Automatic Bike Sharing System Planning from Urban Environment Features

Nicolai A. Weinreich[a,*], Daniel B. van Diepen[a], Federico Chiariotti[b,c], Christophe Biscio[a]

[a] *Department of Mathematical Sciences, Aalborg University, Skjernvej 4A, 9220 Aalborg Øst, Denmark*
[b] *Department of Electronic Systems, Aalborg University, Fredrik Bajers Vej 7C1, 9220 Aalborg Øst, Denmark*
[c] *Department of Information Engineering, University of Padova, Via G. Gradenigo 6B, 35131 Padova, Italy*

Abstract

The planning process for bike sharing systems is often complex, involving multiple stakeholders and several considerations: finding hotspots in the potential demand, and dimensioning the system, requires an intimate knowledge of urban mobility patterns and specific local features of the city. The significant costs associated with dynamic rebalancing of bike sharing systems, i.e., with moving bikes across the city to correct the demand imbalance and ensure that they are available where and when they are needed, make correct planning even more critical for the economic viability of the system. In this work, we consider urban environment data from multiple sources and different cities in Europe and the United States to design an automated planning pipeline to place stations in an area with no direct knowledge of the demand. The first step in the planning is to build models of activity patterns and correlate them with features of the urban environment such as land use and mass transit availability; these statistical models can then be used to expand an existing network or even create an entirely new one in a different city. A use case in New York City shows that our system can effectively plan a bike sharing system expansion, providing a valuable first step for the planning process and allowing system designers to identify gaps in existing systems and the locations of potential demand hotspots.

Keywords: Bike sharing, Urban planning, Traffic prediction
2022 MSC: 00-01, 99-00

*Corresponding author
Email addresses: nicolai.weinreich@gmail.com (Nicolai A. Weinreich), daniel@vandiepen.dk (Daniel B. van Diepen), chiariot@dei.unipd.it (Federico Chiariotti), christophe@math.aau.dk (Christophe Biscio)

1. Introduction

A few years ago, the United Nations projected that 6.7 billion people will be living in urban areas in 2050, compared to 4.2 billion in 2018[1]. One of the main challenges facing urban planners and city governments is to accommodate this growth and reconcile it with sustainability goals and climate change mitigation. Macfarlane (2019) argues that private cars, which already contribute disproportionately to environmental problems such as noise and air pollution, must give way to greener, more socially equitable alternatives for urban mobility, encouraging both mass transit and public bike sharing Shaheen et al. (2010) in an extremely urbanized, zero-carbon future. From the late 2000s and onward, advancements in information technology have made implementing and managing bike sharing systems more feasible and cost-efficient. This has led to a rapid expansion in the amount of bike share systems worldwide: DeMaio et al. (2021) show that around 2000 bike sharing systems are currently in operation, with a global fleet size of more than 10 million bikes as of August 2021.

From the perspective of city planners, bike share systems pose an interesting solution to the objectives of reducing emissions, decreasing congestion and increasing the mobility of city residents, all while dealing with the constraints of a limited budget. A bike sharing system meets all of those requirements while being orders of magnitude cheaper for the city to implement than other mobility solutions. In some cases, cities contribute little more than the public land for parking of bicycles, as Daddio (2012) argues. On the other hand, from the users' perspective, a well-functioning bike sharing system should work well in tandem with other modes of public transport, and convenience is a crucial factor for increasing ridership and convincing people not to drive. Bike sharing is convenient when it manages to effectively cover distances which are too far to walk and too close to warrant other transportation options. One frequent use case for bike sharing is to cover the distance between the origin of the trip and the closest subway station or bus stop, as shown by Ma et al. (2018), sometimes referred to as the *first mile* problem. Conversely, Zhang et al. (2019) show how bike sharing can also serve as a *last mile* solution to cover the part of a trip from the closest public transport station to the final destination. In general, integration with mass transit is critical, as Martin and Shaheen (2014) showed how a poorly planned bike sharing system might end up competing directly with walking and mass transit instead of reducing car usage. However, if bike sharing is seen not in isolation but as one part of the urban mobility arsenal, it can help increase the coverage and flexibility of public transport, making mixed mode commuting a valid mobility solution for more people.

However, planning bike sharing systems is a complex and difficult problem: rebalancing, i.e., moving bikes to high-demand areas before peak usage times to guarantee that the demand will be met, is both critical for user convenience

[1] United Nations, *68% of the world population projected to live in urban areas by 2050, says UN*, `https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html`

and system reliability and the biggest operational expense for system managers. Finding the best locations for bike sharing stations (or dedicated parking spots in free-floating systems), and dimensioning the number of stalls and bikes for each station, can reduce the need for rebalancing and, consequently, operational costs, but requires *a priori* knowledge of spatio-temporal demand patterns.

In this work, we aim at putting together traffic pattern analysis, urban features, and prediction models, in order to build a pipeline for automated planning, either to expand an existing system to new areas or to build a bike sharing system from scratch in an entirely new city. Our framework outputs the complete bike sharing system design on a neighborhood or city-wide scale, and while refinements and community meetings are needed, automating the design process for the proposal to present to citizens and municipal authorities can save money and time. The contribution in this work is the following:

- We define an optimization pipeline to evaluate possible bike sharing station locations in an area with a manageable computational cost. Our automated process improves the planning process by analyzing traffic patterns and building a prediction model to estimate the suitability of a crossing or street side location for a bike sharing station. Additionally, we optimize each station's required capacity, based not only on total demand but also on the daily flow patterns of commuter traffic;

- We compare traffic patterns in several major cities across the United States and Europe, discussing commonalities and regional differences, and their impacts when using the automated planning pipeline for an entirely new bike sharing system;

- We present a use case for our optimization model based on the 2019 expansion of the New York City CitiBike system, showing the differences between the output of our automatic model and the actually installed stations.

All the analysis and optimization code is publicly available on GitHub [2], with a full description of the necessary steps to repeat our results and apply our framework to a different city. The repository also includes an interactive dashboard that can be used to dynamically interact with data from any of the cities in the dataset The data used in this paper include both bike sharing trip data and general information about the built environment (land use, census, and subway station location data), and have been obtained from a variety of open data sources listed in Table 1. The data and their processing are discussed in detail in the Appendix.

The rest of the paper is divided as follows: first, we present a review of the relevant literature in Sec. 2. Then, Sec. 3 presents our automated station planning method, which is tested on the New York City CitiBike system in

[2] `https://github.com/cykelholdet/superbike`

Dataset	Area	Provider	Source link
Trip Data	New York City	Citi Bike	https://ride.citibikenyc.com/system-data
Trip Data	Chicago	Divvy Bikes	https://ride.divvybikes.com/system-data
Trip Data	Washington D.C.	Capital Bikeshare	https://www.capitalbikeshare.com/system-data
Trip Data	Boston	Bluebikes	https://www.bluebikes.com/system-data
Trip Data	London	Transport for London	https://cycling.data.tfl.gov.uk/
Trip Data	Helsinki	Helsinki Region Transport	https://hri.fi/data/en_GB/dataset/helsingin-ja-espoon-kaupunkipyorilla-ajatut-matkat
Trip Data	Oslo	Oslo City Bike	https://oslobysykkel.no/en/open-data/historical
Trip Data	Madrid	BiciMad	https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)
Station Data	Chicago	City of Chicago	https://data.cityofchicago.org/Transportation/Divvy-Bicycle-Stations-All-Map/bk89-9dk7
Station Data	Washingon D.C.	Dept. of Real Estate Services	https://opendata.dc.gov/datasets/DCGIS::capital-bike-share-locations/about
Station Data	London	Transport for London	https://api.tfl.gov.uk/
Station Data	Madrid	BiciMad	https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)
Land Use Data	New York City	NYC Dept. of City Planning	https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-gis-zoning.page
Land Use Data	Chicago	City of Chicago	https://data.cityofchicago.org/Community-Economic-Development/Boundaries-Zoning-Districts-current-/7cve-jgbp
Land Use Data	Washington D.C.	District of Columbia	https://opendata.dc.gov/datasets/DCGIS::zoning-regulations-of-2016/about
Land Use Data	Boston	Boston Planning and Development Agency	https://data.boston.gov/dataset/zoning-subdistricts1
Land Use Data	Europe	European Enviroment Agency	https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018
Census Data	US	US Census Bureau	https://www.census.gov/data.html
Census Data	Europe	European Enviroment Agency	https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018
Transit Data	All cities	OpenStreetMap	http://overpass-api.de/

Table 1: Data sources.

Sec. 4. The benefits and issues of using data from different cities in the planning are presented in Sec. 5, and the paper is concluded with our final remarks and some suggestion for future work in Sec. 6.

2. Literature Review

The widespread adoption of bike sharing and micro-mobility in cities all over the world in the past decade has spurred a significant research effort on analyzing demand patterns and trying to optimize the management of these systems. After the COVID-19 pandemic, Hensher (2020) and others argue that changes in urban mobility have accelerated, paving the way for approaches that include micro-mobility as a crucial component, as it represents an eco-friendly and socially distanced mode of transportation, and can complement mass transit in normal times, as discussed by Saltykova et al. (2022), and providing a much needed last mile service, as shown by Zhang et al. (2019).

In the context of this work, we will focus our literature review on two topics: the first is the prediction of demand patterns in bike sharing systems and its

relation to the built environment and to other variables such as transit stations, while the second is bike sharing system planning and its automation. For a more complete review of the large body of work on bike sharing systems, we refer the reader to Eren and Uz (2020) and Albuquerque et al. (2021).

2.1. Pattern Identification and Prediction

Pattern identification in bike sharing systems is often performed by clustering, i.e., by dividing stations into groups based on spatial or temporal characteristics of the demand. Purely spatial clustering aims at the definition of *neighborhoods* in the bike sharing graph, showing areas with high internal connectivity, i.e., many trips within the neighborhood. On the other hand, spatio-temporal clustering aims at finding stations with similar patterns in the variation of their hourly demand, distinguishing, e.g., groups of stations with a very high demand for bicycles in the morning, or receive a high influx of bikes in the afternoon.

The main objective of spatial clustering is to define cycling neighborhoods, which can then be used to determine the type of mobility enabled by the system: Lee et al. (2021) argue that neighborhoods clustered around mass transit hubs and that consist mostly of shorter trips are consistent with the use of the system as last mile coverage for multimodal trips which also involve public transit, while longer trips across different neighborhoods might indicate a purely cycling commute. Clustering, as well as other graph-based metrics, can be used to improve short-term flow prediction, considering recent demand and common patterns inside and between different neighborhoods, as done by Yang et al. (2020b). The same type of analysis can also be applied, as Zhang et al. (2021) did, to dockless bike sharing systems, determining the mobility between different areas and using local demand clusters as starting point to build the system-wide graph.

Spatio-temporal clustering has been applied on trip data from several cities, often with common results: an analysis in Vienna by Vogel and Mattfeld (2011) used 5 clusters, relating the daily activity patterns to likely user profiles and distinguishing between stations used mostly for leisure and by tourists and stations used by commuters for their daily trips to work. Similar results were found for the Chicago bike sharing system by Zhou (2015), although two of the 5 clusters were characterized by extremely low usage.

The opposite approach to the definition of clusters can be adopted: instead of clustering based on the demand patterns and analyzing the correlations with land use and nearby public transit stations, Côme and Oukhellou (2014) divided the Paris system into clusters by considering the land use features, comparing the resulting patterns for each cluster. While this approach is not an instance of spatio-temporal clustering, as it only uses spatial information to arrive at the cluster definition, the resulting analysis is similar, showing strong differences in the patterns for residential and commercial areas.

It is also possible to predict bike sharing traffic at individual stations, as Yang et al. (2016) did, or for geographical and spatio-temporal clusters of stations, as Li et al. (2015) considered. Li and Zheng (2019) then defined a coarser-grained prediction on the cluster level and a finer-grained one for individual

stations. Chen et al. (2016) analyzed the correlation between close-by clusters to help predict spikes in demand, and graph information in general can be a powerful tool to predict future behavior, as shown by Yang et al. (2020b,a). Hulot et al. (2018) argue that average behavior might not be enough to provide full service availability even in worst-case scenarios of high and unbalanced demand, and Sohrabi et al. (2020) show how risk and extreme value theory can be used to further improve worst-case performance.

However, all these prediction and clustering works assume full knowledge of past demand, which is impossible in the planning phase of the system: in order to properly plan the system, it is crucial to find a way to estimate the potential demand at a given location *without* any direct data. Fortunately, bike sharing traffic patterns are intimately linked with the geographic and social characteristic of the urban environment, such as land use, population density, social and economic inequality, and road infrastructure, and several works in the literature have investigated this relation Osama et al. (2017) show that bike-friendly road infrastructure such as separated bike lanes and flatter, direct routes can improve bike sharing usage, as does the presence of recreational areas. However, commuting accounts for a large percentage of bike sharing trips, as Cervero et al. (2019) found, and Wu et al. (2019) showed that workplaces such as universities, hospitals, or large commercial complexes can also play a role in attracting bike sharing traffic. Finally, the analysis by Araghi et al. (2022) shows how different types of points of interest can attract different populations of users, often with unique usage patterns and travel mode preferences.

Public transit is another key factor in the success of bike sharing systems: Radzimski and Dzięcielski (2021) show that stations close to bus stops or subway stations often have significantly more traffic, an indication of multimodal commuting and the aforementioned use of bike sharing as a first or last mile service in a longer trip. Interestingly, Martin and Shaheen (2014) argue that dense transit networks have synergies with bike sharing in the first mile access, as the city center is well-served by mass transit options and bike sharing only competes with them in that area, while cities with sparser networks can exploit bike sharing both at the first and last mile, shifting larger modal shares from driving to multimodal trips using bike sharing and mass transit. Several other recent works consider different urban, geographic, and socioeconomic features and their value in predicting overall bike sharing demand; for a more thorough overview of the literature on bike sharing traffic patterns and on the factors affecting usage and user behavior, we refer the reader to the survey by Elmashhara et al. (2022).

Predictive models that can infer future demand from basic features of the urban environment play a key role in planning, providing information on where to place new stations when expanding an existing system or designing a new one, as done by Eren and Katanalp (2022). Perhaps the most similar works to our own are the ones by Noland et al. (2016) and Hyland et al. (2018): in the former, Bayesian regression is used to estimate the effect of land use, population, and bicycle infrastructure, with limited success over future patterns, while in the latter, clustering is used to divide stations into 3 classes, over which the

prediction is performed. However, both works are limited to a single city, and therefore not directly applicable to planning in an entirely new city, and the patterns and features used to determine the station classification are hand-designed, making the effectiveness of the prediction highly dependent on design parameters.

