ABSTRACT

Title of Dissertation:	OPERATIONAL CHALLENGES IN DOCKLESS BIKE-SHARES: THE CASE OF HYPERLOCAL-IMBALANCE
	Shivam Mishra Master of Science, 2021
Dissertation Directed by:	Dr. Ashish Kabra Department of Decision, Operations & Information Technologies and Prof. Derek Paley Department of Aerospace Engineering

Recent times have seen a shift from traditional docked to dockless bike-sharing systems. It is popular among consumers as it allows flexibility to drop-off bikes anywhere and solves the last-mile problem of transportation. While convenient for users, the dockless bike-share system's free-floating model introduces the problem of hyperlocal imbalance, about which little or no research is available. The hyperlocal imbalance is the supply-demand disparity created in a small geographical region due to consumer's bias to pick up bikes from some locations compared to others. This paper introduces, demonstrates, and determines the reasons behind the hyperlocal-imbalance in dockless-bike-sharing systems.

The study of hyperlocal-imbalance requires access to fine-grained trip level data, which is not easily accessible to the research community due to privacy or competition issues. To deal with it, in this work, we introduce an algorithm to extract trip-level information from the General Bikeshare Feed Specification (GBFS) feeds, which bike-share companies are obliged to upload as per transportation department regulations across the US. The algorithm is validated against the actual trip data of dockless bikes. It extracts the trip details from the GBFS data with a recall of 77% and precision of 80%.

OPERATIONAL CHALLENGES IN DOCKLESS BIKE-SHARES: THE CASE OF HYPERLOCAL-IMBALANCE

by

Shivam Mishra

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2021

Advisory Committee:
Professor Derek Paley, Chair/Co-Advisor
Dr. Ashish Kabra, Advisor
Dr. Ilya Ryzhov

Acknowledgments

I want to thank my advisor, Dr. Ashish Kabra, for providing me with an opportunity to work on this project. He kickstarted this project last year and allowed me to build on it. Apart from making himself available for help and advice, he also helped me with all the resources I needed for this project: the workstation, the books, or anything else. I would also like to acknowledge the financial support provided by Dr. Kabra while working on this project. It has been a pleasure to work and learn from Dr. Kabra.

I would also like to thank my co-advisor, Dr. Derek Paley. His advice and suggestions helped me greatly to improve and fine-tune this work. I would also like to thank him for being the chair of my thesis committee. I also owe my gratitude to Dr. Ilya Rhyzhov for agreeing to serve on my thesis committee and for sparing the invaluable time reviewing this work.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Dockless Bike-Sharing Systems

Almost every major city worldwide has implemented a bike-share system to make its transportation system smarter and sustainable. In the United States alone, nearly 200 cities have bike-sharing systems, multiple of them in many cases [1]. Over 136 million trips were completed on shared bikes and scooters in 2019 [2]. The bike-sharing systems can be divided into docked and dockless categories based on whether the bikes are picked up from and dropped off at a fixed station or not. Recent years have seen a shift from traditional docked bike-sharing systems to hybrid or completely dockless systems. The dockless system, in 2019, accounted for 70% of all shared micro-mobility trips in the US [2]. This is primarily because the dockless systems solve the last-mile problem of local transportation. Other than that, dockless systems improve user experiences, offer high flexibility to the riders, and provide tighter integration with the public transit system [3]. The ability to drop off bikes anywhere provides dockless systems an inherent advantage over station-based systems. This flexibility, however, introduces some significant problems. The dockless shared bikes are often parked or abandoned in pathways, rivers, and other public spaces that destroy the city's aesthetics and create safety hazards and public nuisance. Because of this reason, the number of bikes in the fleet of any dockless bike-sharing company is regulated by the city's transportation authority. It is revised regularly, consistent with the city's needs and performance of the bike-sharing companies. With a fixed number of bikes in hand, each bike in the fleet needs to be appropriately managed and optimally utilized to meet the city's dynamic bike supply-demands. Bike-providers have to regularly re-distribute the bikes across the city to keep up with the supply-demand imbalance.

1.2 Supply-Demand Imbalance in Dockless Bikesharing Systems

The most critical aspect of meeting the dynamic supply-demands in the city is rebalancing. Rebalancing is the distribution of bikes from areas with a surplus to areas with a shortage of bikes. Rebalancing is critical to meet fluctuations in bike demands over the city and nearly accounts for 30-50% of the operating costs for the bike companies [4]. Rebalancing is more difficult in dockless systems where individual bikes are scattered in an unplanned manner and are often challenging to locate. The number of bikes at any location for these dockless systems depends heavily upon user behavior. Their spatial distribution in the city changes after every trip and results in a significant supply-demand imbalance. This supply-demand imbalance can be divided into local and hyperlocal imbalances. The local imbalance is due to similar bike usage patterns of the users. For example, users mostly ride from their houses to subway stations during morning peak hours, which leads to the accumulation of bikes at the subway station and leaves very few bikes in the residential areas. This type of imbalance is generally countered by manual redistribution of the bikes, mostly overnight and sometimes during the day hours. On the other hand, the hyperlocal imbalance is concentrated in a small geographical area (few blocks) where users show higher dispersion in the choice of dropoff locations but a concentrated choice of pickup locations. This phenomenon occurs because users prefer to pick up from the locations where they are certain of having bikes available. When dropping off bikes, no such concern arises. This results in bikes being scattered in a block or two away from favored pickup locations. The likelihood of bikes being picked up at the other locations is comparatively much lower than the favoured locations. This hyperlocal imbalance phenomenon is a feature of dockless bike-sharing systems. The stations in docked systems, which are often more than a few blocks apart, essentially aggregate pickups and dropoffs and alleviate or eliminate this issue. There is little or no literature available on the hyper-local imbalance in dockless bike-sharing systems. In this work, we introduce the concept of hyper-local imbalance, demonstrate this phenomenon using the trip and GBFS (General Bikeshare Feed Specification) data of dockless bike-sharing companies operating in Washington DC, and explore the primary reasons behind it.

1.3 Trip Data Scarcity

Conducting research on the user-behavior of bike-sharing consumers requires access to trip-level data, which bike-share companies are wary of providing due to competition or privacy issues. Trip-level data is also required by the government authorities to oversee and regulate the bike-sharing operations. Easy access of bike-sharing trip data to researchers can play a pivotal role in solving complex bike-sharing problems like rebalancing. In this work, we introduce an algorithm to extract the trip level information from General Bikeshare Feed Specification (GBFS) data. GBFS is a specification for real-time, read-only data targeted to provide transit information to the shared mobility end user [5]. It mainly provides locations of all parked vehicles at any time and is a must for the bike-providers to make publicly available for authorities oversight. We validate the trips extracted by the algorithm in a time period against the actual trips that took place during the same time period. The data for the actual trips for the brief time period is provided by the DDOTS. We also use this algorithm to extract trips and study hyper-local imbalance among the dockless bike-sharing companies operating in Washington DC.

1.4 Contribution of the Paper

This paper makes the following contributions:

1. Propose and validate an algorithm to extract trip level data of bikesharing systems from GBFS feeds.

2. Introduce the concept of hyperlocal imbalance, demonstrate it using the trip-data of dockless bikeshare systems operating in Washington DC, and explore the primary reasons behind it.

1.5 Thesis Outline

The thesis is divided into five chapters. Chapter 2 provides a background on GBFS data and hyperlocal imbalance. Chapter 3 describes the algorithm to extract trip-level information from the GBFS data feeds. Validation of the algorithm is also done in this chapter 3. We use the algorithm developed in Chapter 3 to extract trip data from the GBFS feeds to study the problem of hyperlocal imbalance in chapter 4.

