S. Food and Drug Administration made COVID-19 vaccines available under an emergency use authorization in December 2020 (U.S. Food and Drug Administration, 2020). After the delivery of vaccines to each state, state departments of health identified their priority groups and planned the distribution of vaccines accordingly. By the end of May 2021, five months since vaccinations began, the number of daily new COVID-19 cases had dipped to a new low of 9,000 cases, roughly a 96% decrease from the levels of infection in January 2021 (Centers for Disease Control and Prevention, 2023). Research on the relationship between COVID-19 vaccines and COVID-19 cases has shown that the vaccines in the U.S. returned positive results for controlling the disease burden, not only with respect to case numbers, but for hospitalization and mortality as well (Patel et al., 2021), under different vaccination coverage scenarios in different regions in the U.S. (Alagoz et al., 2021; Yuan et al., 2021). One finding predicted that if vaccine coverage between 20% and 50% could be achieved in the first half of 2021, the reduction in the total number of COVID-19 infections would be between 30% and 50% by the end of 2021, compared

with a no vaccination scenario (Saldaña & Velasco-Hernández, 2021). These predictions under different scenarios have been validated by researchers who indicated that COVID-19 vaccinations in the first five months of 2021 lead to fewer severe COVID-19 outcomes and prevented additional COVID-19 vases (Sah et al., 2021). Vaccinations were investigated for preventing future COVID-19 waves not only in the U.S., but also in China, Japan, Israel, and Europe (e.g., Italy) as well as other countries (B. Huang et al., 2021; Jabłońska et al., 2021; Kurita et al., 2021; Spinella & Mio, 2021).

Patterns of mobility during the pandemic have been a major topic of research as the COVID-19 pandemic significantly changed the way people travelled. People's daily mobility measured, for example, using metrics for home-dwelling time (X. Huang et al., 2022; X. Huang, Lu, et al., 2021), median travel distance (Gao, Rao, Kang, Liang, & Kruse, 2020; Gao, Rao, Kang, Liang, Kruse, et al., 2020), trips by distance (Truong & Truong, 2021, 2022), and miles traveled per person (Tokey, 2021) showed different mobility patterns in different regions in the U.S., and during the different stages of the COVID-19 pandemic (Kim & Kwan, 2021; Lee et al., 2020; Zhang et al., 2021). For example, trip rates, out-of-county trips, and miles traveled per person decreased significantly during the early pandemic (March-August 2020), and COVID-19 infection rates were found to be negatively correlated with miles traveled and out-of-county trips (Tokey, 2021).

With the rollout of vaccines in the U.S. in 2021 and as the government stringency (e.g., activity restrictions and social distancing measures) were relaxed, people could be expected to feel safer once vaccinated, and be more comfortable about

leaving their homes and traveling more (Fiori & Lacoviello, 2021). People might be willing to travel more and undertake more trips. For this study, we examine mobility from January to May 2021 when a widescale vaccination program was put in place to understand whether a relationship between mobility and ongoing vaccinations existed and how these relations may have varied across space and time. Inferring causal impact of vaccinations on mobility is challenging as there are potentially multiple factors including, for example, infection rates of COVID-19, access to healthcare, and government policy, among others, that could also play a role in the relationship. Understanding the spatiotemporal associations between human mobility and vaccination rates during the rollout of vaccinations is crucial for policymakers, businesses, and individuals since insights about the relationship between mobility and vaccination rates can serve as a guide for local and regional transportation demand, economic recovery, and public health policies as citizens adapt to a post-pandemic world.

Existing research has mainly focused on the overall relationship between mobility and COVID-19 cases at state level or country level. There is a gap regarding our understanding of how ongoing COVID-19 vaccination rates were associated with mobility across the U.S. One study focused on Texas, for example, found that reopening after vaccinations in Texas did not appear to impact changes in social mobility, the rate of new COVID-19 cases, or short-term employment (Dave et al., 2021). Moreover, the factors associated with the spatiotemporal differences in mobility and vaccinations require further attention. Potential factors found in existing research were related, for example, to socioeconomic and demographic factors and healthcare

infrastructure attributes associated with COVID-19 cases and deaths (Hu et al., 2022; Radulescu et al., 2021), where race and ethnicity (e.g., Black Americans), health conditions, income, and education were factors impacting infections (Bhowmik et al., 2021; Lamb et al., 2021; Sun et al., 2022).

For this study, we examined how mobility and vaccination rates varied across the entire U.S. by applying a K-means time-series clustering approach based on county-level trends for both mobility (i.e., relative changes of trips) and COVID-19 vaccination rate, and considered the factors for the clustering patterns. Time-series clustering methods (Aghabozorgi et al., 2015; Singhal & Seborg, 2005; Wang et al., 2006), especially multi-dimensional K-means time-series clustering (Giordano et al., 2021; Siebert et al., 2021), have been used in related COVID-19 and mobility research. For example, Elarde et al. (2021) clustered the changes in time spent in public places from March to December 2020 in all counties in the U.S. and returned 3 clusters with outliers. Changes in home-dwelling time calculated from SafeGraph datasets were clustered at different spatial granularities, such as metropolitan statistical areas and census block groups, along with socioeconomic and demographic variables (e.g., economic status, race and ethnicity, gender and age, education, and transportation) to detect spatial units with similar spatiotemporal trends (X. Huang et al., 2020; X. Huang, Lu, et al., 2021). SaTScan clustering (Kan, Kwan, Huang, et al., 2021; Kan, Kwan, Wong, et al., 2021), shape-based time-series clustering (Cao et al., 2022), and spatial autocorrelation techniques (J. Huang et al., 2021) were applied to detect COVID-19 virus transmission and spread.

The objectives of this study are: (1) reveal the patterns of county-level mobility across the U.S. in the context of increasing vaccination rates by week during 2021 using space-time clustering, (2) examine a set of factors that could serve as drivers for the mobility patterns and how these drivers relate to the different mobility-vaccination clusters across the U.S., and (3) examine county-level mobility by analyzing trip purposes for selected urban and rural counties to better understand how trip purposes such as retail/recreation and workplaces among others may be associated with locations having different clusters that represent different profiles of mobility together with vaccination rates.

3.3 Materials and methods

3.3.1 Mobility data

For this study, we investigated mobility patterns at county level across the contiguous U.S. by week from January 2021 to May 2021. The principal data were from Apple Mobility Trends Reports[1] that captured daily relative changes in routing requests from Apple Maps users in the U.S. for driving, walking, or taking public transit (compared to each county's baseline for January 13, 2020). The Apple dataset, which is publicly available, was considered a good proxy for trips in the U.S. for several COVID-19-related studies (X. Huang, Li, et al., 2021; Kang et al., 2020; Noi et al., 2022). However, due to the application of privacy protocols by users, available county-level trip data for 2021 for the U.S. cover only 2,064 counties in the contiguous U.S. out of 3,108 counties in total, approximately 66% of counties.

[1] https://covid19.apple.com/mobility

Since this dataset did not cover all U.S. counties, another mobility dataset, the Federal Highway Administration (FHWA) Traffic Volume Data[2], was used to supplement the Apple data. The FHWA Traffic Volume Data is unweighted raw continuous traffic count data collected by over 7,000 traffic counting stations nationwide reported by State Highway Agencies that cover 1,944 counties, including 448 counties that were not included in the Apple mobility dataset. As the Apple data captures relative travel volumes over time for each county, we similarly calculated relative traffic volumes by county using the FHWA dataset, allowing us to compare these travel volumes with the Apple data. The correlations between the two datasets were analyzed for counties where data were available for both datasets (M=0.59, SD=0.24, over 91% of the counties had p<0.01 correlations), and the analysis showed that the two datasets represented very similar trip trends and could be used together to provide broader coverage of counties with respect to mobility.

After filtering out records with more than 20 missing dates during the study period, the two datasets were used together and readjusted using the median mobility values for the first two weeks in January 2021 (January 1 to January 14, 2021) as the new baseline reference (i.e., 100) for each county. The newly combined mobility dataset captures relative level of changes in the number of trips for 2,445 counties (approximately 79% of the contiguous U.S. counties), which covered an estimation of 317,186,955 population (approximately 96% of the U.S. population), and is referred to as the *mobility index* in this analysis.

[2] https://catalog.data.gov/dataset/tmas-data-program

Another dataset used for this research was trip purpose data for the U.S. sourced from the Google COVID-19 Community Mobility Report[3]. This dataset reports the percent change of travel to different categories of place, i.e., retail and recreation; grocery and pharmacy; parks; transit stations; workplaces; and residential by county in the U.S. compared to a baseline for January 2020. Due to variability in data availability (many counties were missing for certain categories), only four trip purpose categories were selected and examined for this study: retail and recreation; grocery and pharmacy; workplaces; and residential. On average, there were approximately 1450 counties available in the Google dataset for the study period (about 46% of contiguous U.S. counties).

3.3.2 COVID-19 vaccination data

COVID-19 vaccinations were made available nationwide based on priority groups from late December 2020. By the end of May 2021, all adults (18 years and over) in the U.S. were eligible for vaccines. The U.S. COVID-19 vaccination data was collected from the CDC[4] at county level. The CDC dataset reports the cumulative numbers of people who got partly or fully vaccinated every day by county for the period January-May 2021. There were some missing states and counties, for example, the Texas Department of State Health Services did not report their COVID-19 vaccination-related data to the CDC. Another data source, COVID Act Now[5], a non-profit that provides up-to-date COVID-19 news and alerts, was used to fill in data for the missing

[3] https://www.google.com/covid19/mobility/
[4] https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh
[5] https://apidocs.covidactnow.org/#register

counties in the dataset as much as possible. We merged the data from these two sources and calculated vaccination rates, i.e., the percent of population with at least one dose of COVID-19 vaccine, for the five-month period of study for 2,980 counties (approximately 96% of contiguous U.S. counties). After the mobility data and vaccination data were joined, there were 2,395 counties for the nationwide analysis (approximately 77% of contiguous U.S. counties).

COVID-19 cases and mortality data were also collected from the CDC[6] and used in our analyses. Although there were known issues in the CDC COVID-19 dataset, e.g., where state departments of health applied corrections to update their historic data, the number of new cases and new deaths (per 10,000 people) for the whole study time period for each cluster were calculated and used in the analysis.

3.3.3 Demographic and socioeconomic data

To examine mobility in the context of increasing vaccination rates, a set of demographic and socioeconomic variables were analyzed, including county-level urban-rural classification, gender, age, race and ethnicity, economic factors (e.g., median household income and unemployment rate), education, and work-related factors (e.g., median travel time to work and mean travel time to work) (Table 3.1). The urban-rural classification data was from National Center for Health Statistics (NCHS) Urban-Rural Classification Scheme for Counties[7], which classifies U.S. counties into four metropolitan levels (large central metro, large fringe metro, medium metro, and small metro) and two non-metropolitan levels (micropolitan and non-core).

---

[6] https://covid.cdc.gov/covid-data-tracker/#datatracker-home
[7] https://www.cdc.gov/nchs/data_access/urban_rural.htm

In this study, we aggregated the levels as metropolitan (urban) counties and non-metropolitan (rural) counties. The 2021 unemployment rate data was from the U.S. Bureau of Labor Statistics (BLS) Local Area Unemployment Statistics (LAUS) program[8]. The 2021 county-level annual average unemployment rate data was used to represent unemployment rates for the study period. The remaining variables that were considered to be related to both mobility and vaccination rates, were collected from the ACS 2020[9] and processed to county-level percentage data as needed.

Table 3.1 Categories of demographic and socioeconomic variables and their data sources.

Category	Variables	Data sources
Urban-rural	Urban-rural classification	NCHS Urban-Rural Classification 2013
Gender and age	Percent of male, Percent of female, Age 0-19, Age 20-44, Age 45-64, Age 65 and older	ACS 2020
Race and ethnicity	Percent of White population, Percent of Black population, Percent of Asian population, Percent of Hispanic population	ACS 2020
Economic	Median household income, Unemployment rate, Percent of population below poverty, Population density	ACS 2020, BLS LAUS
Education	Percent of bachelor's degree and above	ACS 2020
Work-related	Percent of population working from home, Mean travel time to work	ACS 2020

[8] https://www.bls.gov/lau/
[9] https://data.census.gov/cedsci/

3.3.4 Spatiotemporal clustering using ML

To investigate the relationship between mobility and vaccination rates and identify different mobility-vaccination clusters across the U.S., we applied a ML clustering method, K-means with Dynamic Time Warping (DTW), that supports both temporal and multivariate analysis, using the *tslearn* Python package. Unlike the standard K-means calculation that considers the corresponding Euclidean distance between two time series elements, DTW also considers the time lag between two time series elements and calculates a temporal alignment that minimizes the Euclidean distances between the two aligned series.

The mobility index, i.e., the relative changes in the number of trips with 100 as a baseline, ranged from 21 to 760 over the course of the study period. Vaccination rates, i.e., the percent of population with at least one dose of a COVID-19 vaccine, ranged from 0% to 80% of U.S. adults over 18 during this same period. These two variables are of incomparable units and different variances and would result in unbalanced weights when clustering. To account for these differences, we performed normalization and feature scaling, so that each variable ranged from 0 to 1. The elbow method (Makwana et al., 2013; Syakur et al., 2018) was used to determine the optimal number of clusters (K) that not only could capture differences between clusters but at the same time would keep the number of classes for mapping to a set that was comprehensible.

The resulting clusters represent counties with similar trends for both mobility and vaccination rates by county. When referring to the clusters in this study, we use *mobility-vaccination profiles.* The barycenter of each cluster was computed to represent the average sequence of values from the two time-series that comprise each

cluster. The barycenters (for both mobility and vaccination rates) were compared with the median for both variables to determine differences between clusters.

After the clusters were obtained, the Silhouette score (Silhouette Coefficient) was used to evaluate the clustering quality, as the Silhouette score could represent the separation distance between the resulting clusters (Shahapure & Nicholas, 2020). The Silhouette score has a range from -1 to 1, while values near 1 indicate better clustering quality, values of 0 indicate the overlapping clusters, and negative values generally indicate that samples have been assigned to the wrong cluster.

3.3.5 Examining the differences among clusters

After the clustering analysis was completed nationwide, the characteristics of the mobility-vaccination profiles were analyzed in the context of COVID-19 confirmed cases and deaths as well as the demographic and socioeconomic variables, to detect differences among clusters using ANOVA. The differences among clusters could provide insights into the different mobility-related behaviors and differences in levels of mobility and vaccination rates. This analysis will shed light on how these variables could be related to the different patterns of mobility-vaccination profiles. We assume that if the distribution of a specific variable for a cluster (e.g., median household income) is distinct or unique, then that variable is more impactful for distinguishing clusters from each other, and could contribute to understanding differences among clusters, as well as providing further insights about the different mobility and vaccination rate patterns captured by the clusters. An ANOVA analysis was applied to check the differences in variance for these variables. A smaller p-value for a variable

reflects more significant differences and indicates that the variable has a higher impact on the clustering results.

After the differences among the clusters were tested using ANOVA, post hoc tests were applied to identify clusters that differed significantly from each other. We used the Tukey HSD test for this analysis (Abdi & Williams, 2010). The post hoc tests were used to guide the analysis of significance levels for comparing between cluster means (i.e., $Cluster_1 - Cluster_2$, $Cluster_1 - Cluster_3$, …, $Cluster_{n-1} - Cluster_n$). The adjusted p-values that statistically reflect the differences between two means for all pairings for all factors were computed using a significance level of 0.05 for this analysis.

3.4 Results

3.4.1 Dynamic trends of mobility and COVID-19 vaccination rates

To analyze dynamic mobility patterns in the U.S., the mobility index for January to May 2021 by week showed spatial and temporal differences across the nation (Figure 3.1). In Week 5 (the end of January), counties in the U.S. showed little to no changes in mobility from the base level, with counties close to Yosemite National Park in California returning higher-than-baseline mobility index values, approximately 110 (Figure 3.1a). By week 9 (the end of February 2021), the median index value for the whole country had risen to 113, returning an overall mobility index increase compared to January (Figure 3.1b). From Weeks 9 to 14 (late February to the end of March 2021), there were greater changes in mobility, with national average index values increasing from 113 to 130. Counties in Utah, Alabama, and South Carolina had

higher values of approximately 180 (Figure 3.1c). Weeks 14 to 18 (April 2021) continued to show an increase in mobility with average index values of around 133, with high values in Montana, Wyoming, and Utah (Figure 3.1d). In Week 22 (the end of May 2021), the greatest changes in mobility were noted (Figure 3.1e). By this week, the mobility index values in over 82% of counties exceeded 200 across the country. Counties in western states, for example, Wyoming, Utah, Arizona, Idaho, and Oregon, tended to have higher index values (e.g., Wyoming of 214 and Oregon of 220) than counties in the east (e.g., Rohde Island of 160 and New Jersey of 167) (Figure 3.1e). Over the entire study period, Goshen County, New Jersey returned the lowest mobility index value of 21 in Week 11, while Cape May County, New Jersey returned the highest mobility index value of 763 in Week 22.

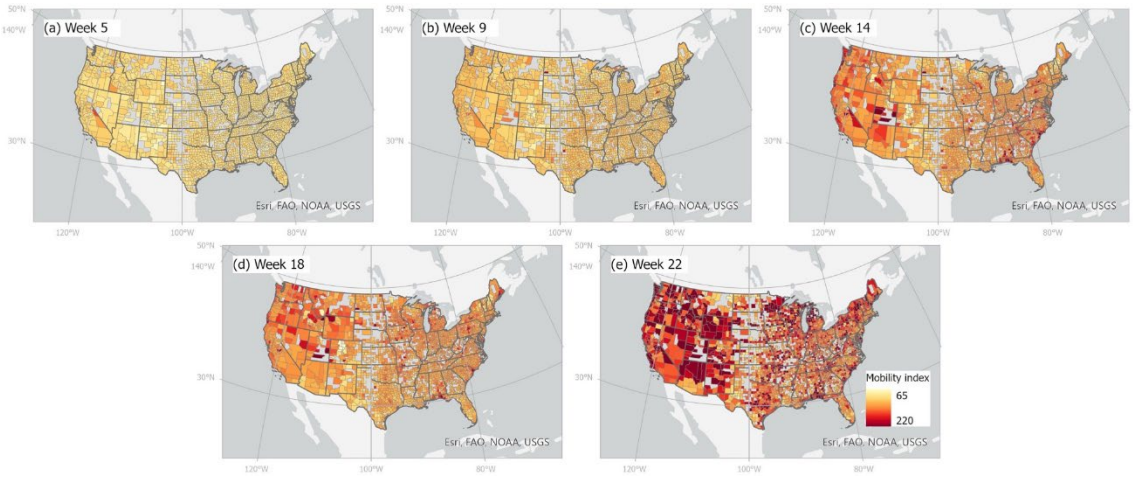

Figure 3.1 Mobility index in the U.S. at county level (01/01-05/31/2021) by week. Weeks 5, 9, 14, 18, and 22 are selected for the end of each month.

