CHICAGO BOOTH
The University of Chicago Booth School of Business

Working Paper No. 15-04

Bike-Share Systems: Accessibility and Availability

Ashish Kabra
INSEAD

Elena Belavina
University of Chicago Booth School of Business

Karan Girotra
INSEAD

All rights reserved. Short sections of text, not to exceed two paragraphs. May be quoted without Explicit permission, provided that full credit including © notice is given to the source.

This paper also can be downloaded without charge from the
Social Science Research Network Electronic Paper Collection.

INSEAD
The Business School for the World®

Working Paper Series
2016/49/TOM
(Revised version of 2015/08/TOM)

Bike-Share Systems: Accessibility and Availability

Ashish Kabra
INSEAD, ashish.kabra@insead.edu

Elena Belavina
Chicago Booth, elena.belavina@chicagobooth.edu

Karan Girotra
INSEAD, karan.girotra@insead.edu

The cities of Paris, London, Chicago, and New York (among many others) have set up large-scale bike-share systems to facilitate the use of bicycles for urban commuting. This paper estimates the impact on bike-share ridership of two facets of system performance: accessibility (how far the user must walk to reach stations) and bike-availability (the likelihood of finding a bicycle). Our analysis is based on a structural demand model for spatially-differentiated products that includes distinct mechanisms for the short and long-term effects of bike-availability (via lost sales and increased user-interest, respectively). The bike-share context, and the distinct mechanisms require us to go beyond past work in incorporating real-time changes in product (bike)-availability information, and including much finer data on potential demand sources. These enhancements render traditional estimation methods computationally infeasible; we transform our estimation from the time domain to the "local-stockout-state" domain to address this. Our estimates for the Vélib' bike-share system in Paris suggest that a 10% increase in station density would increase ridership by 5.09% (±0.45%), while a 10% increase in bike-availability would increase ridership by 12.29% (±0.39%), three-fourths of which comes from fewer lost-sales, and the rest from increased user interest. We illustrate the use of our estimates in identifying neighborhoods and times to target for improvements, and in comparing alternate operational improvements and station-networks.

Electronic copy available at: http://ssrn.com/abstract=2555671

A Working Paper is the author's intellectual property. It is intended as a means to promote research to interested readers. Its content should not be copied or hosted on any server without written permission from publications.fb@insead.edu
Find more INSEAD papers at http://www.insead.edu/facultyresearch/research/search_papers.cfm

BIKE-SHARE SYSTEMS: ACCESSIBILITY AND AVAILABILITY

ABSTRACT. The cities of Paris, London, Chicago, and New York (among many others) have set up large-scale bike-share systems to facilitate the use of bicycles for urban commuting. This paper estimates the impact on bike-share ridership of two facets of system performance: accessibility (how far the user must walk to reach stations) and bike-availability (the likelihood of finding a bicycle). Our analysis is based on a structural demand model for spatially-differentiated products that includes distinct mechanisms for the short and long-term effects of bike-availability (via lost sales and increased user-interest, respectively). The bike-share context, and the distinct mechanisms require us to go beyond past work in incorporating real-time changes in product (bike)-availability information, and including much finer data on potential demand sources. These enhancements render traditional estimation methods computationally infeasible; we transform our estimation from the time domain to the "local-stockout-state" domain to address this. Our estimates for the Vélib' bike-share system in Paris suggest that a 10% increase in station density would increase ridership by 5.09% (±0.45%), while a 10% increase in bike-availability would increase ridership by 12.29% (±0.39%), three-fourths of which comes from fewer lost-sales, and the rest from increased user interest. We illustrate the use of our estimates in identifying neighborhoods and times to target for improvements, and in comparing alternate operational improvements and station-networks.

1. INTRODUCTION

Urban agglomerations across Asia, Europe, and the Americas are faced with unprecedented traffic congestion and poor air quality that threatens their attractiveness to citizens and businesses. An increased use of bicycles for urban commuting can help alleviate both these concerns. The cities of Paris, Barcelona, London, Wuhan, Hangzhou, Shanghai, New York, and Chicago (among many others) have thus set up large-scale *bike-share systems* that facilitate the use of bicycles in cities.[1]

A typical bike-share system includes a communal stock of sturdy, low-maintenance bicycles distributed over a network of parking stations. Each station provides 10–100 automated parking spots, or docking points, and a networked controller interface. A registered user can "check out" any available bicycle from a station and, at the end of her commute, can return the bicycle to *any station* in the network. Typically, the first half hour of use is free or very inexpensive and subsequent use is progressively more expensive. Bike-share systems eliminate barriers to bike-ownership such as the lack of safe parking spaces for bikes in urban dwellings and public transit stations, vandalism and theft of bicycles, and the inconvenience and cost of owning and maintaining a bicycle. They also facilitate

[1]As of June 2014, public bike-share systems were operating in 712 cities with approximately 806,200 bicycles and 37,500 stations (Wikipedia entry on "Bicycle-Sharing system").

one-way trips that make bicycles an effective "last-mile" complement to other public transit systems, such as bus, metro or regional rail.

Although bike-share systems have garnered considerable attention, their promise of urban transformation is far from being fully realized. A key reason is that while providers and operators have focused on bike-design and technology aspects, there is limited rigorous analysis of operational aspects such as station location and service-levels, nor are the user responses to such aspects understood [Tangel, 2014]. The aim of this paper is to identify relationships between ridership and operational performance of a bike-share system, and to illustrate the use of these relationships in designing systems that achieve higher ridership.

In particular, we estimate the impact on ridership of two facets of operational performance: station accessibility, or how far a user must walk to reach a station; and bike-availability, or the likelihood of finding a bicycle at the station. There are, in turn, two aspects of bike-availability. First, if nearby stations don't have bicycles *at the instance* when a user wants to take a trip, users must substitute to farther stations or abandon using bike-share. The extent of this abandonment, "lost-sales" in traditional operations parlance, is the *short-term effect* of availability. The more subtle, *long-term* aspect is that– if users typically expect a higher chance of finding a bicycle in a neighborhood, they are more likely to consider bike-share for their daily commutes, to recommend it to visitors and tourists in the area, etc., and the system will experience increased user interest.

Estimating the effect of accessibility directly requires data on the location at which each idiosyncratic user starts her trip, so as to compute the distance experienced. Neither we nor does any system operator have this data; we only observe aggregate customer choices in terms of station use. As is typically the case when customer preferences must be imputed from heterogenous customers' aggregate choices amongst products with potentially endogenous attributes, we build a random utility based choice model with unobserved customer heterogeneity that follows Davis [2006], which itself extends the celebrated work of Berry et al. [1995] (BLP) to the case of spatially differentiated products. As in Davis [2006], in our model different service locations (stations in our context, movie theaters in Davis, 2006) are the differentiated products, the differentiating characteristic is the distance a user must walk to access the locations, and the unobserved heterogeneity is the user's origin location. Station locations and bike-availabilities are the endogenously determined attributes, akin to theater locations and prices in Davis [2006].

Yet, our desire to include and estimate the distinct short and long-term effects of availability, and the bike-share context require us to go beyond the past work on spatially differentiated products [Davis, 2006, Thomadsen, 2005, Allon et al., 2011] in two respects: (1) we include information on the *actual realized* product (bike)-availability at the time of use rather than assuming full availability

(key to capturing the short-term effects of availability and a known source of bias in past works), and (2) we build and estimate a hyper-local parametric spatial density model for potential-user origins that includes a measure of *average* neighborhood bike-availability to capture the long-term effects of availability and other drivers of neighborhood-level user-interest such as locations and ridership of public-transit stops, tourist attractions and over 70,000 points of interest in the city (cafes, hotels, stores, etc.), over and above the coarser measure of census-unit level population density used in past work.

Past work in consumer choice models (including the seminal works Berry et al., 1995, Nevo, 2001; and Davis, 2006 in the spatially differentiated retail choice context) assumes that all *offered* products are always available. This has been shown to substantially bias parameter estimates in the case of consumer goods [Bruno and Vilcassim, 2008, Conlon and Mortimer, 2013]. Bike-availability is typically ~60-70%, much lower than the 90% or so availability in the case of consumer goods, and arguably more important to users, thus assuming full availability is likely to bias estimates even more in our context. More importantly, such an assumption would run counter to one of our key goals– measuring the impact of bike-availability. Yet, unlike past work, we have almost minute-to-minute information on the actual realizations of product (bike)-availability. We can thus directly include this information in our model by considering the actual choice-sets of users, or the actual set of stations from which they could have chosen bikes at the time of bike-share use. While this captures key information, it precludes aggregation of data in time or space, which makes numerical estimation of the model computationally infeasible. We develop a transformation (described below) to address this challenge.

The second departure arises from the hyper-local nature of the user-interest in bike-share and its drivers. Bike-availability likely influences user-interest via user perception of *typical or average* bike-availabilities in a small neighborhood, likely only a few blocks. Further, given the dense bike-share station networks and the fact that users must walk to these stations (rather than drive to retail-locations of the kind in past work), interest in bike-share can vary significantly from one block to another. We must thus build and estimate a *hyper-local* parametric spatial density model for potential-user origins.

The drivers of user-interest in bike-share are also likely different than those in other contexts. While interest in consumer products in different areas is likely to depend substantially on the residential populations in those areas, this might not be the case for bike-share demand. Bike-share is often used as a last-mile connection in the city by commuters from suburban areas and by the large transient population in the city (tourists, office workers, students, etc.), all of whom are not captured well in the residential populations. Therefore our spatial density model includes precise locations of public transit stops and their ridership (metro, regional rail), tourist locations and their frequentation, locations of

over 70,000 points of interest (supermarkets, cafes, hotels, educational institutions, and ten other such categories), in addition to the census-unit level population density used in past work. To capture the long-term impact of bike-availability, we also include a measure of *neighborhood average* bike-availability in this spatial density model.

Taken together, including *realized* bike-availability information and the hyper-local spatial density model with neighborhood average bike-availability, help us incorporate distinct mechanisms for the short-term and long-term effects of availability and build a full picture of the role of availability or service levels in the bike-share context.

We follow the procedures in Berry et al. [1995], Davis [2006], and Dubé et al. [2012] to estimate our model using a Generalized Method of Moments formulation that exploits the cross-sectional variation between stations for identification. We address potentially endogenous bike-availability and station locations by building the aforementioned hyper-local spatial density model and by including instruments primarily based on Berry et al. [1995] and Davis [2006]. We formulate the estimation problem as a Mathematical Program with Equilibrium Constraints (MPEC, Dubé et al. [2012]) problem. Despite its computational advantages, the scale of our data, our desire to include availability information (specifically its high-frequency changes), and the spatial differentiation of stations makes estimation on our original data many orders of magnitude more computationally intensive than the cases in past work.

The frequently changing bike-availability precludes aggregating times, while the neighborhood-to-neighborhood difference in user-interest precludes aggregating in space. Yet, we notice that in the context of our model, we can include all desired information if we aggregate times with the same choice-sets available to users; that is we transform the data from the time domain to a system-level *stockout-state* domain. This combines times where the choice-sets of all users are the same into one data-point, while retaining all systematic differences. In practice, we can do even better by combining data for a station for all times when just the *nearby* stations are all in the same stockout state, or by considering station-specific *local-stockout states*. We further develop a procedure for consistently including the information on local stockout-state of stations located within focal stations' local-stockout state relevant area. This transformation drastically reduces our computational load and allows us to include real-time bike-availability information and its high-frequency changes in long spans of data, leading to unbiased and precise estimates. We validate the use of our transformation, identifying assumptions, and computational choices in a number of small simulated data-sets akin to our data-set, and find that our procedure recovers the same seed estimates as the untransformed procedure, while being orders of magnitude faster.

Our approach allows us to impute preferences of *heterogenous customers'* (in origin location) from *aggregate* choices (we only observe aggregate station-use), to include instruments for potentially *endogenous attributes* of bike-stations and to efficiently include a large number of fixed effects– all key advantages of BLP-like models. Further, our enhancements allow us to bring these features to the study of the key operational issue of availability (stockouts) in a hyper-local, information-rich, transportation context.

The use of a BLP-like approach also advances the transportation literature by accounting for endogenous system performance. In past work, public-transport performance (service frequency and reliability) was assumed to be exogenously set whereas we allow system performance (bike-availability) to be endogenous, accounting for responsive management by the system operator (for e.g. transshipment in response to high demand), the direct reverse-causality with demand (an instance of high-demand leads to low bike-availability), etc.– features that are important in the study of modern, information-enabled, smart-transportation systems.

We estimate our model using data from the Vélib' bike-share system in Paris, the biggest bike-share system outside of China, and the densest system in the world. Our data is based on observing, every two minutes, 946 bike-stations in central Paris for a period of four months, which cover more than 4.35 million trips. We obtain locations and riderships of metro, tram and regional rail stations in Paris from RATP, the nodal transportation agency for the region. We obtain locations of more than 70,000 *stores, restaurants, bars, hotels and lodges, cafes, groceries and supermarkets, universities, parks, museums, libraries, movie theaters, shopping malls* and other *points of interests* from Google Places data. We obtain locations and frequentation of most popular tourist locations in Paris from the tourism and conventions office (Office du Tourisme et des Congrès de Paris), half-hourly weather data on temperature, humidity, wind speed, and weather conditions from weatherbase.com, and the finest administrative tract-level population density from INSEE, the french national statistics bureau.

Our accessibility estimates imply that a 10% *increase* in station density (or 10% more stations in the city) would increase system-use by 5.09% ($\pm$0.45%). On the other hand, a 10% *increase* in bike-availability can increase system-use by about 12.29% ($\pm$0.39%), of which about three-fourths of the effect (9.40%) arises from fewer lost trips (short-term effect) and the rest (2.64%) is due to increased user interest (long-term effect). We also find that only 6.01% ($\pm$0.71%) of the demand substitutes to nearby stations when confronted with a stockout at the station of choice, consistent with the significant disutility of walking that we estimate. These estimates are robust to multiple alternate model specifications, variable definitions, computational choices, instrument choices, etc.

These estimates arise from the more primitive estimates for distance disutility and the spatial user density. The user disutility for distance is nuanced– we find that the marginal disutility is increasing

(disutility is convex). For the first 300 meters, every additional meter of walking to a station decreases a user's likelihood of using a bike from that station by 0.252% (±0.092%), the effect is higher after the first 300 meters, every additional meter decreases the likelihood by 1.367% (±0.363%). A user that originates 300 meters away from a station is less than half as likely to use the system than one at the station, while a user that originates 500 meters away is highly unlikely to use the system at all. Alternate functional forms of the distance disutility lead to the same conclusions.

The estimated density model tells us that users originating at their residences, public-transit locations, supermarkets and cafes are the most significant users in the daytime, while residences, bars and cafes and are the most significant contributors in the night hours. Finally, combining the user disutility and the density model tells us that the median user walks a distance of 186 meters and only 11.01% of usage comes from users that originate farther than 300 meters of a station.

Our estimated marginal dis-utilities for distance and commuting time are in line with those in other studies on public transport systems and retail networks, and our aggregate substitution patterns are close to those observed from reduced form analysis. Our users (expectedly) walk slightly less than what users do to access other (motorized) public-transit systems, perhaps due to denser station networks and the shorter length of bike-share trips.

We illustrate the use of our estimates in system improvement by providing a number of use cases. Our estimated model can be used to estimate station-level system-use, for any given station network and any realized or average bike-availabilities at the stations in that network. This provides a powerful tool for a system manager to compare alternate station-networks and/or system management policies to arrive at the best improvement opportunities. For example, we find that increasing station density in the younger, diverse and hip districts (viz. 3, 4, 11, 12) is more useful than in other districts. The same investments in improving bike-*availability* can have more than twice the benefit in the hip district 4 than in the residential district 16. We also identify opportunities for allocating mobile availability and accessibility improvement resources at different times, for example, system managers should use transshipment trucks to improve availability in districts 11 and 12 in the morning hours, and assign them to districts 4 and 7 in the evening hours. At a system-wide level we compare the marginal benefits of accessibility or availability improvements (obtained from our model) and identify ranges of costs for which either is preferred.

Our study makes three important contributions. We provide the first large-scale archival data based analysis of user response to accessibility (walking distance) and availability in the context of bike-share systems and illustrate its use in system improvement efforts. Our method accounts for and measures distinct long and short-term effects of availability and adapts methods from the demand-estimation literature to the smart transportation context. Finally, we hope our analysis and

methodology provides a useful template for future research on consumer-behavior in other disruptive models of smart-transportation such as ride-sharing, ride-pooling, app-based ride-hailing, on-demand public-transport, etc.

