*ECAI 2024*
*U. Endriss et al. (Eds.)*

*doi:10.3233/FAIA241056*

# A Data-Driven Approach Supporting Location Decisions for Docking Stations in Bike-Sharing Systems

**Blerina Spahiu[a], Daniela Briola[a], Riccardo Sartori[a] and Giuseppe Vizzari[a,*]**

[a]University of Milano-Bicocca, Italy
ORCID (Blerina Spahiu): https://orcid.org/0000-0002-6958-8215, ORCID (Daniela Briola): https://orcid.org/0000-0003-1994-8929, ORCID (Giuseppe Vizzari): https://orcid.org/0000-0002-7916-6438

**Abstract.** Bike-sharing systems (BSSs) have become integral to urban mobility, improving accessibility, multimodality of transportation, and sustainability. This paper presents a novel approach to supporting decisions on the positioning of docking stations for dock-based BSSs by leveraging real-world historical mobility data to estimate mobility demand. In particular, we used taxi travels data as a proxy to generate a synthetic mobility demand dataset that was exploited to estimate the locations of a dock-based BSS stations through clustering techniques. This study aims to improve the practicality of station positioning. By addressing challenges related to station placement, this research offers insights into the practical implementation of data-driven approaches in BSS planning and management, advancing the efficiency and sustainability of urban bike-sharing systems.

## 1 Introduction

Urban mobility technologies have experienced a substantial change in the last two decades, driven by an increasing attention to environmental sustainability and by a shift of the socio-economic conditions in the cities. Mobility as a Service (MaaS) is a recent innovative transport concept that is still surrounded by ambiguity [11]. This term generally refers to a concept in transportation where various forms of transportation services, such as public transit, ride-sharing, bike-sharing, and more, are integrated into a single, accessible platform.

Bicycle sharing is increasingly popular as a sustainable transport system and as a matter of fact, the number of bike-sharing schemes has grown significantly worldwide in recent years[1]. Bike-sharing systems (BSSs) are bridging gaps in public transportation networks that might be insufficient to serve an entire urban area. They are also serving as catalysts for the concept of "sustainable cities" within urban environments. BSSs offer a multitude of benefits for urban mobility, including reducing emissions, improving user health and lifestyle, alleviating traffic congestion, enhancing traffic systems, and integrating seamlessly with public and multimodal transportation [3].

There are two types of bike-sharing systems: dock-based and dockless. Dock-based systems require designated stations for pick-up and drop-off, demanding careful planning and strategic station placement based on demand and usage patterns. Service users should not experience situations in which the do not find available bikes when they need one, or cannot drop off a used bike when they reach a destination. Rebalancing means moving bikes between under- and over-supplied areas, in an attempt to maximise user satisfaction [19, 22]. In contrast, dockless systems offer the flexibility of parking anywhere within the service area. Traditional methods for station placement rely on existing BSS data [4], which may not be available for new systems or those in developing regions. However, also these services need to be implemented carefully: so-called bike-share graveyards[2] represents an example in which bike supply was unnecessarily high, but additional issues are related to blocked paths and irregular parking of these vehicles [5] that decreases the perception of the service quality not just in the users.

Dock-based bike sharing is therefore likely to remain a reasonably adopted approach (and even more so for electric bikes): dock positioning is a key challenge, minimising unbalanced ridership, where some stations lack bikes while others overflow. Ideally, stations would be evenly used, minimizing the need for manual bike redistribution and maximizing user satisfaction [2]. Despite being a crucial element for BSS success, station location selection is complex due to factors like surrounding environment and public transport networks [13].

This paper addresses this challenge by exploring the potential of leveraging already available dockless MaaS data, such as taxi trip records, as a proxy for estimating mobility demand. Based on demand data and plausible assumptions about users' behaviour, what-if scenarios can be easily created to support dock-based BSS system dimensioning and station locations. To evaluate bike-sharing station placement, this approach creates plausible demand for potential bike trips based on real taxi data. Clustering algorithms are then employed to identify zones with high potential based on trips. Reasonable assumptions are then employed to evaluate the adequacy of the decisions about BSS station positioning, again considering demand data based on real world taxi trips records, also comparing with existing stations. The adopted approach is exemplified in the context of NYC but the adopted methods and techniques are of general applicability, and the only requirements are: (i) a dataset of dockless mobility demand for supporting spatial decisions (i.e. trips for which user can reasonably select where to start without moving towards a fixed station), (b) an estimate (or real world data) of the trend of BSS mobility demand in time, especially for supporting the evaluation of BSS system station placement.

This paper is organised as follows: Section 2 provides an overview

* Corresponding Author. Email: giuseppe.vizzari@unimib.it.
[1] In August 2022, there were 1.914 schemes with 8.967.122 bikes, including 194.351 pedal-assisted ones (e-bikes). https://bikesharingworldmap.com/#/all/2.6/0/51.5/

[2] https://www.theguardian.com/uk-news/2017/nov/25/chinas-bike-share-graveyard-a-monument-to-industrys-arrogance

of existing works for optimizing the positioning of BSS stations. Section 3.3 presents an overview of the data used in this paper. The process of synthetic data generation is discussed in Section 4. Section 5.3 provides the experiments on evaluating the positioning of bike-sharing stations and the analysis of the advantages of our method. Finally, Section 7 concludes and draws future research directions.

## 2 Related Work

Regarding the optimization of docking stations's positioning leveraging dockless MaaS data, the study in [20] discusses the optimization of virtual station locations in dockless BSS to meet user demand during morning and evening rush hours. It presents a mixed-integer linear programming (MILP) model and a clustering algorithm to maximize user demand and to solve the parking disorder caused by the systems. The model considers various parameters such as the number of shared bikes at virtual stations, the distance between adjacent virtual stations, and the capacity of each virtual station. It also ensures that virtual stations are mutually supportive within a certain distance, forming a mesh structure. The MILP model aims to find the optimal design scheme for virtual stations under the condition of maximizing user demand. The proposed method is shown to be superior with respect to existing works in terms of computational efficiency.