The percent of population with at least one dose of a COVID-19 vaccine by county was mapped by week for the period January-May 2021 (Figure 3.2). In Weeks 1-5 (January 2021), when only high-priority healthcare workers were eligible for

vaccines, the percentage of the vaccinated population ranged between 6% and 10% (Figure 3.2a). By Week 9 (end of February 2021), vaccination rates had increased to 15% (Figure 3.2b). By Week 14 (the end of March), states started to show greater variance in the percent vaccinated, for example, Colorado (12.0% on average) and New Hampshire (12.9% on average) had significantly lower vaccination rates than other states at this time, e.g., Connecticut with 39.5% with at least one vaccination, and Rhode Island with 38.2% (Figure 3.2c). Vaccination rates on average for all counties experienced the greatest increase in March 2021, from 15.1% in Week 9 to 28.4% by Week 14. The increase slowed down in April 2021, as vaccination rates on average reached 34.3% by Week 18 (Figure 3.2d). By week 22 (the end of May 2021), vaccination rates had risen on average to 37.1%, and vaccination rate differences among counties can be easily identified (Figure 3.2e), for example, counties in Virginia, Georgia, and Colorado showed rates of approximately 20%, while counties in New England (e.g., New Hampshire, Vermont, and Maine) showed higher rates, approximately 50%-60%. Overall, the percent vaccinated showed greater increases between Week 9 and Week 14 (late February to the end of March 2021), which was the same period when there were greater increases in the mobility index (approximately 113 to 130) (Figure 3.1c). In April and May, differences among states appeared for both mobility and vaccination rates.

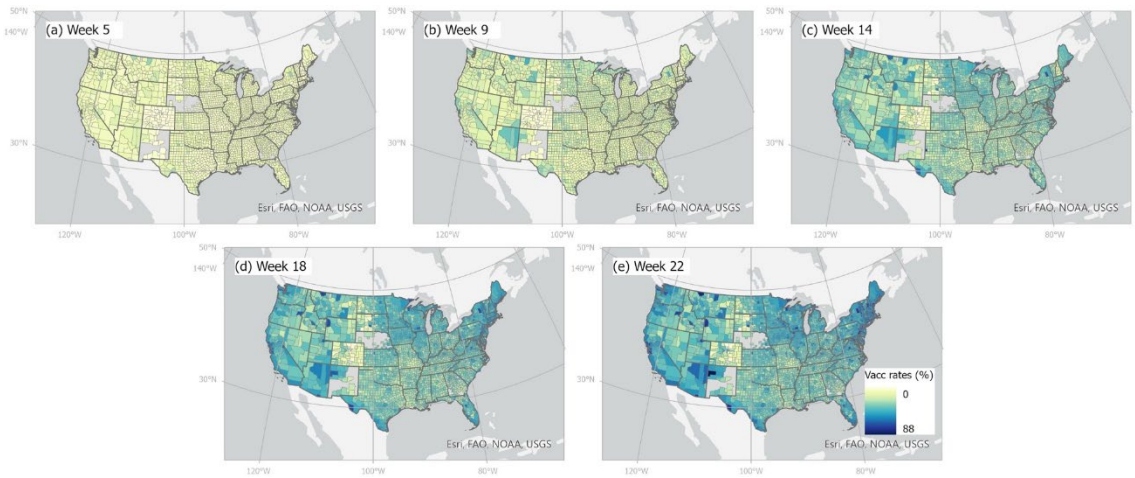

Figure 3.2 Cumulative vaccination rates (percent of population with at least one dose of COVID-19 vaccine) on (a) 01/31/2021, (b) 02/28/2021, (c) 03/31/2021, (d) 04/30/2021, and (e) 05/31/2021.

3.4.2 Spatiotemporal clustering results

Using the elbow method, five clusters ($K = 5$) were selected for visualization and analysis. The Silhouette score for the clustering results was 0.3844, which indicated that the five clusters were generally independent of each other. A closer examination into the Silhouette score showed that there was minimal overlap between clusters and that most counties were a member of a single cluster. The barycenters of mobility-vaccination profiles were compared with the median values (Figure 3.3). The temporal trends of the clusters over the weeks were more similar before Week 9 and started to separate after Week 9 (Figure 3.3a).

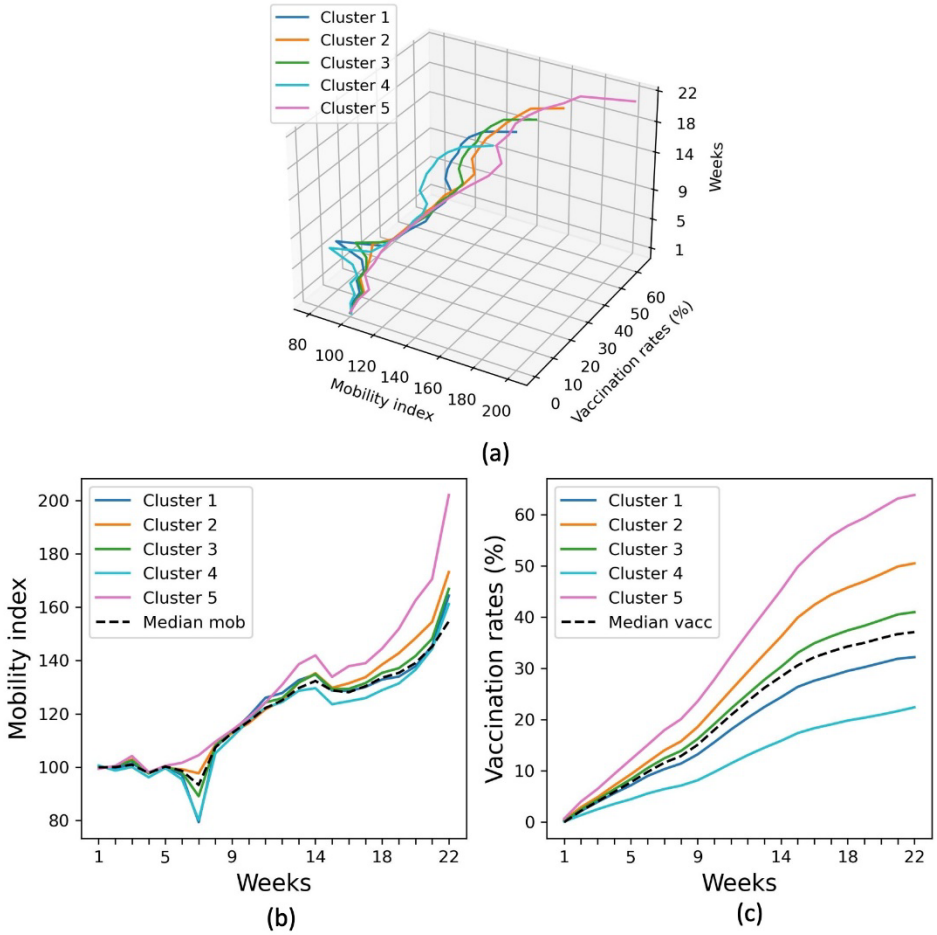

Figure 3.3 (a) Barycenters for five clusters plotted in a 3-dimensional approach. Barycenters for (b) mobility and (c) vaccination rates for each cluster compared to the median mobility and median vaccination rates separately.

With respect to the mobility differences among clusters (Figure 3.3b), in Weeks 1-5 (January 2021), all five clusters fluctuated around a mobility index value of 100 and showed only minor differences between them. In Week 7, a large decrease in the mobility index could be seen in all clusters except for Cluster 5. Mobility index values for Clusters 1 and 4 were approximately 80, while that of Cluster 5 was around 103. The mobility index values for all clusters increased to approximately 110 (just over baseline) by Week 9 (the end of February). All clusters' mobility index values experienced increases from approximately 110 to 130 in Weeks 9-14. From Week 14,

the differences among clusters start to become more noticeable with Cluster 5 showing higher relative mobility and Cluster 4 lower relative mobility. For Weeks 14-22 (April and May 2021), except for a slight decrease in Week 15 (around 7%), mobility for all clusters increased, for example, the mobility index of Cluster 5 increased from 132 to 200, and Cluster 2 went from 128 to 170. By Week 22 (the end of May), Cluster 5, with the highest mobility index value of 200, was 25% higher than Clusters 3 and 4 which had a lower mobility index of 160.

The patterns of vaccination rates among clusters were different from that of mobility as the trend for each cluster was distinct from that of other clusters over the entire time period. Vaccination rates increased the most rapidly during Weeks 9-15 (March to early April 2021) (Figure 3.3c). Cluster 5 had the highest rates throughout the time period and Cluster 4 had the lowest. In Week 22, Cluster 5 counties had approximately 70% higher vaccination rates than Cluster 4 counties.

Differences in percentage vaccinated among clusters could be easily identified, while differences in mobility among clusters were not as marked. For example, Cluster 4 was comprised of counties with approximately 70% lower vaccination rates than Cluster 5, but Cluster 4 had approximately 25% lower mobility index than Cluster 5 by the end of the study time period. Scatter plots of the mobility index and vaccination rates for clusters over the weeks supported this finding as clusters were distinguished mainly by levels of vaccination rates, demonstrated by the varying colors along the y-axis (Figure 3.4).

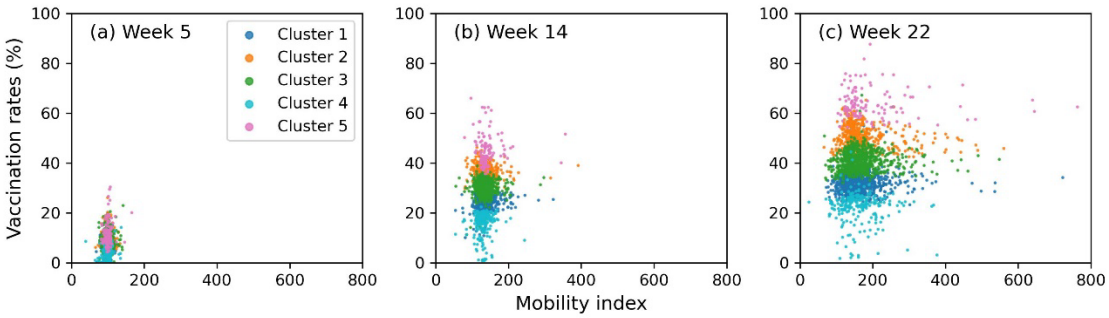

Figure 3.4 Scatter plots of mobility index and vaccination rates for selected weeks (Weeks 5, 14, and 22).

The different clusters were distributed across the U.S. (Figure 3.5). Cluster 5, capturing a profile based on the relatively highest mobility and vaccination rates, could be found in metropolitan areas, for example, Santa Clara County, CA is where the Silicon Valley is and New York County, NY is where Manhattan is. Cluster 5 counties could also be found in tourist destinations, for example, Coconino County, AZ was famous for the Grand Canyon National Park and Mono County, CA is close to Yosemite National Park. Cluster 2 counties, with profiles of high mobility-vaccination, were found in the areas surrounding Cluster 5 counties (over 47% of Cluster 5 counties were surrounded by Cluster 2 counties), for example, Santa Barbara County, CA (close to Los Angeles) and Queens County, NY (right next to Manhattan and part of New York City).

Cluster 1 representing relatively lower mobility-vaccination profiles and Cluster 4 representing the relatively lowest mobility and vaccination rates were mostly found in the Mountain and South census regions, with over 47% of Cluster 4 counties being surrounded by Cluster 1 counties (Figure 3). Cluster 1 counties comprised many of the counties in Arkansas (69%), Mississippi (66%), Tennessee (63%), and Louisiana (61%), while over 80% of counties in Colorado were categorized as Cluster 4.

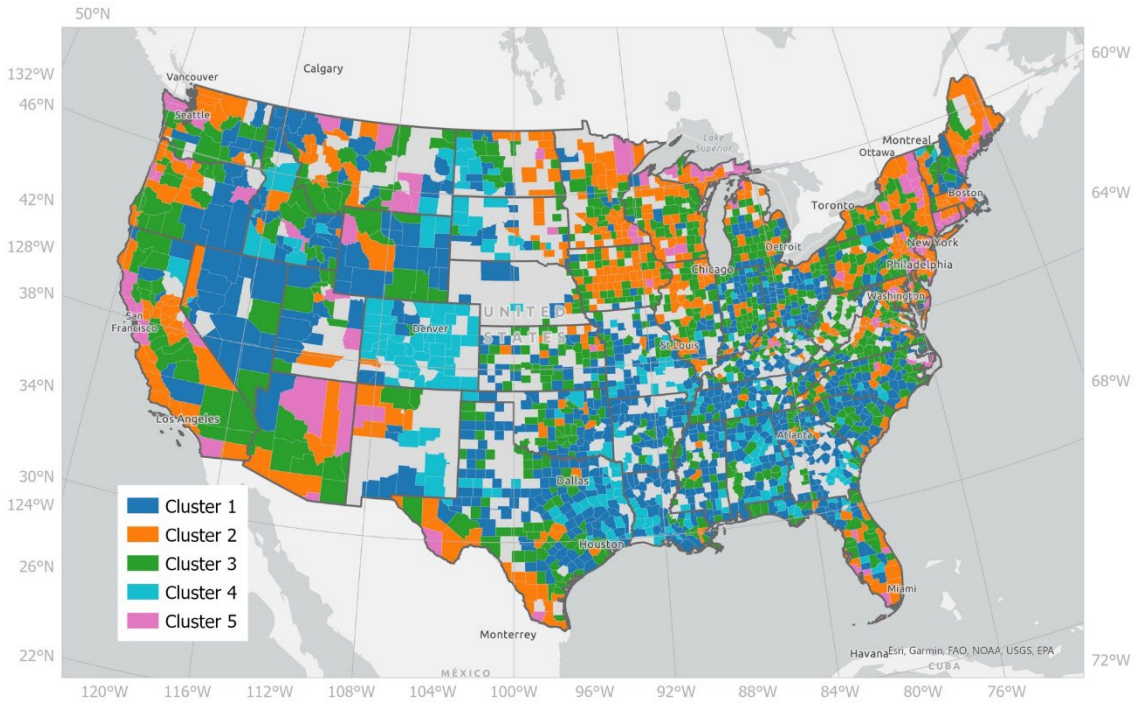

Figure 3.5 Spatial distributions of the clusters of mobility and vaccination rates across the U.S.

3.4.3 Underlying characteristics of clusters

The results of the ANOVA analysis provided further insights into the different clusters. The nineteen sociodemographic variables are listed in ascending order of p-values (Table 3.2), where the top five of these variables were percent of Bachelor's degree and above, percent of population working from home, median household income, percent of Asian population, and percent of population below poverty. These results suggest that the mobility undertaken by individuals during this period as vaccines became more available varied across the U.S. depending on the level of education, work-travel demand, race and ethnicity, and economic status.

Table 3.2 ANOVA analysis of p-values of the COVID-19 cases and deaths as well as demographic and socioeconomic variables in ascending order.

Ranking	Variable	p-value
1	Percent of bachelor's degree and above	5.38×10^{-191}
2	Percent of population working from home	7.72×10^{-100}
3	Median household income	1.59×10^{-92}
4	Percent of Asian population	1.96×10^{-71}
5	Percent of population below poverty	3.55×10^{-41}
6	Age 0-19	1.36×10^{-31}
7	COVID-19 deaths	1.28×10^{-19}
8	Percent of male	6.18×10^{-16}
9	Percent of female	6.18×10^{-16}
10	Percent of Black population	3.25×10^{-12}
11	Population density	2.51×10^{-11}
12	Age 65 and older	1.04×10^{-10}
13	Age 45-64	3.84×10^{-6}
14	COVID-19 cases	2.16×10^{-4}
15	Mean travel time to work	3.06×10^{-3}
16	Unemployment rate	7.15×10^{-3}
17	Percent of Hispanic population	4.80×10^{-2}
18	Percent of White population	6.38×10^{-2}
19	Age 20-44	2.23×10^{-1}

Taking a closer look at these variables, we analyzed the mean values and standard deviations of the variables for each cluster following the ANOVA order (Table A1). The Tukey HSD post hoc test returned values for comparisons between all five clusters, providing a measure for how clusters differed from each other (Table A2). With regard to education, a positive association was found between education and the mobility-vaccination profiles as Cluster 5, with the highest mobility for most weeks and highest vaccination rates, was comprised of counties with an average 28.86% population with a Bachelor's degree and above, significantly higher than the other four clusters, while Cluster 4, with the lowest mobility and vaccination rates (in Weeks 6-8 and Weeks 13-22), had only about 13.33%. For household income, it was significantly different among all five clusters and was positively associated with higher mobility and vaccination rates, for example, the median household income for Cluster 5 on average was $73,606, approximately 43% higher than that of Cluster 4 ($41,408).

With respect to race and ethnicity, the percent of Asian population was ranked 4th in importance. Clusters 5 and 2, with higher mobility and vaccination rates, had 4.88% and 2.69% of Asian population respectively, significantly higher than in Clusters 1 and 4 (lower mobility and vaccination rates) with only 0.85% and 0.80% respectively. As for the percent of Black population, a higher percent of Black population was associated with lower mobility-vaccination profiles. Clusters 4 and 5, with lowest and highest mobility and vaccination rates respectively, had 10.19% and 5.34% percent Black population respectively. We could not draw a clear conclusion between the clustering results and the percent of both Hispanic and White populations as the five clusters were not significantly different from each other, suggesting that further research into these particular population-based differences in mobility-vaccination by county is warranted.

In terms of age, the 20-44 age group, ranked 19th in importance and we did not see great variance among the five clusters (i.e., all comparisons between five clusters were not significantly different). As an age group likely to work, they may have been working at home (so mobility could be lower) and they may have been later in getting vaccinations due to not being in an age-related risk group. For older age groups, the percent of individuals aged 65 and over in Cluster 5 counties was 20.24%, significantly higher than all other clusters, suggesting that for certain locations, this age group was both vaccinated and mobile.

The results also showed that the rates of COVID-19 deaths appeared to be impactful for mobility-vaccination patterns (ranked 7th in importance) where rates of

COVID-19 deaths increased, from 6.13 deaths per 10,000 people (Cluster 5), 7.16 (Cluster 2), to 8.32 (Cluster 4).

We also analyzed the percentage of urban counties by cluster and found differences did exist. For example, Cluster 5 and Cluster 2 were each approximately 60% urban counties. In contrast, Cluster 4 and Cluster 1 were comprised of 70% and 64% rural counties, respectively. Cluster 3 was comprised of 46.65% rural counties, which corresponded the locations of Cluster 3 counties (Figure 3.5).

3.4.4 Trips to different categories of places

Five counties, three in Maryland (MD) and two in Alabama (AL), were selected to further examine mobility by analyzing changing patterns of trips to different categories of places (i.e., retail and recreation, grocery and pharmacy, workplaces, and residential places) among the clusters (Figure 3.6). Montgomery County, MD, and Prince George's County, MD were selected as they were considered representative of Cluster 5, Cluster 2 counties respectively. These two urban counties are adjacent to Washington D.C. and represent urban areas with high population density (both have over 700 people per square mile). Charles County, MD was selected as it was representative of a Cluster 3 county, which is urban and also close to large population centers, but not as populated as Montgomery and Prince George's Counties. In contrast, Walker County, AL and Blount County, AL were selected as examples of Cluster 1 and Cluster 4 counties respectively. These two counties have an economic focus on mining and manufacturing and represent rural areas in the South with lower population density (both around 30 people per square mile).

