Evaluating Bike Share as A Solution to the Last Mile Problem in Public Transit: An Alternative Approach*

Bobae Kang†

Spring 2017

(version first complete draft)

Abstract

In recent years, the worldwide success of bike share programs has attracted much scholarly attention. Researchers have highlighted particularly the potential contribution of bike share to solving what is known as the last mile problem in public transit. However, few studies have explored the connected use of bike share and other public transit services using observational data. The current study contributes to the growing body of research on bike share by examining the connected use of bike share and other public modes of transit. To that end, this study introduces an original notion of potentially multi-modal (PMM) trips as an alternative measure for the effect of public transit accessibility to bike share trips, thus challenging the typical measure based solely on the distance between bike stations and public transit stops. In particular, using 2016 data from Chicago's Divvy program, this study tests the reliability of 300-meter distance as the proximity standard, previously used to assess the effect of public transit accessibility to Divvy trips. The model estimation results suggest that PMM trips with the 300-meter proximity standard constitute a heterogeneous set of trips, thereby offering a ground for reconsidering the use of 300-meter standard to evaluate the effect of access to public transit to Divvy trips.

keywords: Bike share, Divvy Chicago, Last mile problem, Multinomial logit model, Potentially multi-modal trips, Rational choice theory.

*This paper is my term paper for Perspectives in Computational Research course in Spring Quarter 2017 as well as the first complete draft of my thesis for the MA Program in Social Sciences (MAPSS) at the University of Chicago. I express my sincerest gratitude to the thesis advisor and the course co-instructor Benjamin Soltoff, Ph.D. for his instruction and support throughout the process of completing this draft. I also give thanks to the other course co-instructor Richard W. Evans, Ph.D. and the MAPSS preceptor Theresa Anasti, Ph.D. (expected 2017) for their guidance.

†The University of Chicago, MA Program in the Social Science, bobaekang@uchicago.edu.

1 Introduction

Bike share is perhaps one of the most exciting developments in the realm of public transportation in recent years. In the last two decades, bike share has spread rapidly throughout the world. Currently, bike share programs operate in over 800 cities with approximately 950,000 bikes, evolving into a major component of the public transportation networks worldwide (Fishman, 2016). The remarkable success of bike share has attracted much attention from academic researchers and, although bike share is a relatively new topic for academic research, the scholarly literature on bike share is growing quickly as well. Researchers note that main benefits of bike share include flexible mobility, emission reductions, physical activity benefits, reduced congestion and fuel use, individual financial savings and support for multimodal transport connections (Fishman et al., 2013, p. 2). In particular, researchers have exhibited a great interest in the last of the listed benefits, i.e., further enabling trips that incorporate multiple modes of transportation and are thus multi-modal, and proposed bike share as a viable solution to what is known as the last mile problem.

The last mile problem is one of the central questions in transportation literature, especially for studies on public transit networks. In brief, the last mile problem describes a phenomenon where passengers are less likely to choose public modes of transportation that leave the passengers a "mile" from their destination. It must be noted that the "mile" in this expression does not refer to a literal mile in distance but rather is used as a metaphor. Researchers have proposed various solutions to the last mile problem in public transit, including local shuttle bus systems (Xie et al., 2011), automated vehicles (Yap et al., 2016), and land-use improvement (Zellner et al., 2016). Thus far, the most popular solution among researchers appears to be public bike share systems (DeMaio, 2009; Fishman et al., 2013; Faghih-Imani and Eluru, 2015; Griffin and Sener, 2016). Bike share programs, they argue, can encourage passengers to use the public transit by serving as a convenient as well as affordable last-mile access to and from public transit stations.

However, there are two major issues with the existing argument for bike share as a solution to the last mile problem. First, the empirical support for this argument mostly relies on survey results rather than on the actual bike share uses. Although surveys can generate meaningful insights into the behavior of their respondents, the survey method is only an indirect approach to examining actual behaviors and cannot replace a direct analysis of observational data on the behaviors of interest. In fact, historical trip data is published and readily available for many bike share programs. To find concrete evidence for their argument, researchers must incorporate observational data into their investigation. While some studies using such observational data do exist, these studies do not concentrate on the last mile problem, but mainly examine broader patterns in the bike share usage: herein lies the second issue. Lacking a clear focus on the last mile problem, these studies have failed to provide an effective measure to evaluate bike share programs' contribution to solving the problem. The current study seeks to address these two issues, which constitute a critical gap in the academic literature.

More specifically, the current study challenges a commonly-used measure for evaluating the effect of public transit accessibility to bike share trips, which is based solely on the distance between bike share stations and other public transit stops. In fact, the true effect of public transit accessibility to bike share uses cannot be distinguished from the possibility of multi-modal trips as they are in fact two sides of the same coin. Consequently, a good measure for the effect of the access to public transit is also a good measure for the contribution of bike share to solving the last mile problem in public transit. To that end, the current study proposes an alternative measure, namely, the original notion of potentially multi-modal (PMM) trips. This new approach accounts for additional characteristics of each bike share trip, including the duration of each trip as well as the expected availability of public transits based on not only spatial but also temporal dimensions. These conditions for classifying PMM trips are derived from the rational choice framework in the transportation literature, which assumes that passengers would seek to minimize the cost of their trips. Using this novel approach, the current study examines the case of Divvy, a bike share service local to the Chicago area, with a clear focus on how this Divvy trips relate to the accessibility of Chicago Transit Authority (CTA) buses and rails. The main objective of this study is then to test the validity of the 300-meter proximity standard between Divvy stations and public transit stops, the same standard used in a prior study on Divvy bikes, Faghih-Imani and Eluru (2015), to account for the impact of public transit accessibility on Divvy uses. In other words, this study attempts to answer the following question: Does 300-meter standard offer a reliable measure for the effect of access to CTA buses and rails to Divvy trips as well as the contribution of Divvy to CTA ridership in Chicago?

The quantitative analysis was conducted using Divvy trips made in the year of 2016. Following the City of Chicago's open data policies, Divvy has been publishing biannually its trip and station dataset since its launching in 2013 on the official webpage. For each trip, the dataset offers trip start and end time, trip start and end stations, rider type (annual membership subscriber or 24-hour pass user), and, if it is a trip by a subscriber, the subscribers gender and year of birth. The dataset also includes a separate file specifying the geographic information on each station. The Divvy trip and station database is then augmented with a variety of other data to generate features that may help explain the characteristics of PMM trips. To that end, I combine Divvy data with the timetable for CTA bus routes and rail lines as well as the coordinates of all bus and rail stops, geographical boundaries of 77 community areas as well as the central business district in Chicago, selected demographic features of Chicago community areas drawn from the Census data, and meteorological data. In my modeling effort, a multinomial logistic regression is employed to compare the use of different proximity standards: 50 meters, 100 meters, 200 meters and 300 meters. The estimated coefficients of the fitted model makes possible a comparison across the trips that are newly labeled PMM with each proximity standard.

The model estimation results suggest that trips labeled PMM with the 300-meter proximity standard constitute a heterogeneous mix of trips. For example, trips that are labeled PMM with 50-meter standard is distinct from trips that will be newly labeled PMM with the 100-meter standard, and so on. When combined with the expected, cost-minimizing behavior of passengers and the counter-intuitively high proportion of PMM trips with the 300-meter standard, this finding offers a ground for reconsidering the use of 300-meter standard to evaluate the effect of access to public transit to Divvy trips. The estimation results, however, cannot determine which proximity standard is the most effective one because no data is available on true multi-modal trips, against which the effectiveness of each different proximity standard can be tested. Despite of such a limitation, the current study contributes to research on bike share by providing a novel, and theoretically sounder approach to evaluating of the effect of public transit accessibility, which is in accordance with the rational choice framework.

The rest of this paper is organized as follows: Section 2 offers a literature review on the last mile problem and prior research on bike share. Section 3 explains the notion of potentially multi-modal (PMM) trips and describes the structure of multinomial logit model. Section 4 presents the sample formation procedures and descriptive summary of the Divvy trip sample. Section 5 describes the model estimation results. Section 6 discusses the implications of the estimation results. Finally, section 7 concludes a brief review of the current paper.

2 Literature Review

In the following, I first discuss the last mile problem in the traditional transportation literature and how recent studies on the bike share relate to the last mile problem in public transit. Then I explore the bike share literature and its limitations with respect to quantitatively evaluating bike share as a solution to the last mile problem.

2.1 Last mile problem

Advocates of public transportation struggle with the last mile problem, which states that passengers rarely opt for public transit systems that leave them a mile from their destination without any convenient connecting options. Alternatively, the first mile problem expresses a similar situation in which passengers have difficulties in getting to a transit system from their point of departure. Although most studies tend to focus on the last mile problem, some use these two terms in tandem as in "first and last mile problem" (Griffin and Sener, 2016).

The term last mile originates in the telecommunications and other related industries, where it is used to refer to the final connectivity leg of the telecommunication networks that delivers services to the end users (Wikipedia, 2016). In communication networks, the last mile is typically the main source of limitation on the bandwidth of data that can be delivered to the customers. Moreover, though necessary to connect individual users to the backbone lines of communication networks, the "last mile" constitutes the most challenging part of the system to maintain and upgrade because it interfaces with a variety of end-user equipment.

This expression has since been appropriated by the transportation literature to address a similar phenomenon in the filed of public transit. In the transportation literature, the last mile similarly limits the accessibility of the public transit network to passengers. Transportation scholars and policy makers have yet to find any one-size-fits-all solution to the last mile problem because of the diversity of passenger needs and the resulting high cost of providing passengers with personalized connecting service to the main transit network.

In the transportation literature, the last mile problem is often examined in terms of passengers' transportation mode choice (Tilahuna et al., 2016; Zellner et al., 2016). Here, mode choice refers to what mode of transportation a passenger chooses for each trip. This notion of mode choice comes from the standard "four-step" approach to modeling demand for and performance of a transportation system, which consists of the following components: trip generation, trip distribution, mode choice, and route choice (McNally, 2007).

More formally, researchers typically examine transportation mode choice using discrete choice models, in which passengers choose among multiple alternatives. Commonly used models are various forms of logit and probit models that assumes passengers would choose a mode of transportation maximizing their utility. Under this framework, the last mile problem can be formulated as a feature or a combination of features that negatively contributes to passenger utility which consequently decreases the probability that each passenger will choose the public transit system. Therefore, any “solution” to the last mile problem must sufficiently reduce the negative impact of such negative features of the public transportation system, so that any passenger is more likely to choose the public mode of transportation over other alternatives.

Researchers have proposed various solutions to the last mile problem in public transit, including local shuttle bus systems (Xie et al., 2011), automated vehicles (Yap et al., 2016), and land-use improvement (Zellner et al., 2016). Thus far, the most popular solution among researchers appears to be public bike share systems (DeMaio, 2009; Fishman et al., 2013; Faghih-Imani and Eluru, 2015; Griffin and Sener, 2016; Liu et al., 2012). All solutions seek to make public transit more available to a diversity of passengers, thereby maximizing both utility of passengers and the economic and environmental benefits of the increased use of public transit.

2.2 Bike share

The recent popularity of bike share has drawn considerable attention from researchers, who have examined various bike share programs and their impact. However, a closer look into the literature reveals that earlier studies failed to take advantage of the opportunities to quantitatively analyze the bike trip data while more recent, quantitatively oriented studies have not focused on the particular question of how bike share contribute to the solution the last mile problem in public transit. In the following, I provide a brief overview of the history of bike share as well as the academic literature on bike share to provide a proper context for the current study.

There have been at least three generations of bike share system since its beginning in the 1960s (DeMaio, 2009; Shaheen et al., 2010). First-generation programs simply provided with ordinary bicycles for public use. Without any means or structure to ensure people’s responsible use, first-generation public bikes were often stolen and vandalized. As a result, White Bike Plan in Amsterdam, Netherlands, which began in July 1965, was brought to its end soon after its launch. Second-generation bike share programs, born in Denmark in the early 1990s, introduced the use of docking stations and coin-deposit system. Arguably the most famous of the second-generation programs is Copenhagen City Bikes, the world's first large-scale organized bike share program that launched in 1995 with 1,000 bicycles. Despite the improvements, however, the second-generation programs were still susceptible to theft due to the inevitable anonymity of their users.

Third-generation bike share programs are characterized by the use of information technology to track bikes and collect user information. The incorporation of information technology has helped third-generation programs to effectively prevent bike theft, which constituted a major concern of the previous generations of bike share, as riders no longer remain anonymous. Another innovative components of the third-generation models include paid membership system and "smart technology" used for bike check-in and checkout (Shaheen et al., 2010). With these new innovative features, the third-generation model has spread worldwide and the number of cities and municipalities offering bike share programs has increased to over 800 currently (Fishman, 2016).

The soaring popularity of bike share has attracted scholarly attention to the subject. Fishman et al. (2013) and Fishman (2016), together, provide an overview of the relevant literature since the early 2010s. The scope of research ranges from user preferences (Shaheen et al., 2013; Fishman et al., 2014) to characteristics of actual usage (Buck et al., 2013; Fishman et al., 2014) to safety concerns (Fishman et al., 2012) to impacts and policy implications (Fishman et al., 2014) to practical issues such as mandatory helmet uses (Fishman, 2012) and rebalancing (**?**Fishman et al., 2014). As noted by Fishman et al. (2013), however, many previous studies become quickly outdated due to the rapid growth in the industry.

