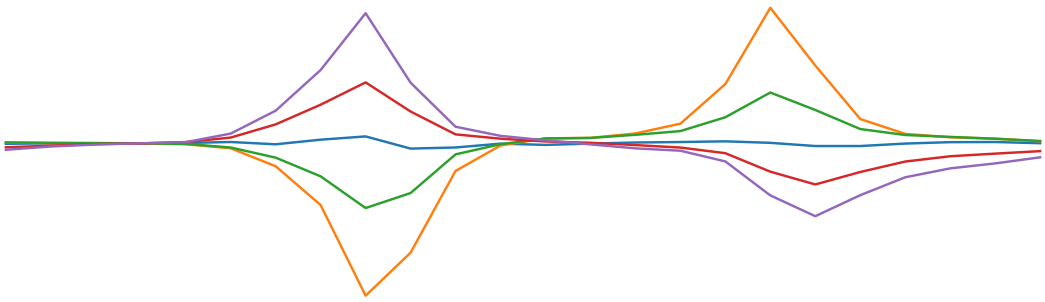
Prediction of Traffic Patterns in Bike Sharing Systems

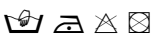
Master's Thesis

Daniel B. van Diepen & Nicolai A. Weinreich



Aalborg University
Mathematical Engineering

This project has been typeset in LaTeX with figures produced in TikZ and Matplotlib unless otherwise stated. Scripts have been made using Python 3.9. In case of stains, please note the washing guidance below.



AALBORG UNIVERSITY
STUDENT REPORT

Mathematical Engineering
Aalborg University
http://www.aau.dk

Title:
Prediction of Traffic Patterns in Bike Sharing Systems

Theme:
Data Analysis and Modelling

Project Period:
2021/2022

Project Group:
MATTEK10 1217a

Participants:
Daniel Bernard van Diepen
Nicolai André Weinreich

Supervisors:
Federico Chiariotti
Christophe Biscio
Petar Popovski

Copies: $\sqrt{-1}$

Number of Pages: 183

Date of Completion:
June 3, 2022

Abstract:

The aim of this project is to analyse patterns of usage in dock-based bike sharing systems in order to distinguish between different types of stations with the end goal of being able to predict the average daily traffic patterns of stations based on spatial factors in their service areas. The analysis is based on trip data from bike sharing systems in New York City, Chicago, Washington DC, Boston, London, Helsinki, Oslo, and Madrid as well as other data external to the systems.

In the analysis, different clustering algorithms are introduced to cluster stations based on the shape of their average daily traffic patterns. It is found that using k-means clustering with five clusters yielded clearly separate types of traffic patterns which are then related to external spatial factors using a logistic regression model. A strong relationship between station type and spatial factors is found for all cities, and variations between the models for different cities are related to differences in commuting culture between cities. Average bike share demand for each station is modelled using a generalised linear model with a logarithmic link function, and coupled with the logistic regression model it is possible to predict average traffic patterns with reasonable precision.

Finally, in a case study of the Citi Bike system expansion in autumn 2019, the demand model is used to optimise station placement.

Preface

This report is a product of the work of the authors in the 9th and 10th semesters of the master's program of mathematical engineering at Aalborg University spanning from September 2021 to June 2022. The authors would like to thank their supervisors whose encouragement and guidance throughout the project period has been invaluable in the development of the project. A special acknowledgement also goes out to the many governing bodies who have embraced and driven the open data movement and provided open resources which made this project possible.

As an aid to the project, the authors have also developed a dashboard application which provides interactive visualisations of the bike sharing systems studied in project and implements some of the core methods presented in the report. The source code of all relevant scripts used in relation to the project as well as a link to the interactive dashboard can be found at `https://github.com/cykelholdet/superbike`.

In conjunction with the project, the authors are planning to publish a paper which sums up the most important findings in the project. The paper is intended to be submitted to Transportation Research Part B. A working draft of the paper can be found in Appendix E.

The referencing style used is the alphabetical IEEE-method with specification of page numbers when relevant. Further information about the sources can be found in the bibliography.

The figures in this project have been made using the Ti*k*Z package in LaTeX, and using `matplotlib.pyplot` in Python 3.9.

