A Data Analysis Method for Estimating Balking Behavior in Bike-Sharing Systems

Aditya Ahire and Ashkan Negahban

1 Introduction and Background

Understanding customer behavior can have significant implications for any business. In particular, bike-sharing systems can benefit substantially from better understanding of rider behavior as it would enhance demand estimation. Virtually all critical short- and long-term decisions related to the design and operation of these systems rely on accurate demand estimates. The main decision-making problems in this context include number of stations and their location, station size (number of docks), fleet size (number of bikes), pre-balancing and rebalancing operations, subscription options/pricing and per trip charges. The more accurate the demand estimates, the more appropriate these decisions. As a result, demand analysis is one of the main research areas in the bike-sharing literature.

1.1 Previous Work on Bike-Sharing Demand Analysis

Existing studies on bike-sharing demand analysis can be classified into the following four categories:

- **Sole bike-sharing demand analysis:** These papers solely analyze usage data without considering other factors (e.g., socio-demographics) or other modes of transportation (taxi, rideshare, subway, etc.) (Bordagaray et al., 2016; Vogel et al., 2011; Oppermann et al., 2018; Bargar et al., 2014; Come et al., 2014; Rudloff & Lackner, 2014).

A. Ahire · A. Negahban (✉)
School of Graduate Professional Studies, The Pennsylvania State University, Malvern, PA, USA
e-mail: anegahban@psu.edu

H. Yang et al. (eds.), *AI and Analytics for Public Health*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-030-75166-1_12

- **Multi-factorial demand analysis:** The papers in this category analyse the demand for bike-sharing systems in conjunction with other factors such as weather, socio-demographic and socio-economic factors, leisure travel and infrastructure (bicycle tracks, etc). For a sample list of these studies, see (Caulfield et al., 2016; El-Assi et al., 2017; Singhvi et al., 2015; Tran et al., 2015).
- **Multi-modal demand analysis:** The papers in this category analyse the demand for bike-sharing systems in conjunction with the demand for other modes of transportation, namely taxi, train/subway, and bus system. For a sample of these studies, see (Singhvi et al., 2015; Tran et al., 2015).
- **Demand censoring:** Data on successful bike pickups censor part of the demand from customers that were unable to pick up a bike due to bike unavailability. These studies propose data cleaning/filtering (O'Mahony & Shmoys, 2015), non-parametric (Albiński et al., 2018), and simulation-based inference (Negahban, 2019) methods to address the censoring problem.

1.2 Contribution of This Paper

Existing demand analysis studies primarily focus on *aggregate* demand patterns at the station, region, or city level. To the best of the authors' knowledge, there is no paper on *individual-level* analysis of balking behavior. This paper contributes to the bike-sharing literature by proposing a data analysis method for inferring the balking threshold and timing of balking decision for individual customers from system-generated data on observed bike pickup times (readily available for virtually any bike-sharing system). Since individuals' true balking behavior is unknown and unobservable, we use simulation to mimic customer behavior and generate synthetic data that are similar to those generated by real-world bike-sharing systems. We then apply our method on the simulated data and assess its efficacy by comparing the estimates with the input parameters used in the simulation model to generate the data.

The remainder of the paper is organized as follows. Section 2 describes the estimation problem. The discrete-event simulation model, simulated data, and the proposed data analysis method are presented in Sect. 3. Section 4 summarizes the results. Finally, conclusions and future research opportunities are discussed in Sect. 5.

2 Problem Description: Estimation of Balking Threshold and Timing of Balking Decision

Our goal is to derive insights on users' balking behavior by estimating their balking threshold (in terms of bike availability) and timing of balking decision. In this section, we frame the estimation problem investigated in this paper.

2.1 The Research Subjects and Their Balking Behavior

We consider a user that regularly picks up a bike from a station according to a relatively fixed schedule. For example, consider a user that has picked up a bike from the same station at around 7:40 AM on many weekdays (say, to ride to work). We let T_A be the random variable representing the user's intended arrival time at the station (or intended pickup time). We assume the user (hereafter referred to as the *subject subscriber*) checks bike availability sometime before her intended pickup time by checking the station status via the service provider's mobile app or website. We use a random variable T_S to represent the time the subject subscriber checks the station status ($T_S \leq T_A$). We also let random variable T_W represent the elapsed time between checking bike availability and arrival at the station (if the subject subscriber decides to pick up a bike). For example, T_W may represent the time it takes the subject subscriber to walk to the bike station. Therefore, we have $T_S + T_W = T_A$. Note that when $T_W = 0$, then $T_S = T_A$, which basically indicates that the subject subscriber does not check bike availability in advance of arrival at the station.

