Model / Data / Initial Results
MACS 30200, Dr. Evans and Dr. Soltoff
Bobae Kang
May 17, 2017

In this assignment, I present a draft of the Model, Data, and Result sections.

1 Model

Currently, to the best of my knowledge, no data exist on actual multi-modal trip behaviors connecting Divvy and CTA. 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 which also involves riding a CTA bus or rail to arrive at the destination. Therefore, until we can directly observe multi-modal trips, we need to use an alternative approach to labeling multi-modal Divvy trips.

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 the utility according to some form of utility function. 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 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 travel time is negatively correlated with the utility of the passenger. Following such reasoning, a trip can be label PMM if it exhibits expected properties of a multi-modal trip and, therefore, 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 about the time when bus or 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. In other words, 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 closet 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.

In this study, I use four different proximity standards, $p \in \{50m, 100m, 200m, 300m\}$. For measuring distances, I use Manhattan distance 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. I fit four separate models for the four proximity standards and compare the results. I also consider two stages of each trip separately, $s \in \{origination, termination\}$, and 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.

The 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 shorty 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 somewhat arbitrarily but conservatively, 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 assume that PMM Divvy trips are 30-minutes long or shorter. This limitation on the duration of Divvy trips is based on the following two factors. 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[1]. 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.

In this study, I use the traditional binomial logistic regression to model PMM Divvy trips. The probability that the stage s of a trip i is PMM at the given proximity standard p is given in the following form:

$$\begin{aligned} Pr(Y_{i,p,s} = y_{i,p,s} | \pi_{i,p,s}) &= \pi_{i,p,s}^{y_{i,p,s}} (1 - \pi_{i,p,s})^{(1-y_{i,p,s})}, \\ \text{where} \quad \pi_{i,p,s} &= \frac{\exp(\beta' \mathbf{X}_{i,p,s})}{1 + \exp(\beta' \mathbf{X}_{i,p,s})} \end{aligned} \tag{1}$$

where $y_{i,p,s} = 1$ if the given trip is PMM at the given stage s and proximity p, and 0 otherwise.

The likelihood function is then given in the following form:

$$L = \prod_{i=1}^{N} \pi_{i,p,s}^{Y_{i,p,s}} (1 - \pi_{i,p,s})^{1-Y_{i,p,s}} \tag{2}$$

And, therefore, the log-likelihood function can be defined as:

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

$$\ln L = \sum_{i=1}^{N} Y_{i,p,s} \ln(\pi_{i,p,s}) + (1 - Y_{i,p,s}) \ln(1 - \pi_{i,p,s}) \tag{3}$$

Here, the model parameters β is estimated by maximizing the log-likelihood function.

2 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 a sample used in the current study. The rest of this section consists of a brief introduction to Divvy datasets, a detailed account of the process of obtaining a trip sample for fitting the model, an explanation on key variables, and a descriptive analysis of the sample.

Divvy data per-trip data consists the following information for each trim: 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 station data, which includes information on total 581 Divvy stations.

Although the focus of current study is different from that of Faghih-Imani and

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

Eluru (2015), Faghih-Imani and Eluru (2015) offers an exemplary case as to forming a sample set for examining Divvy trips. Therefore, I have followed some of their sample formation steps in the current study while adding original steps to better serve the 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 trips). 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). Because 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 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).

Fourth, I have deleted trips longer than 90 minutes in duration (only 1.01% of all trips). Not only trips longer than 90 minutes are atypical bike share rides, but also they could result from misplaced bikes. Fifth, this study do 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 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 25,000 trips, one for trips originating (stage $s = origination$) and the other for trips terminating (stage $s = termination$), are randomly selected from the *clean* dataset obtained from the aforementioned five steps, which consists of total 3,434,320 trips or approximately 95.5% of all observations. For each set, among these 25,000 trips, a randomly selected subset of size 20,000 is used to fit the model while the rest 5,000 are set aside to validate the fitted model.

Tables 2 and 3 offer descriptive summaries of the sample Divvy trips for fitting the model and the Divvy stations relevant to the sample. Table 2 is divided into two panels, *Panel A* on the top for trips originating and *Panel B* on the bottom for trips terminating. In both cases, 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. 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. 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

Table 2: Descriptive summary of Divvy trips (training set)

Panel A. Trips originating

	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	-	-	-	15,563[1]
Age	Count	35.40	10.57	32	77	16	15,563[1]
Weekday	0-1	0.73	0.44	-	-	-	20,000
Rush hour	0-1	0.45	0.50	-	-	-	20,000
Trip duration (min)	Cont.	14.32	10.70	11.52	89.80	1.13	20,000
Trip distance (m)	Cont.	1,457.30	1,251.74	1,135.47	15,618.66	15.16	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	0.89	0.32	-	-	-	20,000
Distance $\leq$ 300m	0-1	0.93	0.26	-	-	-	20,000
Potential multi-modality[4]							
Distance $\leq$ 50m	0-1	0.21	0.40	-	-	-	20,000
Distance $\leq$ 100m	0-1	0.36	0.48	-	-	-	20,000
Distance $\leq$ 200m	0-1	0.53	0.50	-	-	-	20,000
Distance $\leq$ 300m	0-1	0.64	0.48	-	-	-	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	-	-	-	15,571[1]
Age	Count	35.47	10.53	32	77	16	15,571[1]
Weekday	0-1	0.74	0.44	-	-	-	20,000
Rush hour	0-1	0.46	0.50	-	-	-	20,000
Trip duration (min)	Cont.	14.29	10.80	11.5	89.75	1.07	20,000
Trip distance (m)	Cont.	1,451.83	1,246.75	1,132.11	9,828.466	5.43	20,000
Proximity to CTA[3]							
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	0.26	-	-	-	20,000
Potential multi-modality[5]							
Distance $\leq$ 50m	0-1	0.20	0.40	-	-	-	20,000
Distance $\leq$ 100m	0-1	0.37	0.48	-	-	-	20,000
Distance $\leq$ 200m	0-1	0.54	0.50	-	-	-	20,000
Distance $\leq$ 300m	0-1	0.64	0.48	-	-	-	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.