[10] presents a novel approach to predicting the usage and distribution of bikes in a dockless BSS using journey data. The study focuses on the city of Nanjing, China, and proposes a multi-input multi-output (MIMO) prediction model with the Random Forest (RF) approach to accurately forecast the usage and bike distribution of dockless BSS. The paper also introduces a forecasting framework based on the usage gap, which outperforms traditional departure and arrival-based predictions. Similarly to our work, this paper aims to use historical data of dockless MaaS to evaluate the locations of the stations of a dock-based BSS.

Optimization of the location of bicycle stations in urban areas of Malaga, Spain, is studied in [6]. The authors employed various metaheuristic algorithms such as Genetic Algorithm, Particle Swarm Optimization, etc., which were configured using irace package [14] to automatically optimize their parameters. Additionally, the paper aimed to identify models that best cater to the current needs of users. The input for such models included data from population distribution, city maps, geographic station locations, and citizen usage patterns of the current bicycle-sharing system. The study incorporated real-distance data to determine optimal station locations and considered the number of inhabitants per neighborhood to align the results with citizen needs. By comparing the results with an expert's assignment from the city council, the study demonstrated improvements of up to 50% in quality when applying metaheuristic techniques.

The effect of price change and travel behavior is studied in [12], highlighting how when stations are at full capacity users feel discomfort because they can not safely park the bike, while if they are empty users have to walk an additional distance in order to find a usable bike. Also, their results show that casual users and members have different travel patterns and price preferences. Those findings provide valuable insights for bikesharing providers in understanding their user base and tailoring their services to meet their needs.

The demand for transport and mobility information in the Bio-Bio Region of Chile is presented in [18]. The approach utilizes geospatial data wrangling and a binary integer mathematical programming model to identify locations that maximize system coverage based on realistic travel demand data. Implemented in Python using Geopandas and LocalSolver libraries, the system is designed to address peak hour demand and improve overall availability by over 37%. Additionally, the model incorporates user participation in bike relocation, potentially eliminating up to 80% of the CO2 emissions associated with the rebalancing process.

Unlike prior works that rely solely on dockless bike-sharing data or traditional optimization techniques, our approach leverages taxi trip data, which essentially mimics a dockless system, to assess dock-based Citi Bike station locations. This allows us to capture a wider range of user travel patterns beyond the immediate vicinity of stations. Furthermore, we address the limitations of directly applying synthetic data generation by creating a novel partially synthetic method. We strategically combine elements from both taxi and real Citi Bike trip data, ensuring realistic locations, timings, and spatial distribution across Manhattan. This overcomes issues like unrealistic locations and uneven spatial distribution observed with traditional synthetic data generation methods.

## 3 Data Profiling

This section provides an overview of the data used for this paper.

### 3.1 Data Sources

This paper utilizes two publicly available datasets:

- NYC yellow taxi trip data[3]: This dataset contains historical trip information for the iconic NYC yellow cabs. These taxis operate exclusively through street hails, meaning pickups are not pre-arranged but flagged down on the street. The number of yellow cabs is limited by a finite number of medallions issued by the Taxi & Limousine Commission (TLC), but the service is generally considered safe, very affordable, and relatively widespread. We employed several months of data from this service from 2015 and 2016 (afterwards the publicly shared data was much less granular – zone maps instead of GPS locations), as a proxy for mobility demand, especially concerning the spatial dimension (in general one waits for an available taxi where the trip should start).
- Citi Bike trip history data[4]: This dataset provides historical trip records for Citi Bike, a dock-based bike-sharing system serving NYC with a very good coverage. We particularly employed this dataset as an indicator of the amount and timing of BSS demand, in particular for sake of evaluation of the BSS station location choice.

Due to the inherent differences between yellow taxis and dockless bike-sharing systems, a preliminary analysis is necessary to assess the suitability of using taxi data as a proxy for dockless BSS data.

### 3.2 Data Preprocessing

Both datasets provide trip origin and destination points (in WGS84 coordinates) and timestamps. However, Citi Bike data only provides station locations, not trip distances. Location data is transformed from WGS84 (EPSG:4326) to UTM zone 18N (EPSG:3261822) for simplified calculations while maintaining distance precision (all distances are provided in meters).

We restrict our analysis to trips within Manhattan Island in March 2016 (the last month for which GPS data about trips is available) for the yellow taxis data, and July 2022 for Citi Bike (the last month

[3] NYC TLC Trip Record Data. URL: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.
[4] Citi Bike System Data. url: https://citibikenyc.com/system-data

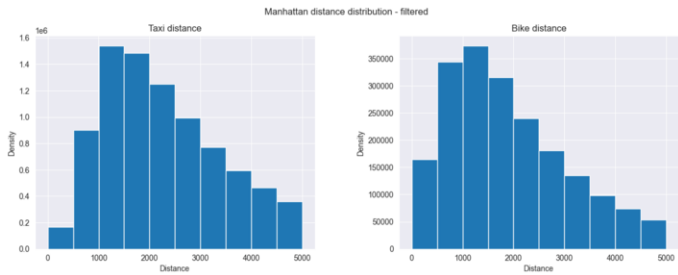


**Figure 1**: Distribution of Trip Distances in Manhattan (Bikes vs. Taxis, Max 5000$m$) - notice scientific notation in taxi demand distribution

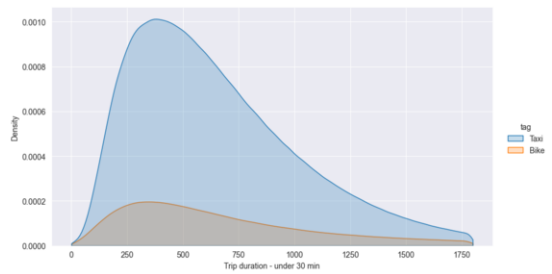


**Figure 2**: Manhattan Trips: Taxi vs. Bike (30 min duration)

available when the present study started, and the month with higher number of travels in the available time frame). This initial filtering yields 10,121,415 taxi trips and 2,294,969 bike trips. To focus on meaningful journeys, we further exclude circular trips and those below a certain distance (details provided in Section 3.3). This additional filtering refines our data to 8,532,937 taxi trips and 1,977,808 bike trips.