Chapter 4 demonstrates the problem of hyperlocal imbalance in dockless bikes using the trip data of dockless bikes operating in Washington DC. The trip data is obtained by extracting trips from the GBFS feed collected for dockless bikes in Washington DC over a period of one year. We further explore the possible reasons behind the hyperlocal imbalance in Chapter 4. Chapter 5 concludes the thesis and discusses the possible future directions.

Chapter 2: Background

2.1 General Bikeshare Feed Specification (GBFS) Data

GBFS (General Bikeshare Feed Data) is a specification for real-time, read-only data targeted to provide transit information to the shared mobility end-user [5]. It requires the bike-sharing companies to publish data feeds that provide real-time information on all currently parked bikes' absolute locations. The data feeds need to be published as JSON files updated at every TTL (Time-to-Live) seconds. The TTL values generally vary from 30 to 300 seconds [5]. Around 290 bike-share systems worldwide have adopted the GBFS data standard since its release in November 2015 [5]. In many cities, bike providers must publish the GBFS data feeds to enable government authorities to oversee and regulate bike companies and uses. GBFS data is also the backbone for many tools that make shared mobility more accessible to the user. In this work we introduce another utility of the GBFS data feeds. We introduce an algorithm in Chapter 3 that can be used to extract trip level information from the GBFS data feeds. Trip level data contains origin and destination pair of bike's trip along with the trip duration. Origin and destination location are specified by it's longitude and latitude values.

2.2 Hyperlocal Imbalance

Hyperlocal imbalance refers to the supply-demand imbalance in a small geographical region (an area of a few blocks) within which the users show a remarkable propensity to pick up bikes only from some preferred locations. It results in bikes being consumed rapidly from these preferred locations, whereas bikes at other locations in the region sit idle for an extended period. Bikesharing companies have to regularly channel bikes at the preferred locations to keep abreast with the demand levels, even though the bikes are available for consumers' use at a small walking distance from the favored spot. The primary reason for this bias is the consumer's preference to pick up the bikes from the locations of high bike availability. The users would know which nearby area would consist of more available bikes to choose from through their experience or the mobile application.

A hyperlocal imbalance example is illustrated in Figure 2.1 and Table 2.1 using the dockless bike-sharing trip data in Washington DC. It maps the region in a radius of 270 yards around the traffic intersection of Connecticut Avenue NW and 24th Street NW in Washington DC. The pickups and dropoffs of dockless bikes in the region are aggregated over the nearest traffic intersection. The figure shows the pickup (right panel) and dropoff (left panel) share for bikes attributed to the region's nearest intersection.

Intersection ID	**1930 (Ref)**	**309**	**5669**	**7812**	**259**	**8004**	**196**
Relative Rate	1	0.17	0.02	0.47	0.22	0.27	0.1
Distance from Ref Int (Yards)	0	250	110	175	170	87	230

Table 2.1: Relative Rate and Distance from Reference Intersection values at intersections in the region. Relative rate is the ratio of pickup rate at any intersection to the pickup rate at the reference intersection (intersection 1930).

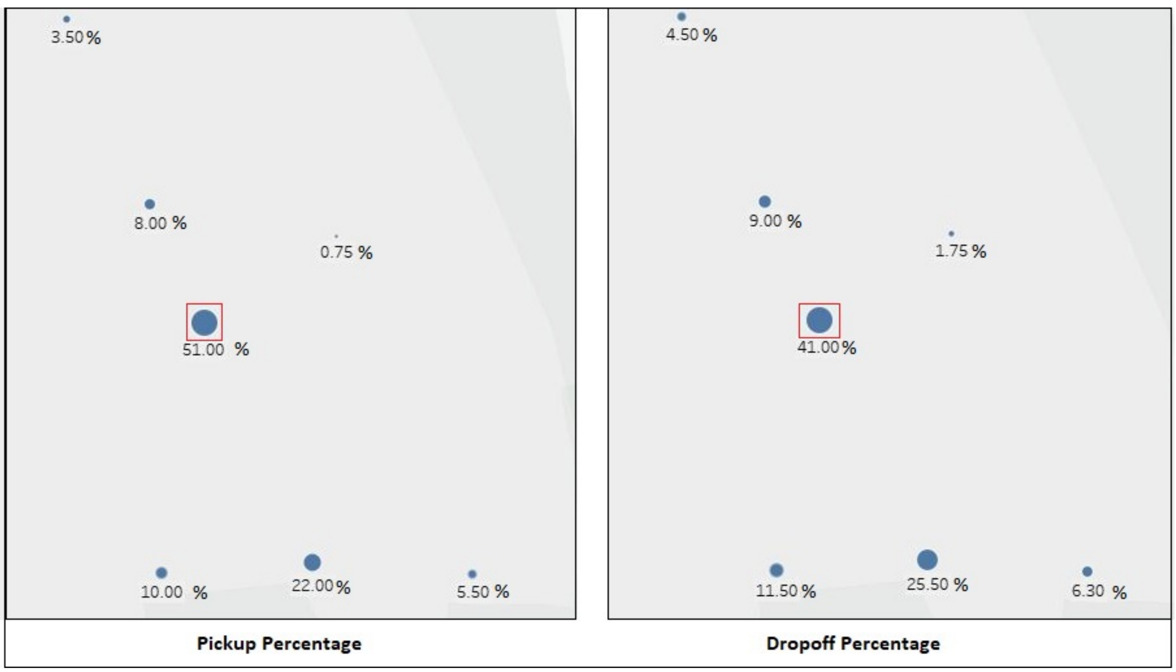

Figure 2.1: Pickup and dropoff percentages at intersections within 270 yards of the intersection of Constitution Avenue NW and 7th Street NW of Washington DC.

We notice that at the marked intersection in Figure 2.1, the pickups are much more concentrated than the dropoffs. The average daily pick-ups at the marked location is 36, whereas the average daily drop-off is just 20. This means that 16 bikes are peddled in daily at the marked location by the bike-providers to meet the actual-demand. We also observe from Table 2.1 that the pickup rate at all unmarked intersections is less than half of the pickup rate at the marked intersection. This points out the under-utilization of the bikes at all unmarked intersections in the region. Bikes at the unmarked intersections have to wait for a considerably longer time before being picked up compared to the bikes placed at the marked location. In Chapter 4, we demonstrate hyperlocal-imbalance in the dockless bike-share systems among the busiest cluster of intersections in Washington DC using the HHI (Herfindahl-Hirschman Index) [12].

Chapter 3: Extracting trip information from GBFS Data

3.1 Introduction

This chapter introduces an algorithm that extracts trip-level-information from the GBFS feeds over a period of time. Extracting trip-level info allows researchers easy access to the fine-grained trip data, which can be used to understand user-behavior, trip-patterns and improve the city's bike-sharing infrastructure and efficiency. Trip level info is generally not available to the research community due to privacy issues. We validate the algorithm and measure its performance by comparing its results with the actual trip info (provided by DDOTS) of bike companies operating in Washington DC.