stops, the proportion of such trips increases with the increasing proximity standard p, which makes intuitive sense. The same holds for the proportion of PMM Divvy

Table 3: Descriptive summary of Divvy stations (training set)

	Type	Mean	Std.Dev.	Median	Max	Min	N
Trips originating[1]	Count	41.49	48.82	25.5	486	1	20,000
Potentially multi-modal							
Distance $\leq$ 50m	Count	18.06	26.16	9	220	1	4,100
Distance $\leq$ 100m	Count	21.77	27.63	11	217	1	7,162
Distance $\leq$ 200m	Count	26.43	33.51	14	277	1	10,809
Distance $\leq$ 300m	Count	28.94	34.59	16	279	1	12,733
Trips terminating[2]	Count	41.67	51.48	27	552	1	20,000
Potentially multi-modal							
Distance $\leq$ 50m	Count	18.17	28.70	9	286	1	4,053
Distance $\leq$ 100m	Count	21.80	29.45	11	257	1	7,348
Distance $\leq$ 200m	Count	26.57	35.95	13	343	1	10,816
Distance $\leq$ 300m	Count	29.44	37.44	16	357	1	12,806
Station capacity	Count	17.76	5.71	15	47	11	500
Located in CBD	0-1	0.17	0.37	-	-	-	500
Presence of CTA stops							
Distance $\leq$ 50m	0-1	0.49	0.50	-	-	-	500
Distance $\leq$ 100m	0-1	0.74	0.44	-	-	-	500
Distance $\leq$ 200m	0-1	0.90	0.31	-	-	-	500
Distance $\leq$ 300m	0-1	0.94	0.24	-	-	-	500
Number of CTA stops							
Distance $\leq$ 50m	Count	0.93	1.22	0	6	0	500
Close stations only	Count	1.91	1.08	2	6	1	245
Distance $\leq$ 100m	Count	2.00	1.81	2	11	0	500
Close stations only	Count	2.70	1.59	2	11	1	370
Distance $\leq$ 200m	Count	4.32	2.95	4	16	0	500
Close stations only	Count	4.82	2.70	4	16	1	448
Distance $\leq$ 300m	Count	7.81	4.63	8	24	0	500
Close stations only	Count	8.32	4.31	8	24	1	469

[1] Total 482 docking stations.
[2] Total 480 docking stations.

trips.

Table 3 presents summary statistics of Divvy stations associated with trips in the training set. Some notable points shown in this table are as follows: The sample trips originated from 480 docking stations and terminated at 480 docking stations. The union of these two sets of docking stations consists of total 500 stations. For trips originating, each station has 41.49 trips on average that started from that station. For trips terminating, each station has 41.67 trips on average that stopped at that station. The mean docking capacity for Divvy stations is 17.76 while the median capacity is 15. 17% of all Divvy stations are located within the central business district (CBD). About a half all all Divvy stations have at least one CTA stop within

$50m$. With different proximity standard, the proportion of Divvy stations that are close to CTA stops increase. With $300m$ 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 1.91 nearby CTA stops for Divvy stations that have at least one CTA stop within the $50m$ range. This number also rises with the increasing proximity standard. With the $300m$ 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.

Following Figures 1 and 2 provide geographical illustrations of PMM trips by proximity standard p, the former for trips originating and the latter for trips terminating. In each map, the size of a 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 CBD. As suggested in Table 2, the number of trips that are marked PMM grow with the proximity standard for both trip stages s. The graphics further illustrate that most of the PMM trips can be found in CBD and the surrounding areas, which makes intuitive sense because of the geographical distribution of docking stations (shown in Figure 7 in Appendix A-1).

3 Results

In this section, I present the results for binomial logit model estimation to understand different factors that characterize potentially multi-modal (PMM) Divvy trips and, by extension, may influence Divvy users' decision to make multi-modal trips in reality. As suggested in Sections 1 and 2 above, the model estimation uses two separately generated sample training 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 have estimated four separate models by proximity standard $p \in \{50m, 100m, 200m, 300m\}$. In addition, I compare the coefficients of fitted models across different p to test the