We also assume that the subject subscriber has a balking threshold (represented by random variable B_T) so that she balks if there are fewer than B_T bikes available at the station at T_S when she checks the station status. In other words, based on her past experience, the subject subscriber expects that the station will be out of bikes by her arrival time (T_A) if there are fewer than B_T bikes available at the station T_W minutes before T_A. Our goal is to estimate the balking threshold (B_T) and time of balking or checking station status (T_S) – and consequently T_W – solely from the observed pickup times and bike availability data, even though the system-generated usage data do not provide any direct information on the events that take place that lead to different possible outcomes. These outcomes are discussed next.

2.2 Possible Scenarios

We can group the realizations of the interval of interest (say, 7:00 AM – 8:00 AM on weekdays) as follows:

- **Group 1 – Days with successful pickup:** Days that the subject subscriber picked up a bike from the station according to her regular schedule (say, around

7:40 AM). Two conditions were met on these days: (a) the station had more than B_T bikes available at time T_S when the subject subscriber checked the station status; and, (b) there was at least one bike available when the subject subscriber arrived at the station at T_A. These days can be directly identified from the system-generated data based on the subject subscriber's pickup time.

- **Group 2 – Days without pickup**: This group includes the days that the subject subscriber did not pick up a bike from the station according to her regular schedule. There are four possibilities/subgroups:
  - **Subgroup 2.1 – No pickup due to insufficient bike availability at T_S:** Days that the subject subscriber balked at T_S because bike availability was less than her balking threshold (B_T) when she checked the station status. These days cannot be identified from usage data since T_S and B_T are unobservable.
  - **Subgroup 2.2 – No pickup due to empty station at T_A:** It is possible that the station had more than B_T bikes available at time T_S when the subject subscriber checked the station status, but then ran out of bikes by the time the subject subscriber arrived at the station at T_A. This group includes such days, which are not identifiable from the system usage data since T_S and B_T are unobservable. Moreover, T_A for these days is also unknown due to censoring. In other words, there is no way to tell whether the subject subscriber actually visited the bike station and failed to pick up a bike due to outage.
  - **Subgroup 2.3 – No pickup due to bad weather:** It has been shown that weather conditions have a significant effect on bike sharing demand (El-Assi et al., 2017). We specifically include days with bad weather conditions as a separate subgroup since we can use the available weather data to distinguish these days (which can include rainy, snowy, and extremely cold or hot days).
  - **Subgroup 2.4 – No pickup due to other reasons:** There are many other possible reasons that may result in the subject subscriber not using the bike-sharing system even on days with sufficient bike availability and pleasant weather conditions. These reasons include but are not limited to illness, random delays and time constraints (say, due to oversleeping), customer mood (say, the subject subscriber may feel lazy to ride a bike to work), being on a work-related or personal travel, random changes in customer's schedule (say, the subject subscriber decided to work from home for the entire or part of the day), and other conflicting commitments (say, a doctor's appointment). Information on these days is also unavailable.

3 Methodology

We employ discrete-event simulation to generate synthetic data and assess the proposed data analysis method. There are two major reasons behind using simulation instead of real-world data:

I. Public data on bike-sharing systems do not include any identifier for users (say, real or censored subscriber ID). While this is mainly for privacy concerns and to prevent tracking individual users, this would prevent us from identifying appropriate subject subscribers to use in our study.
II. Individuals' true balking behavior is unobservable, prohibiting validation of the resulting estimates. Simulation allows us to verify and assess the accuracy of our data analysis method by comparing the resulting estimates with the input distributions for balking threshold (B_T) and time of checking station status (T_S) used in the simulation to generate the synthetic data in the first place.

