CUSTOMER PREFERENCE AND STATION NETWORK IN THE LONDON BIKE SHARE SYSTEM

PU HE
COLUMBIA BUSINESS SCHOOL,

FANYIN ZHENG
COLUMBIA BUSINESS SCHOOL,

ELENA BELAVINA
CORNELL UNIVERSITY, AND

KARAN GIROTRA
CORNELL TECH

Abstract. We study customer preference for the bike share system in the city of London. We estimate a structural demand model on the station network to learn the preference parameters and use the estimated model to provide insights on the design and expansion of the bike share system. We highlight the importance of network effects in understanding customer demand and evaluating expansion strategies of transportation networks. In the particular example of the London bike share system, we find that allocating resources to some areas of the station network can be 10 times more beneficial than others in terms of system usage, and that the currently implemented station density rule is far from optimal. We develop a new method to deal with the endogeneity problem of the choice set in estimating demand for network products. Our method can be applied to other settings, in which the available set of products or services depends on demand.

1. Introduction

Bike share systems have rapidly expanded across major cities of the world. Cities such as New York, London, and Paris have all introduced this new shared transport service in the past few years. There are many benefits associated with the bike share system. Studies have found that

bike sharing systems have positive effects on public health by creating a large cycling population (DeMaio, 2009; Woodcock *et al.*, 2014). Researchers have also shown that there are significant environmental gains from introducing the systems (DeMaio, 2009). From the managing company or the local government's perspective, however, there are a lot of challenges and room for improvement in managing the bike share systems (Midgley, 2011). One of the main challenges is the design and expansion of the docking station network (New York City Department of City Planning, 2009; Transport for London, 2010). For example, if the manager's goal is to maximize bike usage on the network and capture as well as possible the benefits of the new transport service, it is important to know where to expand the network and where to install new stations.

Since bike sharing programs are a relatively recent phenomenon, few studies have focused on the network design and expansion of the stations. In practice, managing companies and local governments have largely relied on ad hoc rules and policies. For example, the city of London has been implementing a 300-meter density rule, which states that two neighboring docking stations should be, at most, 300 meters away from each other across the city (Transport for London, 2010). Given the potential customer demand variation across the city, it is not clear whether the uniform density rule is optimal. In New York city, the Citi bike network focused primarily on building a high-density network in the downtown area in the first few years since its introduction. After many major expansions of the network, the system still covers only the downtown and midtown areas in Manhattan, and very few places in uptown Manhattan and Brooklyn.

In this paper, we provide guidance on the network design and expansion question of the bike share system using the example of the system in London. The analysis is conducted in two steps. First, we estimate customer demand on the station network using system usage data. We take the structural estimation approach for the following reason. Since the objective is to provide network design and expansion recommendations to managers, we need a model of customer behavior to recover the preference parameters in order to evaluate different counterfactual expansion strategies. Without a structural model, simple regression analyses do not recover customer preference parameters, and therefore we would not be able to use the estimated parameters to evaluate counterfactural expansion and design experiments of the network.

To model customer demand, the natural way to go is to treat the docking stations as products and to assume that customers choose the stations based on the utility they gain from using the bikes at those stations (Kabra *et al.*, 2016; Singhvi *et al.*, 2015). This approach is problematic, however, because it leaves out the important network structure between stations. Customers will choose the origination station only if the destination station is also attractive. In other words, customers are choosing the route or the link on the network between stations instead of the individual stations. For instance, if there is only one station in the network, demand is going to be very low because customers will be able to make only trips originating and ending at the same station. When a node has a lot of links on the network, and the links are attractive routes to the customer, however, the demand at that station is going to be substantially higher. This is what we refer to as network effects in this paper. Those effects can be captured only if we take the entire network into account, instead of treating stations as independent, when modeling the choices of the customers.

In light of the importance of the network effects in the current setting, we estimate customer demand for each origination and destination station pair instead of for individual stations. In other words, we study customers' preferences on the routes generated by a station network. This creates two challenges in the estimation. The first one is the endogeneity problem of the choice set. The choice set of the customer is endogenous because whether a station is in the choice set depends on whether it has bikes or docks available. The availability of bikes and docks at a station can be correlated with unobserved characteristics of the station, which then give rise to the endogeneity issue. The problem is particularly difficult to solve in a network setting in which the conventional instrumental variable approach does not apply. Using reduced-form regression evidence, we first show that this problem leads to biased estimates and unreasonable policy recommendations. Then, we propose a novel instrumental variable solution to this problem and show that our solution removes the bias in the parameter estimates and provides reasonable policy recommendations. The second challenge in estimation is the computational difficulty given the extremely high number of routes in the network. We reduce the computational burden by dividing the coverage area into blocks and model demand on the block level. We argue that this approximation is reasonable given the objective of evaluating long-term expansion strategies.

In the second step, we use the estimated model and customer preference parameters to provide insights into the design and expansion of the docking station network. This is done through three counterfactual analyses. In the first counterfactual, we evaluate a particular expansion proposal by the local government and predict network usage increase after the expansion. It demonstrates the practical insight that our study can provide to managers of such bike share systems. In the second and third counterfactual, we generalize the insight and highlight the importance of network effects in studying customer demand on bike share systems. We compute the effect of adding stations and adding bikes or docks to different parts of the network in the two counterfactuals, respectively. First, we show that increasing density in the city center leads to usage increase ten times as high as that from increasing the scope of the network. This shows that the 300m density rule in the London system is far from ideal. Second, we decompose the variation in usage increase at different locations of the network. We identify two types of network effects and show that network effects play a key role in understanding demand variation across the network and evaluating the expansion strategies.

Despite the importance of the design and expansion of the station network, there are very few empirical studies on this topic. The closest to our paper is Kabra *et al.* (2016), who study the demand of the bike sharing system at the station level in Paris. The main difference is that our paper focuses on route-level demand, which brings the network effect into the analysis. This allows us to evaluate the changes to the network of stations by usage throughout the entire system instead of focusing on usage at individual stations. Our paper is also closely related to several studies analyzing the local demand and rebalancing of bikes in the Citi bike system in New York (O'Mahony and Shmoys, 2015; Singhvi *et al.*, 2015). The main difference is that we model customer behavior and recover preference parameters in the structural model, which allows us to compute counterfactual predictions and provide prescriptive recommendations to managers. Our work also contributes to the broader literature of ride sharing and car sharing, with a focus on the spatial network structure of customer demand (Jorge and Correia, 2013).

There are many studies in the transportation literature that estimate origination-destination traffic demand (Ashok and Ben-Akiva, 2000; Ben-Akiva *et al.*, 2001; Barcelö *et al.*, 2010; Toole

et al., 2015). Some of the studies use similar discrete choice modeling tools to study customer behavior (Ashok and Ben-Akiva, 2000; Ben-Akiva *et al.*, 2001). However, those studies have mainly focused on real time traffic predictions, instead of recovering causal customer preference parameters and applying the estimated model to provide long term policy recommendations. In addition, our model is much more flexible as we allow for unobserved route characteristics and customer heterogeneity, compared with the relatively simple models commonly used in this literature (Ashok and Ben-Akiva, 2000; Ben-Akiva *et al.*, 2001).

The main estimation method we use in the analysis is based on the classic demand estimation framework introduced by Berry *et al.* (1995) and the MPEC algorithm in Su and Judd (2012). The new method we develop to account for the endogeneity problem of the choice set relies on the network structure. The intuition of this method is related to that of the identification strategy in Bramoullé *et al.* (2009). The method also contributes to the literature on estimating demand with endogenous stock-out products studied by Musalem *et al.* (2010) in operations management and by Conlon and Mortimer (2013) in economics.

The main contribution of the paper is threefold.

1. Practical guidance Our analysis provides important practical guidance to managers and the local government in evaluating network expansion strategies in London. With similar data from other cities, the framework can also be easily applied to other bike share systems to understand customer demand and evaluate expansion proposals.

2. General insights Our analysis highlights the importance of network effects in studying customer demand on a transportation network or products with a spatial aspect. We provide strong evidence to show that treating bike stations as individual products is far from sufficient. The structure of the network–i.e., where the connecting nodes are and what the weight is on each link–plays a significant role in determining the demand on the each node.

3. Methodology We illustrate the problem of endogenous choice set using regression analyses. We show that the endogeneity problem leads not only to biased estimates, but, more importantly, to unreasonable policy recommendations. We provide a novel instrumental variable approach to

address the problem in a spatial network setting. The method can be easily applied to studying demand for other network or spatial products.

2. Background and data

2.1. **Background.** The bike share system in London, "Boris Bikes", was introduced in 2010. Like many bike share systems in major cities around the world, "Boris Bikes" went through a few major expansions after the initial launch.[1] There are calls and proposals to expand the station network further–for example, to central and south Islington and Hackney, the borough of Southwalk. "Boris Bikes" has been an important part of the local government's urban development program. In the 2016 mayoral election, both Labour and Conservative party candidates pledged to expand "Boris Bikes" and make London more bike-friendly.[2]

2.2. **Data.** The data we use in the analysis consist of four parts.

2.2.1. *Stations and trips.* First, we have data on the stations of the London bike share system and the trips made by customers on the system in 2014. There are 724 bike stations on the network. For each station, we observe the exact location (longitude and latitude coordinates) and the size of the station (total number of docks). The system covers an area of about 8 kilometers by 16 kilometers at the center of the city.[3] For the trip data, we observe the location (longitude and latitude coordinates) of the origination station and the destination station of each trip, as well as the starting and ending time of the trip. There are about nine million trips in total. We find that there is huge variation in the number of trips across routes, where a route is defined as a directional link between two stations. Some routes have thousands of trips, while others have only one trip.[4] It appears that, similar to retail sales data, the distribution of usage has a long tail, where many routes have very low usage, while a few routes have very high usage. There are also many routes with zero usage throughout the year. In the subsequent analysis, we follow the standard practice of

[1]Source: `https://tfl.gov.uk/info-for/media/press-releases/2013/december/mayor-launches-huge-expansion-of-flagship-barclays-cycle-hire-scheme`
[2]Source: `https://www.theguardian.com/environment/bike-blog/2016/apr/14/london-mayoral-election-qa-on-cycling-policy-with-the-main-candidates`
[3]We provide a map of the Greater London administrative area with the system coverage area in the Online Appendix.
[4]See the Online Appendix for detailed summary statistics.

including only routes with positive usage. In the structural estimation, in particular, we aggregate routes across nearby stations, which alleviates the long tail issue.

2.2.2. *Availability Snapshots.* Second, we have station snapshot data for each station every five minutes throughout 2014. The snapshots contain information about the number of bikes and docks available at each station. To demonstrate the variation in the snapshot data, we compute the following availability measures during the busiest time windows for the system: weekday morning rush hour (5:30am-9:30am) and evening rush hour (4:00pm-8:00pm). We define the bike availability measure as the percentage of time a station has at least five bikes available–i.e. $bike\_avail = \frac{\text{num of 5-min intervals with >=5 bikes}}{\text{Total num of intervals}}$; and, similarly, we define the dock availability measure as $dock\_avail = \frac{\text{num of 5-min intervals with >=5 docks}}{\text{Total num of intervals}}$. Similar to Kabra *et al.* (2016), we use five as a threshold to allow for the possibility of broken bikes and docks. We present the summary statistics in Table 1. For all four time windows and two availability measures, there is substantial variation in bike and dock availability across stations. This shows that availability conditions vary a lot at different locations, which is important to take into account in the subsequent analysis.

Time Window	Availability	Mean	sd	Min	25%	median	75%	Max
morning rush hour	bike availability	0.73	0.21	0.10	0.58	0.77	0.90	1.00
evening rush hour	bike availability	0.74	0.16	0.19	0.63	0.76	0.88	1.00
morning rush hour	dock availability	0.80	0.14	0.27	0.71	0.83	0.91	1.00
evening rush hour	dock availability	0.84	0.13	0.28	0.77	0.87	0.94	1.00

Table 1. Summary Statistics for Long-term Availability Measure

2.2.3. *Google Data.* We collect four sets of data from Google API. The first set is the biking distance data. We use ij to denote a route from station i to station j. For each route ij, we collect the biking distance data from Google Maps using the best biking route from location i to j according to Google. We refer to this distance as the biking distance for the rest of the paper.