2.2. Bike Sharing System Planning

The literature on automated planning is extremely limited, and often significantly aided by human oversight and direct control over the variables: the novelty of our planning framework is that it limits the number of human-set parameters in the design, using bike sharing data from existing system areas (in the case of the expansion of an existing system) or from other similar cities (in the case of the design of an entirely new system). For a more thorough review of the literature on the subject, we refer the reader to Shui and Szeto (2020).

One of the first works to propose automated planning, by Vogel and Mattfeld (2011), considered an approach similar to ours in dividing stations into clusters, but did not address the problem of placing new stations, limiting itself to the operation of already existing systems. A more recent work Strauba et al. (2018) also focuses on Vienna, considering point of interest data and public transit access and optimizing the location of station based on these factors: however, it does not include any past data from bike sharing systems, relying only on indirect indicators and not considering the different effects they may have on different cities. The same approach was considered by Kleisarchaki et al. (2022) and García-Palomares et al. (2012), who also included land use data in their model, but still did not directly validate their statistical models on bike sharing demand data.

A recent optimization work by Çelebi et al. (2018) considers an optimization of station placement, but it assumes that the demand at each given point in the map is already known. Frade and Ribeiro (2015) took a similar approach, starting from an area-level survey of potential demand to perform the optimization, and Garcia-Gutierrez et al. (2014) used a traffic department level of service study. Obtaining reliable demand data is, naturally, a major issue, as the only way to determine demand exactly is to have a bike sharing system in place, and the system itself will alter demand. On the other hand, Larsen et al. (2013) focused on the system itself, identifying missing connections and underserved paths by looking at bike lane positions and accident data.

In general, existing works have looked at several sources of data to try and plan a bike sharing system, but the direct use of data from existing systems is still mostly unexplored as a potential avenue for planning. Our work aims to fill that gap in the literature, providing a planning pipeline that is grounded in existing bike sharing data, as well as urban environment features.

3. Automated Station Planning

In this section, we will present our automated station planning system, which is the main focus of our work. Our automated approach to bike sharing system

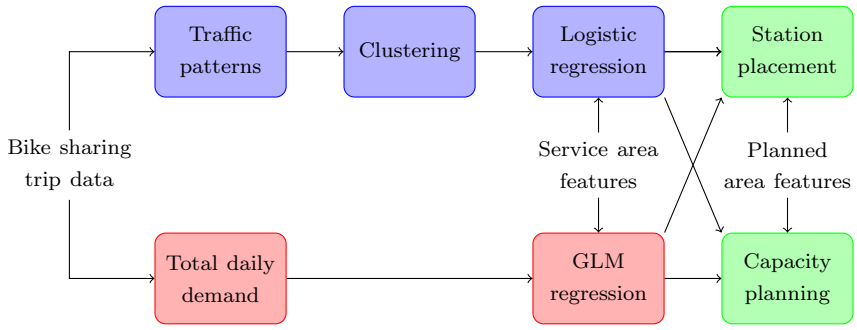


Figure 1: Flowchart of the demand modeling and prediction approach.

planning is divided in two main phases: firstly, we need to establish a demand model for the expansion area or the new bike sharing system, and secondly, we need to place and dimension stations in the area.

Naturally, measuring bike sharing system demand directly is impossible, as it would require having an existing bike sharing system in place before starting to plan: however, it is possible to connect traffic patterns with other features of the urban environment, building a predictive model that can be generalized to other neighborhoods and cities. In the case of the expansion of an existing system, patterns from areas in which the system is up and running can be used to derive the model; however, we will show that patterns across different cities share significant similarities, allowing designers of entirely new systems to draw from existing data.

An overview of the approach we consider in the first step of the station planning process can be seen in Fig. 1. The data obtained from the bike sharing system are used to determine a traffic pattern vector for each station. Stations are clustered into a predetermined number of classes, which represent different types of traffic patterns in each system. Logistic Regression (LR) is then used to determine a station's class based on external features such as land use, population density, and distance from public transit stations and stops. In parallel, the total demand at a station during the whole day is estimated using a Generalized Linear Model (GLM), and finally, the outputs of the two models are combined in the automated planning system.

The flowchart shows that we use two parallel prediction models: the first attempts to classify bike sharing stations by considering daily patterns in the flow of bikes to and from them, finding typical patterns and trying to use urban environment data (e.g., land use, distance from transit stations, etc.) to predict the possible patterns in new locations, while the second is a simple regression model to predict the amount of traffic expected for a given location.

3.1. Data Sources and Processing

We considered bike sharing trip data from 2019, as this was the most recent year in which the systems were unaffected by the recent COVID-19 pandemic,

which affected traffic patterns in significant ways, as Pase et al. (2020) showed. We also limited our analysis to business days: tourist and leisure traffic on holidays is an order of magnitude smaller than commuter traffic, and thus has a negligible effect on planning, as well as depending on local factors and landmarks more significantly. These datasets include the times and location of the beginning and end of each trip, as well as the user ID.

We excluded two kinds of trips from the dataset: loop trips, i.e., trips that had the same departure and arrival point, which are often recreational, as shown by Zhao et al. (2015), and trips taken by temporary users (in cities which have this distinction in the dataset), who Noland et al. (2019) argue are most likely tourists visiting the city for a short period. Finally, trips shorter than 60 seconds were considered as false starts or users ensuring that their bike is locked, so they were removed as well. We also removed stations which are suspected to be test stations or otherwise used for maintenance purposes, as well as stations that have a very low traffic (i.e., fewer than 8 daily trips counting both departures and arrivals), from our analysis.

In order to map each station to a particular area, and consider the urban environment in our prediction of the demand patterns for that station, we need to determine the service or catchment area of each station. We considered a simple approximation, i.e., to map each point in the city to the closest station as the crow flies, with a maximum distance of 500 m. The use of Euclidean distance is a minor approximation in urban areas with a dense street grid, as shown experimentally in O'Brien et al. (2014). The catchment area determination is equivalent to a Voronoi tessellation of the city map, which is a well-known problem that can be solved efficiently. The service areas are further truncated such that they do not span over bodies of water such as seas, rivers, and lakes. Land use, population, and public transit data were obtained from public sources, as described in the Appendix.

We also considered that, as stations are often opened and closed, the service area of each station needs to be determined on a daily basis, depending on which stations in the neighborhood were open on a given day: in order to simplify the calculations, we considered the average of each environmental feature over the whole analyzed period. The service areas in New York City for an example day are shown in Fig. 2, overlaid on a street map of the city.

3.2. Pattern Identification and Clustering

Our automated planning system follows other spatio-temporal clustering works, mapping individual bike sharing stations onto relatively few patterns based on their daily arrivals and departures. Using the bike sharing trip data, we calculated the hourly number of arrivals and departures for each station for every business day in which the station was used. The number of arrivals and departures for a specific hour are counted from the beginning to the end of the hour, e.g., from 16:00:00 to 16:59:59. Let $\mathcal{T}_i$ be the set of days in which station i has been used. We then define the two 24-element vectors $\mathbf{d}_{t,i}$ and $\mathbf{a}_{t,i}$, representing the departures and arrivals from and to station i in each hour of day t, respectively. In order to mitigate the effect of the concentration of trips in the

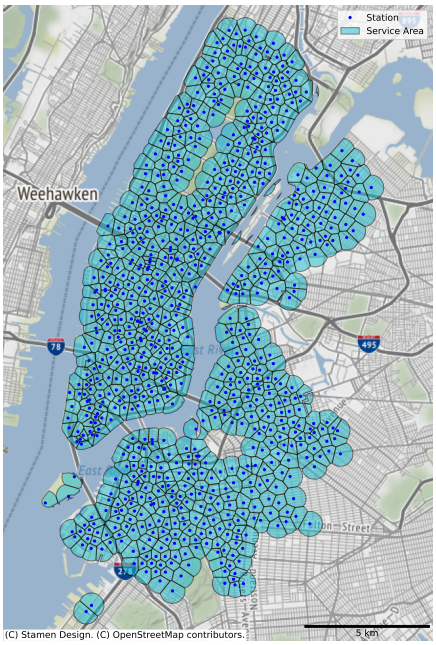


Figure 2: Service areas for New York City on October 23rd 2019.

rush hours on the traffic pattern, we consider the flow to the station, defined as the difference between the number of arrivals and departures:

$$\mathbf{f}_{t,i} = \mathbf{d}_{t,i} - \mathbf{a}_{t,i}. \tag{1}$$

The traffic flow for a given hour is then positive if there are more departures than arrivals, and negative in the opposite case. The hourly traffic flow is then averaged over all days and normalized:

$$\tilde{\mathbf{f}}_i = \frac{\sum_{t \in \mathcal{T}_i} \mathbf{f}_{t,i}}{\sum_{t \in \mathcal{T}_i} ||\mathbf{a}_{t,i}||_1 + ||\mathbf{d}_{t,i}||_1}. \tag{2}$$

The normalization is performed in order to focus the clustering on traffic patterns, not on the absolute number of arrivals and departures to each station. The total traffic demand of a station i is defined as

$$V_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} ||\mathbf{a}_{t,i}||_1 + ||\mathbf{d}_{t,i}||_1, \tag{3}$$

which yields the average number of daily departures and arrivals for the station.

We then used the classical k-means algorithm to divide the stations into classes based on their traffic patterns. The algorithm is a partitioning algorithm which divides the data points into k clusters, minimizing the distance between

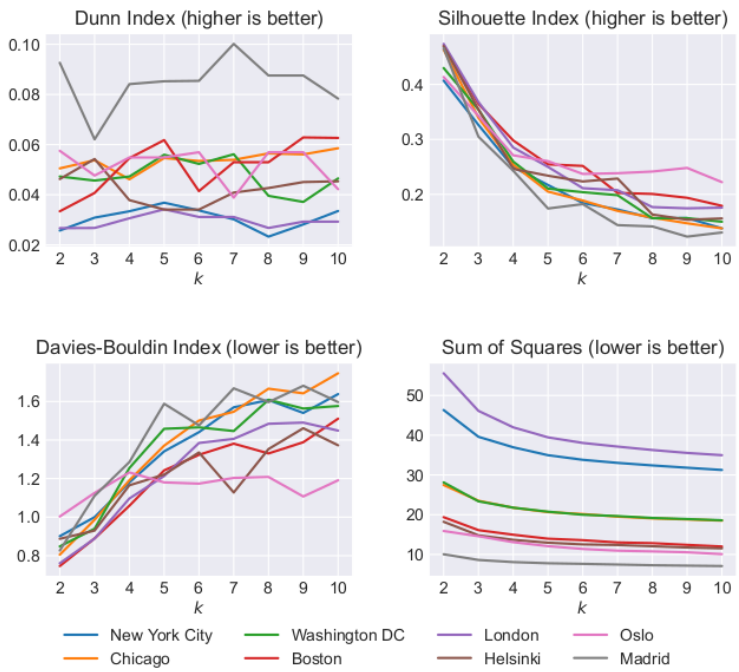


Figure 3: Davies-Bouldin, Dunn, and Silhouette indices and SSE for all cities.

data points in a shared cluster while maximizing the distance between data points in different clusters. While Sarkar et al. (2015) used a distance measure based on the Dynamic Time Warping (DTW) algorithm to account for temporal displacement of traffic patterns, we found that the Euclidean distance led to similar results, while being less computationally demanding, so we adopted that approach.

The choice of the number of clusters k is more complex, as several clustering measures exist. In this paper, we considered the Sum of Squared Errors (SSE), the *Dunn index*, defined by Dunn (1973), the *Davies-Bouldin index*, defined by Davies and Bouldin (1979), and the *silhouette index*, defined by Rousseeuw (1987), to obtain the value $k^* = 5$. The clustering measures for different values of k and different European and US cities can be seen in Fig. 3. The Davies-Bouldin and Silhouette indices are generally better for smaller numbers of clusters. Based on the elbow criterion on SSE, 3 to 5 clusters would seem to be the best choice, while the Dunn index is slightly higher for 5 clusters. We can also note that, while the Davies-Bouldin and Silhouette indices are better for fewer clusters, a lower compactness of the clusters is expected if we can have more nuanced classes.

The resulting cluster centers for all cities are shown in Fig. 4 and the size of each cluster is shown in Table 2. This clustering generally leads to 5 distinct

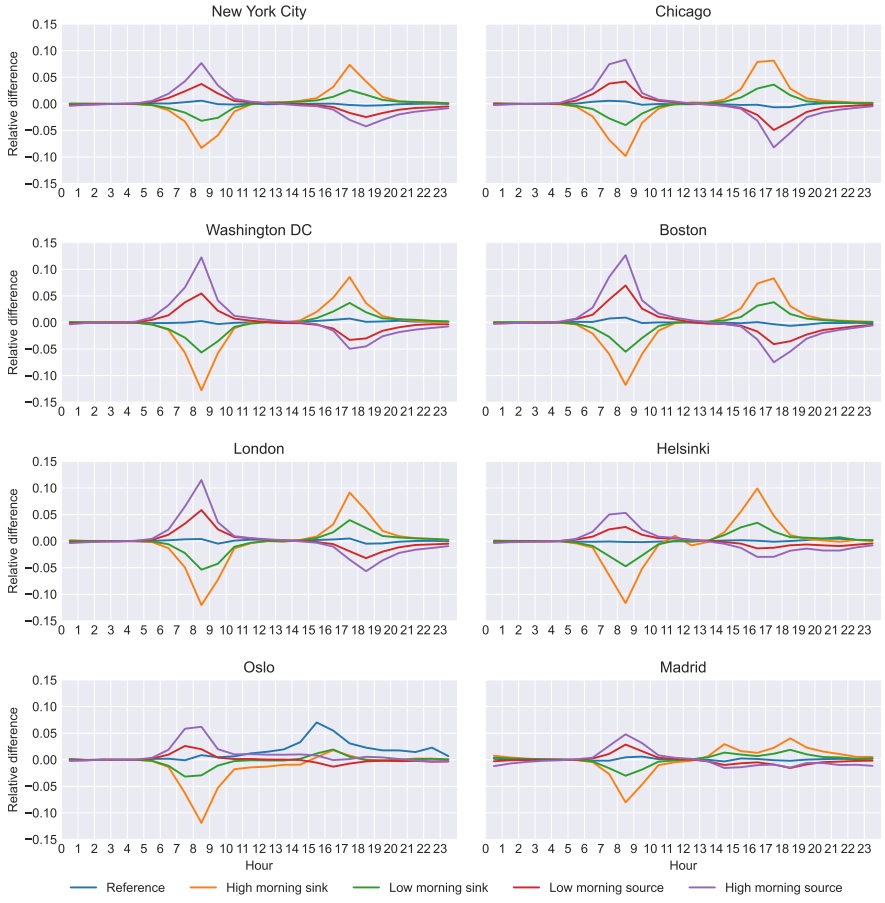


Figure 4: Cluster centers for all cities.

