A Locational Demand Model for Bike-Sharing

Ang Xu
Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720

Chiwei Yan
Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720

Chong Yang Goh
Uber Technologies, Inc., San Francisco, CA 94158*

Patrick Jaillet
Department of Electrical Engineering and Computer Science and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139

Problem Definition: Micro-mobility systems (bike-sharing or scooter-sharing) have been widely adopted across the globe as a sustainable mode of urban transportation. To efficiently plan, operate and monitor such systems, it is crucial to understand the underlying rider demand—where riders come from and the rates of arrivals into the service area. They serve as key inputs for downstream decisions, including capacity planning, location optimization, and rebalancing. Estimating rider demand is nontrivial as most systems only keep track of trip data which is a biased representation of the underlying demand. ***Methodology/Results***: We develop a locational demand model to estimate rider demand only using trip and vehicle status data. We establish conditions under which our model is identifiable. In addition, we devise an expectation-maximization (EM) algorithm for efficient estimation with closed-form updates on location weights. To scale the estimation procedures, this EM algorithm is complemented with a location-discovery procedure that gradually adds new locations in the service region with large improvements to the likelihood. Experiments using both synthetic data and real data from a dockless bike-sharing system in the Seattle area demonstrate the accuracy and scalability of the model and its estimation algorithm. ***Managerial Implications***: Our theoretical results shed light on the quality of the estimates and guide the practical usage of this locational demand model. The model and its estimation algorithm equip municipal agencies and fleet operators with tools to effectively monitor service levels using daily operational data and assess demand shifts due to capacity changes at specific locations.

Key words: locational demand model, bike-sharing, expectation-maximization, location discovery.

1. Introduction

Bike-sharing services have been widely adopted to provide a sustainable mode of urban transportation. In the United States, as of July 2024, there are 54 docked bike-sharing systems operating 8,862 docking stations (Bureau of Transportation Statistics 2024). Perhaps more excitingly, since

* Formerly affiliated with Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, where the co-author's research was conducted.

its debut in Seattle in 2017, dockless bike-sharing and e-scooter systems have been quickly expanding coverage and gaining popularity due to its increased accessibility. As of July 2024, dockless bike-sharing systems serve 49 cities and e-scooters serve 130 cities in the United States (Bureau of Transportation Statistics 2024).

One critical aspect of monitoring and operating bike-sharing systems is to understand the rider demand and how accessible the current service is to different communities in the service region. For example, the city of Seattle has been actively monitoring the usage and accessibility of these services using trip data (see Seattle Department of Transportation 2022 for a comprehensive online dashboard). In addition, these systems often experience supply and demand imbalances that require careful bike allocation and rebalancing. The success of these operations crucially depends on the operator's ability to estimate the ridership demand accurately over the planning horizon. These motivate the topic of our paper.

Demand estimation in bike-sharing systems is non-trivial due to the following difficulties. First, the operator often does not know the exact location of riders but only observes the booking data which indicates the location and time a bike is reserved and picked up. Although technically speaking, rider location data can be accessed through GPS on riders' phones, to the best of the authors' knowledge, it is not an industry standard to record these data (Open Mobility Foundation 2022). These rider location data can be harder to collect and is subject to errors themselves—for example, the place rider is checking her phone can be different from where she departs to pick up a bike. Second, the observed trip locations and trip counts can form biased estimates for the underlying demand—the booking location is likely not the actual rider location and a rider may choose not to book or switch to other transportation services if there are no bikes available in the vicinity of her location, and in such a case this demand will be censored in the observed trip data.

A standard approach for modeling such demand is to use a choice model, which specifies the likelihood that a customer makes a certain choice when presented with a set of alternatives. Analogous to the retail settings to which such models are commonly applied, we can view each bike in the system as a product to be considered among a set of available bikes. One characteristic of a bike-sharing service is that products are horizontally differentiated—they are almost the same in terms of price and quality and are largely differentiated by their proximity to the rider.

We primarily focus on this locational feature to build our demand model which is applicable to both docked and dockless systems. In particular, we consider a model where riders arrive within a set of locations inside the service region according to a Poisson process. Upon arrival, a rider makes a choice of which bike to pick (or leaves the system) based on her walking distance to all available

bikes, governed by a *general* choice model. Our goal is to estimate the set of rider locations and their corresponding arrival rates, using only booking and vehicle location and status data.

We summarize our key contributions as follows. We first study the statistical properties of the demand model. We give general conditions under which our model is identifiable and our estimator is consistent. We also discuss specific identifiability properties when riders make bike choices according to specific choice models including a multinomial logit model and a model based on distance ranking, under both docked and dockless settings. Second, we derive an efficient expectation-maximization (EM) algorithm with closed-form updates on location weights for the estimation problem. To scale this algorithm and handle situations when the set of potential rider locations is large or even not known a priori, we develop a location-discovery procedure that iteratively explores and adds new rider locations in the service region. Lastly, we implement our algorithms on a set of synthetic data and real dockless bike-sharing data in the Seattle area. These experiments demonstrate the scalability and accuracy of our proposed demand model and its estimation algorithm.

Our paper is organized as follows. In Section 2, we discuss relevant literature. In Section 3, we describe the data generating process, our demand model, and discuss its identifiability. In Section 4, we introduce the EM algorithm and the location-discovery procedure. We report the performance of our demand model and its estimation algorithm on an extensive set of numerical experiments in Section 5. All proofs and auxiliary results are provided in the Online Appendix. Data and code to reproduce all experiments in the paper can be found here.

2. Literature Review

In this section, we briefly review related work. There have been numerous works that analyze bike-sharing usage by incorporating data from heterogeneous sources (see, e.g., Rixey 2013, Singhvi et al. 2015, El-Assi et al. 2017), among them being demographic characteristics, built environment factors, weather and usage data of other connecting transportation services (see El-Assi et al. 2017 for a summary of the recent literature and references therein). In contrast, our work is more related to the literature of structurally understanding demand censoring and substitution due to the service (un)availability of bike-sharing systems. Close to our work are O'Mahony and Shmoys (2015), Kabra et al. (2019), Freund et al. (2019a), Freund et al. (2019b) and He et al. (2021). Although the main focuses of O'Mahony and Shmoys (2015), Freund et al. (2019a) and Freund et al. (2019b) are on improving the operations of bike allocation and transshipment, they all emphasize the importance of demand correction. Specifically, O'Mahony and Shmoys (2015) and Freund et al. (2019b) filter out time periods when the dock runs out of bikes to correct for demand

censoring. Kabra et al. (2019) focus primarily on the question of how the accessibility and the availability of a docked bike-sharing service impact ridership. To that end, the authors proposed a structural demand model in which the rider arrival rate at a location is assumed to vary with several covariates, such as the local population density and metro usage. The pick-up model is constructed using a logit choice model where the utility includes a piecewise linear function of walking distance with a break-point at 300 meters. He et al. (2021) study the network effect of bike-sharing demand, i.e., riders choose to pick up a bike because both the origin and destination docks are attractive. They develop an instrumental variable method to tackle the endogeneity issue of choice set in estimating demand and apply the method to a London bike share system to estimate the rider demand for network products. In contrast to previous papers that often use a parametric model to dictate rider arrival patterns across the service region, our paper differs by developing a model that features a nonparametric component capturing the locational aspect of rider arrivals. In particular, rider arrival locations do not have a fixed structure but are gradually discovered within the estimation procedure. A parametric model may perform well in generalizing demand patterns to new service areas using covariates (e.g., predicting demand in a new region prior to launch). In contrast, the proposed location-based nonparametric model is more effective for analyzing demand within an existing service area, such as monitoring service levels or assessing demand shifts resulting from changes in capacity at specific locations.

Specifically, we look at a parsimonious and operational setting where the rider demands are inferred solely based on trip and vehicle status data that fleet operators collect in their daily operations. We reveal statistical properties of estimating rider arrival locations and intensities and develop scalable estimation procedures that are applicable to both docked and dockless systems. These add to the toolkit of fleet operators or municipal agencies to effectively plan, operate and monitor micro-mobility systems. Perhaps the closest work to ours is Paul et al. (2023) who develop an expectation-maximization (EM) algorithm (similar to our all-in algorithm, Algorithm 1) to estimate censored spatial-temporal demand for shared micromobility services in Providence, Rhode Island. Both works are developed independently.

Our work is also closely related to the abundant list of literature on estimating the demand of substitutable products using choice models based on possibly censored transaction data (Anupindi et al. 1998, Talluri and van Ryzin 2004, Vulcano et al. 2012, Newman et al. 2014, Abdallah and Vulcano 2020). Assuming customer demand follows a multinomial logit (MNL) choice model, Vulcano et al. (2012) propose an EM method to estimate customers' demand for substitutable products from (censored) sales transaction data. Newman et al. (2014) develop a two-step strategy that is served

as an alternative to the EM method for parameter estimation. Their approach entails breaking down the log-likelihood function into separate marginal and conditional components. To improve upon these two methods, Abdallah and Vulcano (2020) propose a minorization-maximization (MM) algorithm that can achieve a unique global maximum of the log-likelihood function under certain data requirements of the transaction data. However, as we will discuss later, it is less efficient in our setting due to a lack of closed-form updates. More recently, Steeneck et al. (2022) utilized a nested EM algorithm to estimate lost sales for retailers with uncertain on-shelf availability. van Ryzin and Vulcano (2014) propose a market discovery algorithm to estimate customer demand based on ranking preferences. The algorithm starts with a small set of customer types and enlarges the set by iteratively generating new customer type (ranking preference) that increase the likelihood value. We draw inspiration from this algorithm when developing our location-discovery procedure, though the underlying generative process and the structure of the subproblem to generate new customer type (rider location in our context) are vastly different.

Structurally speaking, our demand model and the EM algorithm share some similarities with the latent-class logit (LCL) model discussed in Bhat (1997) and Greene and Hensher (2003), or the mixed multinomial logit model (MMNL) proposed in McFadden and Train (2000). A latent class or a customer type in our setting corresponds to a particular rider location, but unlike LCL or MMNL, we do not restrict rider choice behaviors to follow MNL models in each class. Another important distinguishing factor of our model is that the features of different alternatives (walking distances to different bikes) depend on the latent class (rider location), which is assumed to be invariant across latent classes in LCL or different customer types in MMNL. This renders existing identifiability results (e.g., Grün and Leisch 2008) or enhanced estimation algorithms (e.g., Jagabathula et al. 2020 and Cho et al. 2023) developed for LCL or MMNL not readily applicable in our setting.

3. Model and Preliminaries

In this section, we introduce our model and preliminaries. We first discuss our data generating process and describe the observed data from an operator's perspective. Then we derive the likelihood function of our statistical model. We conclude the section by discussing its identifiability.

Data Generating Process. We consider a set of rider locations $\mathcal{L} := \{1, \cdots, L\}$ distributed over a bounded space $\mathcal{P} \subset \mathbb{R}^2$ where potential riders come from. We consider a total length of arrival period T and time runs continuously. We assume that riders arrive at the area according to a Poisson process with a total rate λ. For each arriving rider, its arrival location follows a multinomial distribution with probabilities $\boldsymbol{w} = \{w_1, \cdots, w_L\}$ satisfying $\sum_{l \in \mathcal{L}} w_l = 1$. One can think of this arrival process as one that is dedicated to a particular hour of day or week. We define

$\mathcal{B} := \{1, \cdots, B\}$ as the set of bikes in the system. At time $t \in [0, T]$, we define the coordinate of bike $b \in \mathcal{B}$ as $(x_{b,t}, y_{b,t})$. Let $z_{b,t} = 1$ if bike b is available to be booked at time t, and $z_{b,t} = 0$ if bike b is not available at time t because it is booked or occupied. Let $\mathcal{B}_t := \{j \in \mathcal{B} : z_{b,t} = 1\} \subset \mathcal{B}$ be the set of bikes that are available for booking at time t.

We start the discussion with a dockless setting. When a rider arrives at a location $l \in \mathcal{L}$ at time t, she is presented with a *bike pattern* $S_t := \{(x_{b,t}, y_{b,t})\}_{b \in \mathcal{B}_t}$ which contains the coordinates of all available bikes at time t. Let p_{l,b,S_t} be the probability that a rider at location l chooses bike $b \in \mathcal{B}_t$ at time t. In addition, let $p_{l,0,S_t}$ denote the probability that a rider at location l leaves without choosing a bike at time t. We require that $\sum_{b \in \mathcal{B}_t \cup \{0\}} p_{l,b,S_t} = 1$. One example of such riders' choice behavior is a multinomial logit (MNL) choice model. Walking distance to the bike is arguably the most important feature. Let d_{l,b,S_t} be the distance from rider location $l \in \mathcal{L}$ to bike $b \in \mathcal{B}$ in bike pattern S_t. For each rider location $l \in \mathcal{L}$ and available bike $b \in \mathcal{B}_t$,

$$p_{l,b,S_t} = \frac{\exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}{1 + \sum_{b \in \mathcal{B}_t} \exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}, \quad p_{l,0,S_t} = \frac{1}{1 + \sum_{b \in \mathcal{B}_t} \exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}, \tag{1}$$

where $\beta_{0,l} \in \mathbb{R}, \beta_{1,l} \in \mathbb{R}_{<0}$ are parameters measuring rider tolerance for walking distance at different locations. Another example is a distance-ranking choice model in which the preference is determined by the ranking of distances to the available bikes. In specific,

$$p_{l,b,S_t} = \frac{\mathbf{1}\{d_{l,b,S_t} \le d_{l,b',S_t},\ \forall b' \in \mathcal{B}_t,\ d_{l,b,S_t} \le \bar{r}_l\}}{\left|\left\{b' \in \mathcal{B}_t : d_{l,b',S_t} = \min\{d_{l,b'',S_t} : b'' \in \mathcal{B}_t\}\right\}\right|}, \quad p_{l,0,S_t} = \mathbf{1}\{d_{l,b,S_t} > \bar{r}_l,\ \forall b \in \mathcal{B}_t\}, \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $\bar{r}_l$ is the consideration radius at rider location l which specifies how far a rider is willing to walk to a bike. In words, a rider at location l chooses to pick up an available bike $b \in \mathcal{B}_t$ if and only if bike b is the closest available bike to rider location l and its distance does not exceed $\bar{r}_l$. When a tie occurs, each bike with the shortest distance within the consideration radius has the same probability to be booked by a rider. More generally, we assume that the choice probability $p_{l,b,S_t}(\boldsymbol{\beta}_l)$ for riders at location l choosing bike b at time t is a function of location-dependent parameters $\boldsymbol{\beta}_l$ as well as corresponding features such as distances to each available bike. We put $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_L\}$. For notational brevity, we simply write p_{l,b,S_t} and suppress its dependence on $\boldsymbol{\beta}_l$ going forward.

Data and Observations. The data available to the operator consists of records of bookings and returns as well as the statuses and locations of all bikes in real-time. The statuses of a bike include "available" or "occupied". When a rider books a bike, the status changes from "available" to "occupied". When she drops off the bike at her destination, the status changes back to "available".

As a key feature of our problem, the operator cannot observe riders' arriving locations. As a consequence, it cannot distinguish between no arrival with a rider arriving without choosing a bike.

Dock-based and Hybrid Systems. Our model can be adapted to a dock-based system (or a hybrid system including both dock-based and dockless bikes) where bikes have to be picked up and returned to docks. We can accommodate this by similarly defining a *dock pattern* as the set of available docks at some time (a dock is available if it contains at least one available bike). Riders at different locations choose docks to pick up their bikes. Features that influence their choices can be, for example, distances to the available docks or number of bikes in each available dock. We refer readers to Appendix A for more details.

Non-stationary Arrival Rate. Our results can be extended to a setting where the arrival rate varies over time. Specifically, following the setups of Vulcano et al. (2012) and Abdallah and Vulcano (2020), we can segment the arrival period into multiple intervals, each with its own constant arrival rate. As we will derive later in equation (4), one can show that the arrival rate estimator takes a similar form and the rest of the estimation procedure requires minimum change.

3.1. Model Formulation

Given the above data generating process and observations, our goal is to estimate the following quantities: (1) riders' total arrival rate λ into the service region; (2) the probability/weight vector $\boldsymbol{w}$ distributed over the set of rider locations $\mathcal{L}$; (3) the parameters $\boldsymbol{\beta}$ that describe riders' choice behavior. We proceed to derive the likelihood of a given set of observations. To simplify the notation, define the *observed* arrival rate

$$\tilde{\lambda}(t, \boldsymbol{\beta}) := \lambda \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t}\right).$$

This quantity takes out the portion of riders who choose not to pick up any bike upon arrival from the total arrival rate λ. Suppose there are N bookings in total during the arrival period $[0, T]$. We denote the sequence $\boldsymbol{t} := \{t_n\}_{n=1}^N$ where $t_1, \ldots, t_N \in [0, T]$, $t_1 < t_2 < \cdots < t_N$ as the time epochs that bookings occur and $\boldsymbol{b} := \{b_n\}_{n=1}^N$ where $b_1, \ldots, b_N \in \mathcal{B}$ as the bikes booked by the riders in the corresponding booking times. Define $t_0 = 0$ and $t_{N+1} = T$. Consider a short enough time period $\delta > 0$ around each booking times $\{t_n\}_{n=1}^N$ such that bike patterns S_t do not change during these intervals $t \in [t_n, t_n + \delta]$, $n = 1, \cdots, N$. The incomplete data log-likelihood function is then given by

$$\begin{aligned} l_I(\boldsymbol{w}, \lambda, \boldsymbol{\beta}) := \lim_{\delta \downarrow 0} \log \Bigg(& \prod_{n=0}^{N} \mathbb{P}(\text{no rider books bikes from } t_n + \delta \text{ to } t_{n+1}) \\ & \cdot \prod_{n=1}^{N} \frac{\mathbb{P}(\text{a rider books bike } b_n \text{ from } t_n \text{ to } t_n + \delta)}{\delta} \Bigg) \end{aligned}$$

$$
\begin{aligned}
&= \lim_{\delta \downarrow 0} \left(\sum_{n=0}^{N} \log \mathbb{P}(\text{no rider books bikes from } t_n + \delta \text{ to } t_{n+1}) \right. \\
&\left. + \sum_{n=1}^{N} \log \mathbb{P}(\text{a rider books bike } b_n \text{ from } t_n \text{ to } t_n + \delta) - \sum_{n=1}^{N} \log(\delta) \right) \\
&= \lim_{\delta \downarrow 0} \left(\sum_{n=0}^{N} \log \Bigg(\underbrace{\exp\left(- \int_{t_n+\delta}^{t_{n+1}} \tilde{\lambda}(t, \boldsymbol{\beta}) dt \right)}_{\substack{\text{Prob. of no booking} \\ \text{in } [t_n, t_n + \delta]}} \Bigg) \right. \\
&\left. + \sum_{n=1}^{N} \log \Bigg(\underbrace{\int_{t_n}^{t_n+\delta} \tilde{\lambda}(t, \boldsymbol{\beta}) dt \cdot \exp\left(- \int_{t_n}^{t_n+\delta} \tilde{\lambda}(t, \boldsymbol{\beta}) dt \right)}_{\substack{\text{Prob. of one booking} \\ \text{in } [t_n, t_n + \delta]}} \cdot \underbrace{\frac{\sum_{l \in \mathcal{L}} w_l p_{l, b_n, S_{t_n}}}{1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_{t_n}}}}_{\substack{\text{Prob. of booking } b_n \text{ conditional} \\ \text{on one booking in } [t_n, t_n + \delta]}} \Bigg) \right. \\
&\left. - \sum_{n=1}^{N} \log(\delta) \right) \\
&= - \int_0^T \tilde{\lambda}(t, \boldsymbol{\beta}) dt + \sum_{n=1}^{N} \log \tilde{\lambda}(t_n, \boldsymbol{\beta}) + \sum_{n=1}^{N} \log \frac{\sum_{l \in \mathcal{L}} w_l p_{l, b_n, S_{t_n}}}{1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_{t_n}}}. \qquad (3)
\end{aligned}
$$

Equation (3) holds because $\int_{t_n}^{t_n+\delta} \tilde{\lambda}(t, \boldsymbol{\beta}) dt = \tilde{\lambda}(t_n, \boldsymbol{\beta}) \delta$ as S_t does not change within $[t_n, t_n + \delta]$. Taking the first-order condition with respect to λ, it can be seen that the total arrival rate λ has a unique closed-form maximizer.

$$
- \int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_t} \right) dt + \sum_{n=1}^{N} \frac{1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_{t_n}}}{\tilde{\lambda}(t_n, \boldsymbol{\beta})} = 0 \;\Rightarrow\; \lambda = \frac{N}{\int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_t} \right) dt}. \qquad (4)
$$

(As briefly mentioned before, to estimate an arrival rate during time interval $[\underline{t}, \overline{t}]$, its estimator takes a similar form $N_{[\underline{t}, \overline{t}]} / \int_{\underline{t}}^{\overline{t}} \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_t} \right) dt$, where $N_{[\underline{t}, \overline{t}]}$ denotes the number of bookings within $[\underline{t}, \overline{t}]$.) Plugging the closed form of λ into (3), we can rewrite the incomplete log-likelihood function as a function of $\boldsymbol{w}$ and $\boldsymbol{\beta}$ only,

$$
\begin{aligned}
l_I(\boldsymbol{w}, \boldsymbol{\beta}) :&= -N + \left(N \log N + \sum_{n=1}^{N} \log \left(\frac{1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_{t_n}}}{\int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_t} \right) dt} \right) \right) + \sum_{n=1}^{N} \log \frac{\sum_{l \in \mathcal{L}} w_l p_{l, b_n, S_{t_n}}}{1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_{t_n}}} \\
&= -N + N \log N - N \log \int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_t} \right) dt + \sum_{n=1}^{N} \log \left(\sum_{l \in \mathcal{L}} w_l p_{l, b_n, S_{t_n}} \right). \qquad (5)
\end{aligned}
$$

A key quantity involved in the above equation is $\int_0^T (1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_t}) dt$. Remarkably, $\int_0^T (1 - \sum_{l \in \mathcal{L}} w_l p_{l, 0, S_t}) dt / T$ measures the average percentage of riders who enter the system and pick up a bike. Typically, bike pattern S_t changes at events such as: (1) a rider books a bike; (2) a rider drops off her bike at her destination; (3) the operator relocates bikes. On the other hand, for the likelihood function (5) to be valid, there is no requirement on how bike patterns S_t change. Note that since S_t

does not change continuously over time, the integral over t in (5) can be reorganized as a finite sum. Assume that the pattern changes at time epochs $t'_1, \cdots, t'_Q$ where Q is the total number of changes within $[0,T]$. Then we have $\int_0^T \left(1-\sum_{l\in\mathcal{L}} w_l p_{l,0,S_t}\right) dt = \sum_{q=0}^{Q} \left(1-\sum_{l\in\mathcal{L}} w_l p_{l,0,S_{t_q}}\right)\left(t'_{q+1}-t'_q\right)$ with $t'_0=0$ and $t'_{Q+1}=T$.

Given $\boldsymbol{\beta}$, the log-likelihood function $l_I(\boldsymbol{w};\boldsymbol{\beta})$ in terms of $\boldsymbol{w}$ is non-concave in general. The non-concavity stems from the fact that the operator is not able to observe rider arrival locations. In the case of complete data where the operator is able to observe every rider's arrival and their arriving locations, the likelihood function $l_I(\boldsymbol{w};\boldsymbol{\beta})$ becomes strictly concave in $\boldsymbol{w}$, as we will show in Section 4. It is worth noting that $l_I(\boldsymbol{w};\boldsymbol{\beta})$, excluding the constant terms, can also be rearranged into a difference of two concave functions $g(\boldsymbol{w};\boldsymbol{\beta}) - h(\boldsymbol{w};\boldsymbol{\beta})$ where $g(\boldsymbol{w};\boldsymbol{\beta}) = \sum_{n=1}^{N} \log\left(\sum_{l\in\mathcal{L}} w_l p_{l,b_n,S_{t_n}}\right)$ and $h(\boldsymbol{w};\boldsymbol{\beta}) = N\log\int_0^T \left(1-\sum_{l\in\mathcal{L}} w_l p_{l,0,S_{t_n}}\right) dt$. We will discuss their algorithmic implications later in Sections 4 and 5.