3.1 The Discrete-Event Simulation Model

We assume that the nonstationary bike demand for a station can be approximated by a piecewise-constant rate function with a series of smaller intervals (say, hourly) where demand is assumed to be stationary, and independently and identically distributed (IID). A schematic example is provided in Fig. 1. This is a common approach for modeling and generating nonstationary stochastic processes in the simulation literature (Morgan et al., 2016) and bike sharing studies (Negahban, 2019; Patel et al., 2019). There are several methods and tools that can help identify an appropriate piecewise-constant rate function, namely the change-point analysis from (Chen & Gupta, 2011) and visual assessment tools (Vincent, 1998; Ansari et al., 2014; Negahban et al., 2016). In this paper, we consider a situation where these intervals are already determined and the goal is to estimate the balking threshold and timing of balking decision for a particular subject subscriber in one of these intervals (say, from 7:00 AM to 8:00 AM on weekdays).

We develop a discrete-event simulation model in Simio (Smith et al., 2018) to mimic the operation of the bike station during the time window of interest. Table 1 summarizes the parameters of the simulation model. Each replication represents a realization of bike availability trajectories during the interval of interest. In all

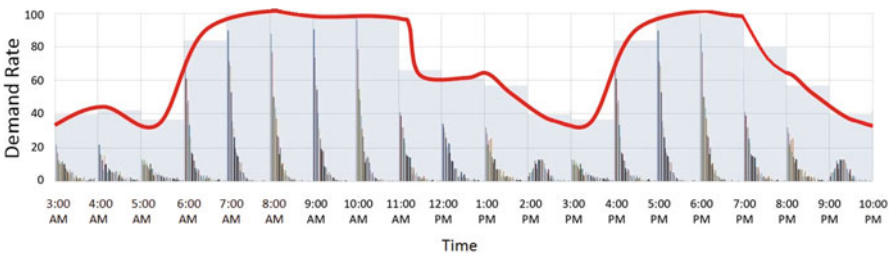

Fig. 1 A piecewise-constant rate function with hourly intervals for approximating the underlying nonstationary bike demand for a station represented by the continuous solid line. The HistoRIA tool developed in (Ansari et al., 2014) is used to generate the piecewise-constant rates as well as the histogram of inter-arrival times within each 1-h block

Table 1 Input parameters of the simulation model of the bike station

Parameter	Description	Value/Range
Input parameters related to the bike station		
N_D	Number of docks at the station	45
$N_B(0)$	Initial number of bikes at station at t = 0	Uniform (5, 30)
CIAT	Customer inter-arrival time	Exponential (0.7) minute
BIAT	Bike inter-arrival time	Exponential (1) minute
P_{BW}	Probability of bad weather conditions	0.15
Input parameters related to subject subscriber (assumed to be unknown)		
T_S	Time of checking station status	Triangular (28, 30, 32)
B_T	Balking threshold	Discrete uniform (10, 11)
T_W	Elapsed time between T_S and arrival time at station (T_A)	Triangular (9, 10, 11)
P_{OR}	Probability of other reasons for not using a bike	0.1

replications, the number of docks at the station is 45 and remains unchanged during the 1-h interval. However, the initial bike inventory at the beginning of the interval on any given day is a random variable and follows a uniform distribution as shown in Table 1. We consider a "busy" station, where the demand rate for bikes is higher than the demand rate for docks (i.e., bike drop-off rate), meaning that the station is likely to have low bike availability during the interval of interest if the initial number of bikes at the beginning of the time interval is small. There are three types of entities in the simulation:

- *Bicycle* entities, which represent riders that attempt to drop-off a bike at the station. Bicycle entities are generated according to the *BIAT* distribution.
- *Subject Subscriber*, which is a marked entity under study, hence there is only one instance of this entity type in any simulation run. This is explained in more detail later in this subsection.
- *Customer* entities represent individuals (other than Subject Subscriber) that attempt to pick up a bike from the station. Customer entities are generated according to the *CIAT* distribution.

The inter-arrival time of customers and bikes into the station both follow an exponential distribution with their respective mean value. Customers will balk (leave the model) if there is no bike available at the station. This represents the scenario that the customer decides to try a nearby station or use an alternative mode of transportation. Similarly, bikes arriving into the station will leave the model if all docks are occupied. This represents the scenario that the rider decides to try a nearby station to drop-off their bike. We assume the time it takes to check-out or drop-off a bike is negligible, i.e., pickups and drop-offs occur in zero simulation time, and that all bikes/docks are functional in the simulation model. For real-world applications, this can be adjusted to account for broken bikes and docks.