Second, in addition to the biking distance, we also collect the data on the change of elevation along each route ij. A standard measure for the degree of inclination in transportation is the slope

grade. Google Maps data allows us to divide each route ij into many short segments, and compute the grade for each segment. We define two features for each route: 1) $AvgAscendGrade_{ij}$: the average grade among the ascending segments of route ij, and 2) $AscendPercentage_{ij}$: the proportion of ascending segments on routes ij. We provide calculation details in the Online Appendix.

Third, we collect the data on other travel options in the city, in order to understand the outside options the customers face when choosing whether to use the bike share systems. We collect the data on two features, the distance and the travel time, for each of the two alternative travel options, driving and public transportation. We find that the travel time is highly correlated with the distance in the data, so we only use the distance feature in the analysis from now on.

Fourth, we collect Google Places data in London. This dataset provides the longitude and latitude coordinates of 97 categories of places that are identified on Google Maps, including subway stations, government office buildings, schools, restaurants, etc.[5] We group 97 Google Place categories into ten general categories, food, health, religion, entertainment, stores, government offices, transportation, education, finance and others, and use these general categories in our analysis. Detailed definition of the groups and the summary statistics are presented in the Online Appendix.

We divide the coverage area of the bike share system into 200 by 200 meter squares and count the number of each Google place category in each square. Since many categories have zero count for more than 40% of the squares, it can be more informative to use the total counts instead of the counts for each category in the analysis. In addition to this observation, we also calculate the correlation between each pair of category counts, including the total counts. [6] We find that the average pairwise category correlation among the ten categories of Google places is 0.29, and the total Google place count is highly correlated with any of the ten category counts, with an average correlation of 0.56. This observation implies that using total Google place count in the analysis can be a good approximation of using all ten category counts.

[5]Both the data content and the category used in Google Place data change over time. Unfortunately, historical data are not available. The set of the data we use was scraped in February 2017.
[6]We present the correlation matrix in the Online Appendix.

2.2.4. *Census.* Finally, we use demographic data from the 2011 Census, which is more complete than the data from more recent years. The data includes population, income, age, gender, and related demographic information, measured for each Lower Super Output Area (LSOA) in London.[7] LSOA is the smallest census unit with accurate data, and there are, in total, 4835 LSOAs in the London city. Our 724 stations in the bike share system cover 430 LSOAs located in the center of the city.[8]

2.3. **Preliminary evidence.** To motivate our analysis, we present several pieces of preliminary evidence about the usage of the system before going into details about the model and the estimation. The evidence comes directly from the data and therefore is model-free.

First, we find that 75% of the total usage of the London bike share system is on weekdays. Since weekday and weekend usage patterns can be very different, and from the local managing company's perspective, the local population's weekday usage matters more, we focus on weekday usage for the analysis. Within the weekday usage data, we find that 60% of the total trips occur during the morning rush hour (5:30am to 9:30am) and the evening rush hour (4pm-8pm) on weekdays. Moreover, the local government clearly makes the commuters the main beneficiaries when discussing plans for expanding the network (Greater London Authority, 2013). For these reasons, we restrict our analysis to rush hour usage.

Next, we provide some evidence on the spatial pattern of the system and its usage. Figure 2.1 shows the usage pattern at 9:30am on a weekday morning in London. Each circle represents a docking station. The shade of the circle corresponds to the number of bikes available divided by the total number of docks at the station, ranging from black, which indicates that the station is full of bikes, to white, which indicates that the station is completely out of bikes.

Figure 2.1 shows that, towards the end of the morning rush hour, many stations around the city center are full, and many of the stations in the more residential areas in the outer part of central London are empty. This implies that people pick up bikes from where they live in the morning, commute to work, and return bikes near where they work in the center of the city. The pattern

[7]The data were downloaded from the local government's website, `https://data.london.gov.uk/dataset/lsoa-atlas`

[8]We provide the summary statistics for both the 430 covered LSOAs and all 4835 LSOAs in the Online Appendix.

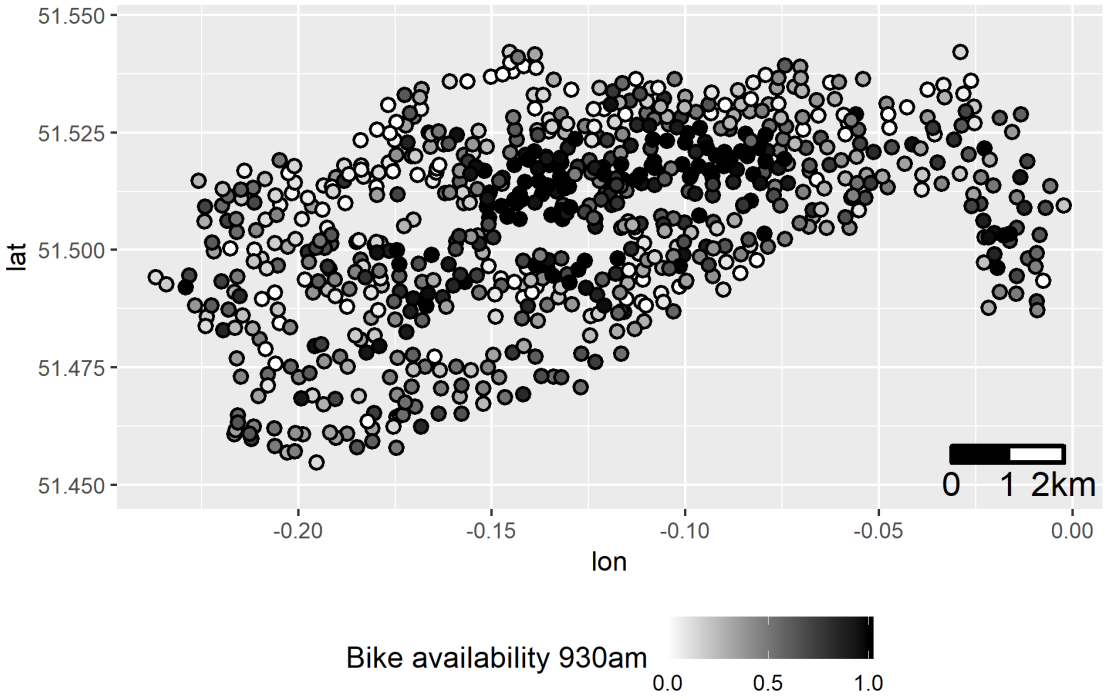

FIGURE 2.1. End of morning rush hour system status

of traffic during morning rush hour is generally from the outer part of central London to the very center of the city. We see the opposite direction in the usage pattern during evening rush hour, captured in Figure 2.2. This is a snapshot of the system around 8pm on the same day. Figure 2.2 shows that a lot of bikes have been picked up in the very center of the city and returned to the more residential areas around the outer part of central London after the evening rush hour commute. This directional pattern of traffic on the network during rush hours will also show up in our estimation results presented later. More importantly, we will rely on the directional pattern to construct our instruments in order to identify the preference parameters in the structural model.

In Figures 2.1 and 2.2, another pattern in the data concerns the density of bike stations. The station density is slightly higher in the very center of the city but is more or less uniform otherwise. This is partly due to the 300m rule between stations imposed by the local transportation department, Transport For London (TFL). We will discuss whether the policy is reasonable, as well as its implications for how to expand the network when discussing the counterfactual analyses.

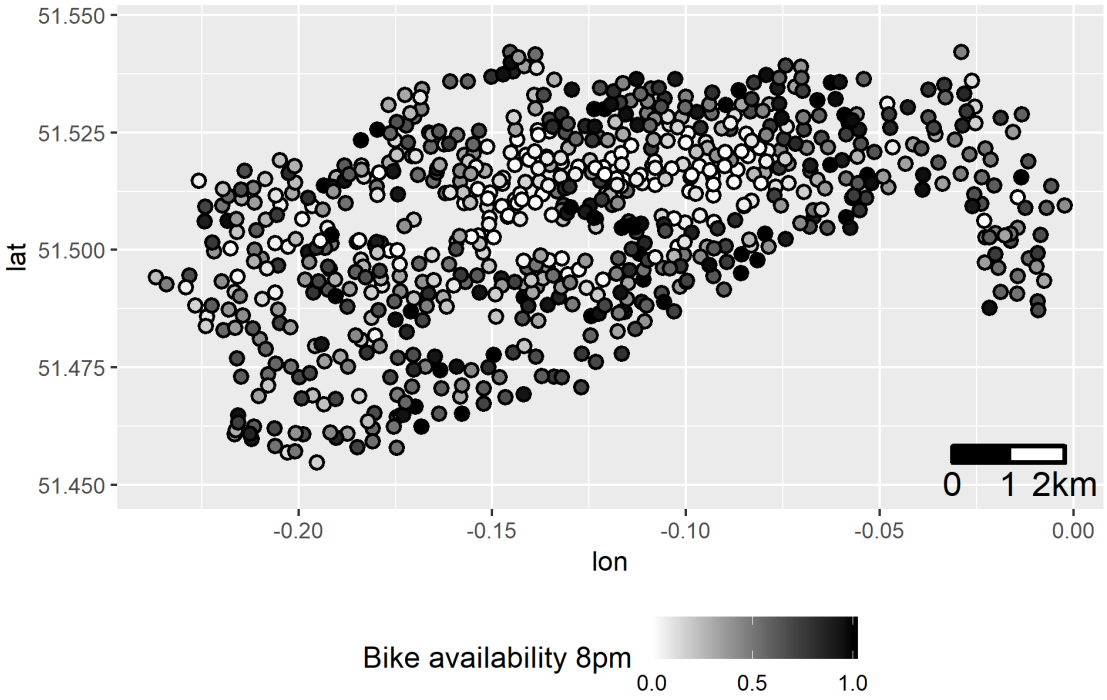

FIGURE 2.2. End of evening rush hour system status

The last piece of evidence concerns the relationship between usage and one key route characteristic that we find in the data: route distance. We plot the average usage per route against route distance in Figure 2.3, which shows that, at first, usage increases very fast as distance increases. This is because as distance increases, more potential customers would prefer biking to walking. The peak usage occurs around 1.5 kilometers (km). As distance increases further, usage decreases quickly, as more potential customers would prefer other forms of transportation to biking. As shown in the estimation results in Section 4, route distance is a key characteristic that determines usage in the bike share system. The increasing and then decreasing usage suggests a non-linear preference pattern for route distance, which will prove to have important implications for network design and expansion. We will revisit this point in the counterfactual discussion.

The above three pieces of model-free evidence provide guidance to the modeling choices in our structural estimation analysis. Note that this evidence alone is not sufficient to provide prescriptive policy recommendations we are interested in. They are merely correlation patterns in the data. To

FIGURE 2.3. Average usage per route vs. route distance
Usage is measured in number of trips per route, and route distance is measured in meters. Peak usage around 1km.

understand customer preferences and therefore predict their choices under different counterfactual scenarios where the network is expanded, we need a choice model and an estimation procedure to recover the preference parameters when describing the model.

3. Choice model

In this section, we present the structural choice model. We do not argue that structural estimation is the only approach to understanding customer demand. But the reason that a structural model is necessary in our setting is as follows. The goal of the analysis is not just to understand customer demand, but, more importantly, to provide the managing company and local government with prescriptive policy recommendations in terms of station network design and expansion. Therefore, we would like to provide out-of-sample predictions that might be far from the observed data. To achieve this goal, we need to rely on a behavioral model that describes how potential customers make commuting decisions, to recover the preference parameters in the model, and to use the model

to compute counterfactual predictions. Without a structural model, the parameters estimated from reduced-form regressions are not interpretable and can not help us make counterfactual predictions.