Faghih-Imani and Eluru (2015) identifies two broad perspectives based on which the quantitative studies on bike share are conducted: first, a systems perspective, and second, a user perspective. The former refers to a set of studies that investigate into the determinants of bike share usage, which is generally characterized by arrivals and departures. Such studies evaluated the significance of bike share infrastructure (e.g., the number, location and capacity of docking stations), environment (e.g., the presence of businesses, public transit stations, and schools), weather (e.g., temperature and precipitation), and other temporal characteristics (e.g., time of day, day of week, and season of the year) in terms of bike share usage. The latter is exemplified by another set of studies focused on how the behavior of users may change in response to bike share programs. These studies examine the impact of different user types (e.g., socio-demographic characteristics, convenience of access, and purpose of usage) as well as policy changes (e.g. opening of new stations, and encouragement of usage) on bike share usage.

Meanwhile, researchers also took note of the potential contribution of bike share with respect to the last-mile problem in public transportation. DeMaio (2004) is one of the early studies that highlighted the positive impact of bike share programs on improving the last mile and first mile connections to other modes of transportation. Although DeMaio (2004) does not use the term "last mile", the paper speculates that bike share programs, or public bikes in its original expression, "may increase trips on other modes of public transportation, as they expand the reach of trains and buses."

In the following years, more studies argued for the contribution of bikes share problems to addressing the last mile problem, providing some empirical evidence. For example, DeMaio (2009) cites two surveys conducted by the City of Paris on the users of its bike share program, called Vélib'; put together, these surveys suggest a growth in the number of bike share riders who use the service to reach or return from the public modes of transportation[1]. Fishman et al. (2013) introduce more evidence from other cities, including Dublin, Ireland, London, UK, and Washington, D.C., that bike share programs may serve as a solution to the last mile problem[2].

As noted by Faghih-Imani and Eluru (2015), not many studies have attempted quantitative evaluation of the actual bike share uses and the integration between bike share and other, traditional modes of public transportation. Instead, researchers often resorted to various survey methods. More recently, however, this trend has

[1] "In 2008, 21 percent of survey respondents used Vélib to reach the subway, train, or bus, and 25 percent used Vélib on the return trip from other transit modes. In 2009, 28 percent used Vélib to begin and to end their multi-leg transit trip" (DeMaio, 2009, p. 45).

[2] Empirical studies on bike share also note that different types of bike share riders have different trip purposes. In fact, only a small portion of bike share riders appear to use the services mainly to support their public transport uses.

changed as bike share programs began to make their trip data publicly accessible. Faghih-Imani et al. (2014) examines trip data of BIXI, the first major bike share program in Montreal, Canada, to identify factors contributing to the increased usage of bike share in the area on the bike station level. Ma et al. (2015) investigates another case of bike share program, namely, Capital Bikeshare (CaBi) in Washington, D.C. Griffin and Sener (2016) develops a multi-method approach that combines "descriptive statistics, plan evaluation techniques, and semi-structured interviews of bike share system planners" for two bike share programs in Chicago, Illinois (Divvy) and Austin, Texas (B-cycle). Finally, Faghih-Imani and Eluru (2015) attempts to adapt the random utility maximization approach used in the transportation literature to the trip-level bike share (Divvy) data.

Although these studies all account for the impact of public transportation on bike share trips and the bike share users' choice of destination, only Griffin and Sener (2016) explicitly focuses on bike share programs' potential impact on facilitating transit trips. In other studies, the existing public transit system constitutes simply another feature that might affect each bike trip. However, the method Griffin and Sener (2016) uses to analyze bike share data involves hardly more than a mere comparison between the number of bike share embarks within 400 meters from a rail station and that of all other trips, thus failing to offer any statistically rigorous analysis of the data.

More importantly, none of the four studies cited above provides a convincing reason for their choices of proximity standard, which marks the maximum distance for the existing public transit stations to influence bike trips. In Griffin and Sener (2016), for example, the choice of 400 meters for the proximity standard is simply drawn from a maximum distance suggested by operators for spacing between bike share stations (p. 11), and its usefulness as the proximity standard is not empirically supported. Ma et al. (2015) uses a quarter mile (or approximately 400 meters) as the proximity standard, again, without explaining why. Faghih-Imani and Eluru (2015) construct spatial variables based on the 300-meter buffer, arguing that it is "to be an appropriate walking distance, considering the distances between Divvy stations."

Lastly, Faghih-Imani et al. (2014) suggests a 250-meter buffer as "an appropriate walking distance considering the distances between BIXI stations," but provides no further rationale.

In fact, if we are to examine how bike share programs may serve as connecting options to public transit networks, neither 250 meter nor 300 meter nor a quarter of mile may be an adequate standard. In downtown Chicago, for example, we may find a dozen bus or rail stops within the 300-meter range from a single Divvy station. In addition, the same 300-meter range will include other docking stations as well. Worse still, a great majority of all Divvy stations in downtown Chicago will have at least one public transit station in its "proximity," making it difficult to properly gauge the effect of access to public transit network on bike trips in the area and, by extension, how bike share contributes to solving the last mile problem in public transit. Furthermore, to my best knowledge, no previous studies have accounted for the time dimension with respect to the accessibility of other public transit options as well as the possibility of multi-modal trips. This is a missed opportunity since the bike trip data are often time-stamped, enabling researchers to incorporate the timing of each bike trip into their analysis. Therefore, we need a better measure to appraise whether and if so, to what extent bike share may serve as a solution to last mile problem in public transportation.

3 Model

In the following paragraphs, I first discuss at length the notion of potentially multi-modal (PMM) trips as an alternative measure for the effect of public transit accessibility to bike share rides. I argue for the advantage of using this alternative approach and present how I operationalize it for the current study. Then I describe the multinomial logit model used in this study.

3.1 Potentially multi-modal (PMM) trips

Currently, to the best of my knowledge, no data exist on actual multi-modal trip behaviors connecting Divvy and CTA.[3] This lack of data makes it impossible to identify a Divvy trip that is truly multi-model, i.e., made as a part of a longer trip that also involves riding a CTA bus or rail to arrive at the destination. Because of this obstacle, both researchers and policy makers have difficulties in evaluating the contribution of Divvy bikes to the existing public transit system in Chicago. Until we can directly observe multi-modal trips, therefore, we need to resort to an alternative approach to labeling multi-modal Divvy trips. Previously, some researchers have relied on surveys to understand the patterns and characteristics of multi-modal trips. Other researchers have sought to assess the influence of access to public transit on bike trip behaviors, mostly by including a simple measure of the proximity between the destinations of bike trips and public transit networks in multivariate regression models examining broader patterns of bike trips. Importantly, both approaches are critically limited as they fail to either account for actual trip behavior or examine explicitly the linkage between bike share and public transit uses.

To address this limitation, the current study introduces the original notion of potentially multi-modal (PMM) trips, which can serve as a useful proxy for actual multi-modal trips. The rationale for using PMM trips as a proxy for actual multi-modal trips is as follows: In the transportation literature, most researchers analyze trips using discrete choice models, which assume that each passenger chooses the modes and routes for her trip to maximize utility according to some form of utility function (Faghih-Imani and Eluru, 2015). In this discrete choice framework, a multi-modal trip is a trip in which a passenger chooses to include multiple modes of transportation to travel from the origin to the destination to maximize her utility, or

[3]In fact, in October 2016, the Federal Transit Adminitration selected CTA and Chicago Department of Transportation (CDOT) to receive a $400,000 grant to enhance the connectivity between CTA rides and Divvy uses through modifying the Ventra App, which is the electronic system to pay for CTA services, to provide customers with better access to Divvy bikes (Unrau, 2016). Once this development is completed, CTA and CDOT will be equipped to collect data on actual multi-modal trips. Until then, however, researchers must rely on a proxy for multi-modal trips between Divvy and CTA such as the one proposed in the current study.

net benefit. The passengers utility maximizing behavior suggests that, in planning her trip, the passenger will seek to minimize the temporal gap for transferring from one mode of transit to the next mode since, as shown in previous research, the overall travel time is negatively correlated with the utility of the passenger. Following such reasoning, a trip can be labelled PMM if it exhibits expected properties of a multi-modal trip and, therefore, is likely such a trip.

In this study, a PMM Divvy trip is defined in the following way: A Divvy trip is labeled PMM if the trip starts from or ends at a station in proximity with any CTA stops approximately the time when a bus or a rail arrives at or departs from such stops. In other words, a trip must meet specific spatial and temporal conditions to be labeled PMM. Firstly, the spatial condition involves the proximity between Divvy stations and CTA stops. That is, a trip must originate from or terminate at a docking station located nearby at least one CTA stop. If the distance between the docking station and the closest CTA stop is too long, passengers' utility maximizing behavior suggests that the trip is highly unlikely multi-modal. Secondly, the temporal condition accounts for the estimated temporal gap between the beginning or ending of a Divvy trip and the availability of buses or trains at nearby CTA stops. For example, if the time between the end of a Divvy trip and the beginning of a following CTA ride is too long, it is improbable that they jointly constitute a single multi-modal trip for any utility maximizing passenger.

To account for the temporal condition, I use and compare four distances as different proximity standards, $d \in \{50m, 100m, 200m, 300m\}$. My choice of these standards are based on the following rationale. First, the proximity standard of 300 meters is used in another study on Divvy trips (Faghih-Imani and Eluru, 2015). Given the goal of the current study to test the existing way of evaluating the impact of access to public transit to bike share, it is a matter of course for the current study to include the 300-meter standard. On the other hand, the choices of 200 meters and 100 meters are based on the dimension of a typical block dimension in the city of Chicago, which is 660 feet by 330 feet or, equivalently, 200 meters by 100 meters (City of Chicago, 2007). Finally, the 50-meter proximity standard is added for a more detailed investigation into the consequences of using different proximity standards. For measuring distances, I use Manhattan distance or the taxicab metric, rather than Euclidean distance, in order to account for the fact that streets in the city of Chicago are generally laid out in a grid.

In addition, to account for the temporal condition, I use the window of 1 minute to 5 minutes to decide whether a trip is PMM or not. That is, a Divvy trip that originates from a station after a bus or rail arrived at a nearby CTA stop based on the given proximity standard within the window of 1 minute to 5 minutes is considered PMM. Similarly, a Divvy trip that terminates at a station before a bus or rail departs from a nearby CTA stop based on the given proximity standard within the window of 1 minute to 5 minutes is labelled PMM. This one-to-five-minute window is based on an assumption that when a bike share user makes a multi-modal trip, she must have planned her trip to be multi-modal and will transfer from one mode of transportation to the next without any significant digression in between. That is, for example, if a passenger wants to ride a Divvy bike to a docking station nearby a bus stop in order to catch a bus, she would plan her trip so that she would arrive at the destination shortly before the time the bus is expected at the bus stop. As soon as the first part of her trip is terminated, the passenger would move on to the next part as soon as possible. The exact time span is chosen rather conservatively, in order to minimize the Type I (false positive) error in labeling multi-modal trips. The first one minute is excluded to account for the time a passenger needs to travel between a docking station and a bus stop and dispatch or return the Divvy bike. Accordingly, the one-to-five-minute window is likely to underestimate the number of actual multi-modal trips.

Finally, I make an additional assumption that PMM Divvy trips are no longer than 30 minutes. This limitation on the duration of PMM Divvy trips is based on the following two considerations: First, under the current pricing structure of Divvy program, trips that are longer than 30 minutes cause additional charges for both annual membership and 24-Hour Passes and, therefore, increase the total cost of making multi-modal trips[4]. Second, riding bikes are generally slower than riding buses or rails so it is cost effective for a passenger to travel by bike as little as possible. Therefore, the limitation on the trip duration is fully in agreement with the overall rational choice framework on which the current study stands.

3.2 Multinomial logit model

In the current study, I use the multinomial logistic regression to model PMM Divvy trips, which is a common choice for modeling a categorical response variable with multiple discrete outcomes. Here, the response variable Y is a categorical variable with five discrete possible outcomes, $y \in \{PMM1, PMM2, PMM3, PMM4, None\}$. Each category is defined as follows: $PMM1$ is for trips that are labeled PMM using the proximity standard $d = 50m$. $PMM2$ trips are those labeled PMM with $d = 100m$, excluding $PMM1$ trips. $PMM3$, then, accounts for PMM trips with $d = 200m$ excluding $PMM1$ and $PMM2$ trips, and $PMM4$, for PMM trips with $d = 300m$ excluding $PMM1$, $PMM2$, and $PMM3$ trips.

The proposed way of categorizing Divvy trips has two key advantages over an alternative way of categorizing trips based simply on the different proximity standards to mark them PMM: one is practical and the other is theoretical. First, the proposed categorization ensures that the possible categories of the response variable are mutually exclusive and exhaustive, a necessary condition for fitting a multinomial logit model as described below. Otherwise, trips that are PMM with, for example, $d = 50m$ are also PMM with $d = 100m, 200m$ and $300m$, trips that are PMM with $d = 100m$ are also PMM with $d = 200m$ and $300m$, and so forth. The result is a nested structure for the response variable, which is more difficult to model and certainly not suitable for a multinomial logistic regression. Second, and perhaps more importantly, the proposed method of categorizing PMM trips allows me to focus specifically on the marginal effect of using a more generous proximity standard. Accordingly, with the proposed way of categorizing the response variable, I can compare between trips that are, for instance, PMM with $d = 50m$ ($PMM1$) and trips that become PMM as

[4]Details on Divvy's pricing structure can be found here: https://www.divvybikes.com/pricing

I extend the proximity standard to $d = 100m$ ($PMM2$). Here, $PMM2$ is precisely the marginal effect of switching from one proximity standard, $d = 50m$ to a more generous one, $d = 100m$. Such focused assessment of the marginal effect of extending proximity standard is not feasible with the alternative categorization.