To focus on the estimation problem, we generate a *marked* customer sometime (say, around 7:30 AM) during the simulation of the interval of interest. This marked entity represents our subject subscriber and the time it is inserted in the simulation

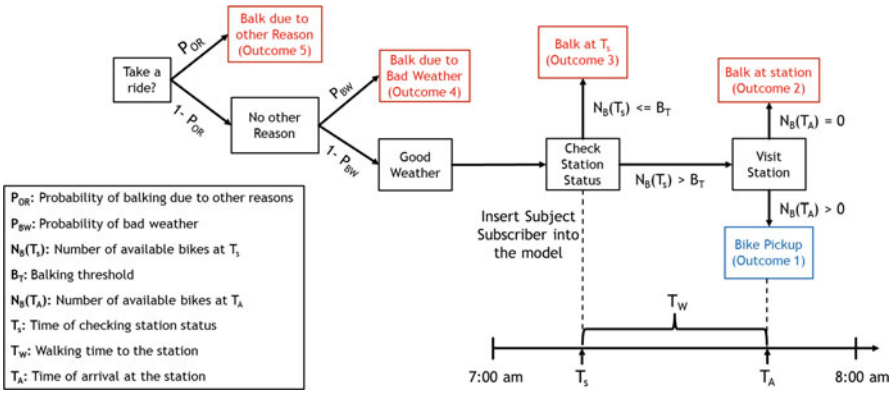

Fig. 2 Different outcomes for the subject subscriber in the simulation

run represents T_S at which she checks bike availability at the station. For example, this can represent a subscriber that picks up a bike at around 7:40 AM on most weekdays to go to work and checks the station status at around 7:30 AM. There are five possible outcomes for the Subject Subscriber entity, which directly correspond to the possible scenarios discussed in Sect. 2.2. The outcomes are summarized in Fig. 2 and can be described as follows:

- Decide to visit the station if $N_B(T_S) > B_T$, where $N_B(T_S)$ denotes the number of bikes available at the station at T_S. Once the marked customer arrives at the bike station at $T_A = T_S + T_W$, there are two possible outcomes:
  - **Outcome 1:** The marked customer will pick up a bike if there is any available.
  - **Outcome 2:** Balk if there is no bike available at the station.
- Decide not to visit the station:
  - **Outcome 3:** Balk after checking station status at T_S due to insufficient bike availability if $N_B(T_S) \leq B_T$.
  - **Outcome 4:** Balk due to bad weather conditions, which occurs with a probability of P_{BW}.
  - **Outcome 5:** Balk due to other reasons, which occurs with a probability of P_{OR}.

It is worth noting that the observations related to the marked customer across n simulation replications will be IID, so standard statistical methods are still applicable. Moreover, inserting arrivals generally carries an inherent risk of distorting the underlying arrival process (in this case, $CIAT \sim$ Exponential (1) minute). However, this effect is negligible in our case as we only insert a single marked customer during a 1-h simulation period.

Dayindex	Pickdrop	Availablebikes	Availabledocks	Objects	Time	Bad Weather
1	1	8	37	Customer	07:38:17	0
1	1	7	38	Customer	07:38:27	0
1	1	6	39	Subject_Subscriber	07:38:29	0
9	2	17	28	Bicycle	07:16:12	1
9	1	16	29	Customer	07:16:28	1
9	2	17	28	Bicycle	07:16:42	1
9	1	16	29	Customer	07:17:31	1

Fig. 3 Sample data generated by the simulation model. The dashed line indicates that there are hidden rows in between