The model consists of two parts. The first part is a classic multinomial logit model that describes how customers choose between biking on a set of routes and using an outside commuting option. The second part consists of feeding the derived choice probabilities from the multinomial logit model into a density model that captures the differences in the number of potential customers in different geographic areas. We allow all parameters in the model to have different values for customer demand during the morning rush hour (MR) and the evening rush hour (ER). But for simplicity, we do not carry the MR and ER subscripts in the model.

We now discuss the model in more detail. Let U_{ijkl} be the utility that a customer who wishes to travel from location k to location l gets from using the bike share system to cycle from station i to station j, if bikes are available at station i, and docks are available at station j. Then, we have

$$U_{ijkl} = X'_{ijkl}\beta + \xi_{ij} + \varepsilon_{ijkl} \qquad \forall ij \in C_{kl} \tag{3.1}$$

X_{ijkl} is a vector of characteristics of route ij for customer traveling from k to l. It includes three types of characteristics in addition to a constant. The first type includes two terms: $\log d(i,j)$, the logarithm of the biking distance between stations i and j, and $\max\{\log d(i,j) - \log b, 0\}$, which captures the potentially non-linear distance preferences for long distance routes. We define long distance routes as those longer than 1.5 km–i.e. $b = 1.5$. This is natural in light of the fact that 1.5 km is the peak point in Figure 2.3. We provide robustness checks for other values of b in the Online Appendix. The second type of covariates are the two elevation features of route ij, $AvgAscendGrade_{ij}$ and $AscendPercentage_{ij}$ as defined in Section 2.2.3. They capture how physically demanding it is to bike from i to j. Third, X_{ijkl} also includes two walking distance variables: $d(k,i)$, the walking distance between the customer's origination location k and the starting station i, and, similarly, $d(j,l)$, the walking distance between the ending station j to the customer's destination location l. Although walking distance is not the focus of our study, we include them since other studies in the literature show that it is an important factor for demand (Kabra *et al.*,

2016). Notice that unlike Kabra *et al.* (2016) who study station level demand and, therefore, only have the walking distance between the origination location of the customer and the starting station in their model, we include walking distance on both ends of the route.

ξ_{ij} are unobserved route characteristics, which can include whether route ij is bike-friendly, whether there is a bike lane, or other features which make ij more or less attractive. ε_{ijkl} are idiosyncratic error terms that are independent and identically distributed and follow extreme value distribution.

Since we differentiate customers only by the origination location k and the destination location l, we index customers by kl. Customer kl chooses from a set of possible routes ij and an outside option to maximize her utility. The choice set of all biking routes for customer kl is denoted by C_{kl}, which includes all routes with origination station i within walking distance of k and destination station j within walking distance of l–if i and j have bikes and docks available. Therefore, the choice set is kl-specific. We specify C_{kl} rigorously in Section 4. We define the utility of customer kl's outside option as

$$U_{0kl} = X'_{0kl}\theta + \varepsilon_{0kl}, \tag{3.2}$$

where X_{0kl} includes the characteristics of customer kl's alternative transportation options. In the main specification, we include the distance of driving and the distance of public transportation from Google Maps data, as described in Section 2.2.3. The choice probability of kl choosing $ij \in C_{kl}$ is then

$$P_{ijkl}(X, \beta, \theta) = \frac{exp(X'_{ijkl}\beta + \xi_{ij})}{exp(X'_{0kl}\theta) + \sum_{i'j' \in C_{kl}} exp(X'_{i'j'kl}\beta + \xi_{i'j'})}. \tag{3.3}$$

For $ij \notin C_{kl}$, we have $P_{ijkl} = 0$. Let q_{ij} be the total number of trips on route ij predicted by the model. Then

$$q_{ij}(X, W, \alpha, \beta, \theta) = \int P_{ijkl}(X, \beta, \theta) dD(W_{kl}, \alpha), \tag{3.4}$$

where $D(W_{kl}, \alpha)$ is a density function that measures the number of potential customers traveling from location k to location l. It is a function of W_{kl}, the characteristics of the commuter origination and destination pair. In W_{kl}, we include the population density and the number of Google place counts at both k and l. Given that the objective of our analysis is to understand long term average customer demand and to evaluate different network expansion strategies which mainly involves comparing usage levels across locations, we do not model the high-frequency variation of usage over time.[9] Instead, we use Equation (3.4) to describe the average usage during the morning and evening rush hours, allowing all the parameter values to be different for the two rush hour time windows. We do not model demand variations on the same route across time or time dimension substitutions on the same route. Therefore, taking the morning rush hour as an example, q_{ij} in Equation (3.4) should be interpreted as the average daily usage of route ij during the morning rush hour, or equivalently, the total usage in a year during the morning rush hour. Similarly $D(W_{kl}, \alpha)$ measures the travel demand from location k to location l on an average day during the morning rush hour, or equivalently, the total travel demand from k to l throughout a year.

The core of the model is multinomial logit. It is known for implying unrealistic substitution patterns or the independence of irrelevant alternatives property. However, Berry *et al.* (1995) illustrate that the multinomial logit model can allow for flexible substitution patterns if one introduces random coefficients. Here, we take an alternative approach and utilize the spatial aspect of the data to generate, through the density function, flexible substitution patterns. The density function plays the role of the random coefficient in the following sense. The options chosen by nearby customers with similar W_{kl} values are closer substitutes than otherwise. This breaks the independence of irrelevant alternatives property of the multinomial logit models. See Ellickson and Misra (2008) and Kabra *et al.* (2016) for similar approaches.

The unknown parameters that we need to estimate in this model are β, θ, and α in Equations (3.3) and (3.4). Note that all three parameters are vectors. We detail the estimation procedure in the

[9]Another reason for not using the temporal variation in the data is that we do not have exogenous covariates that vary over time, which introduces difficulties in estimating the parameters consistently. In fact, if we consider the morning rush hour as an example, let t be different days or weeks in a year, none of the features X in (3.3) and (3.4) depends on t.

following section. With the estimates of β, θ, and α, we will be able to predict the customer choices across the city and, therefore, the overall changes in usage on the network under counterfactual station network expansions.

4. Estimation

In this section, we explain the estimation of the model. We start by introducing the endogenous choice set problem. Then, we explain our proposed solution to the problem and provide evidence showing that, without properly accounting for the endogeneity problem, the parameter values can not be recovered without bias, thus leading to unreasonable expansion policy recommendations. Finally, we provide details on the rest of the estimation procedure and present the structural estimation results.

4.1. **Endogeneity of choice set.**

In a classic discrete choice model, customers choose from a set of products or service options to maximize utility. The choice set is either perfectly observed or pre-set by the researcher. In our setting, the choice set is the set of possible routes ij from which customers can choose to commute from k to l. Compared with the classic setting, the complication here is that whether a particular route ij is in the choice set of a customer, C_{kl}, depends on whether there are bikes available at station i for the customer to pick up, and whether there are docks available at station j for her to return the bike to. As shown in Table 1, some stations frequently run out of bikes and docks, and there is significant variation across stations in terms of bike and dock availability. Of course, whether a station has bikes or docks available is *not* randomly assigned. It depends on the usage level at the station, or, in other words, how popular it is. Indeed, more popular origination stations are more likely to run out of bikes, and more popular destinations are more likely to run out of docks. Therefore, the choice set of the customer depends on how popular or preferable the routes are. This means that the choice set is endogenous to the choice behavior itself.

We now discuss the details of the endogeneity problem and how it biases the estimation results. In our discrete choice model, the utility of customer kl choosing to bike the route ij is given by Equation (3.1). Let X, W, and ξ be the matrices $\{X_{ijkl}\}_{ij=1,\dots,N,kl=1,\dots,M}$, $\{W_{kl}\}_{kl=1,\dots,M}$, and

$\{\xi_{ij}\}_{ij=1,...,N}$, respectively. To obtain consistent estimators of α and β, one necessary condition is that ξ is mean zero, conditioning on X and W–i.e., $\mathbb{E}[\xi|X,W]=0$. Under this condition, one can construct moment conditions $\mathbb{E}[\xi \cdot X|X,W]=\mathbb{E}[\xi \cdot W|X,W]=0$ (Berry *et al.*, 1995, 2004). One important implication of this condition is that ξ_{ij} is uncorrelated with C_{kl} because, although we control for X, there are always route characteristics that are observable to the customer but not to the researcher. These unobserved characteristics, or preferability factors, are captured by the ξ_{ij} term in Equation (3.1). If the choice set C_{kl} is also affected by the unobserved popularity or preferability of the routes, then C_{kl} is correlated with ξ_{ij}. This is the sense in which there is an endogeneity problem with C_{kl}. The endogeneity issue will lead to bias in the estimated parameters.

We now provide the formal reasoning behind the endogeneity problem. To simplify the analysis, we first assume that the choice set C_{kl} of each customer kl is observed. We discuss later the practical implications and feasibility of this assumption. Rewriting Equation (3.3), we have

$$P_{ijkl}(X,\beta,\theta)=\frac{\mathbb{1}_{ij}\cdot\Big[exp(X'_{ijkl}\beta+\xi_{ij})\Big]}{exp(X'_{0kl}\theta)+\sum_{i'j'}\mathbb{1}_{i'j'}\cdot\Big[exp(X'_{i'j'kl}\beta+\xi_{i'j'})\Big]}, \tag{4.1}$$

where $\mathbb{1}_{ij}$ is a variable indicating whether route ij is in customer kl's choice set C_{kl}, i.e., whether station i has bikes and station j has docks available. Since $\mathbb{1}_{ij}$ is precisely observed in the data, we can treat it as a standard route characteristic. Equation (3.4) stays the same. Then, the moment condition we are actually using is

$$\mathbb{E}[\xi|X,W,\mathbb{1}]=0, \tag{4.2}$$

where $\mathbb{1}$ is the vector of $\mathbb{1}_{ij}$, for all ij. Now, one could use Equation (4.2) to obtain estimates for α and β. However, the estimated coefficients will be biased since 4.2 is violated. The reason is that the unobserved product characteristics ξ_{ij} are correlated with the choice set indicator $\mathbb{1}_{ij}$–i.e., $\mathbb{E}[\xi_{ij}\cdot\mathbb{1}_{ij}]\neq 0$. For example, a particular route might be easier or harder to bike, depending on whether it is uphill or downhill, whether it has bike lanes, or whether the traffic along the route is more or less friendly to cyclists. These characteristics are unobserved to the researcher and, as discussed above, captured by ξ_{ij}. If the preferences over such characteristics are correlated across

customers, then a customer arriving later in the time window is more likely to face empty stations with no bikes available, or full stations with no docks available. Therefore, the choice set indicator $\mathbb{1}_{ij}$ is correlated with ξ_{ij}. Moreover, like many other bike share systems, the managing company restocks bikes throughout the day. The reallocation decisions are not random but are optimized by the managing company (O'Mahony and Shmoys, 2015; Freund *et al.*, 2017). The more popular origination and destination stations are also more likely to receive reallocated bikes. Thus, ξ_{ij} can also be correlated with $\mathbb{1}_{ij}$ through the supply of bikes. Therefore, the parameter estimates obtained using moment condition given in Equation (4.2) will be biased due to the endogeneity problem.[10]

It is important to realize that the endogeneity problem of the choice set is not specific to our estimation context. The same logic applies to all customer choice or demand estimation settings. Whenever a product is out of stock, which happens often in practice, it automatically drops out of the customer's choice set. Products are *not* randomly out of stock: more popular products are more likely to run out. As a result, the choice set of the customer is correlated with the unobserved preferability of the products. This leads to the endogeneity problem of the choice set. Since the problem arises whenever there is variability in the choice set, it is common in the context of demand estimation. However, the problem has not been studied extensively in either the operations management or the economics literature. To the authors' best knowledge, only two studies in the operations management literature explicitly discuss this issue. Musalem *et al.* (2010) study the impact of stock-outs on shampoo sales. Kabra *et al.* (2016) look at station-level demand in the bike share system in Paris and take into account the impact of availability. In the economics literature, Conlon and Mortimer (2013) study the vending machine demand for candies, where the customer's choice set can be restricted by stock-out events. We discuss shortly why the methods used in those studies are not applicable in our setting.

