CUSTOMER PREFERENCE AND STATION NETWORK IN THE LONDON BIKE SHARE SYSTEM

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 increasing system usage, and that the currently implemented station density rule is far from ideal. 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.

Keywords: sharing economy, bike-sharing system, network effects, endogenous choice set, commuter preference
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 is 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 300m density rule, which states that two neighboring docking stations should be, at most, 300m 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 has 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 upper 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 counterfactual expansion and design experiments of the network.

To model customer demand, the natural way to go is to treat stations as products and to assume that customers choose the bike 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, with a focus on the spatial network structure of supply and demand (Bimpikis et al., 2016).

The main 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. It 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 products 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.\(^1\) The system currently has around 800 stations, and covers most of central London. 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). We plot the station locations in Figure 2.1 on the map of the Greater London administrative area, with its 33 boroughs. It shows that the bike share system covers only the central boroughs of the city. 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 present the distribution of the trips by route in row 1 of Table 1, where a route is defined as a directional link between two stations. The statistics illustrate that there is huge variation in usage across routes. Some routes have thousands of trips, while others have only one trip. 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. To demonstrate this pattern in more detail, in Figure 2.2 we plot the histogram of trip count for those routes with a trip count ranging between zero to 500 throughout 2014. It clearly shows the long tail pattern in the data. In the subsequent analysis, we follow the standard practice of including only routes with positive usage. In the structural estimation, in particular, we aggregate routes across nearby

\(^2\)Source: https://www.theguardian.com/environment/bike-blog/2016/apr/14/london-mayoral-election-qa-on-cycling-policy-with-the-main-candidates
stations, which alleviates a lot the long tail issue. For those routes with positive usage, we also present summary statistics of route distance measured in meters, using the shortest distance between two stations, in row 2 of Table 1. The statistics show that, while there is large variation in route distance, most routes with any positive usage are less than five kilometers.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trip counts by route</strong></td>
<td>27</td>
<td>76</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>25</td>
<td>6707</td>
</tr>
<tr>
<td><strong>Route distance (m)</strong></td>
<td>5403</td>
<td>3052</td>
<td>27</td>
<td>2341</td>
<td>3620</td>
<td>5130</td>
<td>15998</td>
</tr>
</tbody>
</table>

**Table 1.** Route Level Usage and Distance Distribution

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. \( \text{bike\_avail} = \frac{\text{num of 5-min intervals with } \geq 5 \text{ bikes}}{\text{Total num of intervals}} \); and, similarly, we define the dock availability measure as \( \text{dock\_avail} = \frac{\text{num of 5-min intervals with } \geq 5 \text{ docks}}{\text{Total num of intervals}} \). Similar to Kabra et al. (2016), we use the five bike threshold to allow for the possibility of broken bikes and docks. We present the summary statistics in Table 2. For all four time windows and availability combinations, there is substantial variation in bike and dock availability across stations.
This shows that customers can face a different set of stations with bikes or docks available at different locations, which is important to take into account in the subsequent analysis.

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

Table 2. Summary Statistics for Long-term Availability Measure

2.2.3. Google Places. Third, we collect Google Places data in London through Google Place API. This dataset provides the longitude and latitude coordinates of 97 categories of places that are identifiable on Google Maps, including subway stations, government office buildings, schools, restaurants, etc.\(^3\) 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 are in the online appendix.

We present summary statistics of the ten categories of Google places in Table 3. 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. We present the summary statistics for each category of places and for the sum of all places. Since many categories have more than 40% zeros, it can be more informative to use the total counts instead of the counts for each category. Therefore, we calculate the correlation between each pair of category counts, including the total counts, and we present the correlation matrix in Table 4. We see that the average pairwise category correlation among the ten categories of Google places is only 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.

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.\(^4\) 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. We present in Tables 5 and 6 the summary statistics of population density for both the 430 covered LSOAs and all 4835 LSOAs.

\(^3\)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.