The multinomial logit model assumes a linear relationship between the explanatory variables and the natural log of the odds ratio for the j^{th} category and the J^{th} category. Mathematically, this relationship can be expressed in the following form:

$$\eta_{ijs} = \log\left(\frac{\pi_{ijs}}{\pi_{iJs}}\right) = \mathbf{x}'_{is}\boldsymbol{\beta}_{js}, \tag{1}$$

where π_{ijs} is the probability that the stage s of a trip i is labeled j, for $j = 1, 2, ..., J-1$, or $\pi_{ijs} = \Pr(Y_{is} = j)$. Here, J is the number of all discrete categories, or possible outcomes, of the response variable. For the current study, J equals 5. Note that we omit the J^{th} category for the numerator. This is because, in the multinomial setting, we use the J^{th} category as the baseline category against which we model the log-odds ratio for all the other categories. $\mathbf{x}_{is}$ is the vector of explanatory variables for the stage s of the trip i, in addition to 1 for the constant term. $\boldsymbol{\beta}_{js}$ is the vector of the corresponding regression coefficients, including the constant term. π_{iJs} equals the sum of the probabilities of all $J-1$ categories subtracted from 1. In order to express this more in more explicit terms, the log-odds ratio η_{ijs} in Equation 1 can be rewritten in the following form:

$$\log\left(\frac{\pi_{ijs}}{\pi_{iJs}}\right) = \log\left(\frac{\pi_{ijs}}{1 - \sum_{j=1}^{J-1}\pi_{ijs}}\right) = \sum_{k=0}^{K} x_{iks}\beta_{kjs}, \tag{2}$$

for $i = 1, 2, ..., N$ and $k = 1, 2, ..., K$, where N is the number of observations and K is the number of explanatory variables.

With some algebraic rearrangement, then, we can recover the probability π_{ijs} that the stage s of a trip i is labeled j:

$$\pi_{ijs} = \frac{\exp(\eta_{ijs})}{\sum_{j=1}^{J}\exp(\eta_{ijs})}, \tag{3}$$

for $j = 1, 2, ..., J$. Since the potential outcomes of the explanatory variable Y_{is} are mutually exclusive and exhaustive, the probability π_{ijs} will add up to one for each i and s.

Parameter estimation for the multinomial logit model is based on maximum likelihood estimation. The likelihood function for the multinomial logit model, $L(\boldsymbol{\beta}|\mathbf{y})$, is given in the following form:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1} \exp\left(y_{ijs}\sum_{k=0}^{K} x_{iks}\beta_{kjs}\right)\left(1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=0}^{K} x_{iks}\beta_{kjs}\right)\right)^{n_i}, \quad (4)$$

where $\boldsymbol{\beta}$ is the model parameters vector and $\mathbf{y}$ is the vector of observed values of the response variable. Then, taking the natural log of the likelihood function given in Equation 4, we obtain the log-likelihood function in the following form:

$$\ln L(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{N}\sum_{j=1}^{J-1}\left(y_{ijs}\sum_{k=0}^{K} x_{iks}\beta_{kjs}\right) - n_i \log\left(1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=0}^{K} .x_{iks}\beta_{kjs}\right)\right) \quad (5)$$

Here, the model parameters vector $\boldsymbol{\beta}$ is estimated by maximizing the log-likelihood function in Equation 5.

4 Data

For the current study, I use multiple datasets from different sources to obtain as well as generate key variables. Table 1 lists all datasets used in this study with their sources and brief descriptions. Appendix A-1 of this paper contains detailed information as to how to access each of the datasets and which organizations are involved in collecting and providing the datasets. Appendix A-1 also provides additional tables and visualizations of some datasets, including descriptive statistics for the entire Divvy trip data form which I have obtained the sample used in the current study. The rest of this section consists of a brief introduction to Divvy datasets, a detailed

Table 1: List of datasets used

	Source	Description
Divvy trip	Divvy	Trip and user features of all Divvy trips made in 2016
Divvy station	Divvy	Names and locations of all Divvy stations
CTA stops	CTA	Names and locations of all CTA stops
CTA stop times	CTA	Scheduled arrival and departure times of all CTA routes at all stops
Community area boundaries	City of Chicago Data Portal	Geographical coordinates for the boundaries of Chicago community areas
Central business district boundary	City of Chicago Data Portal	Geographical coordinates for the boundary of the central business district
Demographic features	United States Census Bureau	Population and employment data on the census tract level in 2015
Weather	National Centers for Environmental Information	Daily temperature and precipitation data in Chicago in 2016

account of the process of obtaining a trip sample for fitting the model, an explanation on key variables, and a descriptive summary of the sample data.

4.1 Divvy data and sample formation

Divvy data per-trip data consists the following information for each trip: ID attached to the trip, day and time the trip started, day and time the trip ended, ID attached to the bike used, duration of the trip in seconds, name of the station where the trip started, name of the station where the trip ended, type of the user (e.g., subscriber to Divvy's Annual Membership or 24-Hour Pass user), and, if the rider is a subscriber, the gender and age of the rider. Divvy's station data provides the ID and name of all Divvy stations as well as their locations in longitude and latitude, dock capacity, and the date when the collection of per-trip data from each station first started. The current project uses per-trip observations for the year of 2016, which amounts to total 3,595,383 trips. Also, this project uses the most recent docking station data, which includes information on total 581 Divvy stations.

Although its focus differs from that of the current study, Faghih-Imani and Eluru (2015), which also examines Divvy uses, offers an exemplary case as to forming a sample set for examining Divvy trips. Therefore, I have followed some of its sample formation methods used in Faghih-Imani and Eluru (2015), in addition to adding original steps to better serve the particular objectives of the current study. First, I have removed trips of which user type is other than members of annual subscription or customers of 24-Hour Pass; only 40 instances have been removed as a result. This is to facilitate the comparison between the two major user types: annual subscriber and daily customer. Second, trips made by annual subscribers older than 80 are excluded from the sample for this study (total 1,284 trips; See Figure 6 in Appendix A-1 for the age distribution for all trips by annual subscribers)[5]. Since the original dataset offers the years of birth instead of the ages of subscribers, I have used a simple arithmetic to calculate age, subtracting the birth years from 2016, which is the year the data were collected. Third, I have removed observations where the trip is made by an annual subscriber but the subscriber's gender is missing (total 308 trips). Considering that the process of subscribing to Divvy's annual membership requires the subscriber's gender information, I suspect that such omission is likely due to some error.

Fourth, I have deleted trips longer than 90 minutes in duration (only 1.01% of all trips). Not only are trips longer than 90 minutes atypical for bike share rides, they also could have resulted from misplaced bikes. Fifth, the current study does not consider trips that had the same origin and destination (only 3.4% of all trips). Such trips could result from malfunctioning bikes that users had to return to the origin stations. Furthermore, as noted by Faghih-Imani and Eluru (2015), accommodating trips that were intended to have the same origin and destination requires additional trip purpose information, which is beyond the scope of the current study. Lastly, I have removed trips that are made from or to docking stations that are not located in

[5]This step is to exclude observations whose age information may be mis-coded while preserving as much information as possible. For example, the original trip data includes trips made by, for example, a subscriber who was born in the 19th century. Although it is possible that such trips were indeed made by a 117-year-old, I suspect that these are more likely due to some error in the encoding process.

any of 77 Chicago community areas. This is a necessary step to take since the current study seeks to account for demographic characteristics of the stations one the level of each community area.

To obtain a reasonable sample size for model estimation, two sets of 20,000 trips, one for trips originating (stage $s = origination$) and the other for trips terminating ($s = termination$), are randomly drawn from the *clean* dataset obtained from the aforementioned six steps. This clean dataset consists of total 3,434,320 trips or approximately 95.5% of all Divvy trips made in 2016. Each set of 20,000 trips is than used for model estimation for each trip stage.

Table 2 offers descriptive summaries of the sample Divvy trips used to estimate the model parameters. Table 2 is divided into two panels, *Panel A* on the top for trips originating and *Panel B* on the bottom for trips terminating. For both trip stages s, about four-fifth of all trips are made by annual subscribers, and about three-quarters of all trips by subscribers are made by male subscribers. The age distribution of trips made by annual subscribers is skewed to the right. For each s, three-quarters of all trips are made in weekdays (Monday to Friday) and a little less than a half of all trips are made in rush hours, defined as 6:00-10:00 AM and 4:00-8:00 PM in weekdays. About 40% of these rush-hour trips are made in the morning. Precipitation marks the proportion of trips for each s that are made in rainy days. It must be noted that the weather data is collected on the level of the entire city for each day rather than the level of each station at each specific moment when a trip starts or ends. The mean and median values of both trip duration (in minutes) and trip distance (in meters) suggest that the distributions of both trip duration and trip distance are skewed to left. Note that, just as in the case of obtaining the proximity between Divvy stations and CTA stops, the trip distance is calculated by using Manhattan distance between the origin and destination Divvy stations instead of Euclidean distance. For the trips made from or to stations with access to CTA stops, the proportion of such trips increases with the increasing proximity standard d, which makes intuitive sense. The same holds for the proportion of PMM Divvy trips.

Table 2: Descriptive summary of Divvy trips

Panel A. Trips originating

	Type	Mean	Std.Dev.	Median	Max	Min	N
Annual membership	0-1	0.77	0.42	-	-	-	20,000
Gender (male)	0-1	0.75	0.43	-	-	-	15,480[1]
Age	Count	35.39	10.63	32	77	16	15,480[1]
Weekday	0-1	0.73	0.44	-	-	-	20,000
Rush hour	0-1	0.46	0.50	-	-	-	20,000
Morning (AM)	0-1	0.41	0.49	-	-	-	9,108
Precipitation	0-1	0.27	0.44	-	-	-	20,000
Trip duration (min)	Cont.	14.44	11.10	11.50	89.73	1.00	20,000
Trip distance (km)	Cont.	1.46	1.27	1.13	10.74	0.02	20,000
Proximity to CTA[2]							
Distance $\leq$ 50m	0-1	0.47	0.50	-	-	-	20,000
Distance $\leq$ 100m	0-1	0.73	0.44	-	-	-	20,000
Distance $\leq$ 200m	0-1	0.88	0.32	-	-	-	20,000
Distance $\leq$ 300m	0-1	0.93	0.26	-	-	-	20,000
Potential multi-modality[4]							
Distance $\leq$ 50m	0-1	0.35	0.48	-	-	-	20,000
Distance $\leq$ 100m	0-1	0.58	0.49	-	-	-	20,000
Distance $\leq$ 200m	0-1	0.75	0.43	-	-	-	20,000
Distance $\leq$ 300m	0-1	0.81	0.39	-	-	-	20,000

Panel B. Trips terminating

	Type	Mean	Std.Dev.	Median	Max	Min	N
Annual membership	0-1	0.78	0.42	-	-	-	20,000
Gender (male)	0-1	0.75	0.43	-	-	-	15,648[1]
Age	Count	35.44	10.58	32	78	16	15,648[1]
Weekday	0-1	0.73	0.44	-	-	-	20,000
Rush hour	0-1	0.45	0.50	-	-	-	20,000
Morning (AM)	0-1	0.41	0.49	-	-	-	9,060
Precipitation	0-1	0.28	0.45	-	-	-	20,000
Trip duration (min)	Cont.	14.38	11.05	11.48	89.42	1.13	20,000
Trip distance (km)	Cont.	1.46	1.27	1.14	16.41	0.01	20,000
Proximity to CTA[2]							
Distance $\leq$ 50m	0-1	0.47	0.50	-	-	-	20,000
Distance $\leq$ 100m	0-1	0.73	0.45	-	-	-	20,000
Distance $\leq$ 200m	0-1	0.89	0.32	-	-	-	20,000
Distance $\leq$ 300m	0-1	0.93	0.25	-	-	-	20,000
Potential multi-modality[4]							
Distance $\leq$ 50m	0-1	0.35	0.48	-	-	-	20,000
Distance $\leq$ 100m	0-1	0.58	0.49	-	-	-	20,000
Distance $\leq$ 200m	0-1	0.76	0.43	-	-	-	20,000
Distance $\leq$ 300m	0-1	0.82	0.39	-	-	-	20,000

[1] Subscribers to annual membership only.
[2] The originating docking station is in proximity with CTA stops for the given proximity standard.
[3] The terminating docking station is in proximity with CTA stops for the given proximity standard.
[4] The trip starts as potentially multi-modal for the given proximity standard.
[5] The trip ends as potentially multi-modal for the given proximity standard.