Montgomery County was classified as Cluster 5 (relative highest mobility and vaccination rates). Montgomery County returned an overall increasing trend for trips to retail/recreation and grocery/pharmacy places and showed an overall decreasing trend for trips to residential places over the study period. The trips to workplaces remained at a lower level (around 50% below baseline) until Week 21 when it started to increase and reached a level of approximately 30% below baseline by the end of May 2021 (Figure 3.6a). Prince George's County was found to be a Cluster 4 county (relatively high mobility and vaccination rates). Prince George's County had a very similar overall trip pattern for all four categories of places as Montgomery County, with only minor differences in the relative levels of trips to retail and recreation (reaching 10% below baseline in Week 22) and trips to workplaces (remained around 40% below baseline during Weeks 1-21) (Figure 3.6b). Although these two urban counties in Maryland were identified as Cluster 5 and 2 with relatively higher mobility, they had different patterns regarding trip purpose with generally higher numbers of trips to retail and grocery and lower overall work-related travels likely mainly due to remote work. Charles County (median level mobility and vaccination rates) showed similar trends with Montgomery and Prince George's Counties, particularly Prince George's County (Figure 3.6c). This county had a lower number of trips to workplaces with an overall increase of only about 10% (from 50% to 40% below baseline) during the study period compared to Montgomery and Prince George's Counties that had 20% increases. Walker County (lower mobility and vaccination rates) and Blount County (lowest mobility and vaccination rates) showed different trends in trip purposes than the Maryland counties. Trips to retail and recreation for Walker County experienced a 30%

increase during Weeks 8-14 bringing the mobility to 20% above baseline. Trips to workplaces fluctuated between 20% and 10% below baseline during the study period (Figure 3.6d). Blount County saw an even larger increase of 40% for trips to retail and recreation (30% above baseline) during Weeks 8-14, and slightly lower trips to workplaces (ranging between 30% to 20% below baseline) (Figure 3.6e).

Trips to workplaces had the lowest levels among all categories for all five counties, suggesting people were not going to workplace locations as frequently during the study period. Walker and Blount Counties had lower percentages of population working from home (1.31% and 0.80% respectively) and their main occupations (mining and manufacturing) possibly contributed to higher trips to workplaces (20% below baseline) than Montgomery, Prince George, and Charles Counties (40% below baseline) whose major occupation is professional, scientific, and technical services. Although Walker and Blount Counties had lower vaccination rates (33.7% and 21.3% respectively) than Montgomery, Prince George, and Charles Counties (70.1%, 50.8%, and 45.6% respectively), Walker and Blount Counties had generally higher numbers of trips to retail and recreation. Trips to grocery and pharmacy increased by 20% over the period of time for all five counties.

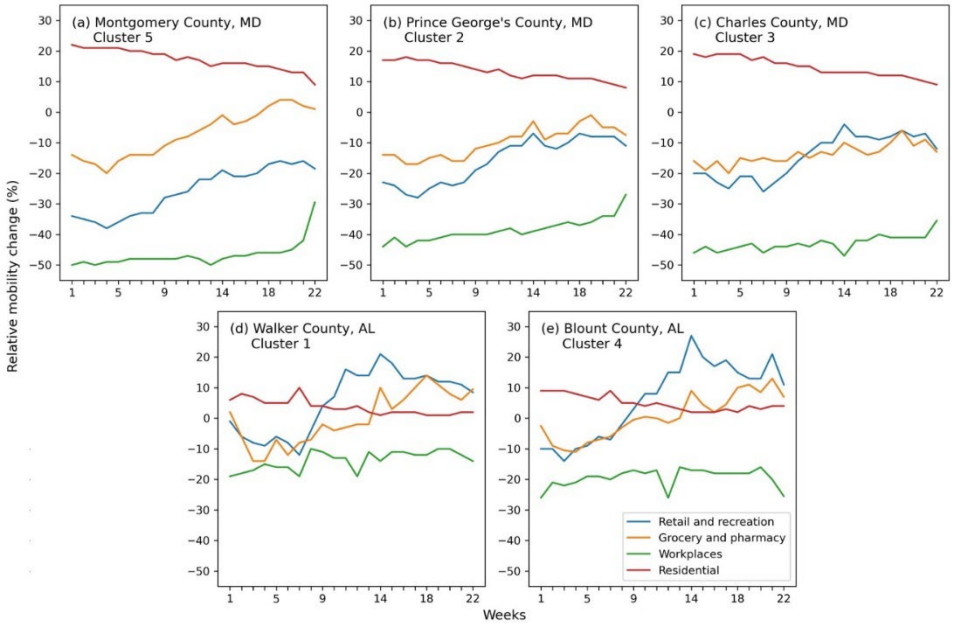

Figure 3.6 Trips to different categories of places for (a) Montgomery County, MD, (b) Prince George's County, MD, (c) Charles County, MD, (d) Walker County, AL, and (e) Blount County, AL.

3.5 Discussion

Our analyses showed that by the end of March 2021 as COVID-19 vaccinations were administered, differences in mobility among U.S. counties were becoming more notable as the number of trips in some locations started to increase from a below-baseline level of travel. This result concurs with other research that also used the Google Mobility Report dataset and the travel behaviors of two groups of people (fast vaccinators and slow vaccinators) across countries (Israel, U.K., U.S., Canada, etc.) diverged after mid-February (Fiori & Lacoviello, 2021). Our spatiotemporal clustering results revealed that differences between mobility and vaccination rates existed across U.S. counties during the study period. Counties with relatively higher mobility and vaccination rates were in populated metropolitan areas and tourist destinations, and counties with relatively lower mobility and vaccination rates were in Mountain and

South U.S. Census regions (e.g., Arkansas, Mississippi, and Louisiana). The five clusters showed different spatial and temporal associations between mobility and vaccination rates, with vaccination rates varying more between clusters than mobility. We saw in Cluster 4, a trend of increasing mobility with vaccination rates lower than in other clusters, while Cluster 5 consistently returned higher (the highest) mobility and vaccination rates.

Statistical tests revealed the top five variables that were considered most related to the mobility-vaccination profiles were percent of Bachelor's degree and above, percent of population working from home, median household income, percent of Asian population, and percent of population below poverty. This result concurs with previous research that found that different income profiles contributed to different mobility patterns, e.g., response to travel restrictions (Lou et al., 2020; Sun et al., 2020, 2022) and pandemic outcomes, e.g., COVID-19 cases and deaths (Lamb et al., 2021; Mollalo et al., 2020; Zhu et al., 2021). Asians and college and/or graduate degree holders have also been found more likely to accept vaccinations (Malik et al., 2020), which confirmed the findings in this study as well. Other researchers found that Black population groups were the most impacted by the transmission of COVID-19 and had the highest mortality rates (Bhowmik et al., 2021), while also having the lowest probability of likely getting a vaccine (Malik et al., 2020; Nguyen et al., 2021; Niño et al., 2021) In our study, the percent of Black population was ranked 10^{th} in the ANOVA analysis also indicating that this demographic was a factor differentiating the mobility-vaccination profiles.

There were limitations with respect to data availability, as the availability of mobility and vaccination rates data at county level was limited or unavailable for some counties. Our county results may be underestimates of both mobility and vaccinations rates, as both the mobility data and COVID-19 vaccination data did not cover all counties in the contiguous U.S. or were suppressed to protect data privacy. Mobile device data has associated biases, which was a challenge to personal data privacy and ethical concerns (Gao, Rao, Kang, Liang, & Kruse, 2020; Li et al., 2021). Although the mobility data were underestimated due to selection bias, the mobility trends (i.e., relative changes of trips) over space and time were considered to represent actual travel trends. In terms of the mobility data accessibility, the mobility data used in this study, i.e., Apple Mobility Trends Report and Google Community Mobility Report, was publicly available at the time of my analysis, however, they stopped updating the mobility data when the COVID-19 Pandemic was mitigated (on April 12, 2022 and October 15, 2022, respectively). Google had its historic data still available online, while the Apple data is no longer available. The FHWA Traffic Volume Data is publicly available, as FHWA poses their data on an U.S. government data sharing platform.

3.6 Conclusions

This study investigated the dynamic patterns of mobility in the context of ongoing COVID-19 vaccinations at county level in the contiguous U.S. by week from January to May 2021 in order to examine whether mobility patterns were responsive to increasing vaccination rates. Differences in mobilities among counties were found as the COVID-19 vaccines were distributed and counties responded differently to vaccine uptakes. Population metropolitan areas and their surrounding counties, as well as

counties with tourist destinations, showed both higher mobility and higher vaccination rates, compared to counties in the Mountain and South U.S. Census regions, including, e.g., Arkansas, Mississippi, and Louisiana, where both mobility and vaccination rates were 25% and 70% lower respectively by the end of the study period. Education, income, and race/ethnicity were sociodemographic variables that contributed to different travel behaviors the most, while percent of Black population and the number of COVID-19 deaths, also contributed to spatiotemporal non-stationarity found with the clusters. Three counties in Maryland and two counties Alabama, representing urban and rural counties respectively, were analyzed with respect to different trip purposes and showed that the two rural Alabama counties comprised mainly of Cluster 1 and 4 (relatively lower mobility and vaccination rates) had higher trips to retail/recreation and workplaces as compared to the urban counties in Maryland comprised of Cluster 2 and 5 (higher mobility and vaccination rates) and Cluster 3 (median level mobility and vaccination rates).

While this research is explanatory research to investigate the associations between mobility and vaccination rates, future work may apply deep learning methods, e.g., Long Short-Term Memory and GeoAI, to infer causal relationships between mobility and vaccination rates and investigate additional underlying drivers, e.g., government policy and vaccination acceptance for dynamic mobility patterns.

Chapter 4: Investigating factors that impact vehicle travel time using Machine Learning approaches

4.1 Abstract

Travelers in the U.S. heavily rely on driving for all purposes, especially for short- and mid-distance trips. It is critical to understand what factors and how these factors contribute to different travel times even for the same origin-destination (O-D). To fill this gap, we computed travel times for two O-Ds in an urban area with different trip lengths/purposes. Two previously understudied factors, driver route choice and driver travel speed behaviors, were derived from GPS data using a clustering method. Factors from three categories, including driver behaviors, built environment and road characteristics, and external (e.g., traffic incidents and weather), were examined for their contributions on travel times for different trip lengths/purposes separately using Random Forest models. Some key results were built environment and road network factors were generally more impactful than driver behaviors and external factors. Travel speed behavior and departure time of the day were commonly important for travel times. Differences between different trip lengths and trip purposes were identified. The results not only filled the gap in research, but also could be contributing to future research in more accurate travel time prediction. The results are also meaningful for urban and transportation planners.

4.2 Introduction

In today's rapidly changing world, transportation is essential to our everyday lives as we rely heavily on driving for all purposes, e.g., commuting to work, traveling

to meet friends and family, and delivering goods. By examining and understanding how different factors impact travel time (i.e., the driving time to a destination) in different contexts, e.g., over different travel distances and for different trip purposes, transportation planners and policymakers can develop more effective strategies to enhance driver awareness, reduce traffic congestion and improve overall transportation efficiency (Carrion & Levinson, 2012; Z. Wang, Fu, et al., 2018). For this study on travel time, factors were categorized into three different categories including driver behavior, built environment and road network, and external factors. Driver behavior factors refer to individual driver characteristics and driving behaviors such as departure time and choice of routes (Feng et al., 2021; Kaplan et al., 2015). Built environment and road network factors refer to characteristics of the road network, for example, road functional class and speed limit (Ahie et al., 2015). External factors refer to different traffic conditions and weather conditions that are present during travel, e.g., precipitation and snow (Ewing et al., 2001; Small, 2012). Existing commercial navigation applications, e.g., Google Maps, provide support for accurate travel time prediction (Derrow-Pinion et al., 2021). After the user O-D is determined, the shortest route between O-D is computed by an A-star (A*) algorithm (Mehta et al., 2019) and the average speed for driving the O-D distance is computed and applied. Speed limits, historical average speed data, actual travel times from previous trips, and real-time traffic information also factor into travel time predictions (Epstein, 2013; Ireland, 2011). However, these data-driven methods do not answer the question of why drivers end up with different travel times even for the same trip. In this study, we analyze the

different factors that contribute to travel time and how these factors vary in their impact with respect to travel time for different trip lengths and purposes.

One challenge in analyzing travel time is to obtain reliable travel time data. Traditional approaches for collecting speed and travel time data include loop detectors, automatic vehicle identification sensors (i.e., automatic plate recognition), probe car data, and travel surveys (Kazagli & Koutsopoulos, 2013). Federal and State departments of transportation primarily capture traffic data from fixed sensors (e.g., loop detectors and automatic vehicle identification sensors), which continuously record data once installed on the roads. However, fixed sensors are relatively expensive to install and maintain, which can limit deploying these sensors on road networks (Hunter et al., 2009; Patire et al., 2015). Traditional travel surveys (e.g., face-to-face interviews, mail-out/mail-back, and telephone surveys) are designed questionnaires that provide transport habits and preferences (Dissanayake & Morikawa, 2010; Z. Li et al., 2020). Traditional travel surveys are known for data quality issues, e.g., under-reporting 20%-30% of trips and for respondents providing inaccurate details of their travel (L. Li et al., 2020; Shen & Stopher, 2014; Z. Wang, He, et al., 2018). Global Positioning System (GPS) data, collected from vehicles or smartphones, provide large volumes of data with rich information and a better way to understand human mobility than both coarse and fine-grained sensor data (Nasri et al., 2019; G. Zhu et al., 2021). GPS trajectories are used in many applications, e.g., inferring transportation modes and purposes (Lu & Zhang, 2015; Xiao et al., 2015; X. Yang et al., 2018), estimating near-real-time road speed (J. Yu et al., 2020; J. J. Q. Yu & Gu, 2019; P. Zhang et al., 2023), and estimating vehicle miles traveled (Blei et al., 2015; Fan et al., 2019; Henao & Marshall, 2019).

GPS data does have some limitations as data sampling frequencies can be unstable and sometimes are lower, e.g., 1 GPS waypoint per several minutes (Bezcioglu et al., 2022; Chen et al., 2019), and inferring the true path between two consecutive GPS points is needed (Rahmani & Koutsopoulos, 2012; S. Sun et al., 2019).

ML methods have been applied for predicting travel time, e.g., Support Vector Machine (Idé & Kato, 2009; Wu et al., 2004), Random Forest (Mendes-Moreira et al., 2012), Artificial Neural Network (Xu et al., 2019), Graph Convolutional Neural Network (Jin et al., 2021), and Gradient Boosting Regression Tree (X. Li & Bai, 2017). This research mainly focuses on improving the accuracy of travel time estimation by applying pre-processing techniques or trying to derive more characteristics and connectivity from the mobility data itself. For example, spatiotemporal relevancy, i.e., the travel time of a target road segment compared to a previously traveled segment or nearby relevant segments, is one factor that has been considered (Xu et al., 2019). Applying other data-processing techniques to trajectory data, e.g., feature selection (i.e., selecting only features that are important to the prediction models) and domain value definition (e.g., categorizing road segment classes based on speed limits), has also been shown to improve travel time prediction accuracy (Mendes-Moreira et al., 2012), and these are factors that we have used in the analysis presented here.

A key objective of this study is to identify different factors that may impact travel time and result in variations in travel time even for the same route. In this analysis, factors were categorized into three categories, driver behavior factors, built environment and road network factors, and external factors. In terms of driver behavior category, driver characteristics have not been widely investigated as these factors are

often hard to obtain for data privacy reasons. Socioeconomic characteristics of drivers (e.g., gender, age, income, and education level) have been collected in travel survey studies (R. M. González et al., 2015; Jang & Ko, 2019), where it has been found that the relationship between sociodemographic characteristics and individual behaviors/attitudes is a complex function. Other drive behavior factors, e.g., departure time, day of week, month, weekday/weekend, and public holidays, have been frequently analyzed (Lenny et al., 1997; X. Li & Bai, 2017; Mendes-Moreira et al., 2012). The built environment and road network category, including road network characteristics (e.g., speed limit, functional class, and intersections), and the external factor category, including weather conditions (e.g., season, wind speed, temperature, and precipitation) and even points of interest (POIs), have been widely investigated in previous research (Jenelius & Koutsopoulos, 2013; Mendes-Moreira et al., 2012; Tang et al., 2016; Y. Wang et al., 2014). Except for the aforementioned factors, some novel factors, e.g., travel speed and driver route choice between O-D, are also considered for this research.

This study identifies factors that impact travel time in a heavily trafficked urban area, and examines differences in factor importance, revealing what factors are key to determining the driving travel time for a trip. Except for the commonly used factors that could be explicitly derived from the GPS waypoints data, we compute driver route choices and driver speeding behaviors using time-series clustering algorithms. We use a ML method to analyze how contributions of different categories of factors to the travel times for two different trip lengths (50 miles and 150 miles) and different trip purposes (daily commuters and vacationers) change.

4.3 Materials and methods

4.3.1 Study area and data

To obtain driving times, this study used raw GPS vehicle trajectory data collected by INRIX[10] and made available for this research by the Center for Advanced Transportation Technology Laboratory (CATT Lab)[11] at the University of Maryland. The INRIX data include vehicle GPS waypoints captured using embedded GPS trackers and the INRIX mobile app providing, device ID, trip ID, longitude, latitude, time stamp, raw speed, and driving profile (i.e., types of vehicles completing the trips, including *consumer vehicles*, *field service/local delivery fleets*, and *for-hire/private trucking fleets*). The GPS sampling intervals (time between two consecutive waypoints) range from 1s to 53min, with an average of 36s. The original data volume for the complete INRIX dataset used in this analysis is large with over 9.6 billion GPS waypoints and 191 million trips in total from January 2018 to September 2019. The trips between the O-D were filtered from the raw dataset using Apache Spark, an open-source analytics engine for large-scale data processing on the Hadoop framework. The road network data is from OpenStreetMap[12] (OSM), a free and open geographic database. OSM road networks for the study areas were downloaded via the OSM API, including road segment length, and road functional classes (e.g., motorways, trunk, primary and secondary roads, tertiary and residential roads).

[10] https://inrix.com
[11] https://www.cattlab.umd.edu/
[12] https://www.openstreetmap.org

The travel times associated with driving on major roads in Maryland were analyzed. There are two pairs of O-D, from Towson to Silver Spring, MD (T-SS) and from Rockville to Ocean City, MD (R-OC) (Figure 4.1). Towson and Silver Spring are two highly populated cities in Maryland (close to Baltimore City and Washington D.C. respectively). Rockville is another city in Maryland, and Ocean City is a resort town, located on the Eastern Shore, on the Atlantic coast. The drive between T-SS represents an O-D trip of approximately 50 miles with highly frequent traffic volumes of commuters. R-OC represents an O-D trip of approximately 150 miles often undertaken by tourists and vacationers.

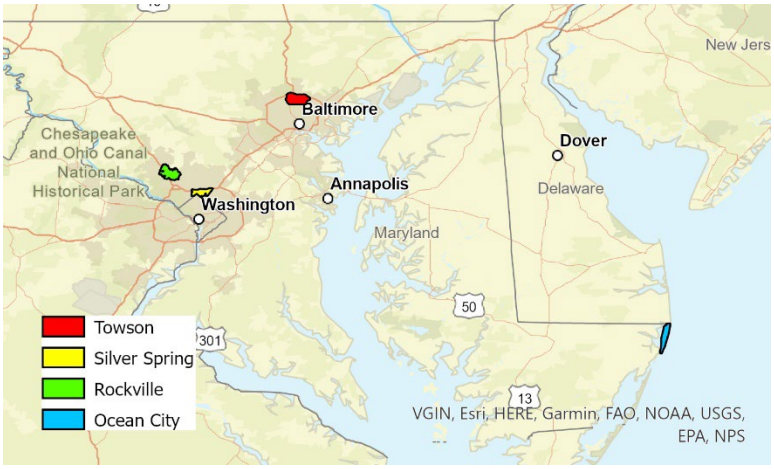

Figure 4.1 The study area of Maryland, U.S., and the locations of Towson, Silver Spring, Rockville, and Ocean City.