3.2 Data Collection

This section details the methods to collect the GBFS data required for the proposed algorithm. We have collected this data for bike-sharing companies operating in Washington DC in the time period starting from the second last week of December in 2019 to the last week of March in 2020. Table 3.1 lists fields scraped for each of the bike companies. The time interval between subsequent scraping for a bike-provider is thirty seconds. The Vehicle IDs uploaded by bike-providers into the GBFS feed can either be static or dynamic. Static Vehicle ID means that the unique identifier for the vehicle would remain the same with time. In contrast, dynamic vehicle ID means that the vehicle's unique identifier would change with every trip or time-step. Table 3.2 provides the list of companies and bike ID types they upload into the GBFS feed. GBFS datasets with static vehicle IDs can be utilized to extract trip-level origin-destination pairs. The trip-level info, then, can be used to determine the trip-distribution in the city. For the GBFS datasets with dynamic bike IDs, it is not possible to extract trip origin and destination pairs, but it can still be used to determine the trip-distribution. By trip-distribution, we mean the total number of pickups and dropoffs from different parts of the city. We save the scraped GBFS data into the CSV file format. CSV file at each time-step contains a list of all the vehicles operating in the city (operated by a particular bike-provider) along with their unique identifier and location. The location for any parked vehicle is expressed by its longitude (lon) and latitude (lat) values.

3.3 Algorithm to extract trip info from GBFS data

This section describes an algorithm to extract trip origin-destination pairs from GBFS data feeds that use static vehicle IDs. GBFS feed reflects a bike's position only when it is idle. It disappears from the GBFS feed as soon as any rider rents it for a trip and re-appears in the feed when the trip is over. The location just before the disappearance from the GBFS-feed is the trip-origin, and the location at

Field	Description
Vehicle ID	Unique Vehicle Identifier
Latitude	Latitude of the parked bike location
Longitude	Longitude of the parked bike location
Battery Level	Battery level of the parked bike
Vehicle Type	e-Scooter or e-Bike
Time Stamp	Time at which the data is scraped
TTL	Time to Live (in seconds)

Table 3.1: Columns extracted from GBFS data.

Bike Provider	Helbiz	Jump	Razor	Skip
Vehicle ID type	Static	Static	Static	Static
Bike Provider	Spin	Lime	Lyft	Bird
Vehicle ID type	Static	Dynamic	Dynamic	Dynamic

Table 3.2: Bike providers and GBFS vehicle ID type.

which it re-appears in the feed is the trip-destination. The time period between the subsequent disappearing and reappearing of the bike from the GBFS feed is the trip duration.

To extract trips for a unique bike, we check its location at each time-step available in the record. If the location of a vehicle changes at $(j+1)^{th}$ time-step and distance between its initial and final location is greater than the threshold distance, we record the j^{th} and $(j+1)^{th}$ timestamp as the *start* and *end* time of the trip. Similarly, we record the location of vehicle at j^{th} and $j+1^{th}$ timestamp as the trip *origin* and *destination*. The algorithm to extract trips from the GBFS feed data is described in Algorithm 2. The helper function, which calculates the distance between two longitude and latitude pairs, is described in Algorithm 1. The distance calculation in Algorithm 1 is based on the spherical law of cosines [6].

Algorithm 1: Function to calculate distance between lon and lat pair.

procedure DISTANCE(lon_1, lat_1, lon_2, lat_2)
  R $\leftarrow$ Radius of Earth (6317 km).
  VS $\leftarrow$ $Sin(lat_1) \times Sin(lat_2)$
  VC $\leftarrow$ $Cos(lat_1) \times Cos(lat_2) \times Cos(lon_1 - lon_2)$
  V $\leftarrow$ VS + VC
  distance $\leftarrow$ R $\times cos^{-1}(V)$
  return distance
end procedure

3.4 Comparison with Actual trips

We validate the trips extracted by Algorithm 2 from the GBFS feed by comparing them against the actual trip-data provided by DDOTS. The validation is done for all the trips occurring in the first week of January 2020 by the vehicles operated by *Skip* (bike-share-company) in Washington DC. For clarity, we refer to the dataset with trips extracted from the GBFS feed as the *gbfs-trip* dataset and the dataset with the trips provided by DDOTS as the *actual-trip* dataset. The validation process includes comparing the origin-destination pairs in the *gbfs-trip* dataset with the origin-destination pairs present in the *actual-trip* dataset.

Algorithm 2: Extract trip info from GBFS data with static Bike IDs.

Input GBFS data scraped in time-period ΔT, ($GBFS_{\Delta T}$).
$ubid \leftarrow$ Unique Bike IDs in $GBFS_{\Delta T}$.
$thr\_dist \leftarrow$ minimum distance that a trip can have (threshold distance).
for $i = 1, 2, \ldots, length(ubid)$ **do**
    $GBFS_I \leftarrow$ Records with vehicle id $= ubid_i$ in $GBFS_{\Delta T}$.
    $GBFS_I \leftarrow$ Sort $GBFS_I$ rows by time (ascending).
    for $j = 1, 2, \ldots, length(GBFS_I) - 1$ **do**
        if $((lon_j \neq lon_{j+1})$ **and** $(lat_j \neq lat_{j+1}))$ **then**
            if distance(lon_j, lat_j, lon_{j+1}, lat_{j+1}) $> thr\_dist$ **then**
                Trip Origin $\leftarrow$ location in j^{th} row
                Trip Destination $\leftarrow$ location in $j + 1^{th}$ row
                Trip Start Time $\leftarrow$ timestamp in j^{th} row
                Trip End Time $\leftarrow$ timestamp in $j + 1^{th}$ row
                Ouput Origin-Destination Pair.
            end if
        end if
    end for
end for

To easily compare the origin-destination pairs, we attribute the origin or destination location to the nearest (in terms of distance) available traffic-intersection. Attributing origin/destination to a traffic-intersection means that we assume that the trip originated or ended at that particular traffic-intersection instead of the nearby location. We then compare the trips in the *gbfs-trip* and *actual-trip* dataset based on the day and hour the trip started, starting nearest traffic-intersection and ending nearest traffic-intersection. [7] provides information about all the traffic-intersections available in Washington DC, including their location (longitude and latitude values) and a unique identification ID. Distance between the vehicle's and traffic intersection's location is calculated based on their longitude and latitude values using the spherical law of cosines [6]. Note that we do not attribute any origin and destination location to a traffic-intersection if the distance between the origin/destination and nearest traffic-intersection location exceeds 270 yards.

The trips in the *gbfs-trip* dataset may contain some invalid origin-destination pairs due to an error in reflecting the bike's position by the on-board GPS module. The common reason includes proximity to buildings, trees, bridges, etc., which results in the GPS signals being blocked or reflected [8]. To eliminate these invalid pairs, we reject all trips that took place in a very short time period (a period of fewer than 2 minutes), or for a very short distance (distance less than 250m), or at an infeasible speed (speed $>$ 20 mph).

Another reason for the existence of invalid trips in the *gbfs-trip* dataset is Rebalancing. It is typically done by collecting vehicles in the wrong place or at a deficient battery level into a truck and distributing them over the town after charging. It generally occurs during the night [9] [10] [11]. To counter the possible invalid trips introduced due to rebalancing, we reject all the origin-destination pairs occurring late-night and early-morning and focus on the trips during the hours when the riders are most active (8 AM to 9 PM). We refer to all the steps taken to eliminate invalid trips from the *gbfs-trip* dataset as the fine-tuning activities.