station types: high morning sources, which yield a large output of trips in the morning, low morning sources, which have the same pattern but with smaller flow peaks, high morning sinks and low morning sinks, which are the symmetric correspondents to high and low morning sources, and a fifth cluster we named "reference", which contains stations that do not fit in these patterns. While the general trends are similar, there are interesting differences between the cities. Firstly, the relative size and position of the peaks is often different, reflecting different urban environments and daily work habits (for example, the evening peak is much earlier in Helsinki than in Madrid), and secondly, Oslo and Madrid show pointedly different patterns, which can be explained by the physical features of the environment, i.e., heights and hills creating preferred directions for cycling. However, the general pattern holds throughout most cities, and US cities are particularly similar to each other: we will discuss the applicability of

City	Reference	High morning sink	Low morning sink	Low morning source	High morning source
New York City	253 (29.5%)	63 (7.4%)	162 (18.9%)	243 (28.4%)	136 (15.9%)
Chicago	84 (22.8%)	45 (12.2%)	63 (17.1%)	99 (26.8%)	78 (21.1%)
Washington DC	86 (25.8%)	43 (12.9%)	57 (17.1%)	75 (22.5%)	72 (21.6%)
Boston	63 (24.8%)	22 (8.7%)	50 (19.7%)	69 (27.2%)	50 (19.7%)
London	190 (24.2%)	82 (10.5%)	135 (17.2%)	221 (28.2%)	156 (19.9%)
Helsinki	108 (31.0%)	12 (3.4%)	45 (12.9%)	113 (32.5%)	70 (20.1%)
Oslo	23 (9.2%)	22 (8.8%)	52 (20.7%)	87 (34.7%)	67 (26.7%)
Madrid	59 (27.7%)	34 (16.0%)	36 (16.9%)	44 (20.7%)	40 (18.8%)

Table 2: Population of the 5 clusters in each city. The population is given both as an absolute number and as a percentage of the total number of stations below.

the pattern identification and prediction model across different cities more in depth in Sec. 5.

3.3. Prediction Model

In order to extend the pattern analysis to different neighborhoods in the same city, or even different cities, the labels obtained from the clustering are used as dependent variables in a multinomial LR model which models the probability of a station being in a specific cluster assuming that the log-odds of being in the cluster with respect to the reference cluster is a linear combination of independent variables.

On the other hand, the total demand can be predicted using a GLM: we can assume that the total traffic demand $V_i = \mu_i + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some variance σ^2. The demand is then modeled using a generalized linear model of the form

$$\ln(\mu_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i, \tag{4}$$

where $\boldsymbol{\beta}$ is the vector of coefficients of the model and $\boldsymbol{x}_i$ is the input vector containing the independent variables. The use of a logarithmic function fits well with a power law distribution of the total traffic demand at different stations, as observed by Li et al. (2021) and Zheng et al. (2021) for Chinese bike sharing systems. Since the power law distribution is extremely similar to the log-normal distribution, this linearization should result in a good fit.

The independent variables for both the LR model of the pattern type and the GLM of the total demand are derived from the urban and transit data for each station, and represent both the attractiveness of the area (in terms of both residential population density and potential points of interest such as commercial or recreational facilities, as well as the distance from the city center) and the mode shift opportunities with public transit. The summary statistics on the independent variables on the service area level are listed in Table 3.

3.4. Station Placement

We can then use the demand model to predict both the overall demand and its temporal pattern for any location, including new neighborhoods or cities.

	New York City				Chicago			
	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
Share of residential use	0.51	0.37	0.00	1.00	0.34	0.31	0.00	0.99
Share of commercial use	0.25	0.34	0.00	1.00	0.16	0.17	0.00	0.95
Share of recreational use	0.07	0.16	0.00	0.84	0.08	0.18	0.00	1.00
Population density [persons/100 m^2]	1.37	0.79	0.00	5.50	0.50	0.28	0.07	1.80
Distance to nearest subway [km]	0.35	0.26	0.00	2.11	0.60	0.47	0.01	2.67
Distance to nearest railway [km]	1.90	0.92	0.07	4.30	1.37	0.84	0.03	3.57
Distance to city center [km]	5.43	2.84	0.12	12.34	5.57	3.99	0.08	21.78
	Washington DC				Boston			
	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
Share of residential use	0.50	0.36	0.00	1.00	0.44	0.30	0.00	1.00
Share of commercial use	0.10	0.22	0.00	1.00	0.18	0.20	0.00	1.00
Share of recreational use	0.13	0.26	0.00	1.00	0.14	0.19	0.00	0.88
Population density [persons/100 m^2]	0.44	0.31	0.00	1.43	0.47	0.25	0.00	1.49
Distance to nearest subway [km]	0.64	0.49	0.02	3.48	0.88	0.81	0.02	4.56
Distance to nearest railway [km]	3.13	1.91	0.14	8.61	0.90	0.67	0.03	2.93
Distance to city center [km]	3.74	2.33	0.32	10.92	3.69	2.07	0.07	8.49
	London				Helsinki			
	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
Share of residential use	0.66	0.29	0.00	1.00	0.41	0.25	0.00	0.95
Share of commercial use	0.19	0.24	0.00	1.00	0.23	0.22	0.00	1.00
Share of recreational use	0.12	0.18	0.00	0.99	0.28	0.19	0.00	0.75
Population density [persons/100 m^2]	1.16	0.67	0.00	3.25	0.59	0.57	0.00	3.44
Distance to nearest subway [km]	0.51	0.40	0.01	2.22	1.78	1.58	0.02	6.44
Distance to nearest railway [km]	0.80	0.50	0.01	2.49	2.64	2.01	0.04	7.17
Distance to city center [km]	3.92	2.05	0.14	9.35	5.76	3.30	0.25	12.30
	Oslo				Madrid			
	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
Share of residential use	0.58	0.30	0.00	1.00	0.69	0.25	0.00	1.00
Share of commercial use	0.23	0.25	0.00	1.00	0.18	0.15	0.00	0.70
Share of recreational use	0.11	0.17	0.00	0.86	0.12	0.19	0.00	0.93
Population density [persons/100 m^2]	1.07	0.88	0.00	4.02	2.67	1.28	0.07	6.44
Distance to nearest subway [km]	0.70	0.51	0.04	3.72	0.24	0.15	0.00	0.82
Distance to nearest railway [km]	1.03	0.63	0.03	3.23	1.15	0.66	0.04	3.32
Distance to city center [km]	1.89	1.08	0.07	4.97	2.13	1.23	0.10	5.66

Table 3: Summary statistics of the variables used in the model.

The second phase of our scheme, i.e., the actual planning of station positions and capacities, can then take place using the model. In order to limit the investigation to a finite subset of points, only the road intersections in the area are used as candidate points.

Let $\mathcal{I} = \{i\}$, with $|\mathcal{I}| = I$, be the set of candidate locations for the placement of a station. The expected demand e_i for each $i \in \mathcal{I}$ can be computed using our model. Let $\mathcal{S} \subseteq \mathcal{I}$ with $|\mathcal{S}| = N$ be the set of chosen candidate locations, and define the indicator variables

$$s_i = \begin{cases} 1, & \text{if } i \in \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The expected demand also depends on the chosen solution, as it is computed over the predicted catchment area of the station. Finally, it is necessary to

introduce a constraint to the optimization problem to ensure a wide coverage and avoid clusters of stations where the expected demand is the highest. One such constraint could be to define a minimum distance $d_{\min}$ between selected stations. Doing this, the more constrained optimization problem is

$$\underset{\mathcal{S}\subseteq\mathcal{I}}{\operatorname{argmax}} \quad \sum_{i\in\mathcal{I}} e_i s_i, \tag{6}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}} s_i = N, \tag{7}$$

$$\min_{i,j\in\mathcal{S},\ i\neq j} d(i,j) > d_{\min}, \tag{8}$$

where $d(\cdot,\cdot)$ is the Euclidean distance between two locations.

As the system includes non-linear effects that depend on the choice of stations, the number of possible placements is combinatorial:

$$|\mathcal{S}| = \binom{I}{N} = \frac{I!}{(I-N)!N!}, \tag{9}$$

which yields 1.7×10^{13} combinations when selecting 10 out of 100 crossroads. For any sizeable area, it is impossible to compute the suitability of all combinations within any reasonable amount of time, and the only practical solution is to use a simplifying heuristic.

The heuristic solution we propose is to divide the problem by splitting the area into smaller subsections with less than 100 intersections. The number of stations to place in each of the areas can then be determined by dividing the total number of stations to place in the expansion area by the population in each sub-area. If the number of stations to be placed in each sub-area is reduced to at most 10, the number of different combinations of station selections in each sub-area can then be limited to a number that allows brute-force search to be computationally feasible.

3.5. Station Capacity Optimization

As a final objective for our planning optimization, we consider the necessary capacity for each station: naturally, increasing the number of docks at a station incurs additional expenses, both in setting up the station and in maintenance and operation. Capacity will then need to be minimized, while still maintaining an acceptable service level. In order to determine this service level, we adopt a Markov-Modulated Birth-Death Process (MM-BDP) model, as defined by Andronov (2011), for the traffic at a station, using the predicted patterns and traffic volumes to determine the hourly rates. The rates of arrivals and departures at station m are then given by vectors $\mathbf{a}_m$ and $\mathbf{d}_m$, respectively, and arrivals and departures follow independent Poisson processes. The state of the MM-BDP is then represented by the number of available bikes at the station: as there can be no arrivals when the station is full, and no departures when the station is empty, the difference between the *bounded* and *unbounded* versions of

the MM-BDP determines the service availability. This model has been exploited by Zhou et al. (2018), Chiariotti et al. (2020), and others in various works on short-term bike sharing prediction and optimization.

If we consider a station i with an initial number of bikes $b(t)$ at hour t, demand vectors $\mathbf{a}_i$ and $\mathbf{d}_i$, and a maximum capacity B_i, we can then consider a 10 minute interval $\tau = \frac{1}{6}$ h and determine the transition matrix $\mathbf{P}^{(1)}(t,i)$, whose element b, b' represents the probability that the station will have b' bikes after a time τ:

$$P^{(1)}_{b,b'}(t,i) \approx \begin{cases} \sum_{n=-\infty}^{-b} p_{\mathrm{Sk}}\left(n; \tau a_i(t), \tau d_i(t)\right), & \text{if } b' = 0; \\ p_{\mathrm{Sk}}\left(b' - b; \tau a_i(t), \tau d_i(t)\right), & \text{if } 0 < b' < B_m; \\ \sum_{n=B_m - b} \infty p_{\mathrm{Sk}}\left(n; \tau a_i(t), \tau d_i(t)\right), & \text{if } b' = B_m, \end{cases} \tag{10}$$

where $p_{\mathrm{Sk}}(n; a, d)$ is the Probability Density Function (PDF) of the Skellam distribution, defined by Skellam (1946), which corresponds to the difference between two independent Poisson random variables with rates a and d, given by:

$$p_{\mathrm{Sk}}(n; a, d) = e^{-(a+d)} \left(\frac{a}{d}\right)^{\frac{n}{2}} I_n(2\sqrt{ad}). \tag{11}$$

$I_n(\cdot)$ is the modified Bessel function of the first kind, as given by (Abramowitz and Stegun, 1964, p. 375). The transition matrix in (10) is approximated, since the Skellam distribution includes all possible sequences of events, including ones that are impossible (e.g., a departure from an empty station, followed by two arrivals, would be counted as a valid transition from $b = 0$ to $b' = 1$). We can also simply extend the transition matrix to consider K timesteps, thanks to the Markov property:

$$\mathbf{P}^{(K)}(t,i) = \left(\mathbf{P}^{(1)}(t,i)\right)^K. \tag{12}$$

We can then write the expected downtime $D^{(K)}_{t,i}(b)$ over a time horizon of K steps, starting from b bikes at hour h:

$$D^{(K)}_{t,i}(b) = \sum_{k=1}^{K} \tau \left(P^{(k)}_{b,0}(\lfloor t + \tau k \rfloor, i) + P^{(k)}_{b,B_m}(\lfloor t + \tau k \rfloor, i)\right). \tag{13}$$

The objective of the optimization is simple:

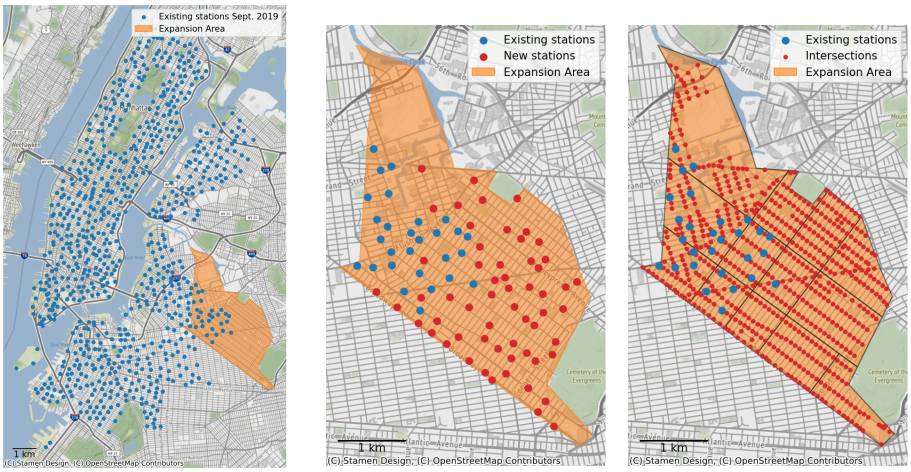
$$\min \qquad B_i, \tag{14}$$

$$\text{s.t.} \qquad \min_{b \in \{0,\dots,B_i\}} \frac{D^{(K)}_{t,i}(b)}{K\tau} \leq p_{\mathrm{thr}} \ \forall t, \tag{15}$$

where p_{thr} is a threshold value set by the system designers.

4. The New York City Use Case

In this section, we will consider New York City's CitiBike system as a representative use case: in particular, we will consider the 2019 expansion of the



(a) New York City station map in September 2019.

(b) Expansion area station map in November 2019.

(c) Expansion area intersections and polygon subdivision.

Figure 5: New York City 2019 expansion map.

system. In the fall of 2019, the CitiBike system in New York City opened several new stations in a new area straddling the boundary between Queens and Brooklyn. The area, consisting of parts of the neighborhoods East Williamsburg, Bushwick and Ridgewood, had been identified as the next area for expansion as part of phase 3 of the New York CitiBike system development.