4.3.2 Map-matching GPS waypoints

Map-matching is the technique to obtain the full GPS waypoint trajectories, on which travel time analysis is based. Map-matching consists of two steps – first, projecting GPS waypoints to their corresponding road segments, and second, inferring and filling the road segment gaps between GPS waypoints. Traditional map-matching

algorithms use only the coordinates of GPS points, e.g., fuzzy logic, Extended Kalman Filter, Bayesian theory, etc. (Quddus et al., 2007). Advanced map-matching algorithms also consider additional travel characteristics, e.g., real-time moving direction and delay at intersections (Hsueh & Chen, 2018; Tang et al., 2016; C. Yang & Gidófalvi, 2018). This study used an original map-matching algorithm that was deployed using Apache Sedona, which was capable of dealing with million-scale big GPS data (P. Zhang et al., 2023). After complete GPS waypoint trajectories were computed, built environment and road network characteristics were calculated, for example, the percentages of highway and residential road usage along the trip, for later analysis.

4.3.3 Travel time impacting factors

The factors that were considered to be impacting travel time were identified and classified into three categories: driver behavior, built environment and road network, and external, for further analysis.

Driver behavior factors

Driver behavior factors refer to driver characteristics and individual driving behaviors. They were calculated based on map-matched GPS waypoints and trips and included departure time of day, day of week, month, driver choices of routes, speed, and driving profile (i.e., types of vehicles completing the trips, provided in the INRIX data). Departure time of day, day of the week, and month are factors that have been considered in studies of travel time prediction (X. Li & Bai, 2017). Different routes taken to reach the same destination may impact travel times and are considered in this study. Given designated travel origins and destinations, drivers may have preferred routes or choose to follow routes suggested by navigation apps. In this study, we used

the *dynamic driving directions* compared to the *overall driving direction* for the entire trip to represent and classify different route choices (Figure 4.2). The overall driving direction was represented using the start and end points of the map-matched GPS trajectory (Figure 4.2a). For each trip, the direction of every road segment (Figure 4.2b) was compared with the overall direction, and the angle between them was calculated (Figure 4.2c). After computing the angles, each trip was converted to a series of dynamic driving directions (i.e., angles), representing each driving route. Since trips have different trip distances and numbers of road segments, the set of dynamic driving directions was normalized against the percentage of trip length that the vehicles completed. The driving directions were later clustered using K-means time-series clustering (Giordano et al., 2021), and using the Elbow method to determine the optimal number of clusters, with the returned clusters representing different route choices (Makwana et al., 2013; Syakur et al., 2018).

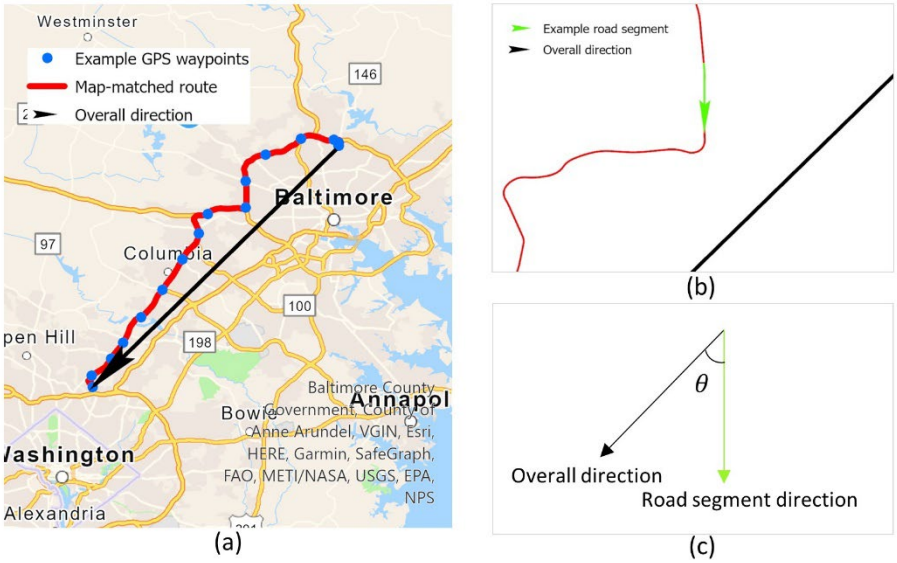

Figure 4.2 Dynamic driving direction compared to the overall driving direction for a trip. (a) An example map-matched trip. (b) An example road segment. (c) The angle between a road segment direction and the overall driving direction.

The travel speed, especially the low travel speed, could reflect both driver behaviors (e.g., stopping for gas/food/etc.) and traffic congestion situations. When the GPS waypoints were collected, their raw speed information was recorded by INRIX. Since there are missing raw speed values in the INRIX data, the missing values were estimated using their neighboring consecutive GPS waypoints. After estimating the missing speed data, trips that still had over 20% missing values were filtered from the dataset. Since trips had different trip distances and numbers of road segments, the series of travel speeds for a trip were normalized against the percentage of travel time the vehicle has completed.

Built environment and road network factors

A second category of factors refers to characteristics of the road network and built environment, including the number of road segments in a trip, the percent of different classes of road segments, the number of changes in road classes, trip distance, and average driving direction. We manually classified and assigned values to the road functional classes based on the importance given by OSM. *Motorways* and *trunk* were assigned as class 1; *primary* and *secondary* were assigned as class 2; *tertiary* and *unclassified* were assigned as class 3; *residential* and *living street* were assigned as class 4. The percent values for the four classes of road segments that vehicles traversed were calculated. When vehicles entered road segments that were different from the previous segments, driving speed also typically changed (e.g., speed will increase when a driver enters a highway from the ramp). The number of changes in road classes was computed and the average driving direction was calculated as the mean of the series of *dynamic driving directions* for a trip.

External factors

The third category of factors refers to properties of the environment that may impact a traveler, including traffic and weather conditions. Traffic incidents are considered to impact overall travel time, and incident data was collected from the CATT Lab, including different types of incidents and events (e.g., accidents, collisions, and road work) and the associated incident durations. Weather data were collected from the Global Historical Climatology Network (GHCN) from the National Oceanic and Atmospheric Administration (NOAA)[13], including daily temperature, wind speed, precipitation, and snow. Holidays are considered to be an external factor that could impact travel time, such as Independence Day (July 4th) and Christmas Day (December 25th). We manually classified the trips into holidays and non-holidays based on the departure date and time. The set of impact factors was used as the explanatory variables that were input to the ML models for determining factor importance (Table 4.1).

[13] https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily

Table 4.1 Factors that may impact travel times. Also the explanatory variables used in ML models.

Category	Factors	Data source
Driver behavior	Time of day	INRIX and
	Day of week	OSM
	Month	
	Route choices	
	Travel speed	
	Driving profile	
Built environment and road network	Number of road segments	INRIX and
	Percent of different classes of roads	OSM
	Changes of road classes	
	Trip distance	
	Average driving direction	
External	Incident duration	CATT Lab and
	Temperature	NOAA GHCN
	Wind speed	
	Precipitation	
	Snow	
	Holiday	

4.3.4 Examining factor importance using ML model

After the data were collected and processed, a Random Forest model was used to incorporate and identify contributing factors that may impact travel time including driver behavior factors, built environment and road network factors, as well as external factors for the two O-Ds T-SS and R-OC separately. A training data set comprising 75% of the data was used to develop the Random Forest model, with the 5-fold cross-validation, while the rest 25% of the data was used to assess model performance (other train-test splits, e.g., 80%-20% and 70%-30%, were also tested), which was evaluated using the coefficient of determination (R^2) and root mean square error ($RMSE$). Parameter tuning techniques, as well as other data processing techniques, e.g., cursive feature elimination (RFE), adding cost-complexity pruning (CCP) for regulation, etc., were applied to achieve better model performance. Finally, factor importance was

returned by the Random Forest models and the ranked list of those factors for travel time for each trip scenario was analyzed.

4.4 Results

4.4.1 Route choice and its impact on travel time

After processing, we obtained 9,103 trips for T-SS and 362 trips for R-OC from January 2018 to September 2019. The travel time for T-SS ($M = 55.27\ minutes$, $SD = 16.78\ minutes$) and R-OC ($M = 189.23\ minutes$, $SD = 25.61\ minutes$) were also computed using the GPS waypoint data.

The different choices of routes by drivers may impact the travel time, and to investigate this, we applied the K-means time-series clustering method to the dynamic driving directions. The Elbow method suggested the optimal number of clusters for T-SS trips should be five. The barycenters of clusters representing the average of each cluster confirmed that the five clusters showed different temporal trends (Figure 4.3a) and the differences among routes could also be recognized from the maps of routes (Figure 4.3c). The ANOVA test ($p = 2.77 \times 10^{-84}$) confirmed the significantly different travel times associated with different route choices between T-SS trips (Figure 4.3b).

The average travel times for the five clusters were $T_{Cluster\ 1} = 56.84\ minutes$, $T_{Cluster\ 2} = 60.87\ minutes$, $T_{Cluster\ 3} = 53.30\ minutes$, $T_{Cluster\ 4} = 51.32\ minutes$, and $T_{Cluster\ 5} = 59.11\ minutes$. Cluster 1 drivers used Interstate-83 (I-83) to travel across Baltimore City, instead of using the Baltimore Beltway (i.e., I-695). Cluster 2 drivers used the east I-695, which included the Baltimore Harbor Tunnel

and Fort McHenry Tunnel. The slowdown at the tunnels could be one possible reason for Cluster 2 having the longest average travel times among the five clusters. Cluster 3 drivers used Exit 16 on I-695 to enter U.S. Route-29, instead of exiting to I-95 using Exit 11 on I-695. Cluster 4 drivers used the west I-695 and I-95, which corresponds to the default route suggested by navigation applications and had the shortest travel time of all clusters. Cluster 5 drivers used "corridor" routes that connected I-95 with other highways, e.g., Route-100, Route-32, and Route-200. Using these "corridor" routes to switch between highways may be the reason why Cluster 5 had the second longest travel times.

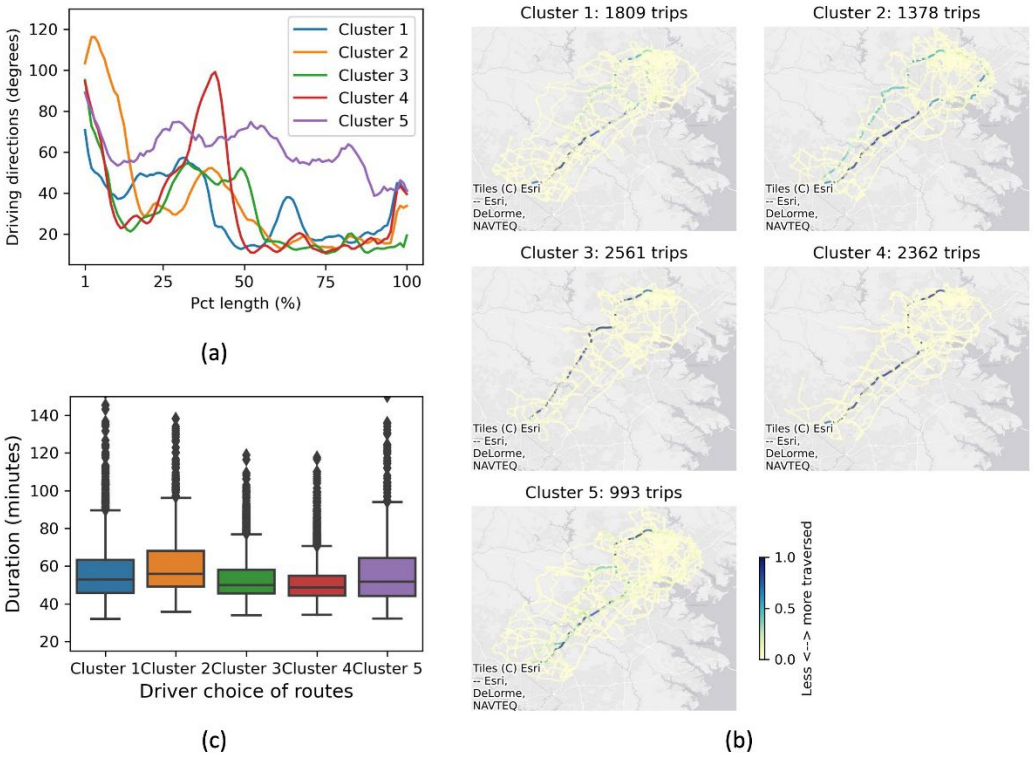

Figure 4.3 Route clustering to determine route choices for Towson – Silver Spring trips. (a) The barycenters of the five clusters of route choices. (b) The maps of the five routes. (c) Box plot of travel time by the five clusters.

The Elbow method suggested three clusters for R-OC trips. The barycenters of each cluster showed that for the first 40% of the trip, Cluster 1 was comprised of different routes to Cluster 2 and Cluster 3, while for the remaining 60% of the trip, Cluster 3 was different from Cluster 1 and Cluster 2 (Figure 4.4a). The ANOVA test ($p = 0.0139$) confirmed the travel times associated with different route choices between R-OC were significantly different (Figure 4.4c). The average travel times for the three clusters were $T_{Cluster\ 1} = 199.15\ minutes$, $T_{Cluster\ 2} = 187.71\ minutes$, and $T_{Cluster\ 3} = 187.65\ minutes$ separately. Cluster 1 drivers mainly used Route-29 and Cluster 2 mainly used I-495 before they drove over the Bay Bridge, located after approximately 40% of the trip lengths had been driven. After going over the bridge, Cluster 3 drivers chose a different route that went through the State of Delaware, compared to Cluster 1 and Cluster 2 drivers who stayed in Maryland for their entire trips (Figure 4.4b). The roads taken before the bridge contributed much to the travel time differences (about 12 minutes), as the road network density was higher and drivers had more choices compared to after the bridge where most of the driving was on the same U.S. Route-50 highway.

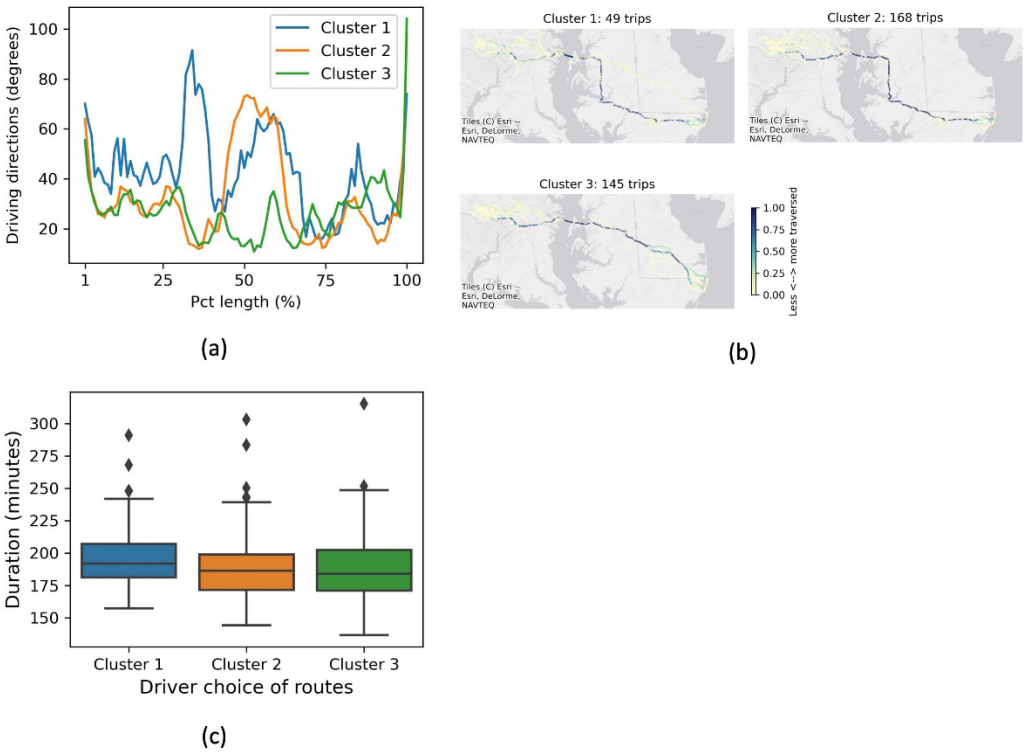

Figure 4.4 Route clustering to determine route choices for Rockville – Ocean City trips. (a) The barycenters of the three clusters of route choices. (b) The maps of the three routes. (c) Box plot of travel time by the three clusters.

4.4.2 Driver travel speed behavior and its impact on travel time

The different travel speed behaviors by drivers may impact travel time. Different patterns could be identified in the histogram of travel speeds for T-SS (Figure 4.5a). Over 70% of T-SS trips were at speeds of 60-76mph during 50%-75% along the trip, while the rest 30% trips were at slower speeds of 30-40mph at the same point along the trip. For R-OC, we could see a clear low-speed period at about the ½ way point of the trip (Figure 4.5b). The slowdown happened at Easton, MD, a popular small town on the Maryland Eastern Shore that is about halfway in the trip. Drivers may stop at this town for multiple reasons, such as food, gas, and rest since R-OC was a relatively long trip of about 150 miles.

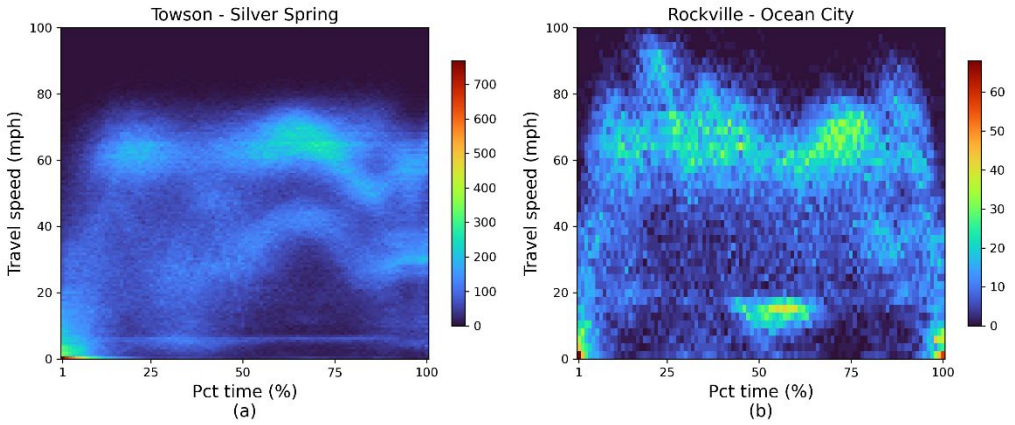

Figure 4.5 Histograms of travel speed (miles per hour) normalized by percentage of time that vehicles completed for (a) Towson – Silver Spring and (b) Rockville – Ocean City.

