.

Bike-Sharing Systems and the Transportation Modal Choice Problem: A Natural Experiment in New York City

A Thesis

Presented to

The Faculty of the Department of Economics

Bates College

In partial fulfillment of the requirements for the

Degree of Bachelor of Arts

by

Christopher J Simard Jr

Lewiston, Maine

December 13, 2019

.

Acknowledgements

I feel the deepest gratitude towards my parents Tara and Chris, my brother Andrew, and my grandparents for their support. It is through your unconditional love that I have the confidence and tenacity to test the limits of my ability. I am grateful for everything I have learned from each of you and I present this thesis to express my thanks.

I also want to thank my family at Bates College. I am grateful for the past three years we have spent together and I am thankful to be a part of your lives. I look forward to our last semester together and I am excited to see where life takes us next.

I owe a debt of gratitude to the many mentors who have helped shape my ambitions to pursue a career in scholarship. I thank Silvia Palano and Zubin Siganporia of Oxford University for the time and intellectual effort they devoted to our weekly meetings. It is because of you both that I will always look back fondly on my experience at the university, and this thesis is a product of that experience.

I express great thanks to Professor Henry Boateng and Professor Julieta Yung for pushing me to cultivate the technical skills that were an integral component of this research. I have succeeded at Bates College because of the high academic standard to which you hold all of your students. I write this thesis as a thank you for your investment in all of your students and for your support and encouragement.

This thesis is dedicated to my advisor, Professor Michael Murray, for whom I express the deepest admiration and gratitude for his support, encouragement, and patience. It is because of my experience working with him that I have the opportunity to pursue a career in economic research. I am greatly indebted to him not only for his invaluable advice and guidance, but for his leading by example to demonstrate what it means to be a scholar.

Contents

1. Introduction

The transportation modal choice problem became relevant in economic research due to the launch of Bay Area Rapid Transit (BART), a public transportation system serving the San Francisco Bay area. The initial phase of the system opened for passenger service in September of 1972 and later phases rolled out through 1974. One of the main concerns for urban planners of the time was the effect that BART would have on existing transportation modes. This problem was first studied by the economist Daniel McFadden (1974), and his findings became known as some of the first major work on discrete consumer choice, a centeral topic in applied microeconomics. His work on the BART, in particular, became known as the first research on the transportation modal choice problem.

The motivation for this thesis is the same as McFadden's motivation for studying the BART. McFadden sought to create a method for planners to perform cost-benefit analysis on the implementation of new transporation systems in existing transportation networks. To perform a cost-benefit analysis in this situation, it is imperative to have accurate ridership forecasts for existing modes following the implementation of the new mode. For McFadden, this meant surveying San Francisco residents to construct a stated-preferences model to forecast changes in ridership following the implementation of the BART.

Since the time of McFadden's research on the BART, the transportation modal choice problem has seen a resurgence in the academic and public interest due to urban expansion. As cities grow large, so does the need for efficient commuting methods as commutes take up a more substantial part of the day. For example, in a recent Boston Globe article, the vehicle insurance company AAA finds that in the city of Boston, the average one-way commute increased from 27 minutes in 2010 to 29 minutes in 2018. There is a clear need for efficient commuting modes to mitigate these observed increases in commute times.

Urban expansion has created a situation in which low-income workers are unable to participate in local urban employment markets due to geographic boundaries preventing easy access to urban centers. In the urban economics literature, this phenomenon is known as the spatial mismatch hypothesis, which is due to Kain (1968). The hypothesis suggests that for urban areas with insufficient fixed-capital public transportation methods, unfacilitated suburban expansion, which he calls "urban sprawl", limits employment opportunities for the poor with limited access to public transportation. Nechyba and Walsh (2004) further suggest that suburban expansion

can create an inequitable distribution of public goods and services that can lead to housing segregation and limit social mobility in the long run. Clearly, employment accessibility is a crucial factor driving social equity.

A closely-related problem to urban sprawl is the first mile, last mile problem, which is best explained through an example. Consider a commuter who wants to travel from point A to point B. For many commuters, the distance between A and B will be too far or inconvenient to be traversed directly. Most commuters in urban areas are forced to employ some form of capital-intensive public transportation mode to cover most of the distance. Many commuters will have trouble closing the distance from point A to the public transportation origin (the first mile), and from the public transportation destination to the point B (the last mile). For many low-income commuters, these distances may be too large or expensive for the potential commute to be feasible.

Bike-sharing is one of the candidate transportation modes being considered by planners to provide commuters with a cost-efficient method to overcome the geographic barriers to employment and make the urban economy more equitable. Bike-sharing programs will only provide a solution to the first mile, last mile problem if commuters actually employ bike-sharing services to complement other forms of public transportation. Understanding the forces that drive the relationship between bike-sharing and public transportation is the goal of this thesis.

With the recent launch of Uber, Lyft, and other ride-sharing services, the transportation modal choice problem has made a resurgence in the public and academic interest. Out of the resurgence, a small body of literature focused on the transportation modal choice effects of bike-sharing emerged.[1] In this thesis, I contribute to this body of academic work by studying the interplay between the MTA subway system and the Citi Bike program in New York City. To do this, I perform an econometric estimation of the effect of the Citi Bike program on subway ridership in a natural experiment environment that exploits the time variation in the subway ridership data. I then perform a similar estimation of subway infrastructure on Citi Bike ridership that exploits the cross-sectional variation across the Citi Bike stations.

I find that subway stations within proximity to bike-sharing infrastructure see more ridership than bike-sharing stations out of proximity to subway stations. This positive effect suggests that some commuters are drawn to subway stations that afford them the opportunity to close the first and last miles of ther commute with a bike.

[1] See Fishman (2016) for a coprehensive overview of the bike-sharing literature.

I perform the same analysis as above with a sub-sample of subway stations located in the urban core of New York City, and find that the Citi Bike program reduces subway ridership. This finding suggests that the Citi Bike program is a substitute for commuters who live in the urban core and for tourists. It also suggests that the overall increase in subway ridership is driven by commuters on the urban periphery, some of whom complement their subway journeys with bike trips. Overall, these results provide evidence that commuters employ bike-sharing to reduce inefficiencies created by the first mile, last mile problem.

To observe the flow of Citi Bike ridership and form a more nuanced understanding of the relationship between bike-sharing programs and subways, I create an econometric model that exploits the cross-sectional variation across Citi Bike stations. To do this, I separate the Citi Bike ridership data into morning and evening times to capture each leg of the daily commute. In the morning, I find that riders depart from residential areas and park near subway stations and in the evening, they get on bikes near subway stations and ride to residential areas. This observed commuting pattern suggests that some riders complement their subway journeys with Citi Bike trips for the first and last miles of their commutes.

The thesis proceeds as follows. Section 2 provides background on bike-sharing programs as well as their benefits and limitations. Section 3 is a review of the relevant literature on bike-sharing programs. Section 4 presents an overview of the methodology I employ to understand the relationship between the two modes. Section 5 discusses the data and provides their sources. Section 6 analyzes the results of the methods outlined in the methodology. I conclude in Section 7 with a discussion of policy implications of the results and areas for further research. Appendex A details the specific processes required to coerce all of the required data into a format conducive to econometrics.

2. Background

What is a Bike-Sharing Program?

Bike-sharing programs are comprised of a fixed network of stations placed throughout an urban area. Bikes freely move about stations as they are used by riders throughout the day. Each station has a fixed number of docks at which the bikes are stored while not in use. Potential riders engage with the system by removing a bike from one station and riding it to another station with an available dock and dropping it off for the next rider to use.

Typical bike-sharing programs can be paid for on a per-ride basis such as a subway or bus, but most riders pay for a yearly subscription out of convenience. The current price for a Citi Bike membership is $149 per year.[2] For the data used in this thesis, 85% of rides are taken by yearly subscribers.[3] A yearly subscription to the Citi Bike program provides customers with unlimited rides under thirty minutes. For every additional fifteen minutes, the customer is charged at a rate of three dollars. Since most customers use this service as a commuting tool, very few rides go over thirty minutes and most trips are covered by the subscription cost.