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

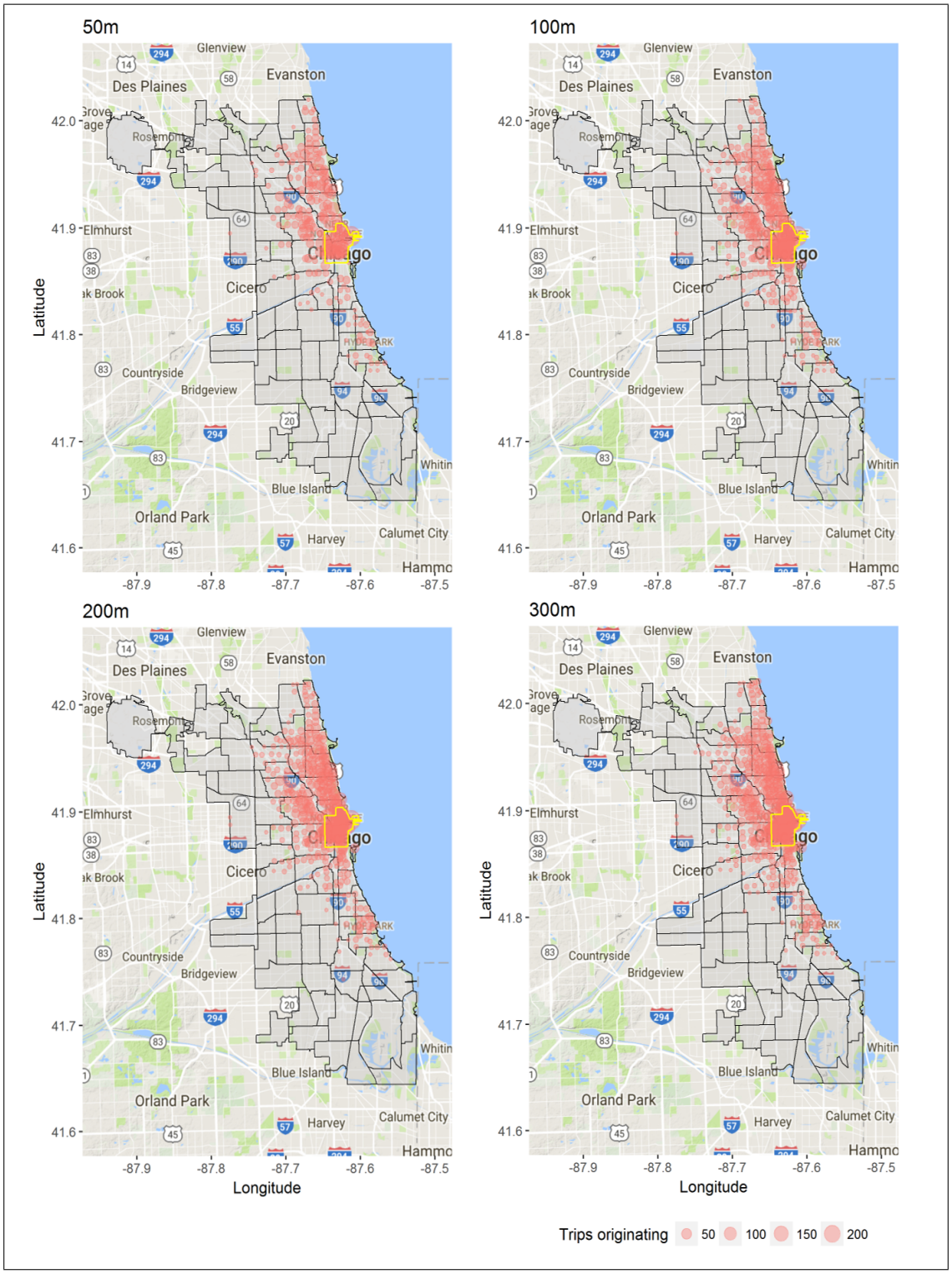
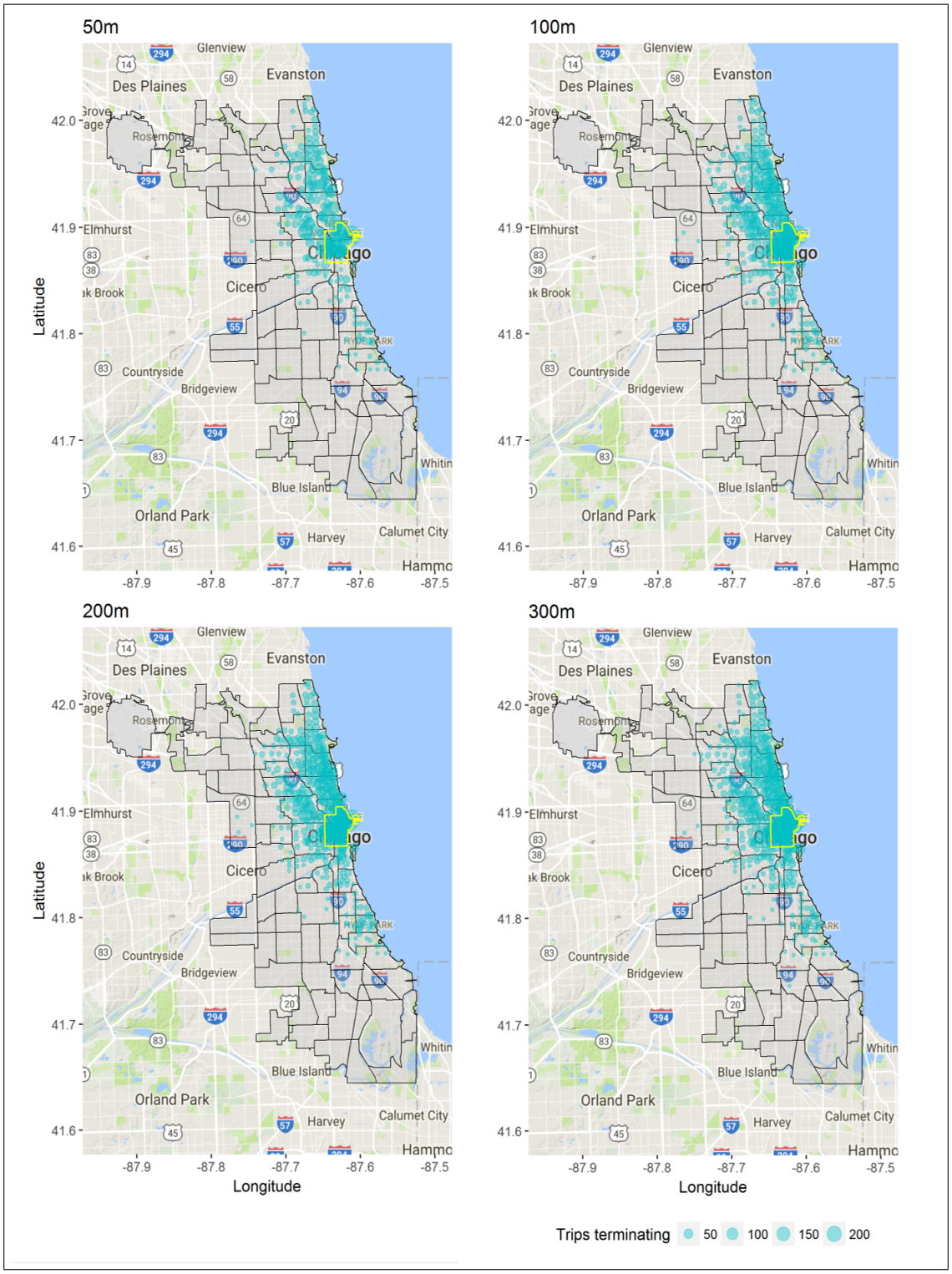


Figure 2: Potentially multi-modal Divvy trips by proximity standard: Trips terminating



validity of the $300m$ standard used in Faghih-Imani and Eluru (2015) to evaluate the impact of access to public transit on Divvy trips.