The actual trip dataset contains 6480 trips (*actual-trips*) occurring in the first week of January 2020 between 8 AM to 9 PM completed by the vehicles operated by skip (bike-share company) in Washington DC. Algorithm 2 extracts 6154 trips after fine-tuning activities from the GBFS feed for the same time-period. On comparing the trips in *gbfs-trip* and *actual-trip* dataset based on the day and hour the trip started, starting, and ending traffic intersection ID, we found that 4944 trips in the *gbfs-trip* dataset matches exactly with the *actual-trip* dataset (*true-gbfs-trips*), the remaining 1210 trips are *false-gbfs-trips* generated by the Algorithm. We evaluate the algorithm's performance by calculating its precision and recall. Precision is defined as the fraction of generated trips (by the algorithm) that matches the actual trips, while its recall is defined as the fraction of actual trips generated by the algorithm. We measure the accuracy of the algorithm using the F1 value. F1 value is the harmonic mean of the calculated precision and recall. Equations 3.1 - 3.3 list the formula to calculate precision, recall, and F1 values for the algorithm.

$$\text{Precision} = \frac{\text{Trips generated by Alg. 2 that matches the actual trips}}{\text{Total trips generated by Alg. 2}} \tag{3.1}$$

$$\text{Recall} = \frac{\text{Trips generated by Alg. 2 that matches the actual trips}}{\text{Total actual trips}} \tag{3.2}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

Table 3.3 lists the value of the metrics measuring the algorithm's performance against the actual trip-data.

Figures 3.1 and 3.2 shows how trips are distributed over the week and day.

Metric	Precision	Recall	F1-value
Value	0.8033	0.763	0.7826

Table 3.3: Metrics measuring performance of Algorithm 2

For the day, it only shows the hours between 8 AM to 9 PM. We notice from these plots, that the trips in both the datasets are similarly distributed over the week and day. The same is true for the distribution of trip-duration in the *gbfs-trip* dataset and *actual-trip* dataset.

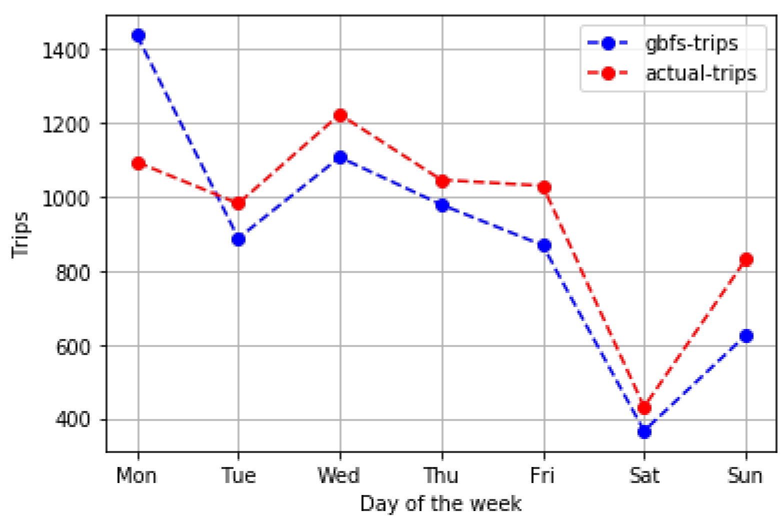

Figure 3.1: Trips vs day of the week.

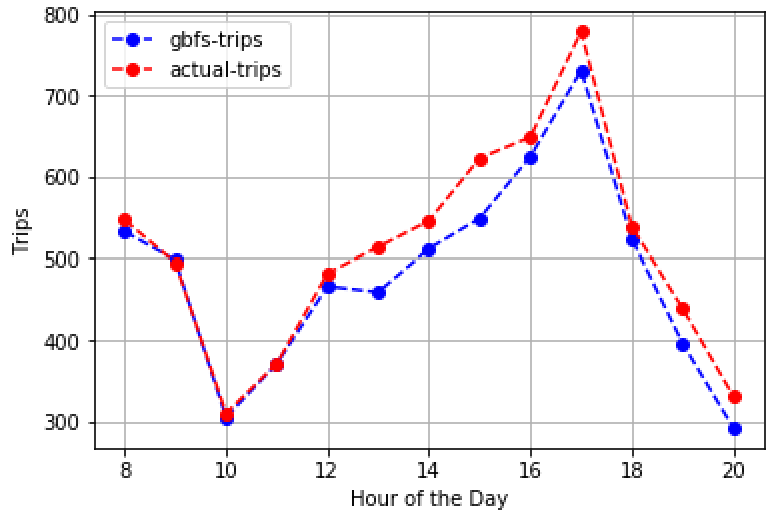

Figure 3.2: Trips vs hour of the day.

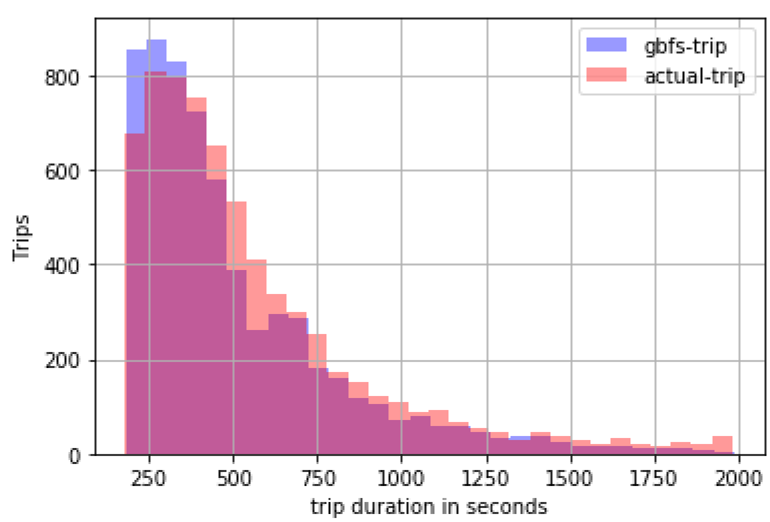

Figure 3.3: Trip duration distribution.

Chapter 4: Hyperlocal Imbalance in Dockless Bike-sharing Systems

4.1 Introduction

Hyperlocal imbalance refers to the supply-demand imbalance in a small geographical region (an area of a few blocks) within which the users show a propensity to pick up bikes only from some preferred locations. In this chapter, we demonstrate hyperlocal-imbalance in the dockless bike-share systems among the busiest cluster of intersections in Washington DC using the HHI (Herfindahl-Hirschman Index) [12]. We further explore the reasons behind the hyperlocal imbalance among the cluster of intersection in Washington DC.

4.2 Data Collection and Processing

4.2.1 Data

The analysis is based on the *GBFS* (General Bikeshare Feed Specification) and *trip* data collected for the nine dockless bike-sharing companies operating in Washington DC in 2020. GBFS publishes the position of all the available bikes/scooters in real-time. The GBFS feeds are non-archival records, and the data available to us has been collected over the entire period of the year 2020. The trip data available has been extracted from the GBFS data using the algorithm described in Chapter 3. For this chapter's purposes, all the analysis is done based on the trip and GBFS data of one month in 2020, starting from the first week of February to the first week of March. The fields of interest extracted from the trip and the GBFS data for processing and analysis in Chapter 4 are listed in Table 4.2.