The role of the company sponsoring the bike-sharing program is to ensure that the stations and bikes are in working order and to redistribute bikes across the system throughout the day to allow the supply of bikes at each station meet the demand. One advantage of bike-sharing programs is their low maintenance cost. The Citi Bike program employs technicians and mechanics to perform routine repairs on bikes and the stations themselves. They also employ truck drivers to move about the city manually redistributing bikes throughout the network overnight when ridership is minimal. Thus, the role of the company is to mitigate the chaos created by routine customer usage through maintenance and redistribution.

The Citi Bike Program

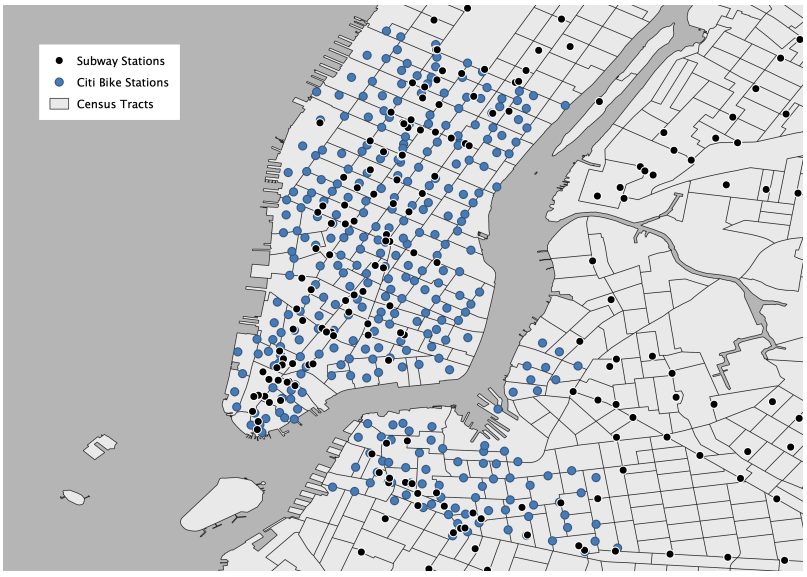

Figure 1: Distribution of Citi Bike Stations at Launch

[2]The yearly subscription price at launch was $95, but was later raised to $149 following the announcement of a system expansion in 2014.

[3]The remaining rides are taken by single riders, one-day pass, three-day pass, and weekly-pass riders.

The initial plan to improve bicycle infrastructure throughout New York City was introduced by the New York Department of Transportation in their Sustainable Streets publication.[4] The Citi Bike program was launched on May 27, 2013 with a network of 332 stations and 6,000 bikes across lower Manhattan and Brooklyn. The program was introduced as a part of a city-wide effort to improve transportation conditions and improve public health. Due to Citi Bike's rapid popularity growth, the program has expanded multiple times.[5] Since launch, the Citi Bike program has expanded into upper Manhattan, Queens, and Jersey City. The program has not grown since 2016, but there are plans in place to expand further into Harlem and Astoria.

Citi Bike's rapid popularity growth allowed the system to quickly integrate into the city's extensive bicycling infrastructure. Since 2013, the program's yearly ridership has grown by 11.8 million trips, increasing from 5.8 million trips in 2013 to 17.6 million trips in 2018. Since the program's launch in 2013, there has only been one death involving a Citi Bike rider which suggests that drivers have made the necessary adjustments to accomodate this new mode of transportation.

The Benefits of Bike-Sharing Programs

They payment structure of the Citi Bike program was a main driver for the system's rapid popularity growth. Unlike the subway system, potential riders can make a single down-payment once a year to cover all of their Citi Bike trips instead of paying per-ride. From the work of Richard Thaler (1999) on mental accounting, we know that consumers prefer making a single lump-sum payment over smaller sequential payments, even when the total costs are equal.[6] In other words, the commuter enjoys the ability to quickly open and close the mental account for their Citi Bike rides once a year rather than re-open it for every bike trip.

Apart from the convenience a subscription-based payment system, bike-sharing programs provide riders with all of the same benefits of owning a bike without the additional payment considerations. Citi Bike docks have an electronic locking system so riders do not have to invest in a personal lock. In addition, routine maintenance is covered by the subscription cost, and so riders do not have to pay for repairs.

[4] Their sustainable streets publication can be found here: http://www.nyc.gov/html/dot/downloads/pdf/stratplan_compplan.pdf.

[5] The sample period for this paper ranges from January 2011 to December 2014, prior to the first program expansion in 2015.

[6] I experiment with distances of 100m and 400m. I do not find the results modestly different at other distances.

Bike-sharing programs have a proven history of improving public health. This is certainly true in New York City, where Babagoli and Kaufman, (2019) perform a cost-benefit study on the expansion of the Citi Bike program in 2015. They find that the expansion resulted in a net economic benefit increase from $18,800,000 to $28,300,000. They also conclude that the health benefits of the Citi Bike program increase with usage and the benefits of the program are not equitably distributed due to inequitable station location. They believe that an equity focus can improve overall health benefits when implementing bike-sharing programs.

The Limitations of Bike-Sharing Programs

Though the yearly subscription cost for a Citi Bike membership is lower than the equivalent number of subway trips for most commuters, the price tag is still substantial and poses a barrier to entry for some low-income workers. To foster equity, the Citi Bike program has a reduced membership rate for New York City residents in public housing. Similar pricing strategies in other cities with bike-sharing programs would make the benefits of the program more evenly distributed across people of different incomes and improve the general welfare since lower-income commuters would get the most use out of the program.

Like any public policy, local governments face the equity-efficiency tradeoff when finding the socially optimal positioning of bike-sharing stations. For bike-sharing programs, the trade-off is between maximizing usage and maximizing coverage. Maximizing usage is when bike-sharing stations are installed in densely populated areas so potential ridership and profit are maximized. This allocation of stations results in the inequitable distribution of benefits analyzed by Babagoli and Kaufman (2019). Maximizing coverage is when bike-sharing stations are installed in locations that cover the largest geographic area. This positioning allows all people across the urban area to have access, though it may cause a shortage of bikes in densely populated areas when demand is high.

Researchers in the urban planning literature have performed computational estimations of the maximal usage and maximal coverage approaches to bike-sharing station allocation as approximations of their efficiency and equity-maximizing allocations. The effiency estimation requires finding the optimal station positioning that maximizes profit subject to a cost constraint on program implementation and maintenance. The equity estimation requires finding the optimal station position that maximizes geographic coverage subject to a profit constraint. It is important to note that these optimizations are approximations of the socially efficient and

equitable allocations. For example, it may not be equitable or efficient to leave some commuter in the urban core without the ability to use a bike due to an insufficient supply of bikes in locations of high demand. In the context of bike-sharing, the equity-efficiency tradeoff is a very complex question that requires an understanding of the program beyond its impact on modal choice.

Another limitation of bike-sharing programs is bike redistribution. As riders use the system throughout the day, the distribution of bikes across the system becomes imbalanced. In the literature, finding the optimal method to redistribute the bikes is called the inventory rebalancing problem. To maintain a proper distribution of bikes throughout the system, bike-sharing program sponsors typically dispatch a number of trucks to move about the network manually redistributing bikes at night to meet the demand for the following day. This process is expensive as it requires substantial capital and labor. To grapple with the cost of redistribution, many programs throughout the world have created a system by which riders can earn free trips by redistributing bikes to stations with insufficient bikes to meet the demand. While this does lower the severity of the problem, many programs like the Citi Bike program still require mass redistribution to ensure bikes are optimally allocated at the start of each day.

3. Literature Review

Here I introduce briefly the literatures on (i) bike-sharing location optimization, (ii) the inventory rebalancing problem, and (iii) bike-sharing and modal choice.

Bike-Sharing Infrastructure Location Optimization

Much of the literature on bike-sharing programs is focused on providing optimization techniques for finding the optimal allocation of bike-sharing stations within an urban area. For cities with a bike-sharing program, these techniques are designed to spot inefficiencies in the system to understand what changes, if any, can be made to improve it. For cities without a bike-sharing program, these same techniques can be applied to estimate the optimal allocation of stations prior to implementation. The techniques employed throughout this area of the literature are GIS intensive and require computational approximations to solve the necessary optimization problems.