Table 4 summarizes model estimation results for all eight models; *Panel A* on the top presents estimation results for trips originating by proximity standard p and *Panel B* on the bottom, for trips terminating by p. All models have the same set of explanatory variables: the first four ("Subcriber," "Weekday," "Rush hour," and "Trip distance (km)") are trip-level features, the next two ("Station in CBD" and "Population density") are station-level features. Standard error for each coefficient estimate is given in parentheses and statistical significance of each estimate is marked by the number of asterisks. The Table also reports the log-likelihood values for the fitted models for each trip stage. In general, a model with higher log-likelihood score is preferred. In the current context, however, comparison between models across proximity standard p or trip stage s cannot be made based on the log-likelihood score since each model has a distinct response variable.

For each trip stage s, estimated coefficients and their statistical significance vary across proximity standard p. The constant term is negative and statistically significant across all p for both s, but the magnitude decreases with increasing p. Conversely the coefficient estimate for "Subscriber" variable, a binary dummy for Divvy user type, is positive and statistically significant across all p for both s, but its size increases with greater p. "Weekday" variable is another binary dummy for trips made during weekdays (Monday to Friday). The coefficient estimate for "Weekday" appear statistically significant only in the models with $p = 200m, 300m$ for trips originating and $p = 300m$ for trips terminating. "Rush hour" variable is a binary dummy for trips made during rush hours as defined in Section 2 above. For trips originating, the coefficient estimate for rush hour are all positive but statistically significant only for $p = 50m, 300m$. For trips terminating, the estimates for $p = 50m, 100m, 200m$ are statistically significant. "Trip distance (km)" is a continuous variable for the Manhattan distance between the originating station and terminating station for each trip in kilometers. The coefficient estimate for "Trip distance (km)" is negative and statistically

Table 4: Model estimation by proximity standard

Panel A. Trips originating

Parameter/fit	50m	100m	200m	300m
Constant	-1.755***	-1.0680***	-0.6331***	-0.3397***
	(0.0582)	(0.0487)	(0.0461)	(0.0465)
Subscriber	0.2660***	0.4170***	0.6400***	0.7917***
	(0.0472)	(0.0395)	(0.0371)	(0.0374)
Weekday	-0.0413	0.0530	0.1404***	0.1135***
	(0.0493)	(0.0411)	(0.0394)	(0.0405)
Rush hour	0.1577***	0.0390	0.0148	0.0763**
	(0.0432)	(0.0361)	(0.0355)	(0.0372)
Trip distance (km)	0.0223	-0.0174	-0.0459***	-0.0686***
	(0.0140)	(0.0120)	(0.0116)	(0.0119)
Station in CBD	0.0477	0.2353***	0.5235***	0.4936***
	(0.0361)	(0.0305)	(0.0300)	(0.0316)
Population density	0.0737***	0.0177	0.0301**	0.0661***
	(0.0165)	(0.0139)	(0.0136)	(0.0143)
Log-likelihood	-10,103.17	-12,936.76	-13,419.65	-12,641.22

Panel B. Trips Terminating

Parameter/fit	50m	100m	200m	300m
Constant	-1.8049***	-0.9827***	-0.5909***	-0.2949***
	(0.0595)	(0.0491)	(0.0469)	(0.0474)
Subscriber	0.2403***	0.4274***	0.6015***	0.7772***
	(0.0472)	(0.0393)	(0.0372)	(0.0376)
Weekday	0.0176	-0.0383	0.0575	0.0795**
	(0.0498)	(0.0413)	(0.0398)	(0.0409)
Rush hour	0.1055**	0.1031***	0.1202***	0.0457
	(0.0429)	(0.0359)	(0.0352)	(0.0369)
Trip distance (km)	0.0029	-0.0389***	-0.0685***	-0.0608***
	(0.0143)	(0.0121)	(0.0117)	(0.0120)
Station in CBD	0.1037***	0.2573***	0.5558***	0.5594***
	(0.0363)	(0.0304)	(0.0301)	(0.0318)
Population density	0.1034***	0.0216	0.0452***	0.0488***
	(0.0165)	(0.0138)	(0.0136)	(0.0142)
Log-likelihood	-10,034.34	-13,029.28	-13,404.87	-12,623.93

p-value * 0.1 ** 0.05 *** 0.001

significant for $p = 200m, 300m$ for trips originating and for $p = 100m, 200m, 300m$ for trips terminating.

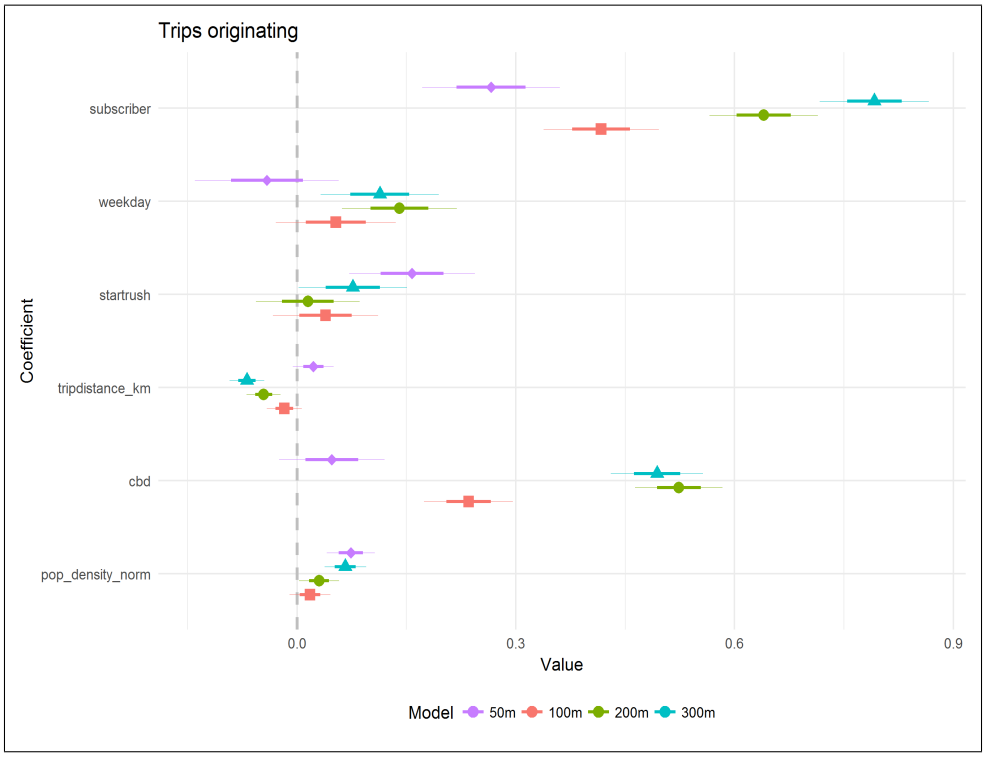
The coefficient estimate for "Station in CBD", the binary dummy that marks whether the station is located within the central business district (CBD), is positive and statistically significant across all models except for the one model with $p = 50m$ for trips originating. The magnitude of these estimates vary somewhat significantly. Lastly, "Population density" is a continuous variable for the standardized population