Trip Data	**GBFS Data**
• Bike ID • Trip Start Time • Trip End Time • Trip Origin Longitude • Trip Origin Latitude • Trip Destination Longitude • Trip Destination Latitude • Bike Provider (Company)	• Bike ID • Time • Bike Location Longitude • Bike Location Latitude • Bike Provider (Company)

Table 4.1: Field of interest in trip and GBFS dataset

4.2.2 Attributing Pick-ups and Drop-offs to Traffic Intersections

Although the dockless bikes are scattered in a region, in this work, we attributed the dockless bikes to the nearest traffic-intersection (referred to as intersection hereinafter) in Washington DC. Data for the intersections in Washington DC is obtained from Open Data DC [7], which lists the name, longitude, latitude, and other attributes for all 8437 intersections in Washington DC. The trip origin and destination are attributed to the nearest available intersection based on the distance calculated between bike location and intersection using their longitude and latitude values. Please note that we do not attribute any pickup and dropoffs to an intersection if the least distance between the pickup/dropoff location and nearest intersection exceeds 270 yards. Attributing pickups/dropoffs to an intersection means that we assume that the trip originated or ended at that particular intersection instead of the location nearby. This provides us with a method to divide Washington DC into different sectors, accumulate dropoffs and pickups in these sectors, and then perform further analysis. The formula used for calculating the distance between bike and traffic intersection location using their longitude and latitude values is listed in Equation 4.1. The abbreviations *lab*, *lob*, *lai*, and *loi* in Equation 4.1 represents Bike Location Latitude, Bike Location Longitude, Intersection Latitude, and Intersection Longitude respectively. We collect the total number of pickups and dropoffs at each of the 8437 intersections present in Washington DC.

$$\text{distance} = \arccos\left(\sin\left(lab\right) sin(lai) + \cos\left(lab\right)\cos\left(lai\right)\cos\left(lob - loi\right)\right) \tag{4.1}$$

Out of the 8437 intersections in DC, not all intersections are attributed to a significant number of pickups. Figure 4.1 plots the cumulative pickups vs. the number of intersections. We notice that out of the total 8437 intersections, approximately 2000 intersections accounts for almost all of the dockless pickups in Washington DC. For our purpose, we would focus only on these 2000 intersections.

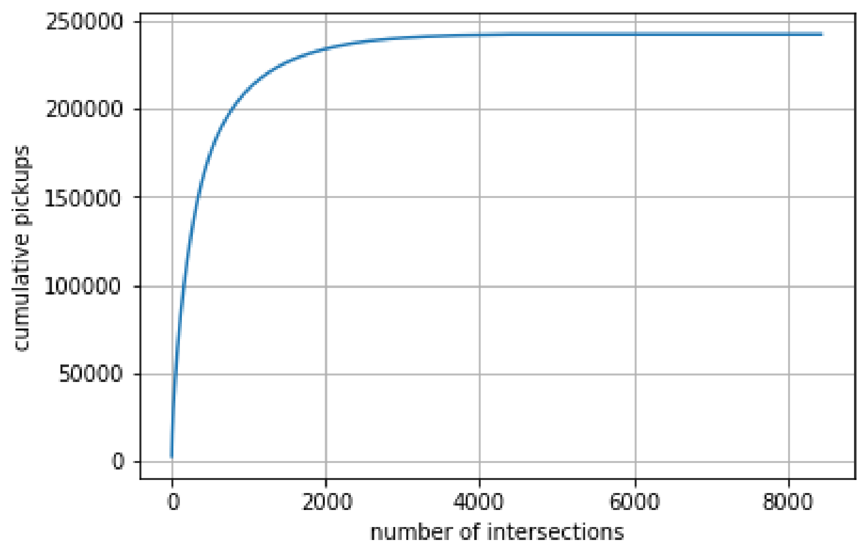

Figure 4.1: Cumulative pickups vs number of intersections

4.2.3 Cluster of Intersections and Metrics of Interest

Many of the intersections in Washington DC, which are very close to each other (at a distance of almost one or two blocks away), are consolidated into a cluster. Almost all of these clusters have one popular intersection that accounts for the majority of the pickups in the cluster; this intersection is referred to as the *reference* or the *central* intersection of the cluster. All the other intersections of the cluster are referred to as auxiliary intersections. The average distance of auxiliary intersections from the reference intersection in their corresponding cluster is around 200 yards (note that the average block length is 220 yards).

At each of the intersections in Washington DC, we use GBFS data to calculate the bikes' mean availability. Please note that the GBFS data provides the real-time location (longitude and latitude values) of all stationary (bikes that are not engaged in any trip) dockless bikes. We again attribute the location of each bike to the nearest available intersection. At each time-step, we calculate the total number of bikes at each intersection. We then use this to calculate the average number of bikes available at each intersection (mean availability). Using the GBFS data, we also calculate the time-period for which at least one bike is available at each intersection (available time). Along with the bike availability and time available, we also calculate pickup rate, relative rate, and relative availability for each intersection. Equations 4.2, 4.3, and 4.4 lists the formulas to calculate the pickup rate, relative rate, and relative availability at intersection k (int_k) belonging to the cluster K ($cluster_K$), which has ref_int_K as the reference/central intersection.

$$\text{Pick-up Rate } (int_k) = \frac{\text{Total Pickups } (int_k)}{\text{Available Time } (int_k)} \tag{4.2}$$

$$\text{Relative Rate } (int_k) = \frac{\text{Pickup Rate } (int_k)}{\text{Pickup Rate } (ref\_int_K)} \tag{4.3}$$

$$\text{Relative Availability } (int_k) = \frac{\text{Mean Availability } (int_k)}{\text{Mean availability } (ref\_int_K)} \tag{4.4}$$

Table 4.2 presents the snapshot of the information accumulated for each of the intersections in Washington DC using the GBFS and trip data.

Int ID	Lon	Lat	Cluster ID	Mean Avail	Pickups	Dropoffs	Pickup Rate	Rel. Rate	Rel. Avail
1930	-77.0523	38.9249	C1930	13	674	427	0.03	1	1
196	-77.0533	38.9266	C1930	1	45	48	0.003	0.1	0.08
259	-77.0526	38.9235	C1930	3	135	118	0.007	0.22	0.230

Table 4.2: Snapshot of information collected for each intersection in Washington DC

4.3 Demonstrating Hyperlocal Imbalance

4.3.1 Measuring Hyperlocal Imbalance in Washington DC

This section focuses on demonstrating the hyper-local imbalance in dockless bike-sharing systems using the information collated in section 4.2.3 about the pickup and dropoff numbers at each intersection in Washington DC. We quantify this imbalance using the HHI (Herfindahl-Hirschman Index) metric [12]. HHI is a commonly accepted metric to calculate the market concentration; it is calculated by squaring each competing firm's share in the market and then summing the resulting number. For our purposes, we calculate the pickup and dropoff HHI by squaring the share of pickups/dropoffs for each intersection in the cluster and then summing the resulting numbers. The pickup and dropoff HHI gives information about the concentration level of pickups and dropoffs in the cluster. At the same time, their ratio quantifies the contrast between pickup and dropoff distribution among the cluster's intersections. Equations 4.5 to 4.7 summarizes the calculation of pickup HHI, dropoff HHI, and the HHI ratio for a cluster with a total of N intersections. p_n and d_n are the pickup and dropoff share of the nth intersection in the cluster.

$$\text{Pickup HHI} = \sum_{n=1}^{N} p_n^2 \tag{4.5}$$

$$\text{Dropoff HHI} = \sum_{n=1}^{N} d_n^2 \tag{4.6}$$

$$\text{HHI Ratio} = \frac{\text{Pickup HHI}}{\text{Dropoff HHI}} = \frac{\sum_{n=1}^{N} p_n^2}{\sum_{n=1}^{N} d_n^2} \tag{4.7}$$

Ideally, pickup and dropoff distribution in any cluster should be equal. The most efficient bike-sharing system would have an equal number of pickups and dropoffs at each intersection, which means an HHI ratio of 1 at each cluster. However, in general, this is not the case. The pickup and dropoff distribution vary substantially. For the hyperlocally-imbalanced cluster, pickups are more concentrated. We can use the HHI ratio to calculate the percentage difference (pd) by which cluster's pickups are more concentrated than the cluster's dropoffs. The formula for the same is listed in equation 4.8.