### *3.3 Data Analysis*

Trip patterns differ by season and time. Taxis are more popular in colder months and at night due to comfort and safety concerns. Bike usage increases in summer. Weekday trips for both taxis and bikes show morning and evening peaks, likely reflecting commute patterns. Weekends see a shift towards afternoon trips, likely for leisure activities. A nighttime peak, especially prominent in taxi data, suggests trips home after evenings out. Both taxi and bike trips concentrate in Manhattan's southern half (both origin and destination points), and they are more sparse in northern areas, with a slight midtown density boost near Central Park. Weekdays and weekends show no significant spatial distribution difference. Due to trip volume and density, identifying additional, and more fine grained, trends from a map view is challenging. However, zooming in reveals specific patterns, such as reversed trends between morning and afternoon demand peaks (e.g respectively starting and ending nearby busy hubs such as certain subway stations). In general, we observed patterns that are in tune with current literature on spatial-temporal analysis of demand patterns [16].

Trip duration and distance are also investigated: the Manhattan distance (city-block or taxicab distance) is employed as the metric to measure and compare trip distances [7]. We excluded circular trips (origin and destination identical) as they provide minimal value and can hinder analysis: they were more frequent in the bike data (over 5%), likely due to false starts or re-docking attempts, as most circular bike trips were very short (under 1 minute). Given the rarity of long bike trips, we retained trips under 5,000 meters (Manhattan distance). This approach preserved around 90% of bike trips and 85% of taxi trips, resulting in more balanced and realistic distributions. (Figure 1).

The results align with expectations: taxi trips take on average 12 minutes, while bikes take nearly 14 minutes. This difference can be attributed to the inherent speed advantage of cars. Additionally, unlike taxi passengers, cyclists may make stops during their journeys, further contributing to the discrepancy between trip distance and duration, particularly for bicycles (Figure 2).

## 4 Synthetic Data Generation

Within this work we explore the use of a dockless dataset describing mobility demand (yellow taxi data, in the performed experiments) to perform analyses aimed at supporting decisions on numbers and location of dock-based BSS stations. To support this work we also have available time, starting, and destination stations from travels related to Citi Bike BSS.

We first of all construct a synthetic BSS demand for a typical (working) day: about 54 thousand trips. We generate a set of synthetic trips based on the actual temporal distribution of Citi Bike trips, as shown in Figure 3. In particular, we generate a synthetic starting time for the trips extracted, and we need to spatially position the trip, so we generate start and end locations and an appropriate arrival time. To do this, we look within the taxi dataset for actual trips whose starting time is closest to the one of the extracted trip. Preliminary tests with this extremely simple stochastic sampling method showed that the uneven spatial distribution of trips within both of the datasets would imply an extremely low number of travels in the northern part of the map, while the vast majority of trips would concentrate in southern Manhattan. To avoid this problem, we divided Manhattan into three zones: *top* (where bike trips outnumber taxi trips, and both are much scarcer than the rest of Manhattan), *mid* (densities are more balanced between taxi and bike trips, and travel density is medium), and *bot* (taxi trips are more frequent compared to bike trips, and mobility demand is very high compared to top and mid areas). The algorithm we devised therefore selects the starting and destination locations from the taxi dataset both considering the starting time, but also preserving a balance between areas, granting a reasonable number of travels in the top and mid zones, to ensures a more realistic spatial distribution for the synthetic data, reflecting the actual usage patterns of dockless bikes across different areas of Manhattan. Since taxi trips represent actual locations on accessible roads, our synthetic trips originate and terminate at realistic spots within Manhattan.

The synthetic data shows a slightly lower number of very short trips, and correspondingly an increase in longer trips compared to the real data (Figure 4). Despite these discrepancies, the overall distributions remain remarkably similar.

To estimate synthetic trip durations, we analyzed speeds from the real bike data. The arrival time is calculated using a Weibull distribution [17] fitted to these speeds (Figure 5). This approach helps us account for abnormally long trips likely caused by pauses during the bike ride, such as stopping for a coffee or waiting at a red light. While perfectly simulating these pauses is difficult, using a low speed in the Weibull distribution provides a close approximation.

The average extracted speed is 3.4 m/s (12 km/h), which is lower than typical cycling speeds. This could be due to factors like traffic congestion or underestimation of actual trip distances in the taxi data. To further evaluate the quality of the synthetic data, a quality report was obtained using the Synthetic Data Metrics library (sdm)[5].

The report reveals a high degree of similarity between the distributions of real and synthetic data for various parameters (92.19%).