density of the community area where the originating or terminating station is located, with mean of 0 and standard deviation of 1. Here, the coefficient estimates are all positive and statistically significant for models with $p = 50m, 200m, 300m$ for both trips originating and trips terminating.

For the current project, perhaps more interesting is whether the coefficients for each variable are sufficiently different from one another across proximity standards. That is, if the variation in estimated coefficients between the $p = 300m$ and the others proximity standards is negligible, we may conclude that the $300m$ is indeed a viable choice. If we observe some significant variation in coefficients, this may suggest that there is a need to reconsider the validity of $300m$ as the proximity standard. Figures 3 and 4 offer graphical comparison of estimated coefficients for all variables across different proximity standards. Careful examination of these figures can shed light on whether the typical proximity standard of $300m$ 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 think horizontal line marks the interval of one standard deviation from the given estimate, and the thin line marks the interval of two standard deviations or corresponds to the 95% confidence interval. The vertical dashed line marks 0, which marks that the given explanatory variable is not correlated with the response variable, i.e. the potential multi-modality of a Divvy trip.

In Figure 3, we observe that, for the trips originating, coefficients of "Subscriber (`subscriber`)," "Trip distance (`tripdistance_km`)," and "Station in CBD (`cbd`)" show sufficient variations across proximity standards. For "Subscriber" variable, the estimated coefficients are sufficiently different (i.e. beyond the 95% confidence interval) between all pairs of p, except for the pair of $50m$ and $300m$. For "Trip distance" variable, only the coefficient estimate for $p = 50m$ sufficiently different from all the others. However, in this case, the estimated coefficient for $p = 50m$ fails to reject the null hypothesis, which is shown in the illustration by the point for $p = 50m$ on the vertical dashed line where the coefficient value equals 0. Meanwhile, the pair of $100m$ and $300m$ shows meaningful difference. In the case of "Station in CBD,"

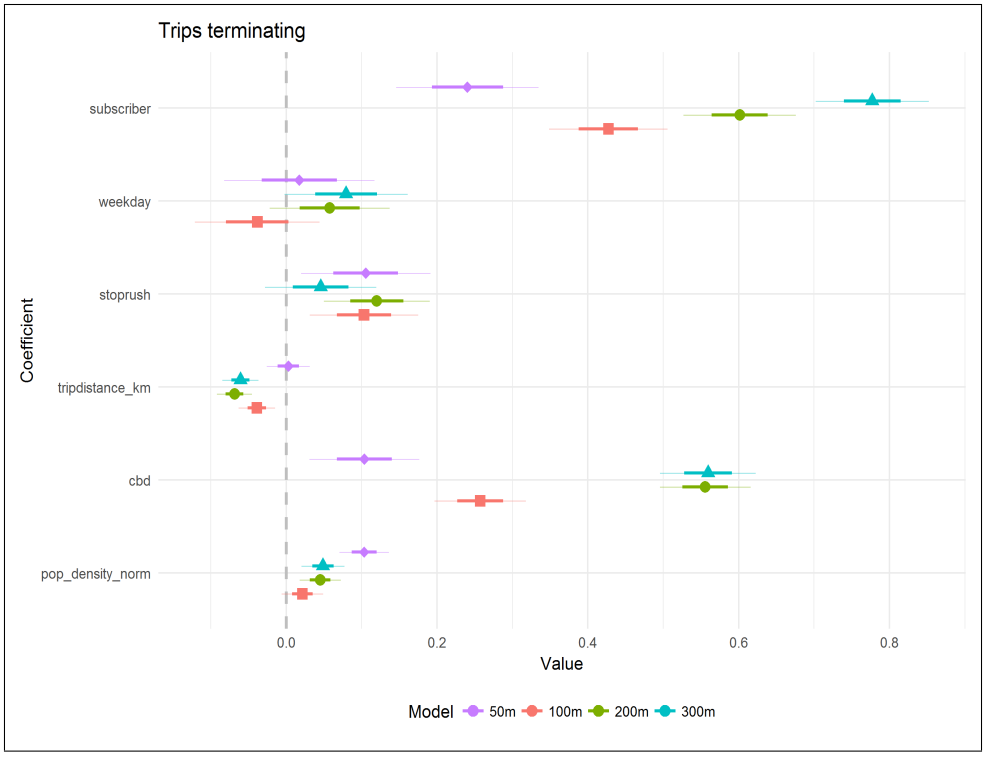
Figure 3: Comparing coefficient estimates across proximity standards: Trips originating



the coefficient for $p = 50m$ again fails to reject the null. The estimated coefficients for both $100m$ and $200m$ shows notable difference from the coefficient for $300m$. The coefficients for the other variables, namely "Weekday (`weekday`)," "Rush hour (`startrush`)," and "Population density (`pop_density_norm`)," do not have any pair of p with sufficient difference.