The NYC Department of City Planning (2009) initially laid out three phases for the roll-out of the CitiBike system when first planning it in 2009. The area identified for phase 1 was built in 2013, while the area of phase 2 was split up as phase 2, which was undertaken between 2015 and 2017, and phase 3 which started construction in 2019 and is expected to be completed in 2023.

While the initial phase 1 was opened all at once to ensure that the system had enough stations from the beginning, phases 2 and 3 were and are rolled out section by section. The section that we are looking at in this specific case study is the first section of phase 3 in 2019. A map showing the expansion area is shown in Fig. 5a.

The NYC Department of City Planning (2009) selected the areas for the different phases primarily basing on estimated demand for a bike sharing system. Factors such as the amount of residents cycling or walking to work, the percentage of workers living within 2.5 miles ($\approx$ 4 km) or 5 miles ($\approx$ 8 km) of their workplace as well as population density influenced the decision. This was done in order to optimize the popularity and thereby the profitability of the system, as the system was designed to operate without government subsidies. The area for the 2019 expansion had high scores for all of the aforementioned factors (NYC Department of City Planning, 2009, pp. 71-73), compared to the other areas included in the phase 3 expansion, which likely contributed to the selection of this area as the first in phase 3. At the eve of the expansion in

September 2019, 28 stations existed in the expansion area.

At the end of November, 58 new stations had been placed all over the expansion area, including the neighborhood in the service area of the bike share system as well as filling in gaps and increasing density between existing stations in its western corner, as Fig. 5b shows.

4.1. Traffic Patterns

The station-level patterns resulting from the clustering are shown in Fig. 6: the reference cluster stations are marked in blue, high and low morning sinks are marked in yellow and green, respectively, while high and low morning sources are marked in purple and red. Stations marked in grey are not considered in the clustering, as their traffic is too low. We can see a relatively clear pattern: stations in Lower and Midtown Manhattan tend to be morning sinks, while the ones placed in Uptown Manhattan, Brooklyn, and Astoria are more often morning sources. This pattern fits with the typical patterns of workers living in residential and suburban areas and commuting into the city center. In New York City, people usually commute to work from 8 to 9 and then leave work between 17 and 18, as Fig. 4 shows. Furthermore, we can analyze the patterns more in depth by looking at the LR model that links the bike sharing demand with the surrounding urban environment.

To better visualize the relationship between the inputs and outputs of the LR model, we constructed a heatmap by dividing the city area into 200 m $\times$ 200 m cells and then calculating the share of different land uses, population density, and the distance from the center of the cell to the nearest subway and railway stations and to the city center. These variables were then used as input for the LR model to obtain the probability of the center of each cell being in each cluster. The resulting heat maps for New York City can be seen in Fig. 7.

Stations in commercial and industrial areas such as the Diamond District, the Financial District, and the Hudson River shoreline are more often morning sink stations, while morning source stations are more likely to be in residential areas and areas with a high population density. In fact, the map of commercial areas in Fig. 7 clearly overlaps with the high morning sink areas, while the map of residential areas is very similar to the residential use map. The heatmaps also show how mixed areas generate the predictions with the lowest level of certainty, represented by darker colors in the cluster probability maps.

This is also readily seen in the coefficients for the model, where high morning sink stations have a significantly higher coefficient for the share of commercial use, while the opposite is true for residential use. When comparing the coefficient for the share of residential use for each cluster type, we can notice a gradual change of this coefficient, with high morning sinks having the lowest value, then low morning sinks, low morning sources, and finally high morning sources with the highest coefficient. It is important to note that there are two effects contributing to the imbalance in morning source stations, both of which can be attributed to the strong separation between residential and commercial areas in most US zoning codes. The first is the abundance of departures in the morning, easily explained by the relatively large number of people living in

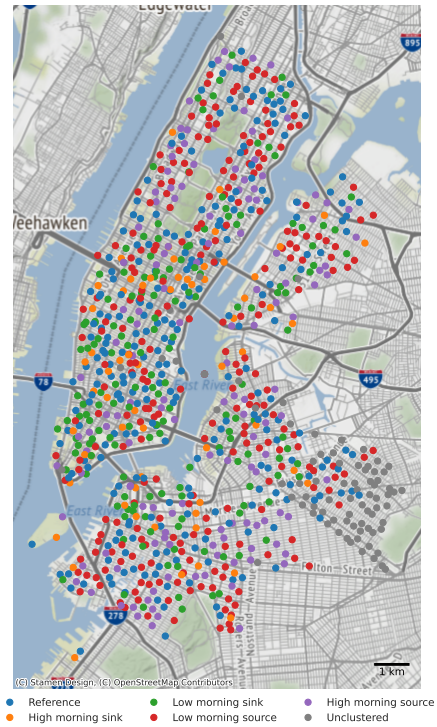


Figure 6: Map showing the demand patterns for each station in New York City.

residential zones compared to other areas, yielding a large amount of potential cyclists who commute towards commercial areas. The second effect is an absence of arrivals: residential areas have little to no commercial purpose, so there is little reason for people to need to visit these areas in the middle of a business day. The high degree of separation between different types of areas and their uses leads to US cities having large areas in which residential and commercial use are mutually exclusive, generating highly unbalanced flows of commuters from residential to commercial areas in the morning, and going the other way at night.

We can also look at the total demand prediction: Fig. 8 shows the heatmap of the demand prediction in New York City, along with the urban environment features. We can see that Lower Manhattan and the center of Brooklyn are the hotspots of demand, with a strong correlation with commercial areas. The Upper East Side, which is the most densely populated area in the city, also corresponds to a demand hotspot, as do a couple of areas in Astoria. The demand prediction map extends outside the current coverage area of the CitiBike system, and shows that another hotspot could exist in the Ocean Hille neighborhood in the south-eastern corner of the map, corresponding roughly to the Broadway Junction metro station and East New York railway station. As this

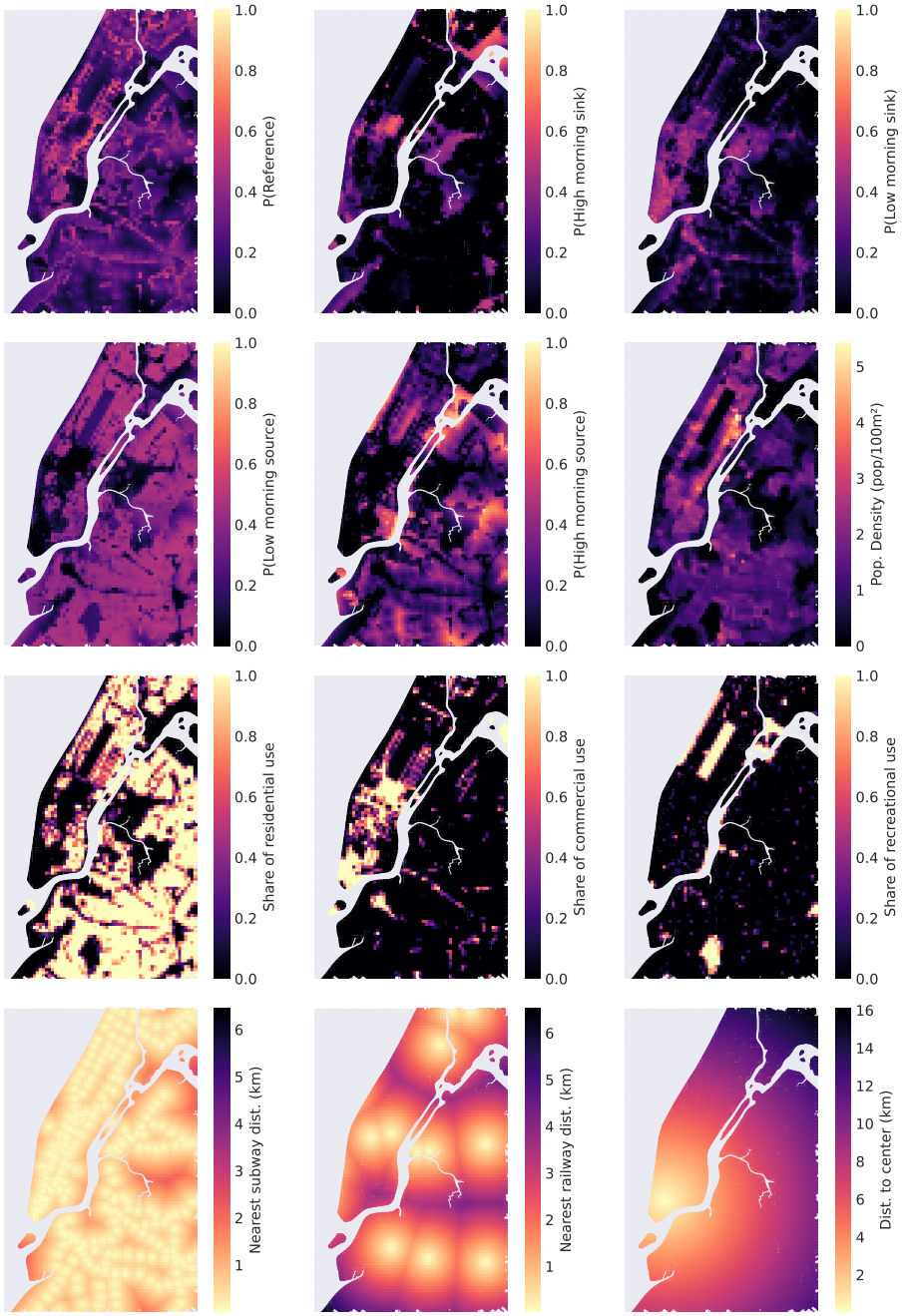


Figure 7: Heatmaps of cluster probabilities and urban features in New York City.

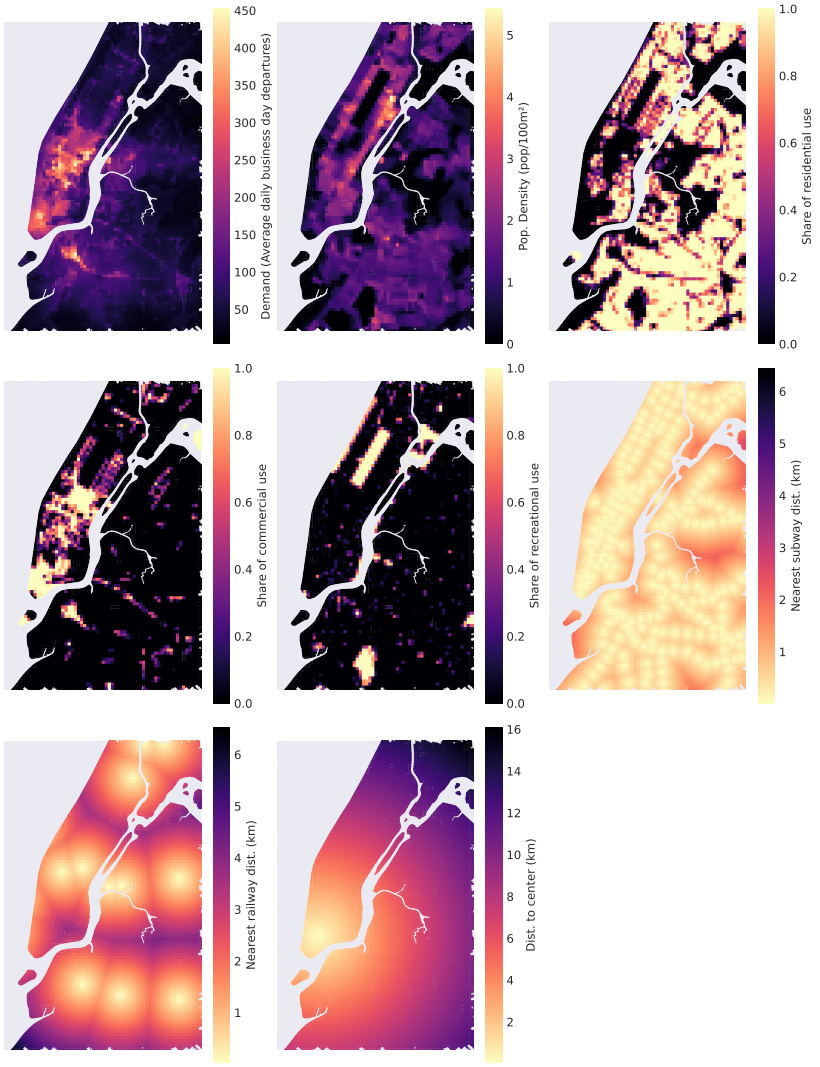
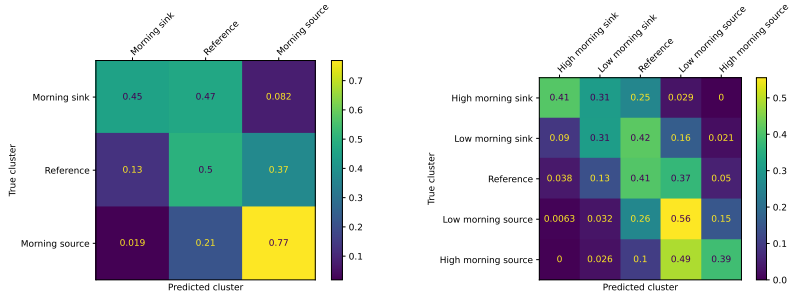


Figure 8: Heat map of predicted demand and external variables for New York City.



(a) 3 clusters. (b) 5 clusters.

Figure 9: LR model confusion matrix for New York City.

area includes both transit hubs and commercial zones in the middle of residential neighborhoods, it is an ideal candidate for a hub of the bike sharing system, concentrating first mile traffic.

Finally, we consider the prediction quality and potential for generalization to different areas by looking at the confusion matrix for the cluster prediction, shown in Fig. 9. In order to fairly evaluate the model, we split the stations randomly into a training set and test set, with the training set having about 80% of the stations. The probability of predicting the correct type is between 30% and 60%, but most of the errors are between clusters of the same type: low morning sinks are often confused by the predictive model as high morning sinks, and vice versa, but morning sinks are almost never confused for morning sources. Confusing morning sinks for morning sources, and vice versa, happens less than 1% of the time, while low morning sinks are confused for low morning sources about 15% of the time, probably due to being in a mixed area. If we consider a simpler clustering with 3 clusters, the prediction is much easier, but morning sink clusters are often confused for reference clusters, due to the large number of reference clusters with this parameter.