3.2. Identifiability

In this subsection, we investigate the identifiability of our model. We start with the case where the choice model parameters $\boldsymbol{\beta}$ and a set of rider locations $\mathcal{L}$ are given. We first consider the following asymptotic regime when the length of the arrival period T gets large. We assume that there are K bike patterns (K is finite), which is denoted by $\mathcal{S} := \{S_1, \cdots, S_K\}$. We assume that $S_k \neq \emptyset$, $\forall k \in \{1,\cdots,K\}$. The long-run average fraction of time that we observe bike pattern S_k follows $\lim_{T\to\infty} \int_0^T \mathbf{1}(S_t = S_k)dt/T = \alpha_k > 0$ for all $k \in \{1,\cdots,K\}$ with $\alpha_1 + \cdots + \alpha_K = 1$. This setting typically models dock-based systems. For example, O'Mahony (2015) and Banerjee et al. (2022) use a continuous-time Markov chain (CTMC) to model state evolution where each state (bike pattern) is defined as the number of bikes in each dock. Then $\{\alpha_1, \cdots, \alpha_K\}$ can be thought of as the steady-state distribution of this CTMC. Let $\boldsymbol{w}^*$ denote the underlying true weight vector. We first establish the identifiability of our estimator $\hat{\boldsymbol{w}}$ (i.e., different parameter values correspond to different data-generating distributions), which is a necessary condition for consistency. Its proof relies on showing that the long-run average expected likelihood function $\lim_{T\to\infty} \mathbb{E}[l_I(\boldsymbol{w};\boldsymbol{\beta})]$ (the expectation is taken over bookings) has a unique maximizer at $\boldsymbol{w} = \boldsymbol{w}^*$, an equivalent condition for identifiability (see Lemma 5.35 in van der Vaart 2000). Note that even if $\boldsymbol{w}$ is not identifiable as a vector, it is possible that the weight of a particular location w_l is identifiable. In Online Appendix EC.1.1, we also provide additional results regarding such *partial identifiability* of location weights (see Proposition EC.1).

THEOREM 1 **(Identifiability)**. *Given $\boldsymbol{\beta}$, the location weights $\boldsymbol{w}$ are identifiable if the set of vectors $\left\{(p_{1,b,S_k}, \cdots, p_{L,b,S_k}) : b \in \mathcal{B},\ k \in \{1,\cdots,K\}\right\}$ spans the vector space $\mathbb{R}^L$. Moreover, the condition becomes necessary and sufficient when we have $w_l^* = 0$ for at most one $l \in \mathcal{L}$.*

Theorem 1 shows that a sufficient condition for identifiability is to have L linearly independent vectors from the vectors of riders' choice probabilities originating from different locations. This condition becomes necessary when the operator has relatively precise prior knowledge of where rider locations are—at most one location can be redundant in the candidate set $\mathcal{L}$. This result also implies that if riders' choice behavior depends only on distances, a minimum of L distinct bike locations across all bike patterns are required to possibly have identifiability of the estimator. In Online Appendix EC.1.1, we provide a few concrete examples to illustrate this result.

In most dockless systems, bikes are located at any place where parking is allowed. We thus provide a generalization to scenarios where bike patterns are continuously distributed over space. Let $(\mathcal{S}, \mathcal{F})$ be a measurable space of bike patterns. Let π and μ both be measures on $\mathcal{F}$. Furthermore, π is an invariant probability measure that is absolutely continuous with respect to μ, which measures the long-run average portion of each bike pattern, $\pi(A) = \lim_{T\to\infty}(1/T)\int_{t=0}^{T}\int_{\mathcal{S}}\mathbb{I}_A(S_t)d\mu(S_t)dt$, $\forall A \in \mathcal{F}$. We have Corollary 1 regarding the identifiability of the maximum likelihood estimator (MLE).

Corollary 1. *Given $\boldsymbol{\beta}$, the location weights $\boldsymbol{w}$ are identifiable if the set of vectors $\{(p_{1,b,S}, \cdots, p_{L,b,S}) : b \in \mathcal{B}, S \in \mathcal{S}'\}$ spans the vector space $\mathbb{R}^L$ for any $\mathcal{S}' \subseteq \mathcal{S}$ such that $\pi(\mathcal{S}') = 1$. The condition becomes necessary and sufficient when we have $w_l^* = 0$ for at most one $l \in \mathcal{L}$.*

We can strengthen our analysis in a dock-based system where riders make choices according to the distance-ranking model described in equation (2). First, we consider a simplified case where the service region is a one-dimensional line segment. Under mild smoothness conditions on the consideration radius over the service region, $\boldsymbol{w}$ is identifiable if for each location $l \in \mathcal{L}$, the sequence of docks within its consideration radius is distinct and uniquely ordered based on the distance to location l (see Theorem 2 in Appendix A for a formalism). In service regions of higher dimensions, a sufficient condition for the identifiability of $\hat{\boldsymbol{w}}$ is that each location has a unique ranking of the first two closest docks (or the only dock) within the consideration radius.

If the choice model parameters $\boldsymbol{\beta}$ are unknown, we need to consider the identifiability of both $\boldsymbol{w}$ and $\boldsymbol{\beta}$. For general choice models, a direct result following the proof of Theorem 1 is that $\boldsymbol{w}$ and $\boldsymbol{\beta}$ are identifiable if, and only if, a system of nonlinear equations (see equations (EC.4) in Online Appendix EC.1.2) has a unique solution that coincides with the underlying true parameters. Under an MNL choice model, we also give specific identifiability results there.

Finally, we comment that in general establishing consistency of the MLE requires not only identifiability but also uniform convergence of the log-likelihood function (see, e.g., Theorem 5.7 in van der Vaart 2000). Let $\hat{\boldsymbol{w}} \in \arg\max_{\boldsymbol{w}\in\Delta^L} l_I(\boldsymbol{w};\boldsymbol{\beta})$ be the corresponding MLE of the location weights given choice model parameters $\boldsymbol{\beta}$. In Proposition EC.2 of Online Appendix EC.1.1, we give

two sufficient conditions that ensure strong consistency, that is $\hat{\boldsymbol{w}} \to \boldsymbol{w}^*$ with probability one as $T \to \infty$.

4. Estimation Procedures

In this section, we show how one can obtain the set of rider locations and the corresponding MLE of their weights $\hat{\boldsymbol{w}}$, as well as the choice model parameters $\hat{\boldsymbol{\beta}}$. First, given a set of candidate rider locations, we show how one can computationally approach the MLE using an expectation-maximization (EM) algorithm with closed-form updates on $\boldsymbol{w}$ (Section 4.1). When rider locations are not known a priori, we introduce a location-discovery procedure to iteratively explore new rider locations until convergence (Section 4.2).

4.1. An Expectation-Maximization (EM) Algorithm

In this subsection, we develop an EM algorithm to optimize the location weight vector $\boldsymbol{w}$ as well as the choice model parameters $\boldsymbol{\beta}$. The EM algorithm, proposed by Dempster et al. (1977), is an iterative algorithm that consists of an *expectation* step (E-step) and a *maximization* step (M-step) in each iteration. The algorithm computes the conditional expected log-likelihood with respect to unobserved data (E-step) and then updates the estimates through maximizing the expected log-likelihood in the E-step (M-step). This procedure is repeated until convergence. In our case, the *observed* data contains sequences of booking times $\boldsymbol{t}$, booked bikes $\boldsymbol{b}$, and bike patterns S_t at any time $t \in [0, T]$. The *unobserved* data includes: (1) the arrival times of riders who leave without choosing a bike (unobserved riders); and (2) the arrival locations of riders. Define the total number of arrivals $\tilde{N} := N + N'$ where N' is the number of unobserved riders. We first consider the *complete data log-likelihood function*, which is the log-likelihood function derived under the full observation of all arrival times, arrival locations and booked bikes. Let $\tilde{\boldsymbol{l}} := \{\tilde{l}_n\}_{n=1}^{\tilde{N}}$ be the arrival locations of all riders in the sequence, $\tilde{\boldsymbol{t}} := \{\tilde{t}_n\}_{n=1}^{\tilde{N}}$ be the arrival times and $\tilde{\boldsymbol{b}} := \{\tilde{b}_n\}_{n=1}^{\tilde{N}}$, $\tilde{b}_n \in \mathcal{B} \cup \{0\}$ be the sequence of bikes booked by the riders. Here, $\tilde{b}_n = 0$ means that the n^{th} rider arrives without picking up a bike. Using these notations, we can write down the complete data log-likelihood function by following a similar procedure as in equation (3).

$$
\begin{aligned}
l_C(\boldsymbol{w}, \lambda, \boldsymbol{\beta}) = \lim_{\delta \downarrow 0} \log \Bigg(& \prod_{n=0}^{\tilde{N}} \mathbb{P}(\text{no rider books bikes from } t_n + \delta \text{ to } t_{n+1}) \\
& \cdot \prod_{n=1}^{\tilde{N}} \frac{\mathbb{P}(\text{a rider arriving at } \tilde{l}_n \text{ books bike } \tilde{b}_n \text{ (or leave) from } t_n \text{ to } t_n + \delta)}{\delta} \Bigg) \\
= \log\left(\exp\left(-\lambda T\right)\right) & + \sum_{n=1}^{\tilde{N}} \log\left(\lambda w_{\tilde{l}_n} p_{\tilde{l}_n, \tilde{b}_n, S_{\tilde{t}_n}}\right)
\end{aligned}
$$

$$= -\lambda T + \tilde{N}\log\lambda + \sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) + \sum_{n=1}^{\tilde{N}} \log\left(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}\right).$$

Similarly, the arrival rate λ can be substituted with the MLE $\hat{\lambda} = \tilde{N}/T$. This simplifies the log-likelihood function to $l_C(\boldsymbol{w},\boldsymbol{\beta})$ which only depends on the location weights $\boldsymbol{w}$ and the choice model parameters $\boldsymbol{\beta}$,

$$l_C(\boldsymbol{w},\boldsymbol{\beta}) = -\tilde{N} + \tilde{N}\log(\tilde{N}/T) + \sum_{n=1}^{\tilde{N}} \log\left(w_{\tilde{l}_n}\right) + \sum_{n=1}^{\tilde{N}} \log\left(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}\right).$$

4.1.1. Estimation of Location Weights For notational brevity, we start with a setting where a set of candidate rider locations $\mathcal{L}$ and the choice model parameters $\boldsymbol{\beta}$ are given, and the goal is to estimate the location weights $\boldsymbol{w}$. We now outline the corresponding EM algorithm.

E-step. In the E-step, we compute the expectation of the complete data log-likelihood conditional on the observed data $\boldsymbol{b}$ and $\boldsymbol{t}$ and the current estimate $\boldsymbol{w}^{(m)}$ at the m^{th} iteration. The expectation is taken over the randomness of unobserved data, which includes: (1) the number of unobserved riders N'; (2) the arrival times of unobserved riders $\tilde{\boldsymbol{t}}\backslash\boldsymbol{t}$; and (3) the arrival locations of all riders $\tilde{\boldsymbol{l}}$.

$$\mathbb{E}\left[l_C(\boldsymbol{w};\boldsymbol{\beta}) \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right] = \mathbb{E}\left[-\tilde{N} + \tilde{N}\log\left(\frac{\tilde{N}}{T}\right) + \sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) + \sum_{n=1}^{\tilde{N}} \log\left(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}\right) \,\middle|\, \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right] \quad (6)$$

Notice that the only part of the expectation in equation (6) that depends on $\boldsymbol{w}$ is $\sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n})$. Thus, it is sufficient to solely focus on the conditional expectation $\mathbb{E}\left[\sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) \,\middle|\, \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right]$. We denote by $\{l_1, l_2, \cdots, l_N\}$ the sequence of the arrival locations of the observed riders, which is a subsequence of $\{\tilde{l}_1, \cdots, \tilde{l}_{\tilde{N}}\}$, the sequence of arrival locations of all riders. For the sake of exposition, we consider two quantities, which are $\mathbb{P}(l_n = l \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)})$ and $\mathbb{E}[N' \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}]$. The first quantity refers to the probability that the n^{th} observed rider arrives at location $l \in \mathcal{L}$, whereas the second quantity refers to the expected number of unobserved riders. They can be computed as follows.

$$\mathbb{P}\left(l_n = l \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right) = \frac{\mathbb{P}\left(b_n \mid l_n = l,\boldsymbol{t},\boldsymbol{w}^{(m)}\right)\mathbb{P}\left(l_n = l \mid \boldsymbol{t},\boldsymbol{w}^{(m)}\right)}{\mathbb{P}\left(b_n \mid \boldsymbol{t},\boldsymbol{w}^{(m)}\right)} = \frac{p_{l,b_n,S_{t_n}} w_l^{(m)}}{\sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',b_n,S_{t_n}}}, \quad (7)$$

$$\begin{aligned}\mathbb{E}\left[N' \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right] =& \mathbb{E}[N \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}] \frac{\int_0^T \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t} dt}{\int_0^T \left(1 - \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t}\right) dt} \\ =& N \frac{\int_0^T \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t} dt}{\int_0^T \left(1 - \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t}\right) dt}. \end{aligned} \quad (8)$$

We now separate the conditional expectation $\mathbb{E}\left[\sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right]$ into two parts, which consist of observed and unobserved data. Let the sequence $\{l'_1, l'_2, \cdots, l'_{N'}\}$ denote the arrival locations of the unobserved riders.

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{N} \log(w_{l_n}) \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] + \mathbb{E}\left[\sum_{n=1}^{N'} \log(w_{l'_n}) \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] \\
&= \sum_{l\in\mathcal{L}} \sum_{n=1}^{N} \mathbb{P}\left(l_n = l \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right) \log(w_l) \\
&\quad + \mathbb{E}\left[N' \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] \sum_{l\in\mathcal{L}} \underbrace{\frac{\int_0^T w_l^{(m)} p_{l,0,S_t} dt}{\int_0^T \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t} dt}}_{\substack{\text{prob. of arriving at } l \\ \text{conditional on leaving}}} \log(w_l) \qquad \text{(Wald's Lemma)} \\
&= \sum_{l\in\mathcal{L}} \left(\sum_{n=1}^{N} \mathbb{P}\left(l_n = l \mid \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right) + \mathbb{E}\left[N' \mid \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] \frac{\int_0^T w_l^{(m)} p_{l,0,S_t} dt}{\int_0^T \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t} dt} \right) \log(w_l) \\
&= \sum_{l\in\mathcal{L}} \underbrace{\left(\sum_{n=1}^{N} \frac{p_{l,b_n,S_{t_n}} w_l^{(m)}}{\sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',b_n,S_{t_n}}} + N \frac{\int_0^T w_l^{(m)} p_{l,0,S_t} dt}{\int_0^T (1 - \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t}) dt} \right)}_{c_l} \log(w_l) \qquad (9) \\
&= \sum_{l\in\mathcal{L}} c_l \log(w_l).
\end{aligned}
$$

Equation (9) holds by replacing the terms with equations (7) and (8). We have thus simplified the conditional expectation into a weighted sum of logarithms, which can be maximized in closed form in the M-step described below.

M-step. Let $c := \sum_{l\in\mathcal{L}} c_l$ be the sum of c_l for all $l \in L$. By weighted AM–GM inequality, we have

$$
\begin{aligned}
& 0 = \log\left(\sum_{l\in\mathcal{L}} \frac{c_l}{c} \left(\frac{c w_l}{c_l}\right)\right) \geq \log\left(\prod_{l=1}^{L} \left(\frac{c w_l}{c_l}\right)^{\frac{c_l}{c}}\right) = \sum_{l\in\mathcal{L}} \frac{c_l}{c} (\log(c w_l) - \log c_l) \\
\iff \quad & \sum_{l\in\mathcal{L}} c_l \log(w_l) \leq \sum_{l\in\mathcal{L}} c_l \log(c_l) - \sum_{l\in\mathcal{L}} c_l \log(c).
\end{aligned}
$$

The equality holds if and only if $w_l = c_l/c$, which shows that the optimal $\boldsymbol{w}$ has a closed-form solution $w_l = c_l/(\sum_{l'\in\mathcal{L}} c_{l'})$ for $l \in \boldsymbol{\mathcal{L}}$. We note that the solution can also be derived from the KKT conditions of the problem $\max_{\boldsymbol{w}\in\Delta^L} \sum_{l\in\mathcal{L}} c_l \log(w_l)$.

Since the closed-form solution $\hat{\boldsymbol{w}}$ derived in the M-step is a unique maximizer, in the EM algorithm, we know the M-step generates a sequence of vectors $\{\boldsymbol{w}^{(m)}, m = 1, 2, \cdots\}$ where $\boldsymbol{w}^{(m)}$ is the unique maximizer for the expected complete log-likelihood function $\mathbb{E}[l_C(\boldsymbol{w}; \boldsymbol{\beta}) \mid \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m-1)}]$. Here, $\boldsymbol{w}^{(0)}$ is the initial weight vector to start the EM algorithm. Moreover, from equation (6), it

Algorithm 1 EM Algorithm with Given Rider Locations

Initialize a location set $\mathcal{L}_0$ with coordinates $(x_1, y_1), \cdots, (x_{L_0}, y_{L_0})$

Initialize the weight vector $\boldsymbol{w} \leftarrow \boldsymbol{w}^{(0)} \in \Delta^L$, $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^{(0)}$.

Initialize the number of iterations $m \leftarrow 0$.

while stopping criteria are not met **do**

Compute $s \leftarrow \int_0^T (1 - \sum_{l \in \mathcal{L}} w_l^{(m)} p_{l,0,S_t}^{(m)}) dt$.

$$\boldsymbol{\beta}^{(m+1)} \in \arg\max_{\boldsymbol{\beta}} \sum_{l \in \mathcal{L}} \left(\sum_{n=1}^{N} \frac{p_{l,b_n,S_{t_n}}^{(m)} w_l^{(m)}}{\sum_{l' \in \mathcal{L}} w_{l'}^{(m)} p_{l',b_n,S_{t_n}}^{(m)}} \log p_{l,b_n,S_{t_n}} + \frac{N}{s} \int_0^T p_{l,0,S_t}^{(m)} w_l^{(m)} \log p_{l,0,S_t} dt \right) \tag{10}$$

Compute $c_l \leftarrow \sum_{n=1}^{N} \left(p_{l,b_n,S_{t_n}}^{(m)} w_l^{(m)} \big/ \left(\sum_{l' \in \mathcal{L}} w_{l'}^{(m)} p_{l',b_n,S_{t_n}}^{(m)} \right) \right) + N(T-s)/s$, for $l \in \mathcal{L}$.

Update $w_l^{(m+1)} \leftarrow c_l / \sum_{l' \in \mathcal{L}} c_{l'}$, for $l \in \mathcal{L}$.

$m \leftarrow m + 1$.

end while

Output $\boldsymbol{w}^{(m)}$, $\boldsymbol{\beta}^{(m)}$.

is clear that the expected complete log-likelihood function $\mathbb{E}[l_C(\boldsymbol{w}; \boldsymbol{\beta}) \mid \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}]$ is continuous in both $\boldsymbol{w}$ and $\boldsymbol{w}^{(m)}$. Then by Theorem 2 in Wu (1983), we can infer that the sequence of likelihood values $\{l(\boldsymbol{w}^{(1)}; \boldsymbol{\beta}), l(\boldsymbol{w}^{(2)}; \boldsymbol{\beta}), \cdots\}$ converges monotonically to $l(\boldsymbol{w}^*; \boldsymbol{\beta})$ for some stationary point $\boldsymbol{w}^*$. Moreover, all limit points of the estimates sequence $\{\boldsymbol{w}^{(m)}, m = 1, 2, \cdots\}$ are stationary points. Though the estimates of our EM method does not guarantee to converge to global maxima, we will numerically show in Section 5 that it empirically gives a close approximation to the ground truth.

Incorporating Prior Information. In some circumstances, the operator may have prior knowledge regarding $\boldsymbol{w}$ (e.g., based on population density or other socioeconomic data). To incorporate this prior, we can use a *maximum a posteriori probability* (MAP) estimation instead of MLE. The key difference here is that MAP uses an augmented objective function that incorporates a prior distribution over location weights. Specifically, we impose a Dirichlet prior with parameters $\gamma_1, \cdots, \gamma_L \geq 0$ for each location $l \in \mathcal{L}$. Let $g(\boldsymbol{w})$ be the corresponding prior density function. Our MAP estimator then becomes $\hat{w}_{\text{MAP}} \in \arg\max_{\boldsymbol{w} \in \Delta^L} l_I(\boldsymbol{w}) + \log g(\boldsymbol{w}) = \arg\max_{\boldsymbol{w} \in \Delta^L} l_I(\boldsymbol{w}) + \sum_{l \in \mathcal{L}} \gamma_l \log w_l$. A highlight here is that our EM algorithm still has a closed-form update $w_l = (c_l + \gamma_l)/(\sum_{l' \in \mathcal{L}} c_{l'} + \gamma_{l'})$. The update is a simple modification of the original form, where we replace c_l with $c_l + \gamma_l$.

4.1.2. Estimation of Choice Model Parameters We now extend the discussion to jointly estimate location weights $\boldsymbol{w}$ and choice model parameters $\boldsymbol{\beta}$. First, we notice that the expectation of the complete data log-likelihood retains the same expression as equation (6), except that the log-likelihood is now conditioned on $\boldsymbol{\beta}^{(m)}$, the estimate of $\boldsymbol{\beta}$ at the m^{th} iteration, as well. We define

$p^{(m)}_{l,b,S_t}$ as the corresponding choice probability under parameters $\boldsymbol{\beta}^{(m)}$. As in equation (6), the first two terms inside the expectation do not depend on $\boldsymbol{w}$ or $\boldsymbol{\beta}$. The third term depends on $\boldsymbol{w}$ but not $\boldsymbol{\beta}$, while the fourth term depends on $\boldsymbol{\beta}$ but not $\boldsymbol{w}$. Hence, $\boldsymbol{w}$ has the same updating process as before with the only difference of using $\boldsymbol{\beta}^{(m)}$ to replace the true value. To optimize $\boldsymbol{\beta}$, we only need to consider the fourth term. Similar to equation (9), this term can be split into parts corresponding to observed and unobserved arrivals. Moreover, the maximization problem for $\boldsymbol{\beta}$ is concave when the choice probability p_{l,b,S_t} is log-concave in $\boldsymbol{\beta}$ (see Online Appendix EC.1.2 for more details). This holds, for example, when rider choice behaviors are governed by an MNL model described in equation (1) and hence the M-step for $\boldsymbol{\beta}$ is efficient. To integrate the estimation of $\boldsymbol{\beta}$, we only need to augment Algorithm 1 with an additional update for $\boldsymbol{\beta}$ at each iteration. Specifically, in the m^{th} iteration, $\boldsymbol{\beta}^{(m)}$ is updated according to $\boldsymbol{\beta}^{(m+1)} \in \arg\max_{\boldsymbol{\beta}} \ \mathbb{E}[\sum_{n=1}^{\tilde{N}} \log(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}) \,|\boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)},\boldsymbol{\beta}^{(m)}]$. This leads to the update in equation (10). We now give our complete EM algorithm in Algorithm 1. To simplify the exposition in the pseudocode, we define $s = \int_0^T \big(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t}\big)dt$. Recall that s/T measures the average percentage of riders who arrive and book a bike.

We conclude this subsection by commenting that a broader class of algorithms for maximizing the log-likelihood used in the related literature is the MM algorithm (Hunter and Lange 2004). MM algorithms maximize simple concave surrogate functions that minorize the log-likelihood function. Different MM algorithms differ in the way the surrogate function is constructed. EM can be viewed as a special case of an MM algorithm using a surrogate function that is constructed by Jensen's inequality (see Section 3.1 of Hunter and Lange 2004). In Section 5, we compare the performance of our EM algorithm with another MM algorithm that is constructed by viewing the log-likelihood function (5) as a difference of two concave functions (see our discussion at the end of Section 3.1), and constructing surrogate functions using supporting hyperplanes of the second concave function. We show that our EM algorithm has superior performance due to its simple closed-form updates.

4.2. A Location-Discovery Procedure

In practice, the operator often does not have full knowledge of the possible rider arrival locations. As a result, the set of candidate rider locations $\mathcal{L}$ is usually underdetermined and has to be estimated from data. A naïve approach is to enumerate a large set of *all possible* candidate locations, for example, all of the residential buildings in a service area. We then implement our EM algorithm on this set to estimate each location's weight. We refer to this algorithm as the *all-in* algorithm. If the underlying arrival locations are entirely contained in the initial set, then by various results in Section 3.2, the MLE can be identifiable and consistent under certain conditions. This algorithm

has the following drawbacks due to a large cardinality of all possible rider locations: (1) it significantly increases the number of iterations necessary for convergence, resulting in a computationally demanding task; (2) it renders the data requirement for identifiability (e.g., Theorem 1) harder to satisfy, making the estimates less accurate.

Motivated by van Ryzin and Vulcano (2014) where they proposed a market-discovery algorithm to estimate ranking-based customer preferences, we develop a *location-discovery* algorithm to address the aforementioned drawbacks of the all-in algorithm. The algorithm iteratively and adaptively explores new locations until convergence. Suppose $\mathcal{L}$ is the set of all possible rider locations. We begin with a parsimonious set of locations $\mathcal{L}_0 \subset \mathcal{L}$, a discovery procedure is employed to gradually enlarge our location set. The main idea is to maximize the log-likelihood by iteratively adding new locations with the largest potential for improvements and then executing the EM algorithm to update their weights.