The endogenous choice set problem is difficult to deal with directly for the following two reasons. First, choice sets are not modeled in classic discrete choice models. Choice models have focused on deriving choice probabilities for a collection of pre-determined options. The set of options is

[10]Manchanda *et al.* (2004) study the endogeneity problem of ξ_{ij} in a different setting: ξ_{ij} is correlated with mixed marketing activities and thus is endogenous. They show that ignoring the endogeneity problem leads to substantial bias in the parameter estimates.

typically not part of the likelihood function or moment conditions that researchers use to estimate the demand model. Second, it is difficult to apply classic instrumental variable approaches directly to the problem. Instrumental variables are classic tools for dealing with endogeneity problems in regression analysis. However, such approaches require that the endogenous variable enters the regression equation in a linear fashion. In our context, as shown in Equation (3.1), the choice set is nonlinear in the utility function. Therefore, it is infeasible to apply the instrumental variable approach directly. Bruno and Vilcassim (2008) proposes an estimation approach when all product features including the choice set are exogenous. The method does not apply to our setting because of the endogenous nature of the choice set.

Our proposed solution consists of two steps. In the first step, we convert the discrete choice set to a continuous average availability measure. Instead of specifying which routes are available in the choice set at different points in time, we allow all possible routes to be in the choice set, and let the long term average availability vary across routes. In other words, we compute, over the one year sample period, the fraction of time a station i has at least five bikes or docks available during the rush hour window, $bike\_avail_i$ and $dock\_avail_i$. We compute the two measures separately for the morning and evening rush hours.[11] By doing so, we ignore the variation of the choice set within the time window and across different days of the year. Effectively we only utilize the cross-sectional variation in the data across different locations and average out the temporal variation. We argue that this is a reasonable restriction for three reasons. First, our model is used to understand the observed and to predict the counterfactual long-term average usage, instead of usage from one hour to the next. Besides, it is natural to assume that the long-term average usage is the key measure which managing companies and local governments care most about when designing the bike share system. Second, as shown in the preliminary evidence section, the London bike share system is commuter-dominant. Therefore, it is reasonable to focus on average availability because most customers are repeat users who care more about average availability within the rush hour

[11]As discussed below, we estimate the morning and evening rush hour usage separately in two models, in light of the very different usage patterns. Therefore, the availability measures are also computed separately for morning and evening rush hours.

commuting time window.[12] Third, we conduct a variance decomposition exercise for the availability measures and find that the cross-sectional variation is the main source of variation in our data. After converting the discrete choice sets to the continuous average availability measures, we rewrite Equation (3.1) as:

$$U_{ijkl} = \beta_1 bike\_avail_i + \beta_2 dock\_avail_j + X'_{ijkl}\beta_3 + \xi_{ij} + \varepsilon_{ijkl}, \quad \forall ij \in C_{kl}, \tag{4.3}$$

where the choice set C_{kl} consists of the feasible routes regardless of their availability. The long-term bike availability measure at the origination station i and the long-term dock availability measure at the destination station j are included in the utility function. As discussed before, these two availability characteristics are likely to be correlated with ξ_{ij}, and, therefore, are endogenous. However, Equation (4.3) shows that the endogeneity problem of choice set can become a familiar endogenous linear characteristic problem in the utility function. The problem is exactly the same as in most demand estimation, where the price of the product or service is the endogenous characteristic, and the usual instrumental variable approach applies.

We note that Equation (4.3) does not correspond to Equation (3.1) integrated over the availability of routes $ij \in C_{kl}$. Instead, it provides an approximation. To illustrate that our approximation is indeed reasonable, we conduct extensive simulation studies comparing our estimates to the true model predictions in the counterfactual scenarios of interest. In particular, we conduct the simulation studies with different cross-sectional vs. temporal variance composition of the variance of the availability measures. Since the goal of our study is to understand long term usage, it is not surprising that the counterfactual predictions calculated using our estimated model are very close to those calculated using the true model. We provide details of the simulation studies in the Online Appendix.

The second step consists of finding valid instruments for the availability measures. In previous studies, Conlon and Mortimer (2013) exploit a quasi-experiment in product restocking time to deal with the endogeneity problem. However, no quasi-experiment is available in our setting. Musalem

[12] On the other hand, if most usage is from casual users like tourists, short term availability would be more important.

et al. (2010) use supply-side instrumental variables to account for the endogeneity of product stock-outs in supermarkets. However, they find that the estimation results are the same with or without the instrumental variables. Kabra *et al.* (2016) use similar types of instrumental variables as in Berry *et al.* (1995), i.e., the characteristics of nearby stations. But these types of instruments have been shown to have weak identification power (Andrews *et al.*, 2017). Moreover, when one treats stations as nodes on a network instead of as independent stations, the characteristics of nearby stations might become even weaker in their ability to identify the parameter of interest. To overcome these difficulties, we propose a novel set of instrumental variables by utilizing the station network structure. We show that properly accounting for the endogeneity of the choice set is key to recovering consistent estimates of the parameters and providing reasonable policy recommendations. We discuss the details of the instrumental variables in Section 4.2. As we will show later in the estimation results, without the instrumental variables, the bias can be substantial, and, more importantly, it can lead to unreasonable policy recommendations.

4.2. **Instrumental variables.** In this section, we discuss the instrumental variables for the bike and dock availability measures. The main difference between our setting and a classic demand estimation setting is that our data are generated from a station network instead of from independent station observations. When the data is generated from a network, the challenge in finding valid instrumental variables is that the exogenous covariates observed in the data are likely to be correlated spatially and throughout the network. The correlations, therefore, might lead to violations of the exclusion restriction for an instrumental variable to be valid. To solve this problem, we introduce a new method of constructing instrumental variables in a network setting. In particular, we utilize the station network structure to construct instrumental variables for the availability measures. We start the discussion by explaining why the exclusion restriction is violated if we apply the commonly used instrumental variables, and we then describe our proposed solution.

The commonly used instrumental variables for the endogenous characteristic (typically, price; in our case, availability measures) in demand estimation are the exogenous characteristics of the products offered in the same market (Berry *et al.*, 1995). These are the so-called BLP instruments

(Berry and Haile, 2014). The reasoning behind the validity of the BLP instruments is that, if a product is more isolated from the other products in the product characteristic space, then it has higher margin which leads to higher price. This provides exogenous variations in prices. The key point for these instruments to work is that the variations in product characteristics are exogenous. Applying the same idea to the availability at station i, for example, one would use station k's characteristics S_k as instruments, where i and k are close to each other. Kabra *et al.* (2016) use these types of instrumental variables in a similar setting. However, when stations are not independent but are nodes on a network, the BLP instruments may not be valid anymore. For example, if k and i are spatially close, S_k and S_i can be highly correlated. This implies that, conditional on S_i, there is little variation in S_k that we can utilize to identify the coefficient of the availability at i. It could also be the case that since k and i are close to each other, kj and ij are similar routes to customers, which implies that the availability at stations k and i can be very much correlated. Then, it is even more difficult to find exogenous variations to identify the availability coefficients. Both examples indicate that the BLP instruments are not appropriate in the current setting, where spatial proximity and network structure are prominent in the data.

Next, we explain how we deal with the challenge of correlations in network data and introduce the proposed instrumental variables. First, we define two stations as "connected" (or "connected on the network") if there are customers using the route between the two stations. To be precise, stations i and j are connected means that $q_{ij} + q_{ji} > 0$. We use the bike availability at station i as an example to illustrate the construction of valid instrumental variables for the availability measures. We want the instruments to be uncorrelated with ξ_{ij} but correlated with $bike\_avail_i$. For that we look for the average of some exogenous characteristic over station $h \in \mathcal{H}_1(ij)$, denoted as $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h$, where station $h \in \mathcal{H}_1(ij)$ must satisfy the following two conditions.

First, station h must be connected to station i. In other words, some customers bike between h and i. If h and i are connected, the exogenous characteristics of station h affect the usage on route hi or ih and, therefore, the bike availability at station i. The *relevance condition* then holds for S_h:

$$\mathbb{E}[bike\_avail_i \cdot S_h] \neq 0, \tag{4.4}$$

for all $h \in \mathcal{H}_1(ij)$.

Second, station h needs to be sufficiently far away from station j, so that no customer bikes the route hj or jh. Formally, we need $d(h, j) \geq D$ where D is a large distance threshold. We discuss the choice of D at the end of this section. If no one bikes the route hj or jh, then route ij and hi (or ih) are *not* substitutable for any customer–i.e., route ij and route hi (or ih) are potential choices of customers who are interested in traveling from and to different locations. In other words, route ij and route hi (or ih) can be considered as products in different markets in standard demand estimation. As a result, the exogenous product characteristic S_h is unlikely to be correlated with the unobserved product characteristic ξ_{ij}. In other words,

$$\mathbb{E}[\xi_{ij}|S_h] = 0, \tag{4.5}$$

for all $h \in \mathcal{H}_1(ij)$. This is the *exclusion restriction.*

If Equations (4.4) and (4.5) are satisfied, $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h$ is a valid instrument for the bike availability at i. Formally, we can write $\mathcal{H}_1(ij) := \{h \in H : d(h, j) \geq D, q_{ih} + q_{hi} > 0\}$, where H denotes the set of all stations. Note that we require the sum of q_{ih} and q_{hi} to be positive only–i.e., we do not require both q_{ih} and q_{hi} to be positive. As long as there are connections between the two stations, Equation (4.4) is satisfied. We use the same criterion to find instruments for the dock availability at station j–i.e., $\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} S_h$ for some station characteristics S, where $\mathcal{H}_2(ij) := \{h \in H : d(h, i) \geq D, q_{hj} + q_{jh} > 0\}$.

Next, we discuss the particular choice of the exogenous characteristics S. In practice, we find that the average station characteristics of h, $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h$, can be very weakly correlated with the bike availability at station i. The main reason is that the traffic goes in and out of station i at the same time. As a result, the average correlation between $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h$ and the availability at i is close to zero. Similarly, the correlation between $\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} S_h$ and $dock\_avail_j$ is very small in absolute value. To solve this problem, we utilize the direction of the traffic during rush hours. As shown in Figures 2.1 and 2.2, the traffic during rush hours is by and large unidirectional. During the morning rush hour, the traffic goes from the outer part

of central London to the less residential city center. During the evening rush hour, the opposite pattern is observed. We rely on the directions of traffic and construct directional instrumental variables to ensure a stronger correlation between our instruments and the availability measures. Take route ij in the evening rush hour as an example. We use the interaction between the number of total Google place counts around station h and the population density around station i as one of our instruments. During the evening rush hours, commuters travel from their places of work to their homes. Then, the interaction between the total number of Google places around h and the population density around i is positively correlated with the bike availability at i. Similarly, the interaction between Google place counts around i and the population density around h is negatively correlated with the bike availability at i. In other words, for the bike availability at station i, our main instrumental variables are $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} GooglePlaces_h * PopDensity_i$, and $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} PopDensity_h * GooglePlaces_i$. We follow the same logic to construct the instruments for the dock availability at the ending station j: namely, $\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} PopDensity_h * GooglePlaces_j$ and $\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} GooglePlaces_h * PopDensity_j$. Recall that $\mathcal{H}_2(ij)$ is the set of stations that satisfy the "opposite" conditions as $\mathcal{H}_1(ij)$–i.e., they are far away from station i but connected to station j. Recall that earlier in this subsection we have established that 1) $\mathbb{E}[\xi_{ij}|S_i] = 0$ and $\mathbb{E}[\xi_{ij}|S_j] = 0$, by the exogeneity of S_i and S_j; and 2) $\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h] = 0$ and $\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} S_h] = 0$, where S is $GooglePlaces$ or $PopDensity$. With these two conditions, we have that the exclusive restriction holds for the interaction instruments:

$$\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h S_i'] = 0,$$

$$\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} S_h S_j'] = 0,$$

where $S, S' \in \{GooglePlaces, PopDensity\}$, and $S \neq S'$. For example,

$$\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} GooglePlaces_h * PopDensity_i] = 0.$$

The interaction instruments also satisfy the relevance condition. Therefore, the interaction terms are another set of valid instrumental variables for the availability measures.