3.2 Synthetic Data Generated by the Simulation Model

We consider B_T and T_S distributions used in the simulation to be unknown and unobservable (as in reality). We strive to estimate B_T and T_S solely from the pickup time and bike availability data generated by the simulation, which mimic real-world data readily available on bike-sharing systems. We simulate 300 realizations of the interval of interest. Figure 3 shows a snapshot of the simulated data. The "Day index" is the index for the realization (replication), hence varies from 1 to 300. The "Pick/drop" column indicates whether the row corresponds to a pickup or drop-off event, indicated by 1 and 2, respectively. The number of "Available bikes" and "Available docks" indicate the value just after the respective pickup/drop-off event. The "Object" column indicates the type of entity corresponding to the event. The "Time" column indicates the time stamp when the corresponding event occurred. Finally, the "Bad Weather" column represents the weather condition for that day (0 = good weather, 1 = bad weather). Before performing the estimation analysis, data pre-processing is performed by converting time stamps to real values. For example, by moving the time origin to 7:00 AM, a time stamp of 7:10:30 AM is converted to 10.5 min.

3.3 The Proposed Heuristic Data Analysis Method

Table 2 presents the notations used in this section. To motivate the proposed method, we consider a simplified version of the problem where the true (unknown) values for balking threshold, time of checking station status, and arrival time at the station, respectively denoted by B_T^*, T_s^*, and T_A^*, are deterministic and constant over all days included in the analysis. Without loss of generality, we set $P_{BW} = 0$. For any B_T and T_S, we define two conditional probabilities:

$$\pi^P_{B_T,T_s} = \Pr\{N_B(T_s) \geq B_T | E\} = \frac{\Pr\{N_B(T_s) \geq B_T \cap E\}}{\Pr\{E\}}$$

Table 2 Notations related to the proposed data analysis method

Notation	Description
E	The event that the subject subscriber picks up a bike
R	The event that the subject subscriber does not use the system due to other reasons
$N_B(t)$	Number of bikes available at the station at time t
s^P	Number of days that subject subscriber picked up a bike
s^{NP}	Number of days that subject subscriber did not pick up a bike
$n^P_{B_T,T_s}$	Number of days that subject subscriber picked up a bike and $N_B(T_S) \geq B_T$
$n^{NP}_{B_T,T_s}$	Number of days that subject subscriber did not pick up a bike and $N_B(T_S) \geq B_T$
$\pi^P_{B_T,T_s}$	Proportion of days that subject subscriber picked up a bike and $N_B(T_S) \geq B_T$
$\pi^{NP}_{B_T,T_s}$	Proportion of days that subject subscriber did not pick up a bike and $N_B(T_S) \geq B_T$

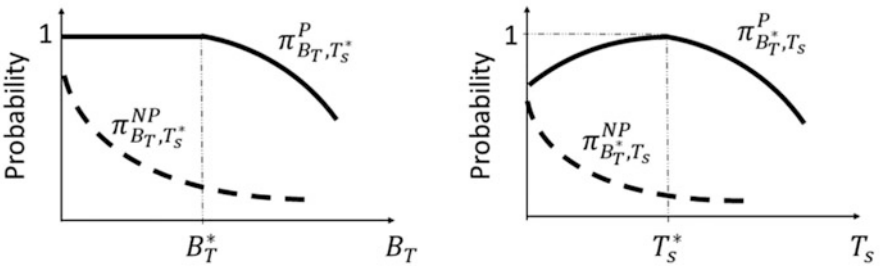

Fig. 4 Possible behavior of the two conditional probabilities around the correct estimates B_T^* and T_s^*. In the left figure, we have $T_s = T_s^*$, and in the figure on the right, $B_T = B_T^*$. The proof for concavity/convexity of these functions is deferred to future research

$$= \frac{\Pr\left\{N_B\left(T_s\right) \geq B_T \cap N_B\left(T_s^*\right) \geq B_T^* \cap N_B\left(T_A^*\right) > 0 \cap R'\right\}}{\Pr\{E\}},$$

$$\pi^{NP}_{B_T,T_s} = \Pr\left\{N_B\left(T_s\right) \geq B_T | E'\right\} = \frac{\Pr\{N_B(T_s) \geq B_T \cap E'\}}{\Pr\{E'\}}$$

$$= \frac{\Pr\{N_B(T_s) \geq B_T \cap [R \cup (N_B(T_s^*) < B_T^* \cap R') \cup (N_B(T_s^*) \geq B_T^* \cap N_B(T_A^*) = 0 \cap R')]\}}{\Pr\{E'\}}.$$

Clearly, for ($B_T \leq B_T^*$, $T_s = T_s^*$), the first conditional probability will take its maximum possible value of 1. A simple assessment of these two conditional probabilities at B_T^* and T_s^* (i.e., correct estimates) suggests the possibility that the magnitude of their difference is maximized at ($B_T = B_T^*, T_s = T_s^*$) as schematically shown in Fig. 4.

