FINAL REPORT

Improving the Service Quality of Bike Sharing Systems via the Analysis of Real-Time User Data

Date of report: October 2018

Jose, Walteros, PhD, Assistant Professor, University at Buffalo

Prepared by:
Jose L. Walteros, PhD
Department of Industrial and Systems Engineering
413 Bell Hall
University at Buffalo
Buffalo, NY, 14260

Prepared for:
Transportation Informatics Tier I University Transportation Center
204 Ketter Hall
University at Buffalo
Buffalo, NY 14260
16. Abstract

The principal objective of this project is to develop a predictive statistical framework to efficiently estimate the ability of a bike-sharing system to serve incoming bike requests. By mining user data collected from the system’s smartphone app, an operator can utilize the proposed models to predict the likelihood that any potential user, who desires to use the system, decides to do so under the given the system’s conditions the user encounters (e.g., the location from where a bike request originates and its proximity to the nearest available bike, the weather conditions, the time of the day, and the customer profile, among others). The proposed statistical models is coupled with an operational bike redistribution model to analyze the cost-effectiveness of triggering quick bike redistribution tours to raise the service quality of the system to desired levels.

17. Key Words
Bikesharing, Shred mobility systems
Final report

Project period:
Sep 1, 2017 – Aug 31, 2018

Improving the Service Quality of Bike Sharing Systems via the Analysis of Real-Time User Data

Prepared by:
Dr. Jose L. Walteros
413 Bell Hall
Buffalo, NY 14260
716-645-2423 (Office)
716-645+3302 (Fax)
josewalt@buffalo.edu

Students who collaborated in this project:
Cai Gao (funded by the grant), Ningji Wei, Hadi Feyzollahi, and Luca Wrabetz
Acknowledgements

We would like to thank Shared Mobility Inc. for supporting this project.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation’s University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.
# Table of Contents

1. **MOTIVATION AND PROJECT OBJECTIVES** ................................................................. 1
2. **PRELIMINARIES** ........................................................................................................ 2
3. **PROPOSED METHODOLOGY** ................................................................................... 5  
   STATION TYPES 6  
   USER PROFILE AND TRIP HISTORY 7  
   FRAMEWORK OVERVIEW 8  
   Clustering component 11  
   Demand forecasting component 13  
   Rejection predictor component 13  
   Bike redistribution component 14  
4. **DOCUMENTATION OF DATA GATHERED** ............................................................... 16  
5. **DISCUSSION AND RECOMMENDATIONS** ............................................................... 18  
6. **CONCLUSIONS** ....................................................................................................... 19  
7. **LIST OF FIGURES** .................................................................................................... 21  
8. **REFERENCES** ........................................................................................................... 22
1. MOTIVATION AND OBJECTIVES

Bike sharing is an innovative urban transportation alternative that provides citizens fast access to bicycles for inner-city commuting. These type of systems are known for bringing significant benefits to its users in the form of a healthy and efficient transportation option, and to cities as an effective way of reducing CO2 emissions and traffic. For bike sharing systems to be considered viable, they must maintain a high-quality service that focuses on coverage and bike accessibility. While other transportation systems are readily available or follow predetermined schedules, bike sharing systems rely heavily on the operator’s ability to efficiently relocate bicycles from low to high demand areas as needed, to attract and serve potential users.

To plan the bike redistribution logistics, most system operators use static models that are strictly based on historical demand averages. These models typically fail to consider valuable user information that is dynamically collected by the system via online queries. Periodic and reactive bike redistribution trips are then exclusively performed when bike stations are deemed near empty or at full capacity, but are rarely planed and guided by predictive models with the ability to anticipate bike shortages given customer behavior. With new technological developments rapidly permeating most bike-sharing systems (e.g., GPS tracked trips, online-locking and paying systems, and smart data collection via smartphone apps), predictive analytic models have the potential to dramatically change the way in which bike sharing systems are operated.

The principal objective of this project is to develop a predictive statistical framework to efficiently estimate the ability of a bike-sharing system to serve incoming bike requests. By mining user data collected from the system’s smartphone app, an operator can utilize the proposed models to predict the likelihood that any potential user who desires to use the system decides to do so under the given the system’s conditions the user encounters (e.g., the location from where a bike request originates and its proximity to the nearest available bike, the weather conditions, the time of the day, and the customer profile, among others). The proposed statistical models are coupled with an operational bike redistribution model to analyze the cost-effectiveness of triggering quick bike redistribution tours to raise the service quality of the system to desired levels.

The proposed statistical models are, nonetheless, strongly dependent of the availability of data; particularly of online data collected daily from smartphone queries. Given the vast amount of user data that can be collected by the system’s smartphone app, it is possible to design the proposed framework to dynamically accept system queries at any point in time during the daily operation of the system. We have observed that if the operators are willing to capture and mine this data, this will enable them to generate an effective statistical map of the system and better inform the operational and strategic decision making process.

Throughout the course of this project however, we have observed most of the information that the operators collect is related to the status of the system, including the number of bikes and parking spots available, as well as descriptive information of the actual bike trips. However,
very little information is kept regarding the queries posed to the system’s smartphone app (even if this information is stored, is rarely made available to researchers). A second objective of this project, is to motivate the operators of bike-sharing systems to collect and mine this data, as it could provide beneficial information to improve the system’s efficiency.

