April 2021
Report No. 21-016

Charles D. Baker
Governor
Karyn E. Polito
Lieutenant Governor
Jamey Tesler
Acting Secretary & CEO

Translating Data Generated by the Transit App into Insights on Transportation Use in Greater Boston

Principal Investigator(s)
Dr. Daniel O'Brien
Dr. Qi Wang
Northeastern University

Research and Technology Transfer Section
MassDOT Office of Transportation Planning

U.S. Department of Transportation
Federal Highway Administration

[This blank, unnumbered page will be the back of your front cover]

Technical Report Document Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
21-016	n/a	n/a

4. Title and Subtitle	5. Report Date
Translating Data Generated by the Transit App into Insights on Transportation Use in Greater Boston	April 2021
	6. Performing Organization Code: n/a

7. Author(s)	8. Performing Organization Report No.
Daniel T. O'Brien, Qi Wang, Justin de Benedictis-Kessner	n/a

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Boston Area Research Initiative, Northeastern University, 1135 Tremont St., Boston, MA 02120	11. Contract or Grant No.: 105598

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Massachusetts Department of Transportation Office of Transportation Planning Ten Park Plaza, Suite 4150, Boston, MA 02116	Final Report - May 2021 [September 2018 – April 2021]
	14. Sponsoring Agency Code: n/a

15. Supplementary Notes
Project Champion - Jen Elise Prescott and Anna Gartsman, MBTA

16. Abstract
Transit is an app that Massachusetts Bay Transit Authority (MBTA) riders can use to navigate the system and compare alternative routes. The data generated by the app are made available to MBTA by Transit and they are a rich resource for understanding the behavior of customers. The purpose of this project was to tap this unprecedented resource through three tasks: (**1) Construct Transit data storage infrastructure** including two subtasks: determining the content and organization of the data and building a storage infrastructure; **(2) Enhanced measures for research and policy**, including a series of derived measures describing sessions (i.e., individual uses of the app), routes, and users that were not explicitly available in the data but could be calculated from the information that was there or otherwise incorporated. **(3) Research illustrating the utility of Transit data** to demonstrate how it might be used in the future. This culminated in an analysis of when and why Transit users selected rideshare options (or TNC) over public transit. This was the most detailed analysis of its kind and the first to leverage real-life decisions as made by transit riders and has been submitted for publication at the journal *Transportation Research Part C: Emerging Technologies*.

17. Key Word	18. Distribution Statement
Transit App, MBTA, Massachusetts Bay Transit Authority, transit data	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
unclassified	unclassified	70	n/a

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

This page left blank intentionally.

Translating Data Generated by the Transit App into Insights on Transportation Use in Greater Boston

Final Report

Prepared By:

Daniel T. O'Brien
Principal Investigator

Qi Wang
Co-Principal Investigator

Justin de Benedictis-Kessner
Co-Principal Investigator

Boston Area Research Initiative
Northeastern University
1135 Tremon St.
Boston, MA 02120

Prepared For:

Massachusetts Department of Transportation
Office of Transportation Planning
Ten Park Plaza, Suite 4150
Boston, MA 02116

April 2021

This page left blank intentionally.

Acknowledgements

Prepared in cooperation with the Massachusetts Department of Transportation, Office of Transportation Planning, and the United States Department of Transportation, Federal Highway Administration.

The Project Team would like to thank partners at the Massachusetts Bay Transit Authority's Office of Performance Management and Innovation, Jen Elise Prescott, Anna Gartsman, and Monisha Reginald, for their collaboration in sharing data, proposing policy priorities for our research, and identifying implications. We also thank representatives from Transit for data sharing and support.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official view or policies of the Massachusetts Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

This page left blank intentionally.

Executive Summary

This study of "Translating Data Generated by the Transit App into Insights on Transportation Use in Greater Boston" was undertaken as part of the Massachusetts Department of Transportation (MassDOT) Research Program. This program is funded with Federal Highway Administration (FHWA) State Planning and Research (SPR) funds. Through this program, applied research is conducted on topics of importance to the Commonwealth of Massachusetts transportation agencies.

The project was completed by the Boston Area Research Initiative (BARI), led by PIs Daniel T. O'Brien, Justin de Benedictis-Kessner, and Qi Wang. Transit is an app that MBTA riders can use to navigate the system and compare alternative routes. These decisions leave "digital breadcrumbs" that are stored in a database that is a rich resource for understanding the behavior of Massachusetts Bay Transit Authority (MBTA) customers, and it is made available to the MBTA through their contract with Transit. The purpose of this project was to tap this unprecedented resource, evaluate and demonstrate its potential, and support the MBTA in using these data for research and analysis moving forward. This work was pursued in collaboration with the Office of Performance Management and Innovation (OPMI).

The project consisted of three tasks that constituted the full data life cycle from description to interpretation to research to the construction of technical processes for sustainable use of the data. We summarize the highlights from each of these tasks.

Task 1. Construct Transit Data Storage Infrastructure included two subtasks: determining the content and organization of the data; and building a storage infrastructure. BARI built a Python script that unfolds the data from their original JSON format into a tabular format that is accessible to analysis; provided variable-by-variable-documentation for the resultant tables; and deployed these on MBTA servers for sustainable use.

Task 2. Enhanced Measures for Research and Policy included a series of derived measures describing sessions (i.e., individual uses of the app), routes, and users that were not explicitly available in the data but could be calculated from the information that was there (e.g., tabulating total walk distance across a trip with multiple segments of walking). We also developed mechanisms for merging the data with general transit feed specification (GTFS) feeds and weather data from the National Oceanic and Atmospheric Administration (NOAA). These variables were included in the full documentation.

Task 3. Research Illustrating the Utility of Transit Data for future use. We determined that the greatest novelty in the data was their ability to capture real-world mode choice decisions. Based on this, we completed an analysis that examined if and when Transit users appeared to select rideshare options (or TNC), as indicated by their decision to "tap" on them when planning a trip. We analyzed the general spatial and temporal distribution of these TNC taps as well as the contextual conditions that were associated with them, many of which were measures drawn from the original data or derived in Task 2 (e.g., wait time, walking distance). The findings illustrated both complementarity and competition between TNC services and public transit. This analysis, which was the most detailed of its kind and the first to leverage real-life decisions made by transit riders, was submitted for publication at *Transportation Research Part C: Emerging Technologies*.

This page left blank intentionally.

Table of Contents

This page left blank intentionally.

List of Tables

List of Figures

This page left blank intentionally.

List of Acronyms

Acronym	Expansion
AdaBoost	Adaptive Boosting
BARI	Boston Area Research Initiative
CART	Classification and Regression Tree
FHWA	Federal Highway Administration
MassDOT	Massachusetts Department of Transportation
NCDC	National Climatic Data Center
NOAA	National Oceanic and Atmospheric Administration
OPMI	Office of Performance Management and Innovation
RF	Random Forest
SP	Stated Preference
SPR	State Planning and Research
SVM	Support Vector Machine
TNC	Transportation Network Companies

This page left blank intentionally.

1.0 Introduction

This study of "Translating Data Generated by the Transit App into Insights on Transportation Use in Greater Boston" was undertaken as part of the Massachusetts Department of Transportation (MassDOT) Research Program. This program is funded with Federal Highway Administration (FHWA) State Planning and Research (SPR) funds. Through this program, applied research is conducted on topics of importance to the Commonwealth of Massachusetts transportation agencies. The work was completed by the Boston Area Research Initiative (BARI).

Transit is a mobile phone application that provides real-time public transit data, as well as information about alternative options, allowing riders to make informed decisions about how to navigate a metro region. It is available internationally in over 200 metro regions, including greater Boston. The Massachusetts Bay Transit Authority (MBTA) has a contract with Transit by which MBTA endorses Transit as its preferred app for navigating the system and Transit provides MBTA with access to the data generated by the app. The purpose of this project was to tap this unprecedented resource, evaluate and demonstrate its potential, and support the MBTA in using these data for their own purposes moving forward.

When riders interact with the Transit app, they leave "digital breadcrumbs" that are collected by Transit and stored in a database. This database is a rich resource for understanding the behavior of customers and unique in that it captures their real-time decision-making and habits. It can in theory be useful for answering a variety of questions. When, where, and why do residents of greater Boston use transit? How are these tendencies affected by the immediate context, policy changes, or disturbances? What factors do individuals prioritize when selecting one of multiple trip options? The raw data generated by Transit (hereon, Transit data), however, are not ready for such analyses. Like many other modern digital (i.e., "big") data generated by administrative systems or internet-based platforms, they lack the structure and documentation necessary to be immediately useful for research, policy, or practice. The first step is to get to know the data better, in order to determine what precise information is contained in the data, what insights they will support, and how we can achieve these insights.

BARI worked in close collaboration with the Office of Performance Management and Innovation (OPMI) to develop an understanding of the content and potential of the Transit. The project consisted of three tasks: **(1) Construct transit data storage infrastructure,** including the two subtasks of determining the content and organization of the data and then building a storage infrastructure for housing the data in tabular form. **(2) Enhanced measures for research and policy**, consisting of derived measures describing sessions (i.e., individual uses of the app), routes, and users that were not explicitly available in the data but could be calculated from the information that was there. **(3) Research illustrating the utility of Transit data** to demonstrate how it might be used in the future. These three tasks constituted the full data life cycle, from description to interpretation to research to the construction of technical processes for sustainable use of the data.

This document summarizes the products of each of the three tasks, with a focus on the culminating research project, which built on and illuminated the value of the first two tasks. Once we completed the initial process of preparing the data for research, we invested much effort in determining what types of research questions Transit data were best suited to answer. A series of initial analyses revealed considerable difficulty in confirming whether the population of Transit app users were demographically or behaviorally representative of the broader population or of those that ride public transit, especially being that app users do not report any demographic information. One promising lead was that a small subset of app users have included their home address in their profile. We used techniques common in the study of mobility data to infer home addresses to see if locations of greatest usage matched onto stated home addresses, but the two rarely agreed. The hope was that if this approach worked we could estimate the demographic composition of Transit users. It turns out that, while people use their cellular devices for communication and social media while at home, allowing the inference of home addresses, they have generally already left the home when they are checking for the schedule of their bus, train, or other travel options. For this reason, we concluded that without a survey of the app users themselves the problem of representativeness seemed intractable (we provide the full analysis in Section 7.1 in the Appendix).

We determined in collaboration with OPMI that one thing that was unique about Transit data that no other data source offered was the ability to reveal the real-time decision-making of its users as they considered different options. This ability partially obviated concerns about representativeness because it allowed for within-person comparisons of options with different characteristics; for instance, if a person is presented with five options for travel, we can determine whether certain modes are more attractive, the influence of wait time or transfers, and other contextual factors without concern as to whether the sample itself was fully representative. Based on this, we undertook an analysis of when and why Transit users select rideshare options (or transportation network companies, or TNC), such as Uber and Lyft. We analyzed the general spatial and temporal distribution of these TNC taps as well as the contextual conditions that were associated with them, many of which were measures drawn from the original data or derived in Task 2 (e.g., wait time, walking distance, time to destination). We believe this is an important advance and we have submitted it for publication at the journal *Transportation Research Part C: Emerging Technologies* with the title 'To Ride-Hail or Not to Ride-Hail? Complementarity and Competition Between Public Transit and TNCs Through the Lens of App Data'. It also makes up the bulk of this report.

The remainder of this document is structured as follows. This chapter continues with background on TNC and current knowledge about when and why transit riders choose it over public transit options. Chapter 2 presents the Research Methodology, which includes a description of Tasks 1 & 2, including the steps for processing the data, the resultant data structure, and the various variables accessed, created, and leveraged. Chapter 3 presents the Results of the study. Chapter 4 presents the Implementation and Technology Transfer, including a more detailed description of the tools for data processing and associated documentation that BARI delivered. Chapter 5 presents the Conclusions from the study on TNC as well as insights on the utility of the Transit data themselves. Chapter 6 includes References from throughout the document. And Chapter 7 contains the multiple Appendices that provide extra detail on process, variables, and descriptive statistics.

1.1 Background on TNC

TNC services are reshaping urban mobility. Uber, for instance, provided 6.9 billion trips in 2019 (Uber Technologies Inc 2019) and is available in over 10,000 cities globally (Inc 2020). The expansion of TNC necessarily means that their riders are shifting many of their trips from other modes of transportation, leading them to have outsized impacts on the dynamics of urban transportation. Scholars and pundits have especially highlighted the consequences this might have for public transportation, but there are differing views on exactly how this relationship is operating. Some see the relationship as competitive, with TNCs and their offers of affordable, door-to-door travel drawing riders away from public transportation (Habib 2019). Others have argued, though, that TNC complements public transportation by filling in the gaps in public transit service. The extent to which each of these mechanisms is at work determines the actual impact TNCs have on cities (Shaheen and Chan 2016). How much do they replace public transit ridership with automobile traffic? What are the implications for the optimal allocation of resources? The answers are in turn crucial for transportation authorities and planners as they shape the urban transportation system of the 21st century.

There have been a number of studies on when and why riders choose TNC, but they have generally relied on self-report surveys (Habib 2019). Such research designs are informative, but their detail and validity are of course limited by the nature of survey work. These studies have highlighted either the potential for competition or complementarity between TNC and public transportation - but rarely have the limited data available given researchers the opportunity to evaluate the two phenomena in tandem. We assess the tradeoff between competition and complementarity directly by analyzing an original data set generated by a transportation services app, *Transit*, that informs riders about options in all available modes of transportation, including public transportation and TNCs, among others. These data allow us to observe the real-time decisions of thousands of riders as they navigate the city - in this case, greater Boston, MA. We examine the geospatial patterns of the TNC and public transit trips selected by users of the app and use machine learning models to determine which contextual and personal conditions lead a rider to select TNC over public transit. From these analyses we find new evidence that, depending on the situation, TNC both complements and competes with public transit.

1.2 Previous Studies

TNCs have become ubiquitous in cities across the world, providing riders with an alternative to more traditional forms of transportation. Most obviously, TNCs draw riders away from taxis, as they offer a nearly identical product but with more convenient, on-demand service. From a transportation planning perspective, however, there has been much discussion of how TNCs might disrupt other modes of transportation in cities, with the greatest attention to its impacts on public transit usage. Two different perspectives on this question have emerged. The predominant narrative, especially in popular media, has been that the same convenience that leads riders to choose TNCs over taxis also draws them away

from public transit (Eluru, Chakour, and El-Geneidy 2012; Bovy and Hoogendoorn-Lanser 2005; Yan, Levine, and Zhao 2019) . Alternatively, some researchers have claimed that TNCs complement public transit services in various ways, possibly even increasing public transit ridership. To date, there is some empirical evidence for each of these dynamics, but as of yet no consensus.

The theory that TNCs might compete with public transit relies on a broader logic and body of research on mode choice. In particular, the kinds of factors that tend to influence riders' mode choices make TNC services appealing. For instance, one study (Carrel, Halvorsen, and Walker 2013) showed that less reliable bus and light rail services - that is, inconsistency in the timing of vehicle arrivals - cause passengers to ride public transit less often. Such unreliability, especially when coupled with real-time information on delays, can make passengers more anxious about their transit mode choice (Chow, Block-Schachter, and Hickey 2014). Other work on public transit mode choice has emphasized the number and wait time of transfers as consistent factors driving transit mode choice. Notably, it was found that passengers consider every additional transfer to be as burdensome as an additional 10 minutes of travel time. Consequently, they estimate that reducing the number of transfers by 1 would increase transit ridership by 9.17% (Eluru, Chakour, and El-Geneidy 2012). While these studies do not explicitly reference TNCs, they highlight the kinds of advantages - including real-time arrival information, no transfers or crowding, and fewer on-board delays - that might offer TNCs a comparative advantage over public transit.

Studies that have directly compared mode choice between public transit and TNCs have reinforced TNCs potential appeal. One major convenience of TNC services is that they offer point-to-point trips without transfers. This suggests that the number of transfers in public transit may lead travelers to choose TNC services in order to minimize transfers (Yan, Levine, and Zhao 2019). Another study (Rayle et al. 2016) found that for most trips taken using TNCs, the public transit trip would have taken significantly more time. Similarly, two other studies (Habib 2019; Clewlow and Mishra 2017) show that longer total travel time for public transit trips can lead passengers to prefer TNCs and subsequently take fewer transit trips. Taken together, these results appear to confirm that the efficiency and convenience offered by TNCs cause some passengers to replace a portion of their public transit trips with ride-hailing. That said, it is possible that the cost of convenience proves too great in certain situations. For instance, very long trips typically served by commuter rail might experience less competition given that the corresponding TNC trip could be prohibitively expensive.