Motivated by the above, we propose the heuristic method shown in Fig. 5 to solve the general case where B_T^*, T_s^*, and T_A^* are random variables. We first average the observed pickup times to compute the subject subscriber's expected arrival time at station, T_A. The subject subscriber checks the station status sometime before or at

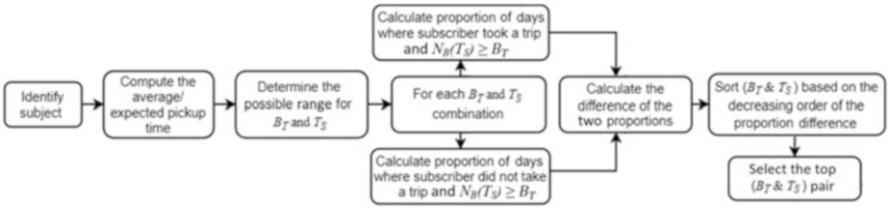

Fig. 5 General steps in the proposed data analysis method for estimating balking threshold and time of checking station status

T_A, hence the T_A value estimated in this fashion is the maximum possible average value for T_S (recall that $T_A = T_S + T_W$). We then start the first loop of the estimation algorithm by setting T_S equal to its upper bound T_A. In other words, we initially assume $T_S = T_A$. We then sequentially decrement the assumed T_S in each subsequent iteration of the search. In the analysis presented in this paper, we decrement T_S by 1 min at a time, although a smaller increment can be used for higher precision. It is unlikely for a customer to make balking decisions a long time before their intended pickup time (say, half an hour before) as bike availability can change drastically by the time they get to the station. The minimum possible value for T_S can be set to the adjusted time origin ($t = 0$) or an arbitrary small value between 0 and the estimated T_A. Here, we set the lower bound for T_S to zero to guarantee that the search covers the true unknown T_S.

Under each potential T_S value, we run a second loop by varying B_T within its possible range. Of course, the minimum possible value for B_T is 0. The maximum B_T value included in the search can be set to an arbitrary large value. In the analysis presented in this paper, we use a conservative maximum B_T value of 20. Although, it is highly unlikely for a customer to balk if the station has 20 bikes available when she checks bike availability (given good weather and no other reason not to ride a bike). By setting this upper bound to 20, we guarantee that our search covers the true unknown B_T. For each potential (T_S, B_T) pair included in the search, we compute two proportions:

$$\pi^{P}_{B_T,T_s} = \frac{n^{P}_{B_T,T_s}}{s^{P}} \text{ and } \pi^{NP}_{B_T,T_s} = \frac{n^{NP}_{B_T,T_s}}{s^{NP}}.$$

We then compute the difference of these two proportions under each potential (T_S, B_T) pair included in the search and sort the searched (T_S, B_T) combinations based on a decreasing order of this difference. The (T_S, B_T) pair for which the difference is maximum would be our estimate of the time of checking station status and balking threshold.

4 Implementation and Results

Figure 6 shows the calculations for some of the (T_S, B_T) pairs searched. Out of the 300 simulated realizations of the interval of interest (7:00 AM to 8:00 AM), there were 68 realizations that the subject subscriber picked up a bike ($s^P = 68$) and 232 that she did not pick up a bike ($s^{NP} = 232$). The average of the observed 68 pickup times was 39.39 min. Therefore, in the first loop of the algorithm, we vary possible T_S values from 39.39 to 0.39 in 1-min increments. In the second loop, we vary B_T from 1 to 20. Consider the first row in Fig. 6 corresponding to $T_S = 0.39$ and $B_T = 1$. There was at least one bike at the station at time 0.39 in 178 days out of the 232 days that the subject subscriber did not pick up a bike ($n^{NP}_{B_T=1,T_s=0.39} = 178$). Similarly, there was at least one bike at the station at time 0.39 in 55 days out of the 68 days that the subject subscriber picked up a bike ($n^{P}_{B_T=1,T_s=0.39} = 55$).