Figure 4 illustrates that the models for trips terminating show a similar pattern. Just as before, the estimated coefficients of "Subscriber (`subscriber`)" and "Station in CBD (`cbd`)" display sufficient variations across proximity standards. However, for trips terminating, the same is no longer the case for the coefficient of "Trip distance (`tripdistance_km`)." Instead, the coefficients of "Population density (`pop_density_norm`)" exhibits a variation beyond the 95% confidence interval for the pair of $50m$ and $300m$.

Figure 4: Comparing coefficient estimates across proximity standards: Trips terminating



References

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.

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

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[2]. Divvy Data is subject to the terms and conditions of the Data License Agreement[3].

Table 5 offers a summary statistics of all divvy trips made in 2016, which amount to total 3,595,383 observations, and Table 6 provides a summary statistics of all 581 divvy stations that were in operation 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. Figure 7 shows all Divvy stations color-coded by their proximity to the closest CTA station.

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

[3] 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[2]		Daily Customers[3]	
	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[4]	Cont.	-	-	11.57	19.87	-	-
female only[5]	Cont.	-	-	13.44	23.12	-	-
Trip distance (m)	Cont.	1407.09	1275.21	1420.73	1263.75	1363.61	1310.14
male only[4]	Cont.	-	-	1410.47	1257.12	-	-
female only[5]	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.
[2] N = 2,7368,69 for trips by annual members.
[3] N = 858,474 for trips by daily customers.
[4] N = 2,047,174 for trips by male members.
[5] N = 689,780 for trips by female members.