An alternative perspective is that TNC can complement public transit, possibly even increasing its effectiveness. There are two main elements to this argument. First, TNCs may solve what is known as “the last mile problem" - passengers’ ability to get to and from the public transit stations where their trip begins and ends (Shaheen and Chan 2016) - and otherwise fill in gaps in public transit. A 2016 joint report from the Shared-Use Mobility Center and American Public Transportation Association stated that because TNC trips commonly occur at times and in places where public transit is not readily available, it often serves as a complement (Murphy 2016). A similar report from Pew Research Center suggested that TNC services complement public transit because TNC users are more likely to use other mobility options such as public transit than non-users of TNCs (Smith 2016). Some indirect evidence for this comes from the finding that public transit ridership, especially rail

transit ridership, has increased in places where TNCs have been introduced (Hall, Palsson, and Price 2018; Malalgoda and Lim 2019) relative to places where they have not been introduced. This parallel rise in both TNC and public transit ridership suggests that TNCs do not replace public transit trips so much as enable individuals to more easily access public transit when stations are not necessarily conveniently located. Others have made the same argument for the timing of rides, as ride-hailing services are more likely to be made in the late afternoon, evening, and night, and on weekends, times at which transit service tends to run at lower frequencies (Feigon and Murphy 2018; Habib 2019). Of course, these aggregate comparisons fail to account for potential trends in more general transportation demand rising over time as well, which could lead to observed parallel increases in TNC and public transit use independent of underlying patterns of trip replacement.

The work on complementarity between TNCs and public transit has been bolstered by evaluations of pilot programs that sought to integrate TNC services with existing fixed-route transit services. These studies are intriguing for two reasons. First, they include one of the only studies (if not the only study) that used actual records of rides to examine mode choice. One study (Terry and Bachmann 2020) found that the majority of ride-hailing trips taken in such a pilot served as "feeder" trips that brought passengers closer to transit stops. Second, however, these ridership patterns are potentially the consequence of sizable discounts on the kinds of TNC trips that complement public transit that are offered by cities partnering with ride-hailing companies (Habib 2019; Westervelt, Schank, and Huang 2017; Curtis et al. 2019). Such subsidies may be unsustainable over the long-term, especially considering the well-known lack of profitability for companies like Uber thus far (Zoepf et al. 2018). If such deliberate price manipulation is a key element of the ability for TNCs to solve last-mile problems, it raises the question of whether such complementarity will persist (Guda and Subramanian 2017).

In summary, there are two well-established perspectives on the relationship between TNC and public transit, one that sees them as complements in a modern urban transportation system, the other that sees them as competitors. The empirical evidence currently comes, however, from only a handful of studies, most of which rely on aggregate inference or survey responses. Further, many of the studies have set out to test only one of these theories and not the other, providing a partial view of the TNC-public transit relationship. There is clearly a need for ride-level data that can reveal passenger choices and the contextual features that inform them - be they wait time, transfers, walking distance, or time of day or week. Such data would help to paint a more conclusive, comprehensive picture of the complementarity or competition between TNCs and public transit (or lack thereof) but have largely been unavailable to researchers to date. Most recently a study (Zhao et al. 2020) utilized stated-preference (SP) survey data and implemented multiple advanced machine learning algorithms, they found that random forest outperforms the logit models in terms of travel mode choice prediction.

1.3 Current Study

The current study uses data from the Transit app to examine when and why travelers choose TNC over public transit services. The app is designed to help riders navigate public transit services by providing them with real-time information on local routes as well as alternative modes of transportation, including TNC. It generates data from user interactions, capturing how riders make mode-choice decisions in real time without the complications and validity issues of survey data. Rather than posing specific trade-offs and comparisons, we are able to model the rider's travel behavior based on their use of the "trip planning interface," which offers the rider multiple options across mode choices for completing a requested trip from one point to another. In doing so, the app produces objective versions of many of the measures used in previous hypothetical survey items, including number of transfers, wait time, and travel time, among others. Alongside these advantages, we do note a downside. Users of the Transit app are not necessarily a random sample of transit riders. Most importantly for our study questions, those who use the app are potentially predisposed to riding public transit, as that is the named purpose of the app and it has been endorsed by multiple public transit agencies. Keeping this in mind, the interpretation of the results should be of the factors that encourage transit riders to select TNC instead. This may in fact be a positive aspect as we highlight the features that are strong enough to encourage such a switch, but it is also important when considering the results and their implications.

Another advancement of using such data is its neutral position with regards to TNCs, which leaves no potential for bias from marketing stimulation in our examination of evidence for either competition and complementary. The precise spatiotemporal nature of the data also allows us to link the information from within the app with contextual information, including time of day and week and real-time weather, which are both believed to affect mode choice. Last, we do note that the specific data we use here only became available at the end of February 2020, meaning our study occurs almost entirely during the COVID pandemic. We include the different phases of the pandemic in Massachusetts (shutdown, Phase 1 reopening, etc.) in our models to control for this, but there is the possibility that some of the results we see are specific to the pandemic. On the other hand, it is unclear how quickly, if ever, transit usage will return to its previous equilibria, meaning that these results could be more informative moving forward. In any case, any interpretation should keep this factor in mind.

2.0 Research Methodology

2.1 Data and Preprocessing

Transit data describes each user session (i.e., discrete time period between opening and closing the app on the phone). This includes information presented to the user by the app and about the actions the user took while in the app. There is no personally identifiable information, but a device identifier represents an anonymous record of a user. These identifiers persist across sessions, meaning it is possible to compare behavior at the individual level over time. In order to directly examine decision-making behavior, our primary scale of analysis is the user session. This report focuses substantively on our analysis of when and why users of the app might choose TNC options instead of or in conjunction with public transit options, but here we describe the data source more generally. Additional information is also provided in Section 4 on Implementation and Technology Transfer and numerous Appendices (specific references made throughout this section).

The two most commonly used features in Transit are the “nearby views” and "trip planning” interfaces. The first is the default when opening the app and informs users of routes that are near to their current location. The second, displayed in Figure 2.1, is the focus of our analysis here as it permits the comparison of preferences across options for trip with established origin and destination. The trip planning interface allows users to input an origin (most often their current location from GPS) and destination. The app then returns multiple potential routes using a variety of different transportation modes by which the user could reach their destination. The app includes all forms of public transportation that are locally available (e.g., subway, commuter rail, bus, light rail, etc.) as well as walking, biking, bike share, and TNC. Routes often combine multiple transportation modes (e.g., walk-subway-walk). In both interfaces, users can “tap” on routes to learn more about them or set alerts for when they need to leave to catch a given vehicle.

The data generated by the app are provided by Transit in a JSON array wherein each element is a nested JSON object containing a user's activity for that day. This format is somewhat difficult for analysis purposes: (1) querying for a slice of the data spanning several days, i.e. a column across all user sessions, requires a full read and parse of each daily data dump; (2) querying for a specific subset of all columns for a specific set of users similarly requires a full read and parse; (3) the data must be flattened into data frame-like structures before being used statistics software packages, given that information within a given data dump reference multiple scales of analysis (e.g., app usage session, user, route) and interfaces (e.g., nearby routes, trip planning searches initiated by the user). For this reason, we “flatten” or “unfold” the data into multiple tables before analysis.

When flattening the data we generate five tables, each describing a different set of objects or actions: user sessions; trip views; favorite locations (from a user’s profile); nearby views; legs (from a suggested route in trip planning). For instance, as shown in the left panel of Figure 2.1, the trip planning interface reports various features for each route, including

estimated departure and arrival times, distance to origin station, and wait time. This information is contained in the trip planning table. We also merge in weather data for the date of the session for Boston provided by the National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Center (NCDC), contained in a sixth table. We also have built a process for merging general transit feed specification (GTFS) data for the MBTA based on spatiotemporal proximity, in order to coordinate the within-app information with the availability of vehicles for routes of interest. We have provided the syntax for the flattening process, the variables in each of these tables, and merger with external data sources to OPMI for use in future work. More detail on these processes is provided in Section 4 on Implementation and Technology Transfer, and the readme guide for the code is included in Section 7.2 in the Appendix.

For the purpose of the analysis that follows, we use data recording 336,873 trip queries generated by 55,163 users between February 29, 2020, and December 15, 2020 that included at least one tap on either a public transit or TNC option. For fuller comparative analyses (Sections 3.2 and 3.3), we limited our data to users who had tapped at least once on a public transit trip in the trip planner and at least once on a TNC trip during the study period. This ensures that we are not making unrealistic comparisons between individuals who would never tap on one mode or the other. Within these users, we only include sessions in which: (1) the user utilized the trip planning feature (i.e., requested a set of trip options from the app); (2) the trip planner generated at least one public transit option and at least one TNC option; (3) the user tapped on at least one of these options; and (4) the option tapped on was not a combination of transit and TNC (see Section 4.1 for more). This amounts to a final sample of 23,475 trip queries generated by 1,918 users for these comparative analyses.

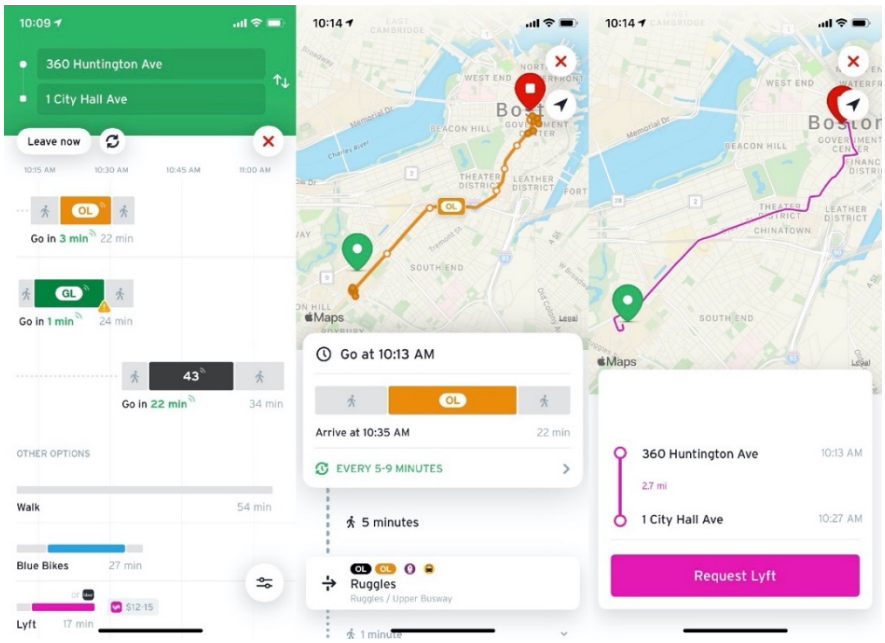

Figure 2.1: Screenshot of Transit app trip planning results (left) and detailed view of one public transportation option (middle) and TNC option (right).

2.2 Measures

Numerous measures are associated with the components of each session and interface. These are described in full detail in Section 7.3 in the Appendix, which is a documentation file that we provided to OPMI to inform their use of the flattened data; Section 7.4 in the Appendix also includes descriptive statistics for select variables capturing basic patterns of usage. To illustrate a particular use case for this content, we concentrate on the measures needed to examine users' tendencies to select options involving TNC.

In order to best measure individuals' choices between public transit and TNC, we treat users' "taps" on individual transportation options as a measure of their interest in - and possible use of - a given mode. We mark the mode choice as TNC if the user tapped on a greater number of TNC trip options during a session than the number of public transit trip options that garnered taps. For analytic purposes, we then remove any sessions in which the user tapped on a different mode choice (e.g., walk) or the tap numbers on Transit and TNC are equal (true for < 1.5\% of sessions with taps).

We use several features to predict the conditions under which users were more or less likely to choose (i.e., tap) on TNC relative to public transit options. These trip characteristics fall into three categories: (1) trip-specific features of the travel options generated by the app (e.g., distance from the user's origin position); (2) the user's ridership habits, based on historical usage patterns (e.g., tendency to ride during commuting hours); and (3) contextual factors (e.g., time of day, precipitation). Before analysis, we log-transformed numerical predictors to account for outliers and skewness and performed One-Hot Encoding for the categorical variables, translating them into a series of dichotomous variables. Section 7.4 in Appendix provides more details on measures, data transformation and correlations.

2.2.1. Trip-specific Features

The trip planning interface provides detailed information on each leg (i.e., the undivided portion of a trip on a single vehicle, like a segment of walking or a ride on a single bus) of each option it generates for a given trip. We extract a series of variables that describe the distance, duration, and modes of the trip and its legs, detailed in Table 2.1. We treat these trip-specific variables as factors that might influence a rider's decision-making. In order to have a direct comparison between TNC and public transit, we select a single "optimal" option for both modes. For both public transit and TNC, if a single trip is selected by a user, then we define that as "optimal" for that mode, as it was clearly the one that most attracted the user's attention. If either more than one or no trips are tapped for that mode, then we define the optimal route as the option with the shortest duration among all candidates with the fewest transfers. This definition is based on previous work indicating that transfers and then length are the most important factors for riders choosing between public transit options. The features of this optimal option are then those used for comparison in the models that follow.

Variable	Computation
Walk Distance	Cumulative distance to reach the origin of the first vehicular leg
Wait Time	Cumulative time before the first vehicular leg (including walking time)
Travel Time	Total duration of the trip
Longest Leg Mode	Vehicle type for the longest non-walking legs (e.g., bus, subway)
Number of Transfers	The number of vehicular legs (excluding walking)

Table 2.1: List of Trip-specific variables

2.2.2. User-specific Features

Individuals' preferences and patterns of behavior could also contribute to their choice of travel mode choice. Some such characteristics can be extracted from the user's usage of the app across the full timespan of the data（February through December 2020), as we summarize in Table 2.2 below.

Variable	Computation
Tap Frequency	Proportion of sessions with taps performed
Usage Frequency	Proportion of days with active sessions
Primary Usage Time (Used to identify "Commuters")	Most frequent usage time window (weekday commuting hours, weekday non-commuting hours or weekends)

Table 2.2: List of User-specific variables

2.2.3. Contextual Factors

Last, we consider contextual factors including weather, time of day, week, and year. Data from the primary Boston weather station, located at Boston's Logan Airport (USW00014739), were retrieved from NOAA' NCDC. These data contain records of precipitation (in inches) for each day. We match the date of the start time of each session in the Transit data to the corresponding weather readings as the amount of precipitation might influence rider behavior.

We also use the time and date of a given session to better understand the context of the user's decision making. First, we note whether the session began during commuting hours from Monday through Friday (i.e., between 7 am and 10 am as well as between 4 pm and 6 pm) or not, and whether it occurred on a weekend. Second, we divided the year according to the evolution of public policies related to the ongoing coronavirus pandemic--namely, pre-shutdown, shutdown, and reopening phases, as defined by emergency declarations by the Governor of Massachusetts. These dates are listed in detail in Table 2.3.

Date	Stage
Before March 13th, 2020	Before Emergency Declaration
March 13th to May 17th	National Emergency
May 18th to June 5th	Phase 1 Reopening
June 6th to July 6th	Phase 2 Reopening
After July 6th, 2020	Phase 3 Reopening

Table 2.3: Temporal Intervals of Lockdown Stage in response to COVID-19 in MA

2.3 Classification Models

We examine whether the features described in the previous section predict whether a rider selected TNC over public transit with four different machine-learning techniques: (1) logistic regression (Scott, Hosmer, and Lemeshow 1991; Peng, Lee, and Ingersoll 2002) uses independent variables to predict the likelihood of an outcome variable being classified as '1'; (2) support vector machine (SVM) (Chih-Wei Hsu, Chih-Chung Chang 2008), which classifies cases by finding the best k-dimensional hyperplane as a decision boundary, with the advantage of reducing overfitting; (3) random forest (RF) (Vladimir Svetnik et al. 2003; Archer and Kimes 2008), which classifies cases using a series of decision trees to control for both the bias and variance of the model; and (4) Adaptive Boosting (AdaBoost) (Schapire 2003), which integrates trees from multiple classifiers into one "great classifier" that is weighted to enhance accuracy on challenging cases, often enhancing overall fit.

For all models the data were randomly split, with 75% samples as training and the remaining 25% samples as testing. Last, the low prevalence of TNC taps creates an imbalanced data set and thus can impact the performance of prediction from conventional classification algorithms (Guo and Viktor 2004) to address this issue, we assign the class weight to each instance based on the inverse ratio of the corresponding label during the training process to balance the label distribution. Additionally, in order to ensure the robustness of the results, we reran analyses 100 times and report both the means and standard deviations for all results.

We evaluate the efficacy of our models with multiple indicators. The most commonly used metric is accuracy, which is the number of accurate predictions divided by total prediction. However, when the data categories are imbalanced, as in the current case, accuracy is a less reliable indicator of the efficacy of a model. In addition, accuracy reflects the entirety of the model, rather than evaluating its performance in predicting a particular class of interest. Therefore, we also consider other metrics that overcome these weaknesses. Precision and recall scores for the minority class are the two most common metrics for dealing with an imbalanced class distribution. The precision score (*true positive/(true positive + false positive)*) captures the proportion of predicted TNC taps that were correct while the recall (*true positive/(true positive + false negative))* reflects the ratio of actual TNC

taps that were classified correctly. Last, the F-1 score considers both metrics, with the formula:

$$F\text{-}1 = 2 * (precision * recall) / (precision + recall)$$

The F1 score is commonly used to evaluate the overall predictive performance for the minority class (Huang et al. 2016; Chawla 2009).