To start the derivation, we define the restricted Lagrangian function $\Theta^{\mathcal{L}_0}(\boldsymbol{w}, \boldsymbol{\beta}, \mu)$ with location set $\mathcal{L}_0$ by relaxing the equality constraint $\sum_{l\in\mathcal{L}_0} w_l = 1$ with a Lagrangian multiplier μ. The restricted Lagrangian can be written as

$$\Theta^{\mathcal{L}_0}(\boldsymbol{w}, \boldsymbol{\beta}, \mu) = -N \log \int_0^T \left(1 - \sum_{l\in\mathcal{L}_0} w_l p_{l,0,S_t}\right) dt + \sum_{n=1}^N \log \sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}} + \mu \left(1 - \sum_{l\in\mathcal{L}_0} w_l\right). \quad (11)$$

Consider a local optimum $(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\beta}})$ of the problem $\max_{\boldsymbol{w}\in\Delta^{|\mathcal{L}_0|}, \boldsymbol{\beta}} l_I(\boldsymbol{w}, \boldsymbol{\beta})$ with location set $\mathcal{L}_0$. Recall that $s = \int_0^T \left(1 - \sum_{l\in\mathcal{L}_0} w_l p_{l,0,S_t}\right) dt$. Suppose that $(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\beta}}, \bar{\mu})$ satisfies KKT conditions of the restricted Lagrangian function $\Theta^{\mathcal{L}_0}(\boldsymbol{w}, \boldsymbol{\beta}, \mu)$. We have for all $l \in \mathcal{L}_0$ such that $\bar{w}_l > 0$,

$$\left.\frac{\partial \Theta^{\mathcal{L}_0}(\boldsymbol{w}, \boldsymbol{\beta}, \mu)}{\partial w_l}\right|_{\boldsymbol{w}=\bar{\boldsymbol{w}}, \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}, \mu=\bar{\mu}} = \frac{N}{s} \int_0^T p_{l,0,S_t} dt + \sum_{n=1}^N \frac{p_{l,b_n,S_{t_n}}}{\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}} - \bar{\mu} = 0,$$

which gives that

$$\bar{\mu} = \frac{N}{s} \int_0^T p_{l,0,S_t} dt + \sum_{n=1}^N \frac{p_{l,b_n,S_{t_n}}}{\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}}, \text{ for any } l \in \mathcal{L}_0 \text{ such that } \bar{w}_l > 0.$$

The essence of our location-discovery procedure is to discover a new location that potentially defines the largest improvement direction for the log-likelihood value. To do so, we consider the Lagrangian function $\Theta^{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{\beta}, \mu)$ with the full set of all possible rider locations $\mathcal{L}$. Define $\bar{\bar{\boldsymbol{w}}} \in \Delta^{|\mathcal{L}|}$ such that $\bar{\bar{w}}_l = \bar{w}_l, \forall l \in \mathcal{L}_0$ and $\bar{\bar{w}}_l = 0, \forall l \in \mathcal{L} \setminus \mathcal{L}_0$. Suppose that

$$\left.\frac{\partial \Theta^{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{\beta}, \mu)}{\partial w_l}\right|_{\boldsymbol{w}=\bar{\bar{\boldsymbol{w}}}, \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}, \mu=\bar{\mu}} = \frac{N}{s} \int_0^T p_{l,0,S_t} dt + \sum_{n=1}^N \frac{p_{l,b_n,S_{t_n}}}{\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}} - \bar{\mu} \le 0, \ \forall l \in \mathcal{L} \setminus \mathcal{L}_0, \quad (12)$$

then $(\bar{\bar{\boldsymbol{w}}}, \bar{\boldsymbol{\beta}}, \bar{\mu})$ satisfies the KKT conditions of the full Lagrangian function $\Theta^{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{\beta}, \mu)$ as well. If not, then there exists some $l \in \mathcal{L} \setminus \mathcal{L}_0$ such that $\partial \Theta^{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{\beta}, \mu)/\partial w_l > 0$ evaluated at $\boldsymbol{w} = \bar{\bar{\boldsymbol{w}}}$, $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}$

and $\mu = \bar{\mu}$. Note that since the log-likelihood function $l_I(\boldsymbol{w}, \boldsymbol{\beta})$ is not jointly concave in $\boldsymbol{w}$ and $\boldsymbol{\beta}$, KKT conditions are not sufficient for local optimum. This means that including such a new location l is not guaranteed but may lead to an improved estimate with greater likelihood. Nevertheless, this gives a principled procedure to gradually include new locations. We summarize the estimation algorithm with a location discovery procedure in Algorithm 2 below.

Finding the location with the largest partial derivative of the Lagrangian. At each iteration of Algorithm 2, a new location is selected by solving

$$\max_{l' \in \mathcal{L} \setminus \mathcal{L}_0} \quad \frac{N}{s} \int_0^T p_{l',0,S_t} dt + \sum_{n=1}^{N} \frac{p_{l',b_n,S_{t_n}}}{\sum_{l \in \mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}} \tag{13}$$

to maximize the partial derivative of the Lagrangian function under the full set of locations $\Theta^{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{\beta}, \mu)$. The terms in the objective function have quite intuitive interpretations. It finds a location that strikes a balance between explaining riders' leaving without picking up bikes (the first term) and the observed booking sequence (the second term). Based on the current estimates, if s is small or $\sum_{l \in \mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}$ is large, then the first term possesses a heavier weight than the second term. The new location l' tends to improve the explanation of riders' leaving behaviors by uplifting the leaving probability $p_{l',0,\mathcal{S}_t}$. On the contrary, with a large s or small booking probability $\sum_{l \in \mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}$, the new location l' focuses more on explaining the booking behaviors by increasing the value of booking probabilities $p_{l',b_n,S_{t_n}}$. Noting that when the ground-truth choice model parameters $\boldsymbol{\beta}$ is unknown, for problem (13) to be well-defined, an implicit assumption for equation (13) is that $\boldsymbol{\beta}$ is not location dependent, i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_L$. Otherwise, the choice model parameter for the new location l', $\boldsymbol{\beta}_{l'}$, will be unknown, which prevents us from computing its corresponding choice probabilities in (13).

As we will illustrate later, this objective as a function of the new location coordinates $(x_{l'}, y_{l'})$ is not concave in general and may exist multiple local optima (see Figure EC.4 in the Online Appendix for an example). Problem (13) thus, unfortunately, does not possess much structure. We could apply a general nonlinear programming algorithm such as gradient-based methods to approach its local optimum. On the other hand, since only two variables $x_{l'}$ and $y_{l'}$ need to be optimized, an alternative simple and effective method is to use *grid search*. The grid search method optimizes the objective function in a full factorial sampling plan, which places a grid of evenly spaced points across the search area. We implement a multi-round variant of the search method to speed up the search process. In each round, the granularity of the search grid becomes finer and finer within a shrinking and more targeted search region. More details are provided in Section 5.

Another benefit of running a grid-search method is to easily enable *batch* addition of new locations. To be specific, instead of exploring one location at a time, we can simultaneously discover a batch of locations in each iteration. For example, all local maxima, in addition to the global maximum, can be included in each iteration. In the grid search procedure, we identify local maxima by discovering new locations whose partial derivatives are greater than those of eight neighboring locations in the search grid. As we will show later in Section 5, batch addition often improves the accuracy of the estimator.

Algorithm 2 Estimation Algorithm with a Location-Discovery Procedure

Initialize a parsimonious location set $\mathcal{L}_0 \subset \mathcal{L}$.

Initialize $\boldsymbol{w}$ and $\boldsymbol{\beta}$ by running Algorithm 1 with location set $\mathcal{L}_0$.

while stopping criteria are not met **do**

$s \leftarrow \int_0^T \left(1 - \sum_{l \in \mathcal{L}_0} w_l p_{l,0,S_t}\right) dt.$

$l^* \in \arg\max_{l' \in \mathcal{L} \setminus \mathcal{L}_0} \left(\frac{N}{s} \int_0^T p_{l',0,S_t} dt + \sum_{n=1}^N \frac{p_{l',b_n,S_{t_n}}}{\sum_{l \in \mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}} \right).$ ▷ discover a new location

$\mathcal{L}_0 \leftarrow \mathcal{L}_0 \cup \{l^*\}.$

Update $\boldsymbol{w}, \boldsymbol{\beta}$ by running Algorithm 1 with location set $\mathcal{L}_0$.

end while

Output the location set $\mathcal{L}_0$, its corresponding weight vector $\boldsymbol{w}$ and choice model parameters $\boldsymbol{\beta}$.

Convergence and stopping criteria. A natural stopping criterion is to terminate Algorithm 2 when conditions (12) hold. In other words, the resulting location set $\mathcal{L}_0$ and its corresponding weight vector $\boldsymbol{w}$, together with the Lagrangian multiplier $\bar{\mu}$, satisfy the KKT conditions. With this stopping criterion, when the set of all possible rider locations $\mathcal{L}$ is finite, it can be seen that Algorithm 2 terminates within a finite number of iterations. This is because, at each iteration, either the stopping criteria are met or a new location will be added. As there are only a finite number of them, the algorithm will converge. However, as we mentioned before, the converging estimate does not necessarily correspond to the MLE estimate as KKT conditions are not sufficient for optimality due to the non-concavity of the log-likelihood function $l_I(\boldsymbol{w})$. Furthermore, with this stopping criterion, the algorithm usually discovers much more locations than necessary since including new location parameters into the likelihood function always improves the log-likelihood value. It is thus practical to explore other stopping criteria that penalize model complexity to prevent over-fitting. A good method is to separate a portion of the data for validation and compute the likelihood of this out-of-sample validation data in each iteration. The stopping criterion can

thus be guided by the out-of-sample likelihood. For example, we can terminate Algorithm 2 when the out-of-sample likelihood decreases after we add new locations.

5. Numerical Experiments

In this section, we present numerical experiments of our demand model and estimation procedures. These are broadly divided into two parts, experiments based on synthetic data and real-world bike-sharing data. In Section 5.1, we demonstrate the performance of the algorithm over synthetically generated data on a square. We benchmark our EM algorithm and give evidence that the location-discovery procedure significantly improves the accuracy of the estimator. In Section 5.2, we illustrate our algorithm and its estimation results on a set of real-world dockless bike-sharing data in Seattle. Based on a comprehensive set of experiments and benchmarks with this data set, we demonstrate the robustness and accuracy of our methods in practice. In addition, the estimation results provide managerial insights regarding bike allocations to increase service levels in the Seattle area. All algorithms are implemented in Python 3.8.10 with NumPy 1.23.3 on a virtual machine with 8 vCPUs using Microsoft Azure. Data and code to reproduce all experiments presented in this section can be found here.

5.1. Experiments Based on Synthetic Data

We describe the generating process of our synthetic data. We consider an arrival period of length T. For each simulation run, we sample the initial locations of each bike (at $t=0$) independently and uniformly from a 10×10 square. In specific, the (x, y) coordinates live in $[-5, 5]^2$. We then sample a sequence of arrivals based on a homogeneous Poisson process with rate λ within $[0, T]$ representing the rider arrival times. We use a discrete-event simulation that processes arrival times one by one in chronological order. Each rider arrives at a location $l \in \mathcal{L}$ sampled from a multinoulli distribution with ground-truth probability w_l^* such that $\sum_{l\in\mathcal{L}} w_l^* = 1$. A rider chooses to book an available bike based on an underlying MNL choice model with ground-truth parameters $\beta_{0,l}^* = 1$ and $\beta_{1,l}^* = -1$ for all $l \in \mathcal{L}$. We use Euclidean distance (the L_2-norm between two location coordinates) as our measurement of distance. When a rider books a bike, we randomly sample her destination uniformly from the 10×10 square. We also generate her booking duration (in hours) according to a rectified Gaussian distribution $\max\{N(\text{walking time} + \text{traveling time}, 0.1), 0.05\}$ where walking time is computed as the distance from the rider's arrival location to the bike location divided by 4 km per hour (walking speed) and the traveling time is computed as the distance from the bike location to the rider's destination divided by 18 km per hour (cycling speed). If the rider chooses to leave, we simply move on to the next arrival. Bike patterns will be updated accordingly. We repeat this procedure until all rider arrivals are processed.

5.1.1. Performance of the EM algorithm We first start with the setting where the choice model parameters $\boldsymbol{\beta}$ are known, and the goal is to estimate location weights $\boldsymbol{w}$. We benchmark the performance of our EM algorithm in maximizing the log-likelihood function by comparing it with an MM algorithm mentioned at the end of Section 4.1. This approach iteratively optimizes a concave surrogate function which minorizes the log-likelihood function. In particular, ignoring the constant, we re-arrange the log-likelihood function (5) into a difference of two concave functions $g(\boldsymbol{w};\boldsymbol{\beta}) - h(\boldsymbol{w};\boldsymbol{\beta})$ where $g(\boldsymbol{w};\boldsymbol{\beta}) = \sum_{n=1}^{N} \log \sum_{l\in\mathcal{L}} w_l p_{l,b_n,S_{t_n}}$ and $h(\boldsymbol{w};\boldsymbol{\beta}) = N \log \int_0^T \left(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_{t_n}}\right) dt$. To find a concave surrogate, we replace $h(\cdot)$ with its first-order approximation. Such technique is also known as the concave-convex procedure in the machine learning community (Yuille and Rangarajan 2003). Then the problem reduces to iteratively solving the following concave program with a simplex constraint

$$\boldsymbol{w}^{(m)} \in \arg\max_{\boldsymbol{w}\in\Delta^L} g(\boldsymbol{w}) - h\left(\boldsymbol{w}^{(m-1)}\right) - \nabla h\left(\boldsymbol{w}^{(m-1)}\right)^T \left(\boldsymbol{w} - \boldsymbol{w}^{(m-1)}\right). \tag{14}$$

Although equation (14) is concave, it does not have a closed-form solution. Observing that the feasible region of equation (14) is a simplex, we use the Frank-Wolfe (FW) algorithm (Frank and Wolfe 1956) to solve this concave maximization problem. At each iteration with current estimate $\boldsymbol{w}'$, FW solves a linear program $\max_{\boldsymbol{w}\in\Delta^L} a^T\boldsymbol{w}$ where $a \in \mathbb{R}^L$ is the gradient of the function $g(\boldsymbol{w}) - \nabla h\left(\boldsymbol{w}^{(m-1)}\right)^T \boldsymbol{w}$ evaluated at $\boldsymbol{w} = \boldsymbol{w}'$. It is easy to check that an optimal solution of this linear program is a standard unit vector e_j for some $j \in \arg\max_l a_l$. Thus each iteration of FW is reduced to a simple `findmax` operation. The current $\boldsymbol{w}'$ is then updated to $\boldsymbol{w}' + \alpha(e_j - \boldsymbol{w}')$ where $\alpha \in [0,1]$ is a step size. To determine α, we perform a line search in $[0,1]$ that maximizes the objective function $g(\boldsymbol{w}' + \alpha(e_j - \boldsymbol{w}')) - \nabla h\left(\boldsymbol{w}^{(m-1)}\right)^T (\boldsymbol{w}' + \alpha(e_j - \boldsymbol{w}'))$.

To compare their performances, we position a $M \times M$ Cartesian grid in the aforementioned 10×10 square and randomly select certain points from the M^2 intersections of the Cartesian grid as the underlying true rider arrival locations. Their corresponding weights are independently sampled from a symmetric Dirichlet distribution with parameter 1. We consider the number of ground-truth rider arrival locations, bikes, and the grid size $(L^*, B, M) \in \{(10,40,5),(25,100,10),(100,400,20)\}$. The length of the arrival period T is set to be 100, and the underlying rider arrival rate is set to be $\lambda = 10$. We consider all M^2 intersections as the set of potential rider arrival locations $\mathcal{L}$. We compare the efficiency of the EM and MM algorithms when they reach comparable near-optimal log-likelihood values. To evaluate the accuracy of the estimator, we compute the Wasserstein distance between the predicted location weights and the true location weights. In our model, given the

estimated rider locations and their weights $(\hat{\mathcal{L}}, \hat{\boldsymbol{w}})$ and the corresponding ground truth $(\mathcal{L}^*, \boldsymbol{w}^*)$, the Wasserstein 2−distance is defined as

$$W\left((\hat{\mathcal{L}}, \hat{\boldsymbol{w}}), (\mathcal{L}^*, \boldsymbol{w}^*)\right) = \inf_{\lambda_{i,j}} \left\{ \sum_{i=1}^{|\hat{\mathcal{L}}|} \sum_{j=1}^{|\mathcal{L}^*|} \lambda_{i,j} \|\hat{l}_i - l_j^*\|_2^2 : \sum_{i=1}^{|\hat{\mathcal{L}}|} \lambda_{i,j} = w_j^*, \sum_{j=1}^{|\mathcal{L}^*|} \lambda_{i,j} = \hat{w}_i, \lambda_{i,j} \geq 0 \right\}^{1/2},$$

where $\|\hat{l}_i - l_j^*\|_2 := \sqrt{(\hat{x}_i - x_j^*)^2 + (\hat{y}_i - y_j^*)^2}$ is the Euclidean distance between locations $\hat{l}_i$ and l_j^*.

Table 1 Performance comparison of the EM and MM algorithms in estimating location weights.

	EM algorithm				MM algorithm			
(L^*, B, M)	Iterations	WD	Lkd	Time	Iterations	WD	Lkd	Time
(10, 40, 5)	120.2	3.45	-4,533	0.02	274.8	3.45	-4,533	20.87
(25, 100, 10)	323.5	3.22	-7,524	0.16	697.9	3.17	-7,524	57.77
(100, 400, 20)	536.3	2.99	-9,807	2.32	1311.2	3.04	-9,807	466.88

Table 1 reports the number of iterations until convergence, Wasserstein distance (WD) between the predicted location weights and the true weights, the log-likelihood values upon convergence (Lkd), as well as the CPU times in seconds of both algorithms averaged over 10 simulation runs. When computing the log-likelihood value, we exclude the constant term $-N + N \log N$. Both algorithms produce very similar prediction accuracy and the converging location weights are in close proximity. However, the EM algorithm requires significantly less computation time to reach similar optimality in terms of likelihood value—it benefits from closed-form updates, which makes each iteration much faster than the MM algorithm. On the other hand, each iteration of the MM algorithm is much more expensive, as it requires tuning a step size by a line search.

5.1.2. Performance of the location-discovery procedure We now evaluate the performance of the location discovery procedure. We randomly generate rider arrival locations and their corresponding weights on a Cartesian grid of size 10×10 in the same square service region. We assume that no prior information is available about rider candidate locations other than they live in a square service region. In other words, we allow rider arrival locations to be arbitrary places within the service region. For each grid size, we consider three scenarios with different numbers of ground-truth rider locations and bikes: $(L^*, B) \in \{(10, 40), (25, 100), (100, 400)\}$. We test with different amount of data by setting the length of the arrival period to $T \in \{100, 500\}$. We compare the following methods.

- **All-in algorithm.** We choose all 100 intersections of the 10×10 grid as the set of candidate rider locations $\mathcal{L}$, respectively. We then run Algorithm 1 to obtain their location weights.

- **Location-discovery algorithm.** We start with two rider arrival locations sampled uniformly from the service region. We implement two variants: a single mode and a batch mode. In the single mode, we begin our search by initializing a coarse Cartesian grid of dimensions 10×10 onto the square service region. We find the rider location in the grid that maximizes the partial derivative. We then perform a second grid search that is confined to a smaller square region whose boundary is defined by the neighboring locations of the one selected from the first round. We again overlay a 10×10 Cartesian grid onto this smaller square region and identify the location that has the largest partial derivative. In the batch mode, we select all local maxima in the first round, and a second grid search is conducted near all locations selected in the first round. As we mentioned in Section 4.2, we use the following stopping criterion for the location-discovery algorithm. We take the first 80% data generated from the arrival period for training and the remaining 20% out-of-sample data for validation. We compute the out-of-sample likelihood in each iteration. In the single mode, the algorithm terminates after two consecutive decreases in likelihood for $(L^*, B) \in \{(10, 40), (25, 100)\}$, and three for $(L^*, B) = (100, 400)$. In the batch mode, the algorithm stops at the first decrease for $(L^*, B) \in \{(10, 40), (25, 100)\}$, and after two consecutive decreases for $(L^*, B) = (100, 400)$.
- **K-means clustering.** We also test a simple baseline method based on K-means clustering that partitions all bike booking locations into different clusters, and each cluster centroid is considered as a rider arrival location. Location weights are then computed by normalizing the number of bookings belonging to each cluster so that they sum up to 1. We also give this K-means clustering some advantages by assuming that the number of underlying rider arrival locations L^* is known beforehand. That is, we set the number of clusters $K = L^*$.

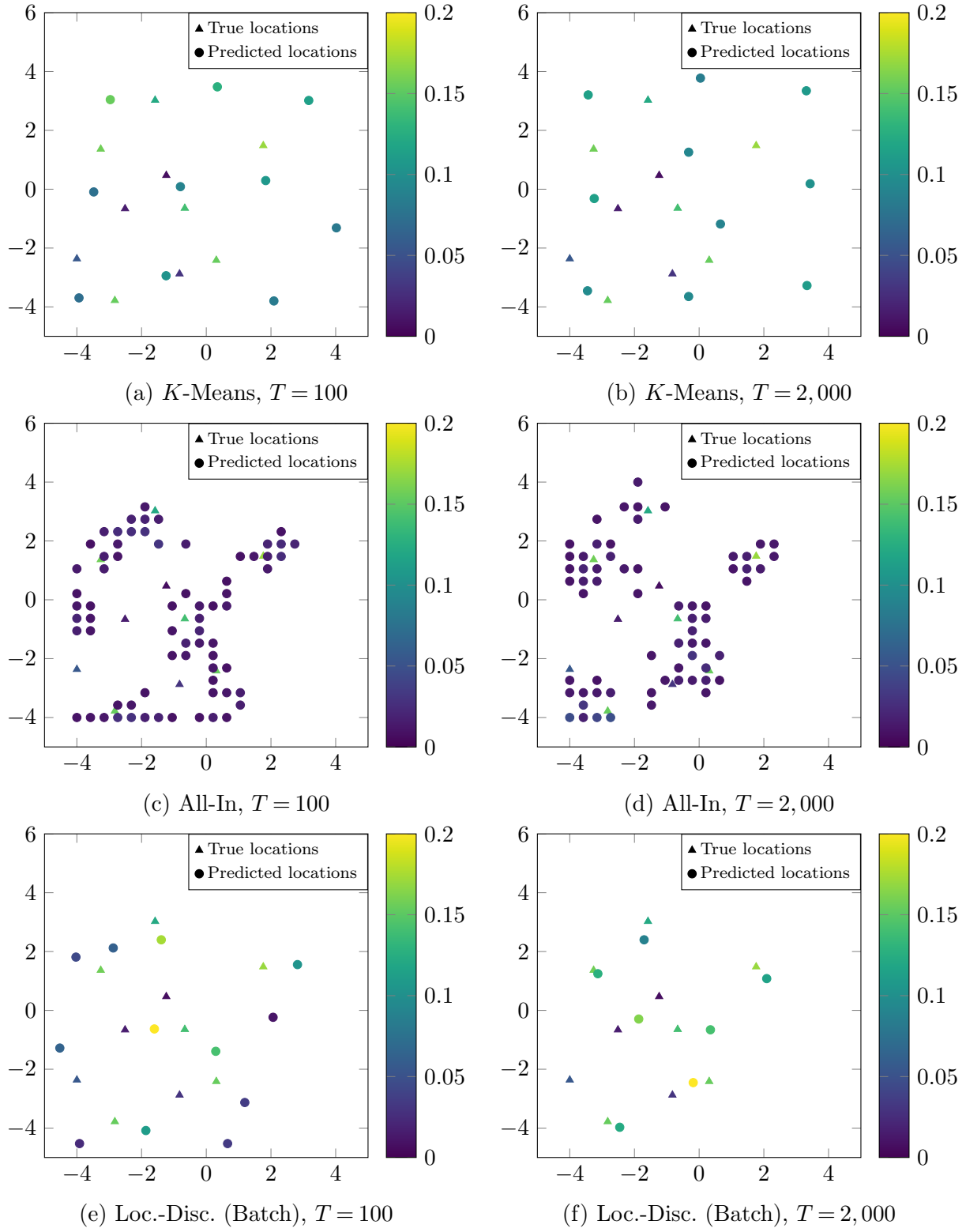
Table 2 reports the performances of different algorithms by averaging over 30 simulation runs. For $(L^*, B) \in \{(10, 40), (25, 100)\}$, we exclude all rider locations with predicted weights smaller than 0.01. The table includes several metrics: "Locs" refers to the number of predicted locations; "WD" refers to the Wasserstein distance between the predicted location weights and the underlying true location weights; "Lkd" refers to the log-likelihood value over the entire arrival period. We also report computation time measured in seconds. We have the following key observations. (1) All algorithms significantly outperform the baseline clustering algorithm in terms of the Wasserstein distance and the log-likelihood value. (2) The all-in algorithm often significantly overestimates the number of rider locations. This leads to worse predictive accuracy in terms of Wasserstein distance and log-likelihood value. Another observation about the all-in algorithm is that the Wasserstein distance does not significantly decrease as one gets more data, which is likely related to stricter identifiability conditions caused by a large set of candidate locations (Theorem 1). In contrast,

our location-discovery algorithms, especially the one with batch addition, achieve improved results with larger sample sizes, as demonstrated by the significantly lower Wasserstein distance obtained for $T = 500$ compared to $T = 100$. (3) The location-discovery algorithms have overall the best performance and the batch mode improves over the single mode as evidenced by the lower Wasserstein distance and similar log-likelihood value. The batch mode typically identifies more locations than the single mode, bringing the estimates closer to the true number of locations in cases where L^* and B are large. This stems from the batch mode's capability to simultaneously detect multiple high-quality locations in each iteration, leading to substantial improvements in likelihood. Remarkably, this increased accuracy is achieved with comparable, and often lower, computation time, highlighting batch mode's efficiency in requiring less time per high-quality location discovered. (4) The location-discovery algorithm under our implementation can exhibit a longer computation time compared to the K-means and all-in algorithms for instances where L^* and B are large. The computation times in these scenarios are primarily spent on evaluating the partial derivatives (13) in the grid search method. Further research can look into methods for accelerating the discovery procedure—e.g., using (multi-start) gradient descent over the coordinates of the new location to approach local optima of (13).