Table 3: Descriptive summary of Divvy stations

	Type	Mean	Std.Dev.	Median	Max	Min	N
Trips originating[1]	Count	41.67	48.66	26.5	477	1	20,000
Potentially multi-modal							
Distance $\leq$ 50m	Count	30.92	41.30	18	364	1	6,958
Distance $\leq$ 100m	Count	32.75	40.63	18.5	364	1	11,527
Distance $\leq$ 200m	Count	35.58	42.63	20	365	1	15,087
Distance $\leq$ 300m	Count	36.69	42.46	22.5	365	1	16,292
Trips terminating[2]	Count	41.93	50.12	26	525	1	20,000
Potentially multi-modal							
Distance $\leq$ 50m	Count	30.68	41.44	18	376	1	6,903
Distance $\leq$ 100m	Count	33.83	41.40	20	376	1	11,537
Distance $\leq$ 200m	Count	36.53	43.71	20	376	1	15,124
Distance $\leq$ 300m	Count	37.51	43.31	23	376	1	16,354
Station capacity	Count	17.72	5.70	15	47	11	503
Located in CBD	0-1	0.17	0.37	-	-	-	503
Presence of CTA stops							
Distance $\leq$ 50m	0-1	0.49	0.50	-	-	-	503
Distance $\leq$ 100m	0-1	0.75	0.44	-	-	-	503
Distance $\leq$ 200m	0-1	0.89	0.31	-	-	-	503
Distance $\leq$ 300m	0-1	0.94	0.24	-	-	-	503
Number of CTA stops							
Distance $\leq$ 50m	Count	0.92	1.21	0	6	0	503
Close stations only	Count	1.88	1.09	2	6	1	248
Distance $\leq$ 100m	Count	2.03	1.84	2	11	0	503
Close stations only	Count	2.73	1.63	2	11	1	375
Distance $\leq$ 200m	Count	4.33	2.97	4	16	0	503
Close stations only	Count	4.84	2.72	4	16	1	450
Distance $\leq$ 300m	Count	7.79	4.66	8	24	0	503
Close stations only	Count	8.32	4.33	8	24	1	471

[1] Total 480 docking stations.
[2] Total 477 docking stations.

Table 3 presents summary statistics of Divvy stations associated with trips in the sample. Some notable points shown in this table are as follows: The sample trips originated from 480 docking stations and terminated at 477 docking stations. The union of these two includes total 503 stations. For trips originating, each station has 41.67 trips on average that started from that station. For trips terminating, each station has 41.93 trips on average that stopped at that station. The mean docking capacity for Divvy stations is 17.72 bikes while the median capacity is 15 bikes. 17% of all Divvy stations are located within the central business district. About a half of all Divvy stations have at least one CTA stop within 50-meter. With different proximity standard, the proportion of Divvy stations that are close to CTA stops increases. With the 300-meter proximity standard, all but 6% of Divvy stations are labeled *close* to CTA stops. Docking stations with the same certain proximity measure differ in the number of CTA stations that are marked close to them. On average, there are 1.88 nearby CTA stops for Divvy stations that have at least one CTA stop within the 50-meter range. This number also rises with the increasing proximity standard. With the 300-meter proximity standard, the average number of nearby CTA stops is as high as 8.32 for those docking stations with at least one close-by CTA stop.

Lastly, Figures 1 and 2 provide geographical illustrations of the sample PMM trips by proximity standard d, the former for trips originating and the latter for trips terminating. In each map, each bubble marks a docking station and the size of the bubble corresponds to the number of PMM trips either originating from or terminating at the given docking station. The yellow box marks the boundary of Chicago's central business district. As suggested in Table 2 above, the number of trips that are marked PMM grows with the proximity standard for both trip stages s. The graphics further illustrate that most of the PMM trips can be found in the central business district and the surrounding areas (especially to the north of the central business district), which makes intuitive sense because of the geographical distribution of docking stations (shown in Figure 7 in Appendix A-1).

5 Results

In this section, I present the results of the multinomial logistic regression and discuss different features that characterize trips in different PMM categories and, by extension, may provide insights into Divvy users' decision to make multi-modal trips in reality. In doing so, I focus on the comparison between the coefficients across different categories of PMM trips, namely $PMM1$, $PMM2$, $PMM3$, and $PMM4$. This comparison allows me to consider the marginal effect of each proximity standard. In addition, through the same comparison, I can assess the reliability of the 300-meter standard used in an earlier study on Divvy trips.

As suggested in Sections 3 and 4 above, the model estimation uses two separately

Figure 1: Potentially multi-modal Divvy trips by proximity standard: Trips originating

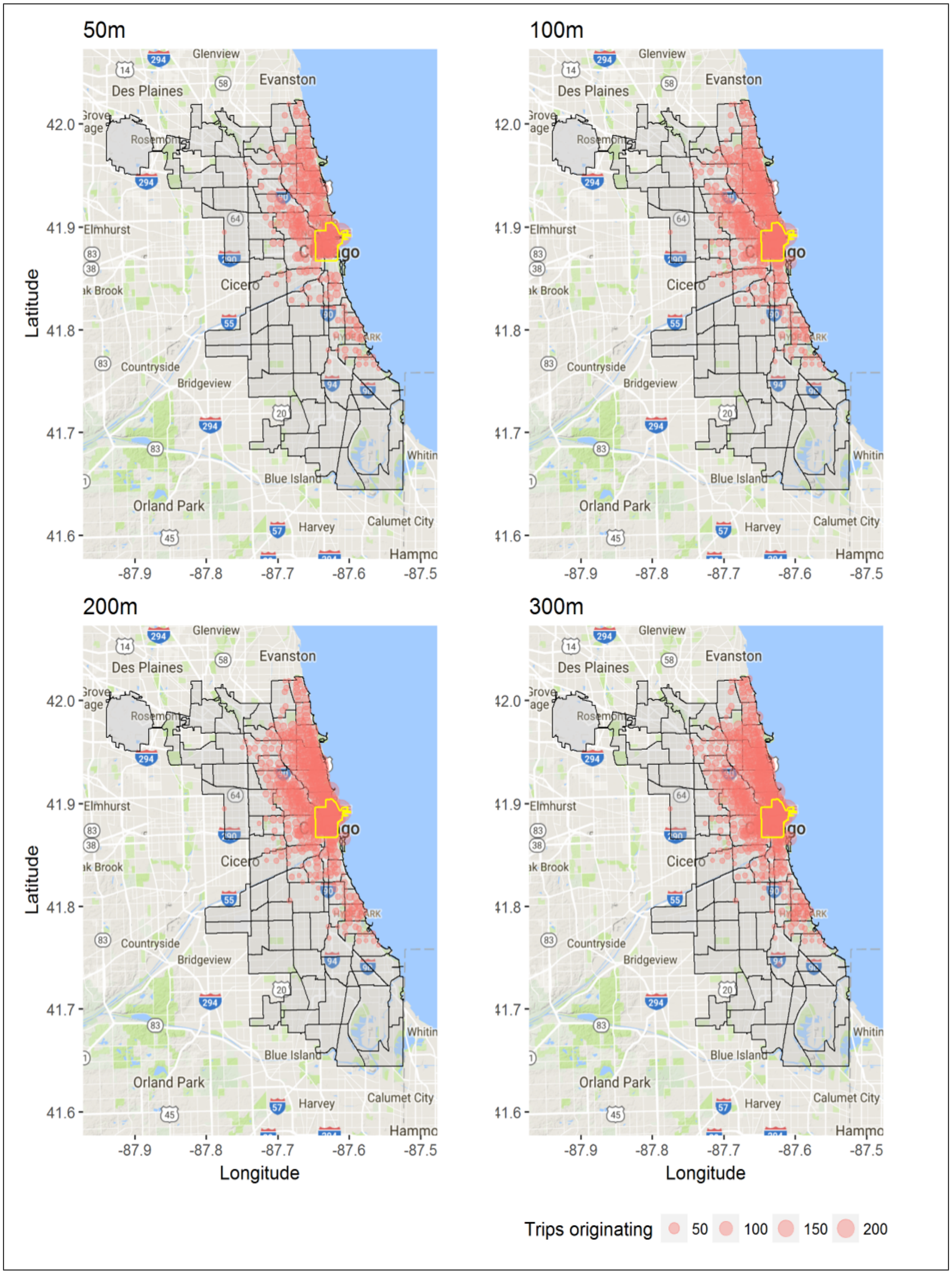
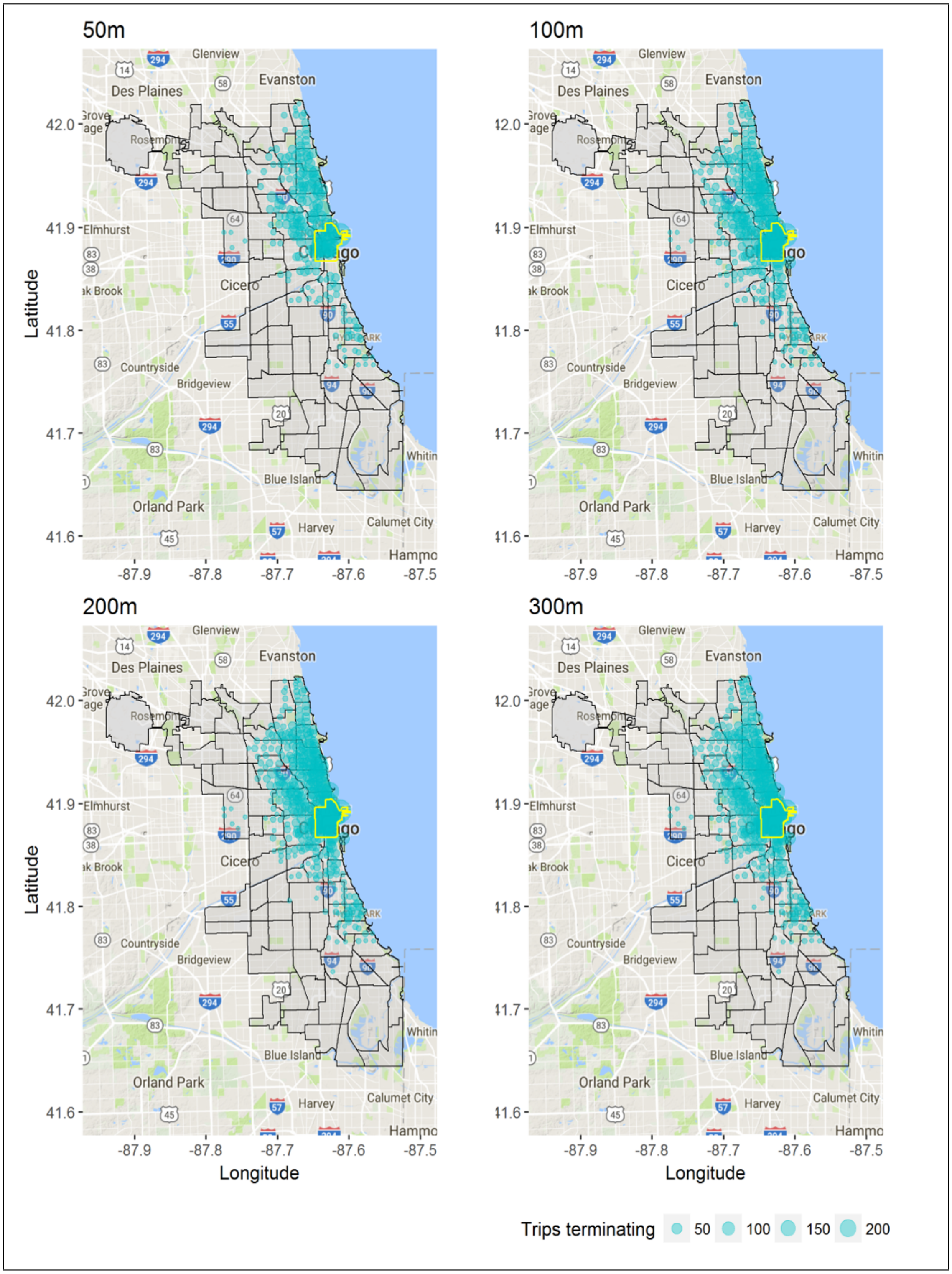


Figure 2: Potentially multi-modal Divvy trips by proximity standard: Trips terminating



generated sample sets of the same size $n = 20,000$, each corresponding to one of the trip stage $s \in \{origination, termination\}$. Then, for each trip stage, I estimate a multinomial logit model where the response variable has five mutually exclusive and exhaustive PMM categories as its possible outcome: (1) $PMM1$ for 0 to 50 meters, (2) $PMM2$ for 50 to 100 meters, (3) $PMM3$ for 100 to 200 meters, (4) $PMM4$ for 200 to 300 meters, and (5) $None$ for trips that are not PMM by any and all proximity standards d (See Section 3.2 above for more on this categorization of the response variable as well as the mathematical description of the multinomial logit model). In addition, I compare the coefficients of fitted models across different PMM categories to test the validity of the 300-meter proximity standard used in the prior study on Divvy (Faghih-Imani and Eluru, 2015) to evaluate the impact of access to public transit on Divvy bike trips.