For example, Garcia-Palomares et al. (2012) provide a GIS location-allocation method to find the optimal allocation of bike-sharing stations based on social, demographic, geographic,

and built-environmental factors. They begin their study by forecasting system usage across the city based on demographic information such as population density and employment. These forecasts allow their GIS model to place stations in locations where potential usage is maximized. They also use the forecasts to set the number of docks per station. They employ commercial and residential land-use information to classify each station as an "attractor" or "generator" based on whether more trips begin or end at each station at different times of day. The setup of their optimization problem allows them to approximate the equity-efficiency trade-off discussed in the previous section by changing the objective function to maximize potential usage or maximize potential access. Their results confirm the intuition that stations should be placed within proximity to metro stations.

The Inventory Rebalancing Problem

In the previous section, I discussed how inventory rebalancing is a substantial cost facing bike-sharing programs. The literature on this problem is split into two areas. The first area deals with optimizing manual rebalancing strategies. These studies investigate static rebalancing strategies in which bicycles are manually reallocated overnight when demand is low. Other studies investigate dynamic rebalancing in which bicycles are strategically redistributed throughout the day when demand is high at some stations and low at others. The second area of research on the inventory rebalancing problem deals with self-rebalancing through price incentives. These authors study how companies that sponsor bike-sharing programs can manipulate prices for bike-sharing services, or even specific routes, to incentivize riders to change their commuting habits and redistribute the bikes on their own to lower the cost of redistribution.

Dell'Amico et al. (2013) provide an analysis of the physical rebalancing problem in the static context for a case study in Reggio Emilia, Italy. They model the truck routes as a multiple traveling salesmen problem in which the routes of some fixed number of trucks are optimized to minimize travel costs. They draw extensively on graph theory to formulate the problem. In particular, they represent the bike-sharing network as a complete digraph for which each station is represented by a vertex and each path between stations as a vertex. Each path between is assigned a weight proportional to the cost of traveling along the edge. Thus, the problem seeks to minimize the total cost subject to a demand constraint. Due to the number of parameters in the problem, they approximate the solution computationally. After constructing their computational method for Reggio Emilia, they expand the results to other urban areas.

Haider et al. (2014) minimize manual rebalancing costs subject to a system profit constraint by manipulating the prices of bike routes to facilitate dynamic rebalancing. By manipulating prices to purposefully make the system imbalanced, they manual rebalancing effort is less costly since the number of manual rebalancing trips is decreased. They model the scenario as a two-agent optimization problem. The first agent is the bike-sharing system operator who sets the prices of each bike route to minimize the number of slack stations (those with too few bikes) and surplus stations (those with too many bikes) at the end of the day. The secod agent represents the riders who minimize their commuting costs subject to time constraints. With this formulation, computationally approximate the solution to this optimization problem and show that if performed well, a dynamic pricing strategy could be employed to lower manual redistribution costs.

Bike-Sharing Programs and Modal Choice

The literature on bike-sharing programs and modal choice is a combination of stated preference studies and revealed preference studies. The revealed preference studies use a similar method to that employed by Daniel McFadden during his study of the BART which consists of retroactively surveying riders to understand how their commuting habits have changed following the implementation of the bike-sharing program. The revealed preference models use econometric methods on ridership data of different modes to understand how ridership of existing modes has changed following the implementation of the bike-sharing program. This thesis is a member of the literature on revealed preferences.

In a survey-based study focused on how commuters have changed their commuting habits following bike-sharing implementation, Martin and Shaheen (2014) study the effect in Washington DC and Minneapols. They find that a higher proportion of commuters in Washington DC decrease rail ridership than those who don't and a higher proportion of commuters in Minneapolis increase ridership than those who don't. In both metro areas, they find that commuters who live in the urban periphery are more likely to complement bike-sharing and metro ridership. They find a substitution effect between modes for commuters living in the urban core.

In a revealed-preferences study, Ma et al. (2019) investigate the metro ridership effect of the Capital Bikeshare program in Washington, DC. They perform an origin-destination analysis with ride-level bike-sharing data and find that over 80% of stations with more than 500 trips per week are located within proximity to subway metro stations. They also perform a panel

regression analysis in which they control for built-environmental and demographic factors and find that a 10% increase in Capital Bikeshare ridership increases metro ridership by 2.8%.

4. Methodology

Subway Ridership Effect of Bike-Sharing Infrastructure

In this section, I analyze the effect of the Citi Bike program on subway ridership in a natural experiment environment. I assume the Citi Bike program is a shock to the New York City transportation system since it is unlikely that people changed their transportation habits prior to its implementation. To measure the program's effect, I estimate a difference-in-differences regression framework at the subway station level that measures the difference in ridership of stations before and after Citi Bike's implementation. In each of the regression analyses, I only consider subway stations within Manhattan and Brooklyn. The reason for this is that many of the stations far from the center of the city, (those in Long Island, for example), have fundamentally different ridership patterns that cannot be captured without a robust measure of subway ridership.

To capture the causal effect of interest, I adapt the work of Campbell and Brakewood (2017) to the study of subway systems. As in their study, I only consider subway stations located in Manhattan and Brooklyn. I let the subway stations out of proximity to a bike-sharing station be the control group and those witihin proximity to a bike-sharing station be the treatment group. I let the parameter α denote the treatment effect and define $a = 1$ if subway station j is within proximity to a bike-sharing station and $a = 0$ otherwise. I define $t = 1$ if the date is on or after May 27, 2013, the first day of operation for the Citi Bike program and $t = 0$ otherwise. With this assignment, the average treatment effect at the station level becomes

$$(\mathbb{E}[ridership_{jat} \mid a = 1, t = 1] - \mathbb{E}[ridership_{jat} \mid a = 1, t = 0]) -$$
$$(\mathbb{E}[ridership_{jat} \mid a = 0, t = 1] - \mathbb{E}[ridership_{jat} \mid a = 0, t = 0]) = \beta$$

Here, β represents the average treatment effect across all subway stations and will be estimated with an OLS panel regression. In this regression framework, the causal effect of interest is the within-station ridership effect of the Citi Bike program. In other words, β corresponds to the expected deviation in subway ridership of a station from its own mean. To find this causal effect, I estimate the following panel regressions:

$$ln(Entries_{jt}) = \alpha + \beta BikeOpen_t \times BikeStations_{jt} + \gamma StationFE_j + \delta DowFE_t + \phi Month_t + \lambda Rain_t + \varepsilon_{jt}$$

$$ln(Exits_{jt}) = \alpha + \beta BikeOpen_t \times ln(BikeDocks_{jt}) + \gamma StationFE_j + \delta DowFE_t + \phi Month_t + \lambda Rain_t + \varepsilon_{jt}$$

Here, β is the difference-in-differences estimator for the within subway ridership effect of the Citi Bike program. The variables, $ln(Entries_{jt})$ and $ln(Exits_{jt})$ are the natural log of subway station entries and exits for subway station j at time t. As a robustness check, I construct two measures of bike-sharing infrastructure. I define $BikeStations_{jt}$ equal to the number of bike-sharing stations within proximity to subway station j. The variable $Docks_{jt}$ represents the total number of bike-sharing docks across all subway stations within proximity to subway station j. Some of the entry, exit, and docks observations are equal to zero, so I add one to every observation to avoid creating undefined observations (since $ln(0) = \emptyset$). Since I am without a robust measure of subway ridership across stations, I proxy for ridership using $ln(Entries_{jt})$ as above and similarly for $ln(Exits_{jt})$. I include the dummy variable $BikeOpen_t = 1$ if the date is on or after May 27, 2013, the first day of operation for the Citi Bike program and $t = 0$ otherwise. All of the regressions use 200m as the proximity for which bike-sharing infrastructure measures are calculated. The results are not changed using modestly different proximity measures.