4.2. Automated Planning

The aim of this case study is to determine the station placement that maximizes the satisfied demand in the expansion area; in order to predict the latter, we apply the demand model to the expansion area. Additionally, the model can also predict the traffic pattern at each candidate location, which can concur in determining the number of required docks for each station position. The model is fitted to average station data for the months of 2019 prior to October, i.e., only to the traffic demand prior to the expansion, ensuring that the implemented station placement does not influence the model coefficients.

In order to limit the investigation to a finite subset of points, only the road intersections in the expansion area are used as candidate points. The locations of the intersections were determined from OpenStreetMap data using the Overpass API via the Python package OSMnx. Due to the way roads are constructed in OpenStreetMap, separated multi-lane roads can give a separate intersection for

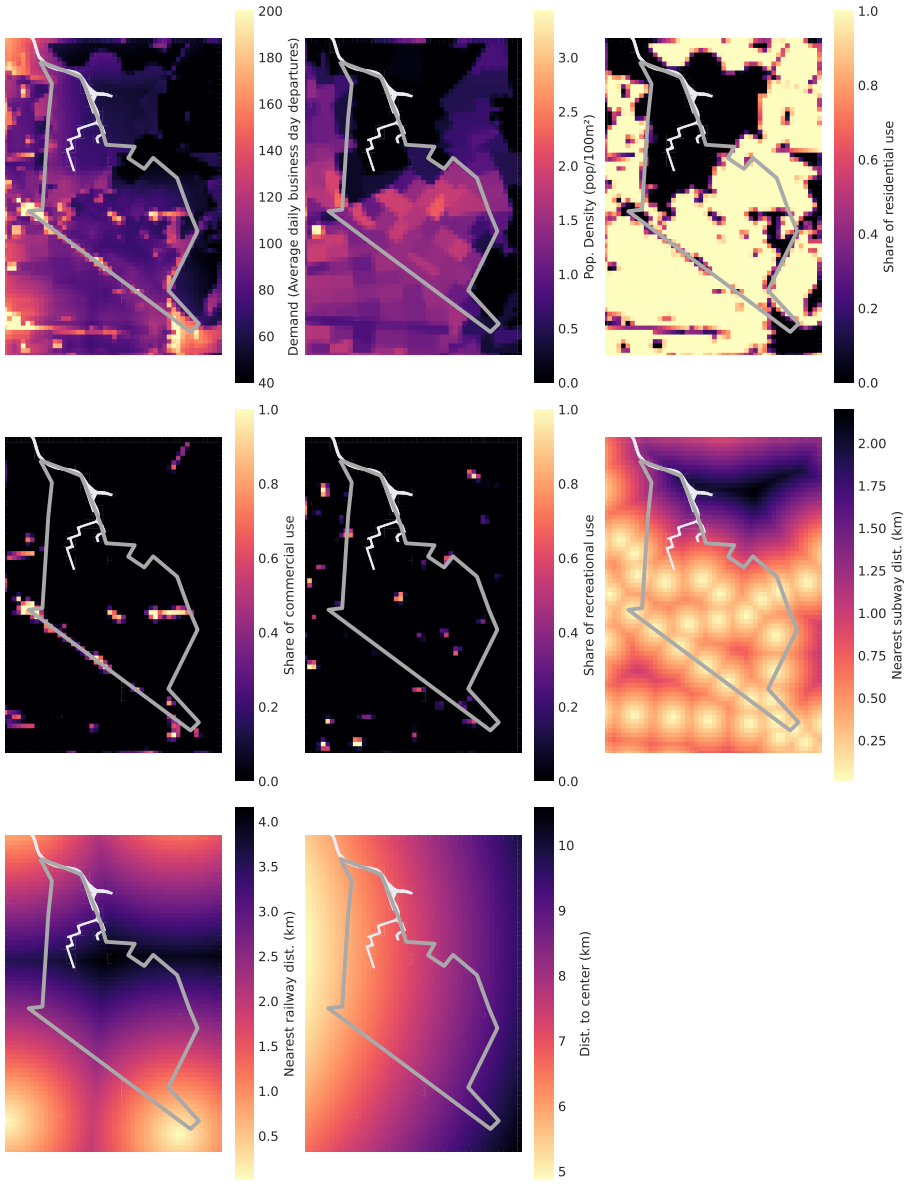


Figure 10: Heatmaps showing the demand model of the expansion area. The grey outline indicates the expansion area.

(a) Selected intersections with predicted cluster and predicted amount of traffic.

(b) Real implemented station locations in November 2019 with traffic data.

(c) Real implemented station locations in February 2020 with traffic data.

Figure 11: CitiBike expansion stations with colors corresponding to traffic pattern cluster type and sizes corresponding to number of trips.

each road lane. Therefore, intersections which are nearer than 20 m to each other are merged to one intersection at the mean point. In the expansion area, this yields 643 candidate locations, as shown in Fig. 5c.

We also show the predicted demand in each point, as well as the external variables considered by the model, in Fig. 10. The demand plot clearly shows that the demand in the northern part of the area, which mostly includes warehouses and logistic infrastructure, has a very low predicted demand. On the other hand, the East to West Myrtle Ave is clearly visible as a high-demand route, as is the south-western corner of the area, close to the East New York train station, a major transit hub. Finally, the western corner of the expansion area is close to the Broadway Triangle, which has both a major subway station and an active commercial area. In 5c, a subdivision of the expansion area into 13 sub-areas is seen. In each of these sub-areas, there are just 33 to 87 intersections, making the station placement problem tractable. We note that our approach divides the total number of stations to place in the expansion area by the population in each sub-area, similarly to how the NYC department of City Planning determined the expansion areas.

With a minimum distance between the stations of $d_{\min} = 250\,\mathrm{m}$ and demand predictions obtained from the model illustrated in Fig. 10, the optimal solution is seen in Fig. 11a. For comparison, the realised solution as implemented by the CitiBike system is seen in Fig. 11b. In order to get an idea of how well the model predictions match with the real world, we use them to predict the traffic patterns, including the predicted cluster type and volume of traffic for each selected station in the expansion area. The model is trained on data from September 2019 and thus has no prior knowledge about the new stations.

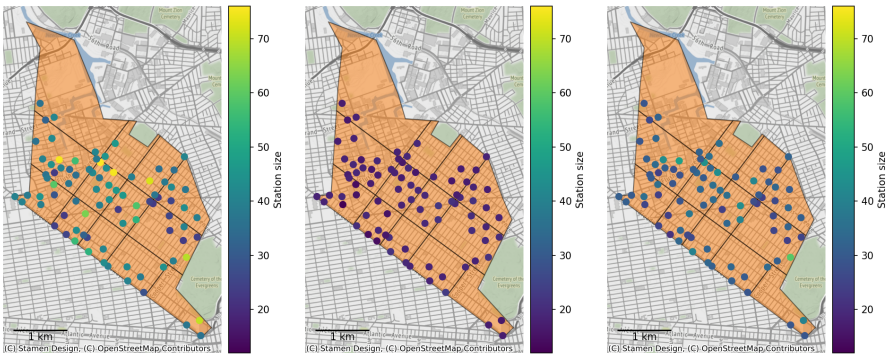
The predicted clusters line up relatively well with the actual traffic types. Reference clusters are placed along the western side, while low morning sources are in the central part of the expansion area. In the actual traffic, there is a mixture of reference stations and low morning sinks on the western side, while the central parts have a mixture of low morning sources and high morning sources as well as reference clusters. While the cluster classification is not completely accurate, for most stations, the actual traffic pattern is either the same as or an adjacent type from the predicted one. However, the prediction consistently underestimates the amount of traffic in the central part of the expansion area. This may be related to the novelty effect of the expansion, which may cause an initial surge of demand, before a subsequent stabilization: the traffic in February 2020, shown in Fig. 11c, is generally lower than in Fig. 11b, and has no stations in the reference stations, with most stations that were in that cluster in November becoming low morning sinks. In general, there are several factors influencing demand, including the weather, but the optimized model is close to reality and can be used as a first step for system design.

Comparing the determined optimal solution to the implemented solution using the objective function for the optimization in (6) shows a clear advantage to the solution determined by the optimization procedure, with a score of 6749.2 for the implemented solution, while the optimized solution scores 8513.6. The determined solution serves a larger fraction of the expected demand than what was really implemented.

However, determining the optimal solution by subdividing the expansion area is vulnerable to border effects: a station in one sub-area can be placed close to a station in another sub-area. This occurs for example in the center of Fig. 11a. This can be mitigated by increasing the size of the areas, at a higher computational costs, or by considering different areas. However, these mitigations are outside the scope of this work, and will be considered in future extensions.

Secondly, the objective function is designed purely to maximize the demand met by the bike share stations. However, the Department of Transportation likely has other considerations which weigh on their decision of station placement, as evidenced by their choice of station locations in 11b. They may, e.g., put more weight on an even distribution of stations throughout the area. If these other considerations are not reflected in the objective function, it is only natural that a solution which is designed to optimize a certain objective function performs better than a solution which is designed to optimize a different objective. The station placement and capacity allocation can be further refined by subsequent practical considerations, followed by feedback rounds to gather input from local residents, as described by Gavin et al. (2016).

Finally, we can see the results of the station dimensioning optimization in Fig. 12. We considered 3 different starting times, i.e., 7:00, 12:00, and 16:00, with $K = 18$ and $\tau = 10$ min, corresponding to a 3-hour interval. Two of the intervals correspond to the morning and evening peaks, while the third considers the midday traffic. The figure clearly shows that the critical moment for the bike sharing system is in the morning: in order to maintain service



(a) Dimensioning for the 7:00-10:00 rush hour.

(b) Dimensioning for the 12:00-15:00 interval.

(c) Dimensioning for the 16:00-19:00 rush hour.

Figure 12: Station dimensioning based on predicted patterns for the 2019 expansion area in New York City.

with probability $p_{\text{thr}} = 0.9$, i.e., have a lower than 10% chance of the station becoming completely full or completely empty, starting from the best possible allocation, 4 stations (2 morning sources, 1 morning sink, and 1 in the reference cluster) require more than 70 docks. In 2019, the largest station was in lower Manhattan, with 79 docks. This means that the system is underdimensioned for the chosen reliability target, as Chiariotti et al. (2018) show: under a similar traffic model, the failure rate of the system with two daily rebalancing trips at 3AM and 3PM is around 10%, meaning that critical stations such as the ones highlighted in the figure will suffer more frequent service outages.

5. Traffic Patterns Across Cities

In the previous section, we discussed the New York City CitiBike expansion as a potential use case for automated planning. Extending a system to a new neighborhood is a relatively simple process: although boundary effects (the edges of the network often behave idiosyncratically) and local peculiarities are always present, similar neighborhoods are often already incorporated in the network and provide a solid basis for pattern characterization. Furthermore, the central areas of a city are usually the first to be served, and expansions of existing systems usually target residential or suburban areas. On the other hand, planning an entirely new system is more complex: local conditions may vary significantly across cities, and factors such as climate, hills, and different urban organization need to be considered. In this section, we examine data from different cities, analyzing the generality of patterns and how data from other cities can be adapted to fit an entirely new system.

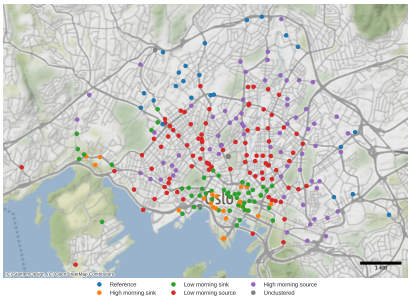
The results of the clustering for Oslo, London, Chicago, and Washington DC are shown in Fig. 13. Oslo, shown in Fig. 13a, is an interesting outlier, as its reference cluster has a balanced traffic in the morning and more departures in the evening: 23 stations, or 9.2% of Oslo's network, belong to this type, and a large

majority of them are located in the north-western part of the network serving the Ullevål district, which contains Oslo University. A possible explanation of this irregular cluster may be that students use conventional public transport, such as buses, trams or the metro, to arrive at the university in the morning, and then use bike sharing to depart from the university in the afternoon. This pattern is probably caused by the difference in elevation: Ullevål is between 50 and 100 m above sea level, so that cycling to it from the city center or other residential areas close to sea level is mostly uphill. Students commuting into the university might prefer to avoid the climb and take the bus in the morning, while cycling becomes convenient in the evening, when the route is mostly downhill.

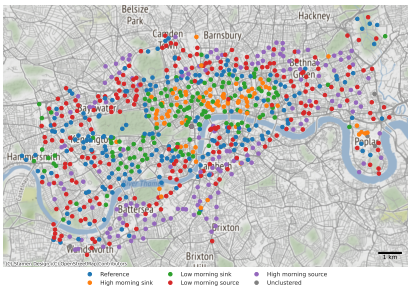
On the other hand, London, Chicago, and Washington DC exhibit a more regular pattern: high morning sinks are clustered in the city center, as commuters get to their workplace in the morning and leave in the afternoon, while stations on the outskirts are often morning sources. This is in contrast with the patterns we showed in New York City, which has a mix of sources and sinks, mostly corresponding to the zoning of different neighborhoods: the five boroughs of the city have a mix of residential and commercial areas, and it is difficult to draw a straightforward distinction between center and periphery. Chicago and Washington DC also have several underused stations, which had fewer than 8 daily trips, on the outskirts of town: stations in these areas, which are often poorer and badly served by mass transit, are far from transit stations or workplaces, and would thus require users to cycle significant distances to get to their destinations. The higher density of London and Oslo, and the limited extension of their bike sharing systems to the suburbs, prevent any such stations from appearing in the maps.

In some cities, including New York City, Washington DC, and London, there appears to be an imbalance in the afternoon rush hours, with the peaks of the morning sinks being considerably larger than the valleys of the morning sources. In Helsinki, Oslo and Madrid, an opposite imbalance is also seen in the morning rush hours. These imbalances are possibly caused by a disparity between the number of morning sinks and morning sources. In Helsinki, where these imbalances are most prominent, 16.3% of the stations are morning sinks, while 55.7% are morning sources, a difference of 39.4%. This indicates that the trips leaving from the many morning source stations are concentrated in a few key areas, which then disperse the trips back to the morning source stations in the afternoon.