2. PRELIMINARIES

Over the last few years, bike-sharing systems (BSS) have become a cost-effective, Environmentally-friendly alternative for public transportation in many cities around the world [9, 21, 27, 37]. From its early emergence about a decade ago to its consolidation in recent years, bike sharing (BS) has rapidly become a mobility paradigm proven to bring significant societal benefits for both its users and the cities adopting them [27]. BSS are known to provide citizens with a fast mode for making short commuting trips, a vibrant option for tourists to visit numerous city landmarks and attractions, and an efficient way for riders to traverse cities in a healthy and exciting way [22]. Similarly, from the perspective of the cities, BSS have become a natural platform for reducing traffic and pollution, for promoting healthy habits within its citizens, and overall, for improving the living conditions of the city [27].

A perfect complement to public transit and other mobility systems, BS has seen an explosive expansion around the world over the past five years [27]. Alone in the U.S., from only a handful of public bike share systems back in 2013, to over 60 systems in 2016, BSS have been gradually changing the daily commute of many users in almost every major American city. States like New York have seen a notable increase in this trend as well. By the end of 2017, six cities in New York are projected to have bike sharing accessible to the public, whereas three years ago only two systems existed in the State. In short, with their proven success, bike-sharing systems will continue to shape the urban transportation landscape in the years to come [24].

A typical bike-sharing system consists of a set of fixed stations that are scattered around a city. In each of these stations, the users can withdraw a bicycle for a given fare, use it for a given amount of time, and return it back to the system at any other station. In addition to the fixed stations, some BSS have recently started adding free-floating capabilities in zones with high demand for bikes. Free-floating areas introduce a convenient option in which the users do not require to walk to the designated fixed stations to pick up or drop off the bikes. Instead, based on the GPS trackers, the riders can use idle bikes parked at any nearby location and drop them anywhere within the free-floating zone [3]. Free floating, provides a good opportunity for BSS to increase the ridership by improving the systems coverage and reducing the extra burden for the users of having to walk longer distances to access the system. In turn, the analysis of demand patterns as well as planning the redistribution tasks become significantly more challenging for the operator.

For BSS to be considered effective transportation alternatives, they must provide their users with a high-quality service in the form of coverage and bike accessibility. While other transportation systems are readily available (e.g., personal vehicles) or follow predetermined schedules (e.g., bus or metro systems), BSS require to maintain
a steady number of available bikes distributed across the service areas, so that potential riders have the possibility to use the system [9]. Providing an acceptable service quality in the form of coverage and bike availability requires the operator to keep detailed track of customer travel behavior and demand patterns, which often poses a heavy operational load [19, 25, 28].

Planning the logistics of bike redistribution is a remarkably difficult task for it requires not only to identify optimal ways to effectively transfer batches of bikes between stations, but also because it depends on the operator’s ability to generate consistent demand patterns and travel behavior estimates [32]. Redistribution trips during operational periods cannot be planned days in advance because the demand for bikes and empty spaces can be quite volatile throughout the day [18, 32]. Since BSS are generally used as alternatives for other mobility systems—like personal vehicles, buses, or metros—users may often decide against using the system if the conditions they encounter are not ideal (i.e., no bikes available, poor weather conditions, among others). Periodic redistribution trips developed from static demand estimation models may fail to properly cope with strong demand fluctuations.

Most of the inference models that are developed for forecasting future ridership of BSS utilize random utility theory and other data mining tools that are commonly used for similar purposes on other types of mobility systems [5, 8, 16]. In the context of BS, the typical technologies that are used for estimating future passenger demand rely mainly on analyzing trip data collected by the system—i.e., the user tracks collected by the bike’s internal GPS as well as station bike levels [1, 6, 13, 15, 16, 18, 32]. Inference models that track user trips to generate detailed OD demand matrices have been quite useful for estimating bike requirements with a relatively good precision [2, 6]. In fact, some of these models are often able to predict surges in traffic given periodical events [1]. Nevertheless, traditional models only consider data of trips that actually happen, and generally neglect data of trips that did not occur because users who wanted used the system could not do so because the system failed to accommodate their needs (e.g., no bike was available near the user).

What is perhaps most interesting, is the fact that tools developed to analyze systems that are highly dependent on service quality, such as BSS, must take into consideration as much information as possible regarding the system’s failures in order to provide an accurate understanding of the system’s performance. For instance, for the BSS case, it is paramount to analyze user rejections; i.e., instances in which a user is interested in using the system, but decides against it due to the current status of the system they encounter. Failing to do so could render any estimation of the system performance to be overly optimistic, or provide misleading results given the “survivor bias” behavior that may occur when only analyzing successful trips.

It is important to notice that system rejections do not only happen when the user arrives at a station and encounters no bikes available to rent. Despite this being the most critical and obvious case, there are other worth-studying scenarios that provide more information about usage patterns and customer behavior. In this context, an important observation is that a large percentage of daily users of these systems are typically registered users. Despite the many casual users and tourists who also benefit from BS, as these systems
keep permeating our society, the number of recurrent (thus registered) users has significantly grown. An interesting advantage from this feature is that registered users have the possibility to observe the status of the system remotely via the smartphone app of the system, providing the operator with the possibility of recording behavioral information that can be used to inform the system operations.