\(^4\)The data were downloaded from the local government’s website, https://data.london.gov.uk/dataset/lsoa-atlas
### Table 3. Summary for Google Places within Uniform 200m by 200m Squares

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Count</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>30</td>
<td>51</td>
<td>0</td>
<td>4</td>
<td>12</td>
<td>33</td>
<td>799</td>
<td></td>
</tr>
<tr>
<td>Food</td>
<td>4</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>5</td>
<td>17</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>641</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entertainment</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Store</td>
<td>5</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finance</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Gov Offices</td>
<td>31</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>680</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transportation</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>6</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>193</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4. Correlation Matrix for Google Places including Total Place Counts

<table>
<thead>
<tr>
<th>Covered LSOAs</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Ages Num</td>
<td>1702</td>
<td>330</td>
<td>985</td>
<td>1480</td>
<td>1664</td>
<td>1892</td>
<td>3081</td>
</tr>
<tr>
<td>Working Ages Num</td>
<td>1294</td>
<td>289</td>
<td>608</td>
<td>1083</td>
<td>1250</td>
<td>1482</td>
<td>2400</td>
</tr>
<tr>
<td>All Ages Den Numer/10^4 m^2</td>
<td>143</td>
<td>67</td>
<td>6</td>
<td>98</td>
<td>139</td>
<td>182</td>
<td>399</td>
</tr>
<tr>
<td>Working Age Den Numer/10^4 m^2</td>
<td>108</td>
<td>50</td>
<td>5</td>
<td>74</td>
<td>105</td>
<td>139</td>
<td>295</td>
</tr>
</tbody>
</table>

### Table 5. Census Data Summary For 430 Covered LSOAs

<table>
<thead>
<tr>
<th>All LSOAs</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Ages Num</td>
<td>1691</td>
<td>264</td>
<td>985</td>
<td>1530</td>
<td>1654</td>
<td>1817</td>
<td>4933</td>
</tr>
<tr>
<td>Working Ages Num</td>
<td>1107</td>
<td>229</td>
<td>608</td>
<td>1010</td>
<td>1128</td>
<td>1281</td>
<td>4235</td>
</tr>
<tr>
<td>All Ages Den Numer/10^4 m^2</td>
<td>96</td>
<td>61</td>
<td>1</td>
<td>52</td>
<td>83</td>
<td>128</td>
<td>685</td>
</tr>
<tr>
<td>Working Age Den Numer/10^4 m^2</td>
<td>68</td>
<td>46</td>
<td>1</td>
<td>34</td>
<td>57</td>
<td>92</td>
<td>440</td>
</tr>
</tbody>
</table>

### Table 6. Census Data Summary For All 4835 LSOAs

We can see from the Area rows that the LSOAs in the coverage areas are generally smaller than the average LSOA. In terms of population density, the LSOAs in the coverage area are denser than the average LSOA. But
comparing the Max columns, we see that, currently, the station network does miss some extremely dense LSOAs, which implies that there would likely be substantial demand increase if the system were expanded.

2.3. **Preliminary evidence.** To motivate our analysis, we present several pieces of preliminary evidence about the usage of the system before going into detail 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. Table 7 provides the trip counts for different time windows. It shows that the majority of the trips happen during rush hours and that the system is heavily commuter dominant. Moreover, the local government clearly makes the commuters the main beneficiaries when discussing plans for expanding the network (Greater London Authority, 2013). For those two reasons, we from now on restrict our analysis to rush hour usage.

<table>
<thead>
<tr>
<th>Number of Trips (in millions)</th>
<th>Share of Total Trips</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Morning Rush Hours (4 hrs)</strong></td>
<td>1.74</td>
</tr>
<tr>
<td><strong>Evening Rush Hour (4 hrs)</strong></td>
<td>2.24</td>
</tr>
<tr>
<td><strong>Rest of the Day (16 hrs)</strong></td>
<td>2.67</td>
</tr>
</tbody>
</table>

**Table 7.** Weekday Trip Counts by Time Window

Next, we provide some evidence on the spatial pattern of the system and its usage. We plot the rush hour trip data in Figures 2.3 and 2.4. Figure 2.3 shows the usage pattern at 9:30am on a weekday morning in London. Each circle represents a bike docking station. The size of the circle indicates the size of the station—i.e., the maximum number of bikes that can be docked at the station. The color of the circle corresponds to the number of bikes available at the station, ranging from black, which indicates that the station is full of bikes, to white, which indicates that the station is out of bikes. Figure 2.3 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 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 2.4. This is a snapshot of the system around 8pm on a weekday. Figure 2.4 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 identify the preference parameters in the structural model.
In Figures 2.3 and 2.4, 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.