2.3.1. Logistic Regression

Logistic regression (Peng, Lee, and Ingersoll 2002) is widely used to model dichotomous dependent variables where log odds of the outcome is modeled as a linear combination of the predictor variables. The basic function that Logistic Regression model relies on is:

$$P(y_i = 1|x_i, w) = sigm(w_0 + w_i \cdot x_i)$$

In this equation, x_i represents the input feature vector and $y_i = 1$ or 0 represents whether it is classified as one particular label or not. The *sigm()* function refers to the sigmoid function and it has been defined as:

$$sigm(x) = \frac{e^x}{e^x + 1}$$

The loss function is the cross-entropy loss which is defined as below:

$$L(f(\hat{x}), y) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \ln f(x_i) + (1 - y_i) \cdot \ln(1 - f(x_i)))$$

Here, *X* is the set of inputs and *Y* represent the corresponding labels. The function $f(x)$ denotes the output of the logistic regression model given input features x_i and binary outcome variable y_i .

2.3.2. Support Vector Machine

Support vector machine (SVM) (Hastie, Tibshirani, and Friedman 2009; Chih-Wei Hsu, Chih-Chung Chang 2008) can be used as a classifier that performs classification by finding the best hyperplane as decision boundary. One main advantage of the SVM is that it implements instinctive complexity control to reduce overfitting. The main idea of SVM is to maximize the margin around the hyperplane and the classifier is determined by a subset of all samples. A simple SVM classifier can be built for a non-separable case via the following constraints:

$$\underset{\beta_0, \beta_1 \ldots \beta_p, \varepsilon_1 \ldots \varepsilon_n}{maximize} M$$

$$subject\ to \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M(1 - \varepsilon_i),$$

$$\varepsilon_i \geq 0, \sum_{i=1}^{n} \varepsilon_i \leq C$$

where M is the width of the margin that is being maximized, C is a tuning parameter and $\varepsilon_{1,2\ldots n}$ are slack variables. Decision function $f(x)$ is defined as:

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

And the classification result is based on the sign of $f(x)$.

The loss function for the SVC is hinge loss, which is also sometimes called the max-margin loss with respect to the optimization function and defined as:

$$L(f(\hat{x}), y) = \max(0, 1 - y \cdot f(\hat{x}))$$

Support Vector Machine is an extension of the SVC with kernels in order to expand the feature space. A simple SVM with Linear Kernel will have the following function:

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i)$$

In addition to the Linear Kernel, the Gaussian Kernel has also been selected and experimented where now the Kernel function has been replaced as:

$$K(x_i, x_{i'}) = \exp(-\gamma \|x_i - x_{i'}\|^2)$$

Note that γ is another tuning parameter that was finalized to be 0.05.

2.3.3. Random Forest

As an extension to the traditional Classification and Regression Tree (CART) method, Random Forest (Breiman 2001) is a tree ensemble non-parametric method that builds many decision trees simultaneously in order to reduce the bias and variance of the mode. The training of Random Forest model combines two statistical techniques: bootstrapping and bagging. First, *N* bootstrapped sample sets are obtained from the entire training set by random sampling with replacement. Each subset is then used to construct a regression tree without pruning, which allows the tree to grow independently to its maximum. Instead of considering all predictors, only a smaller and fixed number *m* of total

predictors would be considered as split candidates, as using only a subset of the predictors can eliminate the correlations among all *N* trees. The variable *m* is generally recommended as the square root of the total number of predictors. Subsequently, bagging essentially averages all *N* trees to reduce the overall model variance:

$$\hat{f}_{RF}^{N}(x) = \frac{1}{N}\sum_{k=1}^{N} T_k\,(x)$$

Where x is the vectored input variable and $T_k(x)$ denotes a single regression tree constructed only based on a subset of variables and the bootstrapped samples. N is the number of trees, which is a key hyperparameter and requires tuning for optimal model performance. Another important feature of RF is its capability in measuring variable importance. The algorithm permutes all the variables and records the mean decrease in prediction accuracy for each, which is then used to assign the relative importance score for each variable.

2.3.4. AdaBoost

AdaBoost (Freund 1999) is an adaptive and ensembled boosting algorithms that essentially integrates multiple classifiers into one "great classifier". The ensembling of many decision stumps and weighting them as emphasizing on the challenging instances and less on those already trained well provides an opportunity to optimize the model performance, which leads to generally higher accuracy compared to linear models. The Pseudo-code for the algorithm (Freund 1999) is shown in Figure 2.2 below.

AdaBoost

input:
training set $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
weak learner WL
number of rounds T
initialize $\mathbf{D}^{(1)} = (\frac{1}{m}, \ldots, \frac{1}{m})$.
for $t = 1, \ldots, T$:
invoke weak learner $h_t = \mathrm{WL}(\mathbf{D}^{(t)}, S)$
compute $\epsilon_t = \sum_{i=1}^{m} D_i^{(t)}\, \mathbb{1}_{[y_i \neq h_t(\mathbf{x}_i)]}$
let $w_t = \frac{1}{2}\log\left(\frac{1}{\epsilon_t} - 1\right)$
update $D_i^{(t+1)} = \frac{D_i^{(t)} \exp(-w_t y_i h_t(\mathbf{x}_i))}{\sum_{j=1}^{m} D_j^{(t)} \exp(-w_t y_j h_t(\mathbf{x}_j))}$ for all $i = 1, \ldots, m$
output the hypothesis $h_s(\mathbf{x}) = \mathrm{sign}\left(\sum_{t=1}^{T} w_t h_t(\mathbf{x})\right)$.

Figure 2.2: Pseudo-code for AdaBoost

This page left blank intentionally.

3.0 Results

3.1 Spatial-temporal Patterns of Trip Taps

We first compare the features of sessions in which a user tapped on a public transit option to those in which they tapped on TNC. We start with sessions with public transit taps to establish a baseline for interpreting the characteristics of sessions with TNC taps.

Figure 3.1 shows the spatial distribution of sessions with taps on public transit. These sessions tended to have origins and destinations that conform with general patterns that match previously-validated measures of public transit usage. Namely, they concentrate in the northern part of Boston, which is home to downtown, and the cities of Cambridge and Somerville to the north, which also have robust transit service. Two major hotspots for both origins and destinations were the areas surrounding North and South Stations, the two primary transit hubs for the metro area. Confirming these patterns, very few trips had an origin or destination that was far from Boston's geographic center (see Figure 3.1c and 3.1d). Last, Figure 3.1e shows that the temporal pattern is consistent with the daily rhythms of public transit ridership. The proportion of trips is fairly smoothly distributed across the daytime hours, with peaks during the morning and afternoon commutes. There is then little use late at night, when public transit services are shut down.

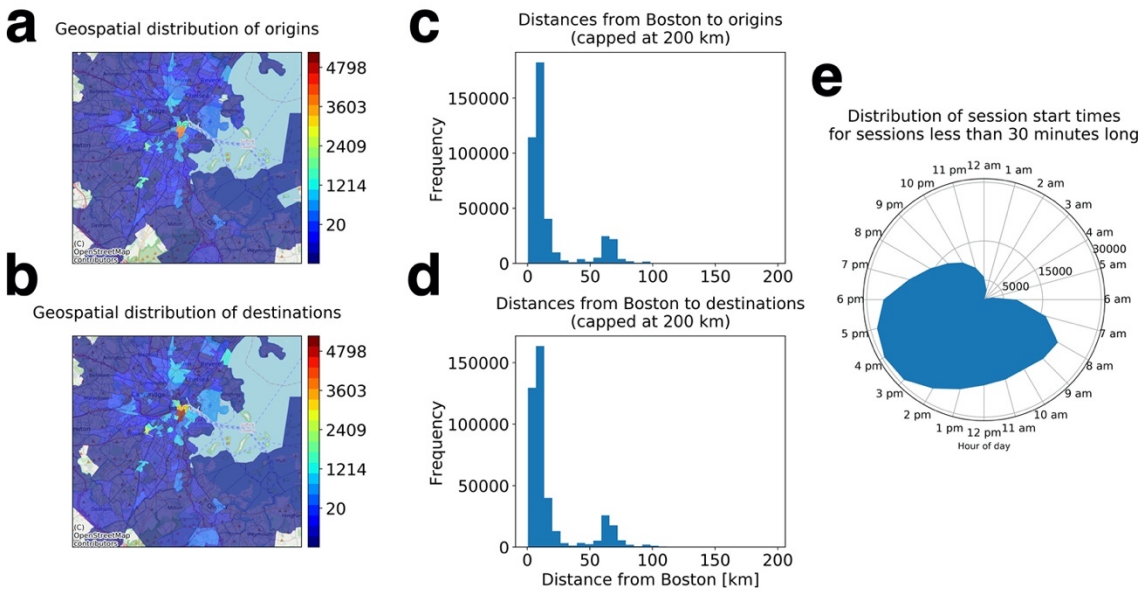

Figure 3.1: The geospatial distribution of (a) trip origins and (b) destinations, by census tract of trip with taps on public transit and the spatiotemporal features of trips with taps on public transit, including the distances of (c) origins and (d) destinations from Boston; and (e) their distribution over the course of the day.

Figure 3.2 shows that the sessions with taps on TNC had a distinct spatiotemporal profile. Though the majority of TNC trips still originated and concluded within the city of Boston, there were substantially more trips starting outside the city and off the main transit corridors (Figure 3.2a and 3.2b). There were also many more trips originating or concluding far from the center of the city (Figure 3.2c and 3.2d). In addition, there is a noticeable hotspot for destinations of trips at Logan Airport (on the northeast shore of the city, shaded dark red in Figure 3.2b), reflecting the popularity of TNC for travel from and, especially, to the airport. Taps on TNC also had their own evident temporal patterns. As with public transit taps, we see peaks during commuting hours (Figure 3.2e), but TNC taps are considerably more common relative to public transit taps after 8 pm, when transit shifts to a reduced, off-peak schedule and then shuts down. This suggests that users may tap on TNC options when transit service is less reliable or entirely unavailable and mostly used for suburban localilties (Acheampong et al. 2020).

A final observation was that approximately half (50.77%) of TNC options that were tapped actually had a combination of TNC and public transit services, typically with TNC as a means for reaching a public transit station or for completing the trip after arrival at a station - in other words, as the proverbial first or last mile. This interpretation of TNC as facilitating public transit use is further supported by the observation that the average trip in this category spends about 20% of its distance and duration in TNC and 80% on public transit. Nearly all such trips spent more time and distance in public transit than TNC. These multimodal trips represent true complementarity between TNC and public transit in the most literal sense. They would in turn be inappropriate for any deeper comparisons between sessions with taps on public transit and those with taps on TNC. Thus, all analyses that follow, which explicitly evaluate the choice between the two options, omit these cases.

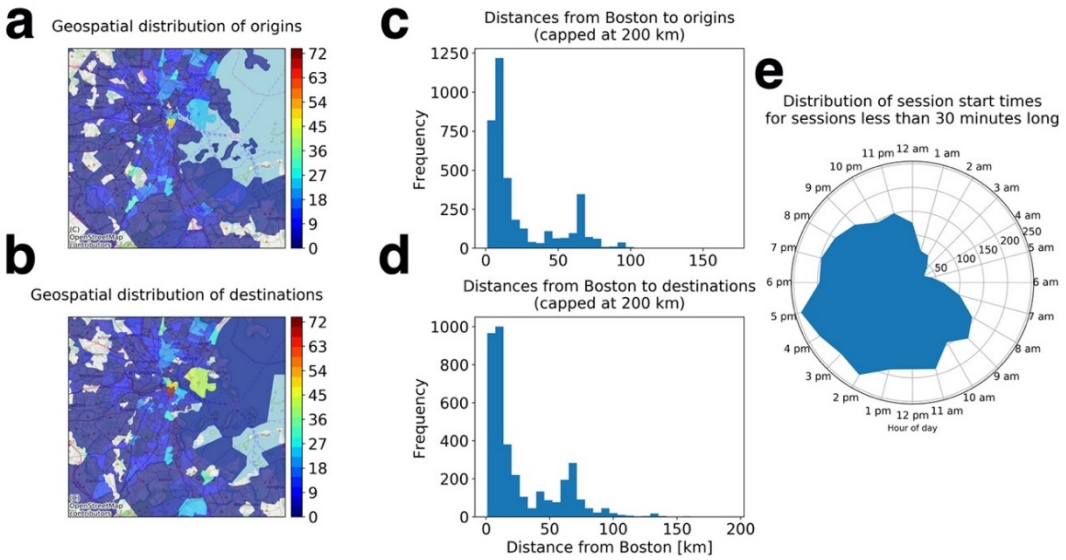

Figure 3.2: The geospatial distribution of (a) trip origins and (b) destinations, by census tract of trip with taps on public transit and the spatiotemporal features of trips with taps on TNC service, including the distances of (c) origins and (d) destinations from Boston; and (e) their distribution over the course of the day.

3.2 Trip Characteristics of Trips with TNC and Public Transit Taps

To further explore the differences between trips in which riders tapped on transit and TNC, we compare the characteristics of trips with taps on TNC to those with taps on public transit graphically in Figure 3.3. A few differences in particular stand out. First, sessions with TNC taps tend to have longer waits for and longer walks to reach the first leg of the optimal public transit option Fig 3.3a and Fig 3.3f). Interestingly, we see no clear association between mode choice and the wait time for the optimal TNC trip or the total travel time for either the optimal TNC or public transit trip (see Figure 3.3b and 3.3d, and 3.3e). One surprising finding is that taps on TNC tended to be in sessions when the distances from origins to destinations were shorter than those with public transit taps (Figure 3.3c). This finding might indicate that the use of TNC is less associated with distance itself - especially considering that many riders of the commuter rail might use the app in advance of long trips - with inconveniences like long wait times and inability to access stations. We will have to probe these patterns more closely, however, in the models that follow.

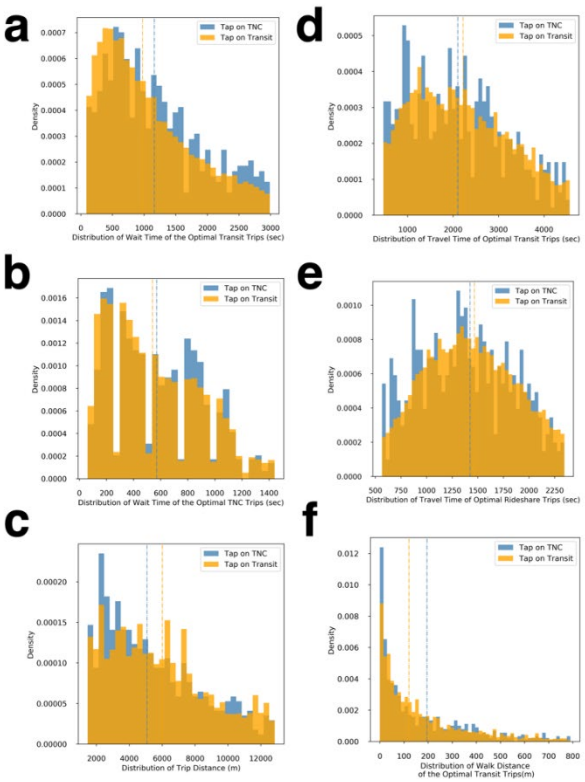

Figure 3.3: Comparisons between sessions with taps on public transit (yellow) and TNC (blue), including distributions and medians (dashed lines) for (a) estimated wait time for optimal transit trip, (b) estimated wait time for optimal TNC trip, (c) trip distance, (d) total travel time for optimal transit trip, (e) total travel time for optimal TNC trip, and (f) walking distance to reach first leg for optimal public transit trip.

3.3 When Do Riders Select TNC over Public Transit?

We next move to the series of machine learning models using trip, user, and contextual characteristics to predict whether a rider tapped on TNC. As we omit those sessions with taps on combined TNC-transit trip options, the final sample for all models is 23,475 trip queries generated by 1,918 users.

Table 3.1 summarizes the overall efficacy of each model, showing accuracy, recall, precision, and F-1 scores. We focus on F-1 scores because of their use as a comprehensive indicator of model performance when predicting a rare outcome (in this case, TNC taps). As Table 3.1 summarizes, Adaboost had the highest F-1 score at 0.38 (s.e. = 0.03), reflecting its more robust combination of multiple decision trees. The other three models were comparable to each other (F-1 scores = 0.22 - 0.25), with logistic regression and SVM especially struggled with precision, that is the proportion of sessions predicted to have TNC taps that in fact had such taps. It is important to note that even AdaBoost was only moderately capable of predicting TNC taps with both recall and precision. The moderate prediction indicates that, at least with these data, TNC taps occur rarely enough and under a sufficient range of conditions - at least through the lens of the features we included in the model - to make prospective prediction difficult. This could be owed to missing variables, the rarity of TNC taps, and other idiosyncrasies of the data (see Section 5 on Discussion for more). That said, the models are still sufficiently strong to merit examining which variables were best able to guide such prediction when it was effective. For parsimony, we proceed in two steps. First, we examine the parameter results of the logistic regression equation because it is the only one of the four methods that provides a clearly interpretable relationship between each predictor and the likelihood of TNC taps. We then describe the results from the AdaBoost and random forest models, which the best predictive efficacy, but due to their complexity are only able to evaluate variable importance in predicting TNC taps, but do not report the magnitude or direction of these relationships. In this way, AdaBoost and random forest provide the strongest evidence of which variables are meaningful, but the logistic regression fills in some ambiguity as to how to interpret these results.