Figure 1: GPS data of user rejections collected by the smartphone app of the BSS of Buffalo (left) and New York City (right)

Capturing consistent data about user rejections is in general a difficult for most mobility systems because it is difficult to track users that abandon the system for the lack of a good service. For the case of BSS however, it is possible to use data collected by the smartphone app of the system to effectively identify user rejections. Every time a user logs into the system’s smartphone app and queries the status of the system, data regarding its location and the proximity to the closest bike, as well as other indicators, such as user demographics, the time and weather can be seamlessly collected by the system (see Figure 1 for a pictorial representation of a potential user rejection given bike proximity and availability). The system can subsequently identify whether the user decides to rent a bike in the near future, collecting a time stamp indicating the time it took the user to reach the desired bike. Since most of the users of these systems must generate a profile in the system and often become members, it is possible then to generate a specific profile for every user, in which historical information regarding trips and rejections is collected. The proposed predictive models can then be trained with these user data to identify the likelihood that a user decides to use the system given a full set of environmental, systemic and demographic conditions.
In this project we develop a predictive statistical framework to efficiently estimate the ability of a bike-sharing system to serve incoming bike requests. The proposed framework will mine the user data collected from the system’s smartphone app, as described above, to predict the likelihood that any potential user, given a specific profile, who desires to use the system decides to do so. The proposed statistical models are coupled with a bike redistribution model for short and reactive redistribution trips, to analyze if triggering a such a trip is cost effective. In other words, the model test whether is preferable for the system’s operator to perform the rebalancing trips, or lose potential users. Given the vast amount of user data that is collected by the system’s smartphone app, it possible to design the proposed framework to dynamically accept system queries at any point in time during the daily operation of the system. This can enable the operator to generate an effective statistical map of the system and better inform the operational and strategic decision making process.

3. PROPOSED METHODOLOGY

The statistical framework we propose in this project is comprised of five analytic components that are in charge of estimating the probability that a potential user who desires to use the system decides to do so. The statistical models are leveraged on the availability of historical data that reflects information regarding successful trips, and what we call trip rejections. In essence, a successful trip, given a user request, happens when a potential user who plans to use the system, logs into the smartphone app (or website) and based on the status of the system and current weather (e.g., the availability of bikes in the stations in their vicinity, the distance to the closest bike, the temperature, and their individual profile) decides to use the system and rent a bike within a predefined time range from the moment they placed the query into the system’s app. If the user refrains from using the system within the expected time frame, we consider such a case a system’s rejection.

System rejections typically happen from different reasons. Clearly, the most common case is bike outages. If a user logs into the system’s app and finds no bikes close to them, the user is directly rejected by the system. However, even if there are potential bikes relatively close to the user, there might be other factors that can persuade them to avoid using the system, including distance and weather conditions. We believe that taking this additional information into consideration can be helpful to better track the system’s performance.

Before describing each of the specific components of the proposed methodology, we first discuss the type of data, and the trips and type of stations we consider in this study. We put a special emphasis on this component because the required input datasets our models expect to receive are often not collected by most system’s operators (or, if collected, the data sets are not clean and difficult to filter). Some of the information can be estimated based on available data, but the results are expected to be less accurate. We first discuss the specific type of stations that will be considered, given the demand fluctuation they receive. The, we discuss the user data required for the models.
3.1 Bike station types

During the daily operation of BSS in cities with active users, one typical phenomenon that often occurs is regarded as the *bike imbalance crisis* of BSS. An imbalance crisis happens when the outflow of bikes from a specific region of a city is significantly larger to the inflow of bikes (or vice versa), resulting in large zones of the system with empty (or full) bike stations for prolonged periods of time. This type of phenomenon is not uncommon in systems like the one of New York City (see, Figure 2 left, where the red dots represent empty stations). What is perhaps more interesting is that even with recurrent redistribution trips between zones with high demands of bikes and parking spots, it is extremely challenging to maintain a balance bike fleet across such zones. This phenomenon is in fact exacerbated in the presence of good weather conditions (summer and early fall months).

System operators have devised different tools to combat major imbalances of this kind by (1) performing continuous redistributions, (2) adding portable bike stations to increase the bike/parking spot supplies, and (3) even paying monetary prices to idle users who want to help with the bike redistribution by taking ad-hoc trips to move bikes from crowded to empty stations (this type of recourse is often called “bike angels”).

Despite being a critical issue for BSS, this project does not consider this type of stations. The main reason is twofold. First, the system is aimed to identify the probability that a rejection occurs under the given conditions of the system. Since such a probability is typically close to one during large time periods throughout most weekdays for those stations, it becomes difficult to properly train machine learning tools to predict rejections in other less problematic stations, without overestimating negative cases because of data imbalances. Second, one of the objectives of the project is to provide a tool that identifies cost effective short redistribution trips to alleviate manageable imbalances on non-critical stations. Since critical stations often require recurrent redistribution trips of many bikes, it is counterproductive to include those in such trips.