We applied K-means time-series clustering method on the raw speed data associated with the INRIX GPS waypoints to investigate the relationship between driver travel speed behaviors and travel times. The Elbow method results returned three clusters for T-SS trips (Figure 4.6a). The ANOVA test ($p = 1.42 \times 10^{-229}$) confirmed the travel times associated with different travel speed behaviors were significantly different for T-SS trips (Figure 4.6c). The average travel times for the three clusters were $T_{Cluster\,1} = 61.61\,minutes$, $T_{Cluster\,2} = 48.21\,minutes$, and $T_{Cluster\,3} = 55.16\,minutes$ separately. Cluster 1 used U.S. Route-29, instead of Interstate-95. The choice of Route-29 meant lower travel speeds than on I-95, which led to Cluster 1 having the longest travel time, 13.40 minutes that was 6.45 minutes longer than times of Cluster 2 and Cluster 3. Cluster 2 and Cluster 3 using the same route may represent trips either without or with traffic congestion (Figure 4.6b). Traffic congestion between T-SS resulted in 6.95 minutes of extra travel time on average.

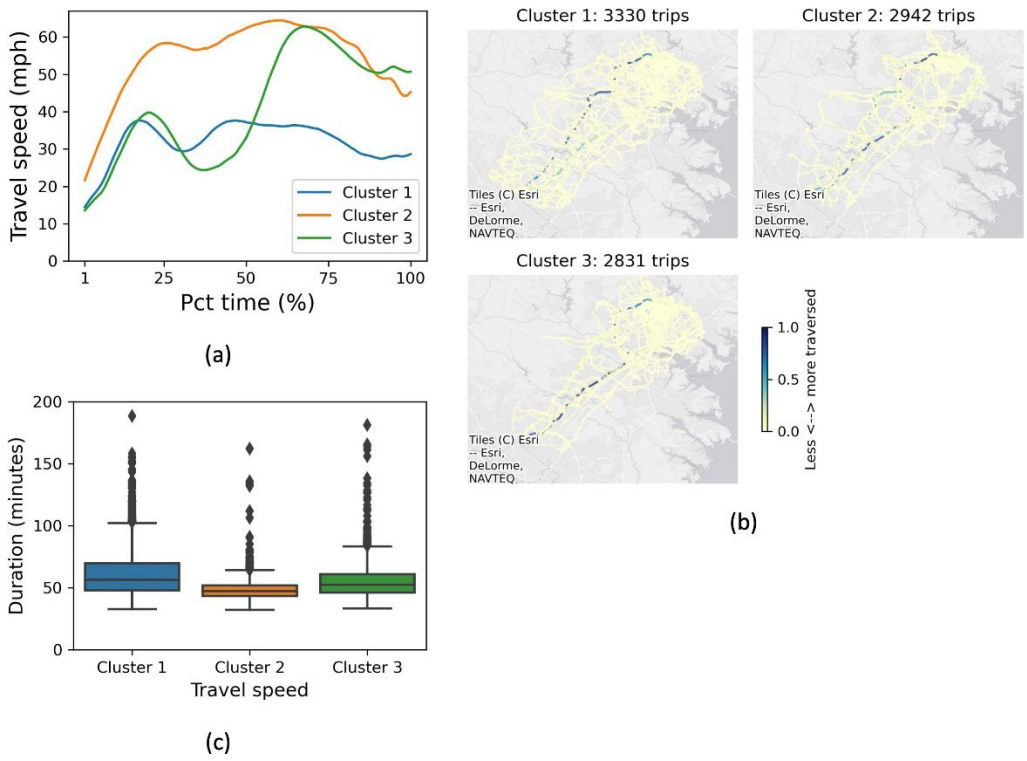

Figure 4.6 Travel speed clustering for Towson – Silver Spring trips. (a) The barycenters of the three clusters. (b) The maps of the three clusters. (c) Box plot of travel time by the three clusters.

The Elbow method results returned three clusters for R-OC trips (Figure 4.7a). The ANOVA test ($p = 1.19 \times 10^{-31}$) confirmed the travel times differences associated with different travel speed behaviors for R-OC were significant (Figure 4.6c). The average travel times for the three clusters were $T_{Cluster\ 1} = 215.97$ minutes, $T_{Cluster\ 2} = 182.25$ minutes, and $T_{Cluster\ 3} = 179.26$ minutes. Cluster 1 drivers tended to drive at 40-55 mph speed throughout the whole trip (I-95 has a speed limit of 55 mph and US Route-50 has a speed limit of 45 mph), leading to the longest travel times, approximately 33-37 minutes longer travel times than the other two clusters. Although Cluster 3 drivers had a slowdown at Easton, MD, they tended to drive at a relatively higher speed, around 70 mph during the trip, resulting in a similar travel time overall as Cluster 2 drivers, who did not stop along the trip (Figure 4.7b).

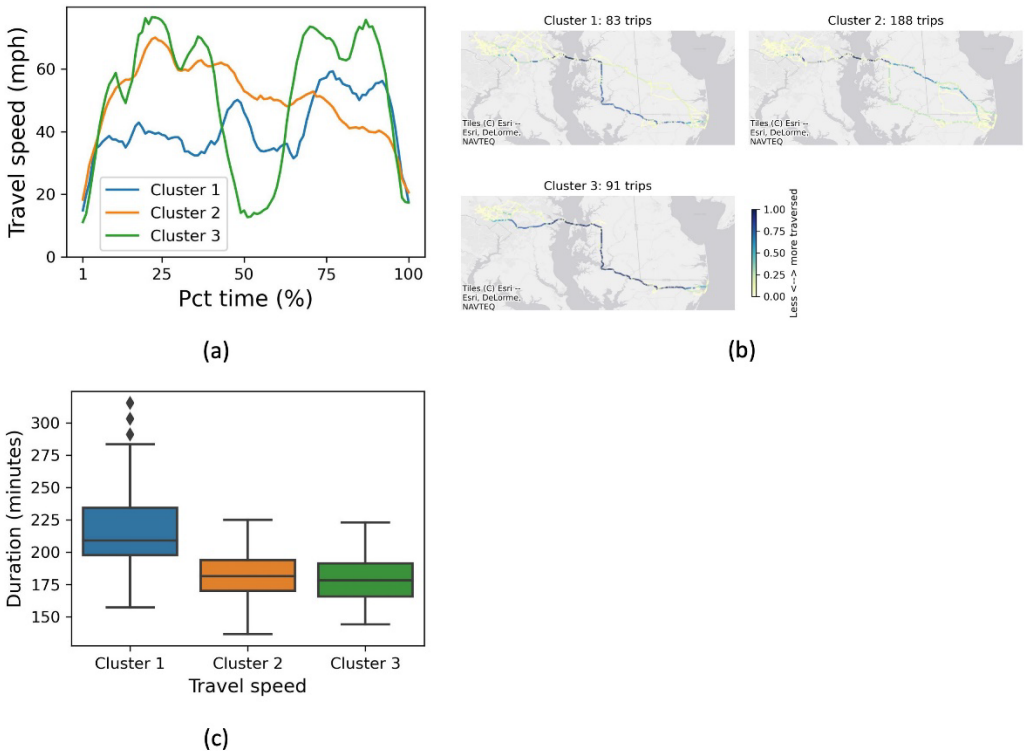

Figure 4.7 Travel speed clustering for Rockville – Ocean City trips. (a) The barycenters of the three clusters. (b) The maps of the three clusters. (c) Box plot of travel time by the three clusters.

4.4.3 Examining the characteristics of trips

The factors in three categories, driver behaviors, built environment and road network, and external factors, and their relationships with travel times were examined for both O-Ds.

With regard to driving profiles (Table 4.2), we found that R-OC trips had a higher percentage of *consumer vehicle* trips (77.07%) and a lower percentage of *field service/local delivery fleet* trips (21.27%) than T-SS, keeping with the case that T-SS trips were urban commuting trips while R-OC trips were likely mainly associated with recreation and vacations. The *for hire/private trucking fleet* trips comprised an extremely low percentage of the total trips for both O-Ds (3.39% and 1.66% respectively). We applied ANOVA tests to investigate the variances across mean travel

times of different driving profiles. The p-value results for T-SS (6.75×10^{-83}) and for R-OC (0.1105) indicated that the travel times for T-SS trips were associated with different driving profiles, while significant differences in travel times for R-OC trips were not detected over different driving profiles. Taking T-SS for example, the average travel times for *field service/local delivery fleet* trips were 60.35 minutes, 7.42 minutes longer than that of *consumer vehicle* trips.

Table 4.2 Numbers of trips by driving profiles for the INRIX data.

Driving profile	Towson – Silver Spring		Rockville – Ocean City	
	Number of trips	Percent	Number of trips	Percent
Consumer vehicles	5,976	65.65%	279	77.07%
Field service/local delivery fleets	2,769	30.42%	77	21.27%
For hire/private trucking fleets	358	3.93%	6	1.66%
Total	9,103		362	

The total numbers of trips and average travel times for driver behavior factors, including departure time of day, day of the week, and month, were investigated for both O-Ds (Figure 4.8). For T-SS, the numbers of trips showed that generally daytime trips, from 10 am to 4 pm, were higher and travel times showed two main peaks, one around 8 am and the other around 5 pm, slightly after the two peaks in number of trips (Figure 4.8a). The departure time of day for T-SS trips was associated with greater variance in travel times, from 50 minutes to over 200 minutes. T-SS had more trips as well as much longer travel times on weekdays than on weekends (Figure 4.8b), reflecting the heavily trafficked commuting patterns that are experienced during weekdays in the Washington DC-Baltimore area (Agarwal, 2004). T-SS showed fewer trips during October-December, but the travel times in these months were relatively longer (Figure 4.8c), which may be related to the holidays held during this time period. More holidays

possibly led to fewer work-commuting trips, but the roads were packed with more holiday-travelers, which resulted in traffic congestion and longer travel times.

R-OC trips, on the other hand, were distributed throughout the daytime with some very early morning peaks (prior to 8 am), showing that some drivers departed early perhaps so that they could arrive at Ocean City by or in the afternoon (Figure 4.8d). However, travel times did not fluctuate very much during the day, suggesting that the departure time did not play an important role in determining the total travel time. A significantly larger number of R-OC trips on Fridays could be seen, together with longer travel times on Fridays (Figure 4.8e), likely due to travelers driving to Ocean City on Fridays for weekend trips and recreation. Over 73% of R-OC trips were made during May-September and the travel times were relatively longer, especially in June (Figure 4.8f), matching the time of the year when better weather and vacations to oceans/beaches increase the number of trips to coastal destinations like Ocean City, MD.

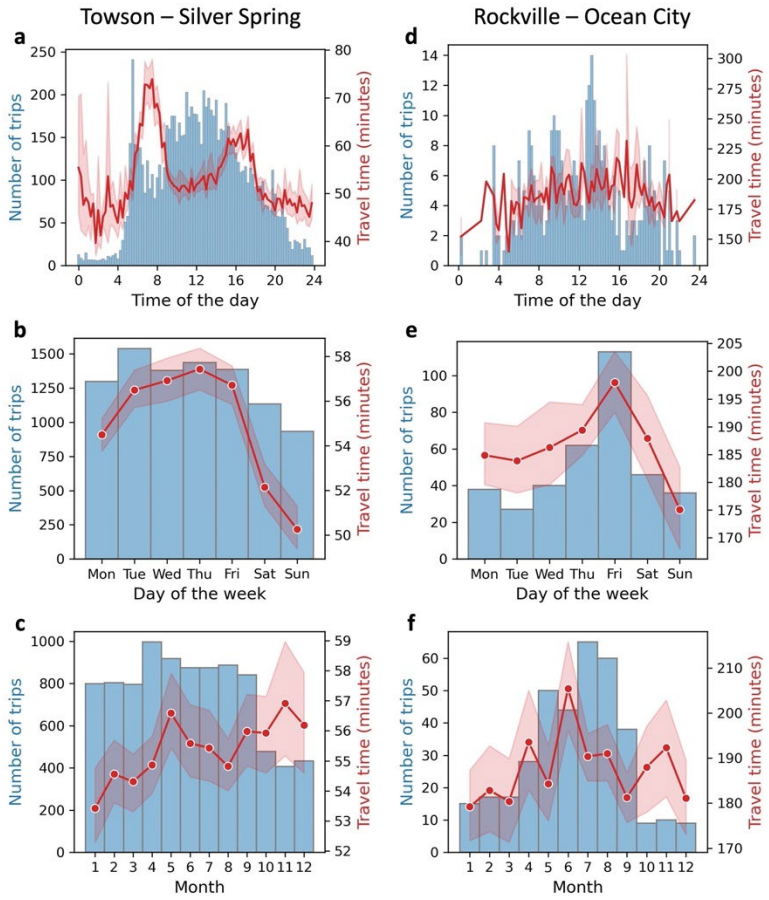

Figure 4.8 Data distributions for driver behavior factors: time of the day, day of the week, and month. Blue bars represent the total number of trips. Red lines represent average travel times (with 95% confidence interval).

The impacts on travel time from external factors, including incident duration, precipitation, snow, wind speed, temperature, and holidays, were also investigated (Figure 4.9). There were many incidents reported for T-SS routes, mostly with incident durations of between 0-20 minutes and sometimes over 40 minutes, which significantly lengthened the overall travel time for T-SS trips (Figure 4.9a). We also found that there were trips with shorter accident times (e.g., 10 minutes), but still long travel times (e.g., 100 minutes), perhaps indicating that once traffic backs up due to an incident, it takes some time for traffic flow to be reestablished (US Department of Transportation, 2020a). Weather-related factors, precipitation (Figure 4.9b), snow (Figure 4.9c), wind

speed (Figure 4.9d), and temperature (Figure 4.9e), had non-linear relationships with travel time. During higher precipitation/snow events, travel times still varied from 50 minutes to 90 minutes, suggesting that traffic was not always impacted by weather factors. The nature of work commuting between T-SS decided there were fewer trips on public holidays, associated with about 6 minutes of shorter travel times (Figure 4.9f). This finding seemed to be the opposite of what we observed in Figure 4.8c that during October-December, when there were more holidays, the travel times for T-SS were even longer. This was possibly due to our definition of "public holidays", where only the public holiday observation days were counted. For example, the 2018 Thanksgiving Day was on Thursday, November 22, 2018. However, the actual Thanksgiving break was from Wednesday, November 21, 2018, to Sunday, November 25, 2018. There were more trips happening before and after the actual holiday observation date.

For R-OC trips, there were few incidents recorded for the study period and the incident durations (2-12 minutes) were shorter than for T-SS incidents showing that total travel times were not adversely impacted by incidents (Figure 4.9g). The relationships between travel times and precipitation (Figure 4.9h), snow (Figure 4.9i), and wind speed (Figure 4.9j) were also non-linear. There was a clear segmentation that more trips happened when the temperature was relatively high, over 60 degrees Fahrenheit (Figure 4.9k). This finding corresponded with the patterns we saw in Figure 4.8f that most R-OC trips happened in summertime, in which temperatures were higher. In contrast to T-SS, R-OC had more trips on holidays with shorter travel times (Figure 4.9l) as people tended to visit recreation/vacation destinations more during holidays. The possible reasons for shorter travel times on holidays were similar to T-SS in that

when we collected the holiday data, only the date when the holidays were observed was counted for this factor.

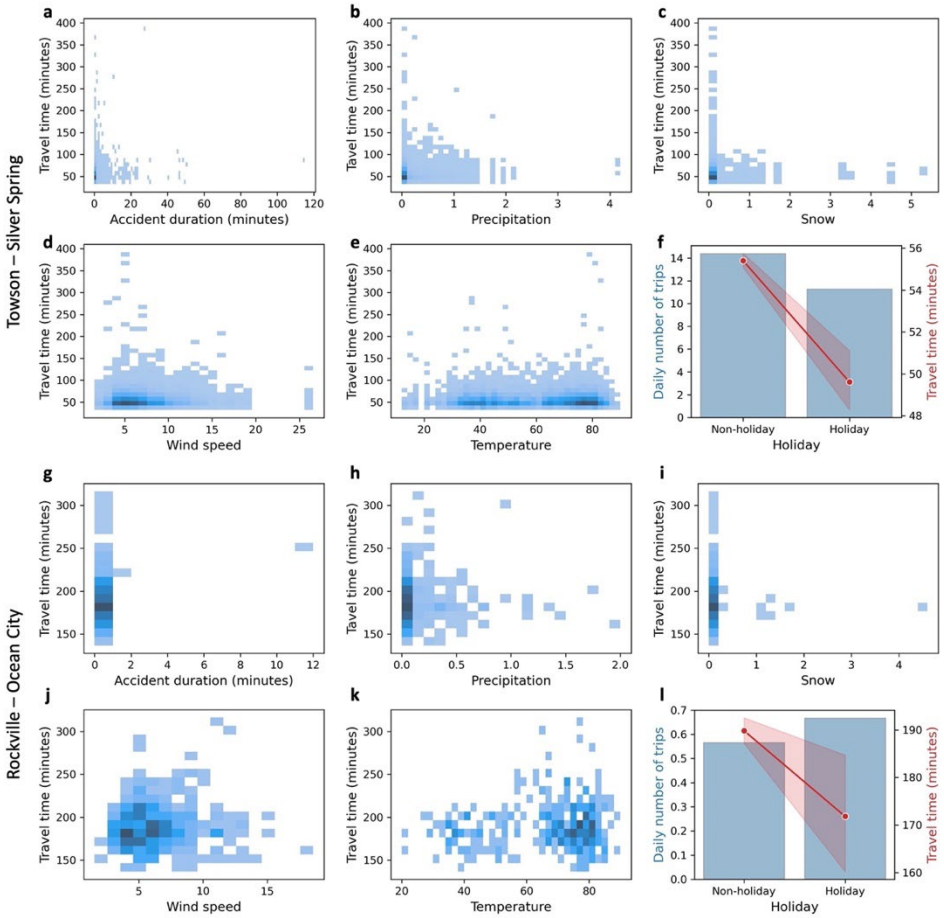

Figure 4.9 Travel time by external factors: incident duration, precipitation, snow, wind speed, temperature, and holidays. (a)-(f) for Towson – Silver Spring and (g)-(l) for Rockville – Ocean City.

4.4.4 Random Forest model results

We applied the Random Forest model with parameter tuning using the collected driver behavior, built environment and road network, and external factors, to examine how these factors impact travel time and how the importance of factors changes with different trip length/purpose (i.e., T-SS and R-OC trips). After testing different train-test split ratios, 70%-30% was used as it returned the best model performance (Table

4.3). The model performance of T-SS was better than R-OC, possibly because a low number of records (362 recorded trips for R-OC) restricted the performance of Random Forest models when building decision trees (Han et al., 2021).

Table 4.3 Random Forest model performance.

	Towson – Silver Spring	Rockville – Ocean City
R^2	0.5357	0.4576
$RMSE$	11.62	19.94

The factors importance rankings represented how importantly explanatory factors contributed to travel time for T-SS (Figure 4.10a). The number of road segments factor and the percent of class 1 roads factor were ranked top for T-SS. The urban road network between T-SS decided that the roads drivers traveled were mostly higher speed limit roads, e.g., motorways and primary roads. The more roads with higher speed limits drivers used, the shorter travel times were. The travel speed behavior factor was ranked 3rd, which was discussed in Section 3.2. The roads with lower speed limits and the traffic congestion conditions both played an important role in the urban work-commuting trips. The driving profile factor was ranked 4th, corresponding with the ANOVA analysis in Section 3.3 that the T-SS travel times were significantly different among different driving profiles. As discussed before, the departure time of the day was associated with different travel times, which was ranked 6th. Driver route choices were ranked at 18th, suggesting different route choices although were confirmed to be associated with different travel times, actually did not greatly impact travel times.