Table 2 Prediction performance under synthetic data.

T	(L^*, B)	Algorithm	Locs	WD	Lkd	Time	Algorithm	Locs	WD	Lkd	Time
100	(10,40)	K-Means	10.0	2.29	-4,654	0.03	All-In	26.7	2.42	-4,533	0.44
	(25,100)		25.0	1.88	-7,498	0.32		31.2	1.62	-7,396	1.51
	(100,400)		100.0	1.49	-10,005	5.96		100.0	1.54	-9,925	6.69
500	(10,40)		10.0	2.29	-27,230	0.18		29.4	2.47	-26,629	2.45
	(25,100)		25.0	1.89	-43,699	1.66		34.3	1.73	-43,189	7.93
	(100,400)		100.0	1.49	-58,044	29.86		100.0	1.57	-57,700	33.65
100	(10,40)	Loc.-Disc.	12.8	2.19	-4,541	0.39	Loc.-Disc.	15.9	2.10	-4,545	0.35
	(25,100)	(Single)	16.9	1.78	-7,404	1.82	(Batch)	20.7	1.64	-7,400	1.61
	(100,400)		34.1	1.40	-9,930	39.46		87.7	0.96	-9,924	49.05
500	(10,40)		15.8	1.95	-26,623	3.42		17.8	1.90	-26,622	2.50
	(25,100)		17.3	1.61	-43,181	11.59		22.3	1.42	-43,174	10.32
	(100,400)		46.7	1.24	-57,703	378.15		90.0	0.88	-57,691	298.60

In Figure 1, we visualize an instance of the predicted locations and their corresponding weights with $L^* = 10$, $B = 40$ and $\lambda = 10$, under the K-means algorithm, the all-in algorithm and the location-discovery algorithm with batch mode when $T = 100$ and $T = 2,000$. We observe that as T increases from 100 to $2,000$, the predicted locations in the location-discovery algorithm approach the true locations more closely; the all-in algorithm also shows some signs of concentration around true locations but not as significant as the location-discovery algorithm; the K-means algorithm, on the other hand, does not exhibit any significant improvement.



(a) K-Means, $T = 100$

(b) K-Means, $T = 2,000$

(c) All-In, $T = 100$

(d) All-In, $T = 2,000$

(e) Loc.-Disc. (Batch), $T = 100$

(f) Loc.-Disc. (Batch), $T = 2,000$

Figure 1 **Predicted and true rider locations and their corresponding weights.** We use triangles to represent the underlying true locations and circles to represent the predicted locations. Different colors represent different weights of the corresponding locations. We use a 20×20 Cartesian grid in both the all-in algorithm and the location-discovery algorithm.

In Section EC.1.1 of the Online Appendix, we display the objective values (13) of all locations in the square service region under the location-discovery algorithm with batch addition. We observe that the objective values become more and more multi-modal iteration by iteration. In Section EC.1.2, we conduct computational experiments when the choice model parameter $\beta_{1,l}$ is *unknown*. By using Algorithms 1 and 2 to jointly estimate the model parameter in each iteration, both the location weights and the value of $\beta_{1,l}$ can be estimated with high accuracy.

5.2. Experiments Based on Seattle Bike-Sharing Data

We now run experiments using data from a real-world dockless bike-sharing system. The data set records all bookings and real-time bike locations and statuses from a dockless bike-sharing company in the Seattle region during July and August 2019. Figure 3 depicts the service region we focus in this experiment. We look at time periods in morning rush hour (from 8 am to 8:30 am) every day. There are 4,024 bookings and 2,317 bikes in total.

We use an MNL model specified in equation (1) to fit rider choice behaviors. Kabra et al. (2019) estimates a structural demand model using Paris bike-sharing data. Their user choice model is specified as an MNL model with walking distance (in km) as the feature and time and location fixed effects. They specify the disutility of walking distance using a piecewise linear structure where the coefficient of walking distance is estimated to be -2.229 for walking distance less than 300 meters and -15.445 for walking distance greater than 300 meters. Similarly, in our MNL model, we let $\beta_{0,l}$ and $\beta_{1,l}$ to be the same across all rider locations. We set $\beta_{0,l} = 1$ and estimate $\beta_{1,l}$ using the EM algorithm with an initial value of -3. We compare the following different algorithms:

(1) *A mixed-effects model*: we develop a parametric mixed-effects model inspired from Kabra et al. (2019) to estimate the number of bookings in different partitions of the service region. The model incorporates key fixed effects such as bike availability, population density, proximity to metro stations, and the presence of tourist attractions (obtained from Google Maps Places API) that influence rider arrival rates at different locations (we include 307 locations from the intersections of a 20×20 grid onto the service region excluding those landed in water). It uses an MNL model to explain rider bookings given the bike availability data. It also imposes a random effect on each block in the partition to explain unobserved heterogeneity in booking numbers. Because the mixed-effect model highly depends on the partition used to train, to examine the robustness of the model accuracy, we train and compare 10 different mixed-effects models using different partitions (see Figure 2 regarding how various partitions are created). Further details about the model setup and its estimated coefficients are documented in Online Appendix EC.1.3.

(2) *All-in algorithm*: initialized with 310 popular demand generating locations including metro stations, major tourist attractions, shopping malls, schools, and parks in our service region using Google Maps Places API. We normalize the population density in each location and use it as prior weights for MAP estimation (see the discussion at the end of Section 4.1.1);

(3) *Location-discovery algorithm*: we use batch addition with a two-round grid search, a granularity of 20×20 for the first round and 10×10 for the second round. We initialize the algorithm by randomly generating 20 coordinates on the service region as initial locations; it is worth noting that this implementation does *not* require any additional data sources, such as point-of-interest and social-economic data used in the previous two algorithms.

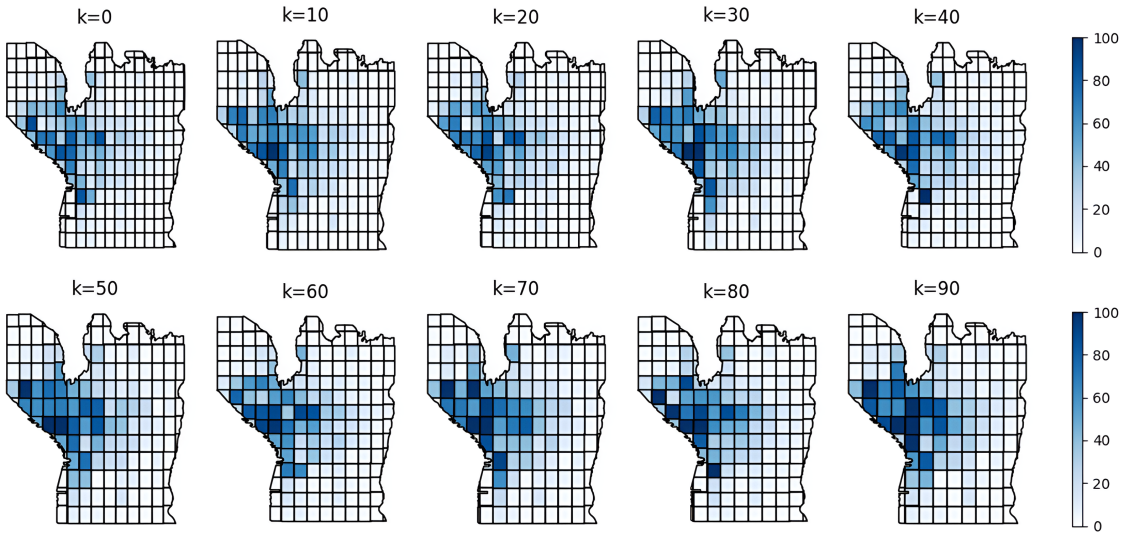


Figure 2 Number of bookings during the training period under different partitions of the service region.

We split our data into training and testing sets. For the all-in algorithm and the mixed-effects model, the training and testing sets consist of bookings ranging from July 1$^{\text{st}}$ to July 31$^{\text{st}}$ and from August 1$^{\text{st}}$ to 15$^{\text{th}}$ respectively. For the location-discovery algorithm, we further split the training set: bookings from July 1$^{\text{st}}$ to July 24$^{\text{th}}$ are used to run the estimation algorithm, and bookings from July 25$^{\text{th}}$ to July 31$^{\text{st}}$ are used as a validation set to guide the stopping criterion—i.e., we stop discovering new locations when the validation likelihood decreases. We implement the aforementioned algorithms and evaluate their performances on the testing set. Based on our predictions of locations $\hat{\mathcal{L}}$, weights $\hat{\boldsymbol{w}}$ and choice model parameters $\hat{\boldsymbol{\beta}}$ using the training set, the arrival rate $\hat{\lambda}$ can be estimated as $\hat{\lambda} = N_{\text{train}} / \int_0^{T_{\text{train}}} \sum_{l \in \hat{\mathcal{L}}} \sum_{b \in \mathcal{B}_t} \hat{w}_l \hat{p}_{l,b,S_t} dt$ by equation (4) where N_{train} is the number of bookings observed in the training set and T_{train} is the length of the training period. Let T_{test} be the length of the testing period. Ideally, we would like to compare the predicted locations and their weights with their corresponding ground truths as what we did in Section 5.1.

However, such ground truths are hard to obtain in real data. Instead, we measure the accuracy in both the training and testing sets by comparing the predicted bookings versus the actual bookings at a granular level—we create 100 different service region partitions, illustrated in Figure 2. In specific, we divide the service region into identical rectangular blocks under each partition. We generate 100 different partitions with different lengths, widths and locational shifts, ranging from 110 to 213 total number of blocks. The lengths and widths of the rectangle in each partition are drawn from $0.005, 0.0055, 0.006, 0.0065$, and 0.007 degrees in longitude and latitude, respectively. This full combination gives $5 \times 5 = 25$ different partitions. For each length-width combination, we also uniformly shift the coordinates of each block by 0.002 degrees, creating 4 different perturbed partitions. This results in a total of $25 \times 4 = 100$ partitions. We index the 100 rectangular partitions from 0 to 99 and parameter $k \in \{0, 10, \cdots, 90\}$ in Figure 2 are the corresponding indices. Let $\mathcal{C}_k$ be the set of blocks under partition $k \in \{0, 1, \cdots, 99\}$. For each partition k, under the all-in and location-discovery algorithms, the predicted number of bookings in block $c \in \mathcal{C}_k$ can be expressed as $\hat{N}_{\text{train},c} = \hat{\lambda} \int_0^{T_{\text{train}}} \sum_{l \in \hat{\mathcal{L}}} \sum_{b \in \mathcal{B}_{t,c}} \hat{w}_l \hat{p}_{l,b,S_t} dt$ and $\hat{N}_{\text{test},c} = \hat{\lambda} \int_{T_{\text{train}}}^{T_{\text{train}}+T_{\text{test}}} \sum_{l \in \hat{\mathcal{L}}} \sum_{b \in \mathcal{B}_{t,c}} \hat{w}_l \hat{p}_{l,b,S_t} dt$ for training and testing sets respectively, where $\mathcal{B}_{t,c}$ is the index set of available bikes in block c at time t. Under the mixed-effects models, we include an additional term that accounts for the random effects specific to each block in the partition k. This is calculated by taking a weighted average of the random effects from the partition used to train the model, where the weights are determined by the amount of overlap between the testing blocks and the training blocks. (If the training and testing use the same partition, we simply apply the trained random effects to each block in the testing partition.) More details are provided in Section EC.1.3 of the Online Appendix. To measure the accuracy, let $N_{\text{train},c}$ and $N_{\text{test},c}$ be the corresponding actual bookings in block c in the training and testing sets. We use the weighted mean absolute percentage error (WMAPE) where the weights are given by the booking volume in each block. Specifically, for each partition k, this accuracy measure can be computed as $\text{WMAPE}_{\text{train},k} = (1/\sum_{c \in C_k} N_{\text{train},c}) \cdot (\sum_{c \in C_k} |N_{\text{train},c} - \widehat{N}_{\text{train},c}|)$ and $\text{WMAPE}_{\text{test},k} = (1/\sum_{c \in C_k} N_{\text{test},c}) \cdot (\sum_{c \in C_k} |N_{\text{test},c} - \widehat{N}_{\text{test},c}|)$ for training and testing accuracy, repsectively. We calculate the $\text{WMAPE}_{\text{train},k}$ and $\text{WMAPE}_{\text{test},k}$ for all 100 partitions and take the average to evaluate the performance of each algorithm. We also record the best and worst WMAPE among the 100 partitions.

In addition, we create a counterfactual scenario by randomly removing 30% of the bikes from the testing data to simulate a situation where a portion of the bike fleet is unavailable (e.g., due to maintenance or malfunctions, or relocation to other communities). Bookings associated with the removed bikes are reassigned to the nearest available bike within a 300-meter radius at the time of

	Average WMAPE			Worst WMAPE			Best WMAPE			Log-likelihood		
	Train	Test	Test (-30%)	Train	Test	Test (-30%)	Train	Test	Test (-30%)	Train	Test	Test (-30%)
All-In	21.33	32.08	29.27	25.40	37.43	34.45	17.96	27.46	23.77	-47,528	-24,114	-23,334
Location-Discovery	18.83	29.61	28.04	22.55	34.59	32.16	13.96	23.79	21.66	-47,428	-24,088	-23,298
Mixed-Effects ($k=0$)	24.15	29.37	29.92	29.79	34.26	36.06	18.41	24.10	26.19	-47,942	-24,250	-23,469
Mixed-Effects ($k=10$)	23.59	30.11	29.96	28.93	36.55	35.24	16.11	23.89	25.44	-47,922	-24,225	-23,431
Mixed-Effects ($k=20$)	30.33	34.51	35.94	34.60	40.15	41.02	24.83	28.72	31.84	-47,990	-24,304	-23,533
Mixed-Effects ($k=30$)	26.19	31.23	31.55	30.89	37.13	36.90	19.44	25.84	26.99	-47,933	-24,241	-23,458
Mixed-Effects ($k=40$)	25.22	30.26	30.92	30.39	35.11	35.82	19.12	24.19	27.02	-47,969	-24,263	-23,483
Mixed-Effects ($k=50$)	29.88	33.34	34.08	34.51	39.72	39.02	23.66	27.68	29.64	-48,013	-24,306	-23,530
Mixed-Effects ($k=60$)	27.05	31.74	32.92	31.16	37.48	39.76	20.37	25.71	28.18	-47,947	-24,269	-23,493
Mixed-Effects ($k=70$)	28.69	32.76	33.12	33.53	37.83	37.82	22.83	26.78	28.12	-47,996	-24,259	-23,471
Mixed-Effects ($k=80$)	24.75	28.50	29.69	30.09	33.20	35.29	14.93	22.75	25.16	-47,880	-24,196	-23,403
Mixed-Effects ($k=90$)	25.13	30.89	30.93	30.72	37.66	35.99	15.75	24.32	26.54	-47,971	-24,291	-23,515

Table 3 Prediction performance on Seattle data. The estimated β_1 values are -4.0 and -4.4 for the all-in and the location discovery algorithms, respectively, and range from -7.8 to -3.6 for the mixed-effects models.

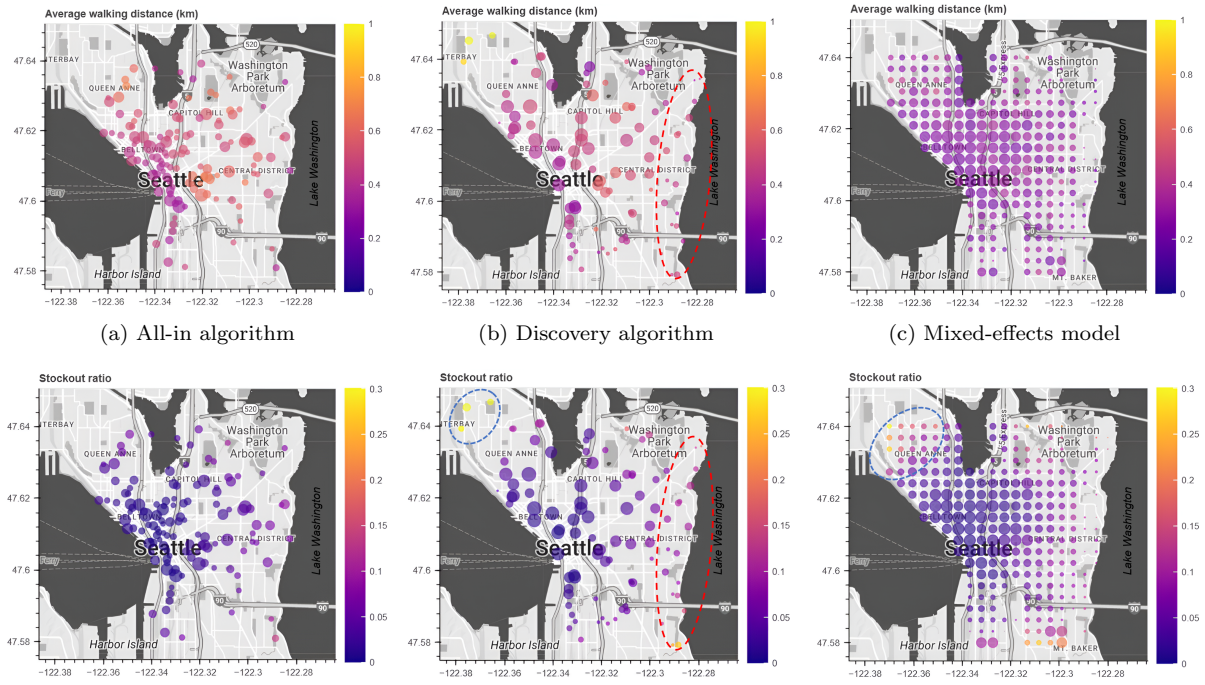
booking. If no suitable bike is found, the booking is considered lost. For each reassigned trip, we assume that the booking time and the rider's destination remain the same as the original trip. As a consequence, for all subsequent trip reassignments, it is equivalent to treat the substituted bike as removed and the originally booked bike as retained. This counterfactual scenario is designed to assess the models' ability to accurately capture shifts in demand resulting from service disruptions or capacity reductions, which offers valuable operational insights.

Table 3 reports the performance of all three algorithms tested. We train 10 different mixed-effects model each using a different partition $k \in \{0, 10, \cdots, 90\}$ based on the partitioning procedure described in Figure 2. However, the trainings of the all-in and location-discovery algorithms do not depend on service region partition. As a consequence, they tend to have the best training accuracy (especially the location-discovery algorithm) across 100 different service region partitions. While the mixed-effects models might be relatively accurate on the partitions they are trained on (see Table EC.3 of Online Appendix EC.1.3), they struggle to generalize to other partitions. For the testing accuracy, the location-discovery algorithm again has excellent performance, while the performance of the mixed-effects models has a lot of variability and is highly dependent on the specific partition it is trained on. Although 2 out of 10 mixed-effects models ($k = 0, 80$) have lower average test WMAPEs compared to the location-discovery algorithm's, it is non-trivial to select the best model a priori without knowing the testing data. (The mixed-effects model with the best training performance ($k = 10$) is worse in testing performance.) The robust performance in predicting bookings across various service region partitions at granular levels demonstrates that the location-discovery algorithm produces estimates closer to the underlying data-generating process. The most notable observation is that, after removing 30% of the bikes from the testing data (column "Test -30%" in Table 3), the discovery algorithm outperforms all others across all evaluation

criteria, followed by the all-in algorithm. This indicates that our locational demand model more effectively captures demand substitution across different locations in response to changes in bike availability, outperforming the mixed-effects models, which show a decline in performance under this counterfactual scenario (with the exception of $k = 10$). These results suggest that the locational demand model is a stronger candidate for counterfactual analysis and for providing reliable inputs for downstream operational decision-making. In Figures EC.5 and EC.6 of Online Appendix EC.1.3, we plot the distributions of block-level relative errors across different algorithms to compare accuracy at an even more granular level. In Table EC.3 of Online Appendix EC.1.3, we report the WMAPE for each mixed-effects model when evaluated on its respective training partition. As expected, their performance improves slightly in these cases but remains subpar overall—4 out of 10 mixed-effects models outperform the location-discovery algorithm on the training data, 5 out of 10 on the testing data, and only 2 out of 10 on the testing data with 30% of bikes removed.

The last three columns of Table 3 also report the log-likelihood values over the training and testing sets. These are independent of the service region partitions. Not surprisingly, the all-in and location-discovery algorithms have similar best performance in this criterion, as their training processes maximize likelihood values. However, the all-in algorithm has noticeably worse predictive accuracy, though comparable to mixed-effects models' in testing set. The all-in algorithm might suffer from identifiability issues due to a large set of candidate locations, however reducing its size might lead to the risk of model misspecification.

Figure 3 depicts the predicted locations as well as two important service level metrics based on the estimation results using the all-in algorithm, the location-discovery algorithm and the best-performing mixed-effects model trained under partition $k = 80$. The locations predicted by the location-discovery algorithm share certain similarities with those under the all-in algorithm that are created with point-of-interest data. Since the mixed-effects model imposes a (linear) relationship between location weights and covariates, it often leads to similar weights for nearby locations when they have similar covariates, e.g., population densities. Figures 3a, 3b and 3c show the average walking distance at each rider's arrival location when the rider decides to book a bike. Formally, for each location $l \in \mathcal{L}$, this quantity can be computed as $(1/N) \cdot \sum_{n=1}^{N} \sum_{b=1}^{B} \hat{p}_{l,b,t_n} d_{l,b,t_n} / (1 - \hat{p}_{l,0,t_n})$. Figure 3d, 3e and 3f depict the stockout ratio of each location, which is the probability that an arriving rider chooses to leave without picking up a bike. These two metrics do not have to be perfectly (positively) correlated. For example, according to Figures 3b and 3e, the arrival locations near Lake Washington (circled in red) have a moderate walking distance (300 to 600 meters) but a relatively high stockout ratio (> 0.1) compared to other locations. This likely suggests that there



(a) All-in algorithm (b) Discovery algorithm (c) Mixed-effects model

(d) All-in algorithm (e) Discovery algorithm (f) Mixed-effects model

Figure 3 **The average walking distance and bike stockout ratio for each arrival location under different algorithms.** Different colors represent different average walking distances (the first row) or stockout ratios (the second row). Each circle represents a discovered location. For a clear view, the radius of each circle is computed as a log-linear transformation of the weight in that location (a larger radius corresponds to a larger weight). Specifically, the radius for location l in the plot is $\log(\hat{w}_l) + 9$.

is a scarcity of bikes but once there is a bike, it is often close to the rider's location. All algorithms reveal a high service level in the downtown area. However, the location-discovery algorithm as well as the mixed-effects model suggest there may exist a shortage near Queen Anne (circled in blue).

6. Concluding Remarks

In this paper, we introduce a locational demand model and estimation algorithm for bike-sharing systems that recover rider arrival locations and intensities using only booking and vehicle availability data. We establish identifiability conditions and develop an efficient EM algorithm, enhanced by a location-discovery procedure for scalability in large metropolitan areas. Our approach is also suitable in other contexts such as ridesharing, car rental or retail where customers are originated over space and the locational aspect of the demand is pronounced.

One might question whether a purely machine-learning-based model (e.g., gradient boosting) could outperform the proposed locational demand model. We believe the answer depends on the specific use case. For instance, machine-learning models often excel at temporal forecasting of future demand or bookings. However, they may lack causal interpretability and produce unreliable

predictions when responding to interventions such as capacity changes—a task where our model demonstrates superior performance, as shown in Table 3. Additionally, a key advantage of the locational demand model is its ability to provide interpretable and reliable inputs for downstream decision-making tasks, such as capacity planning and bike rebalancing. For example, our predicted demand locations and corresponding arrival intensities can serve as pre-specified gathering points (i.e., demand-generating locations) in the framework proposed by Luo et al. (2022), who address the rebalancing challenges faced by Mobike, China's largest dockless bikesharing company.