![Figure 2: Examples of regular and critical stations from the systems of Boston and New York City](image-url)
A station is then considered critical, and as such omitted from the study, if a set comprised by the station and some other neighboring stations remain empty/full for periods of time longer than a predefined threshold $max_{\text{outage\_time}}$. Examples of this predefined parameter are more than one hour, or the expected time it takes to perform a redistribution trip to such a station, given the traffic conditions. A specific example of this type of station is the one depicted in Figure 3. Here, the station remained empty/full for many consecutive periods. If a station is deemed critical all the trip data to and from such a station is removed from the database.

![Figure 3: Examples of a critical station Stony Brook, from the Boston BSS.](image)

### 3.2 User profiles and trip history

The proposed methodology requires analyzing the user’s profiles and generate the predictions based on their successful trips and rejection rates. It is known that multiple users have different tolerances to walking longer distances according to their individual characteristics, including age, gender, as well as the type of trips they typically perform (i.e., casual, regular commute, exercise routine, among others). For this reason, the data training is performed by only studying the trips made by registered users. We believe this assumption is not a strong limitation because in larger cities, given the bike renting prices and the renting “bundles” that
you typically get as a recurrent user, a considerable portion of the trips is performed by registered users.

A key component for this project is aimed to understand the behavioral component of the users. More specifically, what are the conditions of the system, other than a bike or parking spot outage, that dissuade the users to use the system. It is clear that each individual responds differently to the conditions they encounter. However, performing a statistical analysis of the individual rejection probability for all the users is not only impractical but statistically inviable due to the lack of significant data for some users. Based on information obtained from the system of the city of Boston the number of trips and the trips characteristics across all users varies significantly. Equally so is the individual usage of the system smart phone app. To perform a proper statistical analysis, a portion of the framework is aimed to cluster the users into groups based on (1) demographic information, including age and gender, (2) typical information of their trips, like typical initial and final destinations, rejection percentage, average ride distance, average distance to the available bikes for both successful trips and rejections, (3) environmental conditions like, weather, day of the week, and time of service, as well as (4) other the system conditions, including the number of bikes on rack.

Further information on this clustering component will be presented below. There are a few challenges. First, perhaps the more crucial is to be able to identify whether a user who is logging into the system app is interested in using the system in the short term, or is simply interested in the system status or information for future events (i.e., irrelevant queries). Since this distinction is vital for properly forecasting system rejections, it is important to process the data first to remove system queries that are not intended for trips. Due to the lack of comprehensive data, this is still a challenge that must be studied in subsequent projects. In this study, the challenges that arise from potential irrelevant queries where circumvented by the fact that the rejection data was estimated from the available datasets.

### 3.3 Framework overview

As mentioned before the framework is composed of five components. The main objective of the framework is to produce a data classification model with the ability to predict is a user decides to use or not the system given the profile data collected by the system’s app. The idea is that the BSS operator can use the framework to forecast incoming demand for bikes and check whether the number of rejected customers based on the prediction is too high, given the current conditions of the system, so that the cost of short redistribution trips is justified. The list of the major components follows. Additionally, a graphical representation of the interaction between the components is depicted in Figure 4.

**Clustering component**: This component clusters users into a few groups based on their demographics and usage behavior. The idea here is that the availability of individual data is highly unbalanced across users, given the differences in the usage; in particular, the number of successful trips and rejections per user. The clusters attempt to capture similarities in terms of demographics and usage patterns (particularly reasons for trip rejection). The clusters are latter
use in the forecasting component to reproduce surges in demand according to the “typical” (average) usage pattern of users in the respective clusters. This helps, first, to reduce the variability that is prevalent in the trip data, and, second, to have a more accurate prediction on how the people respond to the system conditions. Furthermore, the clusters may also help to mitigate the imbalances of the data, given the user information.

**Demand forecasting component:** We developed a trip-generation model that estimates the OD demand matrix using the trip information collected by the system for each of the clusters generated before. In other words, the initial estimation model uses exclusively trip data of the users within the clusters. The estimated customers who demand bikes of each cluster for the operational period is then assumed to behave homogeneously and thus react equally according to the “typical” characteristics of the cluster. Since the purpose of this model is to exclusively generate a stamp of the OD matrix, we use a variation of the statistical model proposed by [1] to generate the demand estimation.

**Rejection predictor component:** This component is essentially a classifier that identifies the probability that a person from a given cluster decides to use the system (i.e., attempt to rent a bike) given the system and weather conditions they encounter. This is the main engine of the proposed framework. We tested three algorithmic options, logistic regression, decision trees (XGBoost), and neural network for this purpose. As inputs for training, the component receives the information of the clustered users (demographics), usage patterns (averages on distance length, typical origin and destination stations, common hours), successful trip data (information of each successful trip), smartphone app usage data (queries placed to the smartphone app), system information data (availability of bikes and parking spots). And returns a success probability that represents that chances that the user ends up using the system.

**Bike redistribution component:** We implemented a classic bike redistribution Routing Problem as a mixed-integer optimization problem. The proposed model has the typical pick-up and delivery and time-windows characteristics that many static models have [26]. One of the fundamental differences of the proposed model with previous approaches is that in our case, the model is tailored to identify quick and short redistribution tours triggered by the proposed inference model. This component acts as follows. Based on the status of the system, the forecasted demand data can be fed into the rejection predictor model to identify the potential stations that may benefit from a redistribution trip in the short term, to satisfy the predicted demands that can be originated in the near future. The number of such stations should be small as no critical station is considered; therefore, the obtained redistribution routes are expected to be short and quick.