2. Literature Review

This paper is related to fledgling research on bike-share systems and to other studies that measure the customer response to accessibility and availability.

Bike-Share Systems: Recent research has employed operations research methods to optimize bike-share system design and operation, considering key decisions such as number of bikes [George and Xia, 2011], station locations [García-Palomares et al., 2012] and transshipment of bikes (Henderson et al., 2016 and the references therein). Another stream is concerned with predicting ridership using demographic and traffic data (see [Daddio, 2012, Singhvi et al., 2015], and references therein). Pendem and Deshpande [2016] combine the two streams by using an empirical demand model as an input to optimal bike-allocation. Neither stream has explored the empirical consumer response to operational performance, the focus of our work.

Accessibility and Availability: The notion of accessibility has been studied in the context of motorized public-transportation systems via surveys [El-Geneidy et al., 2014], and in the context of retail networks using archival data, like this paper [Davis, 2006, Pancras et al., 2012, Allon et al., 2011, Thomadsen, 2005]. While neither case is directly comparable to our work, we compare our methods and estimates with those in these studies (in Sections 4.2 and 6.3 respectively). The impact of product availability has been studied in the context of consumer goods (eg: Musalem et al., 2010) and in a mail-order catalogue context [Anderson et al., 2006], we again compare methods and estimates.

Transportation Choice: Problems in transportation choice have primarily been modeled using three main approaches– surveys that directly measure user preferences, distances traveled, etc. (see El-Geneidy et al., 2014 and references therein), early work that used gravity models (estimated on city or district level aggregate data, for e.g. Reilly, 1931), and using multinomial logit-like models on individual user-level data (for e.g. McFadden, 1974). Our approach builds on the third class of models by incorporating unobserved consumer heterogeneity to allow estimation from aggregated data and by addressing endogeneity concerns ignored in past work.

Finally, our work is part of a renewed interest in studying facility-location, (sustainable) transportation and spatial competition in the operations management community (see for e.g. Cachon 2014, Lederman et al. 2014, Li et al. 2015, Belavina et al. 2016).

3. Data Description

We estimate our model using data from the Vélib' bike-share system in Paris. Vélib', the biggest bike-share system outside of China and the densest system in the world, includes 946 bike stations located in the city of Paris that house roughly 17,000 bicycles.[2]

Our data is built by capturing the status of stations in the network, every two minutes, via programing interfaces.[3] Each two-minute observation that we collect contains the number of available bikes and the number of empty docking points at each bike-station, we collect these snapshots for a four-month period starting in May 2013.

3.1. **Station-Use, Distances, Choice-Sets and Bike-Availability.**

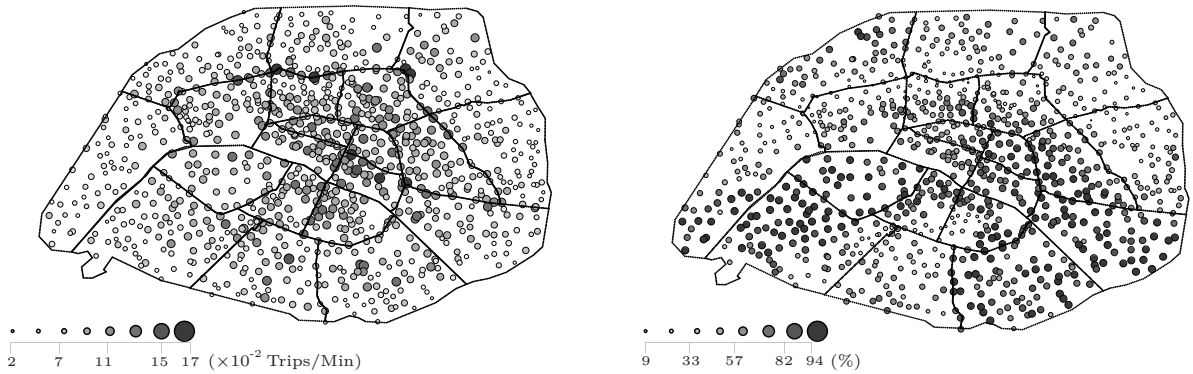
We assume that each decrement in a station's available bikes is an instance of a bicycle being checked out and used. Arguably, a declining number of bikes merely signifies the *net result* of simultaneous check-outs and returns of bikes. However, the average rates of both activities within a two-minute interval are low; check-outs and returns at a station also exhibit a *negative* temporal correlation, which implies that the likelihood of such contemporaneous events is extremely small (observed rates of these activities indicate that such simultaneity occurs at a frequency of less than 1%) and decrements accurately capture bike-use at the station level.

Vélib' system managers regularly transfer bikes from full stations to empty ones, a procedure that could confound our usage data. We therefore omit the data from any two minute period in which *more* than four bikes are checked out or brought in to a station, which we interpret either as transshipment by system managers or as outliers in the usage. This scenario is rare, so even this conservative elimination allows us to retain over 95% of the data. Results of our analysis are unchanged when other nearby thresholds are used for eliminating outliers. This leads to our main dependent variable, *station use, or the number of trips that originate at a station in a unit time* conditional on bikes being available at that station.

Figure 3.1(a) shows the mean use by station overlaid on a map of Paris; Figure 3.2(a)-(c) shows the distribution of station-use as well as inter- and intraday patterns of station-use. We observe ~3.25 Million Weekday trips and ~1.1 Million Weekend Trips. When not stocked out, a typical station in Paris is the starting point for 3.78 rides/hour on average; the rate is doubled during the evening peak hours, and is about one-seventh as much in the early morning hours.

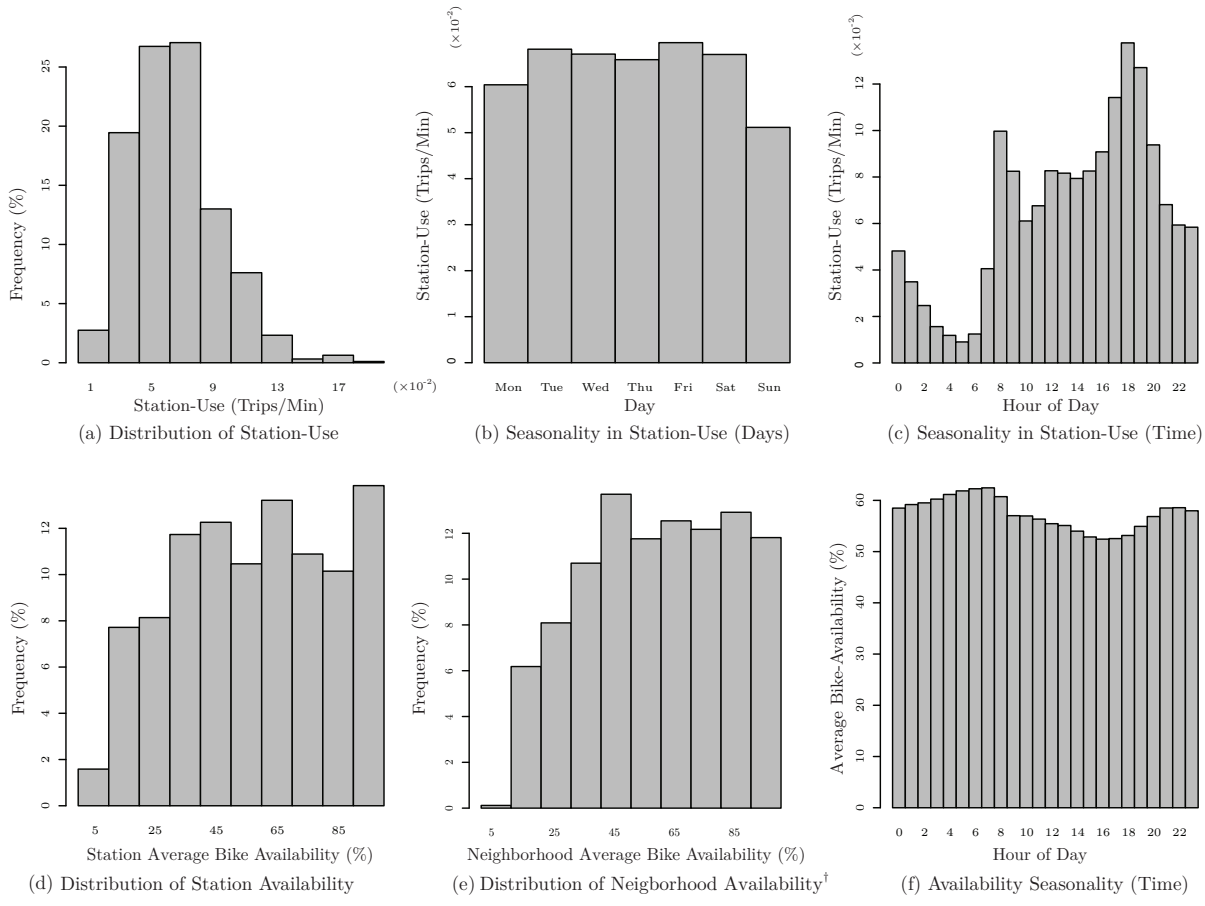
Our study has three main independent variables, the first of which is the distance that a user must walk to reach different stations. We have the GPS coordinates of each station, which allows us to compute the "Manhattan" distance between a station and any point in the city. The Manhattan

[2]"Paris fête les six ans de son Vélib' (en infographie)", Mes Débats, 15 July 2013, http://bit.ly/14Cn6n6
[3]See Oliver O'Brien, "Bike-Share Map", 31 August 2013, http://bikes.oobrien.com/paris/



(a) Average Station-Use (b) Average Bike-Availability

FIGURE 3.1. Vélib' Stations: Usage and Bike-Availability



(a) Distribution of Station-Use (b) Seasonality in Station-Use (Days) (c) Seasonality in Station-Use (Time)

(d) Distribution of Station Availability (e) Distribution of Neigborhood Availability† (f) Availability Seasonality (Time)

†Distribution across ~140,000 neighborhoods spread uniformly in the city.

FIGURE 3.2. Station-Use and Bike-Availability Statistics

distance is simply the l_1 norm distance or the sum of the absolute differences of the co-ordinates of the station and the point in the city, and is often used to capture walking distances in built-up environments [Minkowski, 1967].

The other two main independent variables–the user's choice set and the average bike-availability–both derive from the state of a station: namely, whether or not there are any usable bikes available at the station. Although we observe the number of bikes available at the station every two minutes, some of these bikes are not actually usable. First, bikes in these systems are removed from circulation after a certain number of trips for purposes of preventive maintenance; these are excluded from our analysis. Second, some bikes are officially in circulation but are in an undesirable state (e.g., a bike with a broken chain or with bird droppings on its seat). Most stations have a few such bikes, whose condition is such that they are practically unusable and they tend to be the last remaining bikes at stations. Like in recent analytical work on bike-share [Henderson et al., 2016], we account for this factor by considering a station to have usable bikes in stock only if it has *more than five* available bikes.[4] In addition to accounting for unusable bikes, arguably this specification also better captures how users think of a station's bike-availability. A user who sees only a small number of bikes may often assume that those last few are likely unusable or might well be checked out (by other users) by the time she reaches the station.

Stations that are stocked-in at the start of a two-minute period are candidates for the users' choice set. The fraction of two minute intervals at whose start the station is stocked-in is the station average bike-availability. Figure 3.1(b) shows the station average bike-availability at different stations in the city. To capture a user's perspective, we compute the *neighborhood average* bike-availability as the average of the station average bike-availability for stations within 300 meters of a point (potential user-locations); or at the nearest 3 stations for the small fractions of points where there are fewer than 3 stations in this range.[5] Panels (d), (e), and (f) of Figure 3.2 show (respectively) the distribution of average bike-availability across stations, the distribution of neighborhood average bike-availability and the hourly pattern of average station level bike-availabilities. Our model includes neighborhood average bike-availabilities at the *location* $\times$ *time-window* level, to capture a user's long-term expectation of bike-availability at each neighborhood and different time-windows (morning, evening, etc.). Table 1 provides some summary statistics for the key variables in our data. Stations are located 239 meters apart, on average, from the next nearest station, but there is wide variation in this distance. Bike-availability also exhibits significant cross-sectional variance.

3.2. **Geographic Density Variables (Controls).** The number of potential bike-share trips that originate at a location depends on the time of the day, location specific characteristics and weather.

[4]We try many alternate definitions for in-stock stations, such as stations with more than four or with more than six bikes, stations that are more than 5% or more than 10% full, and stations that have more bikes than the day's or the week's minimum number (if less than 5). Similar results are obtained with each of these alternate specifications; some are reported in Section 8 (on robustness).

[5]Survey literature, anecdotal accounts and best practices from urban planning suggest that users (on foot) typically consider 300 meters as their neighborhood [O'Neill et al., 1992, Zhao et al., 2003, O'Sullivan and Morrall, 1996].

Variables		Mean	Min	1 Qu.	Median	3 Qu.	Max	St. Dev.
Station-Use|Stocked-in	per min	0.063	0.011	0.042	0.061	0.079	0.181	0.028†
Distance to nearest station	kms	0.239	0.029	0.175	0.231	0.299	0.636	0.095†
Av. Distance to 2 nearest stations	kms	0.280	0.074	0.220	0.274	0.335	0.649	0.090†
Station Average Bike-Availability (#bikes>0)	fraction	0.886	0.370	0.838	0.930	0.978	1.000	0.125†
Station Average Bike-Availability (#bikes>5)	fraction	0.574	0.022	0.365	0.582	0.789	1.000	0.261†
Neighborhood Average Bike-Availability (#bikes>5)‡	fraction	0.589	0.046	0.399	0.594	0.798	1.000	0.242

†Across Stations ‡Across ~140,000 neighborhoods distributed uniformly in the city.

TABLE 1. Summary Statistics

While existing research on random-utility based choice models typically assumes the potential demand at a location (or in a market) to depend simply on population density, we build and estimate a hyper-local parametric spatial density model for user origins that includes locations and ridership of public-transit stops, tourist attractions and over 70,000 points of interest in the city (cafes, hotels, stores, etc.), in addition to the population density.

Population Density. Vélib' trips can potentially originate from the residences of individuals; the population-density of different locations in Paris captures this. The city of Paris is divided into *arrondissements* or districts. We allocate each location in the city to a district using precise administrative boundaries made available from Google Maps as Keyhole Markup Language files. We collect population and area data for each district from INSEE, the french national statistics bureau, and use this to compute the relevant population density.

Transit Data: Metro, RER and Tram Lines. Anecdotal accounts suggest that the Vélib' system is used extensively as a last-mile supplement to the existing public-transit systems in Paris, thus locations of transit-stops can potentially be an important origin location of Vélib' users. Paris has one of the densest metro systems in the world, a modern multi-line tram system and is at the center of 5 commuter lines of a regional train network (Réseau Express Régional, RER); together roughly 4.25 million commuters use these systems every day. Using a Google Maps provided API, we collect data on the locations of 245 metro stations, 33 RER stations and 26 tram stations. We supplement these locations with each metro and RER station's annual ridership data collected using the French public transport operator, RATP's open data archive. For inter-change stations (co-located metro and RER stations), we use the higher ridership level among the two.

Points of Interest Data. Likewise, trips of Vélib' users may also originate at different points of interest such as restaurants, museums, etc. Using an API provided by Google Places, we collect the location coordinates of about 70,000 points of interest in the city of Paris and the type of the point of interest. The most significant types of point of interest are *stores or retail locations, restaurants, bars, cafes, other food-service locations, hotels and lodges, groceries and supermarkets, shopping malls, universities,*

parks, museums, libraries and movie theaters, which cover a majority of the points of interest in the city, the remaining categories or unclassified points are included as "other points of interest".

Tourist Frequentation. Paris is the world's most-visited tourist destination and tourists from around the world use the Vélib' system as a convenient way to commute from one tourist spot to another and to also enjoy the beautiful cityscapes. Thus tourist attractions can potentially be an important demand source for Vélib'. We include the locations of the 20 most popular tourist spots in our model and supplement these locations with their annual visitor numbers provided by the Office du Tourisme et des Congrès de Paris (OTCP).

Weather Data. Finally, the potential number of bike-share users could vary with varying weather conditions. We collect half-hourly weather data for the city of Paris, specifically the temperature, humidity, wind speed and "conditions" (clear, mist, cloudy, etc.) from weatherbase.com. There is little weather variation in our study period, with mild temperatures (between 10^oC and 30^oC) for 92.7% of the period. Nevertheless, we deseasonalize the system-use data using the prevalent weather conditions (see Appendix A).