Future studies can be directed toward addressing the following gaps. First, as recognized by He et al. (2021) and Kabra et al. (2019), there can be endogenity issues introduced by unobservable factors such as the political importance of each arrival location affecting its nearby bike availability. Features like walking distance to the closest bike thus can be correlated with these unobserved factors. In addition, our experiments can be enhanced by including additional features such as battery usage and the maintenance status of the bikes which help to explain rider choice behaviors. In a docked-based or hybrid system, features like the number of available bikes in a dock or convenience of the drop-off location (docks are needed for a dock-based system but not for a free-floating system) can also influence rider decisions.

Acknowledgments

The first and second authors gratefully acknowledge the support of Pacific Northwest Transportation Consortium (PacTrans) and the Seattle Department of Transportation (SDOT) which provide research funding and support for this project.

References

Abdallah T, Vulcano G (2020) Demand estimation under the multinomial logit model from sales transaction data. *Manufacturing & Service Operations Management* 23(5):1196–1216.

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

Banerjee S, Freund D, Lykouris T (2022) Pricing and optimization in shared vehicle systems: An approximation framework. *Operations Research* 70(3):1783–1805.

Bhat CR (1997) An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science* 31(1):34–48.

Bureau of Transportation Statistics (2024) Bikeshare and e-scooters in the u.s. `https://data.bts.gov/stories/s/Bikeshare-and-e-scooters-in-the-U-S-/fwcs-jprj/`, accessed: 2024-08-28.

Cho S, Ferguson M, Pekgün P, Vakhutinsky A (2023) Estimating personalized demand with unobserved no-purchases using a mixture model: An application in the hotel industry. *Manufacturing & Service Operations Management* 25(4):1245–1262.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.

El-Assi W, Mahmoud MS, Habib KN (2017) Effects of Built Environment And Weather on Bike Sharing Demand: A Station Level Analysis of Commercial Bike Sharing In Toronto. *Transportation* 44(3):589–613.

Frank M, Wolfe P (1956) An algorithm for quadratic programming. *Naval Research Logistics Quarterly* 3(1-2):95–110.

Freund D, Henderson SG, O'Mahony E, Shmoys DB (2019a) Analytics and bikes: Riding tandem with motivate to improve mobility. *INFORMS Journal on Applied Analytics* 49(5):310–323.

Freund D, Henderson SG, Shmoys DB (2019b) Bike sharing. Hu M, ed., *Sharing Economy: Making Supply Meet Demand*, volume 6 of *Springer Series in Supply Chain Management*, 435 – 459 (Springer).

Greene WH, Hensher DA (2003) A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological* 37(8):681–698.

Grün B, Leisch F (2008) Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification* 25(2):225–247.

He P, Zheng F, Belavina E, Girotra K (2021) Customer preference and station network in the london bike-share system. *Management Science* 67(3):1392–1412.

Hunter DR, Lange K (2004) A tutorial on MM algorithms. *The American Statistician* 58(1):30–37.

Jagabathula S, Subramanian L, Venkataraman A (2020) A conditional gradient approach for nonparametric estimation of mixing distributions. *Management Science* 66(8):3635–3656.

Kabra A, Belavina E, Girotra K (2019) Bike-share systems: Accessibility and availability. *Management Science* 66(9):3803–3824.

Luo X, Li L, Zhao L, Lin J (2022) Dynamic intra-cell repositioning in free-floating bike-sharing systems using approximate dynamic programming. *Transportation Science* 56(4):799–826.

McFadden D, Train K (2000) Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15(5):447–470.

Newman JP, Ferguson ME, Garrow LA, Jacobs TL (2014) Estimation of choice-based models using sales data from a single firm. *Manufacturing & Service Operations Management* 16(2):184–197.

O'Mahony E (2015) *Smarter tools for (Citi) bike sharing*. Ph.D. thesis, Cornell University.

O'Mahony E, Shmoys DB (2015) Data Analysis and Optimization for (Citi) Bike Sharing. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 687–694.

Open Mobility Foundation (2022) Mobility-data-specification: A data standard to enable right-of-way regulation and two-way communication between mobility companies and local governments. `https://github.com/openmobilityfoundation/mobility-data-specification/`, accessed: 2022-11-29.

Paul A, Flynn K, Overney C (2023) Estimating censored spatial-temporal demand with applications to shared micromobility. *arXiv preprint arXiv:2303.09971* .

Rixey RA (2013) Station-level forecasting of bikesharing ridership: Station network effects in three us systems. *Transportation Research Record* 2387(1):46–55.

Seattle Department of Transportation (2022) Scooter share data & permitting. `https://www.seattle.gov/transportation/projects-and-programs/programs/new-mobility-program/scooter-share`, accessed: 2023-05-14.

Singhvi D, Singhvi S, Frazier PI, Henderson SG, O'Mahony E, Shmoys DB, Woodard DB (2015) Predicting Bike Usage for New York City's Bike Sharing System. *AAAI Workshop On Computational Sustainability*.

Steeneck D, Eng-Larsson F, Jauffred F (2022) Estimating lost sales for substitutable products with uncertain on-shelf availability. *Manufacturing & Service Operations Management* 24(3):1578–1594.

Talluri K, van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. *Management Science* 50(1):15–33.

van der Vaart AW (2000) *Asymptotic statistics*, volume 3 (Cambridge University Press).

van Ryzin G, Vulcano G (2014) A Market Discovery Algorithm to Estimate a General Class of Nonparametric Choice Models. *Management Science* 61(2):281–300.

Vulcano G, Van Ryzin G, Ratliff R (2012) Estimating Primary Demand for Substitutable Products from Sales Transaction Data. *Operations Research* 60(2):313–334.

Wu CF (1983) On the convergence properties of the EM algorithm. *The Annals of Statistics* 11(1).

Yuille AL, Rangarajan A (2003) The Concave-Convex Procedure. *Neural Computation* 15(4):915–936.

Appendix A: Dock-based and Hybrid System

We discuss how our methods can be adapted to a dock-based or hybrid system.

Modeling. We define $\mathcal{M} = \{1, 2, \cdots, M\}$ as the set of all docks. We further define a *dock pattern* as the set of available docks at some time (a dock is available if it contains at least one available bike). Instead of observing bike patterns, the operator observes dock patterns during the arrival period. As a result, the choice probability $p_{l,m,S_t}, m \in \mathcal{M}$ becomes the probability that a rider at location l chooses dock m at time t. Note that the structure of the likelihood function remains unchanged given the choice probability. We can then implement both the all-in algorithm and the location-discovery algorithm without much modification. This also extends to a hybrid system where a rider's consideration set may consist of bikes from both docked and free-floating systems. In this case, we modify the definition of bike/dock pattern in a similar way to let it include both docks and free-floating bikes.

Identifiability. We now discuss the identifiability of the model under dock-based systems. We analyze cases where riders make dock choices according to the distance-ranking model (2). These results bypass the general identifiability and consistency results stated in Theorem 1 and Proposition EC.2, and give intuitive conditions under which the MLE of location weights has strong consistency. We first analyze a stylized case where the service region $\mathcal{P} \subset \mathbb{R}$ is a one-dimensional

line segment. For any location $x \in P$, let $r(x)$ be the consideration radius of riders arriving at location x. We assume that $r(\cdot)$ is large enough so that when every dock has at least one available bike, riders will always choose a bike regardless of where they arrive in $\mathcal{P}$. For convenience, we assume that all possible dock patterns are observed with a positive fraction of time as $T \to \infty$, although strictly speaking this can be relaxed as we show in its proof.

THEOREM 2. *In a one-dimensional service region, suppose that the consideration radius $r(\cdot)$ is Lipschitz continuous with constant one, i.e., $\|r(x) - r(x')\| \leq \|x - x'\|$ for all $x, x' \in \mathcal{P}$. Then for each rider location $l \in \mathcal{L}$, if the sequence of docks within its consideration radius is uniquely ordered based on the distance to location l (i.e., without any ties), and this sequence is distinct from the sequences of other rider locations, then with probability one, $\hat{w}_l \to w_l^*$ as $T \to \infty$.*

Theorem 2 shows that in a one-dimensional service region, under smoothly changing consideration radius, we can consistently estimate the location weight whose distance ranking is unique. This result is crisper than the previous ones as we give consistency results for estimating each *individual* rider location weight. The proof of Theorem 2 uses very different techniques from the proofs of Theorem 1 and Proposition EC.2 of Online Appendix and requires deliberately constructing a unique solution of location weights from choice probabilities of certain docks being picked among a set of available docks, which are shown to be consistently estimated. It relies on recovering interesting structures through which the unique distance rankings differ in a one-dimensional space (see Lemmas EC.2 and EC.3 of Online Appendix). Although this one-dimensional result can be stylized in the bike-sharing setting, the model is often used to capture consumer choice behavior in horizontal product differentiation where prices and quality levels are equal across all products, such as yogurt with different flavors and shoes of different colors (see, e.g., Hotelling 1929, Lancaster 1966, 1975 in the economics literature and Gaur and Honhon 2006 in the operations literature on related locational choice models). In service regions of higher dimensions, counter-examples can be established that even rider locations with unique distance rankings can be non-identifiable. Nevertheless, we have the following generic result which states that weights of locations whose first two closest dock is unique can be consistently recovered.

PROPOSITION 1. *For each rider location $l \in \mathcal{L}$, if the ranking of the first two closest docks (or the only dock) in its consideration radius is unique and it is distinct from those of other rider locations, then with probability one, we have $\hat{w}_l \to w_l^*$ as $T \to \infty$.*

Online Appendix

Appendix EC.1: Additional Results

In this section, we give additional results complementing the main text. Section EC.1.1 discusses additional results with respect to the estimation and identifiability of location weights. Section EC.1.2 discusses results regarding the estimation of the choice model parameters. Section EC.1.3 discusses the details of a mixed-effects model predicting the number of bookings across census tracts in Seattle.

EC.1.1. Estimating Location Weights

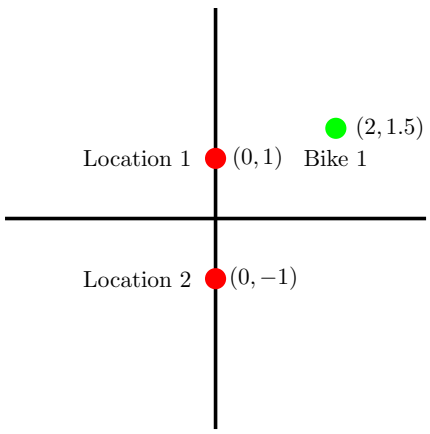
In this subsection, we provide additional results on the indentifiability as well as the consistency of the location weights given choice model parameters. We first give some examples below regarding the identifiability of location weights under the MNL and distance-ranking models to illustrate Theorem 1.

Example EC.1 (Identifiability). Suppose that we have two rider arrival locations shown in Figure EC.1 (red circles). We consider a case with only one bike pattern $\mathcal{S} = \{S\}$. We adopt Euclidean distance when calculating walking distances. In Figure EC.1a, we begin by examining a scenario in which a single bike (green circle) remains fixed at a specific location throughout the arrival period until it is booked. In this scenario, the location weights cannot be consistently estimated with any choice model that only depends on distances. Intuitively, riders only have two options—book the bike or leave. Therefore, the only observation here is the booking time, which results in the weights being non-identifiable. On the other hand, when we have two bikes fixed at two distinct locations, the identification depends on the specification of the choice model. By Theorem 1, the weights can be identified if and only if $p_{1,1,S}p_{2,2,S} \neq p_{1,2,S}p_{2,1,S}$, i.e., vectors $[p_{1,1,S}, p_{2,1,S}]$ and $[p_{1,2,S}, p_{2,2,S}]$ are linearly independent. For the MNL model specified in equation (1), this can be simplified to $\beta_{1,1}(d_{1,1,S} - d_{1,2,S}) \neq \beta_{1,2}(d_{2,1,S} - d_{2,2,S})$, where $d_{l,b,S}$ is the distance from location l to bike b under pattern S. For the distance-ranking choice model specified in equation (2), the necessary and sufficient conditions for the location weights to be identifiable are: (i) the closer bike to locations 1 and 2 are different; (ii) for both locations 1 and 2, at least one bike is within their consideration radiuses. To ease the presentation, we assume that $\beta_{0,1} = \beta_{0,2}$ and $\beta_{1,1} = \beta_{1,2}$ in the MNL choice model, and the consideration radius r is infinite in the distance-ranking choice model. Figure EC.1b illustrates a scenario in which non-identifiability of location weights occurs under any choice model since bikes 1 and 2 both have the same distance to locations 1 and 2. Figure EC.1c gives a scenario where location weights are identifiable under both MNL and distance-ranking choice models. Finally, in Figure EC.1d, location weights are identifiable under the MNL model but not the distance-ranking choice model. This is because, in the latter, a rider always chooses bike 1 over bike 2 regardless of where she arrives. It is not hard to prove that in this example of two rider locations and two bike locations, if location weights are identifiable under the distance-ranking model, they must be identifiable under the MNL model as well. Interestingly, this is *not* true in general. Here, we provide another example that the location weights under the distance-ranking model are identifiable while they are not under the MNL model. Again, we use the Euclidean distance as our distance metric. Consider the bike and rider locations depicted in Figure EC.2. Denote the bike pattern by S. We let $\beta_{1,1} = \beta_{1,2} = -\ln(2)/(2\sqrt{3} - 2)$ and $\beta_{1,3} = (\ln(3) - \ln(4))/(\sqrt{7} - \sqrt{3})$ and consider an asymptotic scenario where $\beta_{0,1}, \beta_{0,2}, \beta_{0,3} \to \infty$. Then for the MNL model, we have $[p_{1,1,S}, p_{2,1,S}, p_{3,1,S}] = [0.4, 0.2, 0.3]$, $[p_{1,2,S}, p_{2,2,S}, p_{3,2,S}] = [0.2, 0.4, 0.3]$ and $[p_{1,3,S}, p_{2,3,S}, p_{3,3,S}] = [0.4, 0.4, 0.4]$, which leads to non-identifiability according to Theorem 1 since these vectors are linearly dependent. On the other hand, consider a distance-ranking model with infinite consideration radius. We have choice probabilities as $[p_{1,1,S}, p_{2,1,S}, p_{3,1,S}] = [0.5, 0, 0]$, $[p_{1,2,S}, p_{2,2,S}, p_{3,2,S}] = [0.0, 0.5, 0]$ and $[p_{1,3,S}, p_{2,3,S}, p_{3,3,S}] = [0.5, 0.5, 1]$ that are linearly independent. This implies that the model is identifiable.

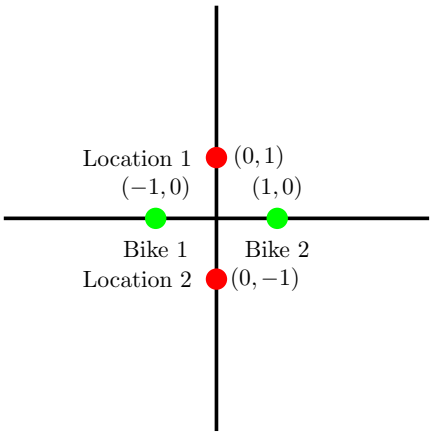
Theorem 1 in the main text described the identifiability of all location weights. Here we give another result regarding partial identifiability of the weight of a particular location $l \in \mathcal{L}$. Let $\boldsymbol{e}_l \in \{0,1\}^L$ be a binary vector whose l^{th} entry is one and zero otherwise and $\mathbb{1}_L$ be an L-dimensional vector with all ones.

Proposition EC.1 **(Partial Identifiability)**. *For any $l \in \mathcal{L}$, if both $\boldsymbol{e}_l$ and $\mathbb{1}_L$ are a linear combination of vectors $\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}] : b \in \mathcal{B}_k,\ k \in \{1, \cdots, K\}\}$, then w_l is identifiable.*

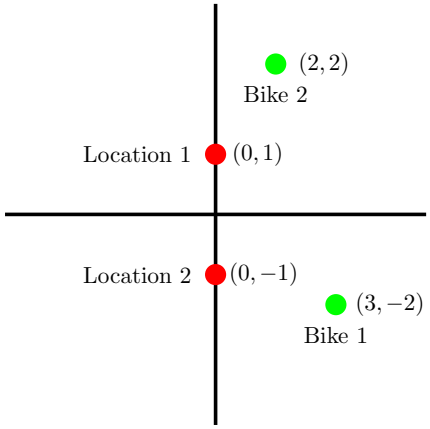
One way to interpret Proposition EC.1 and its proof is to think of all locations other than location l as a whole. Then the problem can be simplified into a scenario only containing two arrival locations with weights w_l and $1 - w_l$, thereby leading to partial identifiability by invoking Theorem 1. Proposition EC.1



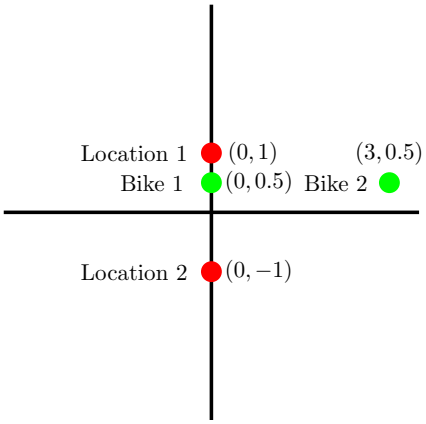
(a) Non-identifiable under any choice model



(b) Non-identifiable under any choice model based on distance



(c) Identifiable under both MNL and ranking model



(d) Identifiable under MNL but not under the ranking model

Figure EC.1 **Examples of identifiability of location weights with only one bike pattern.** In the graphs, red circles represent arrival locations and green circles represent bike locations.

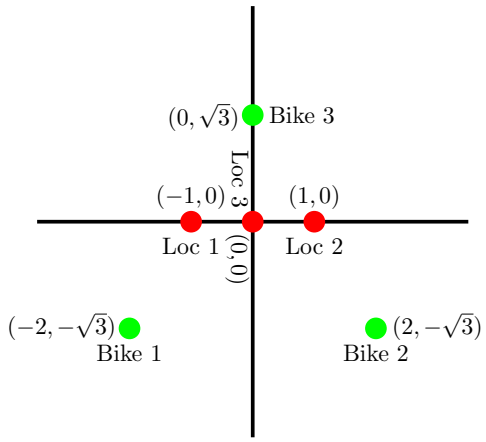


Figure EC.2 **Identifiable under a distance-ranking model but non-identifiable under the MNL model for certain parameter values.**

is particularly useful under a distance-ranking model as the choice probabilities are mostly 0 or 1. We now give an example below to illustrate Proposition EC.1.

EXAMPLE EC.2. Consider a one-dimensional case (Figure EC.3) with three rider arrival locations and one bike. The rider arrival locations are located at -1, 4, and 5 on the axis from left to right. Suppose that there are two possible bike patterns S_1, S_2 where the bike locates at 0 under S_1 and 3 under S_2. Given a constant consideration radius $r = 3$, we have $[p_{1,1,S_1}, p_{2,1,S_1}, p_{3,1,S_1}] = [1, 0, 0]$ and $[p_{1,2,S_1}, p_{2,2,S_1}, p_{3,2,S_1}] = [0, 1, 1]$. Then by Theorem 1, the location weights $\boldsymbol{w}$ are non-identifiable since these vectors cannot span $\mathbb{R}^3$. However, by Proposition EC.1, we know that $\boldsymbol{e}_1$ and $\mathbb{1}_3$ are a linear combination of $(1, 0, 0)$ and $(0, 1, 1)$, which implies that the w_1 is identifiable.

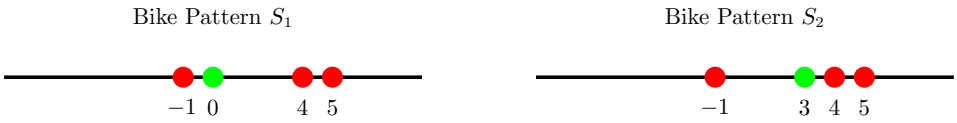


Figure EC.3 **An example of partial identifiability in the distance-ranking model.** As before, red circles represent rider locations and green circles represent bike location. The numbers underneath are coordinates.

We then give sufficient conditions under which the MLE estimator for location weights has strong consistency.

PROPOSITION EC.2 **(Strong Consistency of the MLE)**. *Suppose that the location weights $\boldsymbol{w}$ are identifiable, then the MLE $\hat{\boldsymbol{w}}$ converges to $\boldsymbol{w}^*$ with probability one if one or both of the following conditions hold:*

1. *choice probability $p_{l,b,S_k} > 0$ for all $l \in \mathcal{L}$, $b \in \mathcal{B}_k$, $k \in \{1, \cdots, K\}$;*
2. *there exists $\epsilon > 0$ such that $w_l^* \geq \epsilon$ for all $l \in \mathcal{L}$.*

The first condition holds, for example, under an MNL choice model specified in equation (1). The second condition holds when the operator has precise knowledge of the set of true locations. Either assumption guarantees a finite log-likelihood value for all $\boldsymbol{w}$ in a compact parameter space that contains the true location weights $\boldsymbol{w}^*$, where the log-likelihood function is dominated by an integrable function. Then by the uniform law of large numbers, the uniform convergence of the log-likelihood can be established by the dominance condition together with the continuity of the log-likelihood function (see Theorem 7.48 in Shapiro et al. 2009).

Figure EC.4 displays the objective values in equation (13) of all locations in the square region during the first six iterations of the location discovery algorithm with batch addition, on an instance with $B = 40$, $L^* = 10$, $\lambda = 10$ and $T = 100$. The peak values undergo a sharp slump in the first three iterations and then fluctuate around $1,000$. Moreover, iteration by iteration, the function becomes more multi-modal—the first graph only depicts two local maxima, whereas the last graph exhibits more than five local maxima.