**Data preparation component:** This component, albeit not listed in the diagram of Figure 4, prepares the raw data and extracts the relevant information that is later fed into the statistical framework. It was perhaps the most challenging component to design and implement. One of the biggest challenges we encountered was the fact that the specific data that we required was not directly available and has to be extracted from complex datasets, or estimated based on historical and incomplete sources.
Two particularly challenging tasks involved (1) pairing the available user data with the trip information and (2) the estimation of the system rejections. The data sets that contained the information we used had limited information regarding the registered user demographics. Moreover, as discussed in Section 4, the trip data as well as the few user information that was available did not necessarily matched collection the dates A thorough preprocessing algorithm was implemented to extract as much information as possible from the given data. Furthermore, the trip information had to be matched with the weather data collected from the National weather Service Forecast Office (https://w2.weather.gov/). Gathering and preprocessing all the data consumed a major portion of the project execution time.

A data association algorithm was implemented to estimate the true association between the trips and the user profiles. The information of which of the registered user performed what trip was not available due to privacy concerns. However, some specific demographic information regarding the user was left. We use this data to generate viable associations between trips and users. Data association problems are commonly used to associate signals in sensor-tracking applications as well as in other pattern recognition tasks. The goal of data association is to identify a correspondence between a collection of new sensor measurements and preexisting tracks. In this context, the trips are the analogues of the sensor measurements. The problem in general seeks to find a correspondence between tracks of the same target that were generated by different sensors (in this case trips in the system). In this context, the users are heterogeneous, which complicates the overall association problem. Data association problems are commonly NP-hard, and given the size of the databases at hand, we implemented a few simple heuristic algorithms that were ran multiple times. The best association was finally selected.

An interesting follow-up project can be to improve these association component, or identify a partnership that provides real associated data, even if it is anonymized first.

An overview of the proposed framework is depicted in Figure 4. The intended use of the system is twofold. First, one can feed the rejection predictor component a set of forecasted demands that may materialize in the subsequent hours. The system will then can be used to predict what is the expected value of rejected trips that can occur in the system given the current status and conditions. This could be used to analyze whether triggering a distribution trip is worth it. A second possibility, which was not initially tested, but could be developed as part of future works involves the online analysis of potential requests. The BSS operator can use this code to check, every time a user logs into the system smartphone app, whether the given query may result in a successful trip or a rejection.

Such kind of application can be coupled with other tools, such as the option of reserving bikes for future trips as well.
We now will provide some specific discussion regarding the techniques used in the proposed components.

3.3.1 Clustering component

The expected type of data includes the user profile, the historic trip information of all the registered users, usage patterns (averages on distance length, typical origin and destination stations, common hours), successful trip data (information of each successful trip), smartphone app usage data (queries placed to the smartphone app), system information data (availability of bikes and parking spots) and weather data.

Figure 5 provides an example of the trip database and the login query database and Figure 6 depicts the typical system information that can be found regarding the availability of bikes and
parking spots in the system.

<table>
<thead>
<tr>
<th>tripduration</th>
<th>starttime</th>
<th>stoptime</th>
<th>start station id</th>
<th>end station id</th>
<th>bikeid</th>
<th>birth year</th>
<th>gender</th>
<th>User id</th>
</tr>
</thead>
<tbody>
<tr>
<td>579</td>
<td>3/1/2015 14:27</td>
<td>3/1/2015 14:36</td>
<td>141</td>
<td>75</td>
<td>1300</td>
<td>1979</td>
<td>1</td>
<td>418150434</td>
</tr>
</tbody>
</table>

a) Trip database

<table>
<thead>
<tr>
<th>logtime</th>
<th>User id</th>
<th>station id</th>
<th>login latitude</th>
<th>login longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/1/2015 13:01</td>
<td>341127441</td>
<td>108</td>
<td>42.36356016</td>
<td>-71.08216792</td>
</tr>
<tr>
<td>3/1/2015 2:275</td>
<td>864320595</td>
<td>141</td>
<td>42.356954</td>
<td>-71.113687</td>
</tr>
<tr>
<td>3/1/2015 17:48</td>
<td>526540903</td>
<td>25</td>
<td>42.370803</td>
<td>-71.104412</td>
</tr>
</tbody>
</table>

b) Query database.

Figure 5: examples of the data bases that are used for the clustering component.

Figure 6: bike and parking spot availability information in the system status database

The trip database for each user, the smartphone query database, the system information database, and the weather database, must be combined to produce a table that describes the information of the successful and rejected trips. Independently of whether a trip was successful, the information regarding the local weather, the time of the trip, and the status of the system must be obtained for as many trips as possible per user. One specific column in such dataset must indicate with a 1 if the trip was successful after placing a query in the smartphone app, and with a 0 if the trip was rejected.