$$pd = (\text{HHI ratio} - 1) \times 100 \tag{4.8}$$

Figure 4.2 shows the HHI-ratio distribution among all the clusters in Washington DC. It shows that the pickups are more concentrated for most clusters than the dropoffs (i.e., HHI ratio > 1). We focus only on the clusters where pickups are at least 10 percent more concentrated than the dropoffs. We refer to these clusters as hyperlocally-imbalanced clusters (HIC). Other clusters are referred to as normal clusters (NC). Around 30% of all clusters falls into the hyperlocally imbalanced category. Table 4.3 summarizes the distribution of HHI ratios among these clusters. It shows the Percentage of clusters in a given HHI ratio range. On average, the pickups for these hyper-locally imbalanced clusters are 20% more concentrated than the dropoffs. Table 4.3 also shows the percentage of all pickups coming out from these clusters to manifest their importance. It shows that around 36% of all pickups in Washington DC are from the clusters which suffer from hyperlocal imbalance. We further explore how the disparity in pickup and dropoff concentration varies throughout the day. Table 4.4 lists how the HHI-ratio varies through different parts of the day for the hyperlocally imbalanced clusters. From Table 4.4, it can be noted that there is not much difference in the distribution of HHI ratios throughout the day. The total percentage of hyperlocally imbalanced clusters remains almost the same for each part of the day, highlighting that even with consumer's different pickup behavior with the time of day [15], the problem of hyperlocal imbalance among the cluster of intersections remains the same.

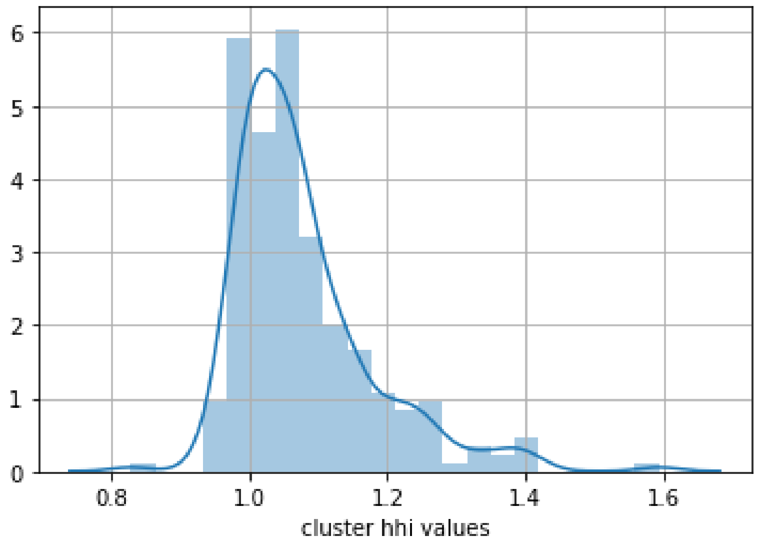

Figure 4.2: pdhhi distribution among the clusters in Washington DC.

HHI-ratio	[1.1, 1.2)	[1.2, 1.3)	[1.3 ++]	Sum
% of clusters	16.12	8.68	4.13	28.93
% of pickups	21	10	5	36

Table 4.3: Percentage of clusters/pickups in the given HHI ratio range.

4.3.2 Hyperlocally Imbalanced Clusters vs. Normal Clusters of Intersections.

As mentioned in Section 4.2.3, the clusters are divided into central and auxiliary intersections, where the central intersection carries the most number of pickups in the cluster. For the hyperlocally imbalanced clusters (HIC), the pickup rate of the bikes at the majority of the auxiliary intersection is meager as compared to the central intersection, which points out that the bikes at the auxiliary intersections of the HIC have to wait for a considerably longer period of time before being picked

HHI ratio	all day	morning	late morning	afternoon	evening	late evening
[1.1, 1.2)	16.12	13.22	19.01	20.25	19.01	16.53
[1.2, 1.3)	8.68	7.44	9.09	7.85	8.68	10.74
[1.3++)	4.13	11.16	6.61	4.96	3.72	2.48
Sum	28.93	31.82	34.71	33.06	31.5	29.75

Table 4.4: Percentage of clusters in the given HHI ratio range for different parts of day

up as compared to the central intersection of the HIC. The low pickup rate of bikes at any location leads to low utilization of bikes, which reduces the system's overall efficiency. To exhibit this phenomenon, we use relative rate (see Section 4.2.3). Relative rate or cluster normalized pickup rate is calculated for any intersection by dividing the intersection's pickup rate by the pickup rate of the central intersection of the cluster to which it belongs. Figure 4.3 shows the distribution of relative rates of auxiliary intersections for both the normal and hyper-locally imbalanced clusters. It can be noted that a high percentage of auxiliary intersections in hyperlocally imbalanced clusters have very low relative pickup rates, unlike normal clusters where we see a symmetric distribution of relative rates. It reveals the selective behavior of users in hyperlocally imbalanced clusters who favors the central intersection more dominantly, unlike consumers in the normal cluster who are open to picking up bikes from the auxiliary intersection as well.

Table 4.5 summarizes the comparison between the hyperlocally imbalanced

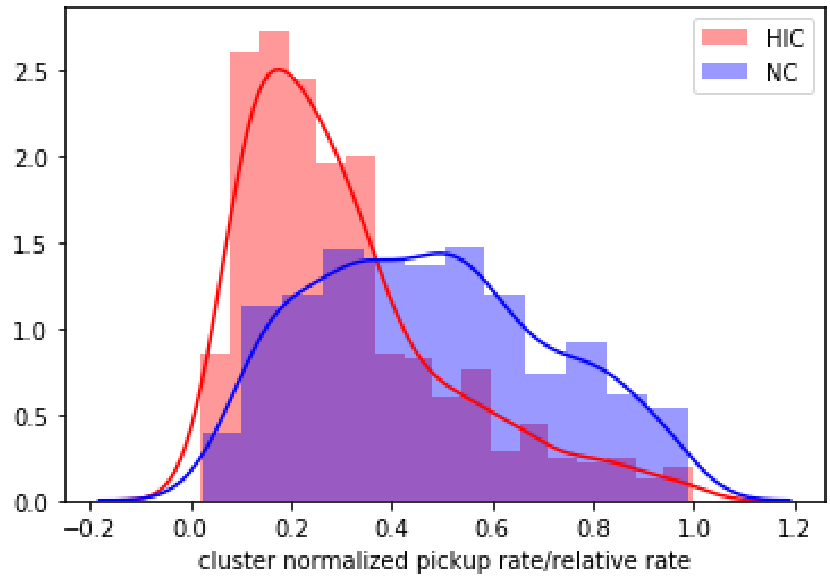

Figure 4.3: cluster normalized pickup rate for intersections among the clusters with hhi greater and less than 1.1.

and normal clusters based on their relative pickup rates at the auxiliary intersections. Note that around 49% of auxiliary intersections among the hyperlocally imbalanced clusters have had a relative pickup rate of less than 0.25. In other words, on average, for each of the hyperlocally imbalanced clusters, almost half of its auxiliary intersections have their pickup rate less than a quarter of the pickup rate at the central intersection, a location which on average is just 200 yards away. We also note that auxiliary intersections in the hyperlocally imbalanced clusters account for around 33% of total dropoffs and around 37% of the total availability. These bikes would have been better utilized if they had been placed just a block away. For hyperlocally imbalanced clusters, the bike operators have to regularly channel-in bikes at the central intersection to meet their demand values even when usable bikes are available just one block away. This is not the case with the normal clusters, as their auxiliary intersections with relative pickup rates less than 0.25 account only for 13% of the total dropoffs compared to the mammoth one-third of the dropoffs for the hyperlocally imbalanced clusters. The users in the hyperlocally imbalanced clusters drop off the bikes at locations from where the probability of being picked up is comparatively minimal.