The factor importance rankings for R-OC were also plotted to examine the differences of these factors between T-SS and R-OC, representing different trip length/purpose (Figure 4.10b). The travel speed behaviors were significantly important

for R-OC trips, outranging all other factors, suggesting that for a long-distance trip, the travel speed drivers took was the determinant factor for the overall travel time. For R-OC, the trips where drivers consistently drove at 40-55 mph (speed limit of 45/55 mph) tended to be 33-37 minutes longer than other trips, some of which even made a stop at Easton, MD. Departure time of the day was ranked at 4th, which was similarly important for travel time as for T-SS trips (ranked at 6th). As there were more local roads (with lower speed limits) on the eastern shore of Maryland, the percent of class 4 roads factor made to the 6th of the list. The driving profile factor was not relatively important (ranked 17th), corresponding to the ANOVA analysis results in Section 3.2 that the driving profiles did not significantly differentiate travel times for R-OC. The incident duration factor was ranked at the bottom for R-OC as there were few incidents recorded and the incident durations (2-12 minutes) were disproportionate to the average R-OC travel time (189 minutes).

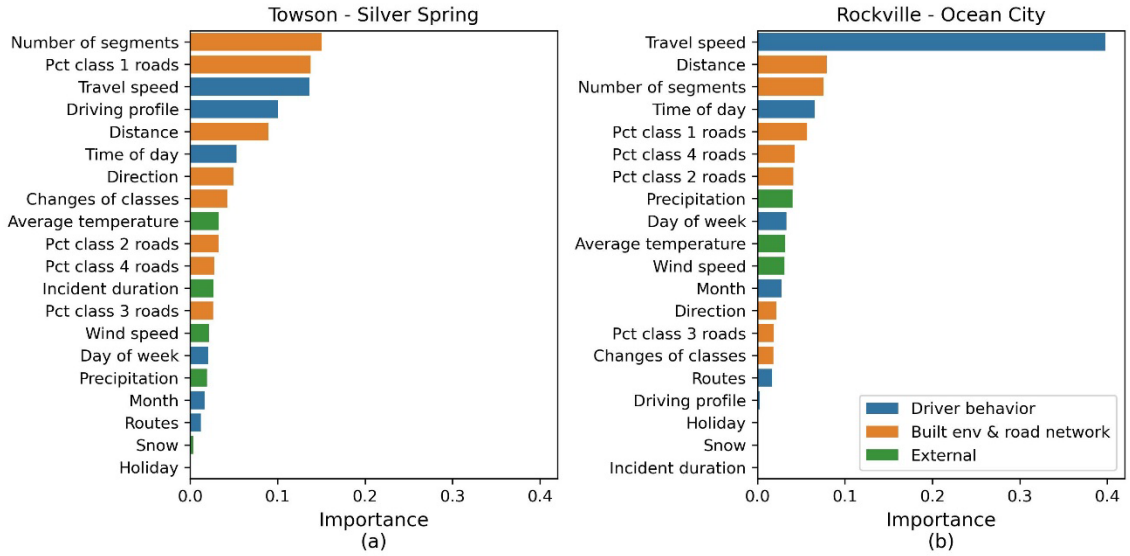

Figure 4.10 Random Forest model importance rankings for the explanatory factors by categories for (a) Towson – Silver Spring and (b) Rockville – Ocean City.

Comparing the two O-Ds, a common conclusion was recognized that the category of built environment and road network factors was the most important for both O-Ds. There were more built environment and road network factors ranking relatively high than driver behavior factors and external factors. This finding was consistent and robust among Random Forest models which we trained with different tuning parameters. When comparing differences between the two O-Ds, we found that the built environment and road network factors (e.g., percent of class 2 roads and percent of class 4 roads) and external factors (e.g., wind speed, precipitation, and temperature) for R-OC were more important than for T-SS, indicating driving on local roads (lower speed limit) and the weather conditions were the factors drivers considered more when traveling for long-distance trips and for vacations and recreations.

4.5 Discussion

This study investigated factors that might impact travel time, including driver behavior factors, built environment and road network factors, and external factors, using the Random Forest model. Two representative trip O-Ds in Maryland, U.S., T-SS (urban commuting driving, about 50 miles) and R-OC (trips to vacation or recreational destination, about 150 miles), were selected. GPS waypoint data were processed and map-matched from INRIX data and OSM road network data. The K-means time-series clustering algorithm was applied to investigate different driver route choices and driver travel speed behaviors. Finally, all factors from the three aforementioned categories were trained using Random Forest models to understand what the key factors were for each O-D and how the factor importance changed over different trip length/purpose.

4.5.1 Driver route choice and travel speed behaviors and their impacts on travel times

We computed two factors that were understudied in previous research, driver route choices and driver speed behaviors, using a time-series clustering method. Although the Random Forest showed different route choices did not greatly contribute to the overall travel times as other factors, the clustering analysis showed that drivers did use different routes and the ANOVA test confirmed that depending on which route drivers chose, their travel time was different. This leaves an interesting topic for future research, what would be dominant when the results from two ML methods do not align with each other.

The driver travel speed factor was used to represent the traffic congestion situations (US Department of Transportation, 2020b). Both the clustering analysis and Random Forest models suggested the significant impact of travel speed behaviors on travel times. For example, Cluster 2 and Cluster 3 for T-SS represented trips using the same route but without and with traffic congestion separately, which had a difference of 6.95 minutes in travel times. Random Forest models confirmed the important role of the travel speed behavior factor for both O-Ds (ranked at 3rd and 1st respectively). Especially for R-OC (long-distance trip and trip for vacation/recreation), the travel speed behavior factor was more important than all other factors combined.

4.5.2 Trip length and trip purpose

The two O-Ds representing different trip lengths and trip purposes were examined in terms of impacting factors by Random Forest models. Except for the common pattern that the category of built environment and road network characteristics

was generally more important than the other two categories, when comparing the two O-Ds, differences between trip lengths and trip purposes could be found.

The clustering analysis suggested that long-distance trips had fewer route choices than short-distance trips. The T-SS trips showed more variance in terms of the routes drivers could choose from. R-OC trips had limited variance before the drivers passed the Bay Bridge, and after passing the bridge, drivers could only choose either U.S. Route-50 or Delaware-404. In addition, the Bay Bridge was the only way for drivers to enter the eastern shore of Maryland. Drivers traveling for long-distance trips may follow the navigation apps, while short-distance trip drivers may be more familiar with the roads and may have their own preferences, which possibly contributed to this finding. There was another difference in the behaviors between long- and short-distance trip drivers, which was stopping along the way. About 25% of R-OC drivers stopped at the small town, Easton, MD, for food/gas/rest/etc. purposes, as the whole trip was over 150 miles and over 3 hours.

With regard to the differences of factors by categories, we found that driving on local roads (lower speed limit) and the weather conditions were the factors drivers considered more when traveling for long-distance trips and for vacations and recreations. Our findings that time of the day factor (i.e., departure time) was ranked relatively high (2^{nd} and 4^{th} respectively) and day of the week factor and month factor were ranked relatively low for both trip purposes, corresponding with the current research findings that departure time is important and weekday/weekend and month are unimportant (Lenny et al., 1997; X. Li & Bai, 2017; Mendes-Moreira et al., 2012). The driver-related characteristics were usually unavailable in travel time-related analysis,

while in this study, the driving profile (what type of vehicle the drivers were using) was included in the analysis. Both the ANOVA analysis and Random Forest models suggested that different driving profiles contributed to differences in travel times for T-SS trips, not for R-OC trips. The nature of trips to vacation/recreation destinations decided that the majority of R-OC trips were *consumer vehicle* trips.

The built environment and road network factors were generally more important than driver behavior factors and external factors for both trip purposes, especially the road class factor. The class 1 roads (*motorways* and *trunk*) played an important role in travel times for both O-Ds, which corresponded with the research conclusion that the travel time of long distances is dominated by the time of "travel-free" sections (Wu et al., 2004). The weather-related external factors were ranked bottom among all factors for both O-Ds by the Random Forest model. This is likely because the Maryland climate was relatively mild, and severe weather was not experienced very much during this period. The incident factor was ranked at 12^{th} and 20^{th} respectively, which were the causes of traffic congestion and unreliable travel and could potentially lengthen travel times (US Department of Transportation, 2020b). This indicated the short-distance trips were more easily impacted by traffic incidents.

4.5.3 Possible data biases present in this research

The biases associated with the data must be acknowledged that the sampling rate of the INRIX dataset (i.e., ratio of recorded trips to the total number of trips actually happening) was about 2%. The INRIX data was passively-collected by mobile devices (cell phones or GPS devices installed on vehicles). Certain population groups might be underrepresented, leading to biased travel patterns for this research. Previous research

investigated multiple mobility datasets, e.g., Apple, Google, SafeGraph, and Descartes Lab, and their sampling rates at different spatial scales demonstrated higher levels of underrepresentation bias in certain population groups, e.g., Black, Hispanic, elder, and low-income, possibly due to lower smartphone and cell phone ownership rates (Abrar et al., 2023; Coston et al., 2021; Z. Li et al., 2023). Different demographic groups were associated with different mobility patterns. For example, the Black population was positively associated with mobility decrease and the median household income was associated with mobility recovery during the COVID-19 Pandemic (G. Zhu et al., 2021; G. Zhu & Stewart, 2023). Compared to the average sampling rate 7.5% for SafeGraph data, the INRIX data with only 2% of sampling rate needs further investigation on which demographic groups were less captured in the dataset and to what degree this selection bias influenced the mobility patterns derived from INRIX data.

4.6 Conclusions

Existing research, including on-market navigation applications (e.g., Google Maps), focuses more on providing accurate travel time prediction using advanced algorithms that take into account shortest path computations, while the focus of this study was to investigate what kinds of factors from driver behaviors, to built environment and road network characteristics and also external factors (e.g., incidents and weather) impact travel time for different trip lengths and different trip purposes. The time-series clustering analysis showed that for each O-D pair, different route choices were being activated by drivers. The Random Forest models confirmed the consistent conclusion over the two O-Ds that factors related to built environment and road network characteristics were significantly influencing travel times in the urban

area of Maryland, as driving on roads with higher speed limits comprised large proportions of trips. Time of the day was another determinant factor for travel time, especially for T-SS, as people commute to work on a daily basis and are more sensitive to traffic congestion and rush hours. On the other hand, the different natures of T-SS and R-OC led to differences in factor importance. For example, T-SS experienced more incidents than R-OC, resulting in the incident duration factor more impacting T-SS travel times than R-OC travel times. The percent of class 4 roads factor was more important for R-OC, possibly due to R-OC trips traversing more local roads on the eastern shore of Maryland than T-SS trips. For long-distance trips and trips to vacation/recreation destinations, the driver travel speed played a determining role in travel times. Other factors, such as driver route choices, holiday, and weather conditions, did not greatly impact the overall travel times.

This study only selected two popular O-Ds in Maryland, which is a mostly urbanized state. Driving conditions for other O-Ds and other suburban areas may not be the same as the two O-Ds in this study. Future work may scale up to the whole road network in an area to investigate the travel time of any trip made within the area, which may eliminate bias as much as possible. The holiday factor could be expanded to the neighboring dates of holidays when more trips happened. Additional data sources with socioeconomic information of drivers and better quality (i.e., less missing data) could help bridge the gap and provide more insights into how these factors generally impact travel time. The two O-Ds, although representing different trip lengths and trip purposes, could not quantitatively differentiate the impacts of trip length and trip purpose. Future research could control the variance and select two O-Ds with the same

trip lengths but different trip purposes (or vice versa) to investigate the solely impact of trip length or trip purpose.

Chapter 5: Conclusions

5.1 Review of dissertation

This dissertation is composed of three studies that investigated the spatiotemporal trends of mobility patterns, specifically vehicle mobility (i.e., vehicle trip changes and travel behaviors), using passively-collected mobile device data in the U.S. These regional (state) and nationwide analyses for trips and driving behaviors were undertaken at different spatial scales, such as census tract and county level, and over different time periods. Different ML models (regression, classification, and clustering) were applied to interpret factors that could contribute to different patterns of mobility, as well as categories of factors, that may impact mobility in different contexts, for example, different stages of the COVID-19 Pandemic. What were the key factors for mobility patterns and how the key factors changed over space and time were investigated by these explanatory analyses using ML models. In addition to providing new insights into how different types of mobility patterns evolved over space and time especially during a major public health crisis, the results are useful for policy and planning implications for local and regional officials, e.g., mobility restriction measurements, decision support for economic recovery, and public health strategies. The integration of diverse data sources (e.g., passively-collected mobility data and other mobility data from different public and private sources) and the utilization of multiple ML models enhanced the interpretability of mobility patterns.

The first study of the dissertation (Chapter 2) examined mobility patterns (inflow trips per person) in Florida during the early COVID-19 Pandemic, when the number of COVID-19 cases started to rise in summer 2020. Three counties in Florida

were selected as the case study area. The temporal changes in the mobility patterns were identified that the number of trips made per person significantly declined in all three counites when new COVID-19 case numbers began to rise in mid-June. An explanatory analysis using Random Forest models to detect the key factors that may impact mobility patterns applied a set of over 30 factors, including sociodemographic, travel-related, and built environment factors, as well as COVID-19 case data. Some of the factors, e.g., distance to beach and bar/restaurant visits, were specially examined for the tri-county region. The Random Forest models revealed differences among counties with respect to factors that contributed to mobility patterns prior to and after the spike in COVID-19 cases. The results not only informed the Florida local health officials the spatiotemporal insights into mobility patterns among the three counties, as the counties were hard hit by COVID-19 infections, but also inspired the stakeholders to pay more attention to certain sociodemographic groups when making mobility-restriction policies.

The second study of the dissertation (Chapter 3) investigated the nationwide mobility patterns (i.e., relative trip changes) at county level in the U.S. in the context of ongoing COVID-19 vaccines. A mobility index representing the relative trip changes at county level was calculated using multiple mobility data sources, including Apple Mobility Trends Report and FHWA Traffic Volume Data, to obtain more county coverage. The patterns of mobility index and vaccination rates were examined for the first five months of 2021. The K-means multivariate time-series clustering algorithm was applied on both mobility and vaccination rates to examine the spatial and temporal distributions of counties with similar mobility-vaccination profiles and returned five

clusters of counties across the U.S. The characteristics of clusters were examined with respect to gender, age, race and ethnicity, economic, education, and work to better understand the drivers for different clusters using ANOVA tests. Results suggested higher education and household income impacted counties' mobility-vaccination profile. The in-depth trip purposes of clusters were also examined for selected counties and higher trips to retail/recreation and workplaces were found to be associated with rural counties with relatively lower mobility-vaccination profile.

The third study of the dissertation (Chapter 4) examined the vehicle mobility behaviors, i.e., vehicle driving time, between two selected O-Ds in Maryland representing different trip purposes and different trip lengths during January 2018-September 2019 using passively-collected GPS waypoint data by INRIX to analyze how travel time varied among trips and the underlying factors for different travel times. Two factors that were understudied in previous research, driver route choice and different travel speed behaviors, were computed using a time-series clustering method, and their impacts on travel times were investigated. These two factors alongside factors from three additional categories, including driver behavior, built environment and road network, and external factors (e.g., incident and weather), were studied for the two O-D pairs using Random Forest models to analyze what were the key factors and whether these key factors changed for different trip purposes and trip lengths.

In this dissertation, I used various sources and types of mobility data. During the COVID-19 Pandemic, human mobility data provided important opportunities for both researchers and policymakers to understand how mobility was related to the spread of COVID-19, among other topics. However, in general, passively-collected

mobile device data is not as widely accessible for academic research as needed. To give some examples, the SafeGraph data in Chapter 2 was accessed via a data sharing agreement that the data could only be used for academic and research purposes and is no longer available from SafeGraph. The Apple Mobility Trends Report and Google Community Mobility Reports used in Chapter 3 were publicly available at the time of my analyses, however, were not updated after April 14, 2022 and October 15, 2022, respectively and while Google has made its historical data still publicly available, Apple data is now no longer provided. The FHWA mobility data used in Chapter 3 was publicly available through the U.S. Federal Government data sharing platform. The INRIX GPS waypoints data used in Chapter 4 was obtained from CATT Lab at the University of Maryland and requires an agreement with the data provider. These different levels of access pose certain disadvantages for mobility researchers. Greater data access would lead to an increased body of research that is more diverse, and would allow more research on mobility topics outside of the Pandemic. It is hoped that more collaboration between academia, mobility companies, and government agencies could be established to improve mobile device data access for academic research.

The biases associated with the mobility data used in these analyses must be acknowledged. Certain population groups might be underrepresented, leading to biased travel patterns for this research. Previous research has found that older and non-White populations were less likely to be captured by mobility data, particularly in the Pandemic context (Coston et al., 2021). The sampling rates of current popular mobility datasets, e.g., Apple, Google, SafeGraph, and Descartes Lab, were examined and demonstrated higher levels of underrepresentation bias in certain population groups,

e.g., Black, Hispanic, elder, and low-income, possibly due to lower smartphone and cell phone ownership rates (Abrar et al., 2023; Coston et al., 2021; Z. Li et al., 2023). The passively-collected mobility data for Chapter 2, which recorded the driving trips in Florida, may leave certain groups underrepresented. The Apple Mobility Report for Chapter 3 captured the navigation app usage only for iPhone users, while the FHWA Traffic Volume Data for Chapter 3, which captured the changing numbers of trips on the road, only counted the number of vehicles detected by the fixed loop detectors installed on certain major highways in each state. The INRIX GPS waypoint data used in Chapter 4 had a sampling rate of 2%, i.e., only 2% of the trips that happened on the Maryland roads were recorded. Compared to the average sampling rate 7.5% for SafeGraph data, the INRIX data with only 2% of sampling rate means certain demographic groups were less captured in the dataset and to what degree this selection bias influenced the mobility patterns derived from INRIX data was left for future research.

5.2 Significant contributions

Chapter 2 - contribution 1: The relationship between mobility and COVID-19 infections were understudied in summer 2020, an early-Pandemic stage, when the analysis for Chapter 2 was conducted. For this research, passively-collected mobile device data was obtained and analyzed in order to extract associations between daily mobility patterns at census tract-level and the daily new COVID-19 cases. Over a three-month period (May-July 2020) after a period of higher levels of mobility, there was a clear change in mid-June, when COVID-19 cases began to rise significantly, and

mobility declined in all three counties. Counties with higher levels of COVID-19 infections experienced larger decrease in mobility.

Chapter 2 - contribution 2: Chapter 2 examined over 30 different factors including sociodemographic, work-related, built environment, and COVID-19 factors, to identify what factors significantly contributed to the different mobility patterns among the three counties and over time using Random Forest models. The analysis revealed different factors were important for different counties and over different time periods, which could support the local officials when making mobility and public health decisions.

Chapter 3 - contribution 1: Chapter 3 examined the mobility patterns in the context of ongoing COVID-19 vaccinations at county level across the U.S. The study was novel for the analysis of compound associations of mobility and COVID-19 vaccination rates using a ML clustering approach. Five clusters of counties representing different mobility-vaccination profiles were identified across the U.S. with different underlying demographic and socioeconomic characteristics. The multivariate time-series clustering method was able to detect the spatial and temporal distributions of counties with different mobility-vaccination profiles and could be applied in the analysis of mobility and other topics.