This information is then aggregated and averaged per user so that each of them gets associated
a specific “behavioral vector”. This vector contains cross-referenced information based on the different sets of raw data. For example, the behavioral vector of each individual user contains information regarding the average number of trips (successful or rejected) when the weather was sunny, when the temperature was low (below 65 degrees), when there were less than (3 bikes left), etc. The cross-referenced data includes averages of the number of rejections and successful trips given a different combination of two and three inputs.

The behavioral vectors are then used to cluster the users based on their similarities. Given its simplicity we implemented a variation of the *k-medoids* clustering algorithms. K-medoids is essentially a variation of k-means clustering when the metrics used to calculate user similarities are not the Euclidean distances associated with the data points. Since not all the information of the behavioral vectors can be averaged (e.g., the gender, common set of stations), we resort to the most flexible k-medoids algorithm.

<table>
<thead>
<tr>
<th>User id</th>
<th>common start stations</th>
<th>common end stations</th>
<th>gender</th>
<th>common start time</th>
<th># trips when sunny</th>
<th># of rejections when sunny</th>
<th># bikes left when rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>129354722</td>
<td>108</td>
<td>96</td>
<td>1</td>
<td>1:01:50 PM</td>
<td>25</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>141</td>
<td>108</td>
<td></td>
<td></td>
<td>2:27:08 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 7: Examples of the combined dataset generated for a user using the raw data

3.3.2 Demand forecasting component

We address the forecasting modeling of the OD demand matrix by using effective off-the-shelf models in literature. We follow similar procedures to the ones presented in [1]. We first tested SVM classifiers with Gaussian kernels, using a one-versus-one approach to multi-classification, as was done in [1]. Several forecasting techniques have been commonly used to generate OD pairs. Traditional SVM tools have been proven to be effective for this purposes and as such, we did not deviate much from the methods that are considered standard.

3.3.3 Rejection predictor component

This component is essentially a classifier that identifies the probability that a person from a given cluster decides to use the system (i.e., attempt to rent a bike) given the system and weather conditions they encounter. This is the main engine of the proposed framework. We tested three algorithmic options, logistic regression, decision trees (XGBoost), and neural network for this purpose.

As will be mentioned later, training this component with the available data was proven rather challenging, as the system requires very accurate information on rejection rates. Due to the lack of fully descriptive data, some of the data sets used were estimated based on the information available, producing mixed results.

One of the difficult challenges of the proposed approach is the imbalance of the data and the presence of false positives. In general, if a BSS is well-maintained, the number of user rejections is significantly smaller than the number of successful trips. The main problem behind data
imbalance is that the models would tend to significantly favor the classification towards successful trips rather than properly predicting user rejections. Furthermore, since a large portion of the user rejection data was estimated, it is possible that the estimations do not emulate real rejections perfectly.

In addition to the issues posed by data imbalance, the presence of false positives in the data could negatively affect the prediction capacity of the model. In the BSS context, a false positive occurs when a user makes a service query to the smartphone app with the simple interest of checking the status of the system, but with no real intention of renting a bike. Training the proposed models with data that contains many queries of this type could impact the rejection rates. We estimated this data based on historical bike availability, but acknowledge that without feeding the framework without proper smartphone data rejection data, as we originally intended, is possible that the rejection probabilities produced by the system yield inaccuracies.

Data imbalance and false positives have been consistent difficulties that arise quite commonly in the machine learning literature, particularly in the context of classification. To combat class imbalance most models relied upon three broad approaches: (1) embedded approaches [7, 10, 20, 38], (2) data preprocessing [4, 11, 14], and (3) ensemble learning [17, 29, 30]. We employ some data preprocessing approaches, including over- and under-sampling approaches to reassemble the training and testing data, but a subsequent refining of the data is highly suggested.

### 3.3.4 Bike redistribution component

The resulting bike redistribution problem is defined as follows. Given its flexibility, we directly implemented a variation of monolithic formulation proposed in [29], which is shown in Figure 8. The input of the problem is a set of stations and the location of the bike depot. For each station
that will be rebalanced, the model requires the initial inventory and the capacity. The problem also receives a travel-time matrix between stations and the capacity information of a set of capacitated repositioning vehicles (assumed to be homogeneous). A solution is defined by a route for each vehicle and the quantity of bicycles to load or unload at each station along the route. The planned routes must satisfy a time constraint given by the urgency of the redistribution. The length of each route is calculated as the sum of the travel times between all pairs of consecutive stations along the route plus the time needed for loading and unloading at the stations, which depends on the quantities of bicycles handled. The goal is to minimize the total operating costs of the vehicles.

\[
\min \sum_{i \in S_F} (y_i^+ + y_i^-)
\]
\[
s.t.
\]
\[
\sum_{k \in K} \sum_{e \in (a',a)} x^e_k - \sum_{k \in K} \sum_{e \in (a,a')} x^e_k - z_e + y^+_e - y^-_e = f_e - C^0_e(s) \quad s \in S_F, \forall (s) = 1
\]
\[
\sum_{k \in K} \sum_{e \in (a',a)} x^e_k - \sum_{k \in K} \sum_{e \in (a,a')} x^e_k + \bar{y}_{pred} - z_e + y^+_e - y^-_e = f_e \quad s \in S_F, \forall (s) \geq 2
\]
\[
\sum_{k \in K} \sum_{e \in (a',a)} w^e_k \leq 1
\]
\[
y^+_e, y^-_e \geq 0
\]
\[
0 \leq z_e \leq C^e_{(s)}
\]
\[
x^e_k \leq Q^e_k w^e_k \quad s \in S_F
\]
\[
\sum_{e \in (a,a')} w^e_k - \sum_{e \in (a',a)} w^e_k = 0
\]
\[
\sum_{e \in (a,a')} w^e_k = 0
\]
\[
\sum_{e \in (a,a')} x^e_k = Q^0_k
\]
\[
x \geq 0
\]
\[
w \text{ binary.}
\]