[5] Synthetic Data Metric Library. URL: https://docs.sdv.dev/sdmetrics

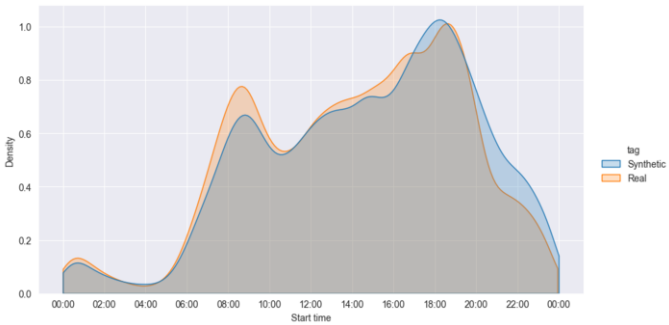


**Figure 3**: Real vs. Synthetic Trips Throughout the Day

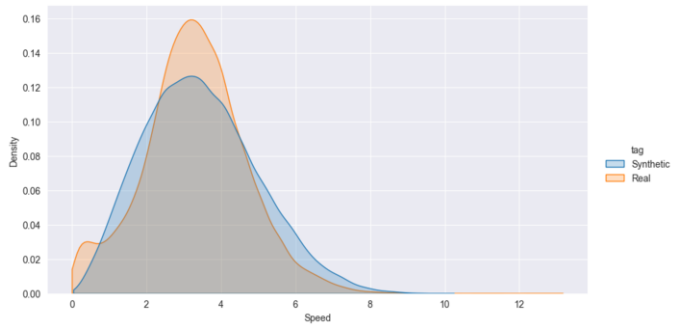


**Figure 5**: Real vs. Synthetic Trips Speed Distributions

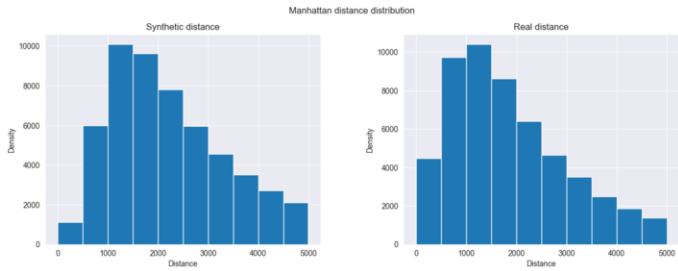


**Figure 4**: Real vs. Synthetic Trips Distance Distributions

Data preparation, including optional steps like time-period selection, can be the most time-consuming part. However, saving intermediate datasets can accelerate this process. Once the data is prepared, the core algorithm execution takes less than a minute.

This partially synthetic data generation method provides a fast and effective way to create data with a realistic spatial distribution and temporal patterns, overcoming the limitations encountered with pure synthetic data generation.

We initially also experimented additional techniques offered by the Synthetic Data Vault (SDV) library[6], experimenting with its synthesizers (statistics and machine learning based functions for generating synthetic data). Two approaches were tested: (i) Gaussian Copula Synthesizer, which uses classical ML technique to learn from real data to generate synthetic data, and (ii) CTGAN Synthetizer, which uses GAN-based approach that promises high-fidelity synthetic data with sufficient training time. Unfortunately, both methods yielded unsatisfactory results. Several issues were apparent:

- Unrealistic locations: Many generated points fell outside Manhattan (and particularly in the Hudson or East rivers) or on inaccessible roads, rendering them unusable.
- Central Park concentration: An unrealistic abundance of starting/ending points appeared in Central Park.
- Southern Manhattan bias: The uneven distribution of points would basically exclude northern Manhattan, that would be essentially not represented in the synthetic dataset.

While the GAN-based technique offered slightly better results, its training time was excessively long.

## 5 Evaluation

To identify proposed BSS station locations, we employed a clustering algorithm (e.g., K-means [8], OPTICS [1]) to strategically group the synthetic mobility demand data. These groupings, termed "virtual station zones," represent areas with a high potential for successful bike-sharing station deployment. Subsequently, these virtual stations are validated by comparing them to the locations of existing stations. This comparison serves a two-fold purpose: firstly, it verifies if the virtual stations align with areas exhibiting strong user demand based on existing bike trips. Secondly, it allows us to identify potential discrepancies between the virtual and physical stations, informing future refinements and optimizations.

**Virtual Station Identification:** Virtual station locations are obtained by applying a clustering algorithm to the starting points of taxi trips. Three algorithms were tested: K-means, OPTICS, and HDBSCAN [23]. The tests were conducted on two subsets of the dataset: one containing trips from 7am to 10am, and another containing trips from 5pm to 8pm, the time frames associated to peak number of travels. K-means was ultimately chosen as the most suitable option due to computational complexity limitations. Running OPTICS and HDBSCAN on the entire dataset proved challenging due to memory constraints. Additionally, K-means offered considerably faster execution times. For instance, the longest K-means execution took 17 seconds, whereas the shortest OPTICS execution took 278 minutes. Another advantage of K-means is that it does not require parameter tuning, unlike OPTICS and HDBSCAN, whose results can vary significantly depending on the chosen parameters. Furthermore, K-means is the only algorithm among the three that allows for the extraction of a fixed number of clusters. It also intrinsically computes the centroids of the clusters, while for OPTICS and HDBSCAN this would require an additional step. Finally, OPTICS and HDBSCAN encountered difficulties in positioning clusters in the northern area of Manhattan, which was not a significant issue for K-means, as illustrated in Figure 6.

**Matching Trips with Nearest Stations in Bike Sharing Systems:** Assigning generated synthetic trips to the nearest bike-sharing station is critical for evaluating station positioning. While seemingly straightforward, optimizing this task becomes crucial when dealing with large datasets. To achieve fast execution times, we leverage SciPy's KDTree class[7]. Stations are organized into a k-dimensional tree, enabling efficient nearest-neighbor lookups. When querying the tree, the method returns both the index of each neighboring station and the Euclidean distance between each queried point and its closest neighbor. This information is then used to join the datasets by associating station details with trip information. Since origin and destination stations require separate analysis, we perform the procedure twice. Running this algorithm on a day's worth of trips (around 54 thousand) and the 648 Citi Bike stations takes only 1 second.