EC.1.2. Estimating Choice Model Parameters

In this subsection, we give more computational details on how the EM algorithm (Algorithm 1) described in Section 4.1 can jointly estimate $\boldsymbol{\beta}$ and $\boldsymbol{w}$. Analogous to equation (6), in the m^{th} iteration, the conditional expectation can be written as

$$\begin{aligned} &\mathbb{E}\left[l_C(\boldsymbol{w}, \boldsymbol{\beta}) \;\middle|\; \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}, \boldsymbol{\beta}^{(m)}\right] \\ =&\mathbb{E}\left[-\tilde{N} + \tilde{N}\log\left(\frac{\tilde{N}}{T}\right) + \sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) + \sum_{n=1}^{\tilde{N}} \log\left(p_{\tilde{l}_n, \tilde{b}_n, S_{\tilde{t}_n}}\right) \;\middle|\; \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}, \boldsymbol{\beta}^{(m)}\right]. \end{aligned} \tag{EC.1}$$

Recall that in equation (EC.1), the first two terms inside the expectation do not depend on $\boldsymbol{w}$ or $\boldsymbol{\beta}$. The third term depends on $\boldsymbol{w}$ but not $\boldsymbol{\beta}$, while the fourth term depends on $\boldsymbol{\beta}$ but not $\boldsymbol{w}$. Hence, $\boldsymbol{w}$ has the same updating process as before with the only difference lies in using $\boldsymbol{\beta}^{(m)}$ to replace the true value. To optimize for $\boldsymbol{\beta}$, we only need to consider the fourth term. To simplify the notation, let $\boldsymbol{\eta}^{(m)} = (\boldsymbol{w}^{(m)}, \boldsymbol{\beta}^{(m)})$

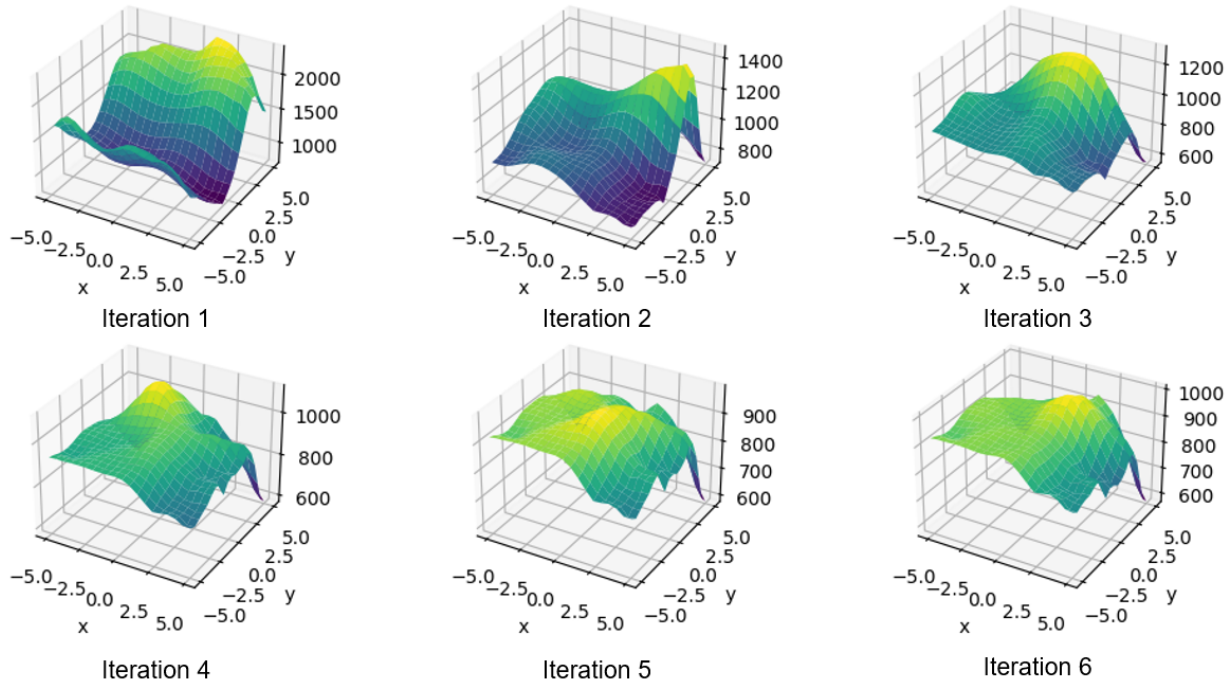


Figure EC.4 **Objective values** (13) **of all locations in the square service region at each iteration of the location-discovery algorithm with batch addition.**

and $\boldsymbol{X}=(\boldsymbol{b},\boldsymbol{t})$. We first derive a few important quantities. The joint conditional density for a rider arriving at location l at time t given that she chooses the leaving option is

$$\begin{aligned} f\left(l,t \;\middle|\; \boldsymbol{\eta}^{(m)}, b=0\right) &= \mathbb{P}\left(l \;\middle|\; t, \boldsymbol{\beta}^{(m)}, \boldsymbol{w}^{(m)}, b=0\right) \cdot f\left(t \;\middle|\; \boldsymbol{\beta}^{(m)}, \boldsymbol{w}^{(m)}, b=0\right) \\ &= \frac{p^{(m)}_{l,0,S_t} w^{(m)}_l}{\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t}} \cdot \frac{\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t}}{\int_0^T \sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t} dt} \\ &= \frac{p^{(m)}_{l,0,S_t} w^{(m)}_l}{\int_0^T \sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t} dt}. \end{aligned}$$

where $p^{(m)}_{l,b,S_t}$, $l\in\mathcal{L}$, $b\in\mathcal{B}$, $t\in[0,T]$ is the rider's choice probability under $\beta^{(m)}_{0,l}$ and $\beta^{(m)}_{1,l}$. Similar to equations (7) and (8), we have

$$\mathbb{P}\left(l_n=l \;\middle|\; \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right) = \frac{p^{(m)}_{l,b_n,S_{t_n}} w^{(m)}_l}{\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',b_n,S_{t_n}}},\ \forall n\in\{1,\cdots,N\},\ \forall l\in\mathcal{L}, \tag{EC.2}$$

$$\mathbb{E}\left[N' \;\middle|\; \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right] = N\frac{\int_0^T \sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t} dt}{\int_0^T \left(1-\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t}\right) dt}. \tag{EC.3}$$

Let the sequence $\{t'_1,t'_2,\cdots,t'_{N'}\}$ denote the arrival time of the unobserved riders. Then we can rewrite the fourth term in the conditional expectation as

$$\begin{aligned} &\mathbb{E}\left[\sum_{n=1}^{\tilde{N}} \log(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}) \;\middle|\; \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right] \\ =&\mathbb{E}\left[\sum_{n=1}^{N} \log(p_{l_n,b_n,S_{t_n}}) \;\middle|\; \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right] + \mathbb{E}\left[\sum_{n=1}^{N'} \log(p_{l'_n,0,S_{t'_n}}) \;\middle|\; \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{l\in\mathcal{L}}\sum_{n=1}^{N}\left(\mathbb{P}\left(l_n = l \mid \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right)\log p_{l,b_n,S_{t_n}}\right) + \mathbb{E}\left[N' \mid \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right]\int_0^T \sum_{l\in\mathcal{L}} f\left(l,t \mid \boldsymbol{\eta}^{(m)}, b=0\right)\log p_{l,0,S_t} dt \\
&= \sum_{l\in\mathcal{L}}\sum_{n=1}^{N}\left(\mathbb{P}\left(l_n = l \mid \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right)\log p_{l,b_n,S_{t_n}}\right) + \mathbb{E}\left[N' \mid \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right]\sum_{l\in\mathcal{L}}\int_0^T f\left(l,t \mid \boldsymbol{\eta}^{(m)}, b=0\right)\log p_{l,0,S_t} dt \\
&= \sum_{l\in\mathcal{L}}\left(\sum_{n=1}^{N}\mathbb{P}\left(l_n = l \mid \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right)\log p_{l,b_n,S_{t_n}} + \mathbb{E}\left[N' \mid \boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right]\int_0^T f\left(l,t \mid \boldsymbol{\eta}^{(m)}, b=0\right)\log p_{l,0,S_t} dt\right).
\end{aligned}
$$

Since the sum of concave functions is concave, we know the above equation is concave as long as the choice probability function p_{l,b,S_t} is log-concave. Substituting equation (EC.2) and (EC.3) into the above equation yields

$$
\mathbb{E}\left[\sum_{n=1}^{\tilde{N}}\log(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}})\middle|\boldsymbol{\eta}^{(m)}, \boldsymbol{X}\right] = \sum_{l\in\mathcal{L}}\left(\sum_{n=1}^{N}\frac{p^{(m)}_{l,b_n,S_{t_n}} w^{(m)}_l}{\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',b_n,S_{t_n}}}\log p_{l,b_n,S_{t_n}} + \frac{N}{s}\int_0^T p^{(m)}_{l,0,S_t} w^{(m)}_l \log p_{l,0,S_t} dt\right),
$$

where $s = \int_0^T (1 - \sum_{l\in\mathcal{L}} w^{(m)}_l p^{(m)}_{l,0,S_t}) dt$.

We generate synthetic data to evaluate this algorithm under an MNL model. We consider two scenarios: $(L^*, B) \in \{(10,40),(25,100)\}$. For simplicity, we assume that $\boldsymbol{\beta}_0$ is known to the operator and $\boldsymbol{\beta}_1$ is to be estimated. We further assume that $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ have the same value across all locations. We thus treat them as scalars (β_0, β_1). We use the same simulation setup described in Section 5.1.2. We test five pairs of ground-truth (β_0^*, β_1^*), which are $(1,-1)$, $(3,-1)$, $(5,-1)$, $(5,-2)$ and $(5,-3)$. When searching the value of β_1, we employ a golden-section search and presume a possible range for β_1 to be $[-10, 0]$. For the K-means algorithm, we utilize a two-step approach. We first estimate the cluster centroids and their weights, and then estimate β_1 using MLE assuming these centroids are rider arrival locations. For the location-discovery algorithm, we terminate when *all* of the following conditions are satisfied. (1) A minimum number of locations N has been discovered. We set $N = 5$ for $(L^*, B) = (10, 40)$ and $N = 10$ for $(L^*, B) = (25, 100)$. (2) Again, we use the first 80% data generated from the arrival period for training and the remaining 20% out-of-sample data for validation. We compute the out-of-sample likelihood in each iteration. In the single mode, the algorithm terminates after two consecutive decreases in likelihood. In the batch mode, the algorithm stops at the first decrease. We generate 10 instances for each scenario and output the average performance among all instances. Performance is reported in Table EC.1. In this table, "MAPE" column reports the mean absolute percentage error between the predicted value of β_1 and its underlying truth, and "Time" column records the CPU times in seconds. The results are summarized in Table EC.1. These observations are similar as those in Table 2. The location-discovery algorithm with batch addition has overall the best performance. It improves over the single mode as evidenced by the relatively lower Wasserstein distance and MAPE of β_1.

We conclude this subsection by discussing the identifiability of $\boldsymbol{\beta}$. By following the proof of Theorem 1, $\boldsymbol{w}$ and $\boldsymbol{\beta}$ are identifiable if and only if the system of nonlinear equations

$$
\sum_{l\in\mathcal{L}} w_l p_{l,b,S_k}/\bar{s} = \sum_{l\in\mathcal{L}} w^*_l p^*_{l,b,S_k}/\bar{s}^*,\ b\in\mathcal{B}_k,\ k\in\{1,2,\cdots,K\} \tag{EC.4}
$$

has a unique solution $\boldsymbol{w} = \boldsymbol{w}^*$, $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. It is expected that this condition has a higher chance to be satisfied as there are more bike patterns. This condition is hard to simplify in general. We provide below a set of simplified sufficient conditions for identifiability of $\boldsymbol{\beta}_1$ under the MNL model when there is only one rider location and $\boldsymbol{\beta}_0$ is assumed to be known.

Proposition EC.3. *Assume that riders only arrive at one given location and riders' choice behavior follows an MNL model with known β_0. β_1 is identifiable if at least one of the following conditions holds:*

1. *At least two bikes have different distances to the rider location in some bike pattern;*
2. *$\beta_1^* < 0$ and one bike has different distances to the rider location in at least two bike patterns where each pattern has the same number of available bikes.*

EC.1.3. A Mixed-effects Model on Seattle Data

In this subsection, we discuss the details of a mixed-effects model that predicts the number of bookings across certain service region partition in Seattle. We construct a 20×20 grid of candidate locations in the service area and exclude those in the water. This gives 307 locations in total. Here, unlike the all-in algorithm,

Table EC.1 Prediction performance when β_1 is unknown.

(β_0^*, β_1^*)	(L^*, B)	Algorithm	Locs	WD	Lkd	Time	MAPE	Algorithm	Locs	WD	Lkd	Time	MAPE
(1, -1)	(10,40)	K-Means	10.0	2.64	-4,519	5	28%	All-In	36.5	2.53	-4,496	196	118%
	(25,100)		25.0	1.93	-7,392	54	23%		37.8	1.79	-7,309	804	91%
(3, -1)	(10,40)		10.0	2.42	-6,543	8	17%		35.4	2.62	-6,472	377	103%
	(25,100)		25.0	1.97	-8,428	66	20%		34.8	1.81	-8,310	1126	61%
(5, -1)	(10,40)		10.0	2.63	-7,178	9	21%		28.4	2.51	-7,174	627	53%
	(25,100)		25.0	1.90	-8,610	67	16%		31.1	1.70	-8,480	1181	26%
(5, -2)	(10,40)		10.0	2.62	-4,108	5	28%		30.0	2.32	-3,888	340	46%
	(25,100)		25.0	1.78	-7,299	58	31%		30.6	1.63	-7,056	1021	25%
(5, -3)	(10,40)		10.0	2.42	-1,821	2	45%		33.1	2.27	-1,647	68	60%
	(25,100)		25.0	1.99	-5,253	36	48%		31.4	1.62	-5,012	513	30%
(1, -1)	(10,40)	Loc.-Disc.	12.0	2.27	-4,389	102	19%	Loc.-Disc.	11.6	2.74	-4,412	52	22%
	(25,100)	(Single)	12.9	1.91	-7,286	294	14%	(Batch)	17.8	1.74	-7,284	229	13%
(3, -1)	(10,40)		11.8	2.23	-6,389	155	13%		12.5	2.20	-6,394	71	20%
	(25,100)		16.9	1.93	-8,308	803	21%		15.8	1.85	-8,312	277	16%
(5, -1)	(10,40)		12.6	2.32	-7,129	175	17%		15.0	2.00	-7,124	87	15%
	(25,100)		15.1	1.85	-8,488	620	11%		15.4	1.83	-8,488	227	11%
(5, -2)	(10,40)		11.2	2.34	-3,855	68	9%		9.0	2.28	-3,903	26	17%
	(25,100)		13.0	1.88	-7,079	270	7%		20.3	1.45	-7,059	291	14%
(5, -3)	(10,40)		7.0	2.66	-1,690	10	20%		7.9	2.68	-1,724	11	27%
	(25,100)		14.4	1.72	-4,992	278	11%		23.8	1.37	-4,973	285	15%

we do not use the popular demand-generating locations because correlations can be introduced with other location-related features as we will discuss below. Inspired by Kabra et al. (2019), we similarly model the arrival rate per half an hour (since the arrival period is half an hour for each day) at candidate location $l \in \mathcal{L}$ using the following linear relationship,

$$\lambda_l := \alpha_0 + \alpha_1 \cdot aba_l + \alpha_2 \cdot rd_l + \alpha_3 \cdot ms_l + \alpha_4 \cdot ts_l, \tag{EC.5}$$

where aba_l is the average number of available bikes within 300 meters of the location over the arrival period and rd_l is the residential population density at the location. We use census tract data from GeoData (2020) to approximate population density. For each candidate location, we compute its population density by taking the total population of the tract it's situated in and dividing it by that tract's area (in square feet). ms_l and ts_l are dummy variables that indicate whether there exists a metro station and a tourist spot, respectively, within 300 meters of the location. Given λ_l, the total arrival rate and location weights are $\lambda = \sum_{l\in\mathcal{L}} \lambda_l$ and $w_l = \lambda_l/\lambda$ respectively. In our model, the response variable is the *total number of bookings* in each partition block during each arrival period (half an hour each day). We recognize that unobserved heterogeneity may arise from inherent variations across different blocks. To reflect this, we add a random effect Z_c to each block $c \in \mathcal{C}_k$ that is drawn from an IID normal distribution. This leads to the following mixed-effects model that estimates the number of bookings $N_{c,d}$ in block $c \in \mathcal{C}_k$ on day d:

$$N_{c,d} = \int_{t\in T_d} \sum_{l\in\mathcal{L}} \sum_{b\in\mathcal{B}_{t,c}} \lambda_l p_{l,b,S_t} dt + Z_c + \epsilon_{c,d}, \tag{EC.6}$$

where $\epsilon_{c,d}$ is the error term for census tract c and day d. To predict the number of bookings under an arbitrary partition k' (which does not necessarily coincide with the partition k used for training), we first approximate the random effects for block $c' \in \mathcal{C}_{k'}$ as $\hat{Z}_{c'} = \sum_{c\in\mathcal{C}_k} \hat{Z}_c A_{c,c'}/A_c$, where A_c is the area of block $c \in \mathcal{C}_k$ and $A_{c,c'}$ is the overlapping area between $c \in \mathcal{C}_k$ and $c' \in \mathcal{C}_{k'}$. Then the number of bookings in $c' \in \mathcal{C}_{k'}$ can be predicted as $\hat{N}_{c'} = \sum_d (\int_{t\in T_d} \sum_{l\in\hat{\mathcal{L}}} \sum_{b\in\mathcal{B}_{t,c'}} \hat{\lambda}_l \hat{p}_{l,b,S_t} dt + \hat{Z}_{c'})$. Again, we set the choice model parameters β_0 and β_1 the same across locations. We choose $\beta_0 = 1$ and estimate the remaining model parameters $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4$ and β_1. Note that since λ_l is a linear function of features, given the choice model parameters, (EC.6) is a linear mixed-effects model. We select β_1 by a line search over $[-3, -8]$ with a step size of 0.2 that minimizes the residual sum of squares of the resulting best-fit linear model.

Figures EC.5 and EC.6 show the weighted distribution of block-level relative errors in the original testing data and the testing data with 30% of bikes randomly removed. For block $c \in \mathcal{C}_k, k \in \{0, 1, \cdots, 99\}$, the relative error is given by $|\hat{N}_{\text{test},c} - N_{\text{test},c}|/N_{\text{test},c}$, and the weight of the frequency is given by $N_{\text{test},c}/N_{\text{test}}$. Here, N_{test} is the total number of bookings during the test period. In the original test data, the location-discovery algorithm has a noticeably shorter tail (highlighted in red) compared to most mixed-effects models

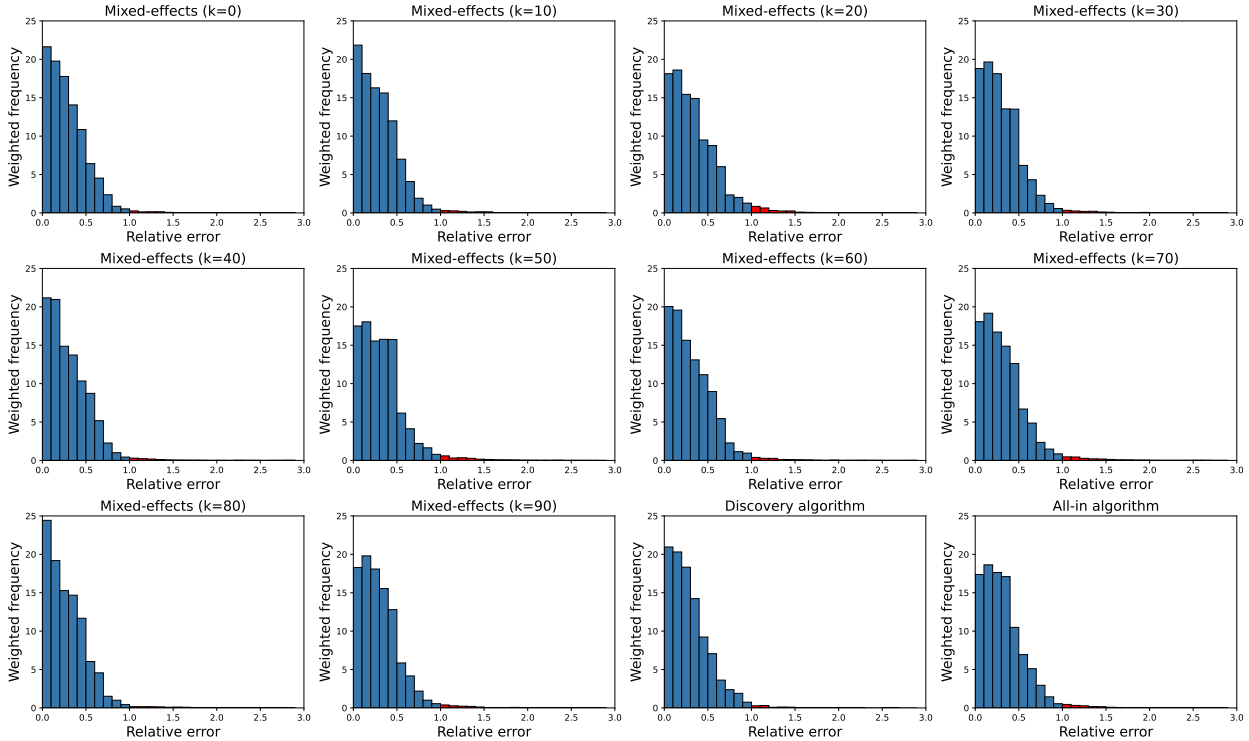


Figure EC.5 Weighted distribution of block-level relative errors in testing data.

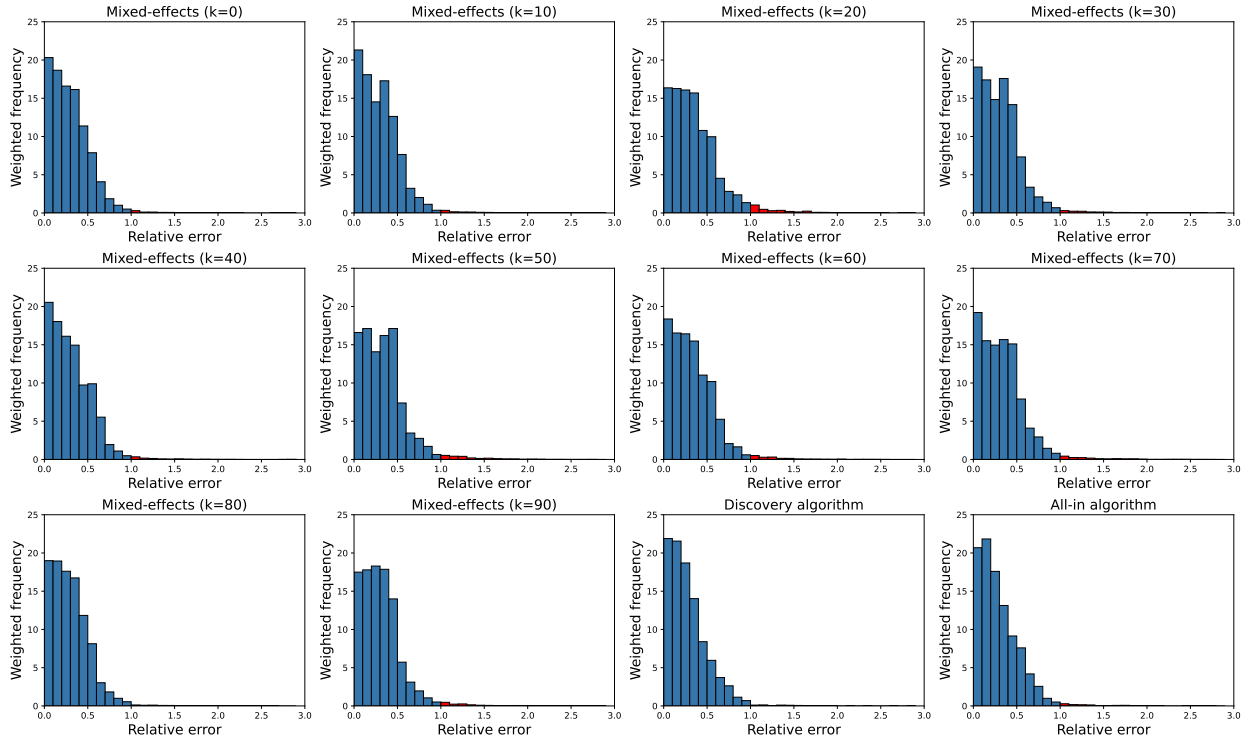


Figure EC.6 Weighted distribution of block-level relative errors in testing data after reducing 30% of bikes.

Feature	Coef.	Std.Err.	z	$P>\|z\|$	[0.025	0.975]
Intercept (α_0)	-0.069	0.022	-3.177	0.001	-0.112	-0.027
Bike availability (α_1)	0.011	0.002	6.587	0.000	0.008	0.015
Population density (α_2)	57.763	6.529	8.848	0.000	44.967	70.559
Metro station (α_3)	0.062	0.087	0.711	0.477	-0.109	0.232
Tourist attraction (α_4)	0.463	0.101	4.567	0.000	0.264	0.662
Group Variance	0.096	0.016				

Table EC.2 Mixed-effects model regression results under the optimal $\beta_1 = -7.8$.

and the all-in algorithm. In the testing data with 30% of bikes removed, the location-discovery algorithm has a higher frequency of small errors compared to all other algorithms. Table EC.2 reports the estimation results of the mixed-effects model under partition $k = 80$. The optimal value of β_1 is -7.8. In the first column, "Group Variance" refers to the estimated variance of the random effect Z_c over all census tracts. The second and third columns report the estimated coefficient and the corresponding standard error for each feature. We observe positive coefficients for all features, which match with intuition. The fourth and fifth columns show the z-score and the p-value for each estimate, and the last two columns provide a 95% confidence interval for the estimated coefficients. We observe the greater significance of bike availability, population density, and tourist attraction, indicating a high impact of these factors on bike usage. On the other hand, the estimated coefficient of the metro station has a relatively large p-value. Table EC.3 shows the WMAPE only on the partition that each mixed-effects is trained on. For example, for a mixed-effect model trained with partition $k = 0$, we compare its WMAPE with the all-in algorithm and the discovery algorithm only on the partition $k = 0$. We observe that 4 out of 10 mixed-effects models outperform the location-discovery algorithm on training data, 5 out of 10 models on the testing data, and 2 out of 10 models on the testing data with 30% of bikes removed.

	All-In			Location-Discovery			Mixed-Effects		
Partition	Train	Test	Test (-30%)	Train	Test	Test (-30%)	Train	Test	Test (-30%)
$k=0$	24.09	35.91	32.81	21.19	33.42	32.09	18.41	32.84	31.73
$k=10$	22.57	31.77	28.95	22.53	29.11	27.40	16.11	26.92	28.52
$k=20$	22.31	33.68	30.28	20.55	31.31	27.81	24.83	35.33	33.73
$k=30$	20.86	32.05	31.10	18.93	30.16	29.05	19.60	29.36	32.20
$k=40$	23.65	32.49	30.08	20.87	30.41	27.36	19.12	29.50	30.29
$k=50$	20.79	29.61	27.82	19.61	28.81	27.74	23.66	30.80	33.58
$k=60$	23.08	33.65	30.33	18.62	31.23	28.21	20.37	31.51	30.35
$k=70$	21.29	28.93	27.08	16.64	26.02	25.42	22.83	30.46	32.29
$k=80$	21.27	33.89	29.26	17.81	31.94	27.75	14.93	26.83	26.63
$k=90$	18.53	31.77	28.59	15.30	26.86	24.41	15.75	29.05	28.90

Table EC.3 WMAPE on the partition that each mixed-effects model is trained on.