Algorithm	Accuracy	Recall	Precison	F-1 Score
Logistic Regression	0.60 ± 0.01	0.52 ± 0.05	0.13 ± 0.01	0.22 ± 0.02
SVM	0.60 ± 0.01	0.53 ± 0.04	0.13 ± 0.01	0.23 ± 0.02
Random Forest	0.78 ± 0.02	0.36 ± 0.04	0.20 ± 0.02	0.25 ± 0.03
AdaBoost	0.75 ± 0.02	0.31 ± 0.03	0.53 ± 0.05	0.38 ± 0.03

Table 3.1: Predictive efficacy of the four algorithms used to predict TNC taps

3.4 What are the Most Influential Factors?

Moving to the variable importance, Figure 3.4 presents the importance of each variable in predicting mode choice with AdaBoost and random forest models. Specifically, variable importance is a quantification of the increases in the error rate with the removal of each variable from the model, with higher scores indicating that a variable improves the predictive efficacy of the model.

Such feature importance is based on expected decrease in performance when a specific predictor is included in the model, which is different from the scale of coefficients. We note that each model had five relatively more important features with scores greater than 0.1. Notably, these were the same five measures in each model. A sixth measure of less importance stood out in the AdaBoost model; the same measure was less prominent but still ranked sixth in the random forest model. In order to make sense of these results, we run logistic regressions with 100 iterations with down-sampling (0.75) at each iteration, in order to maintain consistent interpretation with the Adaboost and random forest methodologies. We concentrate on these six measures, most of which were also significant in the logistic regression model, in the summary that follows.

Five features that were most predictive of TNC taps echoed the descriptive results from the previous section and the logistic regression results, heavily featuring characteristics of the trip options generated by the app. Though these models do not estimate the direction of a given variable's effect, we describe them in this way based on the previous descriptive and logistic regression analyses. A greater walk distance to the first leg of the optimal public transit option and more wait time for public transit predicted a greater likelihood of tapping TNC. Having a greater trip duration on public transit was also predictive of tapping TNC. We also see that overall trip distance still predicted tapping TNC. As seen above, longer distance was associated with a lower likelihood of tapping TNC, which we believe is because these longer trips are largely associated with commuter rail. In each model, the sixth and least prominent of the potentially meaningful predictors was wait time for TNC. If we look back to the descriptive statistics and logistic regression, it would appear that sessions with longer wait times for TNC were more likely to have a tap on TNC. This might be because trips that originate in more remote places - which are also poorly served by public transit - require more wait time for a TNC driver to arrive. Last, the logistic regression models found that if the optimal transit trip contained commuter rail, the user was more likely to tap on a TNC-containing option. This was not an important variable in the other models, however, so we regard this finding cautiously.

One of the top five most important features for predicting TNC taps was a characteristic of the user, not the session. Both models found that those who more often used the app during commuting hours were more likely to tap on TNC. Importantly, this was independent of whether the session occurred during commuting hours, which had negligible importance in the models, suggesting that it indicates more about how the rider's historical habits are associated with future decisions. Apart from this, however, no other non-session characteristics were important predictors. Most strikingly, this includes time of day, time of week, and daily precipitation. This last finding contrasts with previous studies suggesting that

variances in the physical environment could contribute to mode shift (Hyland et al. 2018). Instead, it appears that the characteristics of the session, and to a lesser extent the rider, are more powerful in influencing this decision.

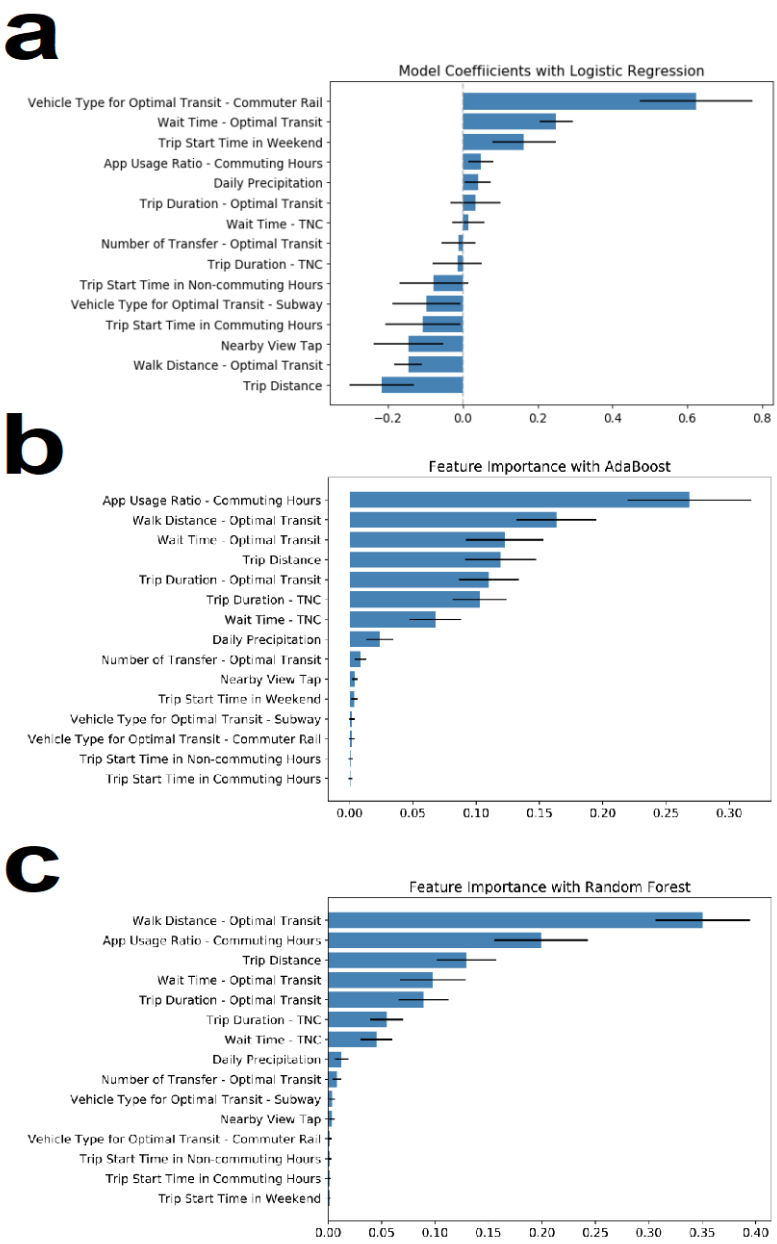

Figure 3.4: Model coefficients for Logistic Regression in (a) and feature importance for all predictor variables included in the models, as determined by the (b) AdaBoost and (c) random forest. For each algorithm, we select only 75\% of the samples and assign the class weight to each instance based on the inverse ratio of the corresponding label to balance the label distribution. We perform 100 iterations and report both the means and standard deviations for all the metrics.This page left blank intentionally.

4.0 Implementation and Technology Transfer

A major focus of the project was to develop processes that not only make the Transit data accessible to analysis, either for research or real-time forecasting, but to codify these processes with replicable code and formalized documentation of the resultant data structure and content. These products were then delivered to OPMI and MBTA so that they could continue to utilize the Transit data past the end of the project. Here we describe the data processing steps and the products that they generate. This information is also spelled out in greater detail in Sections 7.2 and 7.3 in the Appendix, which contain the readme document that accompanies the syntax and the data documentation, respectively.

4.1 Accessing and Preparing Transit Data for Analysis

Transit serves up the data generated by the app in JSON format, wherein each element is a nested JSON object containing a user's activity for that day. This format is efficient for storage but difficult for analysis: (1) querying for a slice of the data spanning several days (e.g., a column across all user sessions) requires a full read and parse of each daily data dump; (2) querying for a specific subset of all columns for a specific set of users similarly requires a full read and parse; the data must be flattened into data frame-like structures before being used statistics software packages. We developed a process for "flattening" or "unfolding" the daily data dumps and loading them into a SQLite database which is both portable and familiar to data scientists. This is described in greater detail in Section 7.2 in the Appendix, the readme document for utilizing the syntax. Here we summarize the main components.

Quick Start advises the user on how to initiate the database and any access keys needed for sourcing information. *Update the Database* explains how to update an already established version of the database. *Transit App Data Structure* describes the structure of the data set when downloaded with some guidance of how these translate to the contents of the final tables generated by the flattening process. *Data Unfolding Process* describes how the nested data structure of each user's data for one day is flattened into records that are then added to the SQLite database. This process leverages a variety of unique identifiers at multiple levels of organization (e.g., sessions nested within users). During this process, multiple additional variables not already present in the data are calculated (e.g., *walk_distance* is a sum of the walking distances of all legs of a single route occurring in a single session). *Technical Notes* summarizes additional details that users should anticipate. *External Data* describes the processes by which we have integrated NOAA's NCDC daily weather data and GTFS real-time arrival times data into the database. These were merged in using spatiotemporal information from the session. The latter content was especially requested by OPMI as giving fuller context to rider decision-making. Trip-Planning Data Processing describes a later addition to the database, though one that forms the basis of the

research study presented here. Data from the trip planning interface did not become available until February 2020, which means we incorporated it at a later stage in the project.

4.2 Data Tables and Variables

After the flattening process, the resultant data structure consists of five tables, each describing a different set of objects or actions. A sixth table containing weather records provided by NOAA is also included in the final output. *sessions* contains information on the unique user session, including the device ID and the start and end times. *trip_views* contains information about any trip presented to the user when entering a search into the trip planner interface, including the unique route ID, the total walking time and distance, total time for the trip, number of transfers, and whether the user tapped on it. *favorite_locations* is a list of stored locations from a user's profile, including the latitude, longitude, and type of location (e.g., home, work). *nearby_views* contains information about each route presented to the user in the nearby views interface, including the unique route ID, whether the user has stored that route as a favorite, and the number of times they have tapped on it historically. *legs* contains a list of all the legs for a given trip, including the mode, distance, and wait time. Importantly, *trip_views* and *nearby_views* are nested in *sessions* though they have additional unique identifiers because a single session can have multiple searches in the trip planner or multiple updates to the nearby routes. *legs* are nested within *trip_views*. Last, *daily_weather* contains the amount of precipitation on the session's date. The *feature_extraction.ipynb* from the repository will read the *.db* file and extract the variables for the predictive modeling of mode choice and the *classficiation_model.ipynb* provides demo code for the pipeline of machine learning analysis. These tables and their contents are described in more detail in Section 7.3 of the Appendix.This page left blank intentionally.

5.0 Conclusions

Data generated by the Transit app are a novel resource for understanding how the riders of public transit navigate their various transportation options. This project set out to do three things: (1) establish a process for making the data accessible for research and analysis; (2) interpret the data and create new variables based on this interpretation to enhance the value of the data; and (3) conduct a research demonstration that illustrated the utility of the data. We discovered early in the project that it would be difficult to confirm whether the data were representative of the transit-riding population that the MBTA serves. As such, it was necessary to concentrate on the features of the database that were sufficiently distinctive to make it stand out from other data sources, and whose analysis might be robust to concerns of representativeness. The best candidate for this was the app's unique ability to capture the real-world decision-making of riders as the consider multiple travel options. This sort of information is not available from other data sources. Further, it permits within-individual analyses that can overcome certain types of representational bias because we are more easily able to establish how individual users tend to prioritize certain characteristics when choosing between routes (e.g., mode of transportation, travel time, number of transfers).

With the focus on real-world decision-making between multiple travel options, we identified a key research question for demonstrating the value of the Transit app database: when and why do riders of public transit select TNC? This question has been of interest to transportation officials, advocates, and others ever since Uber, Lyft, and others became commonplace options for travel. Some have argued that they act as competition for public transit, whereas others have seen them as complementary. The Transit data set acted as an ideal mechanism for testing this question. The remainder of this section walks through the implications of those research findings, but we note that it is just one of many studies that might be conducted on mode choice through Transit's data.

5.1 Research Discussion

The current study used a unique database generated by the Transit navigation app to examine how riders choose between TNC and public transit options when navigating greater Boston. Where previous work has relied on surveys of riders and aggregated ridership data, the app captures rider decisions in real time. We leveraged "taps" that users of the app made on TNC and public transit service options as a proxy for which mode they selected. We then combined these decisions with extensive contextual data stored by the app - including characteristics of the route and the user, time of day and week, and weather conditions - to better understand when, where, and why riders select one mode over the other. A series of descriptive analyses and machine learning models revealed patterns of both competition and complementarity between TNC and public transit, adding evidence and nuance to both sides of an ongoing debate.

The more prominent discussion in the literature has been about how TNC might draw riders away from public transit, especially because TNC offsets or altogether avoids the sorts

of issues that deter transit ridership in general, such as long wait times, long walks to reach a station, and transfers between vehicles (Eluru, Chakour, and El-Geneidy 2012; Bovy and Hoogendoorn-Lanser 2005; Yan, Levine, and Zhao 2019). Such work has often been conjectural, though surveys have found that when the same trip would be substantially shorter by TNC than public transit riders are more likely to opt for the former (Rayle et al. 2016; Habib 2019). In this light, the results of our machine learning models not only reinforce these perspectives but give them greater depth. They found that characteristics of the available public transit options are the most important in predicting whether a rider opts for TNC. When the wait time for transit, the walking distance to the nearest station, and the trip duration on public transit were greater, users of the Transit app were more likely to tap on TNC. It is important to note that these models compared the TNC option to the shortest of the public transit options with the fewest transfers, which we assumed to be the one that riders would be most likely to concentrate on when making deciding between modes. Notably, the total distance of the trip was negatively associated with selecting TNC. This might be better understood when considering that all models also considered expected time length of trip, in which case it could point to the fact that the overall convenience advantage of TNC is what matters, not the distance alone. That is, given a long-distance trip, if the time duration on transit is not so great, then TNC is less attractive; but if a trip takes longer relative to its distance, TNC is more attractive. Overall, given the detailed nature of the data and the fact that they capture riders' decisions when navigating the transit system in real time, these results are likely the most direct confirmation of how and when TNC competes successfully for public transit riders to date.

The machine learning models did generate two additional results that went beyond the characteristics of the TNC and public transit options themselves. First, users who regularly ride public transit during commuting hours were more likely to tap on TNC. This might reflect a certain facility with the app and the varied options that might be available for transit not as present in those who do not use the app as part of their daily routine. Second, other contextual factors, including time of day and week and precipitation, were not predictive of TNC selection. This was somewhat surprising given other literature that suggests that such factors play a substantial role in mode choice (Hyland et al. 2018).

Our results also highlight areas in which TNC services may complement public transit, particularly in addressing the “last-mile problem.” First, the spatial distribution of TNC origins and destinations suggests that people are using TNC service to access large transportation hubs, such as bus and train stations and airports. Further, half of all taps on TNC routes were actually taps on routes in which TNC was carrying the rider either to or from a public transit connection. This indicates that TNCs may help to fill spatial gaps left between public transit service and long-haul shared transportation by inter-city rail and air. Last, as noted above, the use of TNC taps was less frequent for longer trips. As we discussed, this primarily captures riders on commuter rail. But it also might point to one of the downsides of TNC, which is the increased cost of TNCs relative to public transit, which is exacerbated for longer-distance travel (Habib 2019). Users may therefore choose to link shorter trips via TNC (with little cost disadvantage) to longer trips via public transit (where public transit remains drastically cheaper).

As with the evidence on competition between TNC and public transit, this evidence for complementarity between the two modes is some of the first that leverages real-time rider decision-making. Additionally, much of the previous work on complementarity occurred in contexts in which the local transit authority had been working with one or more TNCs to subsidize first-mile and last-mile trips connecting to the public transit system (Terry and Bachmann 2020). This work had been limited by whether such dynamics occurred naturally or were economically sustainable. Because greater Boston's transit system has no such program in place, the results would suggest that such behaviors emerge organically.

5.2 Limitations and Next Steps

There are a few limitations of the study that must be noted. First, we were limited to the information generated by the Transit app. Though this information in many cases was more detailed and precise than previous studies, it lacked certain measures that have been in use. For instance, the data did not include wait time of transfers, which in previous research was found to be a major factor influencing mode choice. This prevented us from verifying its effect, though we were able to verify that number of transfers is a significant factor. Second, it is important to reiterate the interpretation of the Transit data and therefore the study as a whole. The Transit app is positioned as a tool for navigating public transit systems around the world, in which context it informs users of all alternative options. Users of the app, then, are presumably individuals who ride public transit with some frequency. However, the demographics of Transit app users are not guaranteed to match the demographics of transit riders, especially during the pandemic. Further, when riders use the app they are probably already inclined to use public transit. The results here in turn should not necessarily be interpreted as describing the frequency and tendencies with which the whole population of a metropolitan region weigh TNC and public transit against each other so much as when and why regular public transit riders opt to ride TNC. The critical factors are sufficiently consistent with previous research to suggest that they are probably relevant more broadly, but this caveat should be kept in mind.