Table 4 summarizes estimation results for two multinomial logit models. *Panel A* on the top presents estimation results for trips originating and *Panel B* on the bottom, for trips terminating. Both models have largely the same set of explanatory variables: the first six ("Subscriber," "Precipitation," "Weekday," "Rush hour," "Rush hour (AM)," and "Trip distance (km)") are trip-level features, the next five ("Station in CBD," "Population density," "Black (%)," "Employed (%)," and "Income per capita ($1,000)") are station-level features. The only difference is that "Rush hour" and "Rush hour (AM)" variables are based on when the trip starts for trips originating and on when the trip ends for trips terminating. It must be noted that each coefficient estimate cannot be interpreted as the size of marginal effect in probability as in a binomial logit model. Instead, the coefficient estimates are given in log-odds ratio between each PMM category and the baseline category for trips that are not PMM by any and all proximity standards (See Equation 1). Standard error for each coefficient estimate is given in parentheses and statistical significance of each estimate is marked by the number of asterisks. The final row presents the likelihood ratio test statistic for the fitted model against the null model with 44 degrees of freedom, $\chi^2_{(44)}$. The likelihood ratio test compares the goodness of fit (model deviance in this case) between

Table 4: Model estimation results

Panel A. Trips originating[1][2]

	$PMM1$	$PMM2$	$PMM3$	$PMM4$
(Intercept)	-2.7125***	-1.8921***	-0.8136***	-5.6990***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Subscriber	1.2092***	1.4588***	1.4186***	1.1214***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Trip distance (km)	-0.1318***	-0.1564***	-0.1475***	-0.1423***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Precipitation	0.0075***	0.0830***	0.0626***	-0.1058***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Weekday	-0.1118***	-0.0881***	-0.0468***	0.1952***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Rush hour[3]	0.4622***	0.4091***	0.2725***	-0.1078***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Rush hour (AM)[3]	0.0580***	-0.2234***	0.0617***	0.0376***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Station in CBD	1.1103***	0.6511***	0.5875***	0.0718***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Population density (km^2)	0.00001***	-0.00001***	0.00000***	0.00000***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Black (%)	0.0032***	-0.0118***	-0.0021***	0.0335***
	(0.0009)	(0.0004)	(0.0005)	(0.0005)
Employed (%)	0.0496***	0.0236***	-0.0032***	0.0419***
	(0.0007)	(0.0007)	(0.0006)	(0.0006)
Income per capita (\$1K)[4]	-0.00003***	-0.00000	-0.00000***	0.00000***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Likelihood ratio test statistic ($\chi^2_{(44)}$) :	3158.159***			

Panel B. Trips Terminating[1][2]

	$PMM1$	$PMM2$	$PMM3$	$PMM4$
(Intercept)	-3.3207***	-3.5320***	-1.5084***	-4.6695***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Subscriber	1.1184***	1.4169***	1.3715***	1.0152***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Trip distance (km)	-0.1296***	-0.1837***	-0.1207***	-0.1487***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Precipitation	0.0213***	0.0285***	-0.0025***	0.1495***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Weekday	0.0312***	0.0982***	0.1440***	0.1467***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Rush hour[3]	0.4566***	0.2442***	0.2244***	0.0543***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Rush hour (AM)[3]	0.2011***	0.5487***	0.5185***	0.2561***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Station in CBD	1.3007***	0.8086***	0.5749***	0.0628***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Population density (km^2)	0.00001***	-0.00001***	0.00000	0.00000***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Black (%)	0.0050***	-0.0051***	-0.0003	0.0267***
	(0.0009)	(0.0004)	(0.0006)	(0.0006)
Employed (%)	0.0589***	0.0460***	0.0036***	0.0314***
	(0.0007)	(0.0007)	(0.0006)	(0.0006)
Income per capita (\$1K)[4]	-0.00003***	-0.00001***	-0.00000 *	0.00000*
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Likelihood ratio test statistic ($\chi^2_{(44)}$) :	3318.692***			

[1] Significance: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.001$
[2] All coefficient estimates are in log-odds ratios against the baseline category $y = None$.
[3] Whether the trip started/ended during rush hour
[4] In 2015 inflation-adjusted dollars.

the fitted model and the he null model. The degrees of freedom, 44, equals the number of added parameters for eleven explanatory variables and four categories.

For each trip stage s, estimated coefficients vary across PMM categories y although most of them appear statistically significant with extremely small p-values. The constant term is negative for all y and for both s, while the magnitude varies across y. Conversely, the coefficient estimate for "Subscriber" variable, a binary dummy for Divvy user type, is positive and statistically significant across all d for both s, but, again, its magnitude varies with d. "Trip distance (km)" is a continuous variable for the Manhattan distance between the originating station and terminating station for each trip in kilometers (km). The coefficient estimate for "Trip distance (km)" is negative for all y for trips originating and trips terminating. "Precipitation" is another binary dummy for trips made in rainy days, i.e. days with positive daily precipitation. The estimated coefficients are all positive except for $y = PMM4$ for trips originating and $y = PMM3$ for trips terminating. "Weekday" is, again, a binary dummy for trips made during weekdays (Monday to Friday). The coefficient estimates for "Weekday" positive for $y = PMM4$ for trips originating and for all y for trips terminating while the estimates are negative for for $y = PMM1, PMM2,$ and $PMM3$ for trips originating. "Rush hour" is another binary dummy for trips made during rush hours as defined in Section 4 above. For trips originating, the coefficient estimate for rush hour are all positive but for $y = PMM4$. For trips terminating, the estimates are all positive. For "Rush hour (AM)," the estimates are all positive but for $y = PMM2$ for trips originating.

The coefficient estimates for "Station in CBD," a binary dummy that marks whether the station is located within the central business district (CBD), are positive across all y and s while the magnitude decreases as j increases. That is, for both s, the estimate is larger for $y = PMM1$ than $y = PMM2$, and so forth. "Population density (km^2)" is a continuous variable for the population density of each community area in square kilometers. Although the estimates are extremely small, they all still appear statistically significant except for $y = PMM3$ for trips terminating. "Black (%)" is a continuous variable for the proportion of Black residents in the given community area in percentage (0 to 100%), and the coefficient estimates are all statistically significant except for $y = PMM3$ for the trips terminating. The sign of these estimates varies although the difference in magnitude is still small across y. Similarly, "Employed (%)" is for the proportion of employed residents in the given community area. All the estimated coefficients are positive except for $y = PMM3$ for trips originating. Lastly, "Income per capita (\$1K)" is a continuous variable for the per-capita income for each community area in inflation-adjusted dollars for the year of 2015. The estimated coefficients are all statistically significant except for $PMM2$ for trips originating although the magnitude is extremely small for all y and s.

For the current study, certainly a more crucial point is whether the coefficients for each variable are sufficiently different from one another across y, which may not only suggest that the marginal effect of increasing proximity standard d is non-negligible but also possibly challenge the validity of $d = 300m$ as a proximity standard when evaluating the extent to which the public transit availability affects bike share uses. If the variation in estimated coefficients between the $y = PMM4$ and the other PMM categories is negligible, we may conclude that the $d = 300m$ is indeed a viable choice for the proximity standard. If we observe some significant variation in coefficients, however, such a variation implies that researchers may have to reconsider the validity of 300 meters as the proximity standard. Figures 3 and 4 offer graphical comparisons among the estimated coefficients for all variables across PMM categories y. Careful examination of these figures can shed light on whether the typical proximity standard of $d = 300m$ is an adequate choice of evaluating the access to public transit on Divvy trips. For each coefficient, the point marks the estimate given by the fitted model. The short vertical line one each point is in fact a horizontal line with a tick on each side end intended to illustrate the interval of two standard deviations, i.e., the 95% confidence interval for each estimate. However, the extremely small standard errors for all estimates have resulted in a seemingly single vertical line on each point. The grey vertical dashed line marks 0, where the given explanatory variable has no discriminatory power between the given PMM category and the baseline category.

Figure 3: Comparing coefficient estimates across proximity standards (in log-odds ratio): Trips originating

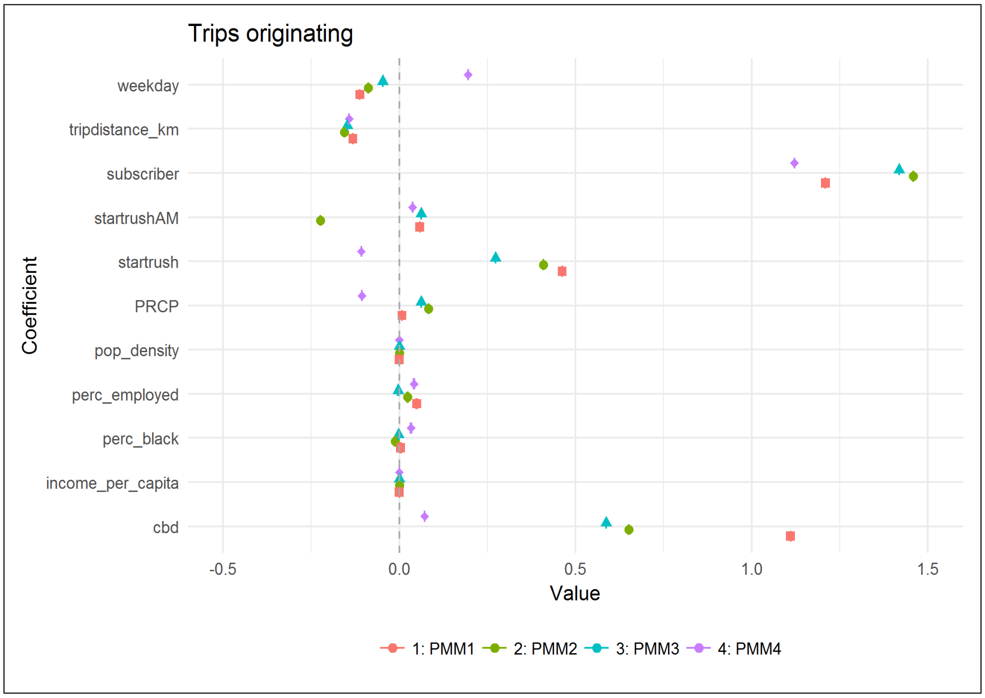
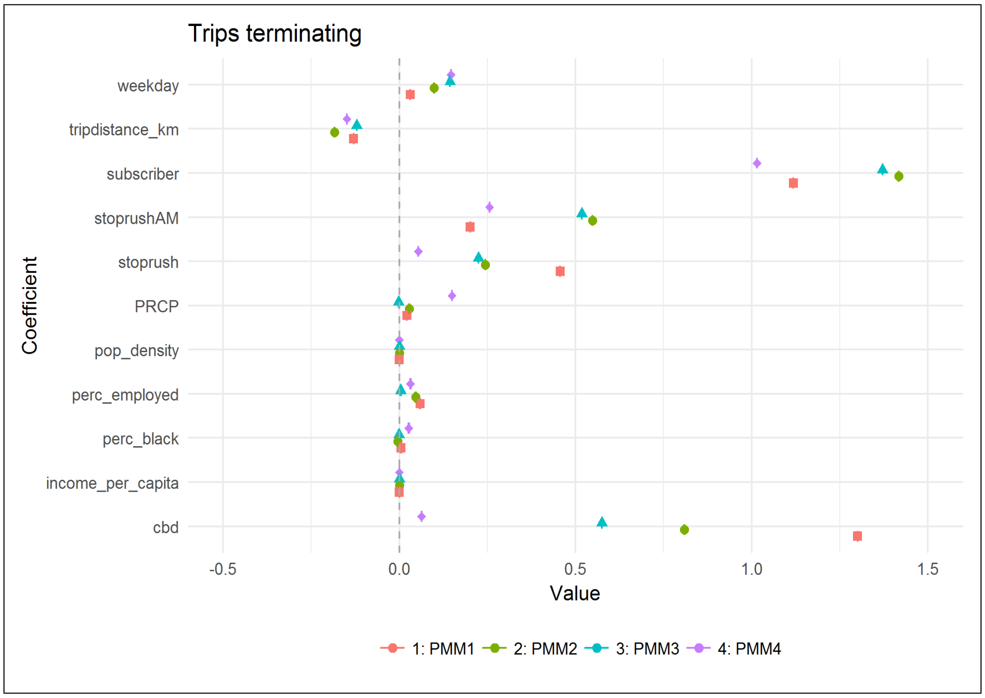


Figure 4: Comparing coefficient estimates across proximity standards (in log-odds ratio): Trips terminating



It must be noted that, largely due to the extremely small standard errors, most estimated coefficients are different across $y \in \{PMM1, PMM2,\ PMM3,\ PMM4\}$ for both trips originating and terminating and these differences are all statistically significant. This point can be easily verified by fitting the same multinomial logit model while switching the baseline category from $y = None$ to any other or manually calculating the p-value using the given coefficient estimates and standard error as presented in Table 4 above. In fact, differences among the estimated coefficients for any given explanatory variable are statistically significant at the 1% level for all pairs but two. These two pairs are, namely, the estimates for "Population density (km^2)" between $PMM3$ and $PMM4$ for the trips originating, which is not statistically significant, and the same pair for the trips terminating, which is still statistically significant at the 10% level.

In addition to the statistical significance, economic significance, or magnitude, of the differences matters as well. Indeed, as the Figures 3 and 4 suggest, some of the differences, regardless of the statistical significance, are indeed too small to have any practical import. Therefore, in the following, I will focus on the explanatory variables for which we observe the notable, or economically significant, differences in estimated coefficients across $y \in \{PMM1, PMM2, PMM3, PMM4\}$ that are equal to or greater than 0.1 in magnitude for any of the pair. A difference of 0.1 in estimated coefficients, which are in log-odds, can be then interpreted as approximately 1.1 in terms of odds ratio (or relative risk ratio) between two categories for a unit change in the given explanatory variable, holding other variables constant. That suggests that a unit increase in the given variable is correlated with an expected increase in the relative risk between y_1 and y_0 (the baseline category) by a factor of 1.1, holding all the other variables in the model constant. That is, given a unit increase in the given variable, the relative risk of being in y_1 category would be 1.1 times more likely than in the baseline category, holding all the other variables constant.