Appendix EC.2: Proofs of Technical Results

Proof of Theorem 1. We consider the data-generating distribution of bookings over any length of arrival period T such that $\int_0^T \mathbf{1}(S_t = S_k)dt/T = \alpha_k > 0, \forall k \in \{1, \cdots, K\}$. To prove the sufficiency, we show that there exists a unique maximizer of $\boldsymbol{w}$ and λ corresponding to the true parameters for the *long-run average expected log-likelihood* function, under the condition that the set of vectors $\left\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}]^\intercal : b \in \mathcal{B}_k, k \in \{1, \cdots, K\}\right\}$ spans $\mathbb{R}^L$. That is, we want to show that any maximizer $(\lambda,\boldsymbol{w})$ of the problem $\max_{\lambda,\boldsymbol{w}} \lim_{T\to\infty}(1/T)\mathbb{E}\left[l_I(\lambda, \boldsymbol{w})\right]$ satisfies $(\lambda, \boldsymbol{w}) = (\lambda^*, \boldsymbol{w}^*)$ where λ^* and $\boldsymbol{w}^*$ are the ground truths of arrival rate and location weight vector. Here, the expectation is taken over bookings. We start with the incomplete data log-likelihood function given bookings over a period of length T,

$$l_I(\lambda, \boldsymbol{w}) = -\lambda \int_0^T \left(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t}\right) dt + N \log \lambda + \sum_{n=1}^{N} \log \sum_{l\in\mathcal{L}} w_l p_{l,b_n,S_{t_n}}. \tag{EC.7}$$

Recall that $s = \int_0^T (1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t})dt = \int_0^T \sum_{l\in\mathcal{L}}\sum_{b\in\mathcal{B}} w_l p_{l,b,S_t} dt$ and we define a similar term $s^* := \int_0^T (1 - \sum_{l\in\mathcal{L}} w_l^* p_{l,0,S_t})dt$ corresponding to the ground-truth weights. The expected log-likelihood can then be computed as follows

$$\begin{aligned}\mathbb{E}\left[l_I(\lambda, \boldsymbol{w})\right] &= -\lambda s + \mathbb{E}[N]\left(\log\lambda + \mathbb{E}\left[\log\sum_{l\in\mathcal{L}} w_l p_{l,b_n,S_{t_n}}\right]\right) && \text{(Wald's Lemma)}\\ &= -\lambda s + \lambda^* s^*\left(\log\lambda + \int_0^T \sum_{b\in\mathcal{B}}\left(\left(\frac{\sum_{l\in\mathcal{L}} w_l^* p_{l,b,S_t}}{s^*}\right)\left(\log\sum_{l\in\mathcal{L}} w_l p_{l,b,S_t}\right)\right)dt\right). && \text{(EC.8)}\end{aligned}$$

By following the first-order condition on λ, the unique maximizer $\hat{\lambda}$ of equation (EC.8) satisfies $\hat{\lambda} = \lambda^* s^*/s$. Again, for convenience, we define $\mathbb{E}[l_I(\boldsymbol{w})] := \mathbb{E}[l_I(\hat{\lambda}, \boldsymbol{w})]$,

$$\begin{aligned}\mathbb{E}[l_I(\boldsymbol{w})] &= -\lambda^* s^* + \lambda^* s^* \log\lambda^* - \lambda^* s^* \log\frac{s}{s^*} + \lambda^*\int_0^T \sum_{b\in\mathcal{B}_t}\left(\left(\sum_{l\in\mathcal{L}} w_l^* p_{l,b,S_t}\right)\log\left(\sum_{l\in\mathcal{L}} w_l p_{l,b,S_t}\right)\right)dt\\ &= \lambda^*\int_0^T \sum_{b\in\mathcal{B}_t}\left(\left(\sum_{l\in\mathcal{L}} w_l^* p_{l,b,S_t}\right)\log\left(s^*\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_t}}{s}\right)\right)dt - \lambda^* s^* + \lambda^* s^*\log\lambda^*\\ &= \lambda^* s^*\int_0^T \sum_{b\in\mathcal{B}_t}\left(\left(\sum_{l\in\mathcal{L}}\frac{w_l^* p_{l,b,S_t}}{s^*}\right)\log\left(s^*\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_t}}{s}\right)\right)dt - \lambda^* s^* + \lambda^* s^*\log\lambda^*.\end{aligned}$$

Recall that for each $k \in \{1, \cdots, K\}$, bike pattern S_k appears with a positive fraction of time $\alpha_k > 0$ such that $\sum_{k=1}^K \alpha_k = 1$. Let $\bar{s} := \lim_{T\to\infty}(1/T)\int_0^T (1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t})dt = \sum_{k=1}^K \alpha_k\left(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_k}\right)$ and $\bar{s}^* := \lim_{T\to\infty}(1/T)\int_0^T (1 - \sum_{l\in\mathcal{L}} w_l^* p_{l,0,S_t})dt = \sum_{k=1}^K \alpha_k\left(1 - \sum_{l\in\mathcal{L}} w_l^* p_{l,0,S_k}\right)$ be the long-run averages. It is clear that $\lim_{T\to\infty} s^*/s = \bar{s}^*/\bar{s}$. Let $\mathcal{B}_k$ be the set of available bikes under bike pattern S_k. We can then write the long-run average expected log-likelihood function as

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[l_I(\boldsymbol{w})\right]\\ &= \lim_{T\to\infty}\lambda^*\bar{s}^*\int_0^T \sum_{b\in\mathcal{B}_k}\left(\left(\sum_{l\in\mathcal{L}}\frac{w_l^* p_{l,b,S_t}}{T\bar{s}^*}\right)\log\left(\bar{s}^*\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_t}}{\bar{s}}\right)\right)dt - \lambda^*\bar{s}^* + \lambda^*\bar{s}^*\log\lambda^*\\ &= \lim_{T\to\infty}\lambda^*\bar{s}^*\frac{1}{T}\int_0^T \sum_{b\in\mathcal{B}_t}^{B}\left(\left(\sum_{l\in\mathcal{L}}\frac{w_l^* p_{l,b,S_t}}{\bar{s}^*}\right)\left(\log\bar{s}^* + \log\left(\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_t}}{\bar{s}}\right)\right)\right)dt - \lambda^*\bar{s}^* + \lambda^*\bar{s}^*\log\lambda^*\\ &= \lim_{T\to\infty}\lambda^*\bar{s}^*\frac{1}{T}\int_0^T \sum_{b\in\mathcal{B}_t}\left(\left(\sum_{l\in\mathcal{L}}\frac{w_l^* p_{l,b,S_t}}{\bar{s}^*}\right)\log\left(\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_t}}{\bar{s}}\right)\right)dt + \underbrace{\lambda^*\left(-\bar{s}^* + \bar{s}^*\log\lambda^* + \bar{s}^*\log\bar{s}^*\right)}_{c_0}\\ &= \lambda^*\bar{s}^*\sum_{k=1}^K\left(\alpha_k\sum_{b\in\mathcal{B}_k}\left(\sum_{l\in\mathcal{L}}\left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*}\right)\log\left(\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_k}}{\bar{s}}\right)\right)\right) + c_0, && \text{(EC.9)}\end{aligned}$$

where c_0 is a constant that is independent of $\boldsymbol{w}$. Due to the fact that

$$\sum_{k=1}^K\left(\alpha_k\sum_{b\in\mathcal{B}_k}\left(\sum_{l\in\mathcal{L}}\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*}\right)\right) = \sum_{k=1}^K\left(\alpha_k\sum_{b\in\mathcal{B}_k}\left(\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_k}}{\bar{s}}\right)\right) = 1,$$

we can use Gibb's inequality to derive an upper bound of the likelihood function by adding and subtracting a term α_k inside the logarithm function in equation (EC.9),

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[l_I(\boldsymbol{w})\right] = \lambda^*\bar{s}^*\sum_{k=1}^K\sum_{b\in\mathcal{B}_k}&\left(\alpha_k\sum_{l\in\mathcal{L}}\left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*}\right)\log\left(\alpha_k\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_k}}{\bar{s}}\right)\right)\\ &- \lambda^*\bar{s}^*\sum_{k=1}^K\sum_{b\in\mathcal{B}_k}\left(\alpha_k\sum_{l\in\mathcal{L}}\left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*}\right)\log\alpha_k\right) + c_0\end{aligned}$$

$$
\begin{aligned}
&\leq \lambda^* \bar{s}^* \sum_{k=1}^{K} \sum_{b \in \mathcal{B}_k} \left(\alpha_k \sum_{l \in \mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \left(\alpha_k \sum_{l \in \mathcal{L}} \frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \right) \\
&\quad - \lambda^* \bar{s}^* \sum_{k=1}^{K} \sum_{b \in \mathcal{B}_k} \left(\alpha_k \sum_{l \in \mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \alpha_k \right) + c_0 \qquad \text{(Gibb's Inequality)} \\
&= \lambda^* \bar{s}^* \sum_{k=1}^{K} \sum_{b \in \mathcal{B}_k} \left(\alpha_k \sum_{l \in \mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \sum_{l \in \mathcal{L}} \frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) + c_0.
\end{aligned}
$$

The equality holds if and only if

$$
\sum_{l \in \mathcal{L}} \frac{w_l p_{l,b,S_k}}{\bar{s}} = \sum_{l \in \mathcal{L}} \frac{w_l^* p_{l,b,S_k}}{\bar{s}^*}, \tag{EC.10}
$$

for all $b \in \mathcal{B}_k$ and $k \in \{1, \cdots, K\}$. Now consider equations (EC.10) as a system of linear equations with respect to $\boldsymbol{w}/\bar{s}$. Since the set of vectors $\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}] : b \in \mathcal{B}_k, k \in \{1, \cdots, K\}\}$ spans the space $\mathbb{R}^L$, then the coefficient matrix for this system of linear equations has full column rank L. Thus, (EC.10) has no solution or exactly one solution. On the other hand, observe that $\boldsymbol{w}/\bar{s} = \boldsymbol{w}^*/\bar{s}^*$ is a solution. This implies that it is a unique solution. Since $1/\bar{s} = \sum_{l \in \mathcal{L}} w_l/\bar{s} = \sum_{l \in \mathcal{L}} w_l^*/\bar{s}^* = 1/\bar{s}^*$, this implies $\bar{s} = \bar{s}^*$ and thus $\boldsymbol{w} = \boldsymbol{w}^*$ is the unique solution to (EC.10) satisfying $\sum_{l \in \mathcal{L}} w_l = 1$. Since the unique maximizer of λ is $\lambda^* s^*/s$, we have proved that $(\lambda^*, \boldsymbol{w}^*)$ is the unique maximizer for $\lim_{T \to \infty}(1/T)\mathbb{E}\left[l_I(\lambda, \boldsymbol{w})\right]$, which implies the identifiability of the model. The sufficiency is trivial. To show the necessity under the condition $w_l^* = 0$ for at most one $l \in \mathcal{L}$, we establish the existence of some $\boldsymbol{w} \neq \boldsymbol{w}^*$ such that $\boldsymbol{w}$ corresponds to the same generating distribution as $\boldsymbol{w}^*$ if the set of vectors $\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}] : b \in \mathcal{B}_k, k \in \{1, \cdots, K\}\}$ cannot span the space $\mathbb{R}^L$. We start by finding a solution set $\{w_1, \cdots, w_N, \bar{s}\}$ other than the true values regarding the following system of linear equations

$$
\begin{cases}
\sum_{l \in \mathcal{L}} w_l p_{l,b,S_k} - \bar{s} \sum_{l \in \mathcal{L}} w_l^* p_{l,b,S_k}/\bar{s}^* = 0, & \forall b \in \mathcal{B}, k \in \{1, \cdots, K\}, \\
\sum_{l \in \mathcal{L}} w_l = 1.
\end{cases} \tag{EC.11}
$$

Note that the constraint $\bar{s} = \sum_{k=1}^{K} \left(\alpha_k \sum_{b \in \mathcal{B}_k} \sum_{l \in \mathcal{L}} w_l p_{l,b,S_k} \right)$ is implied from (EC.11). Since $\sum_{l \in \mathcal{L}} w_l^* p_{l,b,S_k}/\bar{s}^*$ is a linear combination of $p_{1,b,S_k}, \cdots, p_{L,b,S_k}$, we know the dimension of the vector space generated by $\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}, \sum_{l \in \mathcal{L}} w_l^* p_{l,b,S_k}/\bar{s}^*] : b \in \mathcal{B}_k, k \in \{1, \cdots, K\}\}$ is strictly less than L. Hence, the rank of the coefficient matrix of the system of linear equations (EC.11) is strictly less than $L+1$. This implies that the system has infinitely many solutions—there exists a nonzero $\theta \in \mathbb{R}^L$ such that $\boldsymbol{w}^* + \beta\boldsymbol{\theta}$ satisfies (EC.11) for all $\beta \in \mathbb{R}$. Since we have $w_l^* = 0$ for at most one $l \in L$, there always exists some small $\beta \neq 0$ such that the resulting solution $\boldsymbol{w}' = \boldsymbol{w}^* + \beta\boldsymbol{\theta}$ is feasible but differs from the underlying truth $\boldsymbol{w}^*$. We now show that $\boldsymbol{w}'$ and $\boldsymbol{w}^*$ imply the same data-generating distribution of booking data. We consider the data-generating distribution of bookings over any length of arrival period T such that $\int_0^T \mathbf{1}(S_t = S_k)dt/T = \alpha_k > 0$, $\forall k \in \{1, \cdots, K\}$.

$$
\begin{aligned}
l_I(\boldsymbol{w}) &= -N + N \log N - N \log \int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t} \right) dt + \sum_{n=1}^{N} \log \left(\sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}} \right) \\
&= -N + N \log N + \sum_{n=1}^{N} \log \frac{\sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}}}{\int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t} \right) dt} \\
&= -N + N \log N + \sum_{n=1}^{N} \log \frac{\sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}}}{\bar{s}}.
\end{aligned} \tag{EC.12}
$$

Since $\boldsymbol{w}'$ and $\boldsymbol{w}^*$ both satisfy the system of linear equations (EC.11), from (EC.12), we have $l_I(\boldsymbol{w}') = l_I(\boldsymbol{w}^*)$ for any booking data, which violates the identifiability of the model. This completes the proof. □

Proof of Proposition EC.2. We show the result by invoking Theorem 5.3 in Shapiro et al. (2009). In particular, we show that there exists a compact set C that satisfies all four conditions in Theorem 5.3 of Shapiro et al. (2009): (i) the true location weights $\boldsymbol{w}^*$ is contained in C; (ii) the long-run-average expected log-likelihood function $\lim_{T \to \infty}(1/T)\mathbb{E}[l_I(\boldsymbol{w})]$ is finite valued and continuous on C; (iii) the empirical time-averaged log-likelihood function $(1/T)l_I(\boldsymbol{w})$ converges to $\lim_{T \to \infty}(1/T)\mathbb{E}[l_I(\boldsymbol{w})]$ with probability one as $T \to$

∞, uniformly in $\boldsymbol{w} \in C$; and (iv) with probability one, for T large enough the set of maximizers of $(1/T)l_I(\boldsymbol{w})$, $\hat{S}_T$, is nonempty and $\hat{S}_T \subset C$.

If we have $p_{l,b_n,S_{t_n}} > 0$ for all $l \in \mathcal{L}$ and $n \in \{1, \cdots, N\}$, we know that $\sum_{l\in\mathcal{L}} w_l p_{l,b_n,S_{t_n}} > 0$ for all $\boldsymbol{w}$ belonging to the simplex $\{\boldsymbol{w} \geq 0 : \sum_{l\in\mathcal{L}} w_l = 1\}$. Then C can simply be this simplex which is a compact set. Similarly, if we have $w_l^* \geq \epsilon$, $\forall l \in \mathcal{L}$ for some $\epsilon > 0$, the compact set C can be $C := \{w_l \geq \epsilon, \forall l \in \mathcal{L} : \sum_{l\in\mathcal{L}} w_l = 1\}$. Then conditions (i), (ii), and (iv) in Theorem 5.3 are satisfied. It remains to show the uniform convergence of the time-averaged log-likelihood function (condition (iii)).

To show the uniform convergence, we invoke Theorem 7.48 of Shapiro et al. (2009) which gives sufficient conditions of uniform convergence of $(1/T)l_I(\boldsymbol{w})$ to the long-run average expected log-likelihood function $\lim_{T\to\infty}(1/T)\mathbb{E}[l_I(\boldsymbol{w})]$ as $T \to \infty$. It requires that: (a) for any $\boldsymbol{w} \in C$, $l_I(\boldsymbol{w})$ is continuous at $\boldsymbol{w}$ for almost all booking data; (b) $l_I(\boldsymbol{w})$, $\boldsymbol{w} \in C$ is dominated by an integrable function that only depends on data; (c) the sample is IID. Condition (a) clearly holds. We then show that the log-likelihood is dominated by another integrable function on the compact set C. For the case where $p_{l,b,S_k} > 0$ for all $l \in \mathcal{L}, b \in \mathcal{B}_k, k \in \{1, \cdots, K\}$, let $\delta = \min_{l\in\mathcal{L}, b\in\mathcal{B}_k, k\in\{1,\cdots,K\}} p_{l,b,S_k}$. We have

$$\begin{aligned} |l_I(\boldsymbol{w})| &= \left| -N + N\log N - N\log \int_0^T \left(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t}\right) dt + \sum_{n=1}^N \log\left(\sum_{l\in\mathcal{L}} w_l p_{l,b_n,S_{t_n}}\right)\right| \\ &< N + N\log(N) + N|\log T| + N|\log\delta| + N|\log\delta|, \end{aligned} \tag{EC.13}$$

which holds for all $\boldsymbol{w} \in C$. In the case that $w_l^* \geq \epsilon$, $\forall l \in \mathcal{L}$, we have $C = \{w_l \geq \epsilon, \forall l \in \mathcal{L} : \sum_{l\in\mathcal{L}} w_l = 1\}$. With a slight abuse of notation, let $\delta = \min_{l\in\mathcal{L}, b\in\mathcal{B}_k, k\in\{1,\cdots,K\}: p_{l,b,S_k}>0} p_{l,b,S_k}$. We have for all $\boldsymbol{w} \in C$,

$$T\delta\epsilon \leq \int_0^T \left(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t}\right) dt \leq T.$$

This gives that, for all $\boldsymbol{w} \in C$,

$$\begin{aligned} |l_I(\boldsymbol{w})| &= \left| -N + N\log N - N\log \int_0^T \left(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t}\right) dt + \sum_{n=1}^N \log\left(\sum_{l\in\mathcal{L}} w_l p_{l,b_n,S_{t_n}}\right)\right| \\ &< N + N\log(N) + N|\log T| + N|\log\delta\epsilon| + N|\log\delta\epsilon|, \end{aligned} \tag{EC.14}$$

Note that (EC.13) and (EC.14) are both integrable with respect to the data-generating distribution, so the dominance property holds. To meet the IID condition, we can always reshuffle the booking data into IID episodes with equal length T' such that within each episode, the data-generating distribution of bookings satisfies $\int_0^{T'} \mathbf{1}(S_t = S_k)dt/T' = \alpha_k > 0$, $\forall k \in \{1, \cdots, K\}$. In such a way, the asymptotic limits of the length of the arrival period approaching infinity and the number of episodes approaching infinity become equivalent. This completes proving the uniform convergence of the time-averaged log-likelihood function. Since in Theorem 1 we already show that the long-run average expected log-likelihood has a unique maximizer at its true value $\boldsymbol{w}^*$, this completes the proof that $\hat{\boldsymbol{w}}$ converges to $\boldsymbol{w}^*$ with probability one by invoking Theorem 5.3 in Shapiro et al. (2009). □

Proof of Corollary 1. To show the sufficiency, by the Radon–Nikodym Theorem, there exists a density function for the invariant probability measure π, $f(S) > 0, S \in \mathcal{S}$, such that $\pi(A) = \int_A f(S)d\mu$ for any $A \in \mathcal{F}$. With a little abuse of notation, let $\mathcal{B}_S$ be the set of bikes in bike pattern S. Recall that $c_0 = \lambda^*(-\bar{s}^* + \bar{s}^*\log\lambda^* + \bar{s}^*\log\bar{s}^*)$ as defined in the proof of Theorem 1. Similar to equation (EC.9), we have

$$\begin{aligned} \lim_{T\to\infty}\frac{1}{T}\mathbb{E}[l_T(\boldsymbol{w})] &= \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} f(S) \sum_{b\in\mathcal{B}_S} \left(\sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log \sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S}}{\bar{s}}\right) d\mu + c_0 \\ &= \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S) \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log \sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S}}{\bar{s}}\right) d\mu + c_0 \\ &= \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S) \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log \left(f(S) \sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S}}{\bar{s}}\right)\right) d\mu \\ &\quad - \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S) \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log f(S)\right) d\mu + c_0 \end{aligned} \tag{EC.15}$$

Note that the Gibbs inequality is not applicable here due to the continuity of bike pattern distribution. However, a similar result can be derived from the almost positive definite property of Kullback–Leibler divergence. By Lemma 3.1 of Kullback and Leibler (1951),

$$\begin{aligned}(\text{EC.15}) \le \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S) \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*} \right) \log \left(f(S) \sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S}}{\bar{s}^*} \right) \right) d\mu \\ - \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S) \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*} \right) \log f(S) \right) d\mu + c_0 \\ = \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S) \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*} \right) \log \sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S}}{\bar{s}^*} \right) d\mu + c_0.\end{aligned}$$

The equality holds if and only if

$$f(S) \sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S}}{\bar{s}} = f(S) \sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S}}{\bar{s}^*} \iff \sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S}}{\bar{s}} = \sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S}}{\bar{s}^*} \tag{EC.16}$$

holds for all $b\in\mathcal{B}$ and almost all $S\in\mathcal{S}$. By the conditions in the statement of the corollary, we know $\boldsymbol{w}=\boldsymbol{w}^*$ is the unique solution for (EC.16), and thus the sufficiency holds.

To show the necessity, following the proof of Theorem 1, there exists $\boldsymbol{w}'\neq\boldsymbol{w}^*$ such that when the observed period $T\to\infty$, $(1/T)l_I(\boldsymbol{w}')-(1/T)l_I(\boldsymbol{w}^*)\to 0$ with probability 1. This contradicts to the identifiability condition, which completes the proof. □

Proof of Proposition EC.1. For any $l\in\mathcal{L}$, by equation (EC.10), it is sufficient to show that the system of linear equations $\sum_{l\in\mathcal{L}} w_l p_{l,b,S_k}/\bar{s} = \sum_{l\in\mathcal{L}} w_l^* p_{l,b,S_k}/\bar{s}^*$, $\forall k\in\{1,\cdots,K\}$ and $b\in\mathcal{B}_k$ has a unique solution $w_l=w_l^*$. Since $\boldsymbol{e}_l$ is a linear combination of vectors $\{[p_{1,b,S_k},\cdots,p_{N,b,S_k}]: b\in\mathcal{B}_k,\, k\in\{1,\cdots,K\}\}$, we know $w_l/\bar{s}=w_l^*/\bar{s}^*$. If $w_l\neq w_l^*$, we have $\bar{s}\neq\bar{s}^*$. However, since $\mathbb{1}_L$ can also be written as a linear combination of these vectors, we have $\sum_{l\in\mathcal{L}} w_i/\bar{s} = \sum_{l\in\mathcal{L}} w_l^*/\bar{s}^*$. Since $\sum_{l\in\mathcal{L}} w_l = \sum_{l\in\mathcal{L}} w_l^* = 1$, we have $\bar{s}=\bar{s}^*$. This leads to $w_l=w_l^*$. □

Proof of Proposition EC.3. Let $p_{l,b,S_k}, b\in\mathcal{B}_k, k\in\{1,\cdots,K\}$ be the probability that a rider in the only rider location l chooses bike b in bike pattern S_k and p^*_{l,b,S_k} as the corresponding probability under the true value β_1^*. By the proof of Theorem 1, we want to show the equality $p_{l,b,S_k}/\bar{s} = p^*_{l,b,S_k}/\bar{s}^*$ holds for all $b\in\mathcal{B}_k$, $k\in\{1,\cdots,K\}$ only if $\beta_1=\beta_1^*$. To show the necessity, we consider the first case where two distinct bikes $b\neq b'\in S_k$ have different distances $d_{l,b,S_k}\neq d_{l,b',S_k}$ to the rider location l in some bike pattern S_k. Without loss of generality, $d_{l,b,S_k} > d_{l,b,S_{k'}}$. This gives,

$$\frac{p_{l,b,S_k}}{p^*_{l,b,S_k}} = \frac{p_{l,b',S_k}}{p^*_{l,b',S_k}} = \frac{\bar{s}}{\bar{s}^*} \;\Rightarrow\; \beta_1(d_{l,b,S_k}-d_{l,b',S_k}) = \beta_1^*(d_{l,b,S_k}-d_{l,b',S_k}) \;\Rightarrow\; \beta_1=\beta_1^*.$$

We then consider the second case where one bike b in two different bike patterns $\mathcal{B}_k$ and $\mathcal{B}_{k'}$ have different distances $d_{l,b,S_k}\neq d_{l,b,S_{k'}}$ to the rider location l. If there are more than one bike having different distances to the rider location in either pattern, then it returns to the first case and we have $\hat{\beta}_1\to\beta_1^*$. Thus, we only need to show the case when all bikes have the same distance to the rider location in each bike pattern. Let B be the number of available bikes in bike patterns $\mathcal{B}_k$ and $\mathcal{B}_{k'}$. Since each pattern has the same number of available bikes, we have

$$\begin{aligned}&\frac{p_{l,b,S_k}}{p^*_{l,b,S_k}} = \frac{p_{l,b,S_{k'}}}{p^*_{l,b,S_{k'}}} \\ \Rightarrow\quad &\frac{\exp(\beta_0+\beta_1 d_{l,b,S_k})}{B^{-1}+\exp(\beta_0+b_k\beta_1 d_{l,b,S_k})} \frac{B^{-1}+\exp(\beta_0+\beta_1 d_{l,b,S_{k'}})}{\exp(\beta_0+\beta_1 d_{l,b,S_{k'}})} \\ =&\frac{\exp(\beta_0+\beta_1^* d_{l,b,S_k})}{B^{-1}+\exp(\beta_0+\beta_1^* d_{l,b,S_k})} \frac{B^{-1}+\exp(\beta_0+\beta_1^* d_{l,b,S_{k'}})}{\exp(\beta_0+\beta_1^* d_{l,b,S_{k'}})}.\end{aligned}$$

The derivative of the left-hand side of the above equation with respect to β_1 can be written as

$$c_0\exp(\beta_1(d_{l,b,S_k}+d_{l,b,S_{k'}})) \frac{c_0(d_{l,b,S_k}-d_{l,b,S_{k'}}) + d_{l,b,S_k}\exp(\beta_1 d_{l,b,S_{k'}}) - d_{l,b,S_{k'}}\exp(\beta_1 d_{l,b,S_k})}{\left((c_0+\exp(\beta_1 d_{l,b,S_k}))\exp(\beta_1 d_{l,b,S_{k'}})\right)^2},$$

where $c_0 := B^{-1}\exp(-\beta_0)$. Since we have $\beta_1 < 0$, it is not hard to see that the derivative would be strictly positive when $d_{l,b,S_k} > d_{l,b,S_{k'}}$. Therefore, the left-hand side is strictly increasing with β_1, which implies the uniqueness of β_1 that satisfies the above equation. This gives the identifiability of β_1. □

Proof of Theorem 2. We let $\mathcal{M} = \{1, 2, \cdots, M\}$ to be the set of all docks. We first give the formal definition of *distance rankings*. For each rider location $l \in \mathcal{L}$ and dock $s \in \mathcal{M}$, let $d_{l,s}$ be the distance from rider location l to dock s.