To isolate the causal effect of the Citi Bike program β, I include station-level fixed effects to control for cross-sectional, station-level variation. The subway data range from January 2011 to 2014 so I include a $Month_t$ fixed effect to control for system-level time variation. Since many riders use the bike-sharing program as a commuting tool, there is substantial variation in bike-sharing ridership throughout the week which I control for with a day-of-week dummy variable. There is additional variation at the line level that cannot be captured with a subway station-level fixed effect which I control for with a subway line fixed effect. Since weather plays an important role in determining bike-sharing ridership, I control for rain with $Rain_t$ in each regression model.

To consider the possibility that subway ridership is fundamentally different for stations within the urban core, I estimate the above regressions for all subway stations across Manhattan and Brooklyn and again for those within proximity to a Citi Bike station. I use Citi Bike station proximity to proxy for urban core and interpret the results.

To address the first mile, last mile problem, I transition to a between estimation of the effects of subway placement on the flow of Citi Bike ridership ridership. I create a regression framework that exploits the cross-sectional variation in the Citi Bike ridership data. I base my regression framework on the work of Noland et al. (2016). Their regression methodology exploits the cross-sectional variation across a number of important factors including general bike ridership infrastructure, proximity to subway stations, land use mix, demographics, and other built-environmental factors. They estimate their regression framework on individual months of data to understand how these cross-sectional factors influence ridership throughout the year.

The groundwork for this regression methodology is similar to that of the authors, and requires extensive GIS work to coerce the data into a format conducive to econometrics. Since I require information on the cross-sectional factors within proximity to each Citi Bike station, I introduce boundaries around each Citi Bike station within which these parameters are calculated. In GIS, this is accomplished with a Voronoi tessalation in which each node is a different Citi Bike station. Performing this GIS calculation gives the following plot:

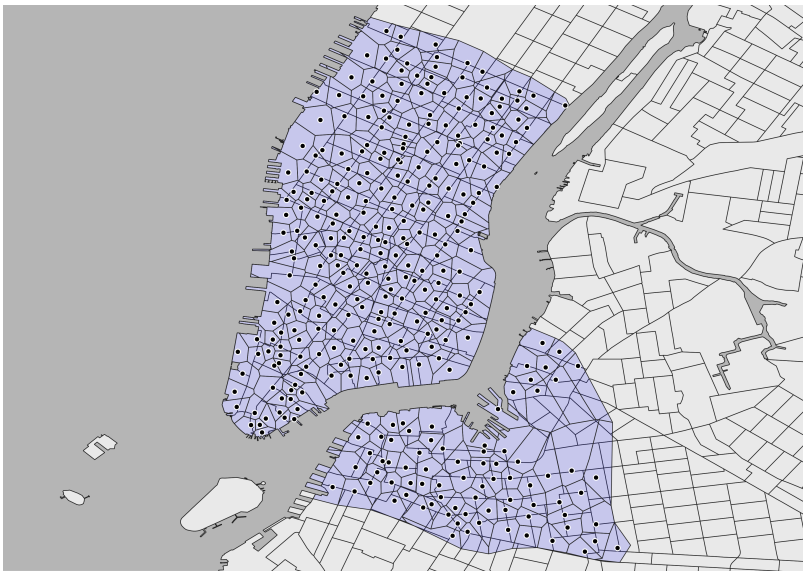

Figure 2: Voronoi Diagram centered on Citi Bike stations

The Voronoi plot creates Citi Bike station regions based on their relative proximities to other stations. Formally, a Voronoi plot is defined on a metric space X with a distance function d. Let K be a set of indices and let $(P_k)_{k\in K}$ be an ordered collection of nonempty subsets (the sites) in the space X.The Voronoi region, R_k, associated with the site P_k is the set of all points

in X whose distance to P_k is not greater than their distances to all other sites P_j where $j \neq k$. In other words, if $d(x, a) = inf\{d(x, a) \mid a \in A\}$ denotes the distance between the point x and the subset A, then

$$R_k = \{x \in X \mid d(x, P_k) \leq d(x, P_j) \ \forall \ j \neq k\}$$

The Voronoi diagram is the collection of Voronoi regions $(R_k)_{k \in K}$. By construction, the regions are disjoint and so they form a partition over the region of interest. In this case, the region of interest is lower Manhattan and Brooklyn. For $x, a \in X$ the distance funtion is the Euclidean distance:

$$l = d[x, a] = d[(x_1, x_2), (a_1, a_2)] = \sqrt{(x_1 - a_1)^2 + (x_2 - a_2)^2}$$

I use the above Voronoi diagram to construct measures of land use mix which I integrate into the following regression methodology.

Since people employ bike-sharing as a commuting tool, there is likely a measurable flow of ridership across the Citi Bike network in the morning that is mirrored in the evening. If we can isolate the commuting variation in ridership, we will be able to determine if subway station placement play a role in determining the flow of ridership for Citi Bike commuters. If commuters complement their subway journeys with Citi Bike trips, we would expect a flow of bikes moving from residential areas to subways in the morning and from subways into residential areas in the evening. To test this claim, I require the following regression.

$$\begin{aligned} OriginsAM_j = \alpha + \psi DestinationsAM_j + \beta Stations_j + \gamma Racks_j \\ + \delta Lanes_j + \phi log(Population_j) + \lambda CommercialShare_j + \\ + \sigma ManufacturingShare_j + \omega ResidentialShare_j + \varepsilon_j \end{aligned}$$

Here, $OriginsAM_j$ denotes the average number of Citi Bike trips starting from station j in the morning. Since we want to understand the effect of subway infrastructure on bike-sharing ridership for the morning and evening commute, it is important to control for Citi Bike round trips in which both legs of the trip occur during the same part of the day. I control for this variation as a proxy for non-commute Citi Bike trips. Thus, the only remaining variation in the data is due to commuting. I believe the factors driving commuting and non-commuting ridership to be fundamentally different, and therefore, controlling for non-commuting round-trips does not

introduce endogeneity into the regression. Therefore, I believe that the remaining coefficients in the regression model provide unbiased estimates of the Citi Bike commuting variation effects of subway infrastructure.

The parameters $Stations_j$ refers to the number of subway stations within 200m of Citi Bike station j. The parameters $Racks_j$ and $Lanes_j$ denote the number of bike racks and bike lanes within the Voronoi polygon containing Citi Bike station j. I include these as regressors to proxy for general bike ridership activity as controls for endogeneity arising from the location of the stations. This controls for the case of supply-driven-demand, which would bias the coefficients upward in areas in which general bike ridership is prevalent. As another proxy for general bike ridership activity, I include $log(Population_j)$ which measures the population of the census tract containing Citi Bike station j. The parameters $CommercialShare_j$, $ManufacturingShare_j$, and $ResidentialShare_j$ represent the proportions of commercial, manufacturing, and residential land use occupying the Voronoi polygon containing Citi Bike station j. I omit park lans use data to avoid perfect multicollinearity. Below is a graphical representation of land use data:

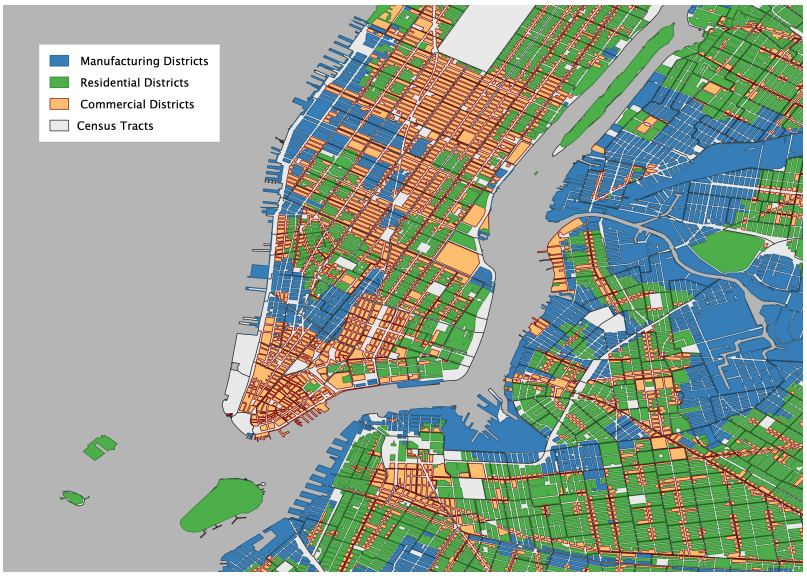

Figure 3: Districts by Land Use Type

In this regression, I use the land use share variables to proxy for commercial and residential areas of the city. The paths of Citi Bike ridership can be inferred based on the directions of the land use variables relative to each other. Since the paths of travel for riders are the main motivator for this analysis, the magnitude of the coefficients on each land use variable is of secondary importance. Thus, we can glean information on the paths of travel for Citi Bike

commuters by comparing the directions of the coefficients on the land use variables to each other.