A third limitation was the overall efficacy of the machine learning models. Specifically, despite decent accuracy for all algorithms used in this paper, they all suffered from moderate-to-poor levels of recall and precision. Practically speaking, this means that Transit app data is not sufficient to generate reliable forecasting mechanisms. While such forecasting would be ideal, there are a number of reasons why it would be difficult to achieve with these data and maybe in general. First, as noted above, we were limited to the measures generated by the Transit app, meaning that the models could not capture the full range of influences on individuals' behavior. These might include environmental features of the mode choice that we do not examine here, such as current traffic, or individual-level features, such as a rider's socioeconomic status or the purpose of their trip. Second, the class imbalance that is inherent in transportation mode choices given the lower prevalence of TNC trips relative to public transit trips may lead to poor accuracy regardless of any unobserved features that could be incorporated into these models. Though we attempt to account for this, imbalance by incorporating down-sampling and weighting techniques to ensure the evenness of data label distribution prior to model training, these problems remain. These results collectively

indicate that such decision-making processes might be more complex to model. While more complex modeling strategies might be fruitful for improving our models' predictive ability, future researchers would be wise to keep in mind that the limitations of measurement and class imbalance might make TNC choices a poor target for predictive modeling.

Another limitation is the absence of cost due to the unavailability of such data for TNC trips due to contract limitations between Transit and TNC companies. The economic aspect of trip mode choice is an important feature which could be a great extension of this research especially during the pandemic. Finally, we note that the data for this study came almost entirely from the period following the onset of COVID-19 in greater Boston, which may limit the generalizability of the results. Namely, individuals might be behaving differently during this period than they typically would when deciding between transportation modes. However, our methodological choices at least partially assuage these types of concerns. In our models, we controlled for the different stages of the shutdown and reopening and found them to have no meaningful effect on mode choice decisions. Because these types of models incorporate both direct and interaction effects, this would mean that the trajectory of the pandemic had little effect on which variables impacted the likelihood of TNC selection. It may also be true that the pandemic has affected ridership behaviors in durable ways such that our data may be more indicative of future patterns than pre-pandemic data would have been. In any case, as we have already noted, the results were sufficiently consistent with the existing theory on the subject that it would not seem as though they were an artifact of the pandemic.

5.3 Final Conclusion

The Transit app data provide a unique view into the factors that influence a rider's choice between public transit and TNC, generating new evidence for both the competition and complementarity between these two modes of transportation. Policymakers and other decision-makers pursuing optimal resource allocation in the realm of transportation would do well to heed these results. The results validated several previously studied features of transit that may be driving mode shift from public transportation to TNC, but also add greater nuance and detail than previous studies. In highlighting features of the transportation network that drive users to choose TNC over public transit, we demonstrate the potential danger for future disinvestment in public transit to exacerbate existing trends in mode shift. Rather than improving efficiency, such policy choices may serve only to push more people towards cars and away from shared buses and trains as a means of transportation. There is some promise, however, in capitalizing on the natural intersections of TNC and public transit that increase the overall efficiency and effectiveness of the transit system as a whole. Further exploring these dynamics as people and transportation systems adapt to shifting services and policies will be critical for shaping the transportation systems of the 21st century. Last, this is one illustration of the potential that Transit app data can have to advance our understanding of the ways public transit riders make decisions when selection between travel options.

6.0 References

Acheampong, Ransford A., Alhassan Siiba, Dennis K. Okyere, and Justice P. Tuffour. 2020. "Mobility-on-Demand: An Empirical Study of Internet-Based Ride-Hailing Adoption Factors, Travel Characteristics and Mode Substitution Effects." *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2020.102638.

Archer, Kellie J., and Ryan V. Kimes. 2008. "Empirical Characterization of Random Forest Variable Importance Measures." *Computational Statistics & Data Analysis* 52 (4): 2249–60. https://doi.org/10.1016/j.csda.2007.08.015.

Bovy, Piet H.L., and Sascha Hoogendoorn-Lanser. 2005. "Modelling Route Choice Behaviour in Multi-Modal Transport Networks." *Transportation*. https://doi.org/10.1007/s11116-004-7963-2.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Carrel, Andre, Anne Halvorsen, and Joan Walker. 2013. "Passengers' Perception of and Behavioral Adaptation to Unreliability in Public Transportation." *Transportation Research Record*. https://doi.org/10.3141/2351-17.

Chawla, Nitesh V. 2009. "Data Mining for Imbalanced Datasets: An Overview." In *Data Mining and Knowledge Discovery Handbook*. https://doi.org/10.1007/978-0-387-09823-4_45.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2008. "A Practical Guide to Support Vector Classification." *BJU International*.

Chow, William, David Block-Schachter, and Samuel Hickey. 2014. "Impacts of Real-Time Passenger Information Signs in Rail Stations at the Massachusetts Bay Transportation Authority." *Transportation Research Record*. https://doi.org/10.3114/2419-01.

Clewlow, Regina R., and Gouri S. Mishra. 2017. "Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States." *Institute of Transportation Studies, University of California, Davis*.

Curtis, Terra, Meg Merritt, Carmen Chen, David Perlmutter, Dan Berez, and Buffy Ellis. 2019. *Partnerships Between Transit Agencies and Transportation Network Companies*. *Partnerships Between Transit Agencies and Transportation Network Companies*. https://doi.org/10.17226/25425.

Eluru, Naveen, Vincent Chakour, and Ahmed M. El-Geneidy. 2012. "Travel Mode Choice and Transit Route Choice Behavior in Montreal: Insights from McGill University Members Commute Patterns." *Public Transport*. https://doi.org/10.1007/s12469-012-0056-2.

Feigon, Sharon, and Colin Murphy. 2018. *Broadening Understanding of the Interplay Between Public Transit, Shared Mobility, and Personal Automobiles*. *Broadening Understanding of the Interplay Between Public Transit, Shared Mobility, and Personal Automobiles*. https://doi.org/10.17226/24996.

Freund, Yoav. 1999. "Adaptive Version of the Boost by Majority Algorithm." In *Proceedings of the Annual ACM Conference on Computational Learning Theory*. https://doi.org/10.1145/307400.307419.

Guda, Harish, and Upender Subramanian. 2017. "Strategic Surge Pricing and Forecast Communication on On-Demand Service Platforms." *SSRN Electronic Journal*.

https://doi.org/10.2139/ssrn.2895227.
Guo, Hongyu, and Herna L. Viktor. 2004. “Learning from Imbalanced Data Sets with Boosting and Data Generation.” *ACM SIGKDD Explorations Newsletter*. https://doi.org/10.1145/1007730.1007736.
Habib, Khandker Nurul. 2019. “Mode Choice Modelling for Hailable Rides: An Investigation of the Competition of Uber with Other Modes by Using an Integrated Non-Compensatory Choice Model with Probabilistic Choice Set Formation.” *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/j.tra.2019.08.014.
Hall, Jonathan D., Craig Palsson, and Joseph Price. 2018. “Is Uber a Substitute or Complement for Public Transit?” *Journal of Urban Economics*. https://doi.org/10.1016/j.jue.2018.09.003.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. *Springer*. Vol. 1. https://doi.org/10.1007/b94608.
Huang, Chen, Yining Li, Chen Change Loy, and Xiaoou Tang. 2016. “Learning Deep Representation for Imbalanced Classification.” In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2016.580.
Hyland, Michael, Charlotte Frei, Andreas Frei, and Hani S. Mahmassani. 2018. “Riders on the Storm: Exploring Weather and Seasonality Effects on Commute Mode Choice in Chicago.” *Travel Behaviour and Society*. https://doi.org/10.1016/j.tbs.2018.05.001.
Inc, Uber Technologies. 2020. “Uber Is Now in over 10,000 Cities Globally.”
Malalgoda, Narendra, and Siew Hoon Lim. 2019. “Do Transportation Network Companies Reduce Public Transit Use in the U.S.?” *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/j.tra.2019.09.051.
Mardia, K V. 1974. “Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies.” *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*.
Murphy, Sharon Feigon and Colin. 2016. *Shared Mobility and the Transformation of Public Transit*. *Shared Mobility and the Transformation of Public Transit*. https://doi.org/10.17226/23578.
Peng, Chao Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. 2002. “An Introduction to Logistic Regression Analysis and Reporting.” *Journal of Educational Research*. https://doi.org/10.1080/00220670209598786.
Rayle, Lisa, Danielle Dai, Nelson Chan, Robert Cervero, and Susan Shaheen. 2016. “Just a Better Taxi? A Survey-Based Comparison of Taxis, Transit, and Ridesourcing Services in San Francisco.” *Transport Policy*. https://doi.org/10.1016/j.tranpol.2015.10.004.
Schapire, Robert E. 2003. “The Boosting Approach to Machine Learning: An Overview.” In . https://doi.org/10.1007/978-0-387-21579-2_9.
Scott, A. J., D. W. Hosmer, and S. Lemeshow. 1991. “Applied Logistic Regression.” *Biometrics*. https://doi.org/10.2307/2532419.
Shaheen, Susan, and Nelson Chan. 2016. “Mobility and the Sharing Economy: Potential to Facilitate the First-and Last-Mile Public Transit Connections.” *Built Environment*. https://doi.org/10.2148/benv.42.4.573.
Smith, Aaron. 2016. “Shared, Collaborative and On Demand: The New Digital Economy.” *Pew Research Center*.

Terry, Jacob, and Chris Bachmann. 2020. “Spatial Characteristics of Transit-Integrated Ridesourcing Trips and Their Competitiveness with Transit and Walking Alternatives.” *Transportation Research Record*. https://doi.org/10.1177/0361198120909842.

Uber Technologies Inc. 2019. “Uber Announces Results for Fourth Quarter and Full Year 2019.” *Online*. https://www.sec.gov/Archives/edgar/data/1543151/000154315120000005/uberq419earningspressrelea.html.

Vladimir Svetnik, *,†, † Andy Liaw, † Christopher Tong, ‡ J. Christopher Culberson, § and Robert P. Sheridan, and Bradley P. Feuston‡. 2003. “Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.” https://doi.org/10.1021/CI034160G.

Westervelt, Marla, Joshua Schank, and Emma Huang. 2017. “Partnerships with Technology-Enabled Mobility Companies: Lessons Learned.” *Transportation Research Record*. https://doi.org/10.3141/2649-12.

Yan, X., Jonathan Levine, and X. Zhao. 2019. “Integrating Ridesourcing Services with Public Transit: An Evaluation of Traveler Responses Combining Revealed and Stated Preference Data.” *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2018.07.029.

Zhao, Xilei, Xiang Yan, Alan Yu, and Pascal Van Hentenryck. 2020. “Prediction and Behavioral Analysis of Travel Mode Choice: A Comparison of Machine Learning and Logit Models.” *Travel Behaviour and Society*. https://doi.org/10.1016/j.tbs.2020.02.003.

Zoepf, Stephen, Stella Chen, Paa Adu, and Gonzalo Pozo. 2018. “The Economics of Ride-Hailing: Driver Revenue, Expenses and Taxes.” *MIT Center for Energy and Environmentla Policy Research*.

This page left blank intentionally

7.0 Appendices

7.1 Attempts to Validate Transit App User Sample

We used an established technique for inferring home locations from mobility data: identifying the location at which a person has the most activity between the hours of 8 pm and 5 am. We pursued two forms of validation. First, we could attach these home locations to census geographies and compare to the residential population. Though transit ridership is not necessarily evenly distributed across the population, we would expect some level of consistency between these measures. Second, approximately 20% of Transit app users have stored a home location, allowing for a direct comparison between known and inferred home location.

These analyses provided conclusive evidence that home locations could not be inferred from Transit app data in this way. We focus here on the results for Suffolk county, which contains Boston, though similar results were found for Norfolk and Middlesex counties. The density of known home locations correlated with population reasonably strongly (r = .59; see Figure 7.1), but when we conducted the same analysis for inferred locations the correlation dropped precipitously (r = .28; see Figure 7.2). These relationships are represented in Figure 7.1.

We concluded that activity between 8 pm and 5 am for an app for navigating public transit might not be an appropriate time window for inferring home locations. We tried again using 5 am to 9 am, anticipating that some people might use the app before leaving home in order to plan their commute. This time, we validated directly against stated home addresses for those who had them saved in their account. We found that the median inferred home address was over 2 km from the known home address of the same individual, meaning the inferences were still far too error prone to allow us to determine where the Transit app population lived, thereby severely limiting our ability to validate its population distribution, whether against census population or ridership statistics.

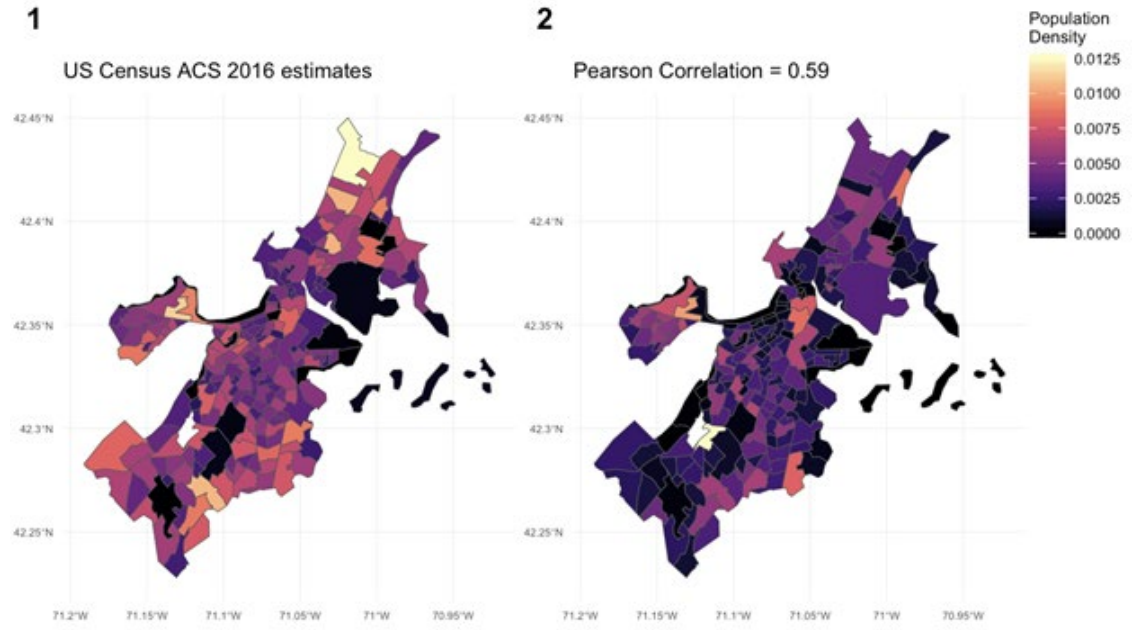

Figure 7.1: Comparison of the Density of Actual Residents (left and Known Home Locations from Transit App Accounts (right) in Suffolk County.

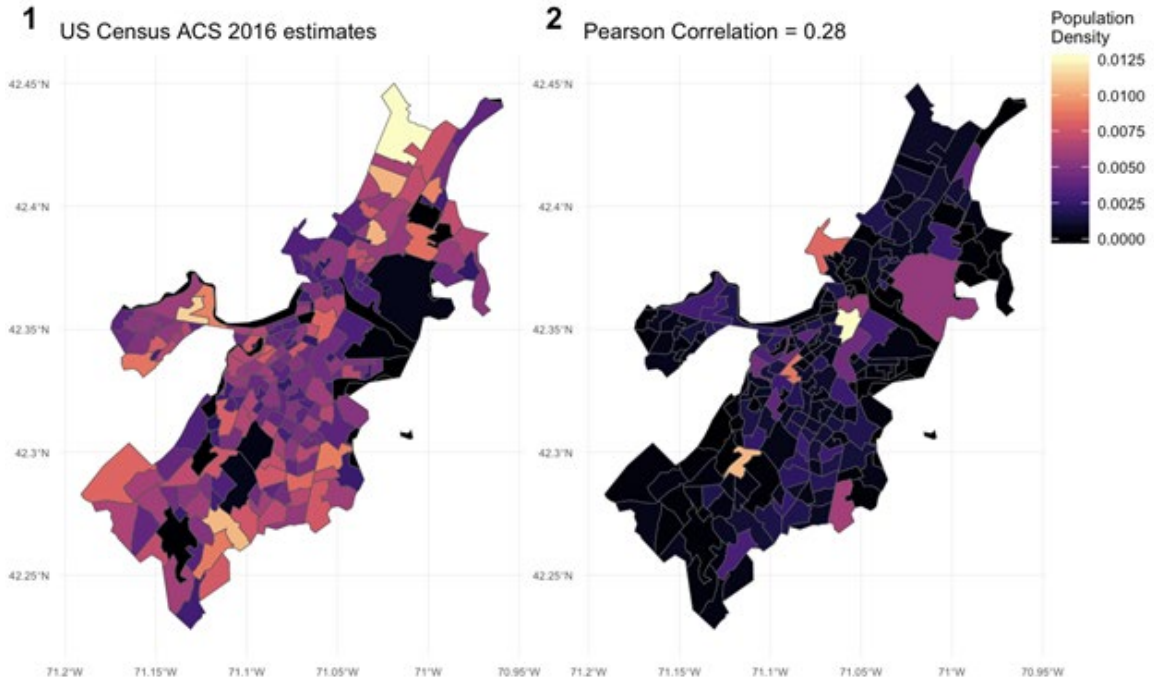

Figure 7.2: Comparison of the Density of Actual Residents (left) and Inferred Home Locations from Transit App Usage (right) in Suffolk County.