[6] Synthetic Data Vault. URL: https://sdv.dev/

[7] SciPy KDTree. url: https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.html

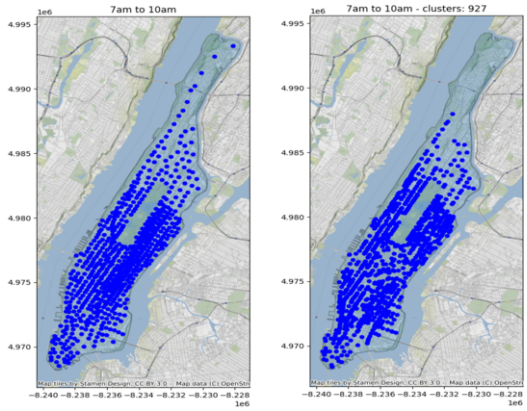


**Figure 6**: K-means clustering tests - k = 648 - time: 17$s$ (left) vs. OPTICS clustering tests - time: 3394$s$ (right)

## 5.1 Evaluation criteria

This section outlines the evaluation metrics used to assess the effectiveness of the proposed virtual station in comparison to existing bike-sharing stations. These metrics will provide insights into factors such as station usage, trip distance, and overall service coverage.

*Station Departures/Arrivals*: These metrics represent the number of trips originating from and ending at each station, respectively. A desirable scenario involves a somewhat correlated number of departures and arrivals, indicating a healthy exchange of bikes. A significant disparity between departures and arrivals would suggest that a station is either running out of available bikes or becoming overloaded.

*Station Usage*: This metric sums the two previously defined metrics, representing the total number of trips originating from or ending at each station. Ideally, every station should have positive usage (with bikes being both picked up and dropped off); a lack of usage could indicate poor station placement. A low average usage might suggest an excessive number of stations, while very high outliers represent overused stations. The number of stations that are never selected is also tracked.

*Distance*: This refers to the Euclidean distance between the trip's origin and destination points and the closest operational bike-sharing stations. While a low average distance is desirable for user convenience, it should not come at the expense of inefficient station distribution. A very low average distance could indicate an overconcentration of stations in a specific area. This creates two problems. First, stations placed too close together become inefficient due to redundancy. Second, other areas of the city might be left underserved, resulting in sporadically high distances for some trips.

*Service coverage*: It refers to the percentage of Manhattan Island accessible by bike within a reasonable distance of a virtual station. We define a location as "served" if it is within 400 meters (quarter-mile) of a station, 5 minutes walk, which is in line with other related work [15]. This distance balances convenience with practicality: a 1.6-kilometer (one-mile) walking radius would render the system impractical for many short trips. It is important to consider some nuances when analyzing coverage. Central Park's vast size might require a higher service radius to achieve full accessibility within the park itself. A high overall coverage indicates a well-distributed network with minimal underserved areas. However, consider that the 400-meter radius represents a sort of worst-case scenario for accessing stations, since they can be much closer one from another if the demand in the area is high.

All of the quantitative evaluations we are going to describe employ data from April 2015 for generating a typical work day starting from

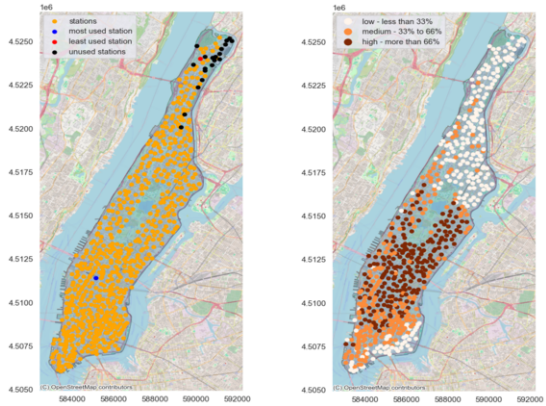


**Figure 7**: Station usage - Citi Bike stations

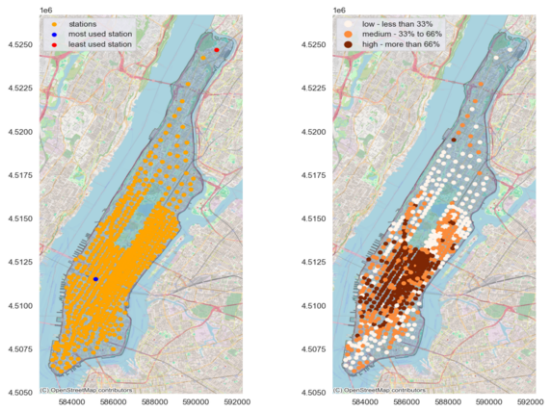


**Figure 8**: Station usage - Virtual stations (648)

the actual usage of the yellow taxi system, for extracting spatial aspects of the demand, and of the Citi Bike BSS system, for the temporal distribution of travels during the day.

## 5.2 Evaluation of the use of existing Citi Bike stations

The evaluation uses the current placement of the actual 648 Citi Bike stations in Manhattan as the benchmark for assessing alternative configurations of stations locations. It is important to acknowledge that Citi Bike's service area extends beyond Manhattan, encompassing other boroughs and nearby cities. While some outlying stations might not be as relevant for this specific analysis, they remain crucial for the overall system's functionality.

***Station usage*:** An interesting aspect of the current Citi Bike station placement is the uneven distribution of daily usage, as shown in