Aalborg University, June 3, 2022

Daniel Bernard van Diepen
<dvandi17@student.aau.dk>

Nicolai André Weinreich
<nweinr17@student.aau.dk>

Contents

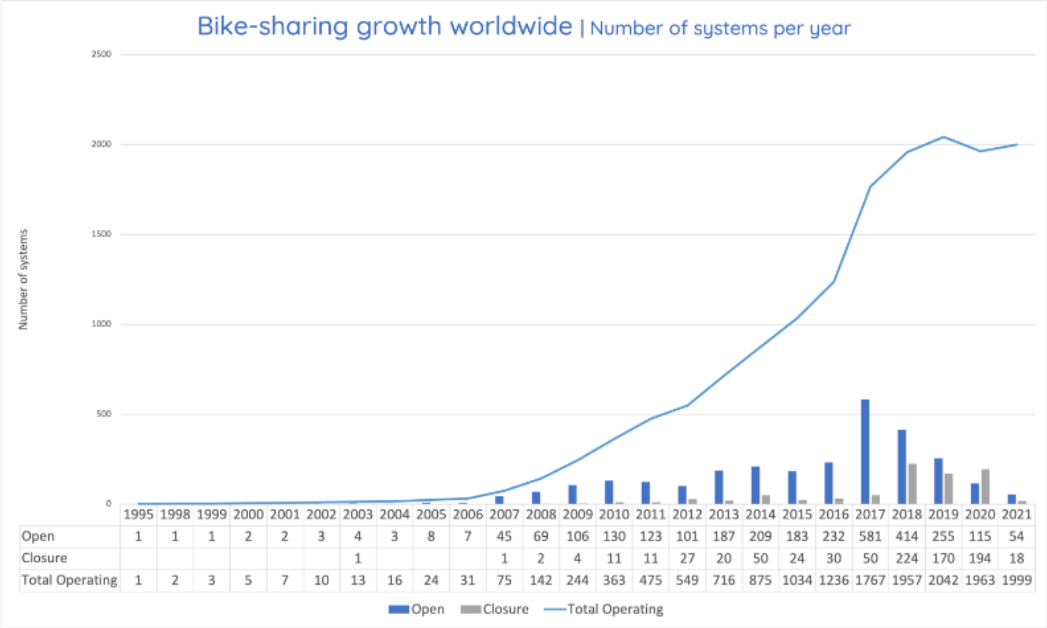
1. Introduction

In 2018, the United Nations projected that 6.7 billion people will be living in urban areas in 2050 compared to 4.2 billion in 2018 [UN]. One of the main challenges in keeping such an urban growth sustainable is developing the transportation infrastructure to meet the needs of the growing population. If an increased population leads to an increase in car-travel within the city, the road infrastructure will become more strained over time, even with heavy investments in its expansion ad maintenance. At the same time, private transportation of people and goods already contributes disproportionately to environmental problems such as noise and air pollution [Eura]. Therefore, as stipulated in the UN sustainable development goal SDG 11, it is essential to the sustainable development of cities that city residents have access to high quality public transport solutions [Uni]. By reducing car-dependence, cities can become a more healthy and livable environment, and by improving sustainable transit options, the mobility of city dwellers can be increased [Eura].

Sustainable urban mobility solutions come in many shapes and sizes, and each mode of transport represents a trade-off between throughput and speed on one hand and flexibility on the other. Where rail transport favors throughput and speed, and walking favors flexibility, there is a niche with higher speed compared to walking and higher flexibility compared to rail transport which is filled by the bicycle. Cycling provides a cheap, efficient and sustainable transportation solution, particularly when coupled with accommodating infrastructure. However, because cycling ordinarily relies on individual bike ownership it can sometimes be difficult to integrate with other mobility solutions. One way for cities to integrate cycling with other mobility solutions and simultaneously promote cycling as a sustainable alternative transport method is by providing access to a bike sharing system.

1.1 Bike Share Systems

The inception of bike sharing is widely attributed to the "White bikes" initiative of 1965 in Amsterdam which was a part the white plans aimed to address social problems and make the city more livable. While the white bikes plan was mainly used to provoke the establishment and promote discussion about social mobility, the plan served as an inspiration for future bike sharing programs.



	1995	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Open	1	1	1	2	2	3	4	3	8	7	45	69	106	130	123	101	187	209	183	232	581	414	255	115	54
Closure							1				1	2	4	11	11	27	20	50	24	30	50	224	170	194	18
Total Operating	1	2	3	5	7	10	13	16	24	31	75	142	244	363	475	549	716	875	1034	1236	1767	1957	2042	1963	1999

Figure 1.1: Growth in number of bike sharing systems worldwide. The figure has been taken from [DeM+21].

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. A recent report shows that around 2000 bike sharing systems are in operation with a global fleet size of more than 10 million bikes as of August 2021, see Fig. 1.1 [DeM+21].

While technological innovation has made it possible to invest more in bike-sharing system infrastructure, there are a number of factors which contribute to the growing popularity of bike sharing.