7.2 Accessing and Preparing Transit Data for Analysis

This repository from GitHub contains code that facilitates the flattening of data from the Transit App into a database that can be easily queried. In its raw form, each Transit App data dump exists as a single JSON array, where each element is a nested JSON object containing a user's activity for that day. This format is somewhat difficult to interface with for analysis:

- Querying for a slice of the data spanning several days, i.e. a feature / column across all user sessions, requires a full read and parse of each daily data dump;
- Querying for a specific subset of all columns for a specific set of users similarly requires a full read and parse;
- The data must be flattened into data frame-like structures before being used statistics software packages.

The code in this repository attempts to address all of the above issues by "flattening" or "unrolling" the daily data dumps and loading them into a SQLite database which is both portable and familiar to data scientists.

7.2.1. Quick Start

For an initial database setup:

1. Create a copy of config.ini.template and rename it to config.ini.
2. Credentials for accessing the Transit App's data feeds have been included in config.ini. Double check that you can log into the data portal with these credentials. Also, if weather data is to be imported, register an NCDC API key and paste it into config.ini. A URL to the registration page has been included in config.ini.
3. Run bootstrap.py, which will download all available data from the Transit App using the URL and credentials specified in config.ini.
4. Run init_db.py to import the data into the database.
    - The default behaviour of init_db.py is to import all daily JSON dumps in the transitapp-data/transit directory. To restrict the import to a subset, simply pass the desired files as command-line arguments.
        - e.g. to import all data on or after 2020 February 28: find transitapp-data/transit/20*.json | tail -n +303 | xargs ./init_db.py
5. Optionally, run weather.py to scrape, import, and create a crosswalk for NOAA NCDC daily weather data.
6. The initialized database will be available at transit.db.

7.2.2. Update the Database

To update the database:

1. Run bootstrap.py.
2. Make a backup of the database. Rationale for this can be found below.
3. Run init_db.py.
4. If updating weather: drop the tables daily_weather and daily_weather_crosswalk then run weather.py.

The database can be updated incrementally with bootstrap.py and init_db.py; imported Transit App files are kept track of with a table in the database and will not be imported again. However, to save on space improve import performance, the database **will not keep track of incomplete imports**. In these cases, the database must be rebuilt. Because of this, it is recommended to back up the database before updating with init_db.py. To update the weather data, the tables daily_weather and daily_weather_crosswalk must be dropped before running weather.py again. Downloaded data is stored locally to avoid redownloading.

7.2.3. Transit App Data Structure

In this section, a brief summary of the Transit App data structure will be presented. Words appearing in bold appear as separate tables in the final database.

The structure of each days' data is in the form of one giant JSON array. Each element of this array corresponds to a user comprising several sessions, which are the primary data unit, and also favorite_locations, which are essentially a user's bookmarks. Favorite locations correspond to a user's bookmarks at the time that the dump occurred and can change between dumps, within users.

A single session corresponds to all activities that occurred between a user opening an app and the app being closed and cleared from the phone's memory. It is important here to note that there is a distinction between simply leaving the app and actually closing it, as the former will not end a session and the session will persist in the background until the user returns. As a result, some sessions can end up being disproportionately long. The rate at which background apps are closed is dependent on the phone's operating system and user settings.

Each session consists, among other things, of several trip_views and nearby_views. Each trip view corresponds to a single "trip" planned by the app, which encompasses a unique multimodal route from the user's input origin and destination and consists of several legs. Each leg has an origin, destination, distance, estimated duration, mode of transportation, and a unique identifier linking the public transit route to the MBTA's GTFS feeds, if applicable.

An important feature of the trip views is that trip views do not explicitly include information of when an origin-destination query was made or which views were presented for which query. However, the former can be reasonably deduced from the session start time

and the latter can be reasonably estimated by examining the coordinate pairs of the origin and destination, as a specific pair of coordinates will generally correspond to a specific query.

Nearby views refer to nearby transit routes that are suggested to the user on the home screen, before they start planning their trip. Each nearby view has information about the number of times the route was tapped by the user during that session, whether or not the route was favorited, and an identifier linking to the MBTA's GTFS feeds, if applicable.

In both sessions and nearby views, ridesharing and bikesharing options are also presented, if applicable.

For a full list of all available variables and types, please refer to the variable SQL_TABLE_INIT in init_db.py, which contains the database schema. This can also be accessed by opening the database in an SQLite database exploration software or by typing .schema in the sqlite3 command line. For more detailed information about each column, please refer to the PDF documentation.

7.2.4. Data Unfolding Process

Overview

The data unfolding process, which is handled by init_db.py, essentially walks through the nested data structure of each user's data for one day and inserts flattened records into the SQLite database. Each element of each array of interest corresponds to a single row in the database.

For elements that are arrays, a unique MD5 hash is generated as a unique identifier for each element within the array by hashing together data that, when combined, are unique for that particular element. The MD5 hash function allows us to reduce each heterogenous and variable-length tuple of values into a single string. For example, a session can be uniquely identified by the user and start time alone, so those are what are hashed.

The unique identifiers are unidirectional in the database, pointing up from the lower levels. For example, legs embed trip view identifiers, trip views embed session identifiers, etc.

In addition to unique identifiers, some other additional data is generated in the unfolding process:

- Several variables in the **trip_views** table are generated as aggregates of constituent **legs**, such as walk_distance which is a sum of the distances of all legs.
- All durations, appearing as e.g. travel_time in the **trip_views** and **legs** tables, have been generated. For legs, this corresponds to the end_time of each leg minus the start_time; for trip views, this corresponds to the end_time of the last leg minus the start_time of the first leg.
- All indicator variables, such as has_rideshare and has_transit in **trip_views**, have been generated.

Technical Notes

Data unfolding takes place in the process_user function of the init_db.py script. Tables have been separated by indentation levels corresponding to loops that unfold each nested array. An SqlRecord object is created for each element and appended to an empty list generated at the start of each user's parsing procedure; the list grows as that user's data is parsed. At the end of parsing, the records are inserted, in the order that they were generated, into the database.

The use of arrays of SqlRecord objects here is to allow for parallel processing of JSON files but sequential ordering of records into the database, which is necessary for preserving the integrity of the data and allowing links across tables. Essentially, users are processed on several different threads, and chunks corresponding to users' processed data are passed back to the main thread in a FIFO queue where they are consumed one-by-one. This eliminates a condition where two users could be simultaneously processed and their data inserted in a staggered manner. This consumption pattern is not a performance issue since the parsing of the JSON data is the main bottleneck.

7.2.5. External Data

Weather

In addition to the data provided by Transit, we have additionally created facilities for merging National Climatic Data Center daily weather data into the database, for assessing route preference under different weather conditoins. Both the scraping and inserting of weather data is handled entirely by the weather.py script, which queries the NOAA NCDC API.

The script additionally generates a crosswalk that links session identifiers to the weather data by comparing session start times to NCDC dates. This is done with an SQL query that is embedded in weather.py. This script can be run without any user input. The default behaviour is to scrape all data from January 1st, 2020 to the present day, though the start date can be changed by changing the GHCND_START_DATE variable.

7.2.5.2 GTFS-realtime Arrival Times

The wip_init_trip_updates.py script is a proof-of-concept showing how enhanced GTFS-realtime archives from the MBTA could be flattened. The behavior of the script is to parse the GTFS-realtime feeds in the transitapp-data/gtfs directory, where GTFS-realtime GZIP-compressed data archives should be dumped (this can be changed by changing the target of the glob that defines the files variable) and outputs flattened data to transitapp-data/gtfs-processed. Similarly to the weather data, a crosswalk for this could be created with a SQL query that compares start times and route identifiers.

7.2.6. Trip-Planning Data Processing

The processing code is available at the *feature-extraction.py.* The main objective is to select all the candidate trip queries where they satify all of the following conditions:

- The user has at least performed taps on trips with both Transit and TNC at least once
- For each trip query, the user was presented with trip candidates containing modes
- The user has to show preference toward one mode over another (i.e more taps)

The code in this repository attempts to address all of the above criteria across all the trip view feeds generated from the previous unflattening process. One notable thing is that, even under the same session, a user could specify multiple origin/destination pairs. Hence we re-assign the trip-id prior to identifying the optimal trips with both modes and extracting the corresponding information as mentioned in Section 2.2. The detailed extraction/aggeration principles are listed below and the code in the repository provides step-by-step extraction code as well as the merge between different features to generate the following variables:

User-level Features

- User's primary usage purpose: the ratio of sessions occuring during different time windows: Weekday Commuting/Weekday Non-commuting/Weekend (***Numerical***)
- User's overall frequency of app usage: the average number of sessions in last week (***Numerical***)
- User's overall tendency of the tapping act: the ratio of sessions with taps to the total number of sessions of the user (***Numerical***)

Trip-level Features

Note: To ensure all spatial and temporal information belong to one single trip rather than extracted from multiple trips, we first define the 'optimal' trip for both Transit and TNC, for both modes. If any trip feed was tapped it automatically becomes an optimal trip candidate. If more then one trip was tapped for each mode, we use the one with the shortest duration from the trips has the minimum transfers. If none of the trips with the specifc mode was tapped, similarly, we select one with the shortest duration from the trips has the minimum transfers from the untapped trips to reflect the 'best' option provided with the corresponding mode.

- User's interaction during the session: Whether the user has tapped on any nearby views during the session, and if yes, if the tapped nearby views contains any of the user's favorite route (***Categorical)***
- Session's start time type: Weekday Commuting/Weekday Non-commuting/Weekend (***Categorical***)
- The initial wait times for both optimal trips (***Numerical***)
- The initial walking distances for both optimal trips (***Numerical***)
- The number of transfers for the optimal Transit trip (***Numerical***)

- The trip durations for both optimal trips (***Numerical***)
- Trip Distance computed based on Origin and Destination information (***Numerical***)
- Vehicle Type of the primary mode(s) of optimal Transit trip (e.g. Bus vs. Subway vs. Commuter Rail (***Categorical***)

Additional Predictors

- Weather information such as the temperature and actual hourly percipitation during the beginning of the session (***Numerical***)
- Stage of the shutdown based on local authority's guidance (***Categorical***)

7.3 Transit App Data

This data documentation describes the schema and contents of the Transit app data once "flattened" or "unfolded" into tabular formats to facilitate analysis. It contains a subsection describing the variables contained in each of the resultant tables. For each, the unique identifier variables that create links across the files (or reflect nesting of one table's rows in the rows of another table) are noted.

7.3.1. sessions

Sessions are unique periods of usage of the app, from opening of the app to closing of the app.

Variable Name	Type	
session_id	TEXT	unique identifier for each session **(PRIMARY KEY)**
device_id	TEXT	unique identifier of the device **(FOREIGN KEY)**
start_dt	INTEGER	start time of the session
end_dt	INTEGER	end time of the session

Table 7.1: List of variables for table: sessions

7.3.2. trip_views

Trip views are all trips presented to a user when conducting a search in the trip planning interface. Trip views are nested within sessions as a given session might include one or more uses of the trip planning interface.

Variable Name	Type	
trip_view_id	TEXT	unique identifier for each trip view **(PRIMARY KEY)**
session_id	TEXT	unique identifier for of the session **(FOREIGN KEY)**
orig_lon	REAL	longitude of origin
orig_lat	REAL	latitude of origin
dest_lon	REAL	longitude of destination
dest_lat	REAL	latitude of destination
mbta_services	TEXT	the global-route-id of all legs
vehicle_types	TEXT	the vehicle type of all legs
walk_distance	REAL	total walk distance of all legs (m)
walk_distance_transit	REAL	the initial walk distance before the first transit leg (m)
walk_time	REAL	the sum of legs with walking (sec)
travel_time	REAL	the sum of all trip legs (sec)
transfers	INTEGER	number of transfers
longest_leg_mode	TEXT	the mode of the longest leg
longest_leg_vehicle_type	TEXT	the vehicle type of the longest leg
tapped	INTEGER	if tapped is 1, else 0
session_start_to_first_vehicle_time	REAL	the time between the session start time and the first vehicular leg
planning_start_to_first_vehicle_time	REAL	the time between the desired departure time and the first vehicular leg
has_transit	INTEGER	if contains Transit mode leg 1 else 0
has_rideshare	INTEGER	if contains TNC mode leg 1 else 0

Table 7.2: List of variables for table: trip views

7.3.3 favorite_locations

Favorite locations are a part of an individual user's account data.

Variable Name	Type	
favorite_id	TEXT	unique identifier for each record **(PRIMARY KEY)**
device_id	TEXT	unique identifier of the device **(FOREIGN KEY)**
address	VARCHAR	address of the favorite location
name	VARCHAR	name chosen by the user to represent this location
favorite_type	INTEGER	type of the favorite location which is a numerical value and there are ten unique values and all records contain such information. Most common entries are home and work.
latitude	REAL	latitude of the favorite location
longitude	REAL	longitude of the favorite location
locality	VARCHAR	locality of the favorite location (such as Boston, Cambridge)
sub_locality	VARCHAR	sub-locality of the favorite location (such as downtown, south end)
postal_code	VARCHAR	postal code of the favorite location
country_code	VARCHAR	country code of the favorite location

Table 7.3: List of variables for table: favorite locations

7.3.4 nearby_views

Nearby views are the routes presented to a user as being nearby (based on GPS-derived location) when the user opens the app. Nearby views are nested within sessions as a user is presented with multiple nearby routes in each session.

Variable Name	Type	
nearby_view_id	TEXT	unique identifier for each nearby view **(PRIMARY KEY)**
session_id	INTEGER	unique identifier of the session **(FOREIGN KEY)**
rideshare_category	TEXT	specifying tnc options (lyft, uber etc)
global_route_id	TEXT	global_route_id of the mbta service shown in the feed if applicable
is_lyft	INTEGER	binary variable indicating whether the nearby view option is by Lyft
is_uber	INTEGER	binary variable indicating whether the nearby view option is by Uber
is_favorite	INTEGER	start time of the session
tap_count	INTEGER	end time of the session

Table 7.4: List of variables for table: nearby views

7.3.5 legs

Legs are the components of a given trip presented by the trip planner interface. Legs are nested within trip views (which, in turn, are nested within sessions), as any given trip generated by the trip planner will have one or more legs.

Variable Name	Type	
leg_id	INTEGER	unique identifier for each leg **(PRIMARY KEY)**
trip_view_id	TEXT	unique identifier of the trip view which the leg belongs to **(FOREIGN KEY)**
leg_idx	INTEGER	leg index within each trip view sorted by the departure time
global_route_id	INTEGER	global route id of the leg if applicable
mode	TEXT	mode of the leg (transit or tnc)
vehicle_type	TEXT	vehicle type of the leg
travel_time	REAL	travel duration of the leg
distance	REAL	travel distance of the leg
wait_time	REAL	the start time of the current leg minus the previous leg for any vehicular leg

Table 7.5: List of variables for table: legs

7.3.6 daily_weather

Weather data are accessed from NOAA's NCDC and merged based on the date of the session (see Section 7.2 for merging process).

Variable Name	Type	
weather_id	INTEGER	unique identifier for each session (**PRIMARY KEY**)
date	TEXT	unique identifier of the device
precipitation_inches	INTEGER	end time of the session

Table 7.6: List of variables for table: daily weather

7.4 Descriptive Statistics of Trip View Data

To provide background on patterns of usage, we present a series of descriptive statistics from the trip view table. This analysis is based on the data obtained immediately after the database was constructed and prior to the feature extraction process. Figure 7.3 shows the top 10 most frequently appeared routes where we see that 7095 (Red Line), 7532 (Green Line) and 7094 (Orange Line) are the top candidates along with the combinations of them. The bus routes appear to be less frequent which could be due to the large number of bus routes compared to subway. Figure 7.4 shows that if we consider only the vehicle type, then bus becomes the most dominant which is most likely caused by the large number of unique bus routes.