DEFINITION EC.1. A distance ranking σ_l for a rider location $l \in \mathcal{L}$ is a sequence of distinct docks $s^{(1)}, s^{(2)}, \cdots, s^{(|\sigma_l|)} \in \mathcal{M}$ defined as follows:
(1) $d_{l,s^{(1)}}, \cdots, d_{l,s^{(|\sigma_l|)}}$ is non-decreasing;
(2) $d_{l,s^{(1)}}, \cdots, d_{l,s^{(|\sigma_l|)}} \le \bar{r}_l$ and $d_{l,s} > \bar{r}_l$ for $s \in \mathcal{M}\backslash\{s^{(1)}, s^{(2)}, \cdots, s^{(|\sigma_l|)}\}$.

We let Σ to be the set of unique distance rankings. We comment that the cardinality $|\Sigma|$ is bounded by $\mathcal{O}(M^{2d+1})$, where d is the dimension of the service region. This is established by Skala (2009) where the author shows that the number of unique distance permutations is bounded by $\mathcal{O}(M^{2d})$. Here we have one degree higher as each distance ranking does not have to be of length M since the consideration radius is limited.

Based on the definition, we further define $\Sigma(\sigma)$ as the set of distance rankings whose first $|\sigma|$ choices are exactly the same as σ. With a slight abuse of notation, we refer to w_σ as the corresponding probability that a rider arrives at a location with distance ranking σ. Then to prove the theorem, it is equivalent to show that the MLE $\hat{w}_\sigma$ is consistent for all $\sigma \in \Sigma$. We use $\mathcal{S}$ to denote all possible dock patterns, i.e., a set of available docks. For each $S \in \mathcal{S}$, we further define $\mathbb{P}_j(S)$ as the probability that a rider chooses to pick up a bike at dock j under some dock pattern S. Similarly, $\mathbb{P}_0(S)$ is defined as the probability that a rider chooses to leave under dock pattern S. Note that $\mathbb{P}_j(S)$ can be written as the sum of all rider location weights who choose to pick up a bike at dock j under dock pattern S. We first prove three useful lemmas. The first one concerns the consistency of the MLE of the arrival rate $\hat{\lambda}$ and choice probabilities $\hat{\mathbb{P}}_j(S)$.

LEMMA EC.1. *The MLEs $\hat{\lambda} \to \lambda^*$ and $\hat{\mathbb{P}}_j(S) \to \mathbb{P}^*_j(S)$ with probability one for all $S \subset \mathcal{M}$ and $j \in S \cup \{0\}$ as $T \to \infty$, where λ^* and $\mathbb{P}^*_j(S)$ are the corresponding true values.*

Proof. Since each dock pattern is observed with a positive fraction when $T \to \infty$, the dock pattern where all bikes are available $S = \mathcal{M}$ is observed with an infinite amount of time. Since $\mathbb{P}_0(\mathcal{M}) = 0$, it is clear that λ can be consistently estimated by simply applying the strong law of large numbers to the renewal process (see, e.g., Proposition 3.3.1 in Ross 1996, note that the MLE of λ is simply the number of total arrivals over the length of the arrival period with $S = \mathcal{M}$). For the same reason, for any other dock pattern $S \in \mathcal{S}$, the observed arrival rate under that pattern $\tilde{\lambda}_S$ can be consistently estimated. We know that $\mathbb{P}_0(S) = 1 - \tilde{\lambda}_S/\lambda$, which can also be consistently estimated. Finally, for any $j \in S$, we have $\mathbb{P}_j(S) = \tilde{\lambda}_{S,j}(1 - \mathbb{P}_0(S))/\tilde{\lambda}_S$ where $\tilde{\lambda}_{S,j}$ is the observed arrival rate for riders who choose dock j under dock pattern S, which can be consistently estimated as well. This completes the proof. □

Lemma EC.1 states that the MLEs of choice probabilities and the arrival rate are strongly consistent. What is left to be shown is thus one can uniquely determine weights of distance rankings from the choice probabilities. We first discuss a few interesting observations regarding how distance rankings vary in a one-dimensional space.

LEMMA EC.2. *Given a subset of docks $S \subset \mathcal{M}$ with $|S| = k$, for any permutation of S, $\left(s^{(1)}, \cdots, s^{(k)}\right)$ such that $\Sigma\left(\left(s^{(1)}, \cdots, s^{(k)}\right)\right) \neq \emptyset$, there exists at most two distinct docks $s^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)}, \cdots, s^{(k)}, s^{(k+1)}\right)\right) \neq \emptyset$. Moreover, there exists at most one permutation $s^{(1)}, \cdots, s^{(k)}$ where $s^{(k+1)}$ can have two distinct options, the rest of the permutations have at most one such $s^{(k+1)}$.*

Proof. Figure EC.7 is an illustration of the proof. Let x_s denote the coordinate of a dock $s \in \mathcal{M}$. It is easy to check the first part of the statement. First of all, for any $s^{(1)} \neq \cdots \neq s^{(k)}$ such that $\Sigma((s^{(1)}, \cdots, s^{(k)})) \neq \emptyset$, they have to form a group such that for all $s \neq s^{(1)}, \cdots, s^{(k)}$, either $x_s < x_{s^{(1)}}, \cdots, x_{s^{(k)}}$ or $x_s > x_{s^{(1)}}, \cdots, x_{s^{(k)}}$. Moreover, $s^{(k+1)}$ has to be the closest one on the left or on the right, i.e., $s^{(k+1)}$ has to be either $s_{\text{right}} = \arg\min_{s \in \mathcal{M}: x_s > x_{s^{(1)}}, \cdots, x_{s^{(k)}}}$ or $s_{\text{left}} = \arg\max_{s \in \mathcal{M}: x_s < x_{s^{(1)}}, \cdots, x_{s^{(k)}}}$.

To see the second part of the statement, suppose that there exists such two choices of $s^{(k+1)} \in \{s_{\text{left}}, s_{\text{right}}\}$ (if not, the statement is trivial). We draw the middle point of the two potential docks of $s^{(k+1)}$, $(s_{\text{left}} + s_{\text{right}})/2$, depicted by the dashed vertical line. The main idea is to show that there exists at most one permutation $(s^{(1)}, s^{(2)}, \cdots, s^{(k)})$, such that the subregion within which riders' first k nearest docks are $(s^{(1)}, s^{(2)}, \cdots, s^{(k)})$, covers the point $(s_{\text{left}} + s_{\text{right}})/2$.

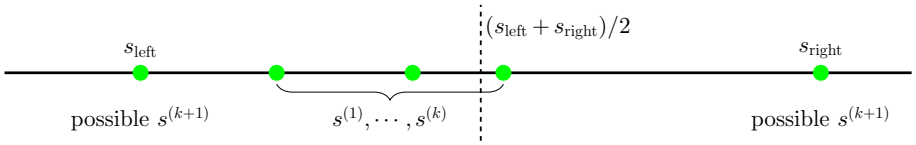


Figure EC.7 Identifiability proof under a one-dimensional space.

For ease of presentation, we define $\mathcal{R}(\sigma)$ as the subregion of the rider locations corresponding to distance ranking σ in the space and $\overline{\mathcal{R}}(\sigma) \supseteq \mathcal{R}(\sigma)$ for the subregion whose first $|\sigma|$ closest docks is exactly σ. There are two situations to discuss:

(1) The middle point $(s_{\text{left}} + s_{\text{right}})/2$ falls into a subregion corresponding to $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ where $s^{(1)}, \cdots, s^{(k)}$ is a particular permutation of S. Then subregion $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ is separated into two parts. The part on the left of $(s_{\text{left}} + s_{\text{right}})/2$ would possibly correspond to distance rankings in $\Sigma((s^{(1)}, \cdots, s^{(k)}, s_{\text{left}}))$ while the part on the right of $(s_{\text{left}} + s_{\text{right}})/2$ would possibly correspond to rankings in $\Sigma((s^{(1)}, \cdots, s^{(k)}, s_{\text{right}}))$. We say "possibly" here because it could be the case that $\Sigma((s^{(1)}, \cdots, s^{(k)}, s_{\text{left}})) = \emptyset$ or $\Sigma((s^{(1)}, \cdots, s^{(k)}, s_{\text{right}})) = \emptyset$ due to limited consideration radius. For any other permutation $s'^{(1)}, \cdots, s'^{(k)}$, such that $\Sigma((s'^{(1)}, \cdots, s'^{(k)})) \neq \emptyset$, subregions corresponding to any $\sigma \in \Sigma((s'^{(1)}, \cdots, s'^{(k)}))$ will be either on the left side of $(s_{\text{left}} + s_{\text{right}})/2$ or on the right side of $(s_{\text{left}} + s_{\text{right}})/2$. Thus $s^{(k+1)}$ will be either s_{left} or s_{right} depending on which side it locates. It can also be the case that $s^{(k+1)}$ does not exist if it is outside their consideration radiuses.

(2) The middle point $(s_{\text{left}} + s_{\text{right}})/2$ does not fall into subregion $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ for any permutation of S: $s^{(1)}, \cdots, s^{(k)}$. Since the union of subregions $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ over all permutations $s^{(1)}, \cdots, s^{(k)}$ form a convex set. This convex set thus locates either on the left side of $(s_{\text{left}} + s_{\text{right}})/2$ or on the right side of $(s_{\text{left}} + s_{\text{right}})/2$, which means that for all permutations, $s^{(k+1)}$ is either s_{left} or s_{right} depending on which side it locates, or $s^{(k+1)}$ simply does not exist due to limited consideration radius. This completes the proof of Lemma EC.2. □

LEMMA EC.3. *For the consideration radius function $r(\cdot)$ such that $\|r(x) - r(x')\| \leq \|x - x'\|, \forall x, x' \in \mathcal{P}$, given a subset of docks $S \subset \mathcal{M}$ with $|S| = k$, there exists at most two distinctive permutations of S, $s^{(1)}, \cdots, s^{(k)}$ and $s'^{(1)}, \cdots, s'^{(k)}$ such that:*

- $\left(s^{(1)}, \cdots, s^{(k)}\right) \in \Sigma$, *and there exists* $s^{(k+1)}$ *with* $\Sigma\left(\left(s^{(1)}, \cdots, s^{(k+1)}\right)\right) \neq \emptyset$.
- $\left(s'^{(1)}, \cdots, s'^{(k)}\right) \in \Sigma$, *and there exists* $s'^{(k+1)}$ *with* $\Sigma\left(\left(s'^{(1)}, \cdots, s'^{(k+1)}\right)\right) \neq \emptyset$.

If the two permutations co-exist, it must be that $s^{(k+1)} \neq s'^{(k+1)}$.

Proof. We prove this by contradiction. We show the second part of the statement first. Suppose that $s^{(k+1)} = s'^{(k+1)}$. Knowing that $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ and $\overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$ are two disjoint convex sets. Without loss of generality, we assume $s^{(k+1)}$ is on the left of $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$. We now consider two situations:

(1) $\overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$ is on the right side of $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$: Because $(s^{(1)}, \cdots, s^{(k)}) \in \Sigma$, there exists $x_0 \in \overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ such that $\|x_0 - x_{s^{(k+1)}}\| > r(x_0)$ (since $\mathcal{R}((s^{(1)}, \cdots, s^{(k)}))$ is non-empty). Let $x^* = \sup_{x \in \overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))} x$ denote the rightmost point in the subregion $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$. Given that $\|r(x^*) - r(x_0)\| = r(x^*) - r(x_0) \leq \|x^* - x_0\| = x^* - x_0$, it is clear that $\|x^* - x_{s^{(k+1)}}\| = x^* - x_{s^{(k+1)}} = \|x^* - x_0\| + \|x_0 - x_{s^{(k+1)}}\| > r(x^*)$. For all $x' \in \overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$:

$$\|x' - x_{s^{(k+1)}}\| \geq \|x^* - x_{s^{(k+1)}}\| > r(x^*).$$

since $\overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$ is on the right-hand side of $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$. In the case of $r(x^*) \geq r(x')$, we immediately get $\|x' - x_{s^{(k+1)}}\| > r(x')$. In the case of $r(x^*) < r(x')$, we know $\|x' - x_{s^{(k+1)}}\| - \|x^* - x_{s^{(k+1)}}\| = \|x' - x^*\| \geq \|r(x') - r(x^*)\| = r(x') - r(x^*)$. Then we have the following:

$$\|x' - x_{s^{(k+1)}}\| - r(x') \geq \|x^* - x_{s^{(k+1)}}\| - r(x^*) > 0.$$

We thus have $\|x' - x_{s^{(k+1)}}\| > r(x')$ for all $x' \in \overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$, which contradicts to the assumption that there exists $s'^{(k+1)}$ with $\Sigma((s'^{(1)}, \cdots, s'^{(k)}, s'^{(k+1)})) \neq \emptyset$.

(2) $\overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$ is on the left-hand side of $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$: This case can be proved by following the exactly same procedure from the first case by symmetry. We only need to swap all notations of $s^{(i)}$ with $s'^{(i)}$ for $i \in \{1, \cdots, k+1\}$.

The first part of Lemma EC.3 can be shown with similar arguments. Suppose that there exists a third different permutation $s''^{(1)},\cdots,s''^{(k)}$ such that $(s''^{(1)},\cdots,s''^{(k)})\in\Sigma$ and there exists $s''^{(k+1)}$ with $\Sigma((s''^{(1)},\cdots,s''^{(k)},s''^{(k+1)}))\neq\emptyset$. From Lemma EC.2 we know that $s''^{(k+1)}\in\{s^{(k+1)},s'^{(k+1)}\}$. However, we just proved that there does not exist two different permutations with the same $(k+1)^{\text{th}}$ dock. This reaches a contradiction and completes the proof. □

We are now ready to prove Theorem 2. We show that there is a unique solution of distance ranking weights from choice probabilities, which can be consistently estimated from Lemma EC.1. In specific, we want to prove for any $1\le k\le M$, and for any $s^{(1)}\neq\cdots\neq s^{(k)}\in\mathcal{M}$, $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)}))}w_\sigma$ and $w_{(s^{(1)},\ldots,s^{(k-1)})}$ are uniquely identifiable by inducting on k. Intuitively, this induction allows us to gradually increase the resolution of the distance rankings we are able to identify. We first prove the initial condition $k=1$. Notice that

$$\sum_{\sigma\in\Sigma((i))}w_\sigma=\mathbb{P}_i(\mathcal{M}),\tag{EC.17}$$

$$\sum_{\sigma\in\Sigma((i,j))}w_\sigma=\sum_{\sigma\in\Sigma((j))}w_\sigma+\sum_{\sigma\in\Sigma((i,j))}w_\sigma-\sum_{\sigma\in\Sigma((j))}w_\sigma=\mathbb{P}_j(\mathcal{M}\backslash\{i\})-\mathbb{P}_j(\mathcal{M}),\tag{EC.18}$$

$$\Longrightarrow\qquad w_{(i)}=\sum_{\sigma\in\Sigma((i))}w_\sigma-\sum_{\sigma\in\Sigma((i,j)),j\in\mathcal{M}}w_\sigma=\mathbb{P}_i(\mathcal{M})-\sum_{j\in\mathcal{M}}\left(\mathbb{P}_j(\mathcal{M}\backslash\{i\})-\mathbb{P}_j(\mathcal{M})\right),\tag{EC.19}$$

for all $i\neq j$ and $i,j\in\{1,\cdots,M\}$. For the induction step at k, suppose that for any positive integer $n\le k$, we have $s^{(1)}\neq\cdots\neq s^{(n)}\in\mathcal{M}$ and $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(n)}))}w_\sigma$ and $w_{(s^{(1)},\cdots,s^{(n-1)})}$ are uniquely identifiable. Then, we want to prove for any $s^{(1)}\neq\cdots\neq s^{(k+1)}\in\mathcal{M}$, $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k+1)}))}w_\sigma$ and $w_{(s^{(1)},\cdots,s^{(k)})}$ are uniquely identifiable. For a given set of k docks $\{s^{(1)},\cdots,s^{(k)}\}$, each permutation of it can be classified into one of the following types:

1. $\left(s^{(1)},\cdots,s^{(k)}\right)\in\Sigma$ and there does not exist $s^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)},\cdots,s^{(k+1)}\right)\right)\neq\emptyset$. Then it is clear that

$$\left(s^{(1)},\cdots,s^{(k)}\right)=\Sigma\left(\left(s^{(1)},\cdots,s^{(k)}\right)\right)$$

Since $\sum_{\sigma\in\Sigma((s^{(1)},\ldots,s^{(k)}))}w_\sigma$ is known according to the induction hypothesis, we thus are able to identify $w_{(s^{(1)},\cdots,s^{(k)})}$ and $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))}w_\sigma$.

2. $\left(s^{(1)},\cdots,s^{(k)}\right)\notin\Sigma$ and there exists only one $s^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)},\cdots,s^{(k+1)}\right)\right)\neq\emptyset$. Then we have

$$\Sigma\left(\left(s^{(1)},\cdots,s^{(k)},s^{(k+1)}\right)\right)=\Sigma\left(\left(s^{(1)},\cdots,s^{(k)}\right)\right)$$

Similarly, we have $w_{(s^{(1)},\cdots,s^{(k)})}=0$ and $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))}w_\sigma=\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)}))}w_\sigma$, which is known according to the induction hypothesis.

3. There exists $s^{(k+1)}$ and $s'^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)},\cdots,s^{(k)},s^{(k+1)}\right)\right)\neq\emptyset$ and $\Sigma\left(\left(s^{(1)},\cdots,s^{(k)},s'^{(k+1)}\right)\right)\neq\emptyset$. With a slight abuse of notation, we define $\Sigma(S,s)$ for some $S\subseteq\mathcal{M}$ and $s\in\mathcal{M}\backslash S$ as the set of distance rankings whose first $|S|$ docks are *any* permutation of S and the $(|S|+1)^{\text{th}}$ dock is s. Put set $S^k=\{s^{(1)},\cdots,s^{(k)}\}$. Then we have the following identity:

$$\begin{aligned}\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))}w_\sigma=&\mathbb{P}_{s^{(k+1)}}\left(\mathcal{M}\backslash S^k\right)-\sum_{S\subsetneq S^k}\sum_{\sigma\in\Sigma((S,s^{(k+1)}))}w_\sigma\\&-\sum_{\sigma\in\Sigma((S^k,s^{(k+1)}))\backslash\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))}w_\sigma.\end{aligned}\tag{EC.20}$$

The second term on the right-hand side is known according to the induction hypothesis. For the third term, by lemma EC.2, since the permutation $(s^{(1)},\cdots,s^{(k)})$ has two distinct options $s^{(k+1)}$ and $s'^{(k+1)}$, any rest of the permutation of S^k, $s''^{(1)},\cdots,s''^{(k)}$, has at most one $s''^{(k+1)}$ such that $\Sigma\left(\left(s''^{(1)},\cdots,s''^{(k+1)}\right)\right)\neq 0$. Clearly, we only need to consider the case when such $s''^{(k+1)}$ exists. Now consider two subcases: 1) if $(s^{(1)},\cdots,s^{(k)})\notin\Sigma$, we claim that $(s''^{(1)},\cdots,s''^{(k)})\notin\Sigma$. We prove by contradiction. Without loss of generality, suppose that $s^{(k+1)}$ and $((s''^{(1)},\cdots,s''^{(k)}))$ are on the left-hand side of $\overline{\mathcal{R}}((s^{(1)},\cdots,s^{(k)}))$. Then if $(s''^{(1)},\cdots,s''^{(k)})\in\Sigma$ it is clear that for all $x\in\overline{\mathcal{R}}((s^{(1)},\cdots,s^{(k)}))$, we have $|x-x_{s^{(k+1)}}|>r(x)$ which contradicts to the fact that $(s^{(1)},\cdots,s^{(k)},s^{(k+1)})\in\Sigma$. This suggests that the permutation $s''^{(1)},\cdots,s''^{(k)}$ belongs to the second type we discussed above and is proven to be identifiable; if $(s^{(1)},\cdots,s^{(k)})\in\Sigma$, then by Lemma EC.3, we know $s'^{(k+1)}\neq$

$s^{(k+1)}$, which implies that the permutation σ in the third term cannot be $(s'^{(1)}, \cdots, s'^{(k)}, s'^{(k+1)})$. Hence, by equation (EC.20), we know that $\sum_{\sigma\in\Sigma\left(\left(s^{(1)},\cdots,s^{(k)},s^{(k+1)}\right)\right)} w_\sigma$ can be uniquely identified. By symmetry, we can show that $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s'^{(k+1)}))} w_\sigma$ is identifiable as well. We can further deduce $w_{(s^{(1)},\cdots,s^{(k)})}$ by

$$w_{(s^{(1)},\cdots,s^{(k)})} = \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)}))} w_\sigma - \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))} w_\sigma - \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s'^{(k+1)}))} w_\sigma.$$

4. The permutation $(s^{(1)}, \cdots, s^{(k)}) \in \Sigma$ and there exists only one $s^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)}, \cdots, s^{(k+1)}\right)\right) \neq \emptyset$. Similarly, we can show that $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k+1)}))} w_\sigma$ is identifiable using the same approach in type 3. We can further deduce $w_{(s^{(1)},\cdots,s^{(k)})}$ by

$$w_{(s^{(1)},\cdots,s^{(k)})} = \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)}))} w_\sigma - \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))} w_\sigma.$$

This completes the induction and proof of Theorem 2. We comment that the proof does not necessarily require all possible dock patterns to be observed with a positive fraction of time in the long run. We only need $\mathcal{M}$ and $\mathcal{M}\setminus S^k$ to be observed for all possible S^k that are constructed by looking at all distance rankings in the set of rider locations $\mathcal{L}$. □

Proof of Proposition 1 The result can be directly shown by equations (EC.17), (EC.18) and (EC.19) in the proof of Theorem 2. □

References

Gaur V, Honhon D (2006) Assortment planning and inventory decisions under a locational choice model. *Management Science* 52(10):1528–1543.

GeoData S (2020) 2020 Census Tract Seattle. `https://data-seattlecitygis.opendata.arcgis.com/datasets/SeattleCityGIS::2020-census-tract-seattle-redistricting-data-1990-2020/`, accessed: 2023-03-17.

Hotelling H (1929) Stability in competition. *The Economic Journal* 39(153):41–57.

Kabra A, Belavina E, Girotra K (2019) Bike-share systems: Accessibility and availability. *Management Science* 66(9):3803–3824.

Kullback S, Leibler RA (1951) On information and sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Lancaster K (1975) Socially optimal product differentiation. *American Economic Review* 65(4):567–585.

Lancaster KJ (1966) A new approach to consumer theory. *Journal of Political Economy* 74(2):132–157.

Ross S (1996) *Stochastic processes*. Wiley series in probability and statistics: Probability and statistics (Wiley).

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures On Stochastic Programming: Modeling and Theory* (SIAM).

Skala M (2009) Counting distance permutations. *Journal of Discrete Algorithms* 7(1):49–61.