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 2.5, 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. Peak usage occurs around 1km. 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 presented Section 4.3 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 choice in our structural estimation analysis. Note that this evidence alone is not sufficient to provide prescriptive the policy recommendations we are interested in. They are merely correlation patterns in the data. To 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 in the 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 demand 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

\[
U_{ijkl} = X_{ijkl}'\beta + \xi_{ij} + \varepsilon_{ijkl} \quad \forall ij \in C_{kl}
\]

\(X_{ijkl}\) is a vector of characteristics of route \(ij\) for customer traveling from \(k\) to \(l\). It includes two types of characteristics in addition to a constant. The first type includes two terms: \(\log d(i,j)\), the logarithm of the distance between stations \(i\) and \(j\), and \(\max\{\log d(i,j) - \log b, 0\}\), which captures potentially non-linear distance preferences for long distance routes. We define long-distance routes as those longer than 1km–i.e. \(b = 1km\). This is natural in light of the fact that 1km is the peak point in Figure 2.5. We provide robustness checks for other values of \(b\) in the Appendix. Distance function \(d(\cdot, \cdot)\) is defined as the shortest distance between two locations \(x\) and \(y\).

The second type 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 in this article, we include them since some 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.
\( \xi_{ij} \) are unobserved route characteristics, which can include whether route \( ij \) is bike friendly, uphill or downhill, or other features, making \( ij \) more or less attractive. \( \varepsilon_{ijkl} \) are idiosyncratic error terms that are independent and identically distributed and follow type I extreme value distribution.

Since we differentiate customers only by the origination location \( k \) and the destination location \( l \), we index customers by \( kl \). Similarly, we index biking routes by starting and ending station pair \( ij \). 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 normalize the utility of the outside option to zero. The choice probability of \( kl \) choosing \( ij \in C_{kl} \) is then

\[
P_{ijkl}(X, \beta) = \frac{\exp(X'_{ijkl}\beta + \xi_{ij})}{1 + \sum_{i'j' \in C_{kl}} \exp(X'_{i'j'kl}\beta + \xi_{i'j'})}.
\]

For \( ij \notin C_{kl} \), we have \( P_{ijkl}(X, \beta) = 0 \). Let \( q_{ij} \) be the total number of trips on route \( ij \). Then

\[
q_{ij} = \int P_{ijkl}(X, \beta) dD(W_{kl}, \alpha),
\]

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 \( W_{kl} \), 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 demand and to evaluate different network expansion strategies, we do not model the high-frequency variation of usage within the rush hour time windows. Instead, we use Equation (3.3) 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. As a result, taking morning rush hour as an example, \( q_{ij} \) in Equation (3.3) should be interpreted as average daily usage of route \( ij \) during morning rush hour, or equivalently, total number of usage in a year during morning rush hour. Similarly \( D(W_{kl}, \alpha) \) measures average daily travel demand from location \( k \) to location \( l \), or total number of travel demand from \( k \) to \( l \) throughout a year.

The core of the model is multinomial logit. It is known for its implied 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 \( \beta \) and \( \alpha \) in Equations (3.2) and (3.3). Note that both are vectors. We detail the estimation procedure in the following section. With the estimates of \( \beta \) and \( \alpha \), 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 2, 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 \( \xi \) be the matrices \( \{X_{ijkl}\}_{ij=1,...,N,kl=1,...,M'} \), \( \{W_{kl}\}_{kl=1,...,M'} \), and \( \{\xi_{ij}\}_{ij=1,...,N} \), respectively. To obtain consistent estimators of \( \alpha \) and \( \beta \), one necessary condition is that \( \xi \) is mean zero, conditioning on \( X \) and \( W \)--i.e., 
\[
E[\xi|X,W] = 0.
\]
Under this condition, one can construct moment conditions 
\[
E[\xi \cdot X|X,W] = E[\xi \cdot W|X,W] = 0.
\]
(Berry et al., 1995, 2004). One important implication of this condition is that \( \xi_{ij} \) 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 \( \xi_{ij} \) 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 \( \xi_{ij} \). 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 (3.2), we have

$$P_{ijkl}(X, \beta) = \frac{1_{ij} \cdot \left[ \exp(X'_{ijkl} \beta + \xi_{ij}) \right]}{\sum_{i'j'} 1_{i'j'} \cdot \left[ \exp(X'_{i'j'kl} \beta + \xi_{i'j'}) \right]},$$