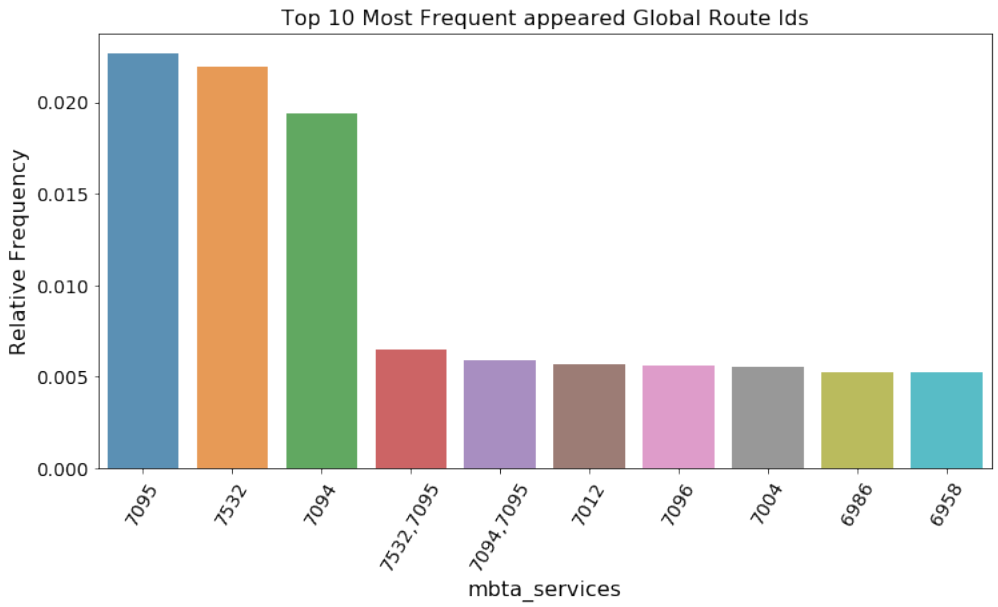

Figure 7.3: Frequency of Trip Global_Route_IDs from Trip View Table

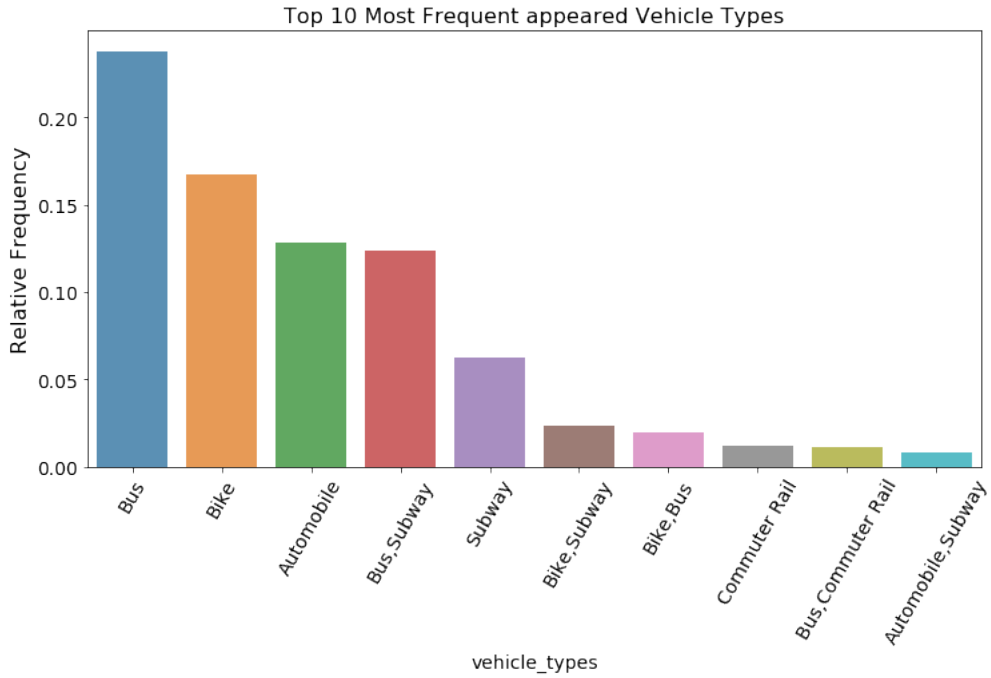

Figure 7.4: Frequency of Vehicle Types from Trip View Table

Figures 7.5 and 7.6 reflect the overall relative frequency of the trip's longest mode. Over 40% of the trips have a public transit mode as the longest leg while the same proportion is 12% for TNC. The most common vehicle type is bus with 22% while for subway is 11%.

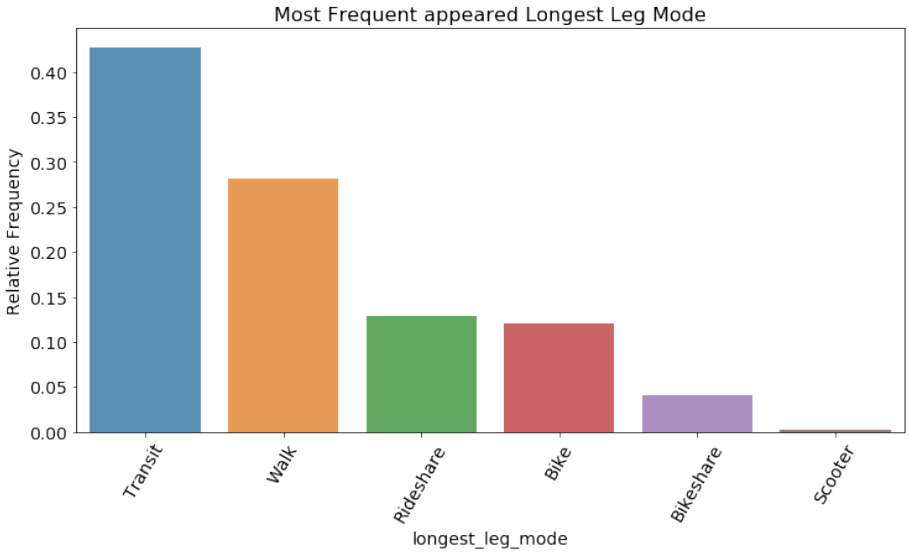

Figure 7.5: Frequency of Longest Leg Mode from Trip View Table

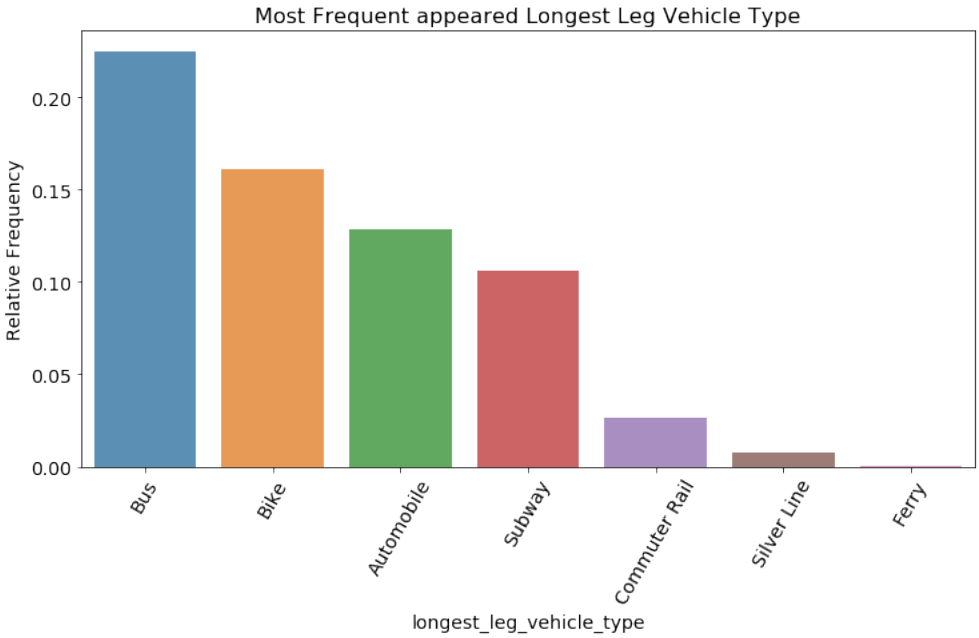

Figure 7.6: Frequency of Longest Leg Vehicle Type from Trip View Table

Figure 7.7 show the percentage of three binary variables included in the table. Where we can observe that 16% of the trips has rideshare service included whereas 67% include public transit. The overall percentage of trips with taps is 7%. Table 7.7 summarizes the quantile information and mean value for the numerical variables. Note these values were obtained after removing extreme or false values (negative value for a distance or time).

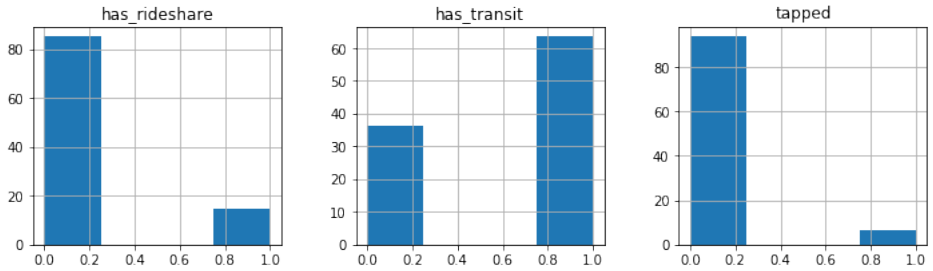

Figure 7.7: Frequency of the Binary Variables from Trip View Table

Numerical Column	Median	Mean	Std	25%	75%
walk_distance (m)	382.9	682.5	1063.3	13.1	906.7
walk_distance_transit (m)	161.4	313.1	437.6	42.3	429.6
walk_distance_rideshare (m)	0	0	135.6	0	0
walk_time (s)	606.0	617.2	303.5	371.0	862.1
travel_time (s)	1887.0	2203.8	2285.3	1261.0	2742.0
transfers	4	3.5	1.7	2	4
session_start_to_first_vehicle_time (s)	856.0	1191.8	1106.2	445.0	1542.0
planning_start_to_first_vehicle_time (s)	795.0	1146.7	1132.7	378.0	1502.0

Table 7.7: Statistics of Numerical Variables from Trip View Table

7.5 Supplementary Information of Analysis

7.5.1. Database Component

As described in the main text, the data used for this paper come from the Transit app, with records of user's interactions with several different aspects of the app stored in separate forms. The main data used in the paper are the trip planning features of the app, which allow us to impute both origins and destinations of users given their input.

Other components of the data include the “nearby view" feature, which is what most users see when first opening the app. This feature allows users to see transit stops and upcoming arrivals in their close vicinity, as well as other alternative transportation options such as bikeshare or TNCs. Our records of users' interactions with this feature of the app include users' taps on each line of transit (or non-transit) transportation option shown to them in this view. Similar to the main data described earlier, we store the records of users' interactions with the ``nearby views" feature in a database consisting of a flattened version of the raw JSON data managed with SQLite3 via Python.

7.5.2. Tap Number Comparison

Figure 7.8 shows that amongst all trip queries, during the vast majority (94.51%) of queries a user only taps on transit-based trips. There is no query during which a user taps on only TNC trips, this might be due to the fact that our data are from Transit App rather than any TNC App, through which many users who intend to use TNC services likely originate their trips. Taps on TNC trips are greater than taps on transit trips for 3.59% of sessions with taps. There are only 1.24% where the total tap count for both modes are equal and we remove those trip queries to capture clear inclinations.

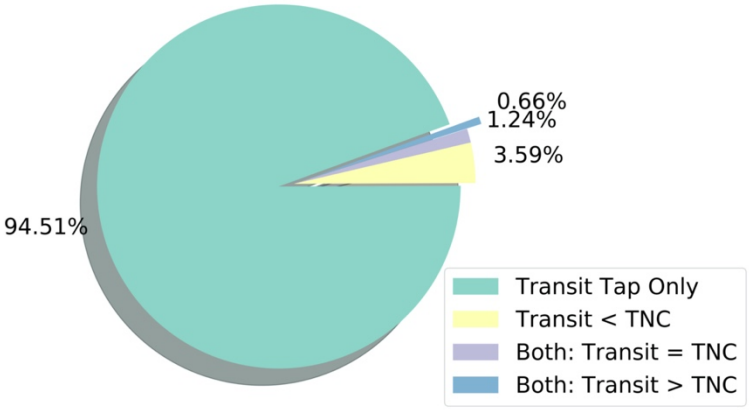

Figure 7.8: Comparison of the Number of Taps on both Transit and TNC across all Queries

7.5.3. Multi-mode Trips with Taps

Another initial finding based on the trip data screening is that a large portion of individuals who have tapped on TNC trips tend to tap on the trips that have a combination of both Transit and TNC services. Based on the historical taps, 50.77% of the tapped trips contains TNC are multi-mode trips and across all users who have previously expressed their interests toward the TNC services, 50.4% of them are solely interested in the multi-mode trips.

This further suggests that a significant fraction of potential riders who select a TNC-containing option from the Transit app are using the TNC service as the first or last mile to the transit stop. Figure 7.9 reveals the spatial and temporal features of trips with both modes. Where it can be observed that the overall hot spots are similar to the TNC-only hot spots including South Station and Logan Airport. The majority of the trips are less than 20km from the city. There are some additional new destination spots such as the Southeastern areas to the city. The temporal distribution shows slight peak around 3pm across the study period. Figure 7.10 shows the ratios between the duration and distance of the TNC portion to the Transit portion, and the ratios are close to 0.25 for both metrics highlighting that the Transit service acts as the major portion of such trips.

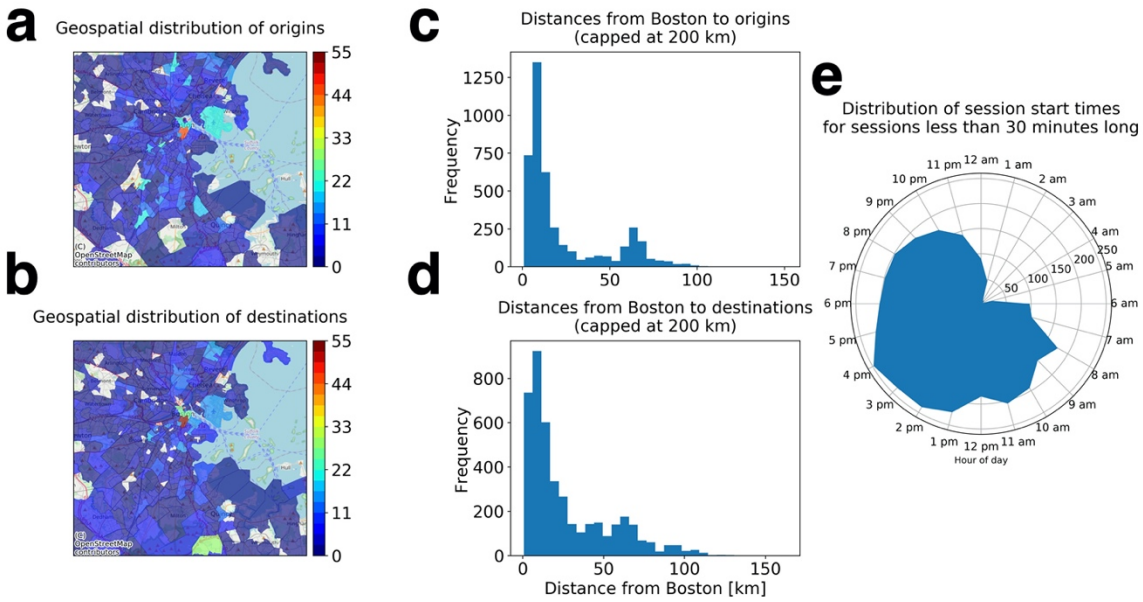

Figure 7.9: Comparison of the Number of Taps on both Transit and TNC across all Queries

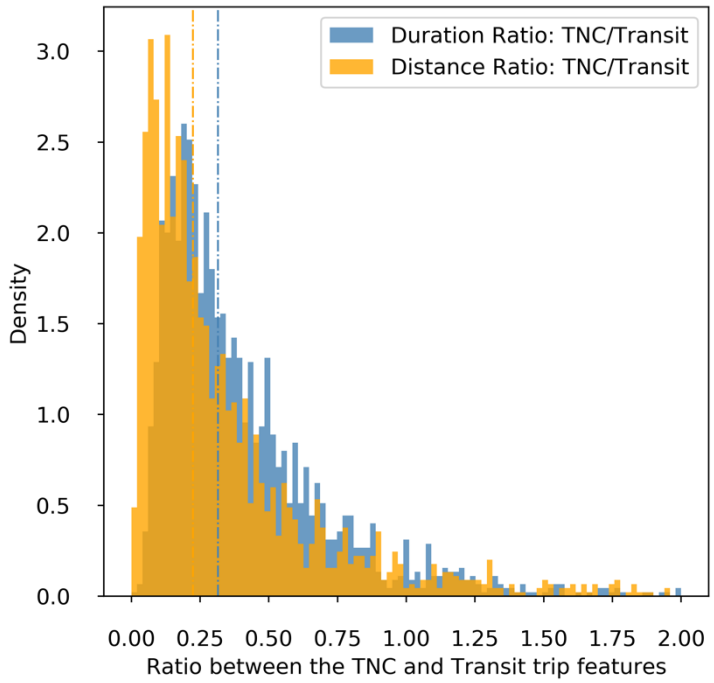

Figure 7.10: Comparison of the Number of Taps on both Transit and TNC across all Queries

7.5.4. Feature Transformation

For all optimal trip features extracted, we visualize the histogram and compute the skewness (Mardia 1974) as shown in Fig 7.11. We use a threshold of 1 as threshold (Mardia 1974) and perform a log-transformation for any variable has a skewness value greater than 1 which is shown in Fig 7.12. We can see that most heavy-tailed features appear to be more bell-shaped after the transformation with significant decrease on the skewness value. However, the walk distance feature appears to be an exception hence we use the original scale for modeling and analysis.