To complete the definition of our instruments, we need to specify a radius R for calculating $GooglePlaces$ and $PopDensity$ of the instrument station h, and the focal stations i and j. To be precise, $PopDensity_h^R$ is the population density integrated over a disk with radius R centered at station h, and $GooglePlaces_h^R$ is the total number of Google places within R meters from station h.

We next discuss the choice of the two hyperparameters R and D. We set $D = 7000m$, which is the 94% quantile of the distribution of route distance for routes with positive trip count in the data. It is a reasonable choice for two reasons. First, it is long enough to ensure the exclusive restriction is satisfied. Moreover, it is not too long, which avoids situations in which $\mathcal{H}_1^D(ij)$ or $\mathcal{H}_2^D(ij)$ has none or very few stations for many routes ij. For the choice of R, we use a range of values. In our main results, we choose $R = 600, 800, 1000m$. Our main results are robust to perturbing the two hyperparameters D and R. The robustness check results are presented in the Online Appendix.

So far, our proposed set of instruments are:

$$
\begin{aligned}
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} GooglePlaces_h^R, \quad \frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} GooglePlaces_h^R * PopDensity_i^R,\\
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} PopDensity_h^R, \quad \frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} PopDensity_h^R * GooglePlaces_i^R,\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} GooglePlaces_h^R, \quad \frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} GooglePlaces_h^R * PopDensity_j^R,\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} PopDensity_h^R, \quad \frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} PopDensity_h^R * GooglePlaces_j^R.
\end{aligned}
\tag{4.6}
$$

In addition to this set of instruments, we also utilize the elevation characteristics which are directional by themselves to construct additional instruments. For example, if the route hi is difficult to bike because it requires heavy climbing, the bike availability at station i is more likely to be low. This is the relevance condition. The exclusive condition can be verified using the same argument as for the interaction instruments proposed before–i.e., if h is far away from j, the elevation characteristics of routes hi and ih should be uncorrelated with ξ_{ij}. As a result, $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{hi}$,

$\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{ih}$, and $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{hi}$ can be used as instruments for the bike availability at the starting station i of route ij.[13] Similarly, we construct three additional instruments for the dock availability at ending station j. We have six additional instruments given by:

$$
\begin{aligned}
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{hi}, \quad \frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} AvgAscendGrade_{hj},\\
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{ih}, \quad \frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} AvgAscendGrade_{jh},\\
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{hi}, \quad \frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} AscendPercentage_{hj}.
\end{aligned}
\tag{4.7}
$$

Equations (4.6) and (4.7) specify the complete set of instruments. We use the exact same set of instruments in Equations (4.6) and (4.7) for both evening and morning rush hour. Although the validity of the instruments follow the same reasoning, the correlation between our instruments specified in Equation (4.6) and the availability measures is reversed. For example, $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} GooglePlaces_h^R * PopDensity_i^R$ is positively correlated with $bike\_avail_i$ in the evening rush hour, but is negatively correlated with $bike\_avail_i$ during the morning rush hour.

4.3. **Digression: reduced-form regressions.** To illustrate the choice set endogeneity problem with data and to show the validity of our instruments, we take a digression from the structural estimation in this section, and provide model-free evidence using reduced form regressions. The reduced form regression results show that ignoring the endogeneity problem leads to biased estimates and unreasonable policy recommendations, and that our proposed instrumental variable approach solves these problems.

[13] We do not use $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{ih}$, since it is the same as $1-\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{hi}$.

4.3.1. *Regression equation.* We run the regressions separately for the morning and evening rush hours. We describe the regression equation using evening rush hour[14]. The dependent variable, $log(\#trip_{ij})$, is the log of total trip count during the evening rush hour in 2014, starting from station i and ending at station j. Then,

$$log(\#trip_{ij}) = \psi_1 bike\_avail_i + \psi_2 dock\_avail_j + Z'_{ij}\phi + \epsilon_{ij}, \tag{4.8}$$

where $bike\_avail_i$ and $dock\_avail_j$ are the availability measures during evening rush hour defined as before. Z_{ij} contains exogenous characteristics for route ij. It has 1) the Google Place counts described in Section 2.2.3, within $R = 600, 800, 1000m$ around both i and j; 2) the population density within $R = 600, 800, 1000m$ from both i and j; 3) a piece-wise linear function of the biking distance capturing the first increasing and then decreasing pattern of the commuter preference for the biking distance: $\log d(i,j)$ and $\max\{\log d(i,j) - \log b, 0\}$, where the kink b is set at 1.5 km. We choose 1.5 km for two reasons. First, Figure 2.3 clearly indicates that the peak of average usage is around 1.5 km. Second, when we vary the kink point, $b = 1.5$ km explains the most variations in log trip counts. We present the results with different b values in the Online Appendix; 4) the elevation characteristics $AvgAscendGrade_{ij}$ and $AscendPercentage_{ij}$ as described in Section 2.2.3. In total, we have $1 + 10 \times 2 \times 3 + 1 \times 2 \times 3 + 2 + 2 = 71$ exogenous characteristics in Z_{ij}. We also have two endogenous covariates for which we need instrumental variables: bike availability and dock availability. Recall that, as shown in (4.6), for each value of $R = 600m, 800m, 1000m$, we have eight instrumental variables. Thus, we have $8 * 3 = 24$ instruments from (4.6). We also have six instruments from (4.7) which makes it 30 instrumental variables in total.

4.3.2. *Regression results.* In Table 2, we present the results of ordinary least squares (OLS) regressions without instrumenting availability measures and the results of the instrumental variable (IV) regressions using our proposed instruments. Table 2 includes the main coefficient estimates and their standard errors in parentheses. For the OLS regressions, we also report the adjusted R^2. For the IV regressions, we estimate Equation (4.8) using the generalized method of moments (GMM).

[14]The analysis for the morning rush hour is the same. For notational simplicity, we do not include an ER superscript for each variable in the regression equation

	Morning rush hour		Evening rush hour	
	OLS	IV	OLS	IV
O. Bike Avail.	0.05	0.69	0.13	1.37
[%]	(0.03)	(0.21)	(0.03)	(0.15)
D. Dock Avail.	0.50	1.47	-0.46	1.89
[%]	(0.04)	(0.16)	(0.03)	(0.13)
Dist. 1	0.12	0.10	0.14	0.15
[log km]	(0.02)	(0.02)	(0.03)	(0.03)
Dist. 2	-0.37	-0.36	-0.42	-0.39
[log km]	(0.03)	(0.03)	(0.03)	(0.03)
R^2	0.20	-	0.33	-
Cragg-Donald Statistic	-	83.64	-	99.37

Table 2. Reduced-form regression results
Standard errors are in parentheses. Variable units are in square brackets.

Moreover, to test the strength of the instruments, we perform the weak instrument test introduced by Stock and Yogo (2005) and report the Cragg-Donald statistics. As discussed above, in our main specification, we set $R = 600, 800, 1000m$ for the station characteristics and $D = 7000m$ for the distance threshold. We check that our results are robust to perturbing values of R and D values and details are provided in the Online Appendix.

Since the results for the morning and evening rush hour are similar, we discuss only the results for the evening rush hour, which are presented in columns 3 and 4 of Table 2. There are four main insights. First of all, in the OLS regression presented in column 3, the ending station dock availability coefficient is negative and significant. This means that commuters get higher utility if it is more difficult to find a dock available at the destination station. Moreover, the policy implication of the result is that, if the managing company or local government wanted to increase the usage of the system, they should make it harder for people to dock their bikes at the destination stations. Obviously, this cannot be the case. But if we acknowledge the endogeneity issue of the availability measures, this result is easy to explain: more popular destinations are more likely to run out of docks-hence the negative coefficient on dock availability. This is the classic "reverse causality"

problem in econometrics. On the other hand, when we run the IV regression with our proposed instruments in column 4, the negative coefficient becomes positive, and it is statistically significant. The interpretation is that higher dock availability at destination stations is preferable to commuters. Therefore, the usage would be higher, if the dock availability improved at the destination stations.

Second, we compare the coefficient on the origination station bike availability in the OLS and IV results. Both of the estimated parameters are positive. One might argue that, in this case, using the IV regression does not benefit the analysis substantially. However, the magnitudes of the two coefficients are very different. The estimated impact of having higher bike availability at the origination station on log trips in the IV regression is more than ten times as high as that in the OLS regression. This difference is expected since the "reverse causality" predicts a negative correlation between bike availability at i and usage on ij. The OLS result combines the "reverse causality" and the correct positive relationship, and, therefore, is smaller than the IV estimate. This comparison shows how important it is to take into account the endogeneity issue of the choice set, and how significant its implications are for policy recommendations derived from the estimation results.

Third, the distance coefficients do not change much between the OLS and IV results. The first distance parameter is positive, which captures the route usage increasing in route distance when the distance is below 1.5 km, as shown in Figure 2.3. The second distance parameter is negative and much bigger in magnitude than the first distance parameter. This is consistent with the pattern in Figure 2.3. The same pattern is also observed in the structural estimation result, which is detailed in Section 4.4.

Fourth, the Cragg-Donald statistics from both the morning and evening rush hour are very high. The critical value corresponding to our setting, in which there are 30 instruments and two endogenous covariates, is 20.86 (for the relative bias level of 0.05, see Stock and Yogo (2005) for details of the calculation). Our statistics, 83.64 and 99.37, are more than four times bigger than the critical value. Therefore, we can strongly reject the weak instrument null hypothesis.

4.4. **Structural estimation.** In this section, we discuss the structural estimation procedure and present the results. We use the generalized method of moments with the proposed instruments to

recover the parameters in Equations (3.3) and (3.4). However, estimating the parameters in the structural model creates computational challenges. The methods introduced in Berry *et al.* (1995) can accommodate non-linear parameters, but they are computationally feasible for only relatively low number of products or services in the choice set (Su and Judd, 2012). Since there are over 500,000 routes in the data, the estimation is computationally infeasible, even using the MPEC algorithm developed by Su and Judd (2012).

To make the estimation feasible, we divide the coverage area into blocks and estimate the model on routes between station blocks instead of routes between stations. Similar methods have been used in the literature studying demand and supply of taxi cabs (Buchholz, 2016; Frechette *et al.*, 2016) and demand predictions of the bike share system in New York (Singhvi *et al.*, 2015). In the block model, customers choose the commuting routes between station blocks instead of the specific route between stations. This reduces the computational burden substantially and makes the estimation possible. Of course, it imposes restrictions on the model and, therefore, on consumer behavior captured by the model. We argue that the modification is reasonable for the following reasons. First, it preserves the average substitution patterns across routes between different blocks. Since the counterfactual predictions we are interested are not the exact longitude and latitude of the location of the stations to be built, but which neighborhoods more stations should be added to or which new areas the network should expand into, understanding the average usage and substitution patterns across blocks is sufficient. Second, only three percent of the total trips in the data are between two stations within a block. As a result, we exclude very few data points by estimating the model at the block level. In other words, summarizing trips across blocks provides a good approximation of the general usage patterns in the network. We also provide additional evidence using simulations in the Online Appendix.