4. A Structural Model for Station Use

As discussed before, estimating the effect of accessibility directly requires data on the location at which each idiosyncratic user starts her trip, so as to compute the distance experienced. Neither we, nor does any system operator have this data; we only observe aggregate customer choices in terms of station use. As is typically the case when customer preferences must be imputed from heterogenous customers' aggregate choices amongst products with potentially endogenous attributes, we build a random utility based choice model with unobserved customer heterogeneity that follows Davis [2006], which itself extends the celebrated work of Berry et al. [1995] to the case of spatially differentiated products.

4.1. A User Choice Model.

Consider a population of utility-maximizing users distributed spatially over a given area. Users choose between different stations of the bike-share system and other modes of transport. The indirect utility of user i from accessing the bike-share system at station $f \in \{1, ..., F\}$ at time $t \in \{1, ...T\}$ is given by

$$u_{ift} = \beta_0 + h\left(\beta_d; d\left(L_i, L_f\right)\right) + \gamma_{w \times di(f)} + \xi_{ft} + \epsilon_{ift}, \tag{4.1}$$

where L_i is user i's origin location, $h()$ is a parametric function that captures the disutility of walking a distance, $d\left(L_i, L_f\right)$ gives the distance between user i and station f (located at L_f). Survey literature, anecdotal accounts and current practices in bike-share network design suggest that users (on foot) are

less sensitive to distance till approximately 300 meters, beyond which users are expectedly much more sensitive to distance [O'Neill et al., 1992, Zhao et al., 2003, O'Sullivan and Morrall, 1996]. We thus assume that the distance disutility function h is piecewise linear with a change in slope or a kink at 300 meters. We examine other kink points and alternate functional forms in Section 8.2.

The operator $w(t)\colon \{1,\ldots,T\} \to \{1,\ldots,6\}$, abbreviated to w wherever possible, maps the time to one of six "time-windows" in a day (05h30–08h00, 08h00–12h00, 12h00–16h00, 16h00–20h00, 20h00–00h30, 00h30–05h30). The windows correspond to the system–operator's internal planning windows: early-morning, morning-rush, afternoon, evening-rush, late-evening and night (metro-closed). $\gamma_{w\times di(f)}$ are the *time-window* $\times$ *district* fixed effects; $di\,(f)$ is the district for station f. The term ξ_{ft} denotes the unobservable components of utility that are common to all users for station f at time t, or the *station* $\times$ *time*-specific shock. The ϵ_{ift} are the idiosyncratic *user* $\times$ *station* $\times$ *time*-specific error terms; we assume that these errors are of type I extreme value, and are independent and identically distributed (i.i.d.). The user's utility from using other modes of transport is

$$u_{i0t} = \xi_{0w} + \epsilon_{i0t};$$

here ξ_{0w} is the unobservable component of this utility that is common to all users in time-windows w, this captures how the other options become more or less attractive over the course of the day. The ϵ_{i0t} are the idiosyncratic utilities that users derive from other means of transport, which we also assume are type I extreme value, and i.i.d.

Users observe the current bike-availability (widely available via the official Vélib' app/website or many third-party apps such as CityMapper), compare the utility of using bikes from different in-stock stations and choose the station that earns the highest utility. The probability of choosing a station now follows the Logit form. Let S_t be the set of stations that are in stock at time t– the choice-set for user i at time t. The probability of a user i using a bike from station $f \in S_t$ at time t is given by

$$p_{ift}\left(\theta, \xi_{\cdot t}\right) = \frac{\exp\left(\beta_0 + h\left(\beta_d; d\left(L_i, L_f\right)\right) + \gamma_{w\times di(f)} + \xi_{ft}\right)}{1 + \sum_{g\in S_t} \exp\left(\beta_0 + h\left(\beta_d; d\left(L_i, L_g\right)\right) + \gamma_{w\times di(g)} + \xi_{gt}\right)}. \tag{4.2}$$

Here $\xi_{\cdot t}$ is the vector of ξ_{ft}, the unobservable characteristics or residuals at time t; and θ represents the parameter values (α, β, and all fixed effects γ).

The net use at station f at time t, or λ_{ft}, is obtained by aggregating choice probabilities of all users in the population:

$$\lambda_{ft}\left(\theta, \xi_{\cdot t}\right) = \int_{L_i} p_{ift}\left(\theta, \xi_{\cdot t}\right) \cdot P_t^D\left(L_i; \alpha\right)\, dL_i. \tag{4.3}$$

Here $P_t^D\left(L_i; \alpha\right)$ is the spatial density of a user's origin location, precisely it is the number of potential users that originate at a location L_i in the two-minute interval t. Specifically,

$$P_t^D\left(L_i;\alpha\right) = \alpha_0 + \alpha_1 \cdot naba_{L_i,w(t)} + \alpha_2 \cdot pd(L_i) + \vec{\alpha}_{3,w(t)} \cdot \vec{V}_{w(t)}\left(L_i\right). \tag{4.4}$$

$naba_{L_i,w(t)}$ is the *neighborhood average* bike-availability at location L_i in the time-window $w(t)$, $pd\left(L_i\right)$ is the population density at location L_i and $\vec{V}_w\left(l\right)$ is a vector of geographic density-relevant variables viz. transit, tourist and points of interests.

Higher *average* bike-availability at stations is likely to encourage users in their catchment areas to consider using bike-share, to incorporate bike-share into their daily commutes and to recommend it to visitors and tourists in the area, among other things. The neighborhood average bike-availability term $\alpha_1 \cdot naba_{L_i,w(t)}$ (defined in Section 3.1) in the spatial density model captures this effect. This effect can be interpreted as the long-term effect of bike-availability.

Beyond bike-availability, our model also allows for the number of potential users to depend on the population density and location specific characteristics. Population density is captured by the second term $pd\left(L_i\right)$, while the vector $\vec{V}_w\left(L_i\right)$ consists of other geographic density-relevant variables. It includes 1) Indicator variables for the presence of different kinds of points of interest at location L_i. Specifically, indicators are included for stores, restaurants, bars, cafes, other food-service locations, hotels and lodges, groceries and supermarkets, shopping malls, universities, parks, museums, libraries, movie theaters, and the catch-all category "others" 2) A variable that takes the value of the annual transit-traffic at location L_i if there is a transit stop at L_i and if the transit system is in operation in time-window w, or is set to 0 otherwise 3) Similarly, a variable that takes the value of the annual tourist-frequentation if there is a tourist location at L_i, and 0 otherwise. We allow the impact of these geographic-density variables to vary between day-hours and the night-time. Formally, the vector $\vec{\alpha}_{3,w(t)}$ is $\vec{\alpha}_3^n$ for the time-window 00h30-05h30 and $\vec{\alpha}_3^d$ otherwise.

4.2. **Comparison with Past work.** As discussed in the introduction, the above model follows Davis [2006]. As in Davis [2006], in our model different service locations (stations in our context, movie theaters in Davis [2006]) are the differentiated products, the differentiating characteristic is the distance a user must walk to access the locations, and the unobserved heterogeneity is the user's origin location. Incorporating this heterogeneity in our model ensures that when a station stocks out, its users are more likely to substitute to nearby rather than distant stations. Station locations and bike-availabilities are the endogenously determined attributes, akin to theater locations and prices in Davis [2006].

Past work in consumer choice models (including the seminal works of [Berry et al., 1995, Nevo, 2001], and Davis [2006] in the spatially differentiated retail choice context) assumes that all *offered* products are always available. This has been shown to substantially bias parameter estimates in the case of consumer goods [Bruno and Vilcassim, 2008, Conlon and Mortimer, 2013]. Bike-availability is typically ~60-70%, much lower than the 90% or so availability in the case of consumer goods, and

arguably more important to users, thus assuming full availability is likely to bias estimates even more in our context. Further, this would run counter to one of our key goals– measuring the distinct short and long-term impacts of bike-availability. Anupindi et al. [1998], Musalem et al. [2010], Bruno and Vilcassim [2008] and Conlon and Mortimer [2013] address product availability issues by developing methods to estimate model parameters in presence of limited or no product availability information to the econometrician. Yet, we have information on the actual realizations of product (bike)-availability and we include it directly in our model.

The real-time bike-availability enters our model indirectly via the relevant choice set that is realized at each time, S_t, in Eq. 4.2. When stations are stocked-out they do not enter any user's choice set and serve no users. The effect of this on system-use depends on the extent of customer substitution to nearby stations– if all customers substituted then there would be no effect on system-use; if none substituted, system-use would change by the same amount as a fraction of time the station is stocked out. We call this the short-term impact of bike-availability and it is estimated through the substitution pattern that is embedded in our choice model.

Demand for bike-share is hyper-local, the typical catchment area that a bike-share station serves is much smaller than that for retail stores considered in past work, as users must walk rather than drive and the networks themselves are much more dense. We thus build and estimate a much finer hyper-local parametric spatial density model for potential-user origins as compared to Davis [2006] (Eq. 4.4), aided by the much freer availability of mapping data today from a variety of Google products. More importantly, this hyper-local density model allows us to include a measure of the typical or average bike-availability in a neighborhood, which likely drives user interest in bike-share, or the above discussed long-term effect of bike-availability.

Taken together, to the best of our knowledge, this is the first model to include and estimate distinct mechanism of short and long-term effects of availability, providing a fuller picture of service-levels in the context of dense urban transportation systems such as bike-share.

5. Model Estimation

We follow the procedures in Berry et al. [1995], Davis [2006], and Dubé et al. [2012] to estimate the model described in Equations 4.2-4.4 using a Generalized Method of Moments formulation that exploits the cross-sectional variation between stations for identification. Like these works, we address the potentially endogenous bike-availability and station locations by including instruments primarily based on Berry et al. [1995], and Davis [2006]. We formulate the estimation problem as a Mathematical Program with Equilibrium Constraints (MPEC, Dubé et al. [2012]) problem. However, our

desire to include availability information (specifically its high-frequency changes) and the neighborhood to neighborhood difference in user interest makes estimation on our original data computationally intensive. We develop a transformation of our data that permits estimation.

5.1. **GMM Estimation.** We estimate our model using the optimal two-step process for the Generalized Method of Moments (GMM, Hansen [1982]). Our moment conditions are:

$$E_{f,t}\left[Z_{fw}\xi_{ft}\left(\theta^*\right)\right] = 0 \text{ , and} \tag{5.1}$$

$$E_w\left[\gamma^*_{w\times di}\right] = 0 \text{ for } \forall di \tag{5.2}$$

The first set of equations restricts the residuals ξ_{ft} to be uncorrelated with instruments Z_{fw} at the true parameter values θ^*. The second set of equations requires all *time-window* $\times$ *district* fixed effects within a district to sum up to 0, so that *district* fixed effects are not implicitly imposed. This allows us to estimate the effect of density variables like population density which are constant at the *district* level.

We use the MPEC method that minimizes an objective function based on the moment conditions 5.1. Following Dubé et al. [2012], we use the balance conditions of Berry et al. [1995] and 5.2 as constraints that must be satisfied at the optimal estimates. In our context, the *balance conditions* or constraints, equate the actual and predicted use rates for each station–time pair; that is the following $F\times T$ equations:

$$\lambda_{ft}\left(\theta, \xi_{\cdot t}\right) = \Lambda_{ft} \quad \forall f, t\,; \tag{5.3}$$

For efficiency, we include an additional condition that matches the total *potential* user-interest in a day to an estimate of this number obtained from external sources.[6]

5.2. **Endogeneity and Instruments.** System managers might choose station locations and bike-availabilities on the basis of neighborhood characteristics. For instance, system managers might provide higher bike-availability in areas with important transit hubs or in areas with politically important stakeholders. There is also reverse-causality, bike-availability influences demand, but the event of high demand realizations also leads to lower bike-availability. Together these can bias our estimates.

[6]Formally, $\int^t \int^{L_i} P_t^D\left(L_i;\theta\right) dL_i dt = T^D$, is the total potential user interest in a day. In principle, T^D could be identified by the above procedure, however we find that there is a broad range of T^D where our parameters of interest are essentially the same and using an external estimate for T^D turns out to be more efficient. More interestingly, this implies that our results are not at all sensitive to a broad range of values around T^D, therefore even a rough estimate of T^D suffices. We base it on the working age population of Paris with each person using the system once per day which gives 1,120,320 potential users per day. As expected, our estimates are very robust to this choice (Section 8).

First, note that our use of a much richer density model (Eq. 4.4)– one that includes transit-information, tourist locations, thousands of point of interests that are classified into over 14 categories that include retail, food and educational establishments among others, over and above the population density that has been used in all past literature on spatially differentiated products much alleviates these concerns. Essentially, we include many more relevant neighborhood characteristics, so there are fewer endogeneities on account of *unobserved* neighborhood characteristics. Nevertheless, we include instruments that follow past work that has applied BLP-like instruments in the context of spatially differentiated products [Davis, 2006, Allon et al., 2011, Thomadsen, 2005]. BLP considers characteristics of competing products as instruments for endogenous characteristics of the focal product. Likewise, following Davis [2006], we use *characteristics of neighboring stations' neighborhood* as instruments for endogenous characteristics of the focal stations. Interestingly, while in principle this strategy follows what is done in past work, our much richer density model also gives us many more such exogenous characteristics, potentially making our instruments more effective.

Formally, we include two sets of instruments in our main model: instruments from Davis [2006] and those from Thomadsen [2005]. For each of density variables in $\vec{V}_w(l)$, each Davis-instrument has the form $V_{fwj}(a,b,c,d)$ where j denotes the indexes of the variables in vector $\vec{V}_w(l)$. $V_{fwj}(a,b,c,d)$ is the sum of the variable $V_{wj}(l)$ for locations that are at a distance (a,b) of all stations within distance (c,d) of the focal station f. $\vec{V}^I_{fw}$ is the vector of these instruments, now defined at the station-level. The parameter sets used for (a,b,c,d) are (in meters) $(0,25,0,25)$, $(25,50,0,50)$, $(50,100,0,100)$, $(0,100,0,100)$, $(100,300,0,300)$, $(300,500,0,500)$, $(0,100,100,300)$ and $(0,100,300,500)$. Note that multiple alternate sets of parameters must be used to capture potential non-linearities. We also include simpler instruments based on those used by Thomadsen [2005], Allon et al. [2011]: distance to the nearest station, average distance to 5 nearest stations and number of stations within 500m of a station. Finally, we also have an instrument based on population density at that station. Together, the vector $\vec{V}^I_{fw}$, the instruments from Thomadsen [2005], $pd(L_f)$, and the model-covariates (intercept term and *time-window*$\times$*district* dummies) constitute the vector Z_{fw} in the moment condition 5.1.

We test the relevance of our instruments by considering the change in the adjusted R^2 of models with bike-availabilities as dependent variables on inclusion of the instruments. Specifically in a linear model with covariates for a station's exogenous characteristics (nearby density variables) and time-window and district fixed effects, inclusion of the instruments increases the adjusted R^2 from 20.4% to 47.3% for station average bike-availability and from 25.1% to 57% for neighborhood average bike availability (see Table 5 in the Online Appendix for more details). This suggests that the proposed instruments are highly relevant in our context.

In extended analysis, we estimate our model with alternate instruments: i) Instruments based only on the focal station's neighborhood characteristics (akin to Davis [2006]'s robustness analysis) ii) A novel instrument based on the average realized rate of incoming bikes at station f in lagged time-window $w-1$, demeaned at the station level; iii) Alternative parameters formulations for above BLP-Like instruments. All instruments provide essentially identical estimates (reported in Table 3) further reinforcing the validity of our instruments.

5.3. **A Computational Challenge.** Despite using the techniques proposed by Dubé et al. [2012], numerically estimating the above model is extremely computationally intensive. There are two reasons that drive the computational burden: first, as is the case with all BLP-like random utility models, the optimal estimates are found by a search algorithm, where more coefficients implies more iterations of the search process, this is effectively handled by the nested fixed point method or the MPEC method. Second, unique to our context, the market share computations for our balance equations require computing station-choice for users at each location and at each time, the former to account for neighborhood-level variation in user interest and the latter to include real-time availability information.