	Relative Pickup Rate	Hyperlocally Imbalanced clusters	Normal clusters
% of intersections in cluster where relative pickup rate is less than x.	x = 0.50 x = 0.25 x = 0.10	83 % 49 % 9.8 %	56 % 23 % 6.9 %
% of dropoffs at the intersections where relative pickup rate is less than x.	x = 0.50 x = 0.25 x = 0.10	64.7 % 32.9 % 4.6 %	44.4 % 13.3 % 2.8 %
% of bikes available at the intersections in the cluster where relative pickup rate is less than x.	x = 0.50 x = 0.25 x = 0.10	68 % 37 % 6.5 %	45.45 % 14.77 % 4.1 %

Table 4.5: Comparison between Central and Auxiliary Intersection of clusters

4.4 Primary Drivers of Hyperlocal Imbalance

4.4.1 Bike Availability vs. Pick-ups

We plot the relative availability distribution among auxiliary intersections for both the clusters to analyze why relative pickup rates are very low for auxiliary intersections in the hyperlocally imbalanced clusters compared to the normal clusters in Figure 4.4. On average, a greater number of auxiliary intersections with very low relative mean availability are present in the hyperlocally imbalanced clusters compared to the normal clusters. Around half of the auxiliary intersections in the hyperlocally imbalanced cluster has relative availability less than 0.2, compared to the quarter of the auxiliary intersections in the normal clusters. The primary reason for this is bikes being more spatially scattered in hyperlocally imbalanced clusters than in normal clusters. Note that, on average, dockless bikes are aggregated over 5 and 8 intersections in normal and hyperlocally imbalanced clusters. Greater spatial scattering in hyperlocally imbalanced clusters does not allow the bikes to reach a threshold bike-availability-level at several intersections necessary for consumers to feel confident about picking up the bikes. People are generally wary of bikes being hard to locate, being non-operational, or someone else reaching before them to pick up the bike first. Hence, they may prefer either to pick up the bikes from some high-availability intersection or use other modes of transport.

We have now identified that the hyperlocal imbalance and mean availability are closely associated. Consumers' over-inclination towards high bike availability

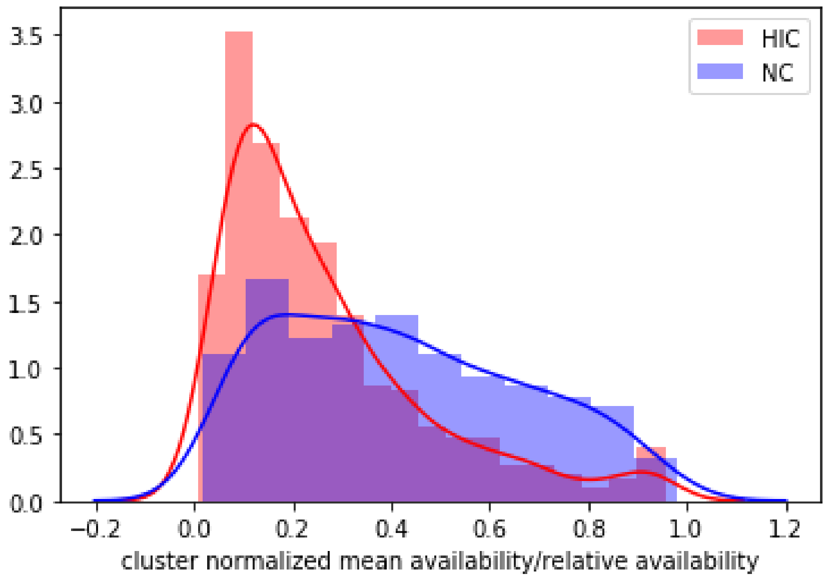

Figure 4.4: Mean availability distribution for intersections among the hyperlocally-imbalanced and normal clusters.

sites is one of the major reasons they skip picking bikes from some locations and favor picking up from other locations. These sites with low mean bikes' availability are known to the user through the bike company's mobile application or experience. The over-inclination prevents the users from picking bikes from intersections in the cluster with low mean availability. This also forces the bike-sharing companies to regularly channel bikes to popular locations, even though bikes may be available in a nearby region (region of one or at max 1.5 blocks away) for pickup. To investigate this relationship between bike availability and pickup numbers at intersections, we conducted regression analysis. We tried multiple regression models with pickups and intersections' mean availability as the dependent and independent variables. We wanted to find a relationship between them, which can help us deduct how the pickups at an intersection vary with their mean availability. The model which stood out in the regression analysis used log-transformed dependent and independent variables. The regression equation is expressed in Equation 4.9.

$$\ln(\text{pickups}) = \alpha \times \ln(\text{mean availability}) + \beta \tag{4.9}$$

No. Observations	R-square	Adj R-square	α	β
1176	0.466	0.465	1.4476* (0.045)	3.2129* (0.047)

Table 4.6: Regression Results with pickups as dependent variable and mean availability as independent variable.
Note : *p < 0.001

The regression analysis is based on the total number of pickups and the mean availability at intersections in one month. Results for the regression are listed in Table 4.6. We modify the results from Table 4.6 in equation 4.10 to evaluate daily pickups. As per equation 4.10, at any intersection, a 10% increase in mean availability results in a 15% increase in the number of pickups. This disproportionate increase in pickups with an increase in mean availability is the root of the hyperlocal imbalance.

$$\text{daily pickups} = \frac{24.85}{30} \times \text{mean availability}^{1.4476} \tag{4.10}$$

4.4.2 Auxiliary and Central Intersections

All the analysis we performed using the pickups, dropoffs, relative rates, and availability indicates that the consumers prefer picking up bikes at the central location instead of the auxiliary location for all intersections. In this section, we exhibit this by deriving a relationship between the pickups at the auxiliary intersections of the cluster with the cluster's central intersection's availability values. The relationship is derived by dividing the data into different time-periods (time-period of fifteen minutes) and collecting the following fields of interest for each time-period using the GBFS and trip data.

1. Total Aux. Pickups at time t (TAP(t)):

   Sum of pickup values at all the auxiliary intersections of the cluster.

2. Total Aux. Avail. at time t (TAA (t)):

   Sum of mean bikes available at all the auxiliary intersections of the cluster.

3. Central Avail. at time t (CA (t)):

   Mean availability at the central intersection of the cluster.

4. Cluster ID:

   Unique identifier identifying the cluster of the intersection to which each row of data belongs.

5. Time:

   Starting time of the fifteen minutes time period for which the data is collated in the row.

Time	TAP(t)	TAA(t)	CA(t)	Cluster ID
2020-02-07 17:45:00	0.0	5.36	2.70	C1930
2020-02-07 18:00:00	1.0	8.15	2.38	C1930
2020-02-07 18:15:00	0.0	8.15	2.10	C1930

Table 4.7: Snapshot of data used for regression in equation 11

$$\text{TAP(t)} = \alpha \ \text{CA(t)} + \beta \ \text{TAA (t)} + \theta \tag{4.11}$$

A snapshot of the data used for the regression is available in Table 4.7. Please note that we rearrange all the available one-month data to get the above-described fields of interest for each cluster at 15 minutes span. The regression equation is listed in Equation 4.11, and the results for the regression are listed in Table 4.8 . Note that we performed the regression after dropping the duplicate rows and with the cluster-ID fixed effects.

No. Observations	R-square	Adj R-square	α	β
36470	0.201	0.200	-0.0124* (0.002)	0.0362* (0.001)

Table 4.8: Regression Results (Note : $^*p < 0.001$) with TAP(t) as dependent variable and CA(t), TAA(t) as indepedent variable.