Next, we discuss the details of the block model. We divide the coverage area into uniform $1000m \times 1000m$ blocks. The stations within each block are treated as a single representative station. The location of this representative station is defined as the center of the stations in that block and used as the location of the station block. We denote the starting station block by capital letter I and the ending station block by J. The availability measures of a station block, ba_I or da_J, are

calculated as the average availability measures for all stations within that block. The route distance of IJ is defined as the average distance across all routes from any station in I to any station in J. The other covariates in the utility function are defined similarly. Compared with the route-level model, we include two additional covariates in the utility function, $\log SC_I$ and $\log SC_J$, which are the log total number of station counts in block I and J. These covariates capture the variation of the number of route options across station blocks.

The origination and destination locations, which commuters are interested in traveling from and to, are modeled as points of a grid. We divide the coverage area into $200m \times 200m$ squares and take the center point of each square as potential origination and destination locations of customers. In total there are 3263 such locations and therefore $3263^2 - 3263$ possible origination-destination pairs kl considered in our model. For commuters traveling from location k to location l, walking distance $d(k, I)$ is defined as the average distance between location k and all stations $i \in I$. $d(J, l)$ is similarly defined. We define C_{kl}, the choice set of commuters kl, as any routes starting from the four closest station blocks to k based on $d(k, I)$, to the four closest station blocks to l based on $d(J, l)$. There are thus 16 block-level routes in C_{kl} for customer kl. Since we do not observe the walking distance directly from the data, we also conduct the estimation using the same model without walking distance. Our conclusions do not change.[15]

The utility of commuter kl choosing route IJ is given by

$$U_{IJkl} = X'_{IJ}\beta + X'_{IJkl}\gamma + \xi_{IJ} + \varepsilon_{IJkl}, \quad \forall IJ \in C_{kl}. \tag{4.9}$$

X_{IJ} includes two sets of covariates. First, it includes a set of variables similar to the ij-level covarites: an intercept, the two endogenous covariates ba_I, da_J, the distance characteristics $\log d(I, J)$ and $\max\{\log d(I, J) - \log b, 0\}$, and the elevation features $AvgAscendGrade_{IJ}$ and $AscendPercentage_{IJ}$. Second, it also includes the log station count for both the starting and the ending clusters, $\log SC_I$, $\log SC_J$. We define X_{IJ} separately from the rest of the covariates because they depends only on IJ but not kl. In other words, β is the vector of linear parameters. X_{IJkl} includes the walking distance

[15] We provide the details of this robustness check in the Online Appendix.

variables $d(k, I)$ and $d(J, l)$. Since X_{IJkl} depends on both IJ and kl, γ is the vector of nonlinear parameters.

The choice probabilities are calculated similarly to Equation (3.3) in the ij-level model:

$$P_{IJkl}(X, \beta, \gamma, \theta) = \frac{exp(X'_{IJ}\beta + X'_{IJkl}\gamma + \xi_{IJ})}{exp(X'_{0kl}\theta) + \sum_{I'J' \in C_{kl}} exp(X'_{I'J'}\beta + X'_{I'J'kl}\gamma + \xi_{I'J'})}. \tag{4.10}$$

Similar to Equations (3.4), the model predicted total number of trips on route IJ is:

$$q_{IJ}(X, W, \alpha, \beta, \gamma, \theta) = \int P_{IJkl}(X, \beta, \gamma, \theta) dD(W_{kl}, \alpha). \tag{4.11}$$

We follow the MPEC algorithm (Su and Judd, 2012) and minimize the GMM loss function while matching the observed usage on the block-level routes with the predicted $q_{IJ}(X, W, \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\theta})$ to recover the parameters α, β, γ, and θ. We have 112 station blocks in total and thus 112^2 potential routes.[16] We use the same set of instruments for the availability measures as in the reduced-form regressions, specified by Equations (4.6) and (4.7).

We present the structural estimation results for the morning and evening rush hours in Table 3 and 4, respectively. We start the interpretations of the results by the linear utility parameters. First, in both the morning and evening rush hour results, the availability measures have the expected signs. The higher the average availability, the higher utility the commuters gain from biking the route. Moreover, on average, the commuters seem to care more about the bike and dock availability during the morning rush hour than the evening rush hour. This is consistent with the fact that commuters have a tighter schedule in the morning than in the evening, and, therefore, value availability more on the way to work in the morning than after work in the evening. Second, the biking distance parameter is positive when the route length is less than 1.5 km, and negative when the route length exceeds 1.5 km. This is consistent with both the model-free preliminary evidence directly observed in the data and the reduced form regression results. The coefficients of both origination and

[16]The dimension of the product space is reduced from around 761^2 in the route-level model to around 112^2 in the current model.

Non-linear para.		Linear para.	
O. Pop. Density	0.021	Intercept	-6.941
[Ppl per hec]	(0.015)		(1.211)
D. Pop. Density	-0.038	S. Station Count	0.899
	(0.019)	[In log]	(0.055)
O. Google Plc.	0.019	E. Station Count	0.756
[Num per 4 hec]	(0.016)		(0.031)
D. Google Plc.	0.286	S. Bike Avail.	1.730
	(0.107)	[%]	(0.263)
O. Walking Dist.	-3.145	E. Dock Avail.	1.790
[1km]	(0.764)		(0.206)
D. Walking Dist.	-0.927	Route Dist. 1	1.610
	(0.514)	[In log]	(0.267)
O.D. Driving Dist.	0.254	Route Dist. 2	-2.953
[1km]	(0.052)		(0.362)
O.D. Transit Dist.	0.039	Avg. ascend grade	-5.549
	(0.056)	$(\tan \alpha)$	(2.953)
		% segments ascending	-1.253
			(0.196)
Pseudo-R^2	0.667		

Table 3. Demand estimates: morning rush hour
Standard errors are in parentheses. Variable units are in square brackets.

destination station counts are positive and statistically significant. This implies that commuters get higher utility when there are more station options in an area. Third, the average ascending grade and the ascending percentage both have negative coefficients and are mostly significant. This implies that the routes with higher ascending grade or higher share of ascending segments are less attractive to commuters which is consistent with our intuition.

Next, we discuss the non-linear parameters in the density model and the utility of the outside option. First, all walking-distance coefficients are negative and significant, except for the walking distance from the ending station block to the destination location in the morning rush hour. This implies that commuters prefer docking stations closer to their origination and destination locations of interest. However, when we compare the pseudo-R^2 of the results to those where we exclude walking distance from the model, the difference is less than .01. In other words, including walking

Non-linear para.		Linear para.	
O. Pop. Density	-0.045	Intercept	-3.494
[Ppl per hec]	(0.004)		(0.586)
D. Pop. Density	0.005	S. Station Count	0.897
	(0.007)	[In log]	(0.024)
O. Google Plc.	0.182	E. Station Count	0.924
[Num per 4 hec]	(0.024)		(0.052)
D. Google Plc.	0.037	S. Bike Avail.	0.978
	(0.011)	[%]	(0.189)
O. Walking Dist.	-2.036	E. Dock Avail.	0.822
[1km]	(0.288)		(0.282)
D. Walking Dist.	-4.174	Route Dist. 1	2.096
	(0.724)	[In log]	(0.201)
O.D. Driving Dist.	0.151	Route Dist. 2	-4.083
[1km]	(0.031)		(0.267)
O.D. Transit Dist.	0.075	Avg. ascent grade	-7.375
	(0.032)	[$\tan\alpha$]	(2.694)
		% segments ascend	-1.223
			(0.149)
Pseudo-R^2	0.792		

Table 4. Demand estimates: evening rush hour
Standard errors are in parentheses. Variable units are in square brackets.

distance does not improve the explanatory power of the model.[17] Second, for the morning rush hour, the population density coefficient is much bigger at the origination location k than at the destination location l. The reverse is true for the Google place count coefficients. This indicates that the commuters are more likely to travel from the residential areas of the city (where they live) to the city center (where they work). The direction of the traffic is consistent with what we observe directly from the data in Figure 2.1. We observe the opposite pattern in the magnitudes of those coefficients for the evening rush hour. This result shows that our model captures the directions of the traffic nicely, which would have been impossible to achieve had we treated stations as independent instead of as nodes linked to each other on a network. Third, the driving distance and transit distance coefficients indicate that commuters find biking a less attractive option when the trip is longer.

[17]The detailed results are presented in the Online Appendix.

Finally, the pseudo-R^2 of both the morning and evening rush hour results, 0.669 and 0.793, respectively, shows that our model fits the data well. Moreover, as shown in the Online Appendix, the pseudo-R^2 remains high with different model specifications. This suggests that our model fit is robust to model specifications.

5. Counterfactuals

Using the estimated model, we conduct three counterfactual experiments related to network design and expansion. In the first counterfactual, we evaluate a specific plan to expand the network into the Islington and Hackney areas, which the local community proposed in 2012. Using our model and estimates, we show that although the expansion benefits the local community in Islington and Hackney, the magnitude of the overall usage increase in the network is not substantial. Our analysis provides the managing company with important insights and guidance for the expansion of the network. In the second counterfactual, we generalize the insights from the first counterfactual and investigate the best locations to add stations in the current network. In particular, we compute the marginal effect of adding one station at different locations in the network. We show that adding stations to the city center leads to ten times more usage increase than adding stations to the peripheries. More importantly, we identify two types of network effects in our findings. Our results highlight the significance of network effects in understanding and evaluating the expansion of the network. In the third counterfactual, we investigate a different type of expanding strategy than in the second counterfactual. Instead of adding stations to the network, we keep the current network and study the best locations to add bikes and docks. Similar to the second counterfactual, we show that the network effects play an important role in determining the usage increase when adding bikes and docks to different locations. Moreover, by comparing the differences between adding bikes and docks, we highlight the interplay of network effects and the direction of commuting traffic in the results.

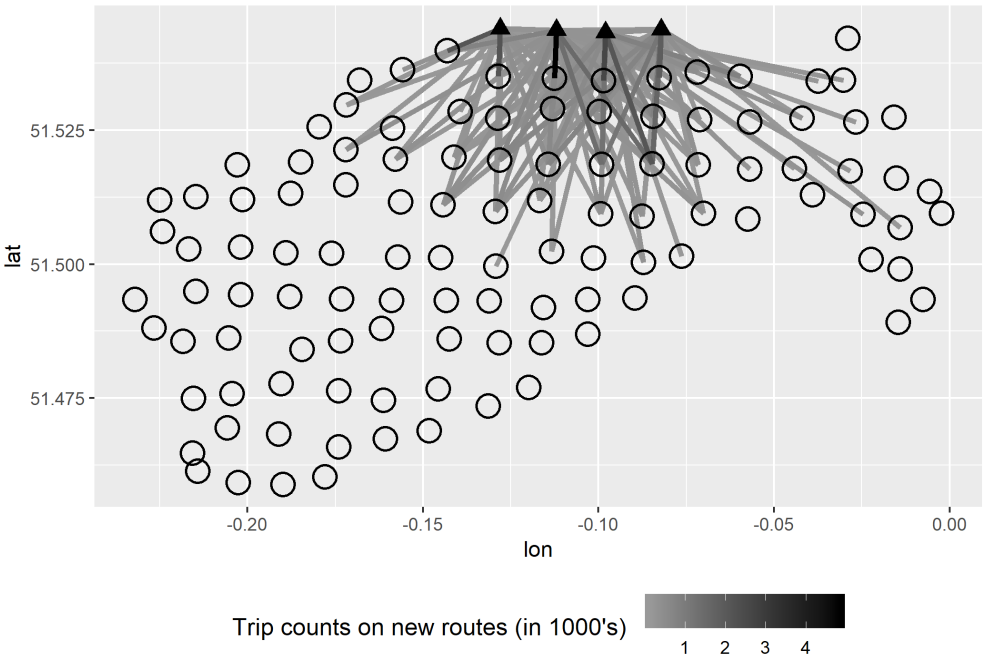

FIGURE 5.1. Predicted new station usage after expansion

5.1. **Expansion to Islington and Hackney.** Like many bike sharing programs in major cities, the "Boris Bikes" went through several major expansions since they were first introduced.[18] The city has been expanding the network continuously, [19] and there are calls and proposals from the local government and communities for further expansions. One of the many proposals is to expand the bike share system to Islington and Hackney.[20] In this counterfactual, we use our estimated model to evaluate this particular expansion proposal.