Since there are two ways in which a commuter can complement a subway journey with a bike trip and each bike-sharing trip has an origin and destination, there are four different regression models that I use to capture each aspect of the first mile, last mile problem. I repeat each regression with $Origins_j$ and $Destinations_j$ switched and with data from the morning and the evening. Thus, all of the regressions are identical and differ only by the dependent variable and the round-trip control variable. The expected direction of each effect of interest for each regression is summarized in the following table.

Dependent Variable	$Stations_j$	$CommercialShare_j$	$ResidentialShare_j$
$OriginsAM_j$			> 0
$OriginsPM_j$	> 0	> 0	
$DestinationsAM_j$	> 0	> 0	
$DestinationsPM_j$			> 0

I expect that since people use the Citi Bike program as a commuting tool, there will be an flow of ridership from residential areas into commercial areas and subway regions in the morning. Therefore, I expect a positive coefficient on $ResidentialShare_j$ and negative coefficients on $CommercialShare_j$ and $Stations_j$ for the first and third regressions, respectively. In the afternoon, I expect that there will be a flow of ridership from commercial areas and subway regions into residential areas. Therefore, I expect positive coefficients on $CommercialShare_j$ and $Stations_j$ and a negative coefficient on $ResidentialShare_j$ for the second and fourth regressions, respectively.

Though it is not necessary, negative signs on the remaining coefficients would provide additional support for the idea that riders employ bike-sharing services as a commuting tool to complement their subway journeys. In the results, I report the regression outputs for each of these regressions and discuss the implications in terms of the first mile, last mile problem.

5. Data

In this thesis, I draw on five sources of data which I detail below. All of the data are publicly available and do not require special permissions, memberships, or subscriptions.

Metropolitan Transportation Authority

Though the MTA does not publish a robust measure of ridership across subway routes, they do publish the number of subway entries and exists across all MTA subway stations which I substitute as a proxy for ridership. In this study, I employ their data between January 1, 2011 and December 31, 2014 to capture the entire time period of interest. The data are disaggregated across turnstiles into four-hour increments and are stored as cumulative entries and exits. These data require extensive pre-processing before they are conducive to panel econometrics. See the data appendix for more information on the data pre-processing steps.

Citi Bike

Since their launch in May 2013, Citi Bike have released disaggregated, open-source data on all Citi Bike trips. These data include the start station coordinates, end station coordinates, start time, end time, ride duration, bike ID, rider type, rider age, and rider gender at the trip level. I download these data from launch on May 27, 2013 to December 31, 2014 to capture the entire treatment period of interest. Citi Bike also maintains a live station feed containing the number of bikes and available docks at every Citi Bike station to provide users with real-time information to guide their commutes. I use their live feed to gather data on the number of bike docks across all Citi Bike stations and construct measures of bike-sharing infrastructure. The abundance of information available for each Citi Bike trips allows for slicing along different parameters to gain a nuanced understanding of the forces driving the system.

New York City Government

The NYC government maintains a number of geographic datasets that I integrate into the analysis. The first of these datasets is a shapefile containing the locations of all subway stations across NYC. In this shapefile, they include the coordinates of each station, the lines for which the stations provide service, and any conditions regarding service modifications throughout the week. They also publish an MTA subway line shapefile which allows me to control for line-level characteristics that are not captured at the station level. In addition, the NYC government maintains a shapefile containing the locations of all bike lanes throughout the city. They include the type of bike lane, the direction it faces, and the date of implementation. Lastly, they publish

a shapefile containing the coordinates of all bike rack locations throughout NYC. Bike rack and bike lane locations are important proxies for cross-sectional bike activity.

New York City Department of City Planning

The NYC Department of City Planning maintains a map of the city broken into land use districts. There are many divisions in the data, though I only incorporate three into the analysis. I employ their shapefiles containing residential districts, commercial districts, and manufacturing districts. Since subways and bike-sharing programs are commuting tools, there is substantial intra-day ridership variation for both transportation modes. Understanding the land use around each subway or bike-sharing station is important to understanding the dynamics governing the complementary and substitution effects between the two modes. In addition, the NYC Department of City planning maintains demographic shapefiles broken down by census tract. I incorporate population and employment shapefiles created with 2010 census data as cross-sectional demographic controls.

National Oceanic and Atmospheric Administration

The NOAA takes online custom historical data requests across all weather observatories throughout the United States. For this study, I performed a custom data request for NYC from January 1, 2011 to December 31, 2014. They provide daily measures of high and low temperatures, wind, snow, rain, thunder, and other weather-related indicators that I integrate as controls.

6. Results

Subway Ridership Effect of Bike-Sharing Infrastructure

I estimate the first regression framework to isolate the MTA subway ridership effect of the Citi Bike program as a natural experiment. Each regression table presents the results for entry and exit data for a geographic proximity measurement of 200m.[7] Table 1 contains the regression output for which the variable $BikeStations_{jt}$ is the measure of bike-sharing infrastructure. Table 2 contains the regression outputs for which the variable $BikeDocks_{jt}$ is the measures of

[7] I experiment with distances of 100m and 400m. I do not find the results modestly different at other distances.

infrastructure. Table 3 contains the regression outputs for a robustness check in which the only subway stations considered are those within a 200m proximity to a bike-sharing station. All regressions include day-of-week, station, and line fixed effects and a system-level time trend. All regressions are performed in R with the plm package. The regression tables are generated in R with the stargazer package.

First, I estimate the regression with $BikeStations_{jt}$ as the measure of bike-sharing infrastructure.

Table 2: Within Estimation with Bike Stations as Infrastructure Measure

	Dependent variable:			
	log(Entries)	log(Exits)	log(Entries)	log(Exits)
	Manhattan Only		Manhattan and Brooklyn	
	(1)	(2)	(3)	(4)
Rain	-0.352^{***}	-0.311^{***}	-0.346^{***}	-0.309^{***}
	(0.017)	(0.016)	(0.014)	(0.013)
Bike Stations within 200m	-0.006	0.012	-0.005	0.017^{**}
	(0.008)	(0.008)	(0.008)	(0.007)
DOW FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Line FE	Yes	Yes	Yes	Yes
Observations	85,162	85,162	104,835	104,835
R^2	0.477	0.512	0.525	0.551
Adjusted R^2	0.476	0.511	0.524	0.550
Note:				*p<0.1; **p<0.05; ***p<0.01

The results in Table 1 suggest that the presense of bike-sharing infrastructure has a significant effect of subway ridership. In particular, the coefficient on $BikeStations_j$ is positive and significant for $ln(Exits_{jt})$, but not for $ln(Entries_{jt}$. This result is robust to the exclusion of Brooklyn-based subway stations. By the subway stations measure of infrastructure, the results suggest a complementary effect between the modes. Since none of the coefficients of interest are negative, the results do not suggest a substitution effect between the modes.

The positive coefficient on $BikeStations_{jt}$ for the $ln(Exits_{jt})$ regressions suggests that commuters choose to exit the subway system at stations where there are bike-sharing opportunities in proximity to the subway station. The magnitude of the coefficients suggests that each additional

bike-sharing station within 200m to a subway station increases the number of exits from that station by about 2%. This result is robust to the exclusion of Brooklyn-based subway stations.

As a robustness check, I estimate the same regressions in Table 2 but with $log(BikeDocks_{jt})$ as the measure of bike-sharing infrastructure.