From the perspective of city planners, bike share systems pose an interesting solution as planners are faced with the demands of reducing emissions, decreasing congestion and increasing the mobility of city residents all while dealing with the constraints of a limited budget. A bike share system serves all of those demands while being many 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. [Dad12]

Among the users of the bike sharing system, an important reason listed is convenience. A well-functioning bike share system works well in tandem with other modes of public transport, and is easy to integrate into the users' daily commute. The pickup-and-go type system is convenient when covering 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 stations for public transport such as a metro system. This is sometimes called the first mile problem. Conversely, bike sharing can also serve as a last mile solution to cover the part of a trip from a public transport station to the final destination. [ZQB19; MS14; Yan+20b]

Research on the usage of bike share systems indicates that bike share trips mainly supplant walking and public transport trips, and only to a lesser degree replace car trips. 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 public transportation a more viable mobility solution for more people. [MS14]

Bike sharing systems can come in many flavours with the two most popular being dock-based systems where the bikes inter-lock with docking stations in fixed locations, and dock-less systems, also called free-floating systems, where bikes can be picked up and dropped freely within a designated area. Both types of systems have advantages and disadvantages both for the user but also for the operator maintaining the system. Dock-less systems give the user more freedom in where they can place the bike at the end of the trip, but this also means that bikes are more sporadically placed and can block sidewalks and create clutter in the streets, while having a nearby bike at the start of the commute may be less reliable. From the operator side, re-balancing the system i.e. moving bikes around to make them more accessible can be more costly when the bikes are not concentrated at set locations [PZ17]. In dock-based systems, it is only necessary to re-balance between the docking stations, but there is also a bigger planning aspect to this type of system in terms of where to put new stations. Modern bike share docks get some flexibility by being able to operate on solar power, and installation is therefore as easy as bolting it into the pavement [NYC09]. While this makes setting up new docking stations cheap and easy relative to other types of transportation methods, they are still an added cost and an added consideration compared to dock-less systems.

1.2 Bike Share Planning

For the users of a bike share system, there is no problem of punctuality or frequency like there is in a typical public transport system. However, a main performance metric which affects the users is whether docking stations are full at the destination, or empty at the origin. Therefore, it is important that the system is balanced with respect to how many bikes are at each station. There are three main dimensions in which unbalance in the bike share system can be addressed [VGM11]:

- Strategic decisions concerning the overall system network including the location and size of bike share stations.

- Tactical design considerations where balance is built into the system by e.g. providing incentives such as discounts to users who drop off their bike at stations low on bikes.
- Operational relocation of bicycles by the operator, e.g. using a truck.

This project will mainly concern planning at the strategic level.

When making planning decisions concerning the strategic level, either for a new bike share system or expanding an existing one, there are many factors which can influence the locations and dimensions of the new stations. In order to make the best possible decisions about locations, it can be useful to apply what was learned from existing systems. This can be done qualitatively by inspecting what works and what doesn't, as was done by the New York City Department of City Planning where they conducted a feasibility study for a bike share system wherein they looked at data from the Parisian bike share system Velib' to get a general idea what kind of traffic could be expected in a bike share system in a major city [New]. However, in order to get a more detailed picture of what type and volume of traffic can be expected in different parts of a city, it is also possible to learn from existing systems quantitatively. If detailed trip data is available from existing systems, this could be used to make an accurate prediction of the expected traffic at different locations across a city.

Traditionally, obtaining detailed traffic data could be difficult. However, in recent years cities around the world have increasingly made their data available for public use under open licensing terms. This open data movement has also been embraced by public and private service providers, including bike share providers who publish anonymised historical data on trips through their systems. The availability of open data has sparked a rise in popularity of using data-mining and data-based modelling of bike sharing systems to analyse behavioural patterns and inform decisions about urban planning.

1.3 Literature Review

Because of the popularity of bike share systems, there exists a wide body of studies concerning various aspects of bike share systems. One popular class of studies concerns the relationship between different social and environmental factors and bike share usage. For an extensive overview of these types of studies, we refer to the review paper by Eren and Uz [EU20]. Other papers concern topics such as environmental impact [ZM18] or system re-balancing strategies at any of the three planning levels [Chi+20].

Of particular relevance for this report are those pertaining to three specific types of bike sharing problems, namely clustering, statistical modelling of station demand,

and planning problems where optimisation methods are used to determine station placement.

1.3.1 Clustering

Clustering is a class of machine learning algorithms which are used for categorising a data set into a number of distinct subsets. In bike sharing, clustering is typically used to classify stations into different categories based on the types of trips they attract. Some stations might be the source of many commuters in the morning rush hour, while other stations may have more leisurely trips spread throughout the day. With the use of clustering methods, stations with similar temporal patterns can be identified, which can be useful both for bike share system planning, re-balancing as well as understanding urban mobility in general.

Different clustering methods have been applied on various bike share systems with the aim of identifying common patterns. Between 3 and 5 clusters were identified as an appropriate amount of clusters in the cities of Vienna [VM11], Chicago [Zho15] and Paris [FAZ17], while one study found 8 clusters to be appropriate [CO14].