Figure 9: Bike redistribution model proposed in [29] and adapted for our application (here depicted verbatim from its source)

Figure 10: Redistribution trip example
The type of solution that is expected is depicted in Figure 9. Contrary to the case of [29] in which a battery of decomposition algorithms is used to speed up the solution time of the formulation, since the expected stations that are to be rebalanced is small in this context, we resort to the compact monolithic formulation of Figure 8. The specifics components can be found in [29].

4. DOCUMENTATION OF DATA GATHERED

One of the principal challenges encountered in this project was the availability of data. To properly train the proposed predictive analytical models, it is fundamental to have actual data that capture system rejections. As mentioned in the previous section, system rejections are not necessarily directly linked to outages. There exist for example a distance threshold the users are willing to walk to find an available bike. Such distance is highly dependent on age, gender, current weather, among other components. Some users are willing to walk more, or wait more than others, so the definition of a bike outage is generally fuzzy and user dependent. Moreover, a user that check the system’s app to find whether there are bike available close to them, may decide not to walk to the system station and use the system if the number of bikes is small enough so that they take the risk of those being rented before they arrive to the station.

In the context of analyzing BS systems in which the usage patterns are highly dependent of other external factors, it could be beneficial to mine this data to identify typical user behaviors and better predict moments in which the system may require a bike rebalancing. Examples of the data required to properly estimate rejections include the user profile information, including demographic data, as well as information about individual system queries, which include the set of stations the user is interested in (per query), the location of the user where the query originated, time and day of the query, as well as whether the user subsequently used the system and the time it took the user to rent the bike immediately after the smartphone app query.

This type of data can be used to cluster the users based on their usage, demographics, and typical location patterns to produce general estimations without the need of testing each individual separately. Moreover, the proximity of the potential user to the closest station, as well as the different time stamps are required to identify typical properties that lead to a system’s rejection, including for instance, the weather and the number of bikes available for the potential user.

Based on discussions that we had with several system’s operators, the app system is capable of collect this type or similar information, but they do not typically store all the specific elements needed, or if they do, the system operator does not share such information due to privacy and logistic issues. Despite the lack of specific data, several systems based in the US that serve different American cities share plenty of data that can be used to infer and estimate the required datasets.
One unanticipated challenge prior to the execution of this project involves the difficulty of finding readily-available data for testing purposes. Nevertheless, despite the additional time required to planning and designing models to produce datasets to train the proposed models. Several further modeling ideas emerged from this challenge that helped us to enrich the discussion of the predictive models.

The initial intention was to test the proposed predictive statistical framework using data collected by the Reddy Bikeshare (RB) System, which currently serves the city of Buffalo, New York. RB operates a system with a fleet of 200 “smart” bicycles powered with GPS capabilities located at 35 stations throughout the city. During its operational months from mid Spring to late Fall, RB registered a total of 11,986 trips covering a staggering 17,614 miles. The average trip distance was 2.1 miles and the longest trip was 66 miles; the average trip duration was 30 minutes. According to our collaborators at RB, preliminary survey findings indicate that most riders replace the use of their personal vehicle to go to work or do personal errands. Similarly, many users indicated the weather and that the fact of not having a bike available close to them were the two principal reasons that prevent them to use the system.

Unfortunately, because of privacy issues, a big portion of the data was not available to generate the tests. Furthermore, given the small size of the BSS of Buffalo, the type of information that was available was not particularly suited to train the models. We decided to proceed with the model testing using open source datasets.

As will be described in subsequent sections, a big portion of the data used was generated by analyzing, mining, and linking information collected from different datasets available online for public use. We would like to highlight the data provided by the bike-sharing of Boston, which is called Blue Bikes (see https://www.bluebikes.com/, and http://www.hubwaytracker.com/). There are other systems like Citi Bikes, which is the system from New York city, which also share information. Nevertheless, we found the data provided by Blue Bikes more complete and better suited for our purposes.

In 2017 Blue Bikes (back then Hubway bikes) open a data challenge inviting researchers across different disciplines to identify potential tools to improve the usability of the system. As part of the data challenge, Blue Bikes provided historical data predominantly about bike usage (i.e., trip data) and station status data (i.e., outage information). Figure 2 provides two examples of the type of data that was available.
Figure 3: Blue Bikes usage data from the 2017 data challenge.

The data challenge concluded before we begin developing this project; nevertheless, the data was still available after completion. As mentioned before, the data we collected from Blue Bikes, was not explicitly what is required by the proposed models; however, it was possible to estimate some small data sets to preliminary test our methodology.