We add four new blocks of stations covering the Islington and Hackney areas. Each block contains four stations, which is the same station density as in the neighboring blocks observed in the current network. We compute the sum of predicted usage during morning and evening rush

[18]Source: https://tfl.gov.uk/info-for/media/press-releases/2013/december/mayor-launches-huge-expansion-of-flagship-barclays-cycle-hire-scheme
[19]Source: http://www.cityam.com/268035/lycra-ready-south-london-santander-cycles-scheme-expanding
[20]Source: http://www.islingtongazette.co.uk/news/environment/plea-for-boris-bikes-to-be-wheeled-out-across-islington-1-1454024
https://www.change.org/p/transport-for-london-and-islington-council-roll-out-the-london-cycle-hire-scheme-to-the-whole-of-islington

hours (throughout a year). We assume that the new station blocks have the same average availability levels to those of the closest existing blocks. The availability of the rest of the station blocks stays the same. Given we are adding a small number of new stations to the network, we think this is a reasonable assumption. The results are presented in Figure 5.1.

The links indicate the routes on which usage increases after the expansions. The shading of the link indicates the level of usage increase: the darker the color, the higher the level of usage. Most of the usage increase comes from trips between the new stations and two areas: one around the new stations, and the other close to the city center. This result is consistent with the intuition that customers are most likely to use bikes to travel to nearby areas or to commute to and from the city center.

However, the key point is that the magnitude of the usage increase is very limited. The total usage increase is about 61,000, which is 1.5% of the total number of trips on the network during the morning and evening rush hours before the expansion. Our model also shows that, if we compare the trip increase per additional station, this usage increase is less than 20% of the usage increase had the new stations been added to the city center instead. It suggests that, if maximizing the overall usage is the objective, adding stations in Islington and Hackney is not optimal. We revisit this point and provide a more general discussion about the optimal expansion location in the next counterfactual.

To summarize, although the expansion would benefit the residents in Islington and Hackney, the usage increase would be much higher if the new stations were allocated, instead, to the city center. The managing company in the city of London faces a trade-off between benefiting more people citywide versus providing service access to a particular community.

5.2. **Adding one station to the system.** In the second counterfactual, we generalize the insights from evaluating the particular expansion proposal in the first counterfactual. We analyze the optimal location to expand the network and highlight the importance of network effects in the efficacy of different types of expansions. The insights are not only important to the bike share system in London, but are also general enough to be applicable to the design of other bike share systems.

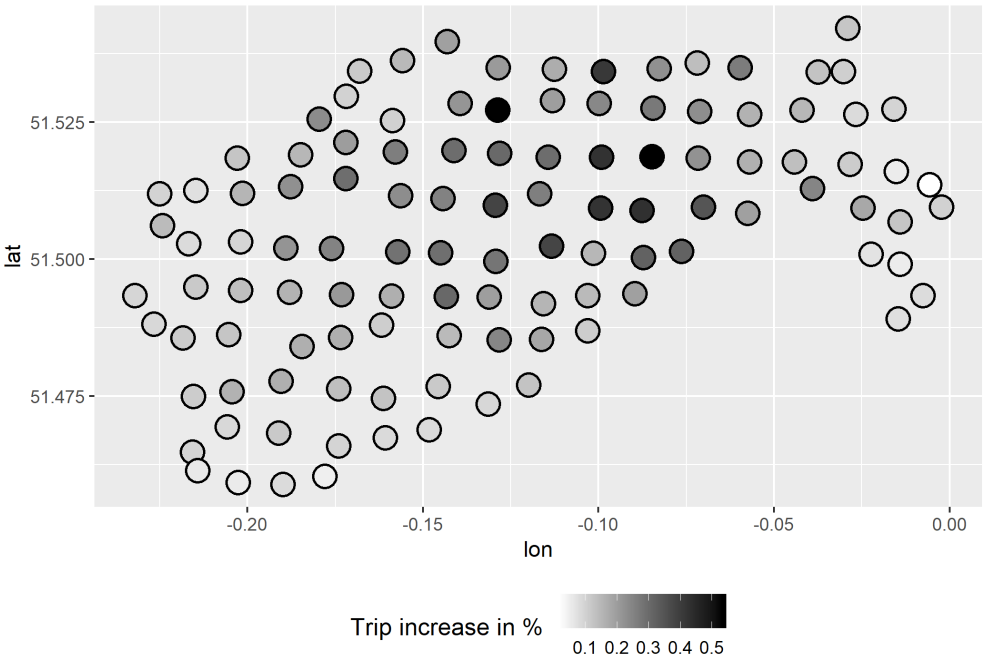

FIGURE 5.2. Predicted usage increase after adding one station

In this experiment, we add one station to different locations of the network. We investigate where the optimal location is–i.e., the location that leads to the highest usage increase in the entire network. Our analysis not only takes into account the usage increase on routes originating from and ending at the new station. It also considers the change in usage in other parts of the network induced by customers substituting between routes. In other words, we calculate the change in usage for all routes on the *entire* network, which, we show, is key to evaluating the expansion of the system.

Adding one station is equivalent to increasing the number of stations by 0.14%. We compute the percentage increase in total usage in the network depending on the block of stations to which the new station is added. Similar to the previous counterfactual, we assume the availability level of all station blocks stays the same. Give the limited impact one station has on the entire network, we think this is a reasonable assumption. The results are presented in Figure 5.2. Each dot represents one of the 112 blocks in our coverage area. The shade of the dot indicates the total usage increase

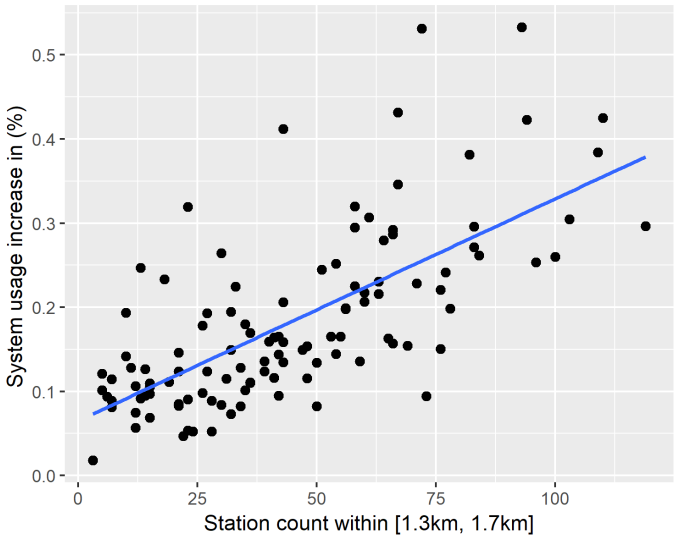

FIGURE 5.3. Predicted usage increase vs. number of stations 1.3km to 1.7km away

in the network when the extra station is added to the block. As shown in the figure, there is large heterogeneity in how effective the additional station is in increasing the overall network usage. The percentage usage increase varies from close to zero to around 0.5%. The usage increase when the station is added to the city center can be ten times as high as when the station is added to the peripheries. This shows that increasing *density in the city center* has much bigger positive effect on the overall network usage than increasing the *scope* of the network. The station density is too low in city center and too high in the peripheries. Thus, the city center is clearly the bottleneck. This finding is partially a result of the implementation of the 300m between-station rule. Our analysis suggests that the uniform density rule is much too rigid. A redistribution of stations from the peripheries of the coverage area to the city center would make the current network far more efficient, if efficiency is judged by overall usage.

We also investigate the importance of network effects in the large heterogeneity of the predicted system usage increase, depending on where the new station is added. First, in Figure 5.3, we plot the usage increase against the number of stations that are approximately 1.5 km away from the focal station block. There is a very strong positive correlation. This suggests that adding station to *central* regions in the network, which we define next, is an important determinant of the usage

increase and the effectiveness of network expansion. This motivates us to look for a precise measure of centrality that captures the network effect associated with it.

We construct a centrality measure using the number of stations approximately 1.5 km away and compare the measure to station-level usage determinants to better understand the network effect. This centrality measure is constructed by counting the number of stations between 1.3 km and 1.7 km away for each station and then averaging over all stations within each block. We regress the total usage increase if one station is added to each block (presented in Figure 5.2), on our centrality measure of the block, the average population density of the block, the average number of Google places within the block, and the total number of stations in the block. We also compare our centrality measure with the conventional eigenvector centrality measure, where the adjacency matrix is defined by one over the distance between each pair of station clusters. We regress the predicted usage increase (when adding one station to each station block) on the three block characteristics and the two centrality measures. We study how much those five factors can explain the variation across the predicted usage increase at different station blocks. We do so by comparing the adjusted R^2 in linear regressions. Notice that the estimated coefficients do not have causal interpretations. Therefore, in the results, we report only whether the coefficients are statistically significant, and the adjusted R^2. We present the results in Table 5.

POPULATION DENSITY	✓	✓	✓	✓
GOOGLE PLACES	✓***	✓***	✓***	✓**
NUMBER OF STATIONS	✓***	✓***	✓	✓***
OUR CENTRALITY	-	✓***	-	✓***
EIGENVECTOR CENTRALITY	-	-	✓***	✓
ADJUSTED R^2	0.30	0.52	0.42	0.52

TABLE 5. Determinants of usage increase and network effect

In column 1 of Table 5, we present the baseline result in which we regress the predicted usage increase (when adding one station to each block) on three block characteristics. The result shows that the station block characteristics explain 30% of the variation in predicted usage increase across

blocks. The regression in column 2 includes our constructed centrality measure as an additional explanatory variable. Comparing column 1 and column 2, we see that including our centrality measure increases the adjusted R^2 by 73%. In other words, using our measure of centrality, being central on the network is key to locating the optimal blocks for the expansion of the network. This is the first type of network effect of relevance to the station network expansion and design, which we identify using the model and the estimation. The regression in column 3 replaces our centrality measure with the conventional eigenvector centrality measure. The comparison between column 2 and column 3 shows that our centrality measure explains the usage increase almost twice as well as the conventional eigenvector centrality measure. Finally, when we include both centrality measures, the conventional measure does not explain any additional variation in the usage increase across station blocks: the adjusted R^2 is still 0.52. Moreover, across all four regressions, the adjusted R^2 is never higher than 0.52. This means that, even when we can identify the right network centrality measure ex ante, the richness of the structural model helps us predict usage increase and identify optimal expansion locations much better than simple reduced-form regressions.

Next, we illustrate a second type of network effect by conducting the following analysis. We recompute the predicted usage increase resulting from adding one station to each block with the following change. While keeping the origination station block characteristics (population density, place counts, number of stations, and bike availability) and route distance the same as in the data, we set the destination station block characteristics (population density, place counts, number of stations, and dock availability) as the median values of those characteristics. Thus, we investigate whether the kind of stations (i.e., stations with similar characteristics) to which a station is connected matters for usage and network expansions. In other words, we examine whether a good match of origination and destination stations matters for usage. Specifically, by setting the destination station characteristics to the median level, we break the match in calculating the predicted usage increase. The difference between the result of this calculation and the original usage increase computed above is the network effect that comes from a good match between an origination station and a destination station.

We find that, when setting the destination station characteristics as the median value, the usage increase goes down by 17%, on average, compared to the original usage increase. This indicates that not only the number of connecting stations approximately 1.5 km away matters, but that the *kind* of connected stations also matters for usage and network design. This is another type of network effect that we find relevant for evaluating network expansions in our analysis. In other words, treating stations as individual products and studying the demand for each on its own is not enough. Including the entire network of stations in the analysis is crucial to understanding customer demand and evaluating system expansion strategies.