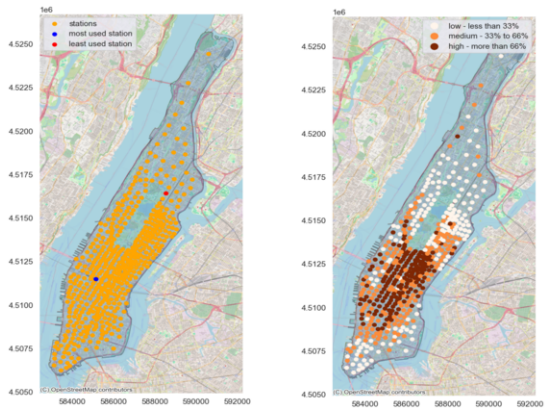


**Figure 9**: Station usage - Virtual stations (500)

| | Citi Bike | | | | | Virtual Stations (648) | | | | | Virtual Stations (500) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Station Usage | | | Distance (in m) | | Station Usage | | | Distance (in m) | | Station Usage | | | Distance (in m) | |
| | Dep. | Arr. | Ov. | Dep. | Arr. | Dep. | Arr | Ov. | Dep. | Arr. | Dep. | Arr. | Ov. | Dep. | Arr. |
| min | 0 | 0 | 0 | 1 | 1 | 3 | 6 | 9 | 1 | 1 | 4 | 15 | 19 | 1 | 1 |
| max | 1085 | 973 | 2058 | 406 | 452 | 440 | 374 | 814 | 850 | 919 | 686 | 469 | 1055 | 973 | 962 |
| mean | - | - | - | 101 | 106 | - | - | - | 61 | 73 | - | - | - | 72 | 84 |
| median | 47 | 52 | 97 | 95 | 101 | 82 | 74 | 148 | 55 | 66 | 91 | 98 | 190 | 66 | 77 |
| skew | 2.77 | 2.55 | 2.68 | 0.51 | 0.57 | 1.78 | 1.69 | 1.63 | 2.81 | 2.79 | 2.11 | 1.71 | 1.88 | 2.64 | 2.62 |
| unused stations | 46 | 31 | 20 | - | - | 0 | 0 | 0 | - | - | 0 | 0 | 0 | - | - |

Dep. (Departures); Arr. (Arrivals); Ov. (Overall)

**Table 1**: Citi Bike vs. Virtual Stations evaluation criteria - usage data is evaluated on a daily basis (April)

Table 1. The median usage sits at a relatively low 97, with a single outlier boasting over 2,000 trips. This highlights a disparity in user demand across different station locations. Further emphasizing this point, is the imbalance between departures and arrivals. A significant number of stations (46) have zero recorded departures, while a smaller number (31) have not seen any arrivals. There are also 20 completely unused stations. Figure 7 shows that stations with high usage are basically clustered in central and southern Manhattan, with the absolute busiest located in Midtown South. Conversely, the northern and southeastern regions see significantly lower usage, with most unused stations concentrated at the very top of the island.

It is worth noting that this distribution could be partially explained by the data only including trips starting and ending within Manhattan. Excluding journeys that traverse boroughs might skew the results towards stations within the island itself. Nevertheless, a clear north-south divide in usage is evident, suggesting potential areas for improvement in station positioning or service offerings. Inevitably, usage metrics are not necessarily in tune with coverage.

**Distance to Stations:** The average Euclidean distance (the distance from the trip's origin and destination points to the selected stations) to a Citi Bike station in Manhattan is approximately 100 meters (Table 1), considered a highly walkable distance. However, the maximum distance is higher than the 400-meter threshold typically used for service coverage calculations. In fact, there are 16 cases where the distance exceeds this limit, with the farthest station reaching 452 meters. While this is a rare event, it highlights that there are potential gaps in service coverage for certain areas.

**Service Coverage:** As expected, the current Citi Bike station placement achieves excellent service coverage, reaching 98% (Figure 10 -left). Nonetheless, as previously said, this comes at the cost of a low (and sometimes even very low) usage of some stations.

### *5.3 Virtual stations evaluation*

This section explores the potential benefits of using virtual stations to optimize Citi Bike's service in Manhattan. The initial analysis employs a configuration with 648 virtual stations, mirroring the current number of physical stations. This experiment was performed as a first what-if analysis trying to understand what would be the implications of just changing the locations, and not the number, of stations in the BSS system. We also considered a hypothetical situation in which only 500 stations were to be positioned, and we estimate the negative implications that the reduction of costs for the creation and maintenance of such infrastructure would imply.

**Station usage patterns:** The usage patterns for virtual stations (648) show a more balanced number of departures from those observed in the existing Citi Bike network (Table 1). Unlike the current situation with numerous unused stations, the virtual configuration eliminates this issue entirely. This placement prioritizes areas with high demand, resulting in a more balanced distribution of service load across stations. The median usage increases to 148 compared to Citi Bike's 97, while the maximum total usage is significantly lower at 814 with respect to 2058 of the overall usage of the current situation. This shift suggests a reduction in heavily overloaded stations and a more equitable distribution of ridership among stations. Figure 8 visually depicts this altered usage distribution. Stations with high usage are more concentrated in areas with high ridership. In contrast, stations near Central Park experience a decrease in usage due to the higher density of stations in that area, leading to a more balanced service offering. Conversely, the northern area now contains a lower number of stations that, however, have a higher usage level, indicating a more cost effective coverage for this region. The situations for the what-if analysis in which just 500 stations were positioned (see the data in Table 1 and the spatial distribution and usage in Figure 9) pushes further the changes already visible in the previous experiment: median station usage is almost double than for the actual Citi Bike infrastructure, station density in the northern area is further (but marginally) reduced, with a slight reduction of the number of stations in situations already served in the upper east side and in the southern east area.

**Distance to stations:** Virtual stations (648) achieve significantly lower average and median distances compared to Citi Bike stations (Table 1). The mean distance drops to 61 meters, with a median value of 54 meters, which is nearly half the distance observed with physical stations. This substantial reduction highlights a noticeable advantage of virtual stations – the ability to strategically position service in areas with high demand, leading to improved accessibility for users. This translates to improved accessibility and potentially increased ridership. The higher skew value further supports this by indicating that the distribution of distances is not symmetrical: the positive skew suggests that more values are in the "lower range" (shorter distances) compared to the "higher range" (longer distances). However, the maximum overall distance observed with virtual stations grows significantly, reaching around 900 meters, which exceeds the acceptable 400-meter limit for service coverage calculations. This growth represents the trade-off inherent in virtual station placement: while accessibility improves in high-demand areas, some sparsely populated areas (or areas with lower demand) might experience a decrease in service coverage. Analogous considerations can be done for 500 virtual stations what-if analysis: while the mean overall distance is still lower than the one associated to the Citi Bike infrastructure, growing marginally (about 10 meters) from the 648 virtual stations configuration, the maximum distance grows about 100 meters, without really changing much an already problematic but infrequent set of situations.