Specifically, for each iteration of parameter-estimates, we need to compute the demand functions $\lambda_{ft}(\theta, \xi_{\cdot t})$. There are over 20 million such functions, on account of $F = 946$ stations, T =22,743 two minute observations for each station. Further, for each λ_{ft} computation, we numerically integrate the choice of users located in the entire city of Paris which is a *105* km^2 area that we discretize into nearly 210,000 points. This implies that about 4 trillion computations are required to compute the demand function for each parameter-iteration. For finding optimal estimates, several iterations of the above computations are required which would take the total number of computations to over a quadrillion, a process we estimate would take over an year.[7]

Note that the computational burden can be reduced by aggregating data over time or by considering less granular spatial models. Aggregating over time implies ignoring the changing real-time bike-availability, as discussed above in Section 4.2, this has been shown to bias estimates and is a key concern in our context. Less granular spatial models are also not desirable as stations can be as close as 50 meters, which requires our density model to be high resolution. Finally, we could also simply consider subsets of our data, but this would reduce the precision of our estimates, especially since the variability in two-minute use is very high; the mean, median and standard deviation of use per minute are 0.065, 0.000, and 0.209 respectively so that coefficient of variation is 3.232. This implies that large spans of data are needed to infer robust estimates. Past work has typically not incorporated either product availability or spatial information, and uses much smaller scales of data; typically having 6

[7]Based on solving smaller instances of this problem using a highly optimized implementation where all core components are implemented and run with C++ binaries and sparse analytical Jacobian and Hessian implementations fed to an IPOPT optimizer.

orders of magnitude fewer computations than us (see Section C.2 for details). In summary, the spatial richness of our model (considering each station and thousands of distinct user-locations separately) and our desire to include the availability information (which requires us to consider each two-minute separately) drive the computational burden of our model.

5.4. **A Transformation.** We propose a transformation of our data that reduces the computational burden while still being able to exploit the bike-availability information. Note that in our model, station-use λ_{ft} is a function of the choice set of stations, the *time-window*$\times$ *district* fixed effects, neighborhood average bike-availabilities, and the variables in the density model. Within a *time-window*, only the choice sets change with time. We can therefore aggregate station data points that are in the same time-window, and have the same choice set of stations, without losing any information relevant to our estimation. That is, if we aggregate data according to *station* $\times$ *system-stockout-state* $\times$ *time-window*, times for which the system-stockout-state is the same, or equivalently times for which choice set for users at all locations is the same are considered as one data-point. We can then include all availability information while potentially decreasing the computational burden.

The computational advantage of estimating our model in the stockout state domain instead of the time domain arises from the fewer distinct stockout states during each time window than there are distinct two-minute time intervals. However, if the stockout state is defined at the system level, there could be as many as $\approx 2^{946}$ distinct values, and while not all distinct values are realized in the data, the number of realized system-states is of the same order as distinct two-minute times; hence such a transformation is only slightly superior to the original model.

We notice that the use at station f is not affected equally by *all* the other stations' stockout states. The stockout state of *neighboring* stations of station f have a much stronger effect than the stockout state of far off stations.[8] The implication is that we can construct a local stockout state for each station and aggregate our data on such local stockout states rather than on the systemwide stockout states. The local stockout state will have lower dimensionality than the systemwide stockout state, so there will be far fewer distinct local stockout states than distinct systemwide stockout states, making the approach computationally feasible while retaining almost all relevant information.

We construct the local stockout state for a station f by working upwards from the choice sets of each user. We limit a user i's choice set to the nearest m_d stations to her; the set of such stations is denoted by N_i. Given this, note that for a station f, the only relevant bike-availability information is the availability at stations close enough to users who are close enough to station f. For any station f, we can write the set of relevant stations N_f as

[8]Note that even though far off stations don't meaningfully affect the use at the focal station, we must estimate the entire system jointly because of the users in the overlapping neighborhoods of different stations.

$$N_f \equiv \bigcup_{i|\, f \in N_i} N_i.$$

The stockout state at time t of stations in N_f is given by a binary vector v_{ft}— it is the "local" stockout state for station f at time t. Let the set of all such realized local stockout states be given by $V_f \equiv \bigcup_t v_{ft}$.

Next we aggregate the use at station f for all times where the local stockout state was v_f, a typical element in V_f . We use Λ_{fwv_f} to denote the average observed use at station f in a time-window w over all times when the local stockout state is v_f. Accounting for the salience of state v_f shall prove useful, so let σ_{fwv_f} denote the number of observations that were averaged to obtain Λ_{fwv_f}; these numbers will serve as weights in subsequent analysis.

Taken together, the transformed model now is: for $f \in N_i \cap S_{v_f}$,

$$\begin{aligned} & p_{ifwv_f}\left(\theta, \xi_{\cdot w \cdot}\right) \\ & = \frac{\exp\left(\beta_0 + h\left(\beta_d; d\left(L_i, L_f\right)\right) + \gamma_{w \times di(f)} + \xi_{fwv_f}\right)}{1 + \sum_{g \in N_i \cap S_{v_f}} \exp\left(\beta_0 + h\left(\beta_d; d\left(L_i, L_g\right)\right) + \gamma_{w \times di(g)} + \xi_{gwv_f}\right)}, \end{aligned} \tag{5.4}$$

where S_{v_f} denotes the set of stations with available bikes in state v_f. Then station-use is given by

$$\lambda_{fwv_f}\left(\theta, \xi_{\cdot w \cdot}\right) = \int_{L_i} p_{ifwv_f}\left(\theta, \xi_{\cdot w \cdot}\right) \cdot P_w^D\left(L_i; \alpha\right) \, dL_i. \tag{5.5}$$

We notice that the above user choice probabilities p_{ifwv_f} depend not only on the utility of using station f but also on the utility of using *other* stations in user $i's$ choice set (i.e., stations g such that $g \in N_i \cap S_{v_f}$, stations that are close by to user i and stocked-in in state v_f). Specifically, the choice probabilities also depend on elements ξ_{gwv_f}, the residual for use of station g in a local state *for station* f. While we determine ξ_{gwv_g}, ξ_{fwv_f} and so on from the balance equations, ξ_{gwv_f} is not directly determined. Furthermore, the set of stations local to station g is not the same as the set of stations local to station f, which means that the local stockout state of station g is not fully determined by v_f (state v_f does not map to any state v_g), which means ξ_{gwv_f} can also not be indirectly determined from other residuals.

Note that the effect of stockouts and therefore the local state is captured in the changed user choice sets and doesn't *systematically* affect the residuals. Specifically, the expected value of residuals ξ_{gwv_g}, is independent of the stockout state that is realized, i.e. $E[\xi_{gwv_g}] = E[\xi_{gw\hat{v}_g}]$, where $\hat{v}_g$ is any other local state of station g. Thus, we can compute a consistent estimate of ξ_{gwv_f}, using weighted averages of terms ξ_{gwv_g}. Accounting for the weight σ_{gwv_g}, which is inversely proportional to the variance of ξ_{gwv_g}, gives us the best consistent estimate for ξ_{gwv_f}, formally given by,

$$\hat{\xi}_{gwv_f} = \frac{\sum_{v_g} \sigma_{gwv_g} \xi_{gwv_g}}{\sum_{v_g} \sigma_{gwv_g}}.$$

We can then compute p_{ifwv_f} by replacing ξ_{gwv_f} by the estimates of $\hat{\xi}_{gwv_f}$. This procedure allows us to consistently include information on the local stockout-state of stations located within focal stations' local-stockout state relevant area.

The transformed estimation procedure now is as follows. The set of moment conditions used by the GMM estimator are,

$$\begin{aligned} E\left[Z_{fw}\sigma_{fwv_f}\xi_{fwv_f}\left(\theta^*\right)\right] = 0 \quad &, \text{ and} \\ E_w\left[\gamma^*_{w\times di}\right] = 0 \text{ for } &\forall di \end{aligned} \tag{5.6}$$

The constraints used to determine values of all $\xi_{\cdot w}$ (ξ_{fwv_f}s) are

$$\lambda_{fwv_f}\left(\theta, \xi_{\cdot w}\right) = \Lambda_{fwv_f} \quad \forall f, w, v_f\,;$$

Note that we use weights σ_{fwv_f} for transformed observations in constructing moment conditions to get efficient estimates (as $Var\left(\xi_{fwv_f}\right) \propto 1/\sigma_{fwv_f}$, refer sec. 6.3.7 Cameron and Trivedi, 2005). We use the MPEC approach both for its computational efficiency and limited error propagation in comparison to the nested fixed point method. The non-linear optimization with constraints is done using the open source package Interior Point Optimizer (IPOPT, Wächter and Biegler [2006]), that is interfaced with R via ipoptr [Ypma, 2010]. The full estimation procedure and the standard error computation is described in the Online Appendix, Section C.1.

5.5. **Computational Choices.** To integrate over the continuous elements of the spatial density model, we discretize the city using a square grid with edge length $\mathscr{D} = 25$ meters. We consider the choice set of each user to be the four nearest stations, $m_d = 4$, this user-level choice drives the size of local stockout state of a station. In the robustness analysis (Section 8), we show the estimates obtained by considering the five nearest stations and the consequent local stockout states.

Considering local stockout states much reduces our computational burden. Yet, for some stations in high station-density neighborhoods, there can still be a large number of local stockout states that are realized in the data. For 75% of the *station*$\times$*time-window*s, more than 30 local stockout states are realized in data and there are a total of 609,858 stockout states across all *station*$\times$*time-window*'s. Note

that the frequency distribution of the realization of each stockout state is expectedly highly skewed.[9] So, while the computational burden increases proportionally with each extra state considered, the relevant and reliable data-observations come from a small subset of higher frequency states. Thus, for each station, we consider the 8 most frequent states, typically these cover more than 60% of the data available for a vast majority of the stations. In Section 8, we also reevaluate our model with the top 16 states, which typically covers about 75% of the data; we find nearly identical estimates.

5.6. **Validation in Simulated Datasets.** While the full validation of our approach remains the subject of a dedicated study that considers many alternate contexts, we provide a limited validation in our context. Specifically, we validate the use of the local-stockout state transformation and our computational choices–the limits on the choice set and the use of top states– on smaller simulated datasets, where both our approach and the full approach (time-domain, no limits on choice sets, all states) are computationally feasible. We created a number of small simulated datasets for demand at 30 stations around the city-center (Hôtel de Ville) for 50 two-minute time-intervals in the evening-rush time window. The detailed data generating process, analysis and the results are provided in Section C.3 of the Online Appendix.

We estimate our model on the simulated datasets in three ways: (1) The benchmark estimation procedure that uses the untransformed time-domain based moment conditions (Eq. 5.1) and places no limits on the choice set of customers. (2) Using the transformed local stockout state based conditions (Eq. 5.6) and imposing a consistent limit on the choice set of the customer and (3) Approach (2) plus focusing on just enough local stockout states to cover 75% of the data for the typical station (the approach of this paper).

We find that all three approaches recover seed estimates from the demand model. Specifically, the recovery from the first procedure provides support for the moment conditions used in our estimation and validates our approach, while the recovery from the last two procedures validates the use of the *top* states among the *local* stockout states. Interestingly, while all three procedures recover the seed estimates, the computational burden of the third approach is an order of magnitude less than that of the untransformed approach, even in these small datasets. We expect the difference to be much larger in a dataset comparable in size to the one in our study. Taken together, while this analysis provides some validation of our approach– a full validation of such transformations in other contexts remains the subject of a future study focused on further developing the methodological ideas here.

The estimation procedure and model described in the preceding sections provide us consistent estimates of the accessibility (distance) effect using the variation across stations and between different

[9]Consider a station-time-window with 5 other stations in the local stockout state, say each of these neighboring stations has a 90% average bike-availability in this time-window. The most likely state is 9^5=59049 times more likely to realize than the least likely state.

Walking Distance (0-300mts)	Walking Distance (>300mts)	Bike-Availability (naba)	Number of observations	$\chi^2(df)$
-2.700 (0.495)***	-15.734 (3.043)***	0.005 (0.001)***	39,302	0.049 (136)

Effects on System Use	%Increase in Demand	
	Mean	95% C.I.
10% decrease in Station Density	5.090%	(4.508%-5.414%)
10% increase in Bike-Availability (Short-term)	9.399%	(9.341%-9.483%)
10% increase in Bike-Availability (Total)	12.293%	(11.914%-12.696%)

TABLE 2. Estimation Results and Effects on System-Use

local-stockout states for stations, while the long-term bike-availability effect is identified using the variation across stations and time-windows.[10] The short-term effect of bike-availability derives from the estimated consumer utility (specifically the marginal disutility of distance) and the structure of the station network.

Together, our approach allows us to impute preferences of *heterogenous customers'* (in origin location) from *aggregate* choices (we only observe aggregate station-use), to efficiently include instruments for potentially *endogenous attributes* of bike-stations (for e.g. bike-availability determined by system managers), and to include a large number of fixed effects in our model (the numerous *station* $\times$ *time-window* $\times$ *local-stockout-state* (ξ_{fwv_f}'s) and *time-window* $\times$ *district* ($\gamma_{w\times di}$'s)). These are all key advantages of BLP-like models. Further, our enhancements allow us to bring these features to the study of the hitherto unexplored operational issue of availability (stockouts) in a hyper-local smart-transportation setting.

6. Results, Interpretation and Comparison

6.1. **Results.** We estimate our models separately for weekdays and weekends to allow for differences in spatial user-origin density and consumer behavior in these two time periods. We focus on the results for weekdays in this section, while those for weekends are reported in Section 8.

Table 2 reports the main effects obtained from estimating our model, while the coefficients from the density model are reported in Table 6 (online appendix). We find a statistically significant effect of distance. Expectedly, users incur a disutility from walking. Interestingly, the *marginal* disutility of walking is lower for the first 300 meters, and is much higher for further distances, in effect we have increasing marginal disutility of distance (or convex disutility).

[10]To ensure cross-section variation in bike-availabilities is meaningful and persistent over time, we look at the correlation in bike-availabilities defined at monthly level for a *station* $\times$ *time-window* and find it to be *0.89*, high enough to justify using it as such.

Next, we compare the likelihood of using the system of a user that originates at the bike-station versus one that originates further away (Figure 6.1(a)). For the first 300 meters, every additional meter of walking to a station decreases a user's likelihood of using a bike from that station by 0.252% (±0.092%), the effect is higher after the first 300 meters, every additional meter decreases the likelihood by 1.367% (±0.363%). A user that originates 300 meters away from a station is less than half as likely to use the system than one at the station, while a user that originates 500 meters away is highly unlikely to use the system at all.

The positive and statistically significant coefficient for neighborhood average bike-availability suggests an important long-term effect of increase in bike-availability or that of reliably finding bikes: higher average bike-availability leads to higher use of the bike-share system in the neighborhood. We next interpret these marginal effects in terms of system-use.

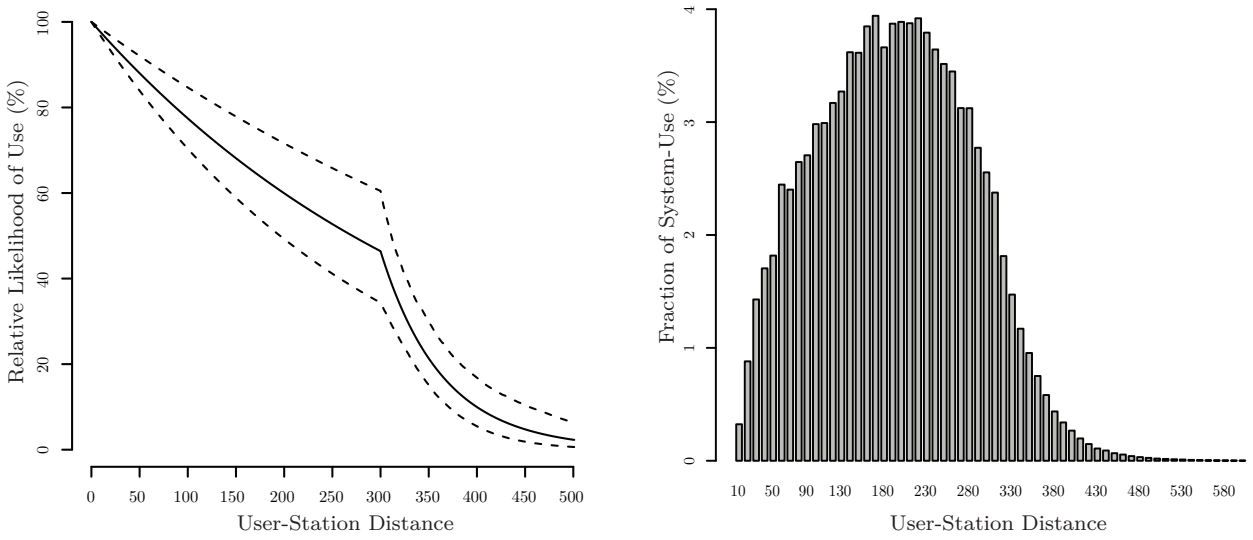
6.2. **Effects on System Use.**

6.2.1. *Accessibility.* Consider the case when the station density is increased by 10% (say, by adding about 95 new stations to the city).[11] Our estimates suggest that such a *10% increase in station density results in a 5.090% (±0.453%) increase in system-use.*

Figure 6.1(b) illustrates the estimated distances traveled by users of the bike-share system, essentially this combines the density of potential users-origin locations with their likelihood of using the system based on their distance from stations and other utility attributes. We find that the median user travels about 186m to reach her preferred station. 6.15% of the system-use comes from users originating within 50m of their preferred station, another 13.18% of system-use comes from the next 50m, about 35.87% of usage comes from users within 100-200m; 33.76% usage comes from users starting within 200-300m and the remaining 11.01% usage comes users that start further than 300m away from a station.