In the last step, we sort (T_S, B_T) combinations based on a decreasing order of their proportion difference. As shown in Fig. 7, we observe that ($T_S = 30$, $B_T = 11$) and ($T_S = 30$, $B_T = 10$) have the two largest proportion difference values, hence would be our top point estimates of T_S and B_T. Based on Table 1, we know these estimates are correct since they match the mean of the T_S and B_T distributions used in the simulation model to generate the data in the first place. It is important to note that we tested the efficacy of the proposed method in 15 other simulated cases with different parameter configurations and were able to estimate the correct balking threshold and time of checking station status in all cases. However, space limitations preclude the inclusion of all results in this paper.

B_T	T_s	$n^{NP}_{B_T,T_s}$	$n^{P}_{B_T,T_s}$	s^{NP}	s^{P}	$\pi^{NP}_{B_T,T_s}$	$\pi^{P}_{B_T,T_s}$	$\pi^{P}_{B_T,T_s} - \pi^{NP}_{B_T,T_s}$
1	0.39	178	55	232	68	0.76724	0.80882	0.04158
2	0.39	177	55	232	68	0.76293	0.80882	0.04589
3	0.39	176	55	232	68	0.75862	0.80882	0.05020
12	12.39	89	57	232	68	0.38362	0.83824	0.45461
13	12.39	82	57	232	68	0.35345	0.83824	0.48479
14	12.39	74	56	232	68	0.31897	0.82353	0.50456
13	24.39	25	50	232	68	0.10776	0.73529	0.62754
14	24.39	19	47	232	68	0.08190	0.69118	0.60928
15	24.39	11	41	232	68	0.04741	0.60294	0.55553

Fig. 6 Sample calculations for different (T_S, B_T) pairs included in the search process. The dashed lines indicate hidden rows

B_T	T_s	$n^{NP}_{B_T,T_s}$	$n^{P}_{B_T,T_s}$	s^{NP}	s^{P}	$\pi^{NP}_{B_T,T_s}$	$\pi^{P}_{B_T,T_s}$	$\lvert\pi^{P}_{B_T,T_s} - \pi^{NP}_{B_T,T_s}\rvert$
10	29.39	21	66	232	68	0.090517241	0.970588235	0.880070994
11	30.39	10	62	232	68	0.043103448	0.911764706	0.868661258
10	30.39	19	64	232	68	0.081896552	0.941176471	0.859279919
9	29.39	33	68	232	68	0.142241379	1	0.857758621
9	30.39	30	65	232	68	0.129310345	0.955882353	0.826572008
10	28.39	30	65	232	68	0.129310345	0.955882353	0.826572008
9	31.39	27	64	232	68	0.11637931	0.941176471	0.82479716
11	29.39	13	59	232	68	0.056034483	0.867647059	0.811612576
8	29.39	44	68	232	68	0.189655172	1	0.810344828

Fig. 7 Final step of the proposed method. Sorted (T_S, B_T) pairs based on a decreasing order of proportion difference

5 Conclusions and Future Research

We propose a simple yet effective heuristic data analysis method for estimating balking behavior of bike-sharing users that visit a station according to a somewhat fixed schedule on a regular basis (e.g., a subscriber that takes a bike to work most weekday mornings). We assume such users check the station status sometime before their intended pickup time to decide whether or not to visit the station based on bike availability at that time. Our heuristic method aims to estimate the time the user checks the station status and their balking threshold in terms of bike availability. We tested and confirmed the efficacy of the proposed method using several simulated scenarios.

An immediate extension of this work involves analytical proof for conditions that determine the concavity or convexity of the two conditional probabilities used in our heuristic method. Another important extension involves validation via real-world data. There are two main obstacles that make validation challenging for researchers: (a) due to privacy concerns and to prevent tracking individuals, service providers do not provide any identifier for subscribers in the data that they make public. Our algorithm needs this information to identify subjects and compute their expected pickup time. Service providers, however, have access to such data; and, (b) collecting information on individual's balking behavior requires requesting information directly from the subjects (say, via a survey or interview).