From Table 3.9, we observe that the coefficient for the CA(t) (availability at the center intersection of the cluster) is negative. This shows that the pickups at the auxiliary intersections decrease with an increase in availability at any cluster's central intersection. The negative coefficient for the CA(t) conforms with the observations and analysis conducted so far in this chapter. It reveals that availability at its central intersection significantly affects the pickups at its auxiliary intersections for a cluster.

Chapter 5: Conclusion

This thesis proposes and validates a novel algorithm to extract the trip-level origin-destination pairs from the GBFS dataset. It also introduces the concept of hyperlocal imbalance in dockless bikeshare, demonstrates its presence using the trip dataset of dockless bikes operating in Washington DC, and explores the primary reasons behind it. Chapter 2 provides a background on the GBFS data and hyperlocal imbalance. Chapter 3 introduces an algorithm to extract trip data from the GBFS feeds with static bike IDs with a great accuracy. Trip level information of dockless bike-sharing companies is not easily available to the research community due to privacy issues. Bike-providers often treat it as a commodity rather than a public resource that could enable users and cities to benefit from city movements and infrastructure planning [13] [14]. The capacity to extract trip information or trip distribution from the GBFS feed greatly improves the accessibility of trip-level data; this enables researchers and government authorities to optimize the bike-sharing systems' overall efficiency.

We also validated the algorithm by comparing the trips extracted from GBFS data with the actual available trip data provided by DDOTS. The algorithm was able to re-generate around 77% of all actual trips (Recall = 77%). Around 20% of the algorithm's trips did not correspond to any of the actual trips (Precision = 80%). sources for wrong trip generation are rebalancing of bikes and error in location reflection by the on-board GPS. We can reduce this number by using more refined tricks to eliminate invalid trips. One approach could be using the battery levels before and after a trip to eliminate the invalid trips, but not all bike-providers upload the battery levels in GBFS feeds. Future work includes refining the fine-tuning tricks to eliminate all invalid trips generated by the algorithm (i.e., to achieve 100% precision).Future work also includes a wide variety of analyses on the generated trip-data; the first is to analyze dockless bike riders' user behavior in Washington DC. We will use the generated trip data to compare the performance of dockless bikes against the docked bikes. We also plan to explore the trip patterns in equity emphasis areas in Washington DC; specifically, we want to explore how bikes' usability in these areas compare against the rest of the city. We also want to find out whether the regulations on bike distribution put forward by DDOTS in Washington DC really help the equity emphasis area's residents or not.

Chapter 4 introduces the concept of Hyperlocal Imbalance in dockless bikes. It is defined as the supply-demand imbalance occurring in an area of few blocks due to the user's bias to pick up bikes from some preferred locations and comfort to drop off bikes anywhere they want. There is no research available at present on the hyperlocal imbalance among dockless bikes. We demonstrate this phenomenon using the dockless bike data of bike-sharing companies operating in Washington DC. We found that almost 30% of all the clusters in Washington DC are suffering from hyperlocal imbalance.

Although the HHI metric is robust in measuring the hyperlocal imbalance of bikes, this thesis doesn't provide any insight about it's statistical significance. Future work includes analysis to determine the significance of HHI values measured for cluster of intersections. Another limitation for the HHI analysis conducted in the thesis comes from the data used for this purpose. The data available for the HHI analysis was just for one month (Feb 2020 to March 2020) and hence the calculated HHI metrics do reveal the state of hyperlocal imbalance during the low temperature months, when comparatively less number of people use dockless bikes. A better calculation of HHI metric should include data from all seasons (preferably the whole year data). Future work includes doing a wider temporal data collection (that includes all seasons) to calculate more robust HHI-metrics.

We also quantified the impact of hyperlocal imbalance on bike utilization using the pickup rate metric. We found that almost one-third of the bikes in the hyperlocally imbalanced clusters are placed at locations where their utilization is less than a quarter of what it would have been if they were placed just a block away. We also displayed that greater scattering of bikes in these clusters prevents bikes' availability at intersections to reach a level necessary for a user to feel confident about picking up the bikes. We also derived a regression relationship between pickups and availability at any intersection which reveals the user's over-preference towards the sites with high mean availability to pick up the bikes. A 10% increase in bike availability at any location results on average increases pickups by 15%.

We demonstrated through Washington, DC data that the auxiliary intersections in hyperlocally imbalanced clusters suffer from very low mean availability of bikes. The low bike-availability-level at these intersections doesn't make consumers comfortable enough to pickup bikes from there. Consumers do not prefer to walk up to these locations to pick up bikes as they are wary of bikes being damaged, dirty, disabled or being picked up by someone else before they reach there. Possible ways to alleviate this issue includes providing a pickup and drop-off incentive; bike-providers can provide incentive to consumers for picking up and dropping off bikes at low availability locations. Pickup incentive would enable the bikes at low availability locations to be better utilized, while dropoff incentive would allow the bikes at low availability locations to reach the threshold level required by the consumers to pick up the bikes. Bike-providers can also allow consumers to *reserve* the bikes from their mobile application so that they won't have to worry about someone else taking that bike before. Future work includes developing more solutions to curb the issue of hyperlocal imbalance.

Bibliography

[1] *Bureau of Transportation Statistics* (US Department of Transportation, 2019)

[2] *Shared Micromobility in the US* (NACTO, 2019)

[3] Zhang, Yuanyuan and Zhang, Yuming *Associations between Public Transit Usage and Bikesharing Behaviors in The United States* (In: Sustainability 10.6 (2018), p. 1868)

[4] *The Bikesharing planning guide* (ITDP, 2018)

[5] *General Bikeshare feed specification* (NABSA, 2020, URL https://github.com/NABSA/gbfs)

[6] Weisstein, Eric W., *Spherical Trigonometry* (Wolfram Web Resource, URL: https://mathworld.wolfram.com/SphericalTrigonometry.html)

[7] *Intersection Points, Washington DC* (Open Data DC, 2019, URL : https://opendata.dc.gov/datasets/96a9bbbb475648769e311d03c78698a7_2)

[8] *How Accurate is GPS* (GPS.gov, National Coordination Office for Space-Based Positioning, Navigation, and Timing, URL: https://www.gps.gov/systems/gps/performance/accuracy/)

[9] Chemla, D.; Meunier, F.; Calvo, R.W., *Bike sharing systems: Solving the static rebalancing problem* (Discrete Optimization 2013, 10, 120–146)

[10] Pal, A.; Zhang, Y., *Free-floating bike sharing: Solving real-life large-scale static rebalancing problems.* (Transportation Research Part C: Emerging Technologies 80 (2017): 92-116)

[11] Erdo ˘gan, G.; Battarra, M.; Calvo, R.W., *An exact algorithm for the static rebalancing problem arising in bicycle sharing systems* (European Journal of Operational Research 245.3 (2015): 667-679)

[12] Rhoades, Stephen A, *The herfindahl-hirschman index* (Fed. Res. Bull. 79(1993):188)

[13] Gössling, S., *ICT and transport behavior: A conceptual review* (International Journal of Sustainable Transportation, 12(3), 153–164, 2018)

[14] Spinney, J., Lin, W.-I, *Are you being shared? Mobility, data and social relations in Shanghai's public bike sharing 2.0 sector* (Applied Mobilities, 3(1), 66–83, 2018)

[15] Zhou, Xiaolu, *Understanding spatio temporal patterns of biking behavior by analyzing massive bike sharing data in Chicago* (PloS one 10.10 (2015): e0137922)