Chapter 3 - contribution 2: The mobility patterns investigated in Chapter 3 were the relative trip changes that were fused from two different mobility data sources, the Apple Mobility Report and the FHWA Traffic Volume Data. Due to data privacy policy, the two datasets only covered a portion of all U.S. counties, which motivated the mobility data fusion. After checking the correlations between the two datasets and

confirming the two datasets had the same trends of mobility, a mobility index was calculated fusing the two datasets, which enlarged the coverage of the U.S. counties. The process of mobility data fusion inspired further research where multiple mobility datasets could be fused to obtain a larger spatial scale and enhance the interpretability of mobility patterns.

Chapter 4 - contribution 1: Chapter 4 investigated the travel (driving) time for two selected O-Ds in Maryland and revealed how different factors contributed to driving time. The two O-Ds represented short and long trips (50 miles and 150 miles) with different trip purposes of work commuting and trips to vacation/recreation destinations respectively. The characteristics of travel times for the two O-Ds were examined by three categories, including driver behavior, built environment and road network characteristics, and external factors. The key factors contributing to travel times between the two O-Ds were identified using Random Forest models. Understanding the importance of the explanatory factors on travel times is critical for transportation planners, researchers, and other stakeholders. For example, the results could better help researchers better select and weigh parameters/factors when designing advanced travel time prediction algorithms.

Chapter 4 - contribution 2: Two factors, driver route choices and driver travel speed behaviors, that could possibly impact travel times and were less studied previously, were computed and investigated in Chapter 3. A time-series ML clustering method with DTW was applied and returned different driving route choices and different travel speed behaviors for both O-Ds, which were associated with statistically significantly different travel times. Travel speed behaviors, particularly, played a

critical role in determining travel times, especially for long-distance trips and trips to vacation/recreation destinations. The two factors provided new elements for future mobility-related research.

5.3 Future work

The three studies in the dissertation are explanatory studies that used multiple ML approaches (regression, classification, and clustering) to interpret mobility patterns in the U.S.

One common future research topic for all three studies is to take advantage of the prediction capabilities of ML methods and perform predictive analysis on mobility-related research. After interpreting the factors that contribute to mobility patterns, the general public would like to know what future mobility patterns will be like utilizing the current explanatory factors. For example, Chapter 2 and Chapter 3 are in the context of COVID-19. Although COVID-19 is no longer a global Pandemic and people have resumed higher levels of mobility, the research methodologies, as well as the selected explanatory, used in Chapter 2 and Chapter 3 are still meaningful for predicting mobility when a special event happens in the future. After a better understanding of how impactful the factors were on travel times for different trip length/purpose in Chapter 3, an advanced ML or Deep Learning algorithm (e.g., Long Short-Term Memory and GeoAI) could possibly improve the accuracy of travel time prediction.

Another topic for future research lies in the mobility data sampling bias, which is discussed in Section 5.1. The COVID-19 Pandemic highlighted the importance of mobility data for e.g., tracking how infections may be spread, and how populations react during a major health crisis among other topics. There are many mobility data

sources, e.g., Google, Apple, and SafeGraph, and the mobility data made available by different providers has a sampling bias. Further study is needed on which demographic groups were underrepresented as the result of the sampling bias, how these underrepresented groups impact the mobility patterns generated by analyses such as the studies in this dissertation, and what methods can be implemented to reduce the sampling bias and account for these biases.

The three studies in this dissertation focused specifically on the vehicle mobility patterns, either the changing numbers of trips or the vehicle driving times. The term *human mobility* contains multiple aspects that are worth investigating, e.g., pedestrian, cyclist, public transit (e.g., bus and subways), air transportation, bike-sharing, web mobility, etc. (Barbosa et al., 2018). It would be interesting to investigate how ML and Deep Learning models could help interpret other human mobility patterns in different contexts (i.e., over difference space and time). The categories of factors impacting other human mobility patterns are definitely different from the categories and factors examined for vehicle mobility patterns in this dissertation. The future research could examine the different categories and factors impacting different mobility types.

Appendices

Table A1 Statistical details of the demographic and socioeconomic characteristics as well as COVID-19 cases and deaths for clusters. The top row is the mean values for each cluster, while standard deviations are in parentheses.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Percent of urban counties (%)	36.14	58.84	46.65	30.53	60.00
Percent of bachelor's degree and above (%)	13.62 (4.55)	22.15 (7.65)	17.01 (5.34)	13.33 (7.15)	28.86 (9.24)
Percent of population working from home (%)	2.01 (1.16)	3.36 (1.35)	2.52 (1.14)	2.22 (1.76)	4.14 (1.40)
Median household income ($)	51,760 (11,028)	65,013 (16,961)	56,560 (11,834)	51,408 (13,734)	73,606 (22,870)
Percent of Asian population (%)	0.85 (0.91)	2.69 (3.55)	1.46 (1.90)	0.80 (0.86)	4.88 (7.03)
Percent of below poverty (%)	15.29 (5.07)	11.56 (4.90)	13.46 (4.94)	15.29 (6.22)	11.04 (6.38)
Age 0-19 (%)	25.64 (3.09)	24.01 (3.23)	24.66 (3.00)	25.85 (3.86)	22.51 (4.23)
COVID-19 deaths (per 10,000 people)	9.48 (4.78)	7.16 (4.58)	8.46 (4.87)	8.32 (4.95)	6.13 (4.94)
Percent of male (%)	49.89 (2.07)	49.61 (1.35)	49.70 (1.81)	50.91 (3.12)	49.32 (1.47)
Percent of female (%)	50.10 (2.07)	50.38 (1.35)	50.29 (1.81)	49.08 (3.12)	50.67 (1.47)
Percent of Black population (%)	11.97 (16.10)	6.29 (9.01)	10.16 (15.55)	10.19 (11.80)	5.34 (6.24)
Population density (per km^2)	42.92 (91.89)	233.55 (645.05)	127.91 (660.02)	34.04 (129.11)	431.27 (1952.63)
Age 65 and over (%)	17.89 (3.52)	18.86 (4.71)	18.50 (4.16)	17.26 (3.93)	20.24 (6.85)
Age 45-64 (%)	26.08 (2.59)	26.69 (2.66)	26.43 (2.70)	25.91 (3.05)	27.13 (2.94)
COVID-19 cases (per 10,000 people)	367.84 (109.77)	359.54 (142.56)	374.58 (114.24)	331.95 (156.23)	346.92 (141.23)
Mean travel time to work (minutes)	24.82 (5.00)	24.11 (5.42)	24.24 (5.06)	25.58 (6.11)	24.52 (5.28)
Unemployment rate (%)	4.69 (1.47)	4.82 (1.73)	4.74 (1.58)	4.77 (1.49)	5.27 (1.85)
Percent of Hispanic population (%)	9.48 (13.01)	11.19 (16.42)	9.61 (13.51)	11.93 (13.21)	11.44 (15.12)
Percent of White population (%)	80.43 (16.15)	82.63 (13.51)	81.32 (16.86)	80.79 (14.08)	78.60 (16.79)
Age 20-44 (%)	30.37 (3.49)	30.41 (4.52)	30.40 (4.17)	31.06 (4.48)	30.09 (5.59)

Table A2 Adjusted p-values for all pair comparisons between 5 clusters from Tukey HSD post hoc test. Significance level is 0.05.

1-1	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
0	0	0.9743	0	0	0	0	0	0	0	0
0	0	0.1997	0	0	0	0	0.0283	0	0	0
0	0	0.9975	0	0	0	0	0	0	0	0
0	0	0.9987	0	0	0	0	0.0093	0	0	0
0	0	1	0	0	0	0.8721	0.0001	0	0	0
0	0	0.9915	0	0.0066	0	0.0001	0.0002	0	0	0
0	0.0005	0.0178	0	0.0001	0.0329	0.2446	0.9957	0	0.001	0
0.0944	0.3209	0	0.0318	0.9375	0	0.6348	0	0.321	0	0.0944
0.0944	0.3209	0	0.0318	0.9375	0	0.6348	0	0.321	0	0.0944
0	0.0944	0.4859	0	0	0.008	0.9673	1	0.0063	0.0261	0
0	0.0808	0.9998	0	0.0473	0.002	0.0275	0.3526	0	0	0
0.0007	0.044	0.3273	0	0.5982	0.0001	0.017	0.0023	0.0005	0	0.0007
0.0011	0.1041	0.9219	0.0011	0.4736	0.005	0.5246	0.1124	0.0751	0.0011	0.0011
0.7846	0.8372	0.0023	0.454	0.2593	0.0644	0.8708	0.0002	0.1853	0.8446	0.7846
0.1284	0.1945	0.3454	0.9795	0.9937	0.0069	0.9409	0.0108	0.9827	0.4178	0.1284
0.567	0.9734	0.9539	0.0027	0.8876	0.9959	0.0617	0.9982	0.0091	0.0641	0.567
0.2243	0.9998	0.1726	0.6362	0.3282	0.9714	0.9998	0.2334	0.7023	0.9983	0.2243
0.1178	0.8148	0.9984	0.7736	0.6349	0.6357	0.1041	0.9936	0.4301	0.7559	0.1178
0.9997	0.9999	0.2076	0.9646	1	0.3404	0.9467	0.2636	0.9498	0.2697	0.9997

Percent of bachelor’s degree and above (%)
Percent of population working from home (%)
Median household income ($)
Percent of Asian population (%)
Percent of below poverty (%)
Age 0-19 (%)
COVID-19 deaths (per 10,000 people)
Percent of male (%)
Percent of female (%)
Percent of Black population (%)
Population density (per km2)
Age 65 and over (%)
Age 45-64 (%)
COVID-19 cases (per 10,000 people)
Mean travel time to work (minutes)
Unemployment rate (%)
Percent of Hispanic population (%)
Percent of White population (%)
Age 20-44 (%)

Bibliography

2019 American Community Survey Single-Year Estimates. (2019). https://www.census.gov/newsroom/press-kits/2020/acs-1year.html

Abdi, H., & Williams, L. J. (2010). Tukey's honestly significant difference (HSD) test. *Encyclopedia of Research Design*, *3*(1), 1–5. http://www.utd.edu/~herve

Abrar, S. M., Awasthi, N., Smolyak, D., & Frias-Martinez, V. (2023). Analysis of performance improvements and bias associated with the use of human mobility data in COVID-19 case prediction models. *ACM Journal on Computing and Sustainable Societies*. https://doi.org/10.1145/3616380

Agarwal, A. (2004). A Comparison of Weekend and Weekday Travel Behavior Characteristics in Urban Areas. *USF Tampa Graduate Theses and Dissertation*. https://digitalcommons.usf.edu/etd

Ahie, L. M., Charlton, S. G., & Starkey, N. J. (2015). The role of preference in speed choice. *Transportation Research Part F: Traffic Psychology and Behaviour*, *30*, 66–73. https://doi.org/10.1016/j.trf.2015.02.007

Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., & Gardner, L. M. (2020). Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases*, *20*(11), 1247–1254. https://doi.org/10.1016/S1473-3099(20)30553-3

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., & Tomasini, M. (2018). Human mobility: Models and applications. In *Physics Reports* (Vol. 734, pp. 1–74). Elsevier B.V. https://doi.org/10.1016/j.physrep.2018.01.001

Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., & Favre, G. (2020). Real estimates of mortality following COVID-19 infection. *The Lancet Infectious Diseases*, *20*(7), 773. https://doi.org/10.1016/S1473-3099(20)30195-X

Belik, V., Geisel, T., & Brockmann, D. (2011). Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases. *Physical Review X*, *1*(1), 1–5. https://doi.org/10.1103/PhysRevX.1.011001

Bezcioglu, M., Yigit, C. O., Mazzoni, A., Fortunato, M., Dindar, A. A., & Karadeniz, B. (2022). High-rate (20 Hz) single-frequency GPS/GALILEO variometric approach for real-time structural health monitoring and rapid risk assessment. *Advances in Space Research*, *70*(5), 1388–1405. https://doi.org/10.1016/j.asr.2022.05.074

Blei, A., Kawamura, K., Javanmardi, M., & Mohammadian, A. (2015). Evaluation methods for estimating vehicle miles traveled with GPS travel survey data. *Transportation Research Record*, *2525*, 112–120. https://doi.org/10.3141/2495-12

Branco, P., Ribeiro, R. P., Torgo, L., Krawczyk, B., & Moniz, N. (2017). SMOGN: a Pre-processing Approach for Imbalanced Regression. *Proceedings of Machine Learning Research*, *74*, 36–50.

Breiman, L. (2001). Random Forests. In *Machine Learning* (Vol. 45, pp. 5–32).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Carrion, C., & Levinson, D. (2012). Value of travel time reliability: A review of current evidence. *Transportation Research Part A: Policy and Practice*, *46*(4), 720–741. https://doi.org/10.1016/j.tra.2012.01.003

Castles, S. (2018). Social Transformation and Human Mobility: Reflections on the Past, Present and Future of Migration. *Journal of Intercultural Studies*, *39*(2), 238–251. https://doi.org/10.1080/07256868.2018.1444351

Centers for Disease Control and Prevention. (2023). *CDC COVID Data Tracker*. https://covid.cdc.gov/covid-data-tracker/#datatracker-home

Chang, M.-C., Kahn, R., Li, Y.-A., Lee, C.-S., Buckee, C. O., & Chang, H.-H. (2020). Variation in human mobility and its impact on the risk of future COVID-19 outbreaks in Taiwan. *MedRxiv*.

Chen, C., Liaw, A., & Breiman, L. (1999). Using Random Forest to Learn Imbalanced Data. *Discovery*, 1–12.

Chen, C., Zhao, X., Zhang, Y., Rong, J., & Liu, X. (2019). A graphical modeling method for individual driving behavior and its application in driving safety analysis using GPS data. *Transportation Research Part F: Traffic Psychology and Behaviour*, *63*, 118–134. https://doi.org/10.1016/j.trf.2019.03.017

Coston, A., Guha, N., Ouyang, D., Lu, L., Chouldechova, A., & Ho, D. E. (2021). Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for COVID-19 Policy. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 173–184. https://doi.org/10.1145/3442188.3445881

Derrow-Pinion, A., She, J., Wong, D., Lange, O., Hester, T., Perez, L., Nunkesser, M., Lee, S., Guo, X., Wiltshire, B., Battaglia, P. W., Gupta, V., Li, A., Xu, Z., Sanchez-Gonzalez, A., Li, Y., & Velickovic, P. (2021). ETA Prediction with Graph Neural Networks in Google Maps. *International Conference on Information and Knowledge Management, Proceedings*, 3767–3776. https://doi.org/10.1145/3459637.3481916

Dissanayake, D., & Morikawa, T. (2010). Investigating household vehicle ownership, mode choice and trip sharing decisions using a combined revealed preference/stated preference Nested Logit model: case study in Bangkok Metropolitan Region. *Journal of Transport Geography*, *18*(3), 402–410. https://doi.org/10.1016/j.jtrangeo.2009.07.003

Elliott, J. R. (2015). Natural hazards and residential mobility: General patterns and racially unequal outcomes in the United States. *Social Forces*, *93*(4), 1723–1747. https://doi.org/10.1093/sf/sou120

Epstein, Z. (2013). *Former Google engineer explains how Google Maps determines your ETA*. https://bgr.com/general/google-maps-eta-calculation-explanation/

Espinoza, B., Castillo-Chavez, C., & Perrings, C. (2020). Mobility Restrictions for the Control of Epidemics: When Do They Work? *SSRN Electronic Journal*, 1–14. https://doi.org/10.2139/ssrn.3496928

Ewing, R., Ewing, R., & Cervero, R. (2001). Travel and the Built Environment: A Synthesis. *Transportation Research Record*, *1780*(1), 87–114.

Fan, J., Fu, C., Stewart, K., & Zhang, L. (2019). Using big GPS trajectory data analytics for vehicle miles traveled estimation. *Transportation Research Part C: Emerging Technologies*, *103*, 298–307. https://doi.org/10.1016/j.trc.2019.04.019

Feng, Y., Duives, D., Daamen, W., & Hoogendoorn, S. (2021). Data collection methods for studying pedestrian behaviour: A systematic review. *Building and Environment*, *187*. https://doi.org/10.1016/j.buildenv.2020.107329

Fiori, G., & Lacoviello, M. (2021). What Did we Learn from 2 billion jabs? Early Cross-Country Evidence on the Effect of COVID-19 Vaccinations on Deaths, Mobility, and Economic Activity. *FEDS Notes*, *2021*(2984), 2–7.

Florida Department of Health. (2021a). *Florida Department of Health Open Data*. https://open-fdoh.hub.arcgis.com/

Florida Department of Health. (2021b). *Florida's COVID-19 Data and Surveillance Dashboard*. https://experience.arcgis.com/experience/96dd742462124fa0b38ddedb9b25e429

Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. *Science of the Total Environment*, *739*, 140033. https://doi.org/10.1016/j.scitotenv.2020.140033

Gao, S., Rao, J., Kang, Y., Liang, Y., & Kruse, J. (2020). Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSPATIAL Special*, *12*(1), 16–26. https://doi.org/10.1145/3404820.3404824

Gao, S., Rao, J., Kang, Y., Liang, Y., Kruse, J., Dopfer, D., Sethi, A. K., Mandujano Reyes, J. F., Yandell, B. S., & Patz, J. A. (2020). Association of Mobile Phone Location Data Indications of Travel and Stay-at-Home Mandates With COVID-

19 Infection Rates in the US. *JAMA Network Open*, *3*(9), e2020485. https://doi.org/10.1001/jamanetworkopen.2020.20485

Ghasri, M., Hossein Rashidi, T., & Waller, S. T. (2017). Developing a disaggregate travel demand system of models using data mining techniques. *Transportation Research Part A: Policy and Practice*, *105*(September), 138–153. https://doi.org/10.1016/j.tra.2017.08.020

Giordano, D., Mellia, M., & Cerquitelli, T. (2021). K-mdtsc: K-multi-dimensional time-series clustering algorithm. *Electronics (Switzerland)*, *10*(10), 1–21. https://doi.org/10.3390/electronics10101166

González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958

González, R. M., Martínez-Budría, E., Díaz-Hernández, J. J., & Esquivel, A. (2015). Explanatory factors of distorted perceptions of travel time in tram. *Transportation Research Part F: Traffic Psychology and Behaviour*, *30*, 107–114. https://doi.org/10.1016/j.trf.2015.02.006

Griffiths, S., Furszyfer Del Rio, D., & Sovacool, B. (2021). Policy mixes to achieve sustainable mobility after the COVID-19 crisis. *Renewable and Sustainable Energy Reviews*, *143*(December 2020), 110919. https://doi.org/10.1016/j.rser.2021.110919

Hagenauer, J., Omrani, H., & Helbich, M. (2019). Assessing the performance of 38 machine learning models: the case of land consumption rates in Bavaria,

Germany. *International Journal of Geographical Information Science*, *33*(7), 1399–1419. https://doi.org/10.1080/13658816.2019.1579333

Han, S., Williamson, B. D., & Fong, Y. (2021). Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01688-3

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348–361. https://doi.org/10.3102/1076998619832248

Henao, A., & Marshall, W. E. (2019). The impact of ride-hailing on vehicle miles traveled. *Transportation*, *46*(6), 2173–2194. https://doi.org/10.1007/s11116-018-9923-2

Hsueh, Y. L., & Chen, H. C. (2018). Map matching for low-sampling-rate GPS trajectories by exploring real-time moving directions. *Information Sciences*, *433–434*, 55–69. https://doi.org/10.1016/j.ins.2017.12.031

Huang, X., Li, Z., Lu, J., Wang, S., Wei, H., & Chen, B. (2020). Time-series clustering for home dwell time during COVID-19: What can we learn from it? *ISPRS International Journal of Geo-Information*, *9*(11). https://doi.org/10.3390/ijgi9110675

Huang, X., Lu, J., Gao, S., Wang, S., Liu, Z., & Wei, H. (2021). Staying at home is a privilege: evidence from fine-grained mobile phone location data in the U.S. during the COVID-19 pandemic. *Annals of the American Association of Geographers*, *March*. https://doi.org/10.1080/24694452.2021.1904819

Hunter, T., Herring, R., Abbeel, P., & Bayen, A. (2009). Path and travel time inference from GPS probe vehicle data. *NIPS Analyzing Networks and Learning with Graphs*.