**Service coverage:** The virtual station (648) configuration prioritizes service quality in congested areas by sacrificing some overall service coverage. As a result, the service coverage drops to around 88%: while this remains an acceptable level, it raises questions about service equity and accessibility in areas with limited virtual stations. Reducing the number of stations to 500, the service coverage further drops but

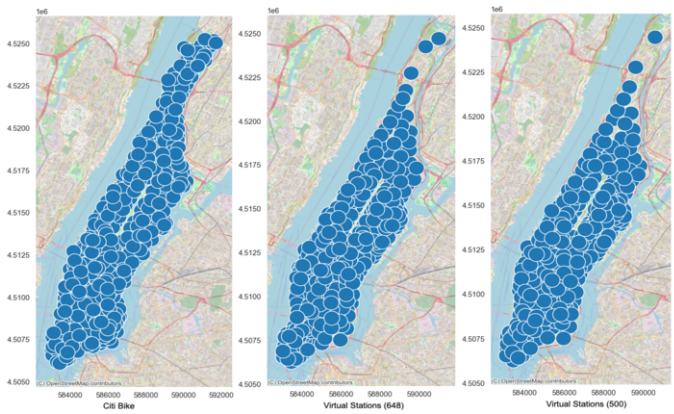


**Figure 10**: Every circle represents the area covered by a station

very marginally, to 85.4%, despite the reduction of almost 23% of the stations. Figure 10 visually compares the coverage of the three configurations.

## 6 Considerations, Limitations, and Lessons Learned

Leveraging multiple datasets, such as taxi trips and bike-sharing records, can provide valuable insights into mobility patterns and demand distribution. In fact, integrating real-world data sources allows for a more comprehensive understanding of urban mobility dynamics, enabling informed decision-making in station placement. The present experiments show how relatively simple techniques can be employed to support what if analyses to explore the implications of choices about the dimensioning and location of BSS stations. The main driver for this kind of work is the *potential for cost reduction*: physical BSS stations have significant expenses not just associated to the initial setup, but also to rebalancing and maintenance [8]. The 500 stations scenario we elaborated was a rather extreme experiment showing that a mostly acceptable service quality can be achieved with a significantly lower number of stations. We do not pretend that this kind of setup would be immediately applicable, but it could represent a serious starting point, produced with a simple automatic procedure, that could be then integrated with a potentially limited budget of additional stations to be tactically located to reduce the issues of underserved areas, reducing maximum travel distance on foot to reach a station. This kind of human intervention would however be necessary, even considering the numerous simplifications that we carried out and the limitations of the proposed approach; to mention a few notable ones (i) we did not make *micro-level considerations about the placement of stations*, and about the practical possibility to set up such an infrastructure on those points; (ii) we assumed an *extremely simple decision making process for service users*, that does not even consider the possibility to opt out due to issues such as bike unavailability or unavailability of space to drop off a bike at the destination station, and we did not consider at all service pricing; (iii) we did not consider *station capacity* and issues related to travels not ending in the planned station due to the fact that it is full (although we expect this approach to reduce the issue, at least keeping the same number of stations); (iv) the evaluation is taken not considering the dynamics of actually setting up the infrastructure, and does not consider competition by other services.

We did not explicitly consider tactical positioning nearby public transportation system hubs, like subway stations, or particular points of attraction, but we expect that the adopted mobility demand data, in particular the taxi dataset, already mostly captures implications of the actual city structure, at least at the time of the data acquisition. This represents an additional limit of the approach: we considered a decision support system in which data describing the spatial side of demand as well as the overall service usage and temporal distribution (or a forecast of them) are available, but we know that a city's infrastructure and citizens' behaviours do change.

The actual value of the present work lies in the exploration of the trade-offs implied by system dimensioning and location decisions, in which sustainability, not just of the overall planet, but also – at a much finer and modest scale – of the provisioning of a BSS service requires balancing principles (equal access of mobility services to citizens) and practical aspects (careful considerations about costs). Offering a capillary service even in areas in which demand is scarce might have an impact on service cost for users that are mostly not interested in having such a high coverage, or it could require increased forms of subsidization by municipalities. From this perspective, this work would require additional analyses in the vein of [9], but this kind of approach and developed simple systems for analyzing data and exploring, also visually, the implications of some choices about the transport and mobility infrastructure can represent a relevant contribution to participatory decision making processes [21].

## 7 Conclusions and Future Work

This paper has presented a data driven method to support decisions about the dimensioning and location of BSS stations. We employed data describing actual trips of a dockless mobility system (yellow taxi data) and an actual BSS system (Citi Bike) to feed a clustering based approach for stations positioning. We exemplified the approach and evaluated it employing data from NYC. We showed that this approach can be used to explore the inherent trade-off between service coverage and cost effectiveness of the BSS infrastructure.

There are different future directions to improve this work. One key area of exploration involves incorporating station size limitations into the analysis. This would allow for a more granular evaluation of how reducing the number of stations actually impacts the system. Beyond station size and rebalancing, there is also the challenge of low-density coverage: how can we improve virtual station distribution in areas with fewer trips? This is a crucial question that future research can address. Another important direction lies in manipulating the historical data itself. Currently, the synthetic dataset generation process is tailored specifically to Manhattan Island. Developing an algorithm that is adaptable to different geographical contexts, and maybe also travel modes, is needed. Finally, the "what-if" scenario potential of the synthetic dataset should not be overlooked: we can easily modify this data to create anticipated real-world situations and test the efficiency of various MaaS systems under those conditions. This would be a powerful tool for future MaaS planning and development.