References

Albiński, S., Fontaine, P., & Minner, S. (2018). Performance analysis of a hybrid bike sharing system: A service-level-based approach under censored demand observations. *Transportation Research Part E: Logistics and Transportation Review, 116*, 59–69.

Ansari, M., Negahban, A., Megahed, F. M., & Smith, J. S. (2014). HistoRIA: A new tool for simulation input analysis In *Proceedings of the 2014 Winter Simulation Conference* (pp. 2702–2713).

Bargar, A., Gupta, A., Gupta, S., & Ma, D. (2014). Interactive visual analytics for multi-city bikeshare data analysis. In *Proceedings of the 3rd Urbcomp*.

Bordagaray, M., dell'Olio, L., Fonzone, A., & Ibeas, A. (2016). Capturing the conditions that introduce systematic variation in bike sharing travel behavior using data mining techniques. *Transportation Research Part C: Emerging Technologies, 71*, 231–248.

Caulfield, B., O'Mahony, M., Brazil, W., & Weldon, P. (2016). Examining usage patterns of a bike-sharing scheme in a medium sized city. *Transportation Research Part A, 100*(2017), 152–116.

Chen, J., & Gupta, A. K. (2011). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance* (2nd ed.).

Come, E., Randriamanamihaga, N. A., Oukhellou, L., & Aknin, P. (2014). Spatio-temporal analysis of dynamic origin-destination data using latent dirichlet allocation: Application to vélib' bike sharing system of Paris. In *TRB 93rd Annual meeting*, France, 19p.

El-Assi, W., Salah Mahmoud, M., & Nurul Habib, K. (2017). Effects of built environment and weather on bike sharing demand: A station level analysis of commercial bike sharing in Toronto. *Transportation, 44*, 589–613.

Morgan, L. E., Titman, A. C., Worthington, D. J., & Nelson, B. L. (2016). Input uncertainty quantification for simulation models with piecewise-constant non-stationary Poisson arrival processes. In *Proceedings of the 2016 Winter Simulation Conference* (pp. 370–381).

Negahban, A. (2019). Simulation-based estimation of the real demand in bike-sharing systems in the presence of censoring. *European Journal of Operational Research, 277*, 317–332.

Negahban, A., Ansari, M., & Smith, J. S. (2016). ADD-MORE: Automated dynamic display of measures of risk and error. In *Proceedings of the 2016 Winter Simulation Conference* (pp. 977–988).

O'Mahony, E., & Shmoys, D. B. (2015). Data analysis and optimization for (citi)bike sharing. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI'15)* (pp. 687–694).

Oppermann, M., Möller, T., & Sedlmair, M. (2018). Bike sharing atlas: Visual analysis of bike-sharing networks. *International Journal of Transportation, 6*(1), 1–14.

Patel, S. J., Qiu, R., & Negahban, A. (2019). Incentive-based rebalancing of bike-sharing systems. In H. Yang & R. Qiu (Eds.), *Advances in service science* (pp. 21–30). Springer.

Rudloff, C., & Lackner, B. (2014). Modeling demand for bikesharing systems: Neighboring stations as source of demand and reason for structural breaks. *Transportation Research Record: Journal of the Transportation Research Board, 2430*, 1–11.

Singhvi, D., Singhvi, S., Frazier, P. I., Henderson, S. G., O' Mahony, E., Shmoys, D. B., & Woodard, D. B. (2015). Predicting bike usage for New York City's bike sharing system. In *Computational sustainability: Papers from the 2015 AAAI Workshop* (pp. 110–114).

Smith, J. S., Sturrock, D. T., & Kelton, W. D. (2018). *Simio and simulation: Modeling, analysis, applications* (5th ed.). Simio LLC.

Tran, T. D., Ovtracht, N., & D'arcier, B. F. (2015). Modeling bike sharing system using built environment factors. In *Procedia CIRP, Elsevier, 2015, 7th Industrial Product-Service Systems Conference - PSS, industry transformation for sustainability and business, 30* (pp. 293–298).

Vincent, S. (1998). Input data analysis. In J. Banks (Ed.), *Handbook of simulation* (pp. 55–90).

Vogel, P., Greiser, T., & Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia – Social and Behavioral Sciences, Elsevier 2011, 20*, 514–523.