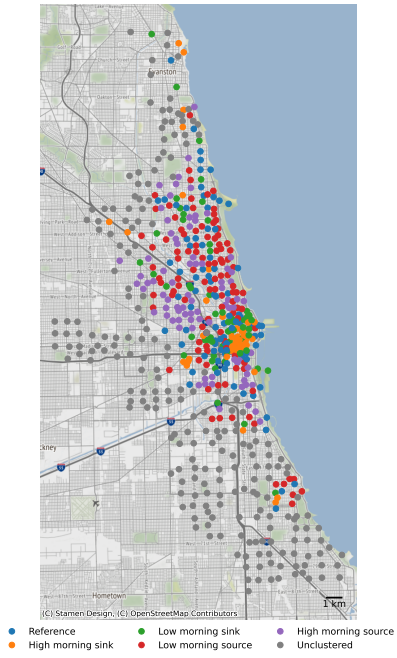
The figures also suggest how well a prediction model could generalize to other cities: despite the minor differences mentioned above, patterns in US cities are very similar. The coefficients obtained from the LR models trained on each city are presented in Table 4. The table confirms our visual intuitions, and the results for New York City, for most other US cities: morning sinks are generally more likely to be in commercial areas, while morning sources are usually in residential areas. On the other hand, if we consider European cities, we note that the patterns are much more idiosyncratic, with wide difference between cities: London is the only capital we analyzed to have a high degree of similarity with US cities, perhaps due to the general commuting patterns from suburban



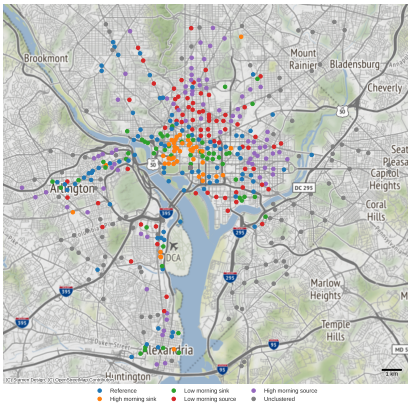
(a) Oslo.



(b) London.



(c) Chicago.



(d) Washington DC.

Figure 13: Station maps for Oslo, London, Chicago, and Washington DC, divided by cluster.

Cluster	Coef. name	NYC	Chicago	DC	Boston	London	Helsinki	Oslo	Madrid
	Const.	-0.247	**2.193**	**2.210**	**3.623**	**-11.171**	-3.792	**19.973**	-2.938
	Residential	**-4.200**	**-3.260**	**-6.538**	-0.185	**15.293**	-2.039	-10.165	**13.670**
	Commercial	**2.016**	-0.511	-0.559	-0.241	**13.703**	1.636	-11.843	4.654
High	Recreational	-1.279	-7.586	**-3.661**	**-3.909**	6.408	2.102	-11.076	3.946
Morn.	Density	**-1.809**	**-1.886**	-1.797	**-6.079**	**-3.521**	-4.598	-1.424	**-2.813**
Sink	Subway distance	-0.502	**-2.349**	0.332	0.014	0.083	-0.345	**6.194**	1.198
	Railway distance	-0.205	-0.383	**0.493**	0.427	0.664	-0.024	**-6.149**	**-2.133**
	Center distance	0.152	0.080	**-0.577**	**-0.609**	**-0.321**	**0.415**	**-3.707**	0.034
	Const.	**1.785**	**2.234**	1.238	**2.787**	**-3.956**	-0.515	**18.862**	-2.188
	Residential	-0.875	-1.261	**-3.797**	-1.147	**6.567**	0.333	-9.182	**9.133**
	Commercial	0.301	-1.798	-0.400	-0.661	**5.758**	1.266	-10.123	3.983
Low	Recreational	-0.093	**-2.628**	**-1.747**	-2.090	1.863	-2.341	-11.704	1.379
Morn.	Density	**-0.769**	**-1.675**	-0.713	**-1.902**	**-1.726**	**-1.625**	-1.760	**-2.052**
Sink	Subway distance	**-1.972**	**-2.140**	-0.341	-0.323	-0.046	-0.172	**2.543**	2.100
	Railway distance	-0.143	0.138	**0.351**	-0.306	**1.001**	0.034	**-3.175**	-0.687
	Center distance	-0.065	0.007	**-0.233**	-0.162	**-0.356**	0.102	**-2.093**	0.040
	Const.	**-2.281**	**-1.507**	**-2.333**	-0.726	-0.004	**-4.019**	15.057	-6.033
	Residential	**1.708**	**3.955**	**2.904**	0.134	-1.814	2.028	-10.586	-1.215
	Commercial	-1.031	0.434	-2.536	-1.062	-2.432	-1.359	-12.577	4.138
Low	Recreational	**1.544**	-0.320	0.231	**-3.259**	-1.298	2.262	-8.750	6.801
Morn.	Density	0.271	-0.243	**1.712**	-0.379	**1.111**	**1.261**	1.739	**1.908**
Source	Subway distance	**1.243**	**1.013**	**0.825**	**0.599**	**0.987**	**0.340**	**3.012**	0.732
	Railway distance	**0.283**	**0.654**	**-0.426**	**0.704**	**-0.924**	**0.385**	-1.313	**-0.868**
	Center distance	0.011	**-0.226**	0.150	0.017	**0.177**	0.075	**-1.880**	0.242
	Const.	**-5.030**	-1.263	**-6.701**	-0.657	-1.461	**-2.356**	10.181	-9.216
	Residential	**3.002**	**6.045**	**5.584**	0.805	-0.633	-0.564	-6.766	-1.206
	Commercial	**-3.195**	-2.686	**-8.969**	-3.322	-2.532	**-7.125**	-9.546	2.369
High	Recreational	**3.623**	**-3.690**	0.668	**-5.676**	-1.905	-1.580	-4.786	5.945
Morn.	Density	**0.693**	-0.101	**2.268**	-1.100	**0.913**	0.867	1.314	**1.941**
Source	Subway distance	**2.950**	**1.441**	**1.230**	0.469	**2.241**	0.222	**2.585**	**4.715**
	Railway distance	**0.413**	0.353	**-0.523**	**0.751**	**-0.777**	0.256	**-1.507**	**-1.167**
	Center distance	-0.061	**-0.376**	**0.686**	0.045	0.112	**0.359**	-0.805	**1.025**

Table 4: Coefficients of LR models trained on different cities. Bold coefficients are statistically significant ($p < 0.05$).

areas into the city center. This unpredictable variation might be due to cultural differences, the local climate, and differences in the urban landscape. In Madrid, there is a monotone decrease in the coefficient for the share of commercial use from morning sources to morning sinks, as was seen in the US cities. However, residential use tends to have a strong correlation with morning sinks as well: in general, it appears that the strongest morning sources are farthest from the center and subway stations, with sinks going both to residential and commercial areas. This can be readily explained by the combination of commuters going into the city using mass transit and using the bike sharing service as the first mile, combined with the more mixed zoning in the city. The enforced separation between residential and commercial areas, common to most US cities, is mostly absent in Europe, except for core business districts, leading to a smaller effect of zoning on the bike sharing traffic patterns.

We can also look at Fig. 14, which shows the cluster probability and urban environment heatmaps for London. The mix of residential and commercial areas is more even, except in the City and the northern shore of the Thames, which coincidentally correspond to the areas with the highest probability of

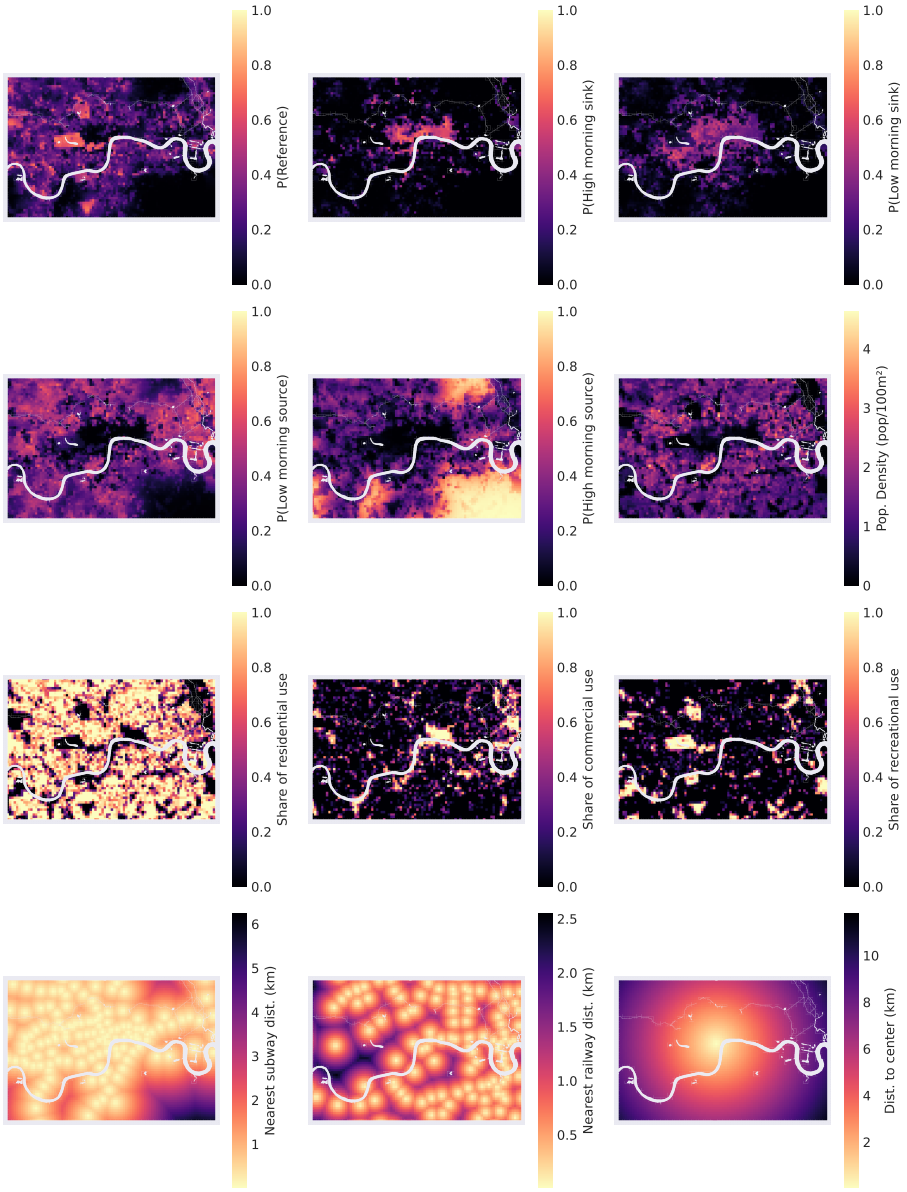


Figure 14: Heatmaps of cluster probabilities and urban features in London.

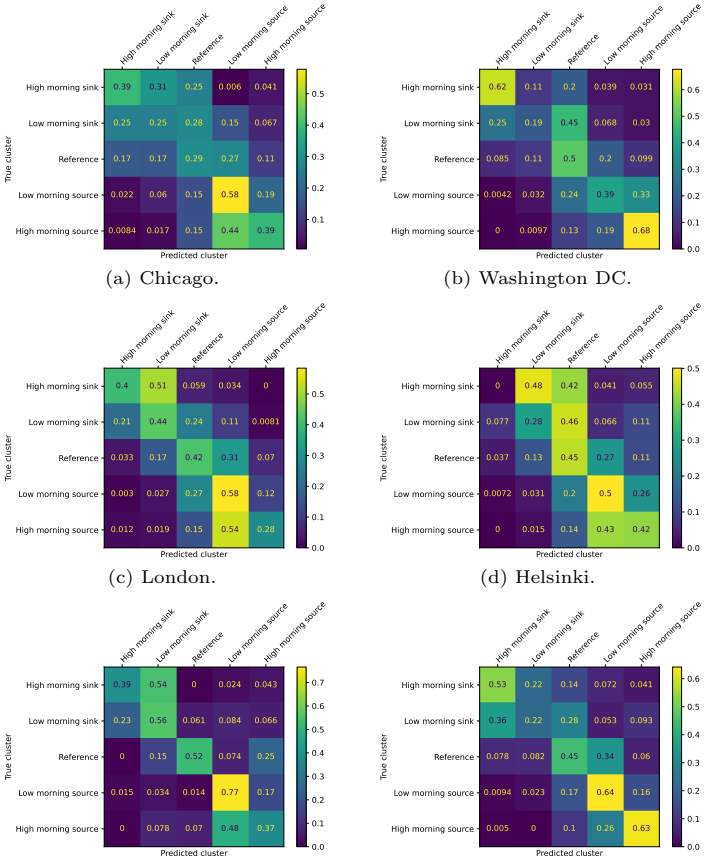
being morning sinks. The south-eastern and northern areas of the map, which mostly contain morning sources, are underserved by the subway, and we can thus infer that first mile service is necessary for those areas in order to reach the closest subway station. By comparing the map to Fig. 13b, we can note that the south-eastern corner and the northern part of the map are not currently served by the bike sharing system: since the predictive model shows that these areas would have a high probability of having morning sources, we can conclude that there is some potentially unserved demand for first mile service in these mostly residential neighborhoods, and that it could be an interesting direction for system expansion.

In general, the distance from public transit (represented by the closest subway and railway stations) is another good predictor of traffic patterns, with two common patterns. In US cities, distance from public transit is strongly correlated with morning sources: as stations closer to transit are less likely to be sources, this indicates that the service is mostly used as the first leg of users' commutes from their homes to public transit. This pattern is common to all US cities, as well as London and Madrid. However, distance from railway stations in London tends to be positively correlated with morning sinks, perhaps indicating the need for last mile service from the central railway hubs. In Helsinki, subway and railway stations are only weakly correlated with traffic pattern types, indicating a smaller share of mixed mode commuting: as the city is the smallest in the dataset, with less than half of the land area of Oslo and less than a third of the area of the other cities, we can expect commuters to just use one mode of commuting.

5.1. Model Generalization Performance

As we did for New York City, we can consider the confusion matrix for each city, shown in Fig. 15. We can note that, for all cities, the probability of predicting the correct type is between 30% and 60%, but most of the errors are between clusters of the same type: low morning sinks are often confused by the predictive model as high morning sinks, and vice versa, but morning sinks are almost never confused for morning sources. This result holds both for European and US cities, and the prediction is particularly accurate in Washington DC, Oslo, and Madrid: in general, confusing morning sinks for morning sources, and vice versa, happens less than 1% of the time in these cities, while morning sinks are confused for morning sources at most 10% of the time. In general, morning sinks are more often misidentified as sources than the reverse, probably due to the type of area they usually are placed in. We would also remark that, while the accuracy of the prediction is only relatively good, we split the stations randomly into a training set and test set, with the training set having about 80% of the stations. The performance also includes the generalization to new areas, and can be used for expanding existing bike sharing systems with the given accuracy.

Finally, we considered the generalization properties of the trained models across different cities: we tested each of the 8 models on the same data used to train the other models. This type of generalization can be extremely useful



(a) Chicago. (b) Washington DC.