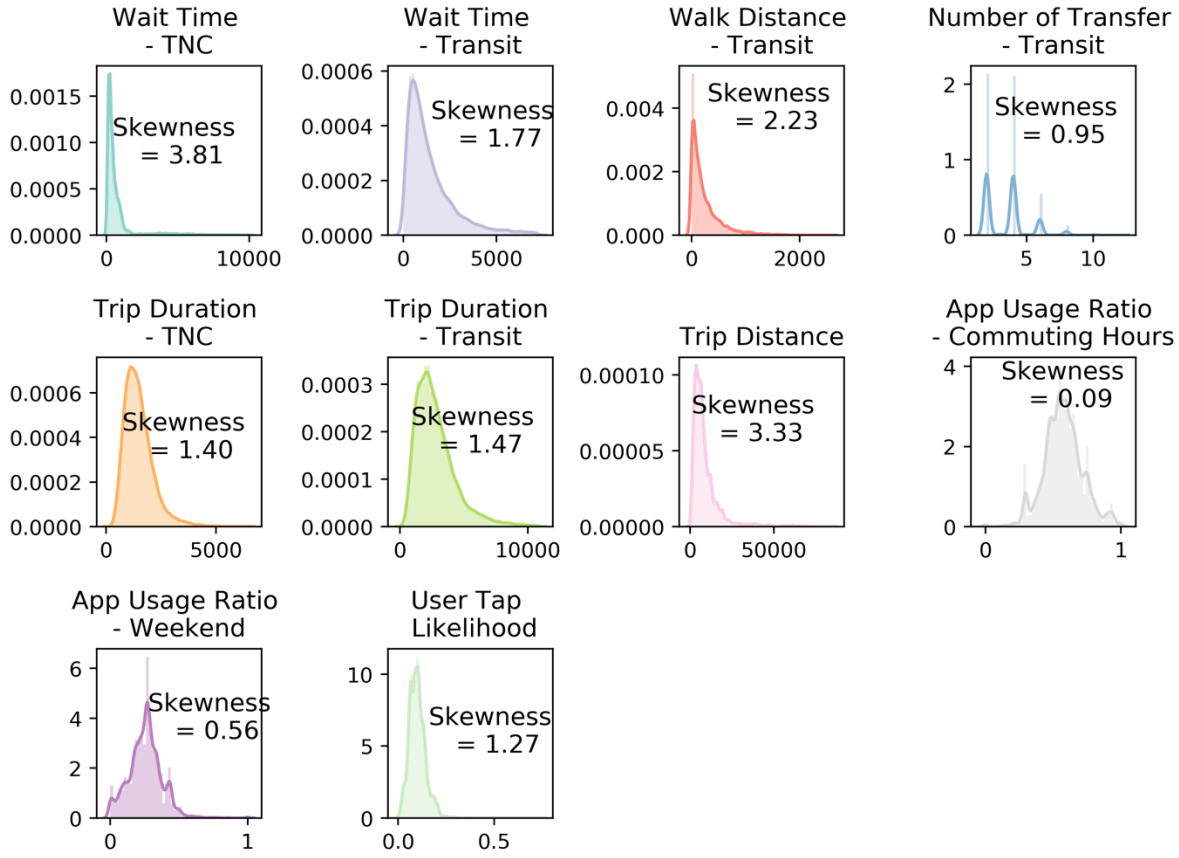

Figure 7.11: Distributions of Values for Original Input Features

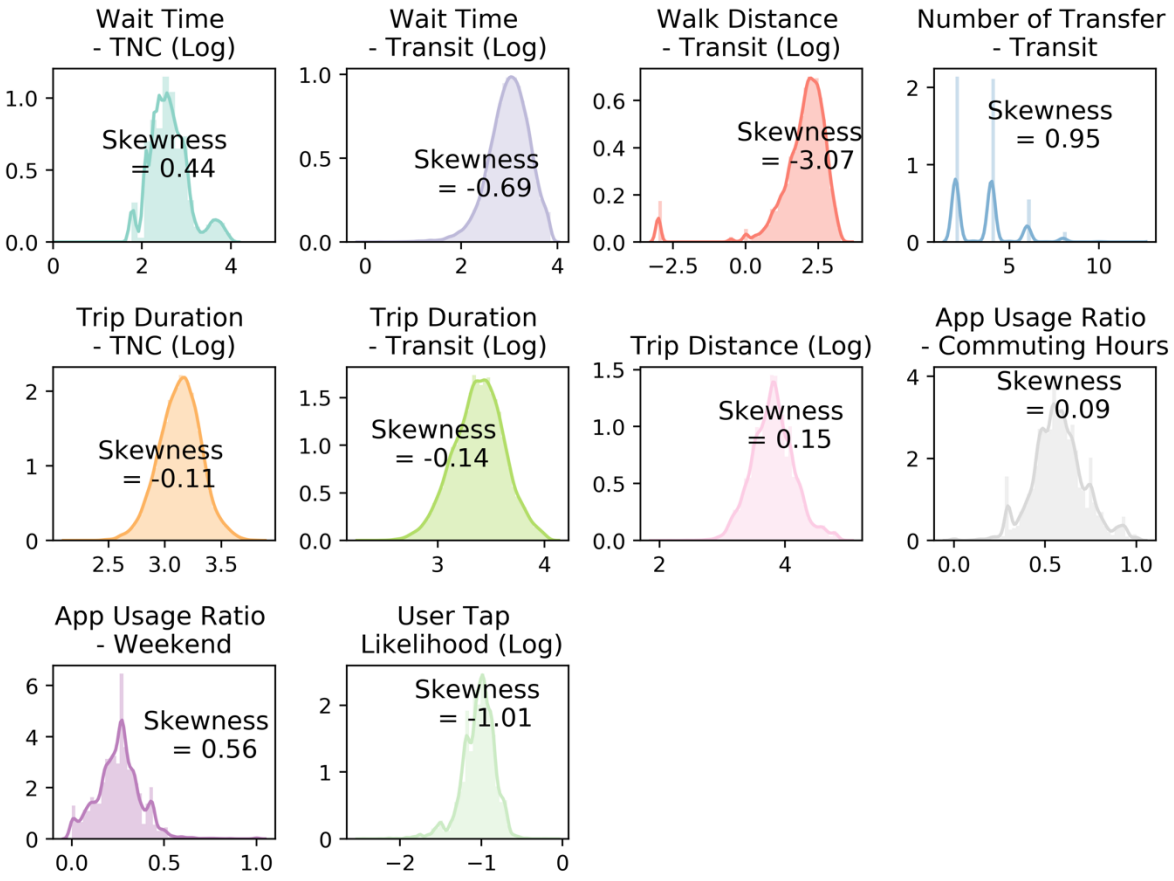

Figure 7.12: Distributions of Values of Input Features with Log-transformation on Selected Features

7.5.5. Precipitation Data

Figure 7.13 summarizes the daily and weekly rainfall based on weather station's historical readings. We can observe some peaks during certain times which are likely caused by storms and such events are quite sparse rather than continuous.

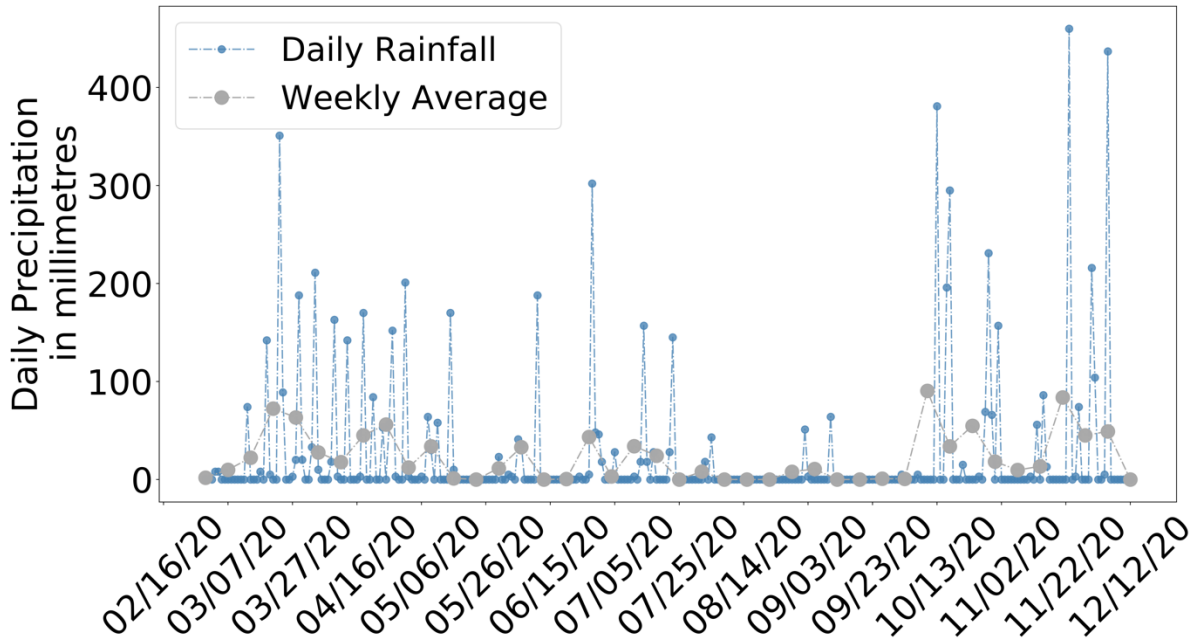

Figure 7.13: Time Series of the Historical Precipitation Data

7.5.6. Temporal Progression of Tap Tendencies

Figure 7.14 shows the temporal progression of daily trip taps from the Transit App users during the study period. The *y*-axis represents the ratio to the annual median value on each day of the week for different type of trip taps (daily value divided by the annual average day-of-week value). We observe that there is a decrease following the national emergency declaration responding to the COVID-19 pandemic, which is visible for both daily and weekly patterns and the usage recovered between late July and late October as the reopening takes place. The overall temporal progression trends for both TNC and Transit are similar to each other with ascending and descending trends at the same intervals.

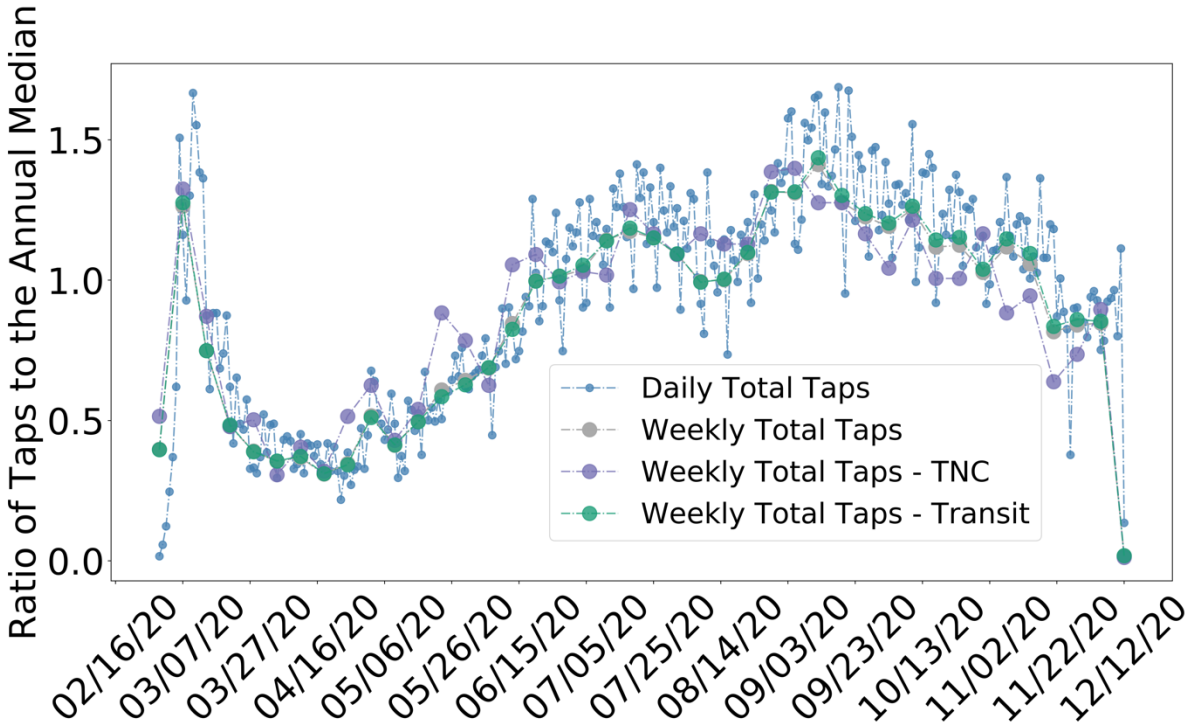

Figure 7.14: Time Series of the Relative Trip Taps on Transit and TNC modes

7.5.7. Correlation Heatmap

We then move towards a predictive approach to more systematically analyze users' choices between TNC and public transit. First, we plot the correlations between the features we use in our predictive models. Figure 7.15 plots the Pearson's correlations between each pair of features used in our models and described earlier. We see that all trip, user, and environmental features of users' trips appear to have no or only weak linear correlation with the output variable (tap choice): no correlation has an absolute value greater than 0.05. However, some trip features are collinear with one another. The difference in trip duration between TNC and public transit is correlated positively with the number of transfers and the trip distance.

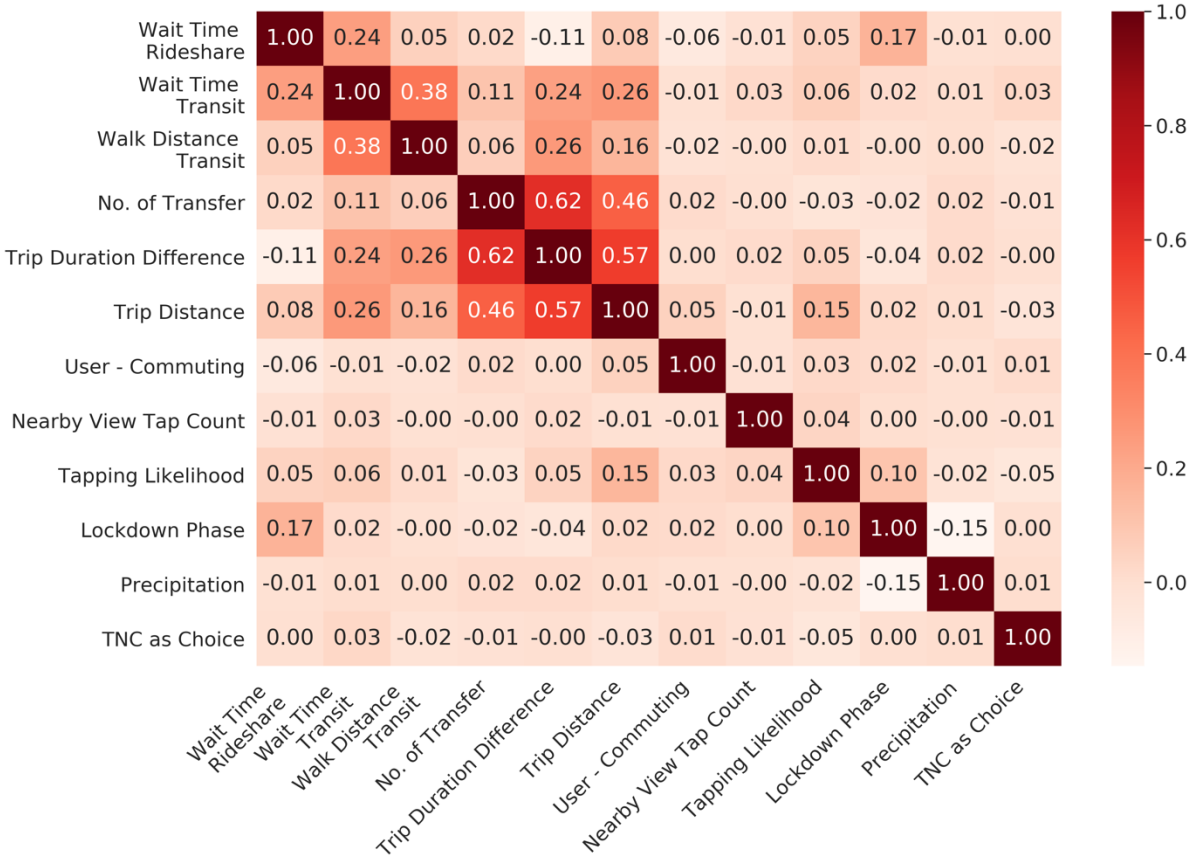

Figure 7.15: Pearson Correlation Heatmap of Predictive Features