In Figure 3, we observe that, for the trips originating, estimated coefficients show notable variations (i.e., equal to or greater than 0.1) across $y \in \{PMM1,\ PMM2,\ PMM3, PMM4\}$ for the following explanatory variables: "Weekday (`weekday`)," "Subscriber (`subscriber`)," "Rush Hour (AM) (`startrushAM`)," "Rush Hour (`start-rush`)," "Precipitation (`PRCP`)," and "Station in CBD (`cbd`)." As mentioned earlier, the precise size of these differences can be verified by comparing estimates presented in Table 4 above or fitting new models using different baseline categories. First, for "Weekday" variable, only the estimate for $PMM4$ is distinctively positive while others are all negative. The estimated coefficients for "Subscriber" are all positive, but they vary noticeably in magnitude. For "Rush hour (AM)" variable, all estimates are positive except for $PMM2$. For both "Rush hour" and "Precipitation" variables, the coefficient estimates are all positive except for $PMM4$. Lastly, the estimated coefficients for "Station in CBD" are positive for all categories although with a high degree of variation.

Figure 4 illustrates that the estimated coefficients for trips terminating show a mostly similar pattern. Just as before, the estimated coefficients for the same set of explanatory variables display sufficient variations across $y \in \{PMM1, PMM2, PMM3, PMM4\}$. However, there are some remarkable differences. First, the estimated coefficients for "Weekday" are all positive. In addition, the difference between the coefficient estimate for $PMM4$ and the estimates for other categories has diminished and is no longer economically significant (less than 0.1 in magnitude). For "Rush Hour (AM)," the coefficient estimates for all categories have increased and are now all positive. The difference in coefficient estimates between $PMM1$ and $PMM3$ is now economically significant and so is the difference between $PMM3$ and $PMM4$. The coefficient estimates for "Rush Hour (AM)" have also increased and now are positive for all PMM categories. Meanwhile, the difference between $PMM1$ and $PMM2$ has become economically significant while the difference between $PMM2$ and $PMM3$ is now no longer greater than 0.1. For "Precipitation," the sign of the estimated coefficient for $PMM4$ has changed from negative to positive. Finally, for "Station in CBD", the difference in coefficient estimates between $PMM2$ and $PMM3$ is now greater than 0.1.

6 Discussion

In the following, I recapitulate the goal of the current study and, in the context of this goal, highlight the key implications of the model estimation results. I also discuss the limitations of the current study and make suggestions as to how to further research on the topic at hand: evaluating the effect of bike share on addressing the last mile problem in public transit on the one hand and the effect of public transit accessibility to the bike share uses, which I argue are in fact two sides of the same coin.

As suggested in the prior discussion on the notion of PMM trips in Sections 2 and 3, including a proximity standard in a model to evaluate the effect of access to public transit on bike share trips makes an unstated assumption that bike trips that are affected by access to public transit are in fact linked to public transit rides. In other words, such trips are multi-modal. Consequently, the distance-only approach in previous studies leads to an inadequate measurement: the distance-only measure is likely to pick up the effect of other features that are correlated with the presence of public transit stop. Therefore, in order to better isolate the true effect of the access to public transit, some extra control is required. To that end, the current study has incorporated two additional criteria (i.e., the timing as well as the duration of each trip) in determining which trips are likely affected by the access to public transit, or, potentially multi-modal (PMM). While keeping these additional criteria fixed, the current study has tested whether the trips labelled as PMM with the 300-meter standard are more or less homogeneous but still distinct from non-PMM trips.

The principal implication of the model estimation results is the following: trips that are labeled PMM with the 300-meter proximity standard (i.e., $PMM1$, $PMM2$, $PMM3$, and $PMM4$ trips combined) constitute a heterogeneous collection of trips. The estimation results suggest that, for both trip stages, the marginal effect of extending the proximity standard from one distance to the next one involves labeling more trips PMM even though the newly labelled PMM trips are notably different from trips that are already PMM with the previous proximity standard. This is to say that, for example, $PMM1$ trips and $PMM2$ trips are sufficiently distinguishable from each other based on certain features, including whether the trips are made by annual Divvy membership subscribers rather than 24-Hour Pass customers ("Subscriber"). The estimation results further suggest that this is also the case for other sequential pairs of PMM categories (i.e., one of $PMM2$ and $PMM3$, and the other of $PMM3$ and $PMM4$).

This finding presents a challenge with regards to using the 300-meter distance to evaluate the impact of public transit accessibility on bike share use as in Faghih-Imani and Eluru (2015), or even a longer distance as in Ma et al. (2015). Based on the results, we may conclude either that there truly exist multiple groups of Divvy trips based on the distance between docking station and public transit stop that are likely multi-modal but distinguishable from one another, or that we must reconsider the use of 300-meter standard. Here, I suggest that the second conclusion is more plausible for at least the following two reasons: First, within the rational choice framework, we expect each passenger to minimize the cost of her trip. The longer the distance is between the bike docking station and the nearby bus or rail stop, the more costly it is for a passenger to link them as part of a multi-modal trip. Given that, if using the longer-distance proximity standard leads to the labelling of trips PMM while the new PMM trips are distinct from the PMM trips with the previous proximity standard, it is a sign that the proximity with a longer distance may be introducing more Type I (false positive) errors. Second, as Table 1 in Section 4 has illustrated, the 300-meter proximity standard results in the suspiciously high proportion of PMM trips. Even when applying the extra criteria regarding the timing and duration of trips, we find that over 80% of all trips would be labelled PMM using the 300-meter proximity standard. Without additional constraints, the proportion is as high as 93%. This is counter-intuitive.

However, interpretation of the estimation results must be made with a great caution. On the fundamental level, trips that are labeled PMM cannot serve as direct measures for multi-modal trips in the real world. Since we are currently missing any data connecting Divvy bike uses and CTA rides, as mentioned earlier in Section 3 above, we are not equipped yet to directly test and compare the proximity standards to determine which one provides the most reliable measure for actual multi-modal trips and, by extension, the true effect of the access to public transit to bike share uses as well as the contribution of bike share to solving the last mile problem in public transit. Although this fundamental limitation does not fully invalidate the finding of the current study as to challenging the legitimacy of the 300-meter proximity measure, the finding cannot be used to make inferences regarding the true characteristics of multi-modal trips.

Aside from the comparison among trips in different PMM categories, the model estimation results suggest that demographic features of the community areas generally do not help us to tell PMM trips apart from non-PMM trips. The only exception is the percentage of residents who are employed in the given community area, but the magnitude is small. When combined with the coefficient estimates for variables related to the timing of Divvy trips, this particular finding may still suggest that PMM trips are likely correlated with the commuting behavior of the working population.

I make two recommendations for future research on how bike share and the existing public transit network influence each other. First and foremost, researchers must reconsider the use of simple distance-based measure. The current study has pointed out the limitation of this approach and proposed an alternative that both accounts for the timing and the duration of each bike share trip and is in accordance with the rational choice framework, which is widely accepted in the transportation literature. Furthermore, the proposed approach can be adjusted and further enhanced based on findings from in-depth qualitative research on multi-modal trips incorporating bike share and other public modes of transportation. Second, if possible, researchers must seek to collaborate with the stakeholders to collect the actual data on multi-modal trips combining a local bike share program and a public transit network. A good sample of true multi-modal trips can serve as the standard against which different measures can be tested and compared, which then can inform the relevant decision makers in the realm of public transportation to better serve the population of their interest.

7 Conclusion

The soaring popularity of bike share programs around the world in the recent years has attracted considerable scholarly attention and resulted in the growing body of research on bike share. In this body of research, we can find many studies highlighting the potential contribution of bike share to solving the last mile problem in public transit. However, few studies have attempted to quantitatively evaluate this claim with observational data on bike share uses. On the other hand, studies that take advantage of observational data have mostly resorted to a rather simplistic measure based solely on distance in order to assess the effect of public transit accessibility, which is essentially equivalent to appraising the possibility of multi-modal trips and, consequently, the contribution of bike share to solving the last mile problem.

The current study addresses this gap and proposes the notion of potentially multi-modal (PMM) trips as an alternative approach to evaluating the effect of public transit accessibility to bike share uses. Based on the rational choice framework, PMM trips account for additional conditions, including the duration of each trip as well as the expected availability of public transit services based on not only spatial but also temporal dimensions, to provide a measure that is theoretically sounder than the distance-only measure for the effect of public transit accessibility. Using a sample of Divvy trips made in 2016 and a selected set of trip-level as well as station-level features as explanatory variables, this study compares trips that would be classified PMM with different proximity standards (50 meters, 100 meters, 200 meters, and 300 meters). Here, the main objective of the comparison is to test the validity of 300-meter proximity standard used in a prior study on Divvy trips (Faghih-Imani and Eluru, 2015). Notably, the comparison focuses on the marginal effect of extending the proximity standard from one distance to the next generous one (e.g., trips that would be labeled PMM when extending the proximity standard from 50 meters to 100 meters, and so forth) instead of comparing all PMM trips by one proximity standard directly with all PMM trips by another standard.

Accordingly, the multinomial logit model in this study takes the categorical response variable with five possible outcomes, each corresponding to the marginal effect of using a more generous proximity standard. The trips that are not labeled PMM by any and all proximity standards constitute the baseline category against which the log-odds ratio for each PMM category can be estimated. The explanatory variables include both trip-level features (rider type, timing of the trip, trip distance, and precipitation) and station-level features (whether the station is located in the central business district and four demographic features of the community area in which the station is located). The resulting model coefficient estimates are given in log-odds ratio between each PMM category and the baseline category. A close examination of the estimation results suggest that Divvy trips that are labeled PMM with the 300-meter proximity standard form a heterogeneous set of trips. When combined with the expected, cost-minimizing behavior of rational passengers and the counter-intuitively high proportion of PMM trips with the 300-meter proximity standard, the finding of this study offers a ground for reconsidering the use of 300-meter standard to gauge the effect of access to public transit to Divvy trips.

With its finding, the current study makes the following contributions to research on bike share. First, it challenges the use of the simple distance-only measure for the effect of public transit accessibility on bike trips. In doing so, it proposes a theoretically better grounded alternative, namely, the notion of potentially multi-modal (PMM) trips. In addition, this study duly emphasizes the need of a new dataset that explicitly connects bike share trips with public transit rides, which can serve as the true standard against which the validity of other measures may be tested. Lacking such data, it is not possible for this study to conclude which distance offers the best measure for the impact of the access to public transit on bike share trips. Nonetheless, once the data is made available, the theoretical notion of PMM trips as well as its practical application that this study exemplifies will prove valuable in better understanding the characteristics of multi-modal bike trips and evaluating bike share as a solution to the last mile problem in a nuanced manner.

References

Buck, Darren, Ralph Buehler, Patricia Happ, Bradley Rawls, Payton Chung, and Natalie Borecki, "Are bikeshare users different from regular cyclists?," *Transportation Research Record: Journal of the Transportation Research Board*, 2013, *2387* (1), 112–119.

City of Chicago, "Street and site plan design standards," 2007. Retrieved 22 April, 2017.

DeMaio, Paul, "Will smart bikes succeed as public transportation in the United States?," *Journal of Public Transportation*, 2004, *7* (2), 1–15.

_ , "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, 2009, *12* (4), 41–56.

Faghih-Imani, Ahmadreza and Naveen Eluru, "Analysing bicycle-sharing system user destination choice preferences: Chicagos Divvy system," *Journal of Transport Geography*, 2015, *44*, 53–64.

_ , _ , Ahmed M. El-Geneidy, Michael Rabbat, and Usama Haq, "How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal," *Journal of Transport Geography*, 2014, *41*, 306–314.

Fishman, Elliot, "Fixing Australian bike share goes beyond helmet laws," 2012. https://theconversation.edu.au/fixing-australian-bike-share-goes-beyond-helmet-laws-10229.

_ , "Bikeshare: A review of recent literature," *Transport Reviews: A Transnational Transdisciplinary Journal*, 2016, *36* (1).

_ , Simon Washington, and Narelle Haworth, "Understanding the fear of bicycle riding in Australia," *Journal of the Australasian College of Road Safety*, 2012, *23* (3), 19–27.

_ , _ , and _ , "Bike share: A synthesis of the literature," *Transport Reviews: A Transnational Transdisciplinary Journal*, 2013, *33* (2).

_ , _ , and _ , "Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia," *Transportation Research Part D: Transport Environment*, 2014, *31* (7).

Google Transit, "Reference: GTFS Specification," 2017. Retrieved 15 May, 2017.

Griffin, Greg Phillip and Ipek Nese Sener, "Planning for bike share connectivity to rail transit," *Journal of Public Transportation*, 2016, *19* (2).

Liu, Zhili, Xudong Jia, and Wen Cheng, "Solving the last mile problem: Ensuring the success of public bicycle system in Beijing," *Procedia - Social and Behavioral Sciences*, 2012, *43*, 73–78.

Ma, Ting, Chao Liu, and Sevgi Erdoğan, "Bicycle Sharing and Transit: Does Capital Bikeshare Affect Metrorail Ridership in Washington, D.C.?," *Transportation Research Record: Journal of the Transportation Research Board*, 2015, *2534*, 1–9.

McNally, Michael G., "The four step model," in David. A. Hensher and Kenneth Button, eds., *Handbook of Transportation Modeling*, 2 ed., Amsterdam, Netherlands: Pergamon, 2007, chapter 3, pp. 35–52.