Table 3: Within Estimation with log(Bike Docks) as Infrastructure Measure

	Dependent variable:			
	log(Entries)	log(Exits)	log(Entries)	log(Exits)
	Manhattan Only		Manhattan and Brooklyn	
	(1)	(2)	(3)	(4)
Rain	-0.352^{***}	-0.311^{***}	-0.347^{***}	-0.309^{***}
	(0.017)	(0.016)	(0.014)	(0.013)
log(Bike Docks) within 200m	0.015^{***}	0.024^{***}	0.007	0.019^{***}
	(0.005)	(0.004)	(0.004)	(0.004)
DOW FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Line FE	Yes	Yes	Yes	Yes
Observations	85,162	85,162	104,835	104,835
R^2	0.477	0.512	0.525	0.551
Adjusted R^2	0.476	0.511	0.524	0.550
Note:				*p<0.1; **p<0.05; ***p<0.01

The positive coefficients on $ln(BikeDocks_{jt})$ for the $ln(Exits_{jt})$ regressions corroborate the results of Table 2 and suggest a complementary effect between the modes for the last mile of the commute. In addition, the coeffients on $ln(BikeDocks_{jt})$ for the $ln(Entries_{jt})$ regressions have become positive and significant. Unlike in Table 2, these results suggest a complementary effect between modes for the first mile of the commute as well. As in Table 2, the results are robust to exclusion of Brooklyn-based subway stations and none of the coefficients are negative which suggests that there is no substitution effect between the modes.

The results suggest that each time the number of bike docks within 200m of a subway station doubles (increases by 100%), the number of entries at that station increase by between 1% and 2%. This range of values is consistent with the coefficients on $BikeStations_{jt}$ form Table 2. To illustrate this, consider a subway station within a 200m proximity to 30 docks. If another bike-sharing station is added with 30 new docks, the coefficients suggest that these new docks

will increase the number of entries at that subway station by between 1% and 2%. Unlike the coefficients in Table 1, the coeffients on $ln(BikeDocks_{jt})$ demonstrate the decreasing marginal change in the effect on subway entries and exits as more subway stations and docks are added.

Next, I estimate the same regression models as in Tables 1 and 2 but only considering the subway stations within a 200m proximity to a bike-sharing station.

Table 4: Within Estimation with Stations in Proximity to Bike-Sharing Infrastructure

	Dependent variable:			
	log(Entries)	log(Exits)	log(Entries)	log(Exits)
	(1)	(2)	(3)	(4)
Rain	-0.359^{***}	-0.314^{***}	-0.359^{***}	-0.314^{***}
	(0.022)	(0.020)	(0.022)	(0.020)
Bike-Sharing Stations within 200m	-0.122^{***}	-0.112^{***}		
	(0.015)	(0.014)		
log(Bike Docks) within 200m			-0.084^{***}	-0.083^{***}
			(0.018)	(0.016)
DOW FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Line FE	Yes	Yes	Yes	Yes
R^2	0.490	0.523	0.490	0.523
Adjusted R^2	0.489	0.522	0.489	0.522
Note:				*p<0.1; **p<0.05; ***p<0.01

The results of Table 4 appear to contradict our earlier findings. However, they add more nuance to our understanding of the relationship between bike-sharing and subway programs. All of the coeffients on each measure of bike-sharing infrastructure are negative which suggests a substitution effect between the modes. Citi Bike stations are located in the urban core, which means the only subway stations being considered in the above regressions are also in the urban core. These subway stations have fundamentally different ridership patterns than those on the urban periphery, which suggests the above regressions are picking up a completely different signal than in the previous regressions.

By dropping subway stations on the urban periphery, we lose the effect of people riding bikes to the subway to commute into and out of the urban core. In effect, we are left observing the changes in subway ridership for people who live in the urban core or for tourists. This suggests

that the effect of bike-sharing on subway ridership depends on where the commuters reside relative to the urban core. Commuters in the urban core appear to employ bike-sharing as a substitute while those on the urban periphery employ bike-sharing as a complement.

The coefficients on $BikeStations_{jt}$ suggest that each additional bike-sharing station within proximity to subway station j decreases subway ridership at that station by between 11% and 12% relative to other stations in the urban core. In addition, the coefficients on $ln(BikeDocks_{jt})$ suggests that if the number of bike docks within proxoimity to subway station j doubles (increases by 100%), subway ridership will decrease by about 8% relative to other stations in the urban core.

The analysis above provides us with a nuanced understand of the interplay between bike-sharing programs and subways. We find that the overall effect of the Citi Bike program was a subway-wide increase in exits, which suggests that commuters employ Citi Bike as a complement to subway ridership in at least one direction. We find this result to be robust across station and dock measures of bike-sharing infrastructure and to the exclusion of Brooklyn-based subway stations.

We also do not observe evidence of a substitution effect between the modes until we exclude subway stations on the urban periphery. After making this exclusion, we find that bike-sharing infrastructure greatly decreases subway ridership relative to other subway stations in the urban core. This suggests that for commuters and tourists living in the urban core, bike-sharing serves as a substitute for their subway journeys. Overall, these results in dialogue with one another imply that the overall increase in subway ridership is driven by commuters on the urban periphery who employ the bike-sharing program as a complement to their subway journeys.

At this point, we have convincing evidence that bike-sharing programs provide a preferred solution to the first mile, last mile problem, since the increase in ridership across the subway system is driven by subway stations on the urban periphery. While the results so far are significant and consistent with our expectations, they only tell half of the story. To form a more complete understanding of the relationship between bike-sharing and subways and the first mile, last mile problem, we need to understand the movement patterns of Citi Bike riders. To do this, I create another regression methodology that exploits the cross-sectional variation across te Citi Bike stations to examine how ridership throughout the system is influenced by subway infrastructure.

Citi Bike Ridership and the First Mile, Last Mile Problem

In this section, I present the results of the first mile, last mile problem analysis with between estimations discussed in the methodology. Table 5 contains the regression results for the morning commute and Table 6 for the evening commute. All regressions include bike rack and bike lane controls, land use controls, and population controls. Since each regression is a between estimation, time variation has been averaged out and so the only variation reflected in the coefficients is cross-sectional at the Citi Bike station level. As in the previous sections, all regressions are performed in R with the plm package and the regression tables are generated in R with the stargazer package.

I begin the cross-sectional study of the first mile, last mile problem with the number of Citi Bike origin and destination trips at each station in the morning as the dependent variables.

Table 5: The First Mile: Cross-Sectional Variation Effects on Citi Bike Ridership

	Dependent variable:	
	AM Origins	AM Destinations
	(1)	(2)
Subway Stations within 200m	-0.675^{**}	0.577^{**}
	(0.341)	(0.291)
Commercial Land Use Share	0.095	-0.113^{**}
	(0.064)	(0.055)
Residential Land Use Share	1.807^{***}	-4.069^{***}
	(0.682)	(0.543)
Observations	332	332
R^2	0.502	0.620
Adjusted R^2	0.489	0.611
Note:	*p<0.1; **p<0.05; ***p<0.01	

For the AM Origins regression, the coefficient on $ResidentialShare_j$ is positive and significant. It implies that a 1% increase in residential land use around the station increases the number of origin Citi Bike Trips from that station by 1.8. This is consistent with our expectations and suggests a flow of ridership out of residential areas in the morning for more residential areas. Furthermore, the coefficient on $Stations_j$ is negative and significant. This provides further evidence of a flow of ridership from residential areas, since fewer riders on average start their

rides near subway stations.

Now that we have an understanding of where riders begin their trips, we must ask where they finish. For the AM Destinations regression, the coefficient on $Stations_j$ is positive and significant. It implies that each additional subway station in a 200m radius of the Citi Bike station increases the number of destination bike trips to that station by 0.58. This result is consistent with our expectations for a flow of Citi Bike ridership into subway regions in the morning, as more subway-dense parts of the city see more ridership in the morning. To further support this claim, the coefficient on $ResidentialShare_j$ is significant and negative which suggests that on average, fewer morning Citi Bike trips end in residential areas.

These regression results imply a clear flow of ridership out of residential areas into subway regions in the morning, which is consistent with our expectations from the previous section. They support the claim that commuters complement subway journeys with bike trips in the morning and therefore, employ the Citi Bike program to make the first mile of the commute more efficient.