where $1_{ij}$ 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 $1_{ij}$ is precisely observed in the data, we can treat it as a standard route characteristic. Equation (3.3) stays the same. Then, the moment condition we are actually using is

$$E[\xi|X, W, \mathbf{1}] = 0,$$

where $\mathbf{1}$ is the vector of $1_{ij}$, for all $ij$. Now, one could use Equation (4.2) to obtain estimates for $\alpha$ and $\beta$. However, the estimated coefficients will be biased since 4.2 is violated. The reason is that the unobserved product characteristics $\xi_{ij}$ are correlated with the choice set indicator $1_{ij}$—i.e., $E[\xi_{ij} \cdot 1_{ij}] \neq 0$. 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 $\xi_{ij}$. 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 $1_{ij}$ is correlated with $\xi_{ij}$. 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, $\xi_{ij}$ can also be correlated with $1_{ij}$ 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.5

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

5Manchanda et al. (2004) study the endogeneity problem of $\xi_{ij}$ in a different setting: $\xi_{ij}$ 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.
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 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.

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 period of the data, the fraction of time a station $i$ has bikes or docks available during morning and evening rush hours. By doing so, we ignore the variation of the choice set within the time window and also the variation of the same time window of the day but from different days in a year. We argue that this is a reasonable restriction for two 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 commuting time window. (On the other hand, if most usage is from casual users like tourists, short term availability would be more important). After converting the discrete choice sets to continuous average availability measure, we rewrite Equation 3.1 as:

\[
U_{ijkl} = \gamma_1 \text{bike}_i \text{ avail} \gamma_2 \text{dock}_j \text{ avail} + X'_{ijkl} \beta + \xi_{ij} + \epsilon_{ijkl},
\]

Long-term bike availability measure at the origination station $i$ and 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 $\xi_{ij}$, 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

---

6As discussed below, we estimate the morning and evening rush hour usage separately, in light of the very different usage patterns. Therefore, the availability measures are also computed separately for morning and evening rush hours.
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.

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 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 and directional travel patterns during rush hours. Properly accounting for the endogeneity of the choice set is key to recovering consistent estimates of the parameters and providing reasonable policy recommendations. 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 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 illustrating the construction of valid instrumental variables for the availability measures. We want the instruments to be uncorrelated with \( \xi_{ij} \) 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, customers are using biking 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 the instrument \( S_h \):

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

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

Second, station \( h \) needs to be sufficiently far away from station \( j \), so that almost no customer bikes the route \( hj \) or \( jh \). To be precise, we require \( d(h, j) \geq D \) where \( D \) is a threshold to be specified later. 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 \( ih \) are products in different markets. As a result, the exogenous product characteristic \( S_h \) is unlikely to be correlated with the unobserved product characteristic \( \xi_{ij} \). Formally, we have

\[
\mathbb{E}[\xi_{ij} \cdot S_h] = 0,
\]

for all \( h \in \mathcal{H}_1(ij) \). This is the exclusion restriction. Note that \( h \) and \( i \) being connected also implies that \( h \) and \( i \) are not within walking distance of each other, which further ensures that Equation (4.5) holds.

If Equations (4.4) and (4.5) are satisfied, then \( \frac{1}{|\mathcal{H}_1(ij)|} \sum_{h \in \mathcal{H}_1(ij)} S_h \) 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; \( d \) denotes the distance function; \( q_{ij} \) denote the total trip count on route \( ij \); and \( D \) denotes a required distance threshold from station \( h \) to station \( j \). We postpone the discussion about the choice of the threshold until the end of this subsection. Note that we require only the sum of \( q_{ih} \) and \( q_{hi} \) to be positive—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 \( \mathcal{H}_2(ij) := \{ h \in H : d(h,i) \geq D, q_{jh} + q_{jh} > 0 \} \).