## Acknowledgements

This work was partly developed within the Spoke 8 — MaaS and and Innovative services of the National Center for Sustainable Mobility (MOST) set up by the "Piano nazionale di ripresa e resilienza (PNRR)" —M4C2, investimento 1.4, "Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S" su alcune Key Enabling Technologies" funded by the European Union. Project code CN00000023, CUP: D93C22000410001. This work was also partially supported by the MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

[8] Here's why it costs $6K$ per Citi Bike bicycle. url: https://www.silive.com/news/2017/04/heres_why_it_costs_6k_per_citi.html

## References


[1] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28 (2):49–60, 1999.

[2] L. Caggiani, R. Camporeale, M. Marinelli, and M. Ottomanelli. User satisfaction based model for resource allocation in bike-sharing systems. *Transport Policy*, 80:117–126, 2019.

[3] L. Caggiani, A. Colovic, and M. Ottomanelli. An equality-based model for bike-sharing stations location in bicycle-public transport multimodal mobility. *Transportation Research Part A: Policy and Practice*, 140: 251–265, 2020.

[4] M. Chen, D. Wang, Y. Sun, E. O. D. Waygood, and W. Yang. A comparison of users' characteristics between station-based bikesharing system and free-floating bikesharing system: Case study in hangzhou, china. *Transportation*, 47:689–704, 2020.

[5] Z. Chen, D. van Lierop, and D. Ettema. Dockless bike-sharing systems: what are the implications? *Transport Reviews*, 40(3):333–353, 2020. ISSN 0144-1647. doi: https://doi.org/10.1080/01441647.2019.1710306. URL https://www.sciencedirect.com/science/article/pii/S0144164722001015.

[6] C. Cintrano, F. Chicano, and E. Alba. Using metaheuristics for the location of bicycle stations. *Expert Systems with Applications*, 161: 113684, 2020.

[7] E. Deza, M. M. Deza, M. M. Deza, and E. Deza. *Encyclopedia of distances*. Springer, 2009.

[8] J. A. Hartigan, M. A. Wong, et al. A k-means clustering algorithm. *Applied statistics*, 28(1):100–108, 1979.

[9] D. Hörcher and D. J. Graham. Maas economics: Should we fight car ownership with subscriptions to alternative modes? *Economics of Transportation*, 22:100167, 2020. ISSN 2212-0122. doi: https://doi.org/10.1016/j.ecotra.2020.100167. URL https://www.sciencedirect.com/science/article/pii/S2212012219300735.

[10] M. Hua, J. Chen, X. Chen, Z. Gan, P. Wang, and D. Zhao. Forecasting usage and bike distribution of dockless bike-sharing using journey data. *IET Intelligent Transport Systems*, 14(12):1647–1656, 2020.

[11] P. Jittrapirom, V. Caiati, A. M. Feneri, S. Ebrahimigharehbaghi, M. J. Alonso-González, and J. Narayan. Mobility as a service: A critical review of definitions, assessments of schemes, and key challenges. *Urban Planning*, 2(2):13–25, 2017.

[12] S. Kaviti, M. M. Venigalla, and K. Lucas. Travel behavior and price preferences of bikesharing members and casual users: A capital bikeshare perspective. *Travel Behaviour and Society*, 15:133–145, 2019.

[13] J. Liu, Q. Li, M. Qu, W. Chen, J. Yang, H. Xiong, H. Zhong, and Y. Fu. Station site optimization in bike sharing systems. In *2015 IEEE International Conference on Data Mining*, pages 883–888. IEEE, 2015.

[14] M. López-Ibáñez, J. Dubois-Lacoste, L. P. Cáceres, M. Birattari, and T. Stützle. The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43–58, 2016.

[15] I. Mateo-Babiano, R. Bean, J. Corcoran, and D. Pojani. How does our natural and built environment affect the use of bicycle sharing? *Transportation research part A: policy and practice*, 94:295–307, 2016.

[16] C. Morton, S. Kelley, F. Monsuur, and T. Hui. A spatial analysis of demand patterns on a bicycle sharing scheme: Evidence from london. *Journal of Transport Geography*, 94:103125, 2021. ISSN 0966-6923. doi: https://doi.org/10.1016/j.jtrangeo.2021.103125. URL https://www.sciencedirect.com/science/article/pii/S0966692321001782.

[17] H. Rinne. *The Weibull distribution: a handbook*. Chapman and Hall/CRC, 2008.

[18] C. Rojas, R. Linfati, R. F. Scherer, and L. Pradenas. Using geopandas for locating virtual stations in a free-floating bike sharing system. *Heliyon*, 9(1), 2023.

[19] K. Saltykova, X. Ma, L. Yao, and H. Kong. Environmental impact assessment of bike-sharing considering the modal shift from public transit. *Transportation Research Part D: Transport and Environment*, 105:103238, 2022.

[20] Z. Sun, Y. Li, Y. Zuo, et al. Optimizing the location of virtual stations in free-floating bike-sharing systems with the user demand during morning and evening rush hours. *Journal of Advanced Transportation*, 2019, 2019.

[21] S. Tori, G. te Boveldt, and I. Keseru. Building scenarios for urban mobility in 2030: The combination of cross-impact balance analysis with participatory stakeholder workshops. *Futures*, 150: 103160, 2023. ISSN 0016-3287. doi: https://doi.org/10.1016/j.futures.2023.103160. URL https://www.sciencedirect.com/science/article/pii/S0016328723000642.

[22] C. M. Vallez, M. Castro, and D. Contreras. Challenges and opportunities in dock-based bike-sharing rebalancing: A systematic review. *Sustainability*, 13(4):1829, 2021.

[23] H. Wickham and H. Wickham. *Data analysis*. Springer, 2016.