Idé, T., & Kato, S. (2009). Travel-time prediction using gaussian process regression: A trajectory-based approach. *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics*, *3*, 1177–1188. https://doi.org/10.1137/1.9781611972795.101

Ireland, K. (2011). *How Does Google Maps Calculate Travel Time?* https://www.techwalla.com/articles/how-does-google-maps-calculate-travel-time

Jahangiri, A., & Rakha, H. A. (2015). Transportation Mode Recognition Using Mobile Phone Sensor Data. *Ieee Transactions on Intelligent Transportation Systems*, *16*(5), 2406–2417.

Jang, J., & Ko, J. (2019). Factors associated with commuter satisfaction across travel time ranges. *Transportation Research Part F: Traffic Psychology and Behaviour*, *66*, 393–405. https://doi.org/10.1016/j.trf.2019.09.019

Jenelius, E., & Koutsopoulos, H. N. (2013). Travel Time Estimation for Urban Road Networks. *Transportation Research Part B*, *March*, pp 2-26.

Ji, Y., Ma, Z., Peppelenbosch, M. P., & Pan, Q. (2020). Potential association between COVID-19 mortality and health-care resource availability. *The Lancet Global Health*, *8*(4), e480. https://doi.org/10.1016/S2214-109X(20)30068-1

Jin, G., Yan, H., Li, F., Li, Y., & Huang, J. (2021). Hierarchical Neural Architecture Search for Travel Time Estimation. *ACM SIGSPATIAL 2021 Conference*, 91–94. https://doi.org/10.1145/3474717.3483913

Kabiri, A., Darzi, A., Zhou, W., Sun, Q., & Zhang, L. (2020). How different age groups responded to the COVID-19 pandemic in terms of mobility behaviors: a case study of the United States. *ArXiv Preprint ArXiv:2007.10436*.

Kang, Y., Gao, S., Liang, Y., Li, M., Rao, J., & Kruse, J. (2020). Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic. *Scientific Data*, *7*(1), 1–13. https://doi.org/10.1038/s41597-020-00734-5

Kaplan, S., Guvensan, M. A., Yavuz, A. G., & Karalurt, Y. (2015). Driver Behavior Analysis for Safe Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *16*(6), 3017–3032. https://doi.org/10.1109/TITS.2015.2462084

Kazagli, E., & Koutsopoulos, H. N. (2013). Arterial Travel Time Estimation from Automatic Number Plate Recognition Data. *Proceedings of the 92nd Annual TRB Meeting*.

Kishore, N., Kiang, M., Engø-Monsen, K., Vembar, N., Balsari, S., & Buckee, C. (2020). *Mobile phone data analysis guidelines: applications to monitoring physical distancing and modeling COVID-19*. https://doi.org/10.31219/osf.io/5arjy

Kraemer, M. U. G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., & Scarpino, S. V. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, *368*(6490), 493–497. https://doi.org/10.1126/science.abb4218

Kreuter, F., Barkay, N., Bilinski, A., Bradford, A., Chiu, S., Eliat, R., Fan, J., Galili, T., Haimovich, D., Kim, B., LaRocca, S., Li, Y., Morris, K., Presser, S., Salomon, J. A., Sarig, T., Stewart, K., Stuart, E. A., & Tibshiran, R. (2020). Partnering with Facebook on a university-based rapid turn-around global survey. *Survey Research Methods*, *14*(2), 159–163. https://doi.org/10.18148/srm/2020.v14i2.7761

Kwan, M. P. (2000). Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: A methodological exploration with a large data set. *Transportation Research Part C: Emerging Technologies*, *8*(1–6), 185–203. https://doi.org/10.1016/S0968-090X(00)00017-6

Kwan, M. P. (2004). GIS methods in time-geographic research: Geocomputation and geovisualization of human activity patterns. *Geografiska Annaler, Series B: Human Geography*, *86*(4), 267–280. https://doi.org/10.1111/j.0435-3684.2004.00167.x

Lamb, M. R., Kandula, S., & Shaman, J. (2021). Differential COVID-19 case positivity in New York City neighborhoods: Socioeconomic factors and mobility. *Influenza and Other Respiratory Viruses*, *15*(2), 209–217. https://doi.org/10.1111/irv.12816

Lee, J. G., Han, J., Li, X., & Gonzalez, H. (2008). TraClass: Trajectory classification using hierarchical region based and trajectory based clustering. *Proceedings of the VLDB Endowment*, *1*(1), 1081–1094. https://doi.org/10.14778/1453856.1453972

Lee, M., Zhao, J., Sun, Q., Pan, Y., Zhou, W., Xiong, C., & Zhang, L. (2020). Human mobility trends during the early stage of the COVID-19 pandemic in the United States. *PLoS ONE*, *15*(11 November), 1–11. https://doi.org/10.1371/journal.pone.0241468

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2015). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *40*(2015), 1–5.

Lenny, M. G., Triggs, T. J., & Redman, J. R. (1997). TIME OF DAY VARIATIONS IN DRIVING PERFORMANCE. In *Accid. Anal. and Prm* (Vol. 29, Issue 4).

Li, L., Jiang, R., He, Z., Chen, X. (Michael), & Zhou, X. (2020). Trajectory data-based traffic flow studies: A revisit. In *Transportation Research Part C: Emerging Technologies* (Vol. 114, pp. 225–240). Elsevier Ltd. https://doi.org/10.1016/j.trc.2020.02.016

Li, X., & Bai, R. (2017). Freight Vehicle Travel Time Prediction Using Gradient Boosting Regression Tree. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1010–1015. https://doi.org/10.1109/icmla.2016.0182

Li, Z., Hensher, D. A., & Ho, C. (2020). An empirical investigation of values of travel time savings from stated preference data and revealed preference data. *Transportation Letters*, *12*(3), 166–171. https://doi.org/10.1080/19427867.2018.1546806

Li, Z., Ning, H., Jing, F., & Lessani, M. N. (2023). Understanding the bias of mobile location data across spatial scales and over time: a comprehensive analysis of SafeGraph data in the United States. *Available at SSRN 4383333*. https://www.researchgate.net/publication/369087412

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lou, J., Shen, X., & Niemeier, D. (2020). Are stay-at-home orders more difficult to follow for low-income groups? *Journal of Transport Geography*.

Lu, Y., & Zhang, L. (2015). Imputing trip purposes for long-distance travel. *Transportation*, *42*(4), 581–595. https://doi.org/10.1007/s11116-015-9595-0

Luca, M., Barlacchi, G., Lepri, B., & Pappalardo, L. (2021). A Survey on Deep Learning for Human Mobility. *ACM Computing Surveys*, *55*(1). https://doi.org/10.1145/3485125

Makwana, P., Kodinariya, T. M., & Makwana, P. R. (2013). Review on Determining of Cluster in K-means Clustering Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6). https://www.researchgate.net/publication/313554124

Mangrum, D., & Niekamp, P. (2020). College Student Contribution to Local COVID-19 Spread: Evidence from University Spring Break Timing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3606811

Mehta, H., Kanani, P., & Lande, P. (2019). Google Maps. *International Journal of Computer Applications*, *178*(8), 41–46. https://doi.org/10.5120/ijca2019918791

Mendes-Moreira, J., Jorge, A. M., De Sousa, J. F., & Soares, C. (2012). Comparing state-of-the-art regression methods for long term travel time prediction. *Intelligent Data Analysis*, *16*(3), 427–449. https://doi.org/10.3233/IDA-2012-0532

Mollalo, A., Rivera, K. M., & Vahedi, B. (2020). Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United

States. *International Journal of Environmental Research and Public Health*, *17*(12), 1–13. https://doi.org/10.3390/ijerph17124204

Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the Total Environment*, *728*(April), 138884. https://doi.org/10.1016/j.scitotenv.2020.138884

Nasri, A., Zhang, L., Fan, J., Stewart, K., Younes, H., Fu, C., & Jessberger, S. (2019). Advanced Vehicle Miles Traveled Estimation Methods for Non-Federal Aid System Roadways Using GPS Vehicle Trajectory Data and Statistical Power Analysis. *Transportation Research Record*. https://doi.org/10.1177/0361198119850790

Nguyen, K. A., Chen, W., Lin, B.-S., & Seeboonruang, U. (2021). Comparison of Ensemble Machine Learning Methods for Soil Erosion Pin Measurements. *ISPRS International Journal of Geo-Information*, *10*(1), 42. https://doi.org/10.3390/ijgi10010042

Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K. E. C., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N. F., Cattarino, L., Cooper, L. V., Coupland, H., Cucunuba, Z. M., Cuomo-Dannenburg, G., Dighe, A., Djaafara, B. A., Dorigatti, I., Eales, O. D., van Elsland, S. L., Nascimento, F. F., … Donnelly, C. A. (2021). Reduction in mobility and COVID-19 transmission. *Nature Communications*, *12*(1), 1–9. https://doi.org/10.1038/s41467-021-21358-2

Patire, A. D., Wright, M., Prodhomme, B., & Bayen, A. M. (2015). How much GPS data do we need? *Transportation Research Part C: Emerging Technologies*, *58*, 325–342. https://doi.org/10.1016/j.trc.2015.02.011

Quddus, M. A., Ochieng, W. Y., & Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, *15*(5), 312–328. https://doi.org/10.1016/j.trc.2007.05.002

Rahman, M. M., Thill, J.-C., & Paul, K. C. (2020). COVID-19 Pandemic Severity, Lockdown Regimes, and People's Mobility: Evidence from 88 Countries. *SSRN Electronic Journal*, 1–17. https://doi.org/10.2139/ssrn.3664131

Rahmani, M., & Koutsopoulos, H. N. (2012). Path Inference of Low-Frequency GPS Probes for Urban Networks. *2012 15th International IEEE Conference on Intelligent Transportation Systems*.

Rasouli, S., & Timmermans, H. J. P. (2012). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *Proceedings of the 17th International Conference of Hong Kong Society for Transportation Studies, HKSTS 2012: Transportation and Logistics Management*, *14*(14), 515–522.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

SafeGraph. (2021). *SafeGraph Social Distancing Metrics*. https://docs.safegraph.com/v4.0/docs/social-distancing-metrics

Sanaullah, I., Quddus, M., & Enoch, M. (2016). Developing travel time estimation methods using sparse GPS data. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *20*(6), 532–544. https://doi.org/10.1080/15472450.2016.1154764

Santos, F., Graw, V., & Bonilla, S. (2019). A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. In *PLoS ONE* (Vol. 14, Issue 12). https://doi.org/10.1371/journal.pone.0226224

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

Shen, L., & Stopher, P. R. (2014). Review of GPS Travel Survey and GPS Data-Processing Methods. In *Transport Reviews* (Vol. 34, Issue 3, pp. 316–334). Routledge. https://doi.org/10.1080/01441647.2014.903530

Siebert, J., Groß, J., & Schroth, C. (2021). A systematic review of Python packages for time series analysis. *ArXiv Preprint ArXiv:2104.07406.* http://arxiv.org/abs/2104.07406

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, *30*(5), 881–906. https://doi.org/10.1080/13658816.2015.1100731

Small, K. A. (2012). Valuation of travel time. *Economics of Transportation*, *1*(1–2), 2–14. https://doi.org/10.1016/j.ecotra.2012.09.002

Song, X., Kanasugi, H., & Shibasaki, R. (2016). DeepTransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2618–2624.

Sun, Q., Zhou, W., Kabiri, A., Darzi, A., Hu, S., Younes, H., & Zhang, L. (2020). COVID-19 and Income Profile: How People in Different Income Groups Responded to Disease Outbreak, Case Study of the United States. *ArXiv Preprint ArXiv:2007.02160*, *(5)2*(2), 285–299. http://arxiv.org/abs/2007.02160

Sun, S., Chen, J., & Sun, J. (2019). Traffic congestion prediction based on GPS trajectory data. *International Journal of Distributed Sensor Networks*, *15*(5). https://doi.org/10.1177/1550147719847440

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012017

Tang, L., Kan, Z., Zhang, X., Yang, X., Huang, F., & Li, Q. (2016). Travel time estimation at intersections based on low-frequency spatial-temporal GPS trajectory big data. *Cartography and Geographic Information Science*, *43*(5), 417–426. https://doi.org/10.1080/15230406.2015.1130649

United States Department of Housing and Urban Development. (2021). *HUD USPS ZIP CODE CROSSWALK FILES*. https://www.huduser.gov/portal/datasets/usps_crosswalk.html

United States Environmental Protection Agency. (2021). *Smart Location Database*. https://www.epa.gov/smartgrowth/smart-location-mapping

U.S. Census Bureau. (2022). *2020 American Community Survey*. https://data.census.gov/cedsci/

US Department of Transportation. (2020a). *Process for Establishing, Implementing, and Institutionalizing a Traffic Incident Management Performance Measurement Program*. https://ops.fhwa.dot.gov/publications/fhwahop15028/step1.htm

US Department of Transportation. (2020b). *Traffic Congestion and Reliability: Trends and Advanced Strategies for Congestion Mitigation*. https://ops.fhwa.dot.gov/congestion_report/chapter2.htm

U.S. Food and Drug Administration. (2020). *FDA Approves First COVID-19 Vaccine*. https://www.fda.gov/news-events/press-announcements/fda-approves-first-covid-19-vaccine

Wang, Q., & Taylor, J. E. (2014). Quantifying human mobility perturbation and resilience in hurricane sandy. *PLoS ONE*, *9*(11), 1–5. https://doi.org/10.1371/journal.pone.0112608

Wang, Y., Zheng, Y., & Xue, Y. (2014). Travel time estimation of a path using sparse trajectories. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *5*, 25–34. https://doi.org/10.1145/2623330.2623656

Wang, Z., Fu, K., & Ye, J. (2018). Learning to estimate the travel time. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 858–866. https://doi.org/10.1145/3219819.3219900

Wang, Z., He, S. Y., & Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, *11*, 141–155. https://doi.org/10.1016/j.tbs.2017.02.005

Warren, M. S., & Skillman, S. W. (2020). Mobility Changes in Response to COVID-19. *ArXiv*.

Wellenius, G. A., Vispute, S., Espinosa, V., Fabrikant, A., Tsai, T. C., Hennessy, J., Dai, A., Williams, B., Gadepalli, K., Boulanger, A., Pearce, A., Kamath, C., Schlosberg, A., Bendebury, C., Mandayam, C., Stanton, C., Bavadekar, S., Pluntke, C., Desfontaines, D., … Gabrilovich, E. (2020). Impacts of US State-Level Social Distancing Policies on Population Mobility and COVID-19 Case Growth During the First Wave of the Pandemic. *ArXiv Preprint ArXiv:2004.10172*.

Wolfe, M. K., McDonald, N. C., & Holmes, G. M. (2020). Transportation Barriers to Health Care in the United States: Findings From the National Health Interview Survey, 1997–2017. *American Journal of Public Health*, *110*(6), 815–822. https://doi.org/10.2105/AJPH.2020.305579

Wu, C. H., Ho, J. M., & Lee, D. T. (2004). Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, *5*(4), 276–281. https://doi.org/10.1109/TITS.2004.837813

Xia, F., Wang, J., Kong, X., Wang, Z., Li, J., & Liu, C. (2018). Exploring Human Mobility Patterns in Urban Scenarios: A Trajectory Data Perspective. *IEEE Communications Magazine*, *56*(3), 142–149. https://doi.org/10.1109/MCOM.2018.1700242

Xiao, G., Juan, Z., & Zhang, C. (2015). Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems*, *54*, 14–22. https://doi.org/10.1016/j.compenvurbsys.2015.05.005

Xiong, C., Hu, S., Yang, M., Luo, W., & Zhang, L. (2020). Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2010836117

Xu, M., Guo, K., Fang, J., & Chen, Z. (2019). Utilizing Artificial Neural Network in GPS-Equipped Probe Vehicles Data- Based Travel Time Estimation. *IEEE Access*, *7*, 89412–89426. https://doi.org/10.1109/ACCESS.2019.2926851

Yang, C., & Gidófalvi, G. (2018). Fast map matching, an algorithm integrating hidden Markov model with precomputation. *International Journal of Geographical Information Science*, *32*(3), 547–570. https://doi.org/10.1080/13658816.2017.1400548

Yang, X., Stewart, K., Tang, L., Xie, Z., & Li, Q. (2018). A review of GPS trajectories classification based on transportation mode. *Sensors (Switzerland)*, *18*(11), 1–20. https://doi.org/10.3390/s18113741

Yu, J. J. Q., & Gu, J. (2019). Real-Time Traffic Speed Estimation with Graph Convolutional Generative Autoencoder. *IEEE Transactions on Intelligent*

Transportation Systems, *20*(10), 3940–3951. https://doi.org/10.1109/TITS.2019.2910560

Yu, J., Stettler, M. E. J., Angeloudis, P., Hu, S., & Chen, X. (Michael). (2020). Urban network-wide traffic speed estimation with massive ride-sourcing GPS traces. *Transportation Research Part C: Emerging Technologies*, *112*, 136–152. https://doi.org/10.1016/j.trc.2020.01.023

Zamir, K., Nasri, A., Baghaei, B., Mahapatra, S., & Zhang, L. (2014). Effects of transit-oriented development on trip generation, distribution, and mode share in Washington, D. C., and Baltimore, Maryland. *Transportation Research Record*, *2413*, 45–53. https://doi.org/10.3141/2413-05

Zhang, L., Ghader, S., Pack, M. L., Xiong, C., Darzi, A., Yang, M., Sun, Q. Q., Kabiri, A. A., & Hu, S. (2020). An interactive COVID-19 mobility impact and social distancing analysis platform. *MedRxiv*, 1–14. https://doi.org/10.1101/2020.04.29.20085472

Zhang, P., Stewart, K., & Li, Y. (2023). Estimating traffic speed and speeding using passively collected big mobility data and a distributed computing framework. *Transactions in GIS*. https://doi.org/10.1111/tgis.13061

Zhu, G., & Stewart, K. (2023). Space-time relationships between COVID-19 vaccinations and human mobility patterns in the United States. *Applied Geography*, *159*, 103086. https://doi.org/10.1016/j.apgeog.2023.103086

Zhu, G., Stewart, K., Niemeier, D., & Fan, J. (2021). Understanding the Drivers of Mobility during the COVID-19 Pandemic in Florida, USA Using a Machine

Learning Approach. *ISPRS International Journal of Geo-Information*, *10*(7), 440. https://doi.org/10.3390/ijgi10070440

Zhu, R., Anselin, L., Batty, M., Kwan, M., Chen, M., Luo, W., Cheng, T., Lim, C. K., Santi, P., Cheng, C., Gu, Q., Wong, M. S., Zhang, K., Lü, G., & Ratti, C. (2021). The effects of different travel modes and travel destinations on COVID-19 transmission in global cities. *Science Bulletin*. https://doi.org/10.1016/j.scib.2021.11.023