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 2.3 and 2.4, 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 main instruments. During the evening rush hours, commuters mainly 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 \( i \) during this time of the day. 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 \( \frac{1}{|\mathcal{H}_1(ij)|} \sum_{h \in \mathcal{H}_1(ij)} GooglePlaces_h \times PopDensity_i \), and \( \frac{1}{|\mathcal{H}_2(ij)|} \sum_{h \in \mathcal{H}_2(ij)} PopDensity_h \times GooglePlaces_i \). We follow the same logic to construct the instruments for the dock availability at the ending station \( j \): namely, \( \frac{1}{|\mathcal{H}_1(ij)|} \sum_{h \in \mathcal{H}_1(ij)} PopDensity_h \times GooglePlaces_j \) and \( \frac{1}{|\mathcal{H}_2(ij)|} \sum_{h \in \mathcal{H}_2(ij)} GooglePlaces_h \times PopDensity_j \). 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) \( \mathbb{E}[\xi_{ij}|X_i] = 0 \) and \( \mathbb{E}[\xi_{ij}|X_j] = 0 \), by the exogeneity of \( X_i \) and \( X_j \); and 2) \( \mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_1(ij)|} \sum_{h \in \mathcal{H}_1(ij)} X_h] = 0 \) and \( \mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_2(ij)|} \sum_{h \in \mathcal{H}_2(ij)} X_h] = 0 \), where \( X \) is GooglePlaces or PopDensity.

With these two conditions, we have that the exclusive restriction holds for the interaction instruments; for example, 

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

The interaction instruments also satisfy the relevance condition trivially. 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 a focal station \( h \), \( i \), or \( j \). To be precise, PopDensity\( h \) is the population density integrated over a disk with radius \( R \) centered at focal station \( h \), and GooglePlaces\( h \) is the total number of Google places within \( R \) meters from the focal station \( h \). The superscript \( R \) denotes the radius of the characteristic calculation. To make set \( \mathcal{H}_1(ij) \) and \( \mathcal{H}_2(ij) \) explicitly depend on threshold \( D \) as well we add a superscript \( D \) to them. As discussed
above, our main instruments are follows,

\[
\begin{align*}
\frac{1}{|\mathcal{H}_1^{D}(ij)|} & \sum_{h \in \mathcal{H}_1^{D}(ij)} \text{GooglePlaces}_h^R \ast \text{PopDensity}_i^R, \\
\frac{1}{|\mathcal{H}_1^{D}(ij)|} & \sum_{h \in \mathcal{H}_1^{D}(ij)} \text{PopDensity}_h^R \ast \text{GooglePlaces}_i^R, \\
\frac{1}{|\mathcal{H}_2^{D}(ij)|} & \sum_{h \in \mathcal{H}_2^{D}(ij)} \text{PopDensity}_h^R \ast \text{GooglePlaces}_j^R, \\
\frac{1}{|\mathcal{H}_2^{D}(ij)|} & \sum_{h \in \mathcal{H}_2^{D}(ij)} \text{GooglePlaces}_h^R \ast \text{PopDensity}_j^R.
\end{align*}
\]

(4.6)

Finally, we 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, and, therefore, the validity of the instrument. 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 \( ij \). On the other hand, we do not have strong views about the exact values to use for \( R \). Thus, we use a range of values. In our main results, we choose \( R = 600, 800, 1000m \). We perturb the choice of the two hyperparameters \( D \) and \( R \) for robustness checks. The robustness check results are presented in the Appendix.

We use exactly the same set of instruments in Equation 4.6 for morning rush hour. Although the validity of the instruments follow the same reasoning, the correlation between our instruments and the availability measures is reversed. For example, \( \frac{1}{|\mathcal{H}_1^{D}(ij)|} \sum_{h \in \mathcal{H}_1^{D}(ij)} \text{GooglePlaces}_h^R \ast \text{PopDensity}_i^R \) is positively correlated with \( \text{bike\_avail}_i \) in the evening rush hour, but is negatively correlated with \( \text{bike\_avail}_i \) during the morning rush hour. In the online appendix, we show model-free evidence of our proposed instruments using reduced-form regressions.

4.3. Structural estimation. In this section, we discuss the structural estimation procedure and present the results. We use generalized method of moments with the instrumental variables discussed above to recover \( \alpha \), \( \beta \), and \( \gamma \). However the high dimensions in our problem creates computational challenges. Since there are over 500,000
routes in the data and the Hessian matrix is not sparse enough, the estimation is computationally infeasible, which takes days to finish one iteration even using the MPEC algorithm developed by Su and Judd (2012) in C++. To make the estimation feasible, we divide the coverage area into blocks and estimate the model on the block level instead of the route level. 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 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.