6.2.2. *Substitution and Short-Term Effect of Bike-Availability.* The estimates from our structural model also allow us to estimate how users behave when a station stocks out. We find that, on average, *93.992% (±0.710%) of a stocked-out station's demand is lost* (so only *6.008% (±0.710%)* of its unserved users substitute to other stations). This figure is calculated by removing one station at a time from the network and then re-computing the total use from our demand model for the remaining stations. We follow this procedure for all stations, one by one; the reported estimate is the average effect, or the effect of removing a typical station. The implication is that a 10% increase in

[11]We effectuate this by reducing all user–station distances by 4.653%, which increases density by 10% while shrinking the city's area by 9.090% ($1 - 0.953^2$), and then scaling-up system-use from this shrunken city to estimate the effect in our actual city. This ensures that we capture solely the distance effect since it preserves all spatial relationships between stations and the nature of the station network design.



(a) Use-Likelihood and Distance (b) System-Use and Distance

FIGURE 6.1. Effects of Distance

bike-availabilities of all stations would lead to an immediate *9.399%* increase in system-use—yet this is only the short-term effect that arises from the increasing odds of stations being present in the choice set.

6.2.3. *Total Effect (Long-term + Short -term) of Bike-Availability.* Our estimated model reveals, in addition, a long-term effect that arises from more users adopting and incorporating bike-share into their lifestyles. We compute this effect by considering a network where each station has a 10% higher average bike-availability than the status quo. We then use our user-level choice model (again with estimated parameters) to compute the new level of system-use. In sum we find: *increasing the bike-availability of all stations by 10% would increase system-use by 12.293% (±0.391%)*; of this, three-quarters of the gains (9.399%) arise immediately on account of a reduced "lost trips", while the rest (2.645%) will be achieved over the long-term on account of increased user interest in the system.

6.3. **Comparison of Estimates.** We compare our estimates with estimates (or in some cases decisions implied by those estimates) from reduced-form analysis and other studies in the bike-share, public-transport and retail-store network design contexts.

Comparison of Distance Estimates. Reduced-Form Analysis: While it is hard to measure the accessibility effect directly from a reduced form analysis,[12] we can use some reduced form analysis for the substitution effect which derives directly from the marginal disutility of distance and the existing station network design. We look at the difference between the use at a station when its neighboring

[12]The effect of increasing distances between stations from any reduced form station-level model necessarily confounds the larger catchment area effect with user response effect

stations have available bikes and when they do not. In a regression model, we explain the log of station-use using the stockout-state of neighboring stations, specifically we include a dummy variable based on whether or not the nearest station to the focal station has any bikes available. We find an average *6.056%* increase in use at a station when the nearest station runs out of available bikes, a number very close to the substitution percentage implied by the distance estimates in our model.

Bike-Share Systems: We are aware of no other econometric study (survey or archival-data based) that has attempted to estimate the disutility of distance in the context of bike-share systems. In practice, European bike-share system designers follow a common handbook (Büttner and Petersen, 2011) which suggests that very few users walk further than 300 meters and provides station location guidelines based on this assumption. This guideline is squarely in line with our estimates, our distance estimate also implies that only 11.019% of the use of a station comes from users who walk further than 300 meters.

Other Public Transport Systems: Several studies survey the walking distance of users of bus, light rail and metro systems. [O'Neill et al., 1992, Zhao et al., 2003] report that 75-80% users of public-transit systems walk less than 400 meters. El-Geneidy et al. [2014] finds that, in Montreal, the 85th percentile walking distance to the bus (resp., rail) transit system is about 524m (resp., 1,259 m). O'Sullivan and Morrall [1996] reports that transit planners in several Canadian and American cities consider the catchment area to be no further than 300-900m, with the median light rail user in Calgary, Canada walking 320m. Alshalalfah and Shalaby [2007] report that median access distance of bus users in Toronto is about 200m and that of subway users is 350m. In comparison, our estimates are marginally lower, our median user walks about 186m, and almost 90% of the demand comes from the first 300m. Since bike-share systems are used for much shorter trips than those taken by other public transport systems, bike-share systems exist in more densely populated areas, and have a much denser station network; it is expected that our users walk less.

Zhao et al. [2003], based on survey data of about one thousand users' transit use (bus or rail) in southeast Florida, determines that usage decreased exponentially with a coefficient of *-4.265/km.* Gutiérrez et al. [2011] uses survey data from the Madrid metro network to estimate the effect of walking distance and finds an average distance disutility coefficient of about *-1.689/km.* Our comparable estimate, *-2.7/km,* is squarely in line with these observed coefficients.

Retail Store Networks: While the context of retail store networks provides us with multiple past studies to compare our estimates, these estimates typically consider the disutility of distance for users who *drive* to the retail locations. One way to compare with these estimates is to convert distances to commuting time using average walking and driving speeds. We compare our estimates with those in Davis [2006] (driving to movie theaters), Pancras et al. [2012] (driving to grocery stores), and

[Thomadsen, 2005, Allon et al., 2011] (driving to drive-through fast food outlets). Our estimate is higher than that of Davis [2006] and Pancras et al. [2012], comparable to those in Allon et al. [2011] and lower than that in Thomadsen [2005]. Together, our estimate is again squarely in line with the average of these past studies. The average disutility of commuting time from these studies is *14.49$/hr* while, for majority of the users in our study, it is *16.87$/hr*.[13]

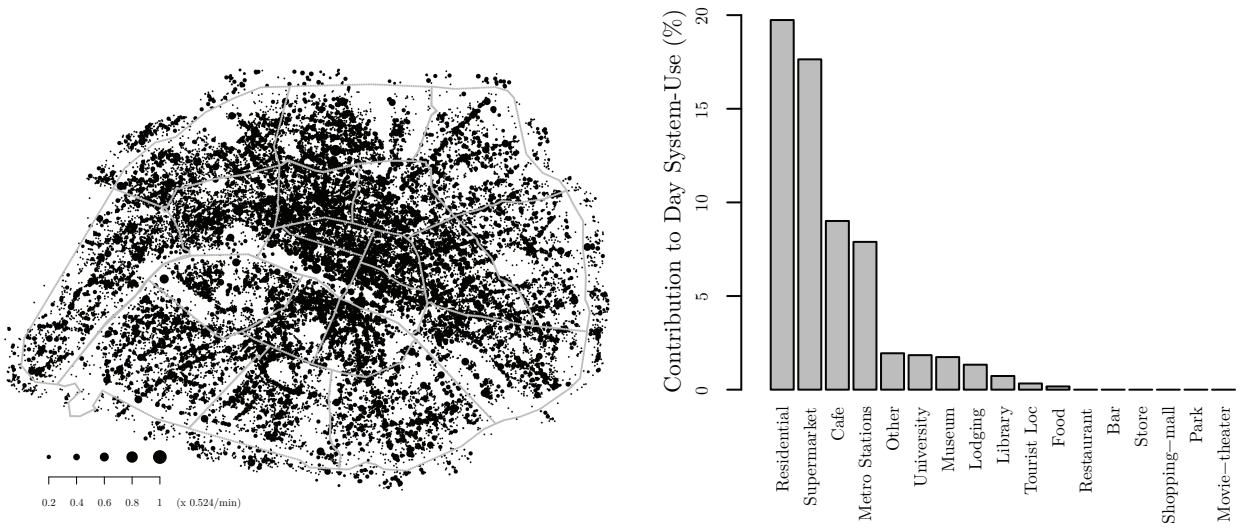
Comparison of Bike-Availability Estimates. We are not aware of any study that has looked at the impact of availability in the context of bike-share systems. Availability in bike-share systems is not directly comparable with that of other public-transportation systems where it concerns the frequency or reliability of a service. The only somewhat comparable estimates are from the long-term and short-term effects of product availability in the context of consumer goods. Note however that demand for customer goods is much less time-sensitive than that for transportation, products are not modeled as spatially differentiated, and as such availability is expected to play a much smaller role.

Anderson et al. [2006] in their study of a home-bedding catalogue retailer find that a 10% decrease in stockouts leads to a 7.2% short-term increase in product sales. The lost-demand in their case is much lower than ours (28% in their case compared to 94% in our case) probably because users are more willing to wait for bedding ordered via a catalogue, than for bikes to get somewhere. The long-term impact of all items ordered by a customer in their setting being out of stock compared to none is 22% lower future demand, i.e. a 10% decrease in stockouts leads to a 2.2% long-term increase in product demand. This estimate is comparable to our estimate of a 2.645% increase in long-term demand due to a 10% higher bike-availability.

Musalem et al. [2010] in their study on estimating the effect of stockouts in the shampoo category find that almost no sales are lost when a few brands stock out, but as much as 20.02% of sales might be lost when multiple brands stockout, suggesting there is more than 80% substitution to adjacent shampoo brands. Not surprisingly, there is much less substitution to adjacent stations in our context (only about 6%), perhaps because users must walk to other stations rather than just simply switch to comparable, adjacent brands.

6.4. **Density Model.** Figure 6.2(a), shows the estimated user interest at origin-locations. Examination of the estimated density shows that it is highly granular and varies significantly both across and within each of the districts, validating our estimation approach of using cross-sectional variation. Figure 6.2(b) shows the relative contribution of users originating at different kinds of points of interest to bike-share use. Users originating at their residences, public-transit locations, supermarkets and cafes are the most significant users in the daytime, while residences, bars and cafes and are the most

[13]We assume dense city-driving speeds of *25 km/hr* and walking speeds of *4 km/hr*.



(a) User Interest at Origin Locations[†]

(b) Origin Locations Type and System-use

[†]For visual clarity residential locations have been excluded

FIGURE 6.2. Estimates of Density Variables

significant contributors in the night hours. Full estimates of our density model are provided in Table 6, Online Appendix.

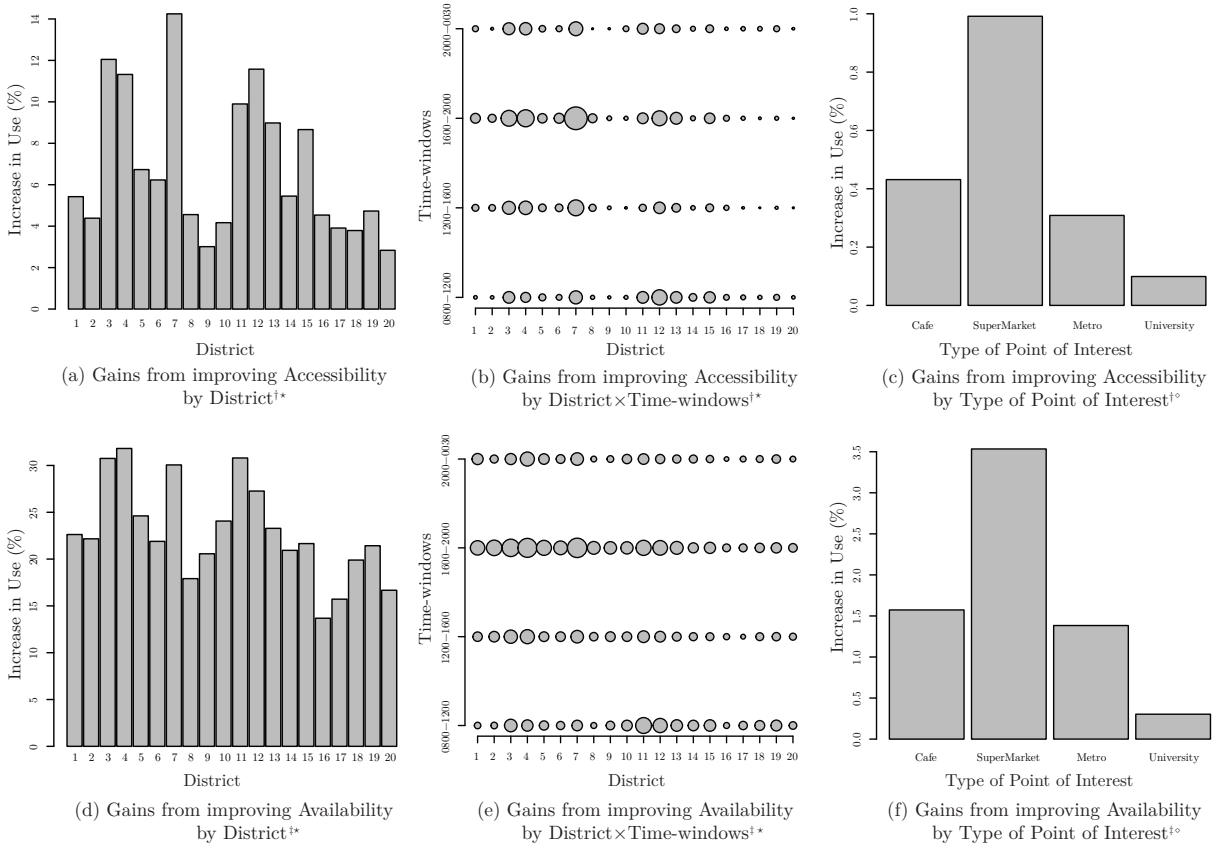
7. Managerial Use Cases

The estimated model can be used to provide station-level system use for any given station network, and any realized or average bike-availabilities at the stations in that network. This provides us with a powerful tool to compare alternate station-networks and/or system management policies (and the associated accessibilities and availabilities) and identify the best improvement opportunities. In this section, we provide an illustration of a subset of ways in which our estimates are used to improve ridership in bike-share systems.

The analyses provided are intended to be illustrative of the potential different uses of our estimates, nevertheless in the interest of simplicity, these analyses necessarily exclude a number of other factors not considered in our study such as political and geographical constraints on station locations and sizes, management challenges in increasing availability, etc. A full, rigorous, careful analysis of the below issues remains an open subject for future study; what follows is simply indicative.

7.1. **Identifying the Best Areas for Improvement.**

Accessibility. System managers can improve the accessibility of stations by adding more stations to the network. While the analysis of Section 6 estimated the advantages of system-wide improvements, in this section we compare different targeted improvements to identify the best areas for improvement.



(a) Gains from improving Accessibility by District†⋆

(b) Gains from improving Accessibility by District×Time-windows†⋆

(c) Gains from improving Accessibility by Type of Point of Interest†◦

(d) Gains from improving Availability by District‡⋆

(e) Gains from improving Availability by District×Time-windows‡⋆

(f) Gains from improving Availability by Type of Point of Interest‡◦

†(a),(b),(c) show e ect of 10% increase in station-density ‡(d),(e),(f) show e ect of 0.1 increase in station average bike-availability ◦Increase in use in (c),(f) is relative to system-use. ⋆(a),(b),(d),(e) show increase in station-use in a District or District×Time-windows relative to average station-use in the system. These e ects are normalized by number of stations a ected in each case.

FIGURE 7.1. Different Impacts of Improving Station Density and Bike Availability

Figure 7.1(a) shows the effect of increasing station density in different districts, normalized by the number of stations in the district. The effect of increasing density is generally higher in the younger, diverse and hip districts (viz. 3, 4, 11, 12) and districts where the station density is currently low (district 7). Interestingly, districts 1 and 2, although quite busy and densely populated, reap lower benefits of increasing density; likely these districts are already saturated with stations. This analysis also suggests that station density in some of the outer districts (16 to 20) could be reduced while investing those resources in the more popular districts. Perhaps, due to equal access concerns, system managers have over-invested in them at the expense of more popular districts.