Shaheen, Susan, Adam Cohen, and Elliot Martin, "Public bikesharing in North America: Early operator understanding and emerging trends," *Transportation Research Record: Journal of the Transportation Research Board*, 2013, *2387*, 8392.

_ , Stacy Guzman, and Hua Zhang, "Bikesharing in Europe, the Americas, and Asia," *Transportation Research Record: Journal of the Transportation Research Board*, 2010, *2143*, 159–167.

Tilahuna, Nebiyou, Piyushimita (Vonu) Thakuriahb, Moyin Lia, and Yaye Keitaa, "Transit use and the work commute: Analyzing the role of last mile issues," *Journal of Transport Geography*, 2016, *54*, 359–368.

Unrau, Reuben, "Plan to merge Ventra App with Divvy receives federal grant," *Chicago Tonight*, 12 October 2016. Retrieved 25 May, 2017.

Wikipedia, "Last mile," 2016. Retrieved 22 April, 2017.

Xie, Chao, Huili Gong, and Fahui Wang, "A solution for the last mile problem of the Beijing rapid transit network: Local shuttle bus system," in "Geoinformatics, 2010 18th International Conference on" IEEE Beijing, China 2011, pp. 18–20.

Yap, Menno D., Gonalo Correia, and Bart van Arem, "Preferences of travellers for using automated vehicles as last mile public transport of multimodal train trips," *Transportation Research Part A: Policy and Practice*, 2016, *94*, 1–16.

Zellner, Moira, Dean Massey, Yoram Shiftan, Jonathan Levine, and Maria Josefa Arquero, "Overcoming the last-mile problem with transportation and land-use improvements: An agent-based approach," *International Journal of Transportation*, 2016, *4* (1), 1–26.

APPENDIX

A-1 Data

This appendix is prepared to provide the following:

- For each dataset, directions as to how to access it, where it is stored, and who curates it.
- Additional tables and visualizations for further exploring the data.

A-1.1 Divvy data

Divvy, a program of the Chicago Department of Transportation, follows the City of Chicago's open data policies and releases twice a year its historical trip data and station data since its launching in 2013. The City of Chicago owns all right, title, and interest in the data. At the time of this writing, the dataset for the third and fourth quarters of 2016 is the most recently published dataset. Anyone can access these datasets on Divvy's website for public access[6]. Divvy Data is subject to the terms and conditions of the Data License Agreement[7].

Table 5 offers a summary statistics of all divvy trips made in 2016, which amount to total 3,595,383 observations. The trips made by the subscribers of annual membership, or members, amount to total 2,7368,69 trips and trips by Daily 24-Hour Pass customers amount to total 858,474 trips. Among trips by members, 2,047,174 for trips were by male members and 689,780 trips were by female members. Table 6 provides a summary statistics of all 581 divvy stations that were operating in 2016.

Figure 5 offers two maps illustrating all Divvy trips at each station. The top map shows from which stations the trips originated, and the bottom map shows at which stations the trips terminated. The black lines mark the boundaries of Chicago community areas and the yellow line illustrates the central business area. Figure 6 presents a histogram of the age distribution of trips by annual subscribers. The age was calculated by using a simple arithmetic, i.e. subtracting the years of birth from 2016 or. the year the trips were made. The dashed vertical line marks age 80. Finally, Figure 7 shows all Divvy stations color-coded by their proximity to the closest CTA station.

[6] https://www.divvybikes.com/system-data

[7] The full content of the Agreement can be found here: https://www.divvybikes.com/data-license-agreement

Table 5: Descriptive summary of all Divvy trips

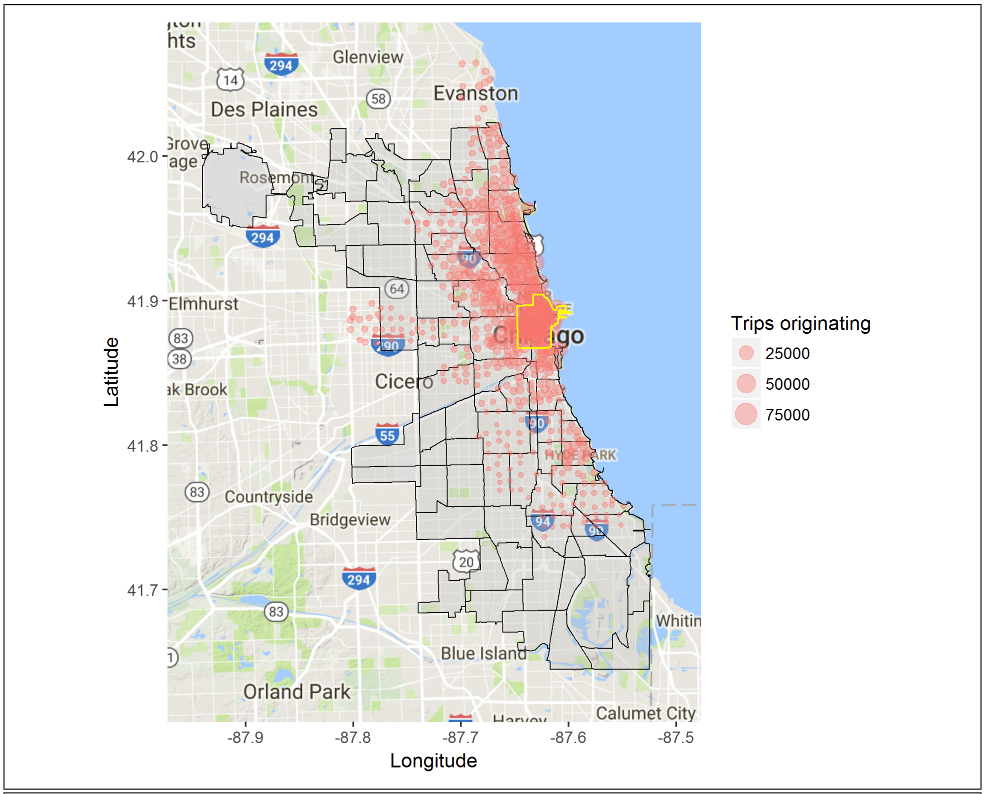
		All Users[1]		Members		Daily Customers	
	Type	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Annual membership	0-1	0.761	-	-	-	-	-
Gender (male)	0-1	-	-	0.748	-	-	-
Age	Count	-	-	35.52	10.75	-	-
Weekday	0-1	0.725	-	0.79	-	0.496	-
Rush hour	0-1	0.445	-	0.527	-	0.184	-
Trip duration (min)	Cont.	16.56	31.54	12.04	20.76	30.96	50.18
male only	Cont.	-	-	11.57	19.87	-	-
female only	Cont.	-	-	13.44	23.12	-	-
Trip distance (m)	Cont.	1407.09	1275.21	1420.73	1263.75	1363.61	1310.14
male only	Cont.	-	-	1410.47	1257.12	-	-
female only	Cont.	-	-	1451.31	1282.65	-	-
Proximity, from (50m)	0-1	0.468	-	0.483	-	0.420	-
from (100m)	0-1	0.725	-	0.744	-	0.665	-
from (200m)	0-1	0.880	-	0.903	-	0.809	-
from (300m)	0-1	0.925	-	0.943	-	0.871	-
Proximity, to (50m)	0-1	0.470	-	0.483	-	0.429	-
to (100m)	0-1	0.726	-	0.744	-	0.668	-
to (200m)	0-1	0.882	-	0.904	-	0.811	-
to (300m)	0-1	0.926	-	0.943	-	0.869	-

[1] N = 3,595,383 for all trips; there are in fact three types of users: `Subscriber` type refers to subscribers to the annual membership, `Customer` refers to customers of the 24-hour daily pass, and `Dependent` type refers to members who are younger than 16 and whose parents purchased the membership. Since there are only 40 trips by the last user type in the current dataset, trips of this type will be mostly ignored in this study.

Table 6: Descriptive summary of all Divvy stations

	Type	Mean	Std.Dev.	Median	Max	Min	N
Trips originating	Count	6,188.27	8,487.86	3,058	90,042	1	581
Trips terminating	Count	6,188.27	8,655.73	3,151	99,495	1	581
Station capacity	Count	17.19	5.56	15	47	11	581
Presence of CTA stops							
Distance $\leq$ 50m	0-1	0.484	-	-	-	-	581
Distance $\leq$ 100m	0-1	0.731	-	-	-	-	581
Distance $\leq$ 200m	0-1	0.888	-	-	-	-	581
Distance $\leq$ 300m	0-1	0.926	-	-	-	-	581
Number of CTA stops							
Distance $\leq$ 50m	Count	0.91	1.19	0	6	0	581
(only stations in proximity)	Count	1.88	1.05	2	6	1	281
Distance $\leq$ 100m	Count	1.98	1.82	2	11	0	581
(only stations in proximity)	Count	2.70	1.60	2	11	1	425
Distance $\leq$ 200m	Count	4.30	2.94	4	16	0	581
(only stations in proximity)	Count	4.85	2.67	4	16	1	516
Distance $\leq$ 300m	Count	7.66	4.60	8	24	0	581
(only stations in proximity)	Count	8.27	4.22	8	24	1	538