Next, we discuss the details of the block model. We divide the coverage area into uniform 1000m × 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. We define the location of the block as the location of the representative station. We denote the starting station block by \( I \) and the ending station block by \( J \). The availability measure of a station block is calculated as the average availability 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 \( J \), which is similar to distance between block centers. Compared with the route-level model, we include two additional covariates in the utility function—namely, \( \log \text{StationCount}_I \) and \( \log \text{StationCount}_J \), which are the log total number of stations 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 × 200m squares and take the center point of each square as potential origination and destination locations. 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 \( d(J,l) \).
The utility of commuters $kl$ choosing route $IJ$ is given by

$$U_{IJKl} = \gamma_1 \text{bike}_{-\text{avail}}_I + \gamma_2 \text{dock}_{-\text{avail}}_J + \gamma_3 \log \text{StationCount}_I + \gamma_4 \log \text{StationCount}_J$$

$$+ X_{IJKl}'\beta + \xi_{IJ} + \epsilon_{IJKl} \quad \forall IJ \in C_{kl}$$

$X_{IJKl}$, similarly to $X_{ijkl}$ in the station route model, includes two types of covariates in addition to intercept. First we have the same route length terms as in the station-level model (but terms here are piece-wise linear in the log of distance between station blocks $I$ and $J$, as discussed above). The second type of covaraites in $X_{IJKl}$ includes the walking distance $d(k, I)$ between origination location $k$ and starting cluster $I$, and walking distance $d(J, l)$, the walking distance from ending cluster $J$ to destination location $l$.

The choice probabilities and the density derivations are the same as in the route-level model, given by Equations (3.2) and (3.3), respectively. When calculating the integral in Equation (3.3), we discretize it into sums over commuters $kl$ that have $IJ$ in their choice set. See Kabra et al. (2016) for similar approaches.

We use the MPEC algorithm (Su and Judd, 2012) and match the observed usage at the block level $Y$ with the predicted $Y(X, W, \hat{\alpha}, \hat{\beta}, \hat{\gamma})$ to recover the parameters $\alpha$, $\beta$, and $\gamma$. We have 112 station blocks in total. 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.\footnote{The exact number of product is different from that since not all routes have positive trip count—i.e. $q_{ij} > 0.$}

We use the same set of instruments for the availability measures as in the reduced-form regressions.

<table>
<thead>
<tr>
<th>NON-LINEAR PARA.</th>
<th>LINEAR PARA.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INTERCEPT</td>
</tr>
<tr>
<td>O. WALKING DIST.</td>
<td>-3.351</td>
</tr>
<tr>
<td>(PER 1KM)</td>
<td>(0.779)</td>
</tr>
<tr>
<td>D. WALKING DIST.</td>
<td>-0.452</td>
</tr>
<tr>
<td>(0.485)</td>
<td></td>
</tr>
<tr>
<td>O. POP. DENSITY</td>
<td>0.018</td>
</tr>
<tr>
<td>(PPL PER HEC)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>D. POP. DENSITY</td>
<td>-0.048</td>
</tr>
<tr>
<td>(0.014)</td>
<td></td>
</tr>
<tr>
<td>O. GOOGLE PLC.</td>
<td>0.022</td>
</tr>
<tr>
<td>(NUM PER 4 HEC)</td>
<td>(0.15)</td>
</tr>
<tr>
<td>D. GOOGLE PLC.</td>
<td>0.307</td>
</tr>
<tr>
<td>(0.117)</td>
<td></td>
</tr>
</tbody>
</table>

| PSEUDO-$R^2$    | 0.666       |

Table 8. Demand estimates: morning rush hour
We present the structural estimation results for the morning and evening rush hours in Tables 8 and 9, 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 dock availability at the destination station than bike availability at the origination station. Indeed, for the morning rush hour, bike availability is a tiny bit higher than dock availability while for the evening rush hour, dock availability is a lot higher. This finding is consistent with the results in the reduced-form regressions. Second, the distance parameter is positive when route length is less than one kilometer, and negative when route length exceeds one kilometer. 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 destination station counts are positive and statistically significant. This implies that commuters get higher utility when there are more station options in an area, which is a sensible result.