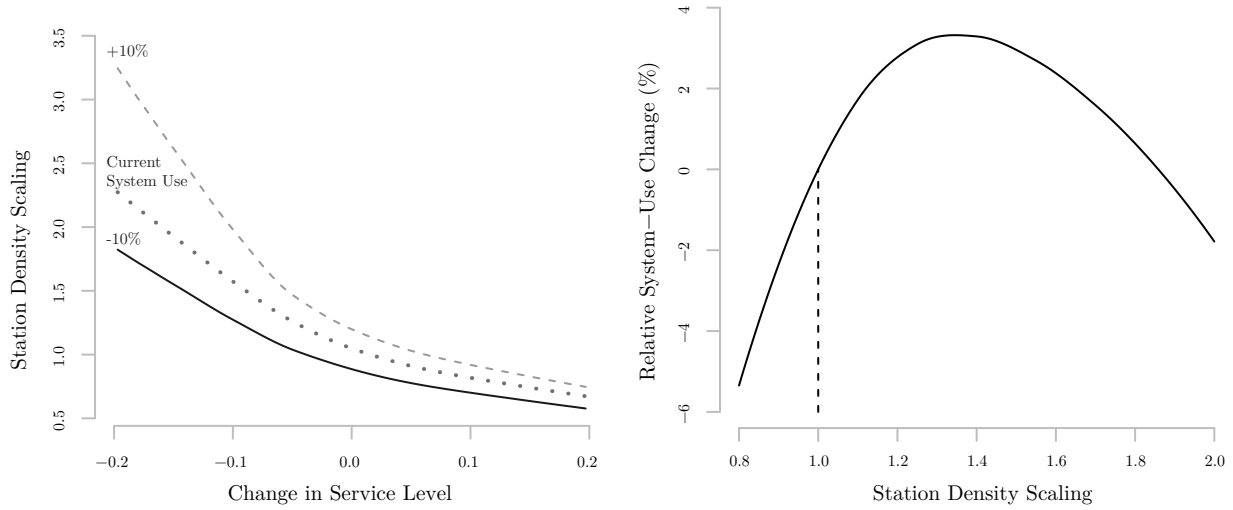
System operators also have the ability to temporarily increase station density in some time-windows by the use of so-called mobile stations or "Valet/manned" stations. Interestingly, comparing different districts in different time-windows (Figure 7.1(b)) helps us identify opportunities for the use of such mobile resources–for e.g. mobile stations could be employed in district 12 in the morning, and moved to district 7 in the evening.

Finally, we consider if there are specific locations where adding stations would be most useful. Figure 7.1(c) provides increases in system-use that come from increasing the accessibility for users originating from a particular kind of location, mathematically by decreasing access distances of specific users by 10%. We find that bringing stations closer to users originating at *supermarkets* has significantly higher impact than bringing stations closer to metro, cafe or university locations. Perhaps, users originating at supermarkets are carrying their shopping and it is most worthwhile providing them with accessibility improvements, or perhaps these are locations which the current network serves least well.

Bike-Availability. System managers can improve bike-availability at specific stations by giving higher priority to these stations in trans-shipments, scheduling of preventive maintenance, etc. Figure 7.1(d) considers improvements in different districts, normalized by the number of stations in each district; the same investments in improving availability have more than twice the benefit in the hip district 4 than in the residential, district 16. Improving availability in the evening time-windows (1600-2000) is the most useful. Considering different districts in different time-windows reveals further opportunities for improvement. System managers should allocate availability-improving resources (transshipment trucks, etc.) to districts 11 and 12 in the morning hours, and move them to districts 4 and 7 in the evening hours (Figure 7.1(e)). Finally, we compare the effect of improving availability for users from different origins, we again find that the effect of improving availability is most prominent for users who originate at supermarkets (Figure 7.1(f)).

7.2. **Accessibility or Availability Improvements.** System-managers are often given targets by city-managers to improve ridership. Figure 7.2(a) plots iso-ridership curves, which provide different combinations of accessibility and availability improvements that lead to desired improvements. For example, a 10% increase in system-use can be achieved by all changes along the dashed curve—by increasing station density by 10%, or by increasing bike-availability by 0.05 and decreasing station density by 3%, and so forth. Such iso-riderships curves are used by system-managers to translate the policy goal of improved ridership into operational performance targets.

Figure 7.3 identifies which improvements–on accessibility or on availability–are preferred, by combining our estimated benefits of these improvements with the potential costs of achieving these improvements. Panel (a) considers system-level costs, while Panel (b) considers costs under the assumption that accessibility improvements are achieved by adding more bikes at new stations while availability is best increased by extra transshipments. While the preferred strategy would depend on the precise

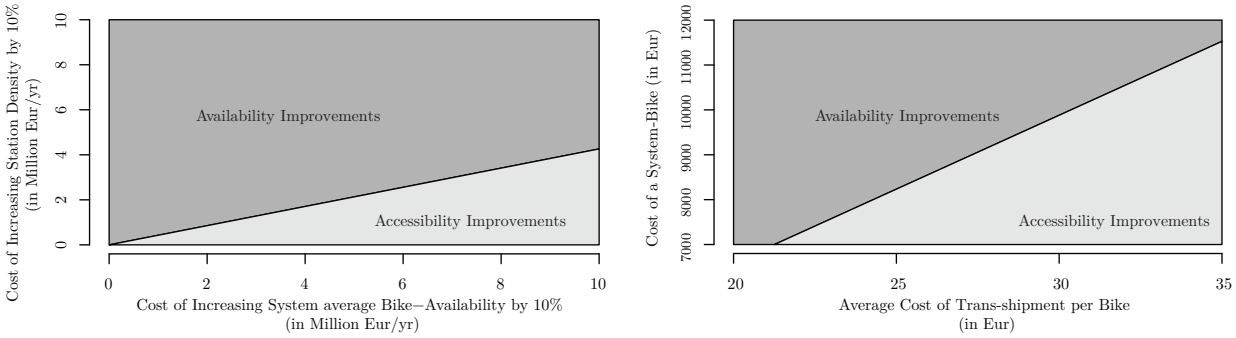


(a) Iso-Ridership Curves

(b) Counterfactual System-use for Alternate Station Designs†

†Arrival and departure processes assumed to be compound poisson with negative binomial jumps. Alternate station designs are constructed by splitting/combining stations and proportionally scaling use-rates (mean and variance). System-use for alternate designs is computed by scaling the distance of each user to all the stations by a factor $\sigma^{-0.5}$ for station density scaling σ and using simulated availability.

FIGURE 7.2. Iso-Ridership Curves



(a) System Costs

(b) Component Costs†

†Bike-Availability increases are obtained by transhipping bikes needed to achieve gains. Density increases are obtained by adding System-bikes, that is bicycles and a proportional number of stations, docking points and management infrastructure.

FIGURE 7.3. Preferred Improvements

cost estimates, anecdotally, for the city of Paris, the costs are in the ranges where this analysis predicts availability improvements are much preferred.[14]

7.3. **Alternate Network Designs with the Same Number of Bikes.** It is interesting to compare station networks that have the same capital costs as the current system. Practically this means that they utilize the same number of bikes (system costs are almost directly proportional to number of bikes), and the system-designer has to trade-off between high accessibility or high-availability systems. On the one hand, a high-density network with many distributed stations but relatively fewer bikes

[14]Anecdotal cost estimates provided by the operator are: for adding new stations, 10,000€/system-bike/yr, and for transshipments cost roughly 25€/bike-transshipment.

at each station reduces user distances to stations, which increases accessibility. On the other hand, a low-density network with fewer stations but with more bikes at each station can achieve higher bike-availability owing to the well-known statistical benefits of holding pooled inventory in systems with demand variability [Cachon and Terwiesch, 2009].

Figure 7.2(b) shows the estimated ridership for a continuum of different station designs. The station network gets denser with the increasing horizontal axis. Our analysis reveals that substantial gains might be achieved with denser station networks than the status quo, that is with networks that are denser with more smaller stations. Specifically, a network that would have had more stations (1272 stations instead of 946), with each station being 75% smaller (~24 dock points instead of 32 currently), would have had 10.168% lower availability, but higher accessibility–and as a result would have achieved the highest ridership, while using the same number of bikes.[15] While such networks improve system-use on account of the availability-accessibility trade-off, they might be less preferred due to the increased management costs of more stations, more complicated IT and control infrastructure, costs of maintenance, etc. It may also simply not be feasible to increase density.

Interestingly, the above policy prescriptions would be the same even if the true parameters were much different than estimated parameters. The finding that denser networks of smaller stations would have had higher demand would hold even if the true marginal disutility of distance was as much as 51% lower than our current estimates, while the preference for using additional resources to improve bike-availability via trans-shipments rather than adding stations would be robust if the cost of bikes was as much as 20% lower or even if the cost of trans-shipments were as much as 20% higher.

8. Robustness

8.1. **Variable Definitions, Model Specification, Computational Choices and Instruments.**

We test the robustness of our effect sizes to alternate variable definitions, model specifications, and to computational choices made in model estimation. Table 3 reports the results of our estimation under many alternate assumptions; row (1) replicates our original estimates (from Table 2) for easy comparison. Rows (2) and (3) of the table report the estimates obtained under alternate definitions of bike-availability. Row (2) gives estimates from a model where a station is said to be in-stock or have bikes available if there are *more than four* bikes available at the station (versus five bikes in the original estimation), Row (3) considers a station stocked in if it has *more than six* bikes available at the station. The estimates are similar to those obtained under our original regressions.

[15]An alternate interpretation of this analysis might be that the existing system would be optimal in terms of the accessibility-availability trade-off if the system managers believed the marginal disutility of distance were roughly half our current estimate. Or if the bike-availability coefficient were 2.65 times the current estimate and distance estimates 0.60 times and so on.

		Primary variables							
		Walking Distance (0-300mts)	Walking Distance (>300mts)	Bike-Availability (naba)	10% increase in Station Density	10% increase in Bike-Availability		Number of observations	$\chi^2(df)$
						Short-term	Total		
(1)	Original Estimates	-2.700 (0.495)***	-15.734 (3.043)***	0.005 (0.001)***	5.090%	9.399%	12.293%	39,302	0.049 (136)
(2)	Stockout: ≤ 4 Bicycles	-2.810 (0.494)***	-15.375 (3.039)***	0.005 (0.001)***	5.079%	9.404%	12.443%	39,320	0.047 (136)
(3)	Stockout: ≤ 6 Bicycles	-2.611 (0.493)***	-15.398 (2.830)***	0.005 (0.000)***	5.170%	9.407%	12.103%	39,090	0.052 (136)
(4)	Weekends only	-3.963 (0.495)***	-14.814 (3.099)***	0.003 (0.000)***	5.295%	9.477%	11.137%	34,374	0.051 (136)
(5)	Metro in outside option	0.737 (0.689)	-14.800 (2.574)***	0.005 (0.001)***	4.721%	9.364%	12.509%	39,302	0.048 (136)
		Metro outside option							
			0.770 (0.166)***						
(6)	Var. of Bike-Availability	-3.534 (0.521)***	-11.816 (2.308)***	-0.036 (0.002)***	5.577%	9.447%	11.714%	39,302	0.045 (136)
(7)	Finer Grid Size (20 meters)	-2.633 (0.501)***	-15.732 (3.061)***	0.003 (0.000)***	5.084%	9.399%	12.265%	39,302	0.048 (136)
(8)	16 Top states considered	-2.359 (0.437)***	-14.823 (2.700)***	0.004 (0.000)***	5.077%	9.434%	11.821%	71,994	0.040 (136)
(9)	Target density -10%	-2.753 (0.494)***	-15.786 (3.071)***	0.004 (0.000)***	5.038%	9.331%	12.221%	39,302	0.049 (136)
(10)	Target density $+10\%$	-2.681 (0.495)***	-15.694 (3.030)***	0.005 (0.001)***	5.137%	9.453%	12.347%	39,302	0.049 (136)
(11)	Choice set size $= 5$	-3.728 (0.516)***	-9.959 (1.760)***	0.005 (0.001)***	5.220%	9.392%	12.675%	38,602	0.046 (136)
(12)	Focal station Instruments	-4.33 (0.518)***	-16.967 (4.561)***	0.005 (0.001)***	5.278%	9.404%	12.219%	39,302	0.028 (71)
(13)	Demeaned Bikes-in$_{w\text{-}1}$	-3.105 (0.453)***	-16.311 (3.236)***	0.005 (0.000)***	5.055%	9.412%	12.571%	39,302	0.058 (137)
(14)	Alt. Instrument Parameters - I	-1.894 (0.590)**	-15.974 (3.328)***	0.005 (0.001)***	4.890%	9.394%	12.442%	39,302	0.039 (70)
(15)	Alt. Instrument Parameters - II	-3.325 (0.461)***	-15.368 (2.984)***	0.005 (0.000)***	5.009%	9.403%	12.469%	39,302	0.053 (169)

*(p-value<0.05) **(p-value<0.01) ***(p-value<0.001)

TABLE 3. Robustness Tests

Row (4) replicates our analysis for data from weekends only. We find that the impact of increasing station density and the short-term effects of increasing availability are marginally higher on weekend days as compared to weekdays (5.295% v/s 5.090%, 9.477% v/s 9.399%) while the total (or long-term effect) is noticeably lower (11.137% v/s 12.293%). There are some interesting differences in density variables. As one would expect, users originating from bars, lodging, museums, residences, tourist locations, cafes and other food locations (at nights only) account for a higher proportion of system use on weekends than on weekdays; while universities, libraries and grocery stores (all in the day time only) are higher demand drivers on weekdays than on weekends.

In row (5) we include the distance to the nearest metro station as a covariate in the outside option of each user. Metro stations in addition to acting as feeders to use of bike-share stations, can also act as substitutes to bike-share. While, the feeder effect is already captured in the density parameters, inclusion of metro stations as the outside option can further capture any substitution effects. Although we had hoped to find a substitution effect, we find the opposite effect in the model in Row (5). This positive coefficient suggests that irrespective of where the metro variable is included, the net effect of the presence of a metro station is an increase in bike-share use, i.e. metro stations feed demand to bike-share stations rather than act as substitutes. In row (6) we use the variance of bike-availability instead of its mean value as the variable in the density model. We use the same instruments as those in the original model. We find that variability of bike-availability has a long-term negative effect on bike-share use, consistent with our expectations.

Next, we investigate the role of various computational choices made in estimation. In row (7) we provide estimates obtained by using a finer grid for our numerical integration (viz., one that covers 1.5 times as many simulation points for continuous spatial elements) this produces no qualitative change in the estimated effects. In row (8) we consider 16 top states for each *station*$\times$*time-window*. In row (9) and (10) we test the sensitivity of our estimates with respect to the choice of total market density. In row (11) we increase the definition of local choice set of a user to nearby *five* stations. The estimates are exceptionally robust to these choices.

Finally, we investigate the effect of different instruments on our estimates. In row (12) we use only the focal station's neighborhood characteristics as instruments as used in Davis [2006]. In row (13) we use a novel instrument which is the average realized rate of incoming bikes at station f in lagged time-window $w-1$, demeaned at station level. This instrument affects the starting number of bikes at a *station* $\times$ *time-window* and therefore its bike-availability. It however does not affect the unobserved factors influencing station-use at f in time-window w, after controlling for demand sources in the density model and station level factors removed in demeaning process. In row (14) and (15) we use alternative parameters for construction of instruments $V_{wj}\,(a,b,c,d)$. The set used for

row (14) is $\{(0,100,0,100),\ (100,300,0,300),\ (300,500,0,500)\}$ and for row (15) is $\{(0,100,0,100),\ (100,200,0,200),\ (200,300,0,300),\ (300,500,0,500),\ (0,100,100,300),\ (0,100,300,500)\}$. The marginal effects are again very close to the original model.

In short, we find that our estimates are robust to various model specifications, variable definitions, computational, and instrument choices.

8.2. **Robustness of Distance Disutility Function.** Our main model assumes a piecewise linear form for the disutility of distance. We reran our model with linear and quadratic disutility functions, and a variety of different kink points for the piecewise linear form– 100, 200, 250, 275, 325, 350m instead of 300m in original model.

Irrespective of the specification used for the distance disutility, our estimate for the effect of *bike-availability* is essentially identical. In so far as the effect of *distance* is concerned, depending on the specification we get different estimates for parameters of the distance disutility function, but remarkably irrespective of the specification, the the effects of walking distance on ridership is essentially the same– a 10% increase in station density always leads to between 4.56%-5.19% increase in system use, with all but 3 specifications returning an effect within 2% of our original estimate. Also notably, irrespective of the functional form– all specifications imply the disutility from distance is convex. The extent of substitution or the short-term effect of bike-availability which derives from the accessibility effect is again essentially identical in all specifications ranging from 9.399% to 9.435%. Finally, the total effect ranges from 11.970% to 12.293%, The full results for all these alternate specifications are reported in Table 7 of the Appendix.

9. Discussion

Each use of a bike-share system involves two transactions: the user must choose a station with available bikes; and she must also be able to return the bicycle to a station with empty docking points. Thus each station features two streams of use—outgoing and incoming—and so there are two kinds of availabilities, bike-availability and docking-point availability. System-use presumably depends on both kinds of availability, but our analysis has focused on *outgoing* use and bike-availability.