(c) London. (d) Helsinki.

(e) Oslo. (f) Madrid.

Figure 15: Confusion matrices of models trained and tested on the same city.

for planning entirely new bike sharing systems, as the prediction is performed entirely based on urban features, and does not require any data about traffic demand in that city. The rate at which the models predicted the cluster types correctly in each city can be seen in Fig. 16. The results show that models trained on US cities tend to perform well on other US cities, with the exception of New York City, for which other models perform significantly worse. However, US cities generally have similar urban structures, with strict zoning regulation and similar commuting cultures, so it is relatively easy to transfer results from one to another.

When looking at the European cities separately, we can immediately notice that Madrid is an outlier, which would be expected due to its warmer climate, since the success rates of the model trained in Madrid range from 11% to 20% when testing on other cities, only at least as good as random guessing. However, models trained on other cities seem to perform slightly better when tested on

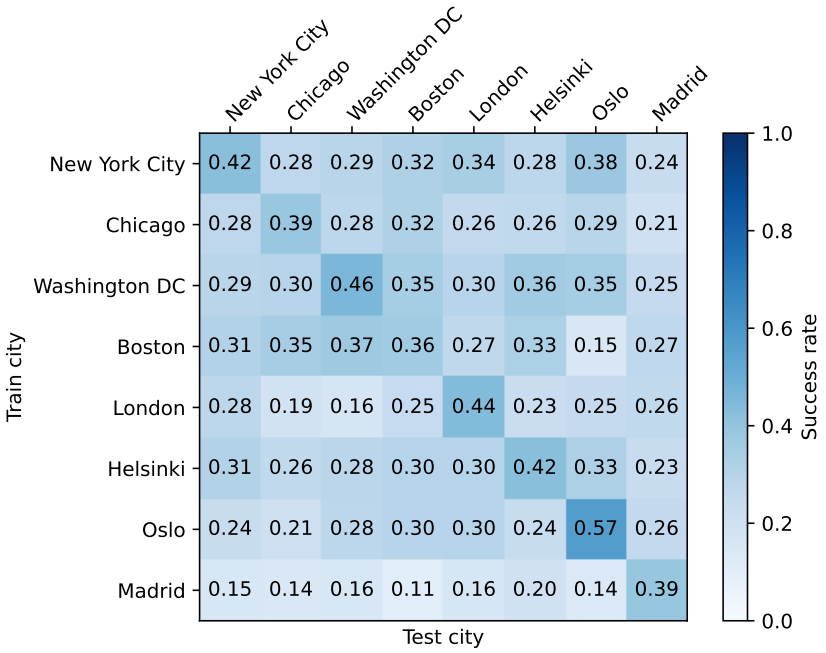


Figure 16: Success rates of models trained and tested on different cities.

Madrid, with success rates ranging from 21% to 27%; this is still far from the 39% accuracy obtained when training and testing on Madrid. The model trained in Helsinki seems to be the most general, with good results both for London and Oslo, while the model trained in London cannot capture the patterns in the two Scandinavian cities as well. When comparing between US cities and European cities, Washington DC and Helsinki stand out as cities whose models perform particularly well on the opposite side of the Atlantic. On the other hand, European models (excluding Madrid) tend to perform better in New York City and Boston, and underperform in Chicago. Interestingly, US models tend to be more robust in Scandinavian cities, as the Washington DC model has better accuracy than any other European model on Helsinki and Oslo. There is also a very low similarity between the two cities, perhaps because of their different physical geography. Interestingly, in Boston, models trained on other cities (namely, Chicago and Washington DC) perform as well or better than the model trained on Boston itself: this might be because the areas of the city are non-homogeneous, and the selection of test stations favored the model in the other cities.

5.2. Traffic Demand Regression

We can then analyze the results of the demand prediction, which is necessary for the dimensioning of stations and the optimization of station placement.

Coef. name	NYC	Chicago	DC	Boston	London	Helsinki	Oslo	Madrid
Const.	**5.851**	**4.349**	**4.687**	**4.586**	**7.244**	**6.204**	**5.082**	**5.055**
Residential	**−0.691**	**−0.553**	0.084	**0.716**	**−2.327**	**−0.599**	**−0.553**	−0.261
Commercial	**0.298**	**0.497**	0.082	0.304	**−2.135**	0.015	0.019	−0.130
Recreational	0.147	−0.381	0.035	−0.157	**−2.11**	−0.230	0.251	0.211
Pop. density	**0.264**	**0.728**	**0.638**	−0.200	**−0.180**	0.049	**0.220**	**0.109**
Dist. subway	**−0.340**	**−0.401**	**−0.381**	−0.154	**−0.459**	**−0.323**	0.191	−0.296
Dist. railway	**−0.277**	**−0.286**	**−0.080**	−0.176	0.080	**−0.175**	−0.139	**−0.116**
Dist. center	**−0.113**	**−0.085**	**−0.237**	**−0.205**	**−0.127**	**−0.084**	**−0.366**	**−0.076**
Pseudo-R^2 (CS)	0.5323	0.5916	0.6861	0.2549	0.5503	0.7249	0.3561	0.1405

Table 5: Coefficients and estimate quality of demand regression model on different cities. Bold coefficients are statistically significant ($p < 0.05$).

Table 5 lists the coefficients of the model for the examined cities. We can easily note that residential areas have a strong negative correlation with demand in most cities, and particularly in London, while the correlation is neutral in Washington DC and positive in Boston.On the other hand, commercial areas in New York City and Chicago have more traffic, while they have less in London, and almost no effect in other cities. The table also reports the Cox-Snell pseudo-R^2 values for all cities: these are not directly comparable, as they might have different ranges, but they represent the improvement over the null model. Cities with a higher value of the pseudo-R^2 also tend to have statistically significant coefficients: New York City, Chicago, Washington DC, Helsinki, and London tend to perform better, while the models are less accurate for Boston, Oslo, and Madrid.

As we discussed for the traffic patterns, this can be related to the urban environment in these cities: in New York City and Chicago, bike sharing acts as a last mile service towards commercial areas in the center, which receive traffic from multiple locations. In Washington DC, the capillary subway service and the different land use, with many administrative services and public buildings in the center, change the traffic patterns significantly, as highlighted by Martin and Shaheen (2014). Population density also tends to be a strong predictor in US cities, with the exception of Boston, and closeness to public transit stations and to the city center is almost universally correlated with more traffic demand. The only exceptions to this trend are Oslo and London, which also showed idiosyncratic traffic patterns for those stations, as we discussed above.

Fig. 17 shows the comparison between the true and predicted traffic in various cities. While the error increases with the number of trips, indicating that the data is probably heteroscedastic, the fit is generally good in most cities, with a relative error on the order of 20-30%. We can then look at generalization performance by comparing the relative error. If we consider the test set $\mathcal{T}_i$ of stations for city i, the relative error of the model trained in city j is given by:

$$e_{i,j} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \frac{e^{\boldsymbol{\beta}_j \mathbf{x}_t} - \mu_t}{\mu_t}. \tag{16}$$

The results are listed in Table 6. Interestingly, models fitted on other US cities do better than the model trained on New York City, even on New York City itself: this might be because of the large number of outliers with a very high traffic

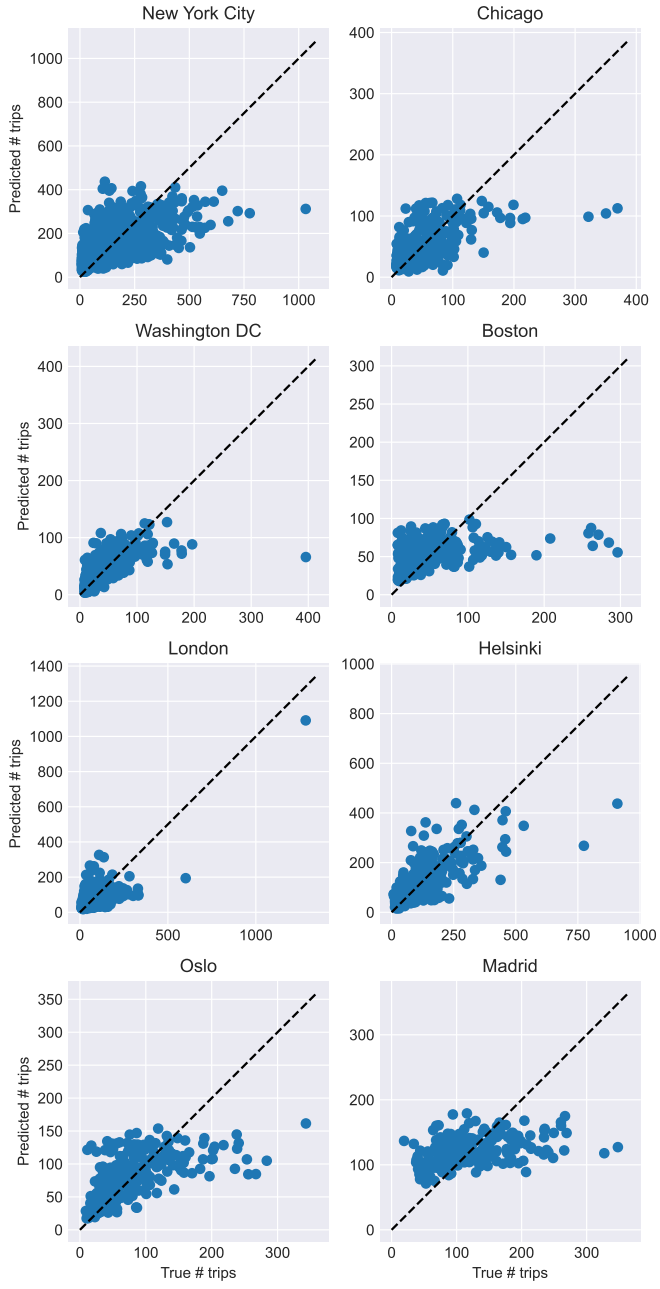


Figure 17: True number of trips vs. predicted number of trips on models trained on each city. Points closer to the dashed line have lower error.

		Test city							
		NYC	Chicago	Wash. DC	Boston	London	Helsinki	Oslo	Madrid
Train city	NYC	0.9	2.9	1.5	3.1	1.7	-0.1	2.0	1.4
	Chicago	0.1	0.4	-0.2	0.4	0.4	-0.7	0.3	1.3
	Wash. DC	0.2	0.3	0.3	0.6	0.7	-0.6	1.2	3.5
	Boston	-0.3	0.4	0.5	0.7	0.0	-0.5	0.3	-0.4
	London	2.1	6.5	5.3	3.8	0.3	0.0	1.0	-0.2
	Helsinki	1.9	5.1	3.2	4.9	2.4	0.5	2.9	1.4
	Oslo	-0.6	0.2	-0.0	0.5	-0.3	-0.6	0.3	-0.1
	Madrid	0.4	1.5	1.1	1.5	0.6	-0.2	0.8	0.3

Table 6: Average relative error when training and testing the demand model across cities. A lower absolute value is better.

load in New York City, which skew the model towards higher predictions. In fact, the New York City model overestimates traffic in all cities except Helsinki. We can note that the Helsinki model also tends to overestimate traffic, while the Oslo, Chicago, and Boston models tend to perform best for all other cities. An aggregate model from multiple cities might give even better results, but it is important to consider both outliers and the overall urban fabric when applying models to a new city. Most likely, US cities will show a closer similarity, as for the traffic patterns, while European cities can have important regional differences due to different laws and local customs.

6. Conclusion

In this paper, we have presented a cross-city method for predicting traffic patterns in docked bike sharing systems from features of the urban environment, giving some considerations on the similarity between different cities and the applicability of models from existing bike sharing systems to new neighborhoods and cities.

The ultimate objective of our analysis is to automate the first stage of bike sharing system planning, i.e., the first proposal for station placement and dimensioning, which can then be submitted to public officials and other stakeholders. We have showcased our automated design pipeline by using the 2019 New York City bike sharing system expansion, providing a comparison with the actually built stations and discussing how our predictive allocation can be used and improved upon.

Future work on the subject may include a deeper analysis of daily and seasonal patterns, in order to account for variations in the usage of the system and climate difference between cities: a wider model could take in information about the city as a whole as well as catchment area information, using data from several cities at once. Another interesting extension is the integration with system maintenance and network data: since operators continuously add or remove docks and stations in different areas of cities, an automatic tool to set up new expansions and highlight overdimensioned stations and underserved areas can be very helpful, as well as furthering understanding of traffic flows in the city. In smaller systems with less than 100 stations, the topology of the bike sharing network graph could also matter, as the density of the system and the precise

location of station affects demand in a more significant way. The automated planning pipeline can also be further improved with the addition of new cities and features, enabling a better prediction and, in turn, a more efficient station placement and dimensioning.

Acknowledgments

This study received funding from the European Union Next-GenerationEU, as part of the Italian National Recovery and Resilience Plan (NRRP), within the MOST – Sustainable Mobility National Research Center (CN00000023).

References

Abramowitz, M., Stegun, I.A., 1964. Handbook of mathematical functions with formulas, graphs, and mathematical tables. volume 55 of *Applied Mathematics Series*. US National Bureau of Standards.

Albuquerque, V., Sales Dias, M., Bacao, F., 2021. Machine learning approaches to bike-sharing systems: A systematic literature review. ISPRS International Journal of Geo-Information 10, 62.

Andronov, A.M., 2011. Markov-modulated birth-death processes. Automatic Control and Computer Sciences 45, 123–132.

Araghi, Y., van Oort, N., Hoogendoorn, S., et al., 2022. Passengers preferences for using emerging modes as first/last mile transport to and from a multimodal hub case study Delft Campus railway station. Case Studies on Transport Policy 10, 300–314.

Çelebi, D., Yörüsün, A., Işık, H., 2018. Bicycle sharing system design with capacity allocations. Transportation research part B: methodological 114, 86–98.

Cervero, R., Denman, S., Jin, Y., 2019. Network design, built and natural environments, and bicycle commuting: Evidence from british cities and towns. Transport Policy 74, 153–164.

Chen, L., Zhang, D., Wang, L., Yang, D., Ma, X., Li, S., Wu, Z., Pan, G., Nguyen, T.M.T., Jakubowicz, J., 2016. Dynamic cluster-based over-demand prediction in bike sharing systems, in: International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), ACM. pp. 841–852.

Chiariotti, F., Pielli, C., Zanella, A., Zorzi, M., 2018. A dynamic approach to rebalancing bike-sharing systems. Sensors 18, 512.

Chiariotti, F., Pielli, C., Zanella, A., Zorzi, M., 2020. A bike-sharing optimization framework combining dynamic rebalancing and user incentives. ACM Transactions on Autonomous and Adaptive Systems 14, 11.