Next, we discuss the non-linear parameters: the density model and walking distances. First, all walking-distance coefficients are negative and significant, except for walking distance from ending station block to destination in the morning rush hour. This implies that commuters are more likely to choose a station closer to their origination and destination locations of interest. However, when we compare the pseudo-$R^2$ here to those in online appendix, where we exclude walking distance from the model for robustness check, the difference is less than .01. In other words, including walking distance does not improve the explanatory power of the model substantially. See online appendix for details on the robustness checks excluding walking distance. 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 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 2.3. 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. The pseudo-$R^2$ of both the morning and evening rush hour results, 0.666 and 0.790, respectively, shows that our model fits the data well.

5. Counterfactuals

Using the estimated model, we conduct three counterfactual experiments. All three counterfactuals provide important insights and practical guidance for 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 are able to compute the predicted usage increase of this expansion in the network. 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. In other words, the counterfactual experiment evaluates different types of expansions of 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.

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. The city has been expanding the network continuously, and there are calls and proposals from the local government and communities for further expansions.

---

9Source: http://www.cityam.com/268035/lycra-ready-south-london-santander-cycles-scheme-expanding
One of the many proposals is to expand the bike share system to Islington and Hackney.\textsuperscript{10} 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 overall predicted usage increase during both the morning and evening rush hours. Results are presented in Figure 5.1.

The links indicate the routes on which usage increases after adding the new stations. 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 per station trip increase, the usage

\textsuperscript{10}Source: \url{http://www.islingtongazette.co.uk/news/environment/plea-for-boris-bikes-to-be-wheeled-out-across-islington-1-1454024}
\url{https://www.change.org/p/transport-for-london-and-islington-council-roll-out-the-london-cycle-hire-scheme-to-the-whole-of-islington}
increase is only one tenth of the usage increase that would be observed if the new station were added, instead, to the city center. It suggests that, if maximizing the overall usage is the criterion, adding stations in Islington and Hackney is not optimal. The next counterfactual provides a general discussion of the optimal locations in which to add stations, according to this criterion.

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.

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 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.

We compute the percentage increase in total usage in the network after adding one new station to each block. 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 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 usage increase varies from close to zero to above 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 300-meter 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.

Adding one station is equivalent to increasing the number of stations by 0.14%. Assuming on average each station costs the same amount of resources to set up and operational efforts to run, we could think of adding one station as expanding the system by spend 0.14% more resource. We observe an interesting economy of network in
Figure 5.2: Predicted usage increase after adding one station

Figure 5.3: Predicted usage increase vs. number of stations 800m to 1200m away

Figure 5.2: the best blocks can easily yield 0.5% more system usage, which is three times more than the 0.14% additional resource spent.

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 1km away from the focal station. There is a very strong positive correlation. This suggests that being central 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 1km 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 800m and 1200m away for each station and then averaging over all stations within each block. We regress the usage increase at each block computed above on our centrality measure of the block, the average population density of the block, the average number of Google places on the block, and the total number of stations on 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 Table 10 we report only whether the coefficients are statistically significant, and the adjusted $R^2$.

<table>
<thead>
<tr>
<th>Population Density</th>
<th>✓</th>
<th>✓</th>
<th>✓</th>
<th>✓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Places</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Number of Stations</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Our Centrality</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Eigenvector Centrality</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

| Adjusted $R^2$ | 0.34 | 0.55 | 0.44 | 0.55 |

Table 10. Determinants of usage increase and network effect

In column 1 of Table 10, we present the baseline result in which we regress the usage increase (when adding one station to each block) on three block characteristics. The result shows that the station block characteristics explain 34% of the variation in 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 62%. 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 more than two times better than 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 0.55. Moreover, across all four regressions, the adjusted $R^2$ is never higher than 0.55. 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 of adding one station to each block using the estimated model 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 16%, on average, compared to the original usage increase. This indicates that not only the number of connecting stations approximately 1km 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 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 percentage usage increase 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 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 center of the network. However, for dock availability, from 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 more scattering around the map, not as clustered in the center of the network, but definitely not clustered in the peripheries either.

Regarding the relative magnitude of bike availability vs. dock availability, we can see that adding dock availability in the most appropriate clusters show a bigger trip increase than adding bike availability in the best clusters in which to boost bike availability. This is consistent with the higher dock availability coefficient estimated in the utility for the evening rush hour.

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

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


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