To continue the study of the first mile, last mile problem and to provide a robustness check for the methods employed above, I run the same regressions again with Citi Bike trips taken in the evening.

Table 6: The Last Mile: Cross-Sectional Variation Effects on Citi Bike Ridership

	Dependent variable:	
	PM Origins	PM Destinations
	(1)	(2)
Subway Stations within 200m	1.024^{***}	-1.130^{***}
	(0.299)	(0.317)
Commercial Land Use Share	-0.112^{**}	0.062
	(0.056)	(0.060)
Residential Land Use Share	-3.720^{***}	2.509^{***}
	(0.574)	(0.632)
Observations	332	332
R^2	0.900	0.888
Adjusted R^2	0.897	0.885
Note:	*p<0.1; **p<0.05; ***p<0.01	

For the PM Origins regression, the coefficient on $Stations_j$ is positive and significant. It implies, on average, that each additional subway station within 200m of a Citi Bike station increases the number of evening origin bike trips from that station by 1. This is consistent with a flow of Citi Bike ridership out of subway station areas in the evening. Furthermore, the coefficient on $ResidentialShare_j$ is negative and significant, which suggests that more residential neighborhoods see fewer origin trips in the evening.

Once again, we must ask where commuters finish their evening Citi Bike trips with the same method as before. For the PM destinations regression, the coefficient on $ResidentialShare_j$ is positive and significant. It implies, on average, that a 1% increase in residential land use around the a Citi Bike stations increases the number of destination bike trips taken by commuters by 2.5. Furthermore, the coefficient on $Stations_j$ is negative and significant, which suggests that fewer commuter Citi Bike trips begin in residential areas.

Taken together, the regression results from Table 6 imply that commuters complement subway journeys with bike trips in the evening. This finding is consistent with our expectations from the methodology. There is strong evidence to suggest that commuters employ the Citi Bike program to make the last mile of their commute more efficient in addition to the first mile.

As an additional robustness check, I estimate the regressions in Tables 5 and 6 in which I convert the $Stations_j$ is binary indicator that takes the value one if there exists at least one subway station wihin 200m of the Citi Bike station and zero otherwise. The results of this robustness check are not modestly different with this change in the variable definition.

The regression results in Tables 5 and 6 suggest that the commuting variation in Citi Bike ridership can be explained as a flow of ridership from residential areas to subway regions in the morning and from subway regions to residential areas in the evening. This result is robust to the inclusion of bike ridership activity controls to mitigate possible endogeneity from the placement of the Citi Bike stations and across multiple definitions of subway infrastructure. Since the directions of the coefficients on $ResidentialShare_j$ and $Stations_j$ are consistent with our expectations for a flow of commuters into and out of the urban core, the results also suggest that controlling for non-commuter Citi Bike rides as in the methodology is a suitable method to isolate the commuting variation of ridership.

7. Conclusion

I begin by summarizing each of the major findings and relate them to the literature on bike-sharing programs. I conclude the section with a policy discussion of the results.

Summary

I found that the overall effect of the Citi Bike program was an increase in ridership across the subway system and the result was robust across multiple measures of bike-sharing infrastructure and to the exclusion of Brooklyn-based subway stations. I also found that excluding subway stations on the urban periphery, the effect of the Citi Bike program becomes a decrease in subway ridership. This has two important implications for the relationship between subway systems and bike-sharing programs. It suggests that for residents within the urban core, bike-sharing is a substitute for subway journeys. It also suggests that the observed increase in subway ridership due to the Citi Bike program is driven by commuters on the urban periphery who complement the two modes.

There is additional evidence supporting the claim that commuters complement the modes. The commuting variation in Citi Bike ridership can be explained as a flow of ridership from residential areas to subway regions in the morning and from subway regions to residential areas in the evening. This finding is robust to the inclusion of controls for general bike-ridership activity to mitigate possible endogeneity from Citi Bike station placement.

The result that commuters in the urban core substitute the modes is consistent with the findings of Campbell and Brakewood (2017) who employ a similar method and find that the Citi bike program decreased bus ridership. The same result is consistent with Martin and Shaheen (2014) who approach the same question with a stated-preferences approach. The implication that commuters on the urban periphery drive the complementary effects of the modes is also consistent with their findings. Lastly, the conclusion that subway ridership drives bike-sharing ridership is consistent with Noland et al. (2016), who find that bike-sharing stations in proximity to subway stations see more ridership on average than others.

The general finding that commuters complement subway journeys with bike-sharing trips is consistent with many other papers in the literature including Ting et al. (2015), Graehler et al. (2019), and Shaheen et al. (2013). The findings of this thesis corroborte the results of research that employ both stated-preference models and revealed-preference models.

Further Research

I have created an econometric modal choice framework to understand the interplay between bike-sharing programs and subway systems that can be generalized into many areas of urban research. For example, by understanding the substitution effects between the two modes, it would be interesting to examine the effects of bike-sharing programs on subway crowding. It would also be effective to extend the analysis to taxi ridership and personal car ridership to form a nuanced understanding of the bike-sharing effect on carbon emissions. The many different directions in which this research can be taken are the first steps towards a formal valuation study of the Citi Bike program.

Policy Implications

This paper's main contribution the literature is a data-driven method to understanding the relationship between bike-sharing programs and subways on each leg of the daily commute. By controlling for non-commuter variation in Citi Bike ridership, we can observe how ridership is driven by subway station locations.

The implications of this paper are clear from a policy perspective. I have shown on a micro-level that commuters do complement their subway journeys with bike trips. This supports the claim that bike-sharing programs can make commutes more efficient by providing commuters on the urban periphery a solution to the first mile, last mile problem. By lessening the geographic barriers to urban employment opportunities, bike-sharing programs have the potential to make urban labor markets more equitable for those who cannot afford to live close to the urban core.

Appendix A. Data

MTA Subway Data

The data provided on the MTA website are broken into one-week segments disaggregated at the turnstile level into four-hour increments and require aggregation across stations and across days. The datasets are very large (~2GB each) which makes downloading each of them as a .csv file infeasible. Scraping is a big data technique used to quickly work around these memory restrictions in computers. To gather pre-process the turnstile data, I run a sequence of python scripts developed by github user piratefsh to scrape the data from the MTA website and convert the cumulative turnstile data into individual measures of entries and exits.

There are three scripts that must be used to perform the pre-processing on these data. The first is the turnstile scraper which downloads the raw data into an sqlite3 database, a file from which the tables containing the turnsitle data can be accessed. After the sqlite3 database is created, the turnstile cleaner is used to convert the cumulative entries and exits into individual entry and exit measures. Once this process is complete, the turnstile cleaner is used a second time to remove any outliers in the dataset which may be due to technical failure in the subway system or user error. The script defines an outlier as any observation for any turnstile that is greater than five standard deviations above or below the mean number of entries or exits. Once the sqlite3 databases are cleaned, they must be converted to .csv files so Stata can read them. The easiest way to do this is to use SQLite Studio to convert each database from a .db to a .csv file.

Once the data are imported into Stata, each turnstile observation can be matched to each subway station using the Remote Unit/Control Area/Station Name Key available on the MTA website by performing a many-to-one dataset merge. The next pre-processing step is to aggregate the turnstile data by station and by day. Before this can be done, the date-time string variables must be converted to date objects. Once this is complete, the aggregation can be done by collapsing the dataset by each station-day pair. This shrinks the data into a panel dataset where the individual index is the subway station and the time index is the day. Once the aggregation process is complete for all time periods, the datasets can be appended to one another. The final result is the subway ridership panel dataset across all subway stations ranging from January 2011 to December 2014.

Bike-Sharing Infrastructure Data

Incorporating the Citi Bike data into the subway ridership dataset requires spatial calculations which can be conveniently performed in QGIS. To do this, the data provided by the NYC government on subway station locations are the base for creating measures of bike-sharing infrastructure. Since their data on subway station locations are stored as a shapefile, they can be imported into QGIS as a geographic layer. The second necessary layer contains the locations of all Citi-Bike stations and the number of docks across each station. These data can be found at the Citi Bike live feed. The data are stored as a .json file, which must be converted to a .csv to be usable. Since the bike-sharing dataset is small, an online .json to .csv converter is the most convenient method to make the appropriate conversion.