Figure 5: All Divvy trips by station



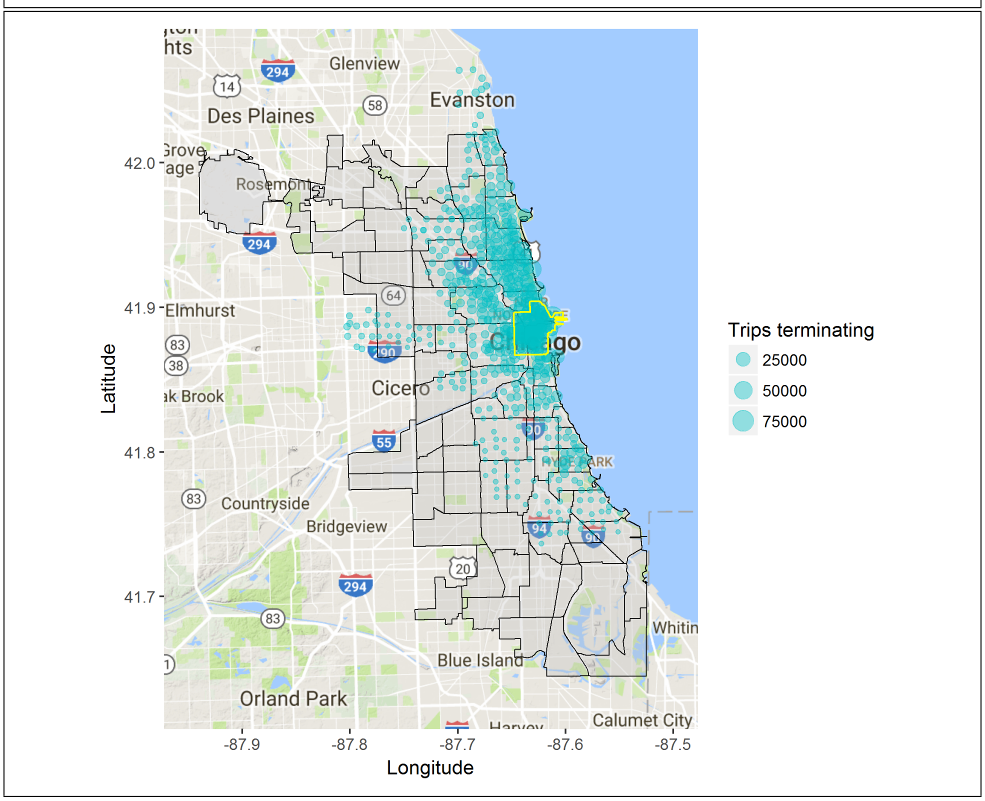


Figure 6: The age distribution of all Divvy trips by annual subscribers

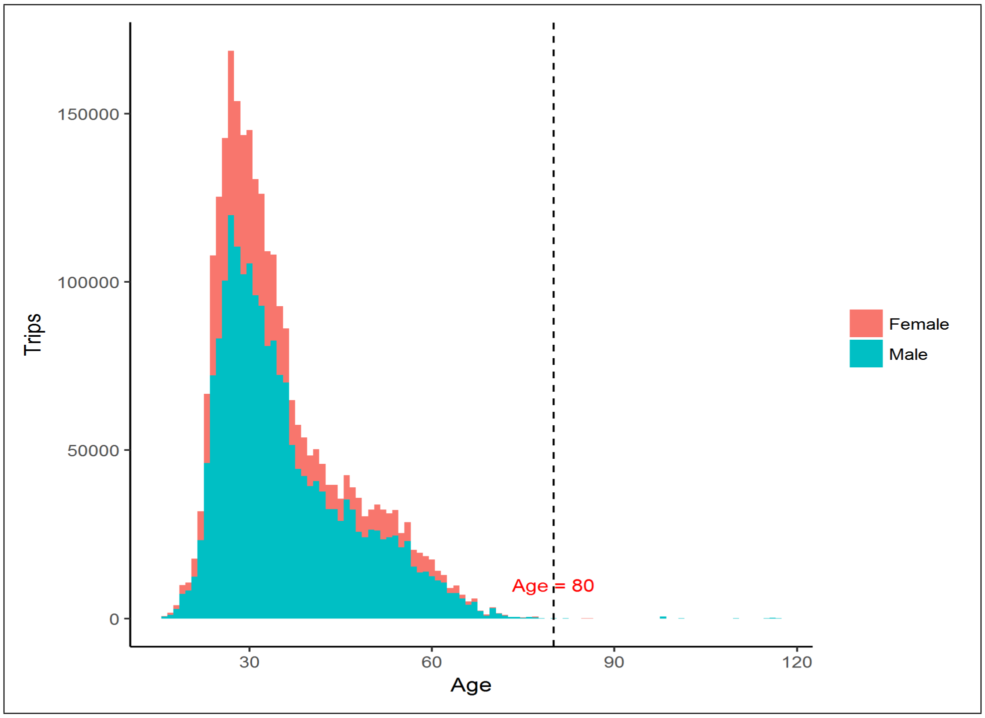
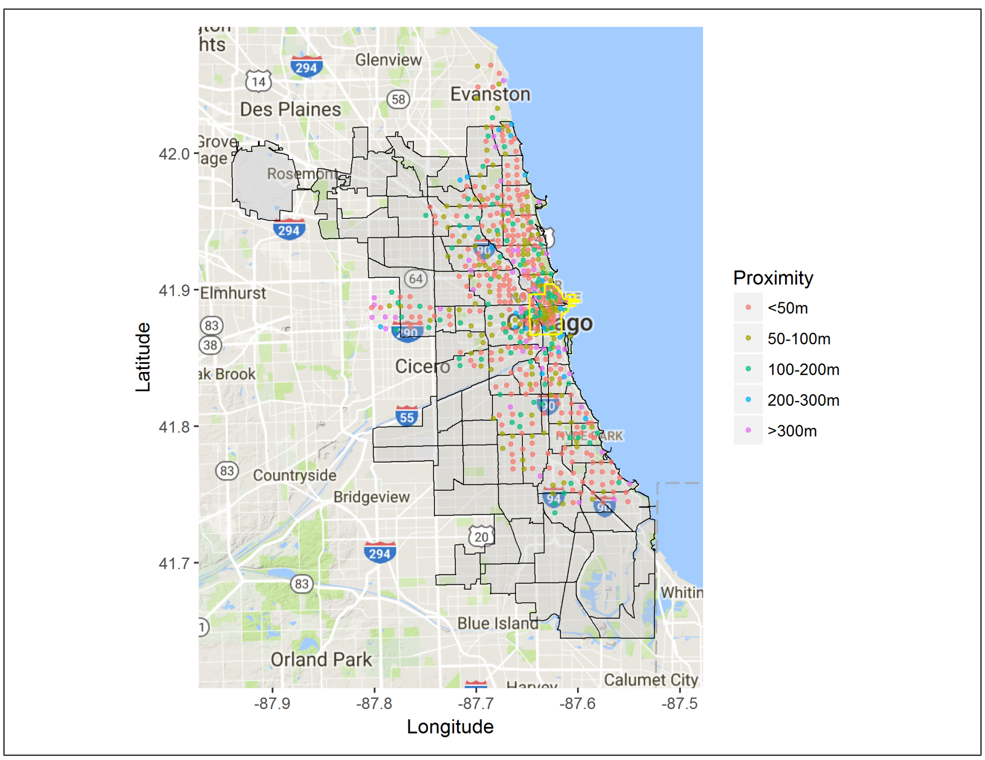


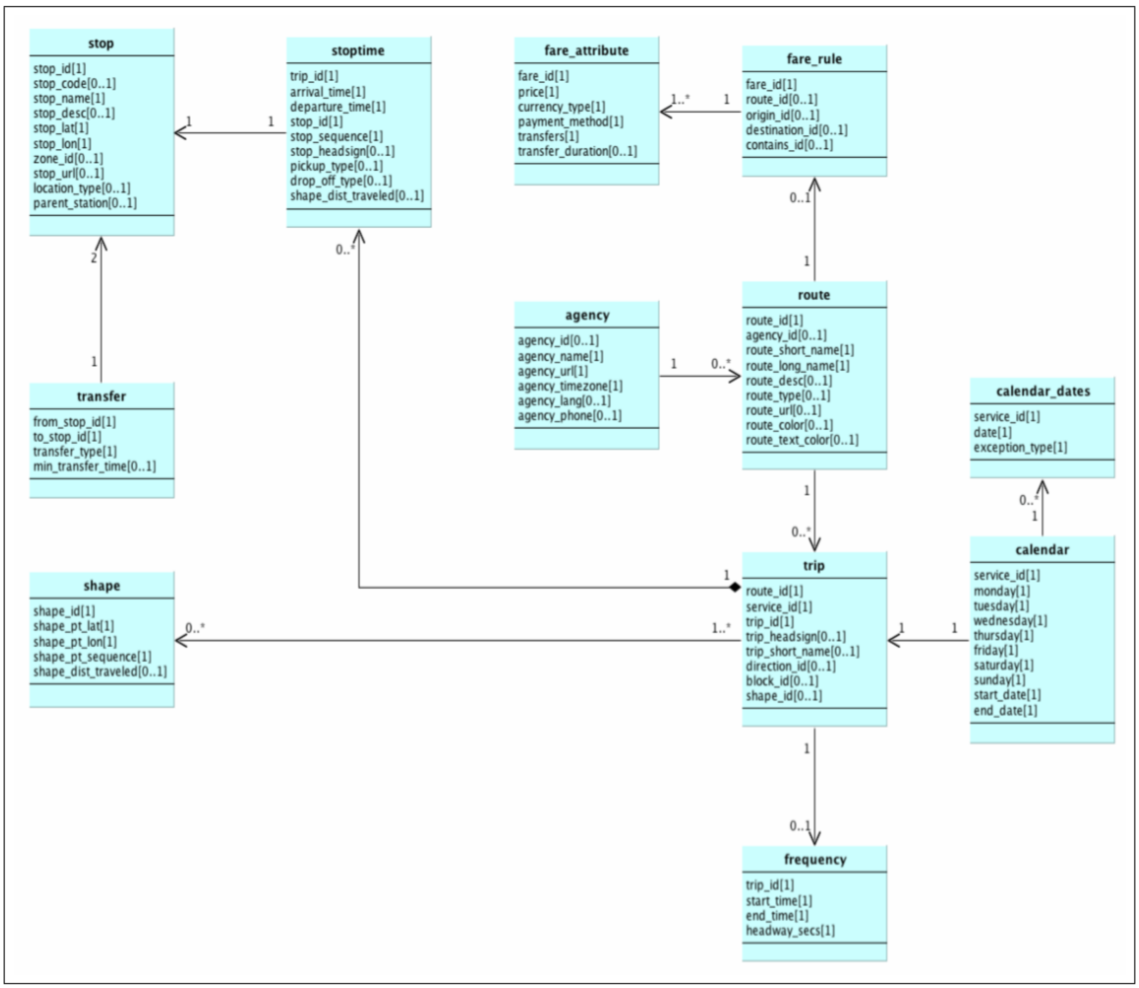
Figure 7: All Divvy stations by proximity



A-1.2 Chicago Transit Authority (CTA) data

CTA provides its data in the General Transit Feed Specification (GTFS) format, first developed by Google and now commonly used for organizing and publishing public transportation data and associated geographic information[8]. In case of CTA data, the full dataset is a relational dataset consisting of eight separate tables. CTA data can be access through its website[9]. For the current study, I use only two tables: one for stop locations data (`stops.txt`) and the other for scheduled arrival and departure times data (`stop_times.txt`). Figure 8 provides an illustration of the relationships among all GTFS format files.

Figure 8: A diagram of the General Transit Feed Specification format: The relationships among the various tables (Google Transit, 2017)



[8]The full reference on the GTFS format can be found here: https://developers.google.com/transit/gtfs/reference/

[9]http://www.transitchicago.com/developers/gtfs.aspx

A-1.3 City of Chicago Data Portal boundaries data

The City of Chicago Data Portal is an online archive of over 200 government datasets about the City's departments, neighborhoods, facilities and services[10]. Any dataset can be freely accessed either programmatically using the Portal's application programming interface (API), or by browsing the Portal website's graphic user interface (GUI). The Portal offers boundaries datasets for all Chicago community areas, census tracts, and the central business district. Each dataset can be found and downloaded in different formats, including KML, KMZ, GeoJSON, and shapefile.

A-1.4 National Centers of Environmental Information (NCEI) data

NCEI serves as the comprehensive archive of all sorts of environmental data for the National Oceanic Atmospheric Administration, a scientific agency within the United States Department of Commerce. As the United States' leading authority for environmental information, NCIE hosts and provides access to comprehensive oceanic, atmospheric, and geophysical data. The NCEI website offers access to its GHCN (Global Historical Climatology Network)-Daily database, which contains historical records for daily temperature, precipitation, snowfall, wind movement, and more[11][12].

Figures 9 summarizes the daily weather in the Chicago area for the year of 2016. The red and blue lines mark daily maximum and minimum temperature, respectively. The dashed horizontal line marks the freezing degree, 32°F. The gray vertical lines mark days with precipitation. The highest maximum atmospheric temperature was 94°F on June 10th, July 24th, August 11th, and September 7th. The lowest minimum temperature was -6°F on December 19th. Out of 366 days, total 116 days (31.69%) had positive precipitation.

A-1.5 American Community Survey (ACS) data

The ACS is a nationwide survey conducted annually by the United States Census Bureau to provide up-to-date social, economic and demographic characteristics about the American population[13]. The Census Bureau maintains American FactFinder website, through which the public can access particular slices of the ACS data[14]. For the current study, I have downloaded the census-tract level population data (B01003) and employment status data (S2301) for Cook County, Illinois.

---

[10] https://data.cityofchicago.org/

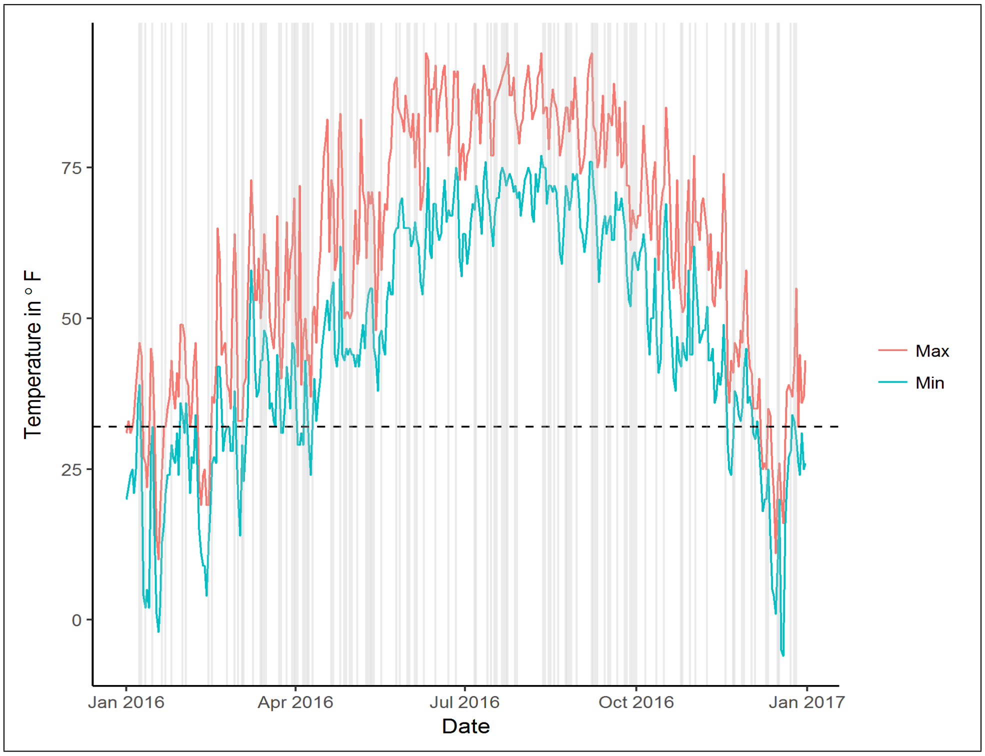
[11] https://www.ncdc.noaa.gov/cdo-web/datasets

[12] The complete documentation of the GHCN-Daily database can be found here: https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/GHCND_documentation.pdf

[13] The official comprehensive guide to the ACS in both English and Spanish languages can be found here: https://www.census.gov/programs-surveys/acs/about/information-guide.html

[14] https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Figure 9: Daily weather in Chicago in 2016



A-1.6 Census tract to Chicago community area data

Because Chicago community areas do not serve as units of data collection for the ACS or any other Census datasets, it is necessary to spatially aggregate Census tract-level data to the level of Chicago community areas. For the current study, I used an unofficial data on converting 2010 Census tracts into Chicago community areas because, unfortunately, I could not find any official source for the same data[15]. Table 7 below presents a quick summary statistics for the characteristics of Chicago community areas.

Table 7: Descriptive summary of Chicago community areas

	Mean	Std.Dev.	Median	Max	Min	N
Population	117,218.62	77,840.28	97,662.00	400,778.00	4,914.00	77
Population density (km^2)	22,390.93	25,723.93	16,760.65	155,017.49	689.84	77
Population, black (%)	39.03	40.44	13.34	100.00	0.32	77
Population, employed (%)	55.00	11.40	56.69	79.53	32.84	77
Income per capita ($)[1]	26,917.55	16,144.58	22,471.57	92,452.44	7,098.20	77

[1] In 2015 Inflation-adjusted dollars.

[15] The unofficial source of data for converting 2010 Census tracts into Chicago community areas can be found here: http://robparal.blogspot.com/2012/04/census-tracts-in-chicago-community.html