Observe that at the system level, incoming and outgoing use must be equal and each corresponds to the number of trips; therefore, either use type can be analyzed to develop important prescriptions for system-use. Yet bike-availability and dock-availability can have different and significant effects on system-use. There are two important differences between these effects that make the analysis of bike-availability far more relevant. First, when bikes are not available, the user has the option of either seeking out another station or forgoing the bike-share system entirely. However, the same cannot be said when docking points are not available: the affected user does not have the option of abandoning

the bicycle and she can complete her trip only by finding another station (users using Vélib' get an extra 15 free minutes when the preferred station has no available docking points). Note that in this case the user can ride the bicycle to an alternate station, which is presumably easier than walking there. So in the short term, use is affected more by bike-availability than by the availability of docking points.

Second, bike-share systems are designed with many more docking points than bikes (to accommodate demand asymmetries at different times of the day, etc.); there are usually almost twice as many docking points as bikes. Hence not finding an available dock is much rarer (in our data) than not finding an available bicycle. So even though an under-supply of docking points will degrade the user experience and, in the long run, have a negative effect on system-use, from a practical standpoint we expect that docking point availability has a much weaker impact. Together these trends indicate that, in the short run and over the long run, system-use is much more likely to be affected by bike-availability than dock-availability; hence our analysis focuses on the former. It is theoretically possible to extend our model so that it includes docking point availability, but by doing so, we expect to find no significant differences than our current model despite much higher computational complexity.

This paper provides the first empirical estimates of user response to accessibility and availability in the context of bike-share systems. We build and estimate distinct mechanisms for the short and long-terms effects of availability and illustrate the use of our estimates in supporting a number of different system improvement efforts. Furthermore, the methodology developed here can be used in a variety of demand estimation contexts where products are spatially differentiated and with choice sets that change frequently. It is important to highlight that sufficing of nearby choices is a unique feature of spatially differentiated markets and might not be applicable to traditional demand estimation problems like choice of products. But given the proliferation of spatially differentiated markets in form of car-sharing, cab-hailing, food delivery platforms, these ideas might be generalizable and used in these other contexts.

In future work, we hope to address the limitations of this study. First, a more detailed data set on user starting locations would improve the precision of estimates of the effects we study. Second, a larger study comparing many cities could provide insight not only into how user preferences vary by city but also into how those preferences might be driven by different demographic and/or geographic factors. Such analyses could help bike-share systems fully deliver on their promise of transforming urban lifestyles.

References

G. Allon, A. Federgruen, and M. Pierson. How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management*, 13(4):489–507, 2011.

B. W. Alshalalfah and A. S. Shalaby. Case study: Relationship of walk access distance to transit with service, travel, and personal characteristics. *Journal of urban planning and development*, 133, no. 2 (114-118), 2007.

E. T. Anderson, G. J. Fitzsimons, and D. Simester. Measuring and mitigating the costs of stockouts. *Management Science*, 52(11):1751–1763, 2006.

R. Anupindi, M. Dada, and S. Gupta. Estimation of consumer demand with stock-out based substitution: An application to vending machine products. *Marketing Science*, 17(4):406–423, 1998.

E. Belavina, K. Girotra, and A. Kabra. Online grocery retail: Revenue models and environmental impact. *Management Science*, 2016.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica*, pages 841–890, 1995.

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

J. Büttner and T. Petersen. Optimising bike sharing in european cities-a handbook. 2011.

G. Cachon. Retail store density and the cost of greenhouse gas emissions. *Management Science*, 2014.

G. Cachon and C. Terwiesch. *Matching supply with demand*, volume 2. McGraw-Hill Singapore, 2009.

A. C. Cameron and P. K. Trivedi. Microeconometrics: Methods and applications. *Cambridge university press*, 2005.

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

D. W. Daddio. Maximizing bicycle sharing: An empirical analysis of capital bikeshare usage. Master's thesis, University of North Carolina at Chapel Hill, 2012.

P. Davis. Spatial competition in retail markets: Movie theaters. *The RAND Journal of Economics*, 37(4):964–982, 2006.

J.-P. Dubé, J. T. Fox, and C.-L. Su. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–2267, 2012.

A. El-Geneidy, M. Grimsrud, R. Wasfi, P. Tétreault, and J. Surprenant-Legault. New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service areas. *Transportation*, 41(1):193–210, 2014.

J. C. García-Palomares, J. Gutiérrez, and M. Latorre. Optimizing the location of stations in bike-sharing programs: A GIS approach. *Applied Geography*, 35(1), 2012.

D. K. George and C. H. Xia. Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research*, 211(1):198–207, 2011.

J. Gutiérrez, O. D. Cardozo, and J. C. García-Palomares. Transit ridership forecasting at station level: An approach based on distance-decay weighted regression. *Journal of Transport Geography*, 19(6) (1081-1092), 2011.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

S. Henderson, E. O'Mahony, and D. B. Shmoys. (Citi)Bike sharing. *Submitted*, 2016.

R. Lederman, M. Olivares, and G. Van Ryzin. Identifying competitors in markets with fixed product offerings. *Working Paper, SSRN 2374497*, 2014.

J. Li, S. Netessine, and S. Koulayev. Price to compete... with many: How to identify price competition in high dimensional space. 2015.

D. McFadden. The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328, 1974.

H. Minkowski. Gesammelte abhandlungen. *Chelsea Publishing Co., New York*, 1967.

A. Musalem, M. Olivares, E. T. Bradlow, C. Terwiesch, and D. Corsten. Structural estimation of the effect of out-of-stocks. *Management Science*, 56(7):1180–1197, 2010.

A. Nevo. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548, 2000.

A. Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

W. A. O'Neill, R. D. Ramsey, and J. Chou. Analysis of transit service areas using geographic information systems. *Transportation Research Record*, 1364, 1992.

S. O'Sullivan and J. Morrall. Walking distances to and from light-rail transit stations. *Transportation research record: journal of the transportation research board*, 1538:19–26, 1996.

J. Pancras, S. Sriram, and V. Kumar. Empirical investigation of retail expansion and cannibalization in a dynamic environment. *Management Science*, 2012.

P. Pendem and V. Deshpande. Maximizing ridership in bike-sharing systems using empirical data and stochastic models. *UNC Working Paper*, 2016.

W. J. Reilly. *The law of retail gravitation.* WJ Reilly, 1931.

D. Singhvi, S. Singhvi, P. I. Frazier, S. G. Henderson, E. O'Mahony, D. B. Shmoys, and D. B. Woodard. Predicting bike usage for new york citys bike sharing system. *In AAAI 2015 Workshop*

on Computational Sustainability., 2015.

A. Tangel. City bike-sharing programs hit speed bumps. *The Wall Street Journal*, July 9, 2014. URL `http://on.wsj.com/1oJWjuv`.

R. Thomadsen. The effect of ownership structure on prices in geographically differentiated industries. *RAND Journal of Economics*, pages 908–929, 2005.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

J. Ypma. Introduction to ipoptr: an R interface to Ipopt. 2010.

F. Zhao, L.-F. Chow, M.-T. Li, I. Ubaka, , and A. Gan. Forecasting transit walk accessibility: Regression model alternative to buffer method. *Transportation Research Record*, 1835:34–41, 2003.

Appendix A. De-Seasonalizing Weather Effects

We report here the effect of de-seasonalizing the system-use data with prevalent weather conditions.

The weather data is collected at a half-hourly frequency for the city of Paris, specifically the Temperature, Humidity, Wind Speed and "Conditions" (clear, mist, cloudy, etc.) from weatherbase.com. We incorporate each weather condition as dummy variables of ranges of their different expected impacts.

The Temperature variable is divided in three ranges of: ≤ 10, $(10, 30]$, and > 30; Humidity in the ranges of: ≤ 40, $(40, 80]$, and > 80; Wind Speed in the ranges of ≤ 20, $(20, 30]$, and > 30, while weather conditions are classified into clear, fog, heavy fog, heavy rain showers, light drizzle, light rain, light rain showers, light thunderstorms and rain, mist, mostly cloudy, overcast, partial fog, partly cloudy, rain, scattered clouds, shallow fog, heavy thunderstorms and rain, light thunderstorm, thunderstorm, thunderstorms and rain, light fog, and patches of fog.

The regression model is given by,

$$\ln(\Lambda_t) = \rho_0 + \vec{\rho}_1 Temp_t + \vec{\rho}_2 Humidity_t + \vec{\rho}_3 Wind\_Speed_t + \vec{\rho}_4 Condition_t + \rho_{h(t)} + \epsilon_t \qquad \text{(A.1)}$$

where t denotes a two minute interval, and $h(t)$ denotes half-hourly index within a day (48 in total) corresponding to t. Each of the $\vec{\rho}_1, \vec{\rho}_2, \vec{\rho}_3$, and $\vec{\rho}_4$ is a vector of effects of different levels (dummies) for our weather variables. Finally, we also include half-hourly fixed effects, $\rho_{h(t)}$.

Table 4 shows the impact of weather variables. Based on the estimated weather effects, station-use Λ_{ft} is de-seasonalized as follows.

Define the net weather effect at time t as,

$$\rho_t^w = \hat{\vec{\rho}}_1 Temp_t + \hat{\vec{\rho}}_2 Humidity_t + \hat{\vec{\rho}}_3 Wind\_Speed_t + \hat{\vec{\rho}}_4 Condition_t$$

	Value	Std. error
Temperature		
Range [,10)	0.000	
Range [10, 30]	0.301	(0.036)***
Range (30,]	0.211	(0.113).
Humidity		
Range (,40]	0.000	
Range (40, 80]	-0.114	(0.046)*
Range (80,]	-0.225	(0.054)***
Wind Speed		
Range (, 20]	0.000	
Range (20, 30]	0.060	(0.03)*
Range (30,)	-0.074	(0.257)
Conditions		
Clear	0.000	
Fog	-1.497	(0.161)
Heavy Fog	-1.064	(0.671)
Heavy Rain Showers	-0.134	(0.487)
Light Drizzle	0.097	(0.126)
Light Rain	-0.651	(0.046)
Light Rain Showers	-0.238	(0.084)**
Light Thunderstorms and Rain	-0.441	(0.306)
Mist	-1.029	(0.346)**
Mostly Cloudy	-0.142	(0.024)***
Overcast	-0.391	(0.085)***
Partial Fog	-1.161	(1.362)
Partly Cloudy	0.024	(0.043)
Rain	-1.804	(0.22)***
Scattered Clouds	-0.058	(0.039)
Shallow Fog	0.163	(0.141)
Heavy Thunderstorms and Rain	-1.073	(0.507)*
Light Thunderstorm	0.108	(0.367)
Thunderstorm	-0.263	(0.361)
Thunderstorms and Rain	0.125	(0.579)
Light Fog	-0.166	(0.52)
Patches of Fog	-0.594	(0.429)

*(p-value<0.05) **(p-value<0.01) ***(p-value<0.001)

TABLE 4. Weather Variables Effect

The de-seasonalized station-use, $\hat{\Lambda}_{ft}$ is given by, $\hat{\Lambda}_{ft} = \Lambda_{ft}/\exp(\rho_t^w)$. For all our further analysis in paper, we have used this de-seasonalized station-use $\hat{\Lambda}_{ft}$ in place of Λ_{ft}.

APPENDIX B. ADDITIONAL RESULTS

B.1. **Relevance Test for Instruments.** In Table 5 we test for the relevance of instruments. We apply the tests for both– the *station average* bike-availability and the *neighborhood average* bike-availability. Specifically, we compare the change in Adjusted R^2 when the proposed instruments are included as dependent variables to explain bike-availability at the *station* $\times$ *time-window* level (Eq.

Dependent Variable:	Station Bike-Availability			Neighbourhood Bike-Availability		
	(1)	(2)	(3)	(4)	(5)	(6)
TwXDistrict F.E.†	Yes	Yes	Yes	Yes	Yes	Yes
Focal Station's Neighbourhood Instruments		Yes	Yes		Yes	Yes
Neighbouring Station's Neighbourhood Instruments			Yes			Yes
Adjusted R^2	0.174	0.204	0.473	0.217	0.251	0.570
Number of observations	5676	5676	5676	5676	5676	5676

†TwXDistrict F.E. sum upto 0 for each District like in structural model

TABLE 5. Relevance test of Instruments

B.1). We divide our instruments in two sets, the ones based on the neighborhood characteristics of the focal station and the ones based on neighboring station's neighborhood characteristics (or equivalently whether or not c is 0 in $V_{fwj}(a,b,c,d)$). The population density variable $pd(L_f)$ is included in the focal station's neighborhood characteristics.

$$ba_{fw} = \eta_0 + \eta_1 \cdot pd(L_f) + \vec{\eta}_2 \cdot \vec{V}_{fw}(a,b,cd)) \tag{B.1}$$

where (a,b,c,d) take values of $(0,25,0,25)$, $(25,50,0,50)$, $(50,100,0,100)$, $(0,100,0,100)$, $(100,300,0,300)$, $(300,500,0,500)$, $(0,100,100,300)$, and $(0,100,300,500)$. A similar regression is run with $naba_{fw}$ as dependent variable.

We see that both the neighborhood characteristics of the focal station, and those of the neighboring stations are effective instruments.

B.2. **Estimates from the Density Model.** Recall the spatial density distribution we have used is given by,

$$P_w^D(L_i;\alpha) = \alpha_0 + \alpha_1 \cdot naba_{L_i,w} + \alpha_2 \cdot pd(L_i) + \vec{\alpha}_{3,w} \cdot \vec{V}_w(L_i).$$

We report the estimates (α) of our density model in Table 6. We observe that bike-availability, residential users, metro stations, supermarkets (day time only) and cafes are the major contributors of bike-share use.

B.3. **Robustness of the Distance Disutility Function.** Table 7 reports the estimates with a variety of alternate distance disutility functions– with alternate kink points in the piecewise linear form– 100, 200, 250, 275, 325, 350 meters instead of 300 meters in the original model, a simple linear function, and a quadratic function. Section 8.2 provides a discussion of these estimates.

	Value	Std. error
Residential Users	0.004	(0.000)***
Metro Stations	0.499	(0.211)*
Intercept	0.000	0.000
Bike-Availability	0.020	(0.002)***
Non Night Hours		
Store	0.000	(0.000)
Food	0.005	(0.033)
Restaurant	0.000	(0.000)
Bar	0.000	(0.000)
Lodging	0.018	(0.017)
Cafe	0.270	(0.039)***
Supermarket	0.391	(0.028)***
University	0.236	(0.053)***
Park	0.000	(0.000)
Museum	0.278	(0.08)***
Library	0.333	(0.183)
Tourist Locations	4.198	(1.099)***
Movie-theater	0.000	(0.000)
Shopping-mall	0.000	(0.000)
Other points of interest	0.005	(0.005)
Tram Line 3a	0.000	(0.000)
Tram Line 3b	0.000	(0.000)
Night Hours		
Store	0.000	(0.000)
Food	0.000	(0.000)
Restaurant	0.000	(0.000)
Bar	0.694	(0.148)***
Lodging	0.139	(0.099)
Cafe	1.316	(0.223)***
Supermarket	0.000	(0.000)
University	0.036	(0.224)
Park	0.000	(0.000)
Museum	0.000	(0.000)
Library	0.000	(0.000)
Tourist Locations	6.253	(1.369)***
Movie-theater	0.000	(0.000)
Shopping-mall	0.000	(0.000)
Other points of interest	0.084	(0.031)**

*(p-value<0.05) **(p-value<0.01) ***(p-value<0.001)

TABLE 6. Density Variables Effect

APPENDIX C. ESTIMATION DETAILS

C.1. **Estimation Procedure.**

		Primary variables							
		Walking Distance (until kink)	Walking Distance (after kink)	Bike-Availability (naba)	10% increase in Station Density	10% increase in Bike-Availability		Number of observations	$\chi^2(df)$
						Short-term	Total		
(1)	Original Estimates	-2.700 (0.495)***	-15.734 (3.043)***	0.005 (0.001)***	5.090%	9.399%	12.293%	39,302	0.049 (136)
(2)	Kink at 100mts	-0.148 (2.115)	-7.186 (0.558)***	0.004 (0.000)***	4.830%	9.427%	12.021%	39,302	0.050 (136)
(3)	Kink at 200mts	-2.220 (0.779)**	-9.058 (0.988)***	0.005 (0.000)***	5.010%	9.409%	12.154%	39,302	0.050 (136)
(4)	Kink at 250mts	-2.342 (0.595)***	-11.472 (1.563)***	0.005 (0.000)***	5.103%	9.401%	12.266%	39,302	0.050 (136)
(5)	Kink at 275mts	-2.628 (0.535)***	-13.094 (2.085)***	0.005 (0.001)***	5.103%	9.400%	12.280%	39,302	0.049 (136)
(6)	Kink at 325mts	-2.676 (0.468)***	-19.579 (4.799)***	0.005 (0.001)***	5.009%	9.398%	12.286%	39,302	0.048 (136)
(7)	Kink at 350mts	-2.639 (0.455)***	-25.513 (8.619)**	0.005 (0.001)***	4.817%	9.397%	12.246%	39,302	0.048 (136)
		Walking Distance	Walking Distance Square	Bike-Availability (naba)					
(8)	Linear Model	-6.459 (0.413)***		0.004 (0.000)***	4.560%	9.435%	11.970%	39,302	0.050 (136)
(9)	Quadratic Model	2.158 (1.468)	-16.099 (3.602)***	0.005 (0.001)***	5.193%	9.422%	12.245%	39,302	0.049 (136)

*(p-value<0.05) **(p-value<0.01) ***(p-value<0.001)

TABLE 7. Robustness of Distance Effect

Estimation. The estimation procedure introduced in section 5.3 is as follows. The set of moment conditions are given by,

$$E\left[Z_{fw}\sigma_{fwv_f}\xi_{fwv_f}\left(\theta^*\right)\right] = 0 \quad \text{, and}$$
$$E_w\left[\gamma^*_{w\times di}\right] = 0 \text{ for } \forall di$$

The constraints used to determine values of all $\xi_{\cdot w}$ $\left(\xi'_{fwv_f}s\right)$ are,

$$\lambda_{fwv_f}\left(\theta, \xi_{\cdot w}\right) = \Lambda_{fwv_f} \quad \forall f, w, v_f\,.$$

We rewrite above moment conditions and constraints in a GMM formulation in Eq's. C.1 below.