Côme, E., Oukhellou, L., 2014. Model-based count series clustering for bike sharing system usage mining: A case study with the Vélib' system of Paris. ACM Transactions on Intelligent Systems and Technology 5, 39.

Daddio, D.W., 2012. Maximizing Bicycle Sharing: An Empirical Analysis of Capital Bikeshare Usage. Master's thesis. University of North Carolina. Chapel Hill.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 224–227.

DeMaio, P., Yu, C., O'Brien, O., Rabello, R., Chou, S., Benicchio, T., 2021. The Meddin Bike-sharing World Map - Mid-2021 Report. Technical Report. URL: https://bikesharingworldmap.com/reports/bswm_mid2021report.pdf.

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32–57.

EEA, 2020. Mapping Guide v6.2 for a European Urban Atlas. Technical Report. URL: https://land.copernicus.eu/user-corner/technical-library//urban_atlas_2012_2018_mapping_guide.

Elmashhara, M.G., Silva, J., Sá, E., Carvalho, A., Rezazadeh, A., 2022. Factors influencing user behaviour in micromobility sharing systems: A systematic literature review and research directions. Travel Behaviour and Society 27, 1–25.

Eren, E., Katanalp, B.Y., 2022. Fuzzy-based GIS approach with new MCDM method for bike-sharing station site selection according to land-use types. Sustainable Cities and Society 76, 103434.

Eren, E., Uz, V.E., 2020. A review on bike-sharing: The factors affecting bike-sharing demand. Sustainable Cities and Society 54, 101882.

Frade, I., Ribeiro, A., 2015. Bike-sharing stations: A maximal covering location approach. Transportation Research Part A: Policy and Practice 82, 216–227.

Garcia-Gutierrez, J., Romero-Torres, J., Gaytan-Iniestra, J., 2014. Dimensioning of a bike sharing system (BSS): a study case in Nezahualcoyotl, Mexico. Procedia-Social and Behavioral Sciences 162, 253–262.

García-Palomares, J.C., Gutiérrez, J., Latorre, M., 2012. Optimizing the location of stations in bike-sharing programs: A GIS approach. Applied Geography 35, 235–246.

Gavin, K., Bennett, A., Auchincloss, A.H., Katenta, A., 2016. A brief study exploring social equity within bicycle share programs. Transportation Letters 8, 177–180.

Hensher, D.A., 2020. What might Covid-19 mean for mobility as a service (MaaS)? Transport Reviews 40, 551–556.

Hulot, P., Aloise, D., Jena, S.D., 2018. Towards station-level demand prediction for effective rebalancing in bike-sharing systems, in: 24th SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), ACM. pp. 378–386.

Hyland, M., Hong, Z., Pinto, H.K.R.d.F., Chen, Y., 2018. Hybrid cluster-regression approach to model bikeshare station usage. Transportation Research Part A: Policy and Practice 115, 71–89.

Kleisarchaki, S., Gürgen, L., Kassa, Y.M., Krystek, M., Vidal, D.G., 2022. Optimization of soft mobility localization with sustainable policies and open data, in: 18th International Conference on Intelligent Environments (IE), IEEE.

Larsen, J., Patterson, Z., El-Geneidy, A., 2013. Build it. but where? the use of geographic information systems in identifying locations for new cycling infrastructure. International Journal of Sustainable Transportation 7, 299–317.

Lee, M., Hwang, S., Park, Y., Choi, B., 2021. Factors affecting bike-sharing system demand by inferred trip purpose: Integration of clustering of travel patterns and geospatial data analysis. International Journal of Sustainable Transportation , 1–14.

Li, R., Gao, S., Luo, A., Yao, Q., Chen, B., Shang, F., Jiang, R., Stanley, H.E., 2021. Gravity model in dockless bike-sharing systems within cities. Physical Review E 103, 012312.

Li, Y., Zheng, Y., 2019. Citywide bike usage prediction in a bike-sharing system. IEEE Transactions on Knowledge and Data Engineering 32, 1079–1091.

Li, Y., Zheng, Y., Zhang, H., Chen, L., 2015. Traffic prediction in a bike-sharing system, in: 23rd International Conference on Advances in Geographic Information Systems (SIGSPATIAL), ACM. p. 33.

Ma, X., Ji, Y., Yang, M., Jin, Y., Tan, X., 2018. Understanding bikeshare mode as a feeder to metro by isolating metro-bikeshare transfers from smart card data. Transport Policy 71, 57–69.

Macfarlane, J., 2019. The Transforming Transportation Ecosystem – A Call to Action. Technical Report. Institute of Transportation Studies, UC Berkeley.

Martin, E.W., Shaheen, S.A., 2014. Evaluating public transit modal shift dynamics in response to bikesharing: a tale of two US cities. Journal of Transport Geography 41, 315–324.

Noland, R.B., Smart, M.J., Guo, Z., 2016. Bikeshare trip generation in New York City. Transportation Research Part A: Policy and Practice 94, 164–181.

Noland, R.B., Smart, M.J., Guo, Z., 2019. Bikesharing trip patterns in New York City: Associations with land use, subways, and bicycle lanes. International Journal of Sustainable Transportation 13, 664–674.

NYC Department of City Planning, 2009. Bike-Share Opportunities in New York City. Technical Report.

Osama, A., Sayed, T., Bigazzi, A.Y., 2017. Models for estimating zone-level bike kilometers traveled using bike network, land use, and road facility variables. Transportation Research Part A: Policy and Practice 96, 14–28.

O'Brien, O., Cheshire, J., Batty, M., 2014. Mining bicycle sharing data for generating insights into sustainable transport systems. Journal of Transport Geography 34, 262–273.

Pase, F., Chiariotti, F., Zanella, A., Zorzi, M., 2020. Bike sharing and urban mobility in a post-pandemic world. IEEE Access 8, 187291–187306.

Radzimski, A., Dzięcielski, M., 2021. Exploring the relationship between bike-sharing and public transport in Poznań, Poland. Transportation Research Part A: Policy and Practice 145, 189–202.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.

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

Sarkar, A., Lathia, N., Mascolo, C., 2015. Comparing cities' cycling patterns using online shared bicycle maps. Transportation 42, 541–559.

Shaheen, S., Guzman, S., Zhang, H., 2010. Bikesharing in Europe, the Americas, and Asia: past, present, and future. Transportation Research Record: Journal of the Transportation Research Board , 159–167.

Shui, C., Szeto, W., 2020. A review of bicycle-sharing service planning problems. Transportation Research Part C: Emerging Technologies 117, 102648.

Skellam, J., 1946. The frequency distribution of the difference between two Poisson variates belonging to different populations. Journal of the Royal Statistical Society: Series A (General) 109, 296.

Sohrabi, S., Paleti, R., Balan, L., Cetin, M., 2020. Real-time prediction of public bike sharing system demand using generalized extreme value count model. Transportation Research Part A: Policy and Practice 133, 325–336.

Strauba, M., Rudloffa, C., Grasera, A., Kloimüllnerb, C., Raidlb, G.R., Pajonesd, M., Beyere, F., 2018. Semi-automated location planning for urban bike-sharing systems. Proceedings of the 7th Transport Research Arena (TRA 2018) , 1–10.

Vogel, P., Mattfeld, D.C., 2011. Strategic and operational planning of bike-sharing systems by data mining – A case study, in: Böse, J.W., Hu, H., Jahn, C., Shi, X., Stahlbock, R., Voß, S. (Eds.), Computational Logistics, Springer, Berlin, Heidelberg. pp. 127–141.

Wu, Y.H., Kang, L., Hsu, Y.T., Wang, P.C., 2019. Exploring trip characteristics of bike-sharing system uses: Effects of land-use patterns and pricing scheme change. International journal of transportation science and technology 8, 318–331.

Yang, H., Zhang, Y., Zhong, L., Zhang, X., Ling, Z., 2020a. Exploring spatial variation of bike sharing trip production and attraction: A study based on Chicago's Divvy system. Applied Geography 115, 102130.

Yang, Y., Heppenstall, A., Turner, A., Comber, A., 2020b. Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems. Computers, Environment and Urban Systems 83, 101521.

Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., Moscibroda, T., 2016. Mobility modeling and prediction in bike-sharing systems, in: 14th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys), ACM. pp. 165–178.

Zhang, H., Zhuge, C., Jia, J., Shi, B., Wang, W., 2021. Green travel mobility of dockless bike-sharing based on trip data in big cities: a spatial network analysis. Journal of Cleaner Production 313, 127930.

Zhang, Z., Qian, C., Bian, Y., 2019. Bicycle–metro integration for the 'last mile': Visualizing cycling in Shanghai. Environment and Planning A: Economy and Space 51, 1420–1423.

Zhao, J., Wang, J., Deng, W., 2015. Exploring bikesharing travel time and trip chain by gender and day of the week. Transportation Research Part C: Emerging Technologies 58, 251–264.

Zheng, Z., Chen, Y., Zhu, D., Sun, H., Wu, J., Pan, X., Li, D., 2021. Extreme unbalanced mobility network in bike sharing system. Physica A: Statistical Mechanics and its Applications 563, 125444.

Zhou, X., 2015. Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago. PLOS ONE 10, e0137922.

Zhou, Y., Wang, L., Zhong, R., Tan, Y., 2018. A Markov chain based demand prediction model for stations in bike sharing systems. Mathematical Problems in Engineering , 8028714.

Appendix: Data Sources and Processing

The bike sharing trip data were obtained directly from the websites of the individual bike sharing providers or from city data portals. All of the datasets used contain data on every individual trip made in the network including trip duration, time of departure from the start station, start station ID, start station name, time of arrival on the end station, end station ID, and end station name. Not all cities provide the location of the stations in their trip data. For these cities, station data has been obtained from other official open data sources such as station occupancy APIs as shown in Table 1. For cities in the United States, the datasets also include the type of user which used the bicycle, primarily split between subscribers, who pay an annual fee to use the system for the whole year, and casual users, who pay for individual trips or to use the system for a short period of time (typically less than a week).

A great deal of care has to be taken when determining the span of time in which the service areas are calculated, since the number and locations of stations vary over time. For instance, in New York City, 938 unique stations have been used in the network over the year 2019. However, at no point in time have these 938 stations been used simultaneously, since some stations have been created, relocated and/or removed entirely. Thus, calculating 938 service areas will give an unrepresentative view of the network and how it was used. To account for this, we calculate the service areas of the network in each day of the year. The change in the service areas due to relocation and removal of stations affects other variables that we consider, such as the population around the station, land use, and distance to nearest transit points. To alleviate this, all variables used in the model for each station are calculated for each day the station has been used and then averaged over those days. This includes not only variables derived from the placement of the station and its service area, but also the daily number of trips at the station.

For US cities, land use data was obtained from zoning data provided by the city governments. The data contains polygons defining each zone, along with a corresponding zone code. We classify each zone as either residential, commercial, recreational, industrial or mixed, depending on the zone code and its stated use in the zoning ordinance. Since no historical zoning data were found, we use the most recent data provided by the cities as of April 2022. It is possible that the zoning has changed since 2019, but we assume that the changes in this time-frame were relatively minor and insignificant to the general ridership of the bike sharing networks.

For European cities, zoning data is not available in a standardized form, as land use regulations differ between areas. Instead, we use land use data from Urban Atlas 2018 in the Copernicus Land Monitoring Service provided by the European Environment Agency (EEA). This data includes polygons of different land areas, each with a designated zone code. The EEA provides a guide containing a description of each zone code EEA (2020). Each area was classified into one of the same land use groups used for the US cities by matching the provided description of each zone code. For instance, “urban fabric” zones were

classified as residential areas, since these areas are predominantly for residential use. We note that the land use data provided by the EEA have a lower resolution regarding the specific land use associated with each zone code. While areas described as urban fabric are predominantly residential, they also include central business areas and downtown areas with only partial residential use. The data also does not distinguish between commercial and industrial use. As above, we assume no significant land use changes occurred between 2018 and 2019.

For each station, we calculated the share of each type of land use within the service area of the station. The European land use data also contains polygons of the cities' road network. While the roads are a part of the stations' service areas, they were not included when calculating the share of land use within the service area. Historical census data for US cities in 2019 is provided by the United States Census Bureau on the census tract level, along with polygons of the census tracts. We used these data to calculate the population density of each census tract, measured in persons/100 m^2. For Helsinki, Oslo and Madrid, population estimates are provided for each polygon in the land use data from the Urban Atlas 2018. For London, population estimates from Urban Atlas 2012 were used instead due to discrepancies found in the estimates from Urban Atlas 2018. We calculated the population density of each station's service area as an average of the population densities of the census tracts or land use polygons within the service area, weighted by their share of the service area. Finally, locations of subway and railway stations as well as city centers were obtained using the Overpass API from OpenStreetMap.

The modeling of the traffic patterns needs to take several factors into account: firstly, the weekly cycle has a strong effect on user behavior, with distinct patterns on weekdays and weekends. Since weekday traffic is significantly more intense, with a correspondingly stronger impact on planning and management considerations, we only considered business days in our analysis. This also simplifies the comparison between different cities, as tourist and leisure traffic is much more unpredictable and strongly depends on individual landmarks and attractions, which are naturally different for different cities. Trips which started on a public holiday were also removed from the dataset. Furthermore, we excluded two more kinds of trips: loop trips, i.e., trips that had the same departure and arrival point, which are often recreational, as shown by Zhao et al. (2015), and trips taken by temporary users (in cities which have this distinction in the dataset), who Noland et al. (2019) argue are most likely tourists visiting the city for a short period. Finally, trips shorter than 60 seconds were considered as false starts or users ensuring that their bike is locked, so they are removed as well. We also removed stations which are suspected to be test stations or otherwise used for maintenance purposes, as well as stations that have a very low traffic (i.e., fewer than 8 daily trips counting both departures and arrivals), from our analysis. The number of trips and stations removed in our data processing can be seen in Table 7.

City	Pre-cleaning		Post-cleaning		Data Retained (%)	
	Trips	Stations	Trips	Stations	Trips	Stations
New York City	14869054	938	13168086	857	88.56	91.36
Chicago	2663558	593	2153584	369	80.85	62.23
Washington D.C.	2588852	429	2285881	333	88.30	77.62
Boston	1865013	335	1547643	254	82.98	75.82
London	7719768	788	7522951	784	97.45	99.49
Helsinki	2755144	348	2677641	348	97.19	100.00
Oslo	1729194	253	1682360	251	97.29	99.21
Madrid	3015679	213	2781463	213	92.23	100.00

Table 7: Number of trips and stations retained after removing low-traffic stations.