5.3. **Improving long-term availability in the system.** Adding stations to the network is one way of expanding the bike share system. Another way of expanding the network is to add bikes or docks to the existing stations in the network. This type of expansion can be captured by improving the average bike and dock availability in our model. In this counterfactual, we analyze the optimal locations for the operating company to add bikes and docks. This is similar to the long-term effect of improving availability computed in Kabra *et al.* (2016), but our focus is on the optimal area to improve the availability measures in terms of long-term system usage. Moreover, we study the effects of improving both bike availability and dock availability, while Kabra *et al.* (2016) only considers bike availability. We present the results for the evening rush hour in this section. The analysis of the morning rush hour is done exactly the same way, and the results do not change qualitatively. Therefore, we leave the results for the morning rush hour to the Online Appendix.

In this experiment, we first calculate the improvement of system usage in the evening rush hour by adding 0.05 to the average bike availability of each cluster–that is, the predicted usage increase of the entire system if the operating company makes the probability of finding an available bike for each cluster 5 percentage points higher during the evening rush hour. Similarly to the previous counterfactual, we present in Figure 5.4 the predicted system usage increase in percentage when adding 0.05 to the average bike availability of every cluster, and in Figure 5.5, the improvement when adding dock availability. Again, each dot represents one of the 112 blocks in our coverage area, and the shading of the dot indicates the total percentage usage increase in the network. We

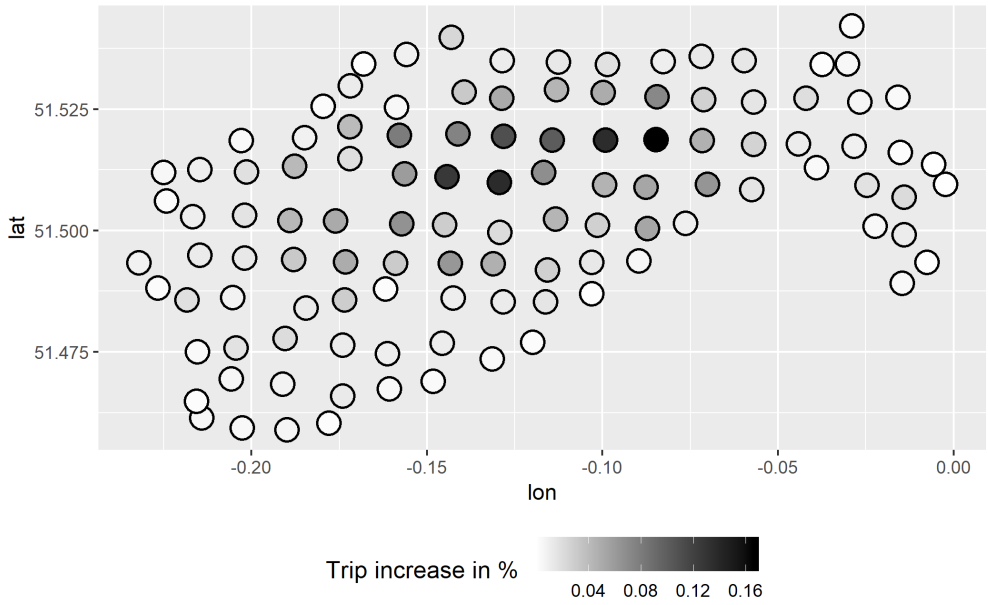

FIGURE 5.4. Predicted evening rush hour usage increase after improving bike availability

can see that there is also a wide gap in predicted trip increase between the most and least beneficial areas in which to boost availability during the evening rush hour.

The results also illustrate the interplay of the unidirectional travel pattern and the network effect. For the evening rush hour, we can see that people generally move from the city center to more residential areas and, therefore, the best places to improve bike availability are all in the very center of the network. However, for dock availability, based on the direction of commuting flow, we should focus on residential stations around the peripheries since, on average, there will be more travelers going to those stations. However, as a destination, those stations in the peripheries tend to have fewer connecting stations that people can bike from–i.e., from the network effect point of view, they are not the perfect choice. The two contrasting factors result in the best clusters for boosting dock availability. Figure 5.5 shows that the highest usage increase stations are a little more scattered around the map, not as concentrated in the very center of the network as in Figure 5.4.

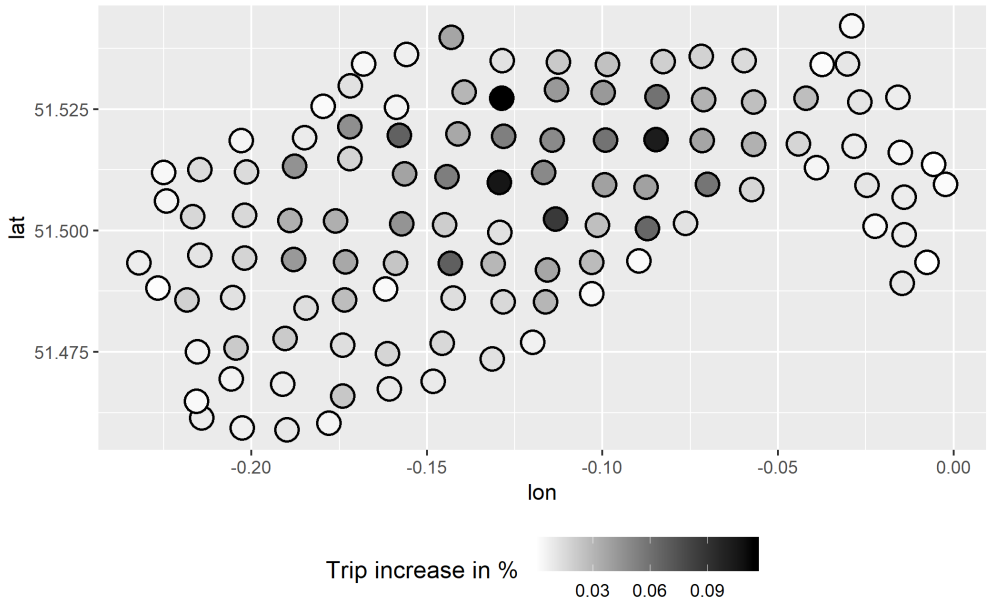

FIGURE 5.5. Predicted evening rush hour usage increase after improving dock availability

6. CONCLUSION

We study customer demand in the station network of the London bike share system. Using our model and the estimated preference parameters, we provide guidance on the network design and expansion of the system. We highlight the tradeoff between the density and scope of a network in increasing overall usage. We also evaluate a particular proposal of expansion to the Islington and Hackney areas of the city. Our empirical results provide important insights and policy recommendations for the managing company and the local government.

In the estimation of the structural demand model, we develop an instrumental variable approach to deal with the endogeneity problem of the choice set. We show that properly taking into account the endogeneity of choice set in the demand estimation is important for both the estimated parameters and the policy recommendations derived from the estimated model. The method can

be applied to other empirical settings in which understanding customer demand is important and products and services can go out of stock.

References

Andrews, Isaiah, Gentzkow, Matthew, and Shapiro, Jesse M. 2017. Measuring the Sensitivity of Parameter Estimates to Estimation Moments. *The Quarterly Journal of Economics*, 1553–1592.

Ashok, Kalidas, and Ben-Akiva, Moshe E. 2000. Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows. *Transportation Science*, **34**(1), 21–36.

Barcelö, Jaume, Montero, Lidin, Marqués, Laura, and Carmona, Carlos. 2010. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation research record*, **2175**(1), 19–27.

Ben-Akiva, Moshe, Bierlaire, Michel, Burton, Didier, Koutsopoulos, Haris N, and Mishalani, Rabi. 2001. Network state estimation and prediction for real-time traffic management. *Networks and spatial economics*, **1**(3-4), 293–318.

Berry, Steve, Linton, Oliver B, and Pakes, Ariel. 2004. Limit theorems for estimating the parameters of differentiated product demand systems. *The Review of Economic Studies*, **71**(3), 613–654.

Berry, Steven, Levinsohn, James, and Pakes, Ariel. 1995. Automobile Prices in Market Equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Berry, Steven T., and Haile, Philip A. 2014. Identification in Differentiated Products Markets Using Market Level Data. *Econometrica*, **82**(5), 1749–1797.

Bramoullé, Yann, Djebbari, Habiba, and Fortin, Bernard. 2009. Identification of peer effects through social networks. *Journal of econometrics*, **150**(1), 41–55.

Bruno, Hernan A, and Vilcassim, Naufel J. 2008. Research note: structural demand estimation with varying product availability. *Marketing Science*, **27**(6), 1126–1131.

Buchholz, Nicholas. 2016. Spatial equilibrium, search frictions and efficient regulation in the taxi industry. *Working Paper*.

Conlon, Christopher T, and Mortimer, Julie Holland. 2013. Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics*, **5**(4), 1–30.

DeMaio, Paul. 2009. Bike-sharing: History, impacts, models of provision, and future. *Journal of public transportation*, **12**(4), 3.

Ellickson, Paul B, and Misra, Sanjog. 2008. Supermarket pricing strategies. *Marketing science*, **27**(5), 811–828.

Frechette, Guillaume R, Lizzeri, Alessandro, and Salz, Tobias. 2016. Frictions in a Competitive, Regulated Market Evidence from Taxis. *Working Paper*.

Freund, Daniel, Henderson, Shane G, and Shmoys, David B. 2017. Minimizing Multimodular Functions and Allocating Capacity in Bike-Sharing Systems. *Pages 186–198 of: International Conference on Integer Programming and Combinatorial Optimization*. Springer.

Greater London Authority. 2013. The mayor's vision for cycling in London: an olympic legacy for all Londoners. *Greater London Authority, London*.

Jorge, Diana, and Correia, Gonçalo. 2013. Carsharing systems demand estimation and defined operations: a literature review. *European Journal of Transport and Infrastructure Research*, **13**(3).

Kabra, Ashish, Belavina, Elena, and Girotra, Karan. 2016. Bike-share systems: Accessibility and availability. *Working Paper*.

Manchanda, Puneet, Rossi, Peter E, and Chintagunta, Pradeep K. 2004. Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research*, **41**(4), 467–478.

Midgley, Peter. 2011. Bicycle-sharing schemes: enhancing sustainable mobility in urban areas. *United Nations, Department of Economic and Social Affairs*, 1–12.

Musalem, Andrés, Olivares, Marcelo, Bradlow, Eric T, Terwiesch, Christian, and Corsten, Daniel. 2010. Structural estimation of the effect of out-of-stocks. *Management Science*, **56**(7), 1180–1197.

New York City Department of City Planning. 2009. *Bike-Share: Opportunities in New York City*. Tech. rept. City of New York.

O'Mahony, Eoin, and Shmoys, David B. 2015. Data Analysis and Optimization for (Citi) Bike Sharing. *Pages 687–694 of: AAAI*.

Singhvi, Divya, Singhvi, Somya, Frazier, Peter I, Henderson, Shane G, O'Mahony, Eoin, Shmoys, David B, and Woodard, Dawn B. 2015. Predicting Bike Usage for New York City's Bike Sharing System. *In: AAAI Workshop: Computational Sustainability*.

Stock, James H, and Yogo, Motohiro. 2005. Testing for Weak Instruments in Linear IV Regression. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80.

Su, Che-Lin, and Judd, Kenneth L. 2012. Constrained optimization approaches to estimation of structural models. *Econometrica*, **80**(5), 2213–2230.

Toole, Jameson L, Colak, Serdar, Sturt, Bradley, Alexander, Lauren P, Evsukoff, Alexandre, and González, Marta C. 2015. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, **58**, 162–177.

Transport for London. 2010. *Cycling Revolution London*. Tech. rept. Mayor of London.

Woodcock, James, Tainio, Marko, Cheshire, James, O'Brien, Oliver, and Goodman, Anna. 2014. Health effects of the London bicycle sharing system: health impact modelling study. *Bmj*, **348**, g425.