During the development of this project, we notice that other approaches that mine smartphone data collected by mobility systems apps are currently put in practice by companies ridesharing like Uber and Lift, as well as by other researcher interesting in analyzing turban transportation usage data. Perhaps one of the most useful data sets for this purpose is the taxi and limousine trip data of New York City collected and maintained by the NYC Taxi & Limousine Commission (see, http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml). The databases therein are often studied to analyze different transportation systems in the city of New York.

5. DISCUSSION AND RECOMMENDATIONS

Predictive inference models involving “Big Data” analytics have the possibility to learn from user behavior and help the operator of any shared-mobility system to make better strategic decisions. By effectively predicting the ability of the system to handle future demands, the system’s operators can better react to potential surges in bike requests, thus increasing the system’s efficiency. In particular, models that can dynamically operate using real-time data on the spot have a higher potential to produce significant improvements on service quality. The principal advantage of utilizing data from online requests and bike availability queries—in addition to the typical user trip data—is that by mining requests that may or not resulted in
actual user trips, the operator gains the ability to understand the reasons why potential users decide not to use the system. By combining these tools with other trip prediction models, the system’s operators would be able to: (1) better plan the redistribution process, focusing on areas of the system with larger probabilities of producing user rejections; (2) fine tune the OD demand estimations to better inform tactical decisions regarding the location of bike stations and the boundaries of free-floating zones; (3) monitor and assess the service quality of the system given various operational strategies, particularly during high-demand events.

With new technological developments—e.g., GPS tracked bikes, online-locking and paying systems, and smart data collection via smartphone apps—rapidly becoming standard of most bike-sharing systems, predictive analytic models like the one proposed in this project have the potential to dramatically improve the way in which bike sharing systems are operated. It is important to note that methods are general in the sense that they are not limited to shared mobility systems. Similar type of models can be used of other systems, like electric-car rentals, and online transportation networks such as Uber and Lift. For the field of Transportation Operations Research, this project can help to increase the adoption of Big Data tools and data-driven decision making framework.

The process of estimating and gathering the input data is still a work in progress. One of the principal challenges that was not anticipated before the conception of the project was the lack of available data, which required the implementation of additional tools to combine, prune, and filter incomplete datasets. The proposed models thus have been trained and tested in preliminary runs with estimated data, and in some occasions simulated data.

The preliminary tests for the training models have yielded mixed results. In particular, some of the individual features of the user demographics are often deemed as statistically insignificant for some data sets but not for others, when testing the clustering component. There seem to be two particular explanations. First, the association models used to link the user data with the trips data may be producing associations that are too dissimilar. Second, the models identify rejected trips could be producing too many false positives, affecting the veracity of the data.

6. CONCLUSIONS

The principal objective of this project was to develop a predictive statistical framework to efficiently estimate the ability of a bike-sharing system to serve incoming bike requests. The proposed framework mines the user data collected from the system’s smartphone app, to predict the likelihood that any potential user, given a specific profile, who desires to use the system decides to do so. The proposed statistical models are paired with a bike redistribution model for short and reactive redistribution trips, to analyze if triggering a such a trip is cost effective.

The statistical framework is comprised of five comprehensive components. First, a clustering component groups users based on their demographics and usage behavior. The clusters are
latter use in the forecasting component to reproduce surges in demand according to the “typical” (average) usage pattern of users in the respective clusters. Second, a demand forecasting component that estimates the OD demand matrix using the trip information collected by the system for each of the clusters generated before. Third, a rejection predictor component that acts as a classifier in order to identify the probability that a person from a given cluster decides to use the system (i.e., attempt to rent a bike) given the system and weather conditions they encounter. Fourth, a bike redistribution component modeled as a mixed-integer optimization problem that identifies fast redistribution routes to test whether performing a redistribution trip is cost effective, based on the current unsatisfied demand. Ad, Fifth, a data preparation component that prepares the raw data and extracts the relevant information that is later fed into the statistical framework.

One unanticipated challenge prior to the execution of this project was the difficulty of finding readily-available data for testing purposes. The initial intention was to test the proposed predictive statistical framework using data collected by the Reddy Bikeshare (RB) System, which currently serves the city of Buffalo, New York. Unfortunately, because of privacy issues the data was not available to generate the tests and given the small size of the BSS of Buffalo, the type of information that was available was not particularly suited to train the models. We decided to proceed with the model testing using open source datasets.

A portion of the technical implementation of the proposed framework, particularly the training component is still an ongoing endeavor that requires further considerations. We are currently refining the data sets to better train the rejection predictor, which is regarded as the main core of the proposed methodology.
7. LIST OF FIGURES

- Figure 1: GPS data of user rejections collected by the smartphone app
- Figure 2: examples of regular and critical stations from the systems of Boston and New York city
- Figure 3: examples of a critical station stony brook, from the Boston BSS.
- Figure 4: Proposed system framework.
- Figure 5: examples of the data bases that are used for the clustering component.
- Figure 6: bike and parking spot availability information in the system status database
- Figure 7: Examples of the combined dataset generated for a user using the raw data
- Figure 9: Rejection predictor component
- Figure 10: Bike redistribution model proposed in [29]
- Figure 11. Redistribution trip example
8. REFERENCES