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

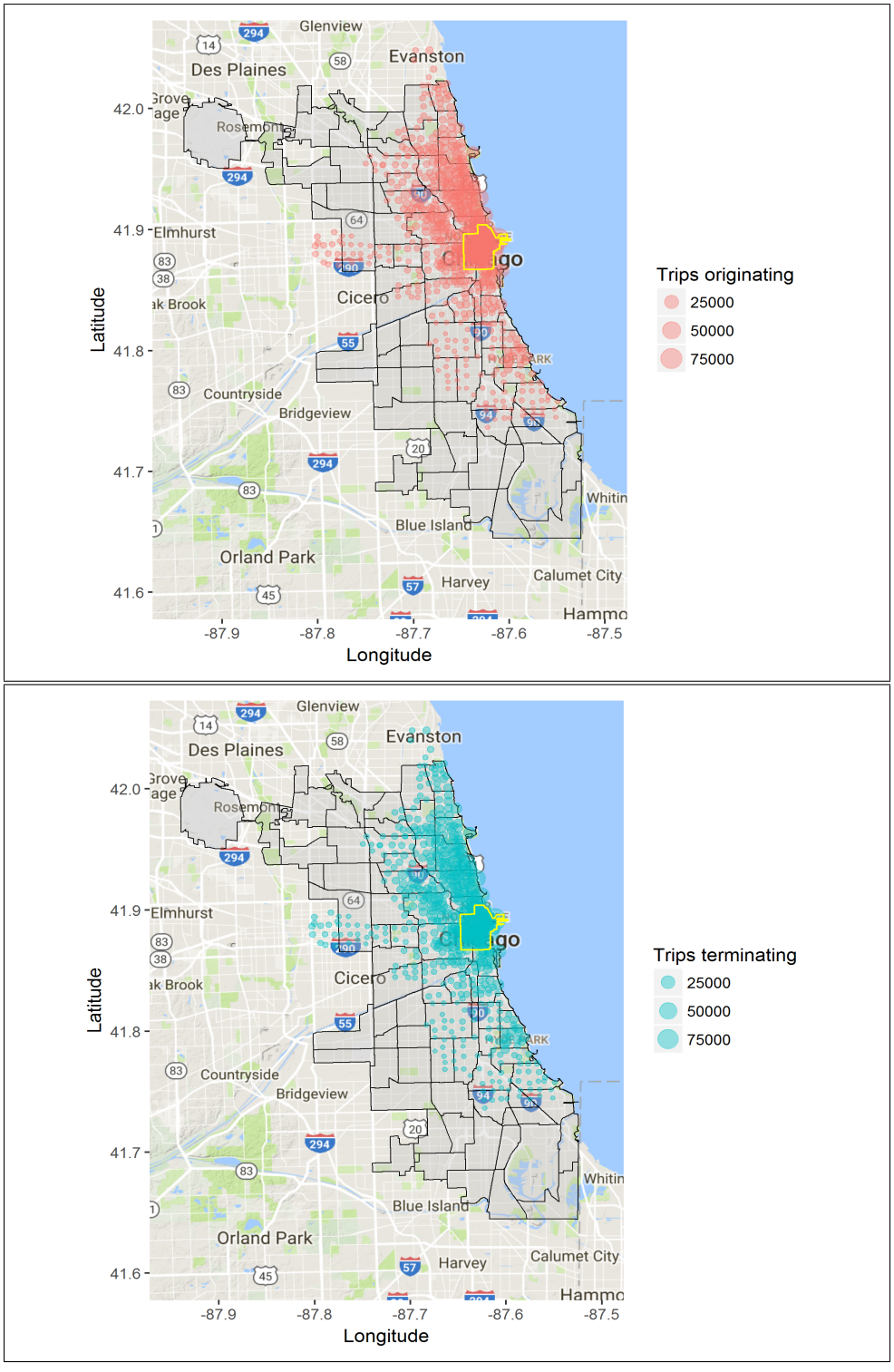


Figure 6: The age distribution of all Divvy trips by annual subscribers

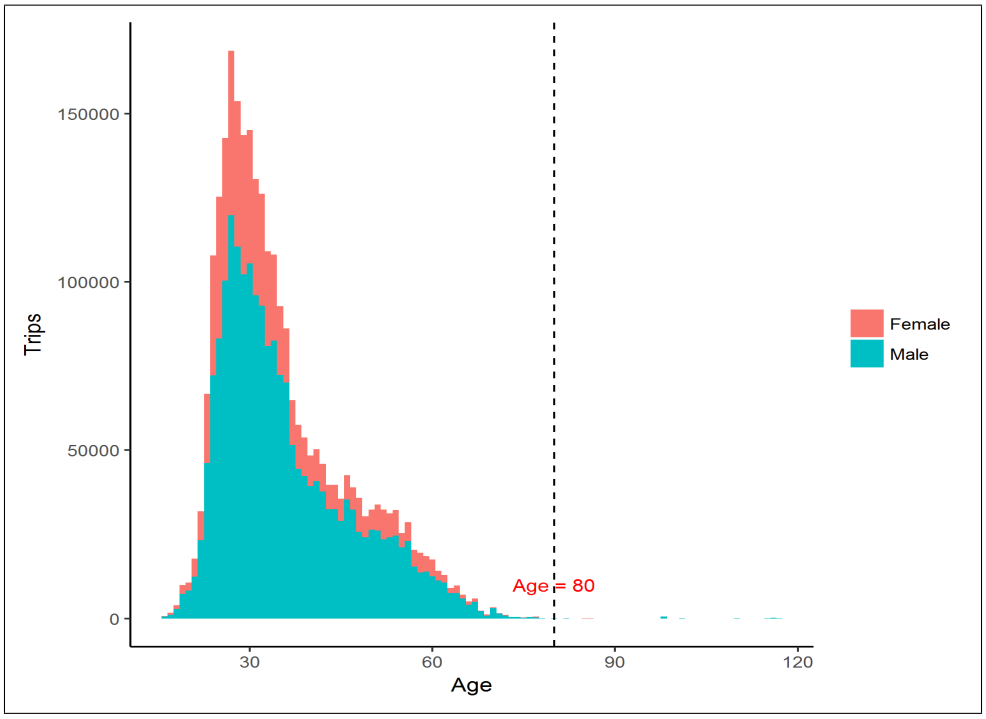
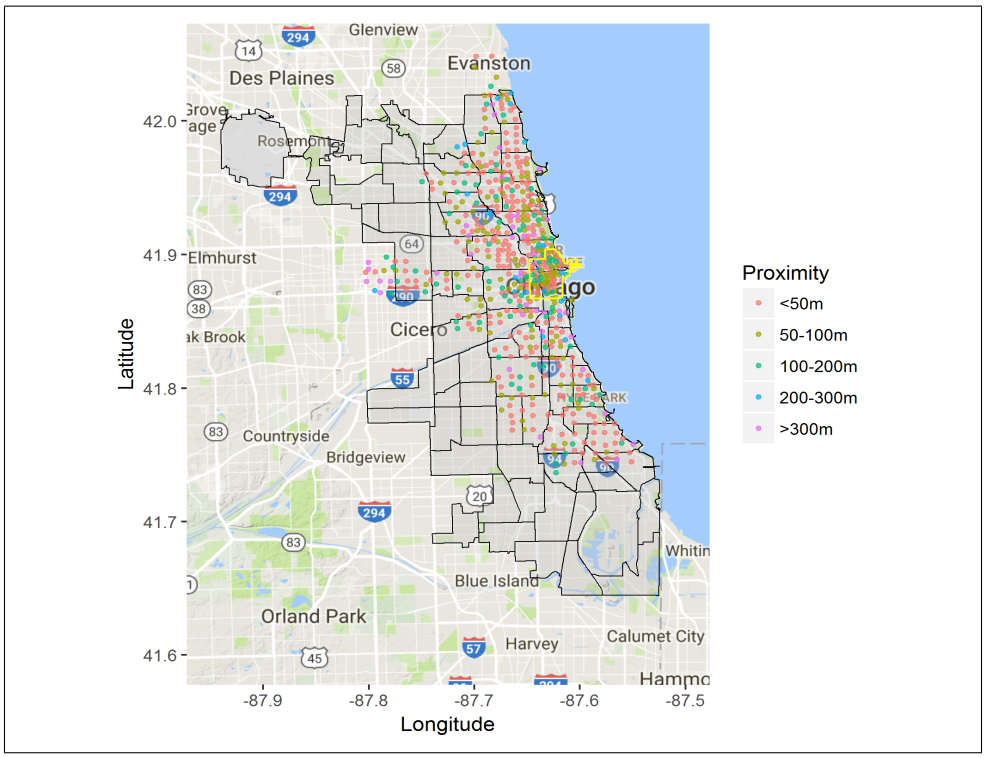


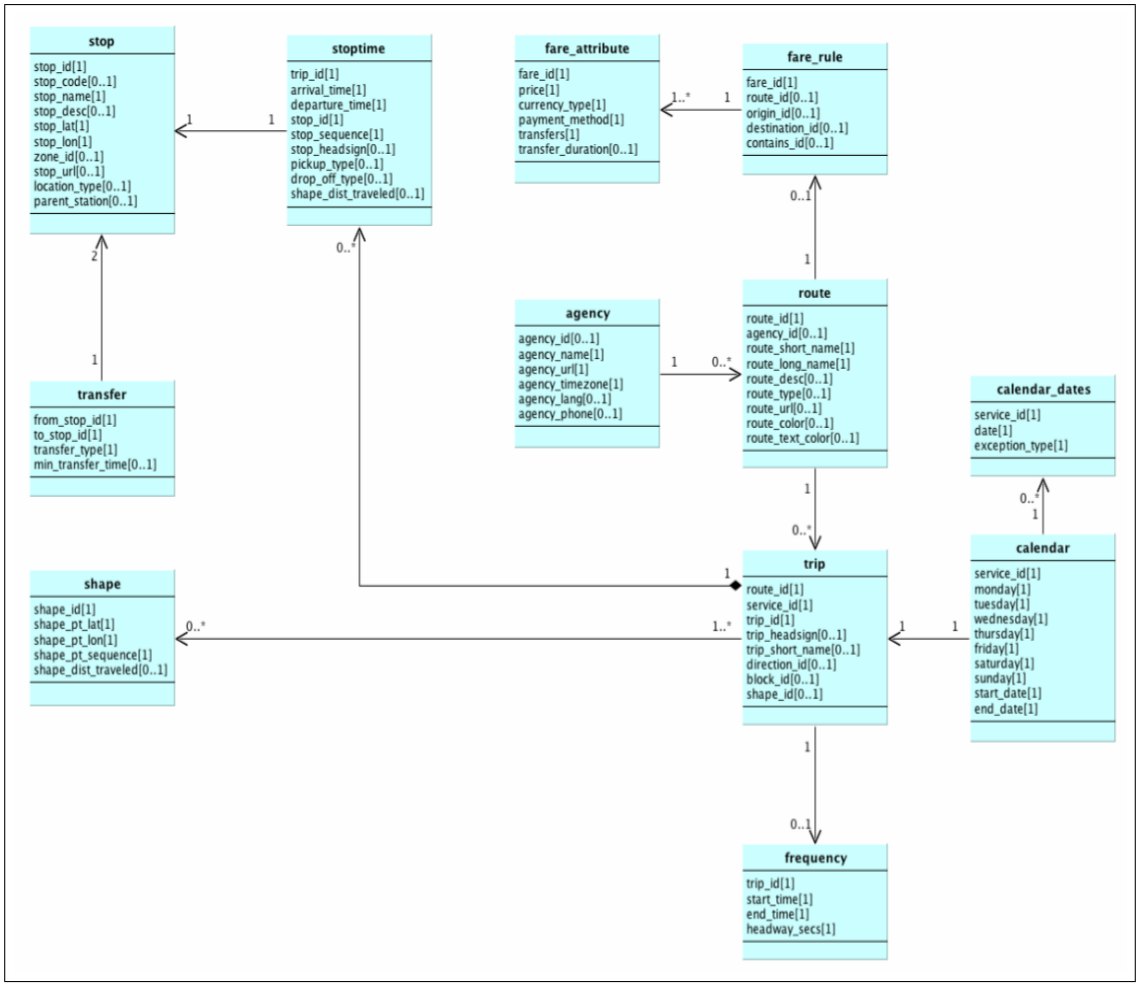
Figure 7: All Divvy stations by proximity