The moment conditions vector given by $G(\theta, \xi)$ is,

$$\begin{aligned} G(\theta, \xi) &= \frac{1}{N} \sum_{f,w,v_f} Z_{fw} \cdot \sigma_{fwv_f} \cdot \xi_{fwv_f} \\ &= \frac{1}{N} Z^T \Sigma \xi \end{aligned}$$

where Z, Σ, and ξ are matrix and vector notations for Z_{fw}, σ_{f,w,v_f} and ξ_{fwv_f} respectively over all observations. Note that Σ is a diagonal matrix and N is number of observations f, w, v_f.

We also introduce a change of variable, so that the moment conditions are treated as additional parameters as suggested by Dubé et al. [2012], which makes the hessian matrix sparse.

The GMM estimator is given by,

$$\hat{\theta}^* = \arg\min_{\theta} \eta^{'} A \eta \tag{C.1}$$

$$\text{s.t.}$$

$$\sum_{w} \gamma_{w \times di} = 0 \quad \forall di$$

$$\lambda_{fwv_f}(\theta, \xi_{.w.}) = \Lambda_{fwv_f}$$

$$G(\theta, \xi) = \eta$$

$$\int_{w} \int_{L_i} P_w^D(L_i; \theta)\, dL_i = T^D$$

where A is the GMM weighing matrix.

Note that each computation of λ_{fwv_f} as per Eq. 5.5 involves integrating over the spatial density of users. We divide the density elements into two components, the discrete density elements, $\vec{V}_w$, such as metro stations, movie theaters, etc. and the continuous density elements which are the population density, bike-availability and intercept term. The integration over latter density elements is performed numerically. We discretize the physical area of the city of Paris into a grid composed of squares with length $\mathscr{D}$ meters; we consider the center of each such square to be a point mass of users. Predicted use is then

$$\begin{aligned} \lambda_{fwv_f}(\theta, \xi_{.w.}) = &\sum_{i \in Points\_of\_Interests} p_{ifwv_f}(\theta, \xi_{.w.}) \cdot \left(\vec{\alpha}_{3,w} \cdot \vec{V}_w(L_i) \right) + \\ &\sum_{j \in Grid(\mathscr{D})} p_{jfwv_f}(\theta, \xi_{.w.}) \cdot \left(\alpha_0 + \alpha_1 \cdot naba_{L_j,w} + \alpha_2 \cdot pd(L_j) \right) \cdot \mathscr{D}^2, \end{aligned}$$

where $\mathscr{D}^2$ is the area of each grid square.

The simplest way of estimating our model would be to search over the parameters θ for values that provide the best fit. This would require a search over a space with as many dimensions as parameters (including numerous fixed-effects parameters), resulting in several search iterations each of which is computationally expensive. We instead estimate our model using a process that relies on some parameters (all except density model parameters $\vec{\alpha}$ and distance coefficient β_d) entering our model in a "user-location–agnostic" way (Berry et al. [1995]). We thus group our parameters in two classes, first as $\theta_1 = (\alpha,\, \beta_d)$, and the parameters that are "linear" (in ξ_{fwv_f}) as $\theta_2 = (\beta_0, \vec{\gamma})$.

We rewrite the p_{ifwv_f} and λ_{fwv_f} in terms of composite terms δ_{fwv_f} :

$$\delta_{fwv_f} = \beta_0 + \gamma_{w\times di(f)} + \xi_{fwv_f}$$

The user choice probabilities and station-use are now written as a function of θ_1 and δ. The user choice probabilities p_{ifwv_f} is now given as,

$$p_{ifwv_f}\left(\theta_1, \delta_{\cdot w\cdot}\right) = \frac{\exp\left(h\left(\beta_d; d\left(L_i, L_f\right)\right) + \delta_{fwv_f}\right)}{1 + \sum_{g\in N_i\cap S_{v_f}} \exp\left(h\left(\beta_d; d\left(L_i, L_g\right)\right) + \delta_{gwv_f}\right)},$$

where S_{v_f} denotes the set of stations with available bikes in state v_f. δ_{gwv_f} are estimated as

$$\hat{\delta}_{gwv_f} = \frac{\sum_{v_g} \sigma_{gwv_g}\delta_{gwv_g}}{\sum_{v_g} \sigma_{gwv_g}}.$$

Then station-use λ_{fwv_f} is given by,

$$\lambda_{fwv_f}\left(\theta_1, \delta_{\cdot w\cdot}\right) = \int_{L_i} p_{ifwv_f}\left(\theta, \delta_{\cdot w\cdot}\right) \cdot P_w^D\left(L_i; \alpha\right)\, dL_i\,.$$

The estimation process searches over values of θ_1 and δ. Given the values of θ_1 and δ, the values of coefficients θ_2 are determined non-iteratively from the closed-form expression below which follows from our moment conditions in Eq. C.1:

$$\hat{\theta}_2\left(\delta_{fwv_f}\right) = \left(\left(X_2^T\Sigma Z\right) A \left(Z^T\Sigma X_2\right)\right)^{-1} \left(\left(X_2^T\Sigma Z\right) A \left(Z^T\Sigma\right)\right) \delta_{fwv_f}. \tag{C.2}$$

where X_2 is the co-variate matrix corresponding to the equation, $\delta_{fwv_f} = \beta_0 + \gamma_{w,di(f)} + \xi_{fwv_f}$, consisting of an intercept column and *time-window*×*district* dummies. Thus, in each iteration, values of θ and ξ are obtained for given values of θ_1 and δ, and GMM objective function is computed.

In the first step of the GMM, we use $\left(Z^T\Sigma^2 Z\right)^{-1}$ as the weighing matrix A. We find the condition number of the matrix inversion step in Eq. C.2 to be low when using this weighing matrix, in

comparison to say an identity matrix which renders the conditions number quite high. This is analogous to the weighing matrix used in the 2SLS procedure.

Standard error. The variance estimate of $\hat{\theta}^*$ is given by,

$$V\left[\hat{\theta}^*\right] = \frac{1}{N}\left(\hat{G}^T\tilde{S}^{-1}\hat{G}\right)^{-1}$$

where, $\hat{G} = \frac{\partial G}{\partial \theta}|_{\theta=\hat{\theta}^*}$ is the first derivative of moment conditions G and $\tilde{S}^{-1}$ is the optimal GMM weighing matrix (sec 6.3.5. Cameron and Trivedi [2005]).

$\tilde{S}$ is given by,

$$\tilde{S} = \frac{1}{N}\sum_{f,w,v_f}\left(Z_{fw}\cdot\sigma_{fwv_f}\cdot\xi_{fwv_f}\right)\cdot\left(Z_{fw}\cdot\sigma_{fwv_f}\cdot\xi_{fwv_f}\right)^T.$$

Implementation Details. The procedure was implemented in R. The open-source package IPOPT (Interior Point Optimizer) (interfaced with R via "ipoptr" [Ypma, 2010]) was used for nonlinear optimization with constraints. The "ffdf" class in R was employed to accommodate the large scale of our data set. Even though we transformed our problem from the time domain to the local stockout state domain, computing the choice probabilities for each user, and then summing over them, was computationally expensive; the initial runtime was of the order of tens of days on a contemporary computer of the workstation class. Implementing the station-use computation function (Eq. 5.5) in C++ and then interfacing with R reduced the computation almost 100 times, to about 70 hours for the Paris data set.

C.2. **Comparison of the Computation Challenge.** We noted two modeling choices that result in extremely large computational burden necessitating us to devise our local-stockout state based transformation procedure. These were

1) the rapidly changing choice sets of stations available in the high frequency data-set, and

2) the spatial nature of the product which requires a fine grained spatial user density model.

Bruno and Vilcassim [2008], Conlon and Mortimer [2013] have shown that the estimates could be substantially biased because of not taking into account real-time availability information (as is implicitly the case in BLP and most applications of it). A fine grained user density model is also necessary in our context because the usage of bike-share systems tends to be quite local in nature, i.e. the the size of potential users could substantially change within a matter of 100-200 meters.

Comparison with work that has accounted for availability information (None in spatial context). Bruno and Vilcassim [2008] have access to only average product availability information. Assuming independent availabilities, Bruno and Vilcassim [2008] extend the BLP model to account for them. In presence of exact availability information, their model resembles our model in time-domain.

	Number of Products	Number of Time periods	(Full/Limited) Availability Information	Spatial Density
Bruno and Vilcassim [2008]	24	113	Yes	No
Conlon and Mortimer [2013]	44	44,458	Yes	No
Musalem et al. [2010]	24	15	Yes	No
Davis [2006]	607	7	No	Population Density
Thomadsen [2005]	103	1	No	Population Density
Allon et al. [2011]	388	1	No	Population Density
Our Model	946	22,743	Yes	Several Demand Sources

TABLE 8. Comparison of data size

Bruno and Vilcassim [2008] consider 24 products (as compared to 946 in our case) for 113 four week periods (as compared to over 22,000 in our case). Conlon and Mortimer [2013] consider 44 products in a vending machine application in 44,458 four-hour time periods. Musalem et al. [2010] consider 24 products for 15 days of data. These papers have users which were not spatially differentiated. There is heterogeneity in user tastes due to normally distributed random coefficients, however the number of draws required to aggregate over these heterogeneous users is much lower. For example, the supplementary code in Nevo [2000] uses 20 draws, and Dubé et al. [2012] use 1000 draws as compared to the more than 210,000 spatially heterogenous users in our case.

Comparison with work in spatial context (None account for availability). On the other hand, models that have accounted for spatially different users (Davis [2006], Thomadsen [2005], Allon et al. [2011]) have not accounted for product availabilities. Davis [2006] considers daily data for 607 theaters for a period of 7 days; Thomadsen [2005] considers 103 fast food locations in Santa Clara county with a single observation per location; Allon et al. [2011] considers 388 fast food outlets in Cook County with one observation per location. Note that in absence of sales data, Thomadsen [2005] and Allon et al. [2011] estimate parameters based on observed prices and other outlet characteristics.

The comparison with past work is summarized in Table 8. The comparison illustrates how the combination of rapidly changing choice sets and a fine grained user density model, in a relatively large scale data set, leads to an explosion in computational requirements.

C.3. **Validation in Simulated Datasets.** While the full validation of our approach remains the subject of a dedicated study that considers many alternate contexts, we provide a limited validation

Estimation Method	Moment Conditions	Limit on Choice Set	Number of States Considered	City Discretization
Data Generation	N.A.	None	N.A.	25 m
No Transformation or Computational Choices (Benchmark)	Time-Domain	None	N.A.	25 m
Local Stockout State	Local Stockout-State domain	Closest 4, $m_d = 4$	All	25 m
Top Local Stockout States	Local Stockout-State domain	Closest 4, $m_d = 4$	75% data	25 m

TABLE 9. Alternate Computational Models

in our context. Specifically, we validate the use of the local-stockout state transformation and our other computational choices- the use of top states and limits on the choice set– on smaller simulated datasets, where both our approach and the full approach (time-domain, all states, no limits on choice sets) are computationally feasible.

Data Generation. We created a number of small simulated datasets for demand at 30 stations around the city-center (Hôtel de Ville) for 50 two-minute time-intervals in the evening-rush time window. In particular, using the average bike-availabilities for each of these stations, we simulate the real time bike-availabilities for each station and 2 minute-interval, resulting in the relevant choice set information. The unobserved station-time characteristics ξ_{ft} are drawn from a Normal distribution with mean 0 and variance of 0.1. We combine these with the distance coefficients $(-2.700, -15.734)$ from our estimated structural model to obtain the choice probabilities for infinitesimal users located at each point in the city. Finally, for each of the 6400 density-relevant grid-points (points of interest, transit, etc.), we generate the rate of a potential trip originating in a 2 minute interval using a lognormal distribution with mean 1 and variance 0.1.[16] The choice probabilities combined with this density model that incorporates point of interest location data give us the simulated station-level demand for our station network for each two-minute interval.

Alternate Estimation Methods. We estimate our model on the simulated datasets in three ways illustrated in Table 9:

(1) The benchmark estimation procedure that uses the untransformed time-domain based moment conditions (Eq. 5.1) and places no limits on the choice set of customers. (2) Using the transformed local stockout state based conditions (Eq. 5.6) and imposing a consistent limit on the choice set of

[16]We also test alternate (more and less dispersed) distributions for the unobserved station-time characteristics and the rate of potential trips originating.

	No Transformation (Benchmark)		Local Stockout State		Top Local Stockout States	
Simulated Dataset #	Walking Distance (0-300mts)	Walking Distance (>300mts)	Walking Distance (0-300mts)	Walking Distance (>300mts)	Walking Distance (0-300mts)	Walking Distance (>300mts)
(1)	-3.487	-15.122	-3.341	-15.106	-3.486	-14.858
(2)	-3.026	-15.680	-2.987	-15.317	-2.673	-16.068
(3)	-2.598	-16.868	-2.594	-16.529	-2.376	-17.428
(4)	-3.116	-14.731	-3.145	-14.215	-2.781	-14.972
(5)	-2.723	-14.415	-2.725	-13.869	-2.417	-14.999
(6)	-2.392	-18.289	-2.263	-18.567	-2.204	-18.352
(7)	-3.275	-14.585	-3.251	-14.228	-3.270	-14.153
(8)	-2.729	-15.163	-2.728	-14.698	-2.547	-15.377
(9)	-2.151	-15.904	-2.077	-15.644	-2.139	-15.244
(10)	-3.712	-14.962	-3.756	-14.261	-3.701	-14.114
Mean	-2.921	-15.572	-2.887	-15.244	-2.760	-15.556
Std. Dev.	0.491	1.199	0.511	1.419	0.546	1.370
Computation Time	41534 sec		4869 sec		2709 sec	

TABLE 10. Simulation analysis results

the customer and (3) Approach (2) plus focusing on just enough local stockout states to cover 75% of the data for the typical station (the approach of this paper). The results for 10 sample datasets are reported in Table 10. We find that all three approaches recover estimates that are reasonably close to our seed estimates. Specifically, note that the time-domain based estimation procedure is able to recover the seed estimates from the demand model (see mean estimate from columns "Time-Domain" in Table 10) thus providing support for the moment conditions used in our estimation and generally validating our approach. The recovery of estimates by the local stockout state based estimation (columns "Local-stockout state" in Table 10) validates the use of our local stockout state transformation and the computational choices of limiting the choice set to 4 closest stations embedded therein. Finally, the estimates obtained by looking at just the top local stockout states (column "Top Local Stockout States" i.e. the approach in our paper) demonstrates that using the top states is sufficient.

Interestingly, while all three procedures recover the seed estimates, the computation burden of the third approach is an order of magnitude less than that of the untransformed approach, even in this small dataset. We expect the difference to be much larger in a dataset of the scale of our study. Taken together, while this analysis provides some validation of our approach– a full validation of such

transformations in other contexts remains the subject of a future study focused on further developing the methodological ideas here.