Since the .csv file contains the coordinate pairs of every Citi Bike station, the file can be imported as a new layer in QGIS over the existing subway station location layer. To calculate the measures of bike-sharing infrastructure, I create a buffer of 200m around each subway station. For every subway station, QGIS calculates the number of bike-sharing stations and docks within the radius and stores the result as a new attribute in the subway station shapefile. The new shapefile can be exported as a .csv file which can then be matched to the existing station names within the subway ridership dataset.

Since there is no standardized way of writing the names of subway stations between MTA dataset and NYC City government dataset, merging the bike-sharing infrastructure data into the subway ridership by subway station name is impossible. For example, in the GIS shapefile, the 1st Avenue station is recorded as "1 AVE" whereas in the MTA dataset it is recorded as "1st Avenue". While these stations are clearly identical, a computer program like Stata would not recognize this relationship. Therefore, the subway station names must be matched manually by a unique index. Then the datasets can be merged by the index as if they were merged by station.

Citi Bike Origin-Destination Data

The raw Citi Bike data can be downloaded from the Citi Bike website. Each data file is broken down by month and disaggregated at the ride level. First, the files should be downloaded and appended to one another in Stata. Once the data are appended, they must be aggregated by start station-day and end station-day pair to create two station-level panel datasets of Citi

Bike ridership over time. To do this, the day-time string variables must be converted to day objects. For the aggregation step, I consider the day of each Citi Bike trip to be the day on which it begins. After aggregation, each dataset contains the the number of Citi Bike trips that begin and end at each bike-sharing station across each day from July 1, 2013 to December 31, 2014. To make the panel data as balanced as possible, I drop every observation for which the origin or destination information are missing.

For the dataset used in the between analysis of the first mile, last mile problem, I import the Citi Bike station location shapefile into QGIS and count the number of subway stations that fall within a 200m radius of each Citi Bike station. I use these counts as the main measure of subway infrastructure. For the Citi Bike data, I aggregate ridership by station in the same way as above but separate the origin and destination counts into morning and afternoon time periods. Doing this allows me to assess how Citi Bike ridership patterns vary across stations throughout the day.

Land Use Data

The cross-sectional analysis of the relationship between bike-sharing and subways requires extensive GIS processing. The first step is to construct a geographic layer containing the spatial influence of each bike-sharing station. To do this, I employ the existing bike-sharing station layer and construct a layer of Voronoi polygons (also known as Thiessen polygons), which partitions the plane of New York City into regions based on the relative locations of bike-sharing stations. This can be completed with one command in QGIS. After creating the polygons, they must be trimmed with the vertex editor to only cover southern Manhattan and Brooklyn.

The land use data emplyed in the analysis come from a variety of sources and therefore, must be merged in QGIS. The land use data are maintained by the NYC Department of City Planning and links to their data can be found at their website. For this study, I use their commercial district, residential district, and manufacturing district geographic datasets They are available for download as shapefiles which can be imported directly into QGIS. The raw shapefile contain polygons with numerous overlaps which makes spatial calculations impossible. I work around this problem by creating a zero-length buffer around each polygon in the shapefiles. This eliminates the overlaps within and between the polygons.

To incorporate the land use data into the cross-sectional bike-sharing station data, I calculate the proportion of each Voroni region covered by commercial, residential, and manufacturing

districts. To do this, I slice the land use polygons along the Voronoi region boundaries for each land use district shapefile. This divides the land use polygons and prevents them from overlapping multiple Voronoi regions. To calculate the share of each Voronoi region occupied by each land use type, I divide each land use area into each Voronoi region area.

The remaining land use data required are included to proxy for general bicycling activity. To proxy for bicycling activity throughout NYC, I procure shapefiles of bike rack and bike lane locations. These data are maintained by the NYC Department of Transportation and can be downloaded from NYC Open Data. These data are provided as shapefiles which can be imported into QGIS. To incorporate these data into the cross-sectional bike-sharing station data, I sum the number of bike racks, bike lanes, and length of all bike lanes that intersect each Voronoi region. After all of the land use data are merged in QGIS, the final layer can be saved and exported as a .csv file. This file can be imported into stata and merged with the Citi Bike ridership panel data by station from the previous section.

Demographic Data

I require cross-sectional demographic information to proxy for bike-sharing opportunities, since more bike-sharing opportunites occur in places with more people. These datasets along with many other related shapefiles can be found at NYC Open Data, a free online source for that houses datasets from numerous public organizations. For this thesis, I procure population data from the 2010 census in a shapefile format maintained by the NYC Department of City Planning. This file is broken down at the census tract level and can be imported directly into QGIS and merged into the existing Citi Bike station locations layer.

Weather Data

There is substantial seasonal and weather-related variation in bike-sharing ridership which makes weather controls necessary for isolating the causal effects of interest. The NOAA maintains an online database containing daily weather information for all observatories across the country. The data used in this thesis were gathered by sumbitting a custom historical data request at the NOAA website. The requests only take a few minutes to process, which makes this a very convenient method to gather a large volume of weather-related data. The data are sent to an email of choice in a .csv format, which makes merging the information into the MTA dataset

simple in Stata. They limit the number of days for which one may download data at one time. For this project, two custom requests were required to gather weather data spanning from January 2011 to December 2014.

References

Babagoli, Masih A., Tanya K. Kaufman, Philip Noyes, and Perry E. Sheffield. "Exploring the health and spatial equity implications of the New York City Bike share system." Journal of Transport & Health 13 (2019): 200-209.

Campbell, Kayleigh B., and Candace Brakewood. "Sharing riders: How bikesharing impacts bus ridership in New York City." Transportation Research Part A: Policy and Practice 100 (2017): 264-282.

Dell'Amico, Mauro, Eleni Hadjicostantinou, Manuel Iori, and Stefano Novellani. "The bike sharing rebalancing problem: Mathematical formulations and benchmark instances." Omega 45 (2014): 7-19.

Fishman, Elliot. "Bikeshare: A review of recent literature." Transport Reviews 36, no. 1 (2016): 92-113.

Garcia-Palomares, Juan Carlos, Javier Gutiérrez, and Marta Latorre. "Optimizing the location of stations in bike-sharing programs: A GIS approach." Applied Geography 35, no. 1-2 (2012): 235-246.

Graehler, Michael, Alexander Mucci, and Gregory D. Erhardt. "Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes?." In Transportation Research Board 98th Annual Meeting, Washington, DC, January. 2019.

Haider, Zulqarnain, Alexander Nikolaev, Jee Eun Kang, and Changhyun Kwon. "Inventory rebalancing through pricing in public bike sharing systems." State University of New York at Buffalo (2014).

Kain, John F. "Housing segregation, negro employment, and metropolitan decentralization." The quarterly journal of economics 82, no. 2 (1968): 175-197.

Ma, Ting, Chao Liu, and Sevgi Erdoğan. "Bicycle sharing and transit: does Capital Bikeshare affect metrorail ridership in Washington, DC." In 83rd Annual Meeting of the Transportation Research Board. 2015.

Martin, Elliot W., and Susan A. Shaheen. "Evaluating public transit modal shift dynamics in

response to bikesharing: a tale of two US cities." Journal of Transport Geography 41 (2014): 315-324.

McFadden, Daniel. "The Measurement of Urban Travel Demand." Journal of public economics 3, no. 4 (1974): 303-328.

Nechyba, Thomas J., and Randall P. Walsh. "Urban sprawl." Journal of economic perspectives 18, no. 4 (2004): 177-200.

Noland, Robert B., Michael J. Smart, and Ziye Guo. "Bikeshare trip generation in New York city." Transportation Research Part A: Policy and Practice 94 (2016): 164-181.

Shaheen, Susan, Elliot Martin, and Adam Cohen. "Public bikesharing and modal shift behavior: a comparative study of early bikesharing systems in North America." (2013).

Thaler, Richard H. "Mental accounting matters." Journal of Behavioral decision making 12, no. 3 (1999): 183-206.