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[4]. 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[5]. 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`). Image 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)



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

[5]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[6]. 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' learning 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[7][8].

Figures 9 summarizes the daily weather in the Chicago area for the year of 2016. The highest maximum atmospheric temperature was 94F on June 10th, July 24th, August 11th, and September 7th. The lowest minimum temperature was -6F 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 an nationwide survey conducted annually by the United States Census Bureau to provide up-to-date social, economic and demographic characteristics about the American population[9]. The Census Bureau maintains American FactFinder website, through which the public can access particular slices of the ACS data[10]. For the current study, I downloaded the census-tract level population data (B01003) and employment status data (S2301) for Cook County, Illinois.

A-1.6 Census tract to Chicago community area data

Because Chicago community areas are not units of data collection for the ACS or any other Census datasets, it is necessary to spatially aggregate Census tract-level

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

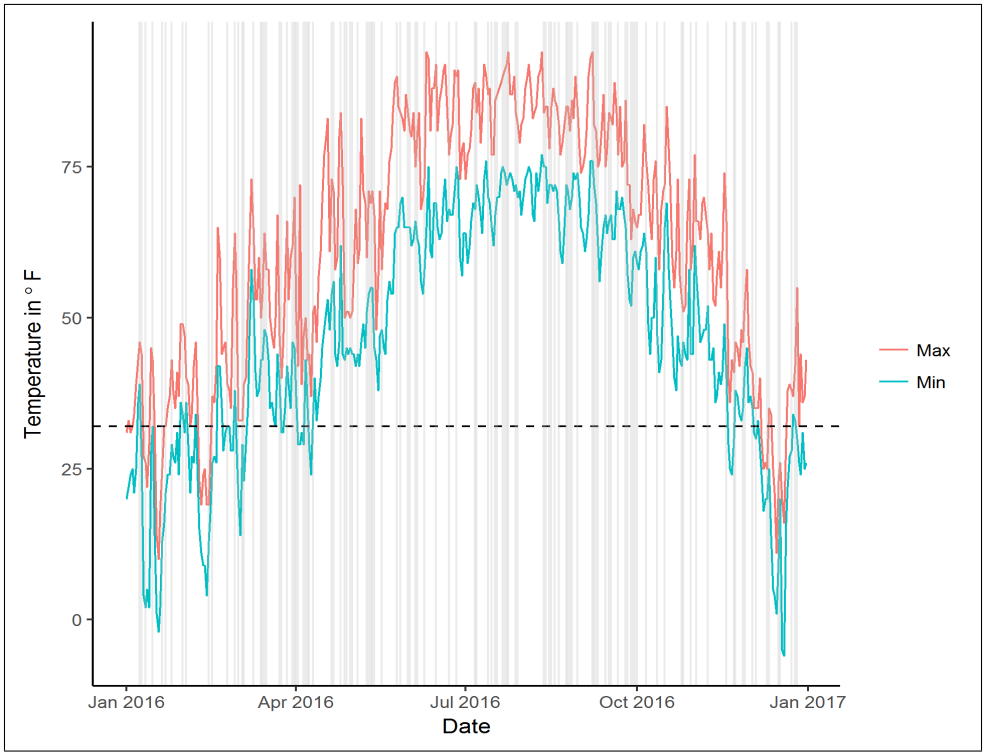
[7] https://www.ncdc.noaa.gov/cdo-web/datasets

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

[9] 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

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

Figure 9: Daily weather in Chicago in 2016
The red and blue lines mark daily maximum and minimum temperature, respectively. The dashed horizontal line marks the freezing degree, 32F. The gray vertical lines mark days with precipitation.



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[11]. Table 7 below presents a quick summary statistics for the characteristics of Chicago community areas.

[11]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

Table 7: Descriptive summary of Chicago community areas

	Mean	Std.Dev.	Median	Max	Min	N
Area (sq.ft.)	83,614,533	54,946,255	79,635,753	371,835,608	16,913,961	77
Population, all	35,867	22,381.46	32,156	98,212	2,457	77
Population, employed	16,781.9	13,475.26	13153.5	70,746	683.6	77
Normalized population density, all	0	1	-0.207	2.702	-1.728	77
Normalized population density, employed	0	1	-0.307	3.644	-1.286	77