Clusters with different types of traffic also commonly attract different types of users, as evidenced by the correlation between cluster type and various demographic data such as age and gender in Chicago in [Zho15], while [VM11] and [CO14] compared the cluster types with visual examinations of the station surroundings.

Different approaches to clustering of bike share traffic data include the determination of spatial communities of stations which are highly inter-connected [Bor+11], the clustering of station occupancy data into temporal occupancy patterns [SLM15], and the use of latent Dirichlet allocation to associate each station with one or more categories [Cô+14].

1.3.2 Traffic Prediction

Being able to effectively model and predict the traffic in a system can be crucial when planning to set up stations in a new area, whether establishing a whole new system or expanding an existing system. Traffic prediction is also important in terms of forecasting when specific stations need to be re-balanced such as to minimise customer grievances when there are no available bikes or no space to park one. Thus, in order to minimise customer dissatisfaction and thereby increase profitability, traffic prediction has been under intensive study in recent years.

One of the main goals of traffic prediction is being able to identify key factors which have a large impact on the demand [EU20]. The most simple models use historical data from a station in the bike sharing system to forecast the future demand of that station [Yan+16], although since this relies on prior data from the station, the application of this is restricted to re-balancing purposes. Being able to associate bike share usage with data external from the bike sharing system is far more interesting in terms of planning of future endeavours. There are two dimensions to this

problem: The spatial factors which relate bike share usage to the *where* and temporal factors which relates it to the *when*. Statistical regression has been shown to be effective in both associating bike share usage with static spatial features such as the land use of the surrounding area and city infrastructure [NSG19; NSG16; BB12] as well as demographic factors [Yan+20a], and temporal effects such as precipitation, temperature and wind speeds [XC20] or some combination of both [FIE16]. Another way to model station usage is by using a two-tier modelling approach, which first clusters the stations and then relates external data to each cluster [Hyl+18].

Dock-based systems also lend themselves well to graph-modelling with docking stations serving as nodes connected by the trips between them. Graph modelling has been shown to be useful for modelling the traffic flow between stations in order to predict demand [Li+19] or to aid more complicated models such as Long-Short Term Memory (LSTM) neural networks [Yan+20b]. While neural networks may not be geared to explain the relations between variables like conventional statistical models, they are proven effective at using both spatial and temporal data to predict short-term demand [Ma+22] as well a incorporating more advanced types of data such as traffic data from other public transportation systems [Zha+18].

1.3.3 Optimisation for Bike Share Planning

The location of stations is one of the essential bike share planning problems. By establishing a set of candidate station locations as well as a suitable objective function, it is possible to use optimisation methods to obtain a configuration of station locations which is in some sense optimal.

Factors which have been used in an objective function include maximising the coverage of pre-specified points of high demand and points of interest by counting the number of demand points covered by the bike share station configuration [CYI18; AB21], as well as maximising the coverage of bikeway segments and population centers [CMF18].

Alternative factors which have also been considered include maximising bicycle modal share in a simulated population using a genetic algorithm [Rom+12] and maximising the demand that is served within city zones [FR15]. Instead of maximising desirable factors, another option is minimising undesirable factors, for example travel distance to and from bike share stations to ensure that bike sharing is convenient for as many users as possible [PS17].

In addition to the objective function, there is a variety of different constraints which can be used to set up an optimisation problem. First and foremost is the budget constraint. This can be expressed in different ways, either as the number of stations to be placed, the number of bicycles distributed among stations or a combination of both [CYI18; FR15]. Other constraints primarily concern either coverage of a certain area such as being situated near a bikeway [AB21].

1.4 Problem Statement

In the literature, the concepts of clustering and statistical modelling of bike share trip data have been treated

However, there was found to be a certain lack of a combined analysis of traffic patterns using both clustering and modelling of demand. In addition, the analyses in most of the studies are restricted to a single city. It is the opinion of the authors that a discussion pertaining to the usability of traffic prediction models between cities could be informative in both about the generalisation ability of the models as well as to identify differences and similarities between cities in terms of bike share usage. In order to narrow the scope of the project, the systems which will be analysed will be entirely dock-based. The project will also be further limited to a spatial analysis of traffic patterns.

Based on these considerations, we have identified a problem statement which builds upon the literature by answering the following question:

How can spatial information from a city be used to predict the shape and volume of a station's average daily traffic pattern in a dock-based bike sharing system, and how does this differ between cities?

In order to answer the problem statement, we will first present theory about different clustering methods in Chapters 2 to 4, followed by measures for cluster validation in Chapter 5. Then, theory about generalised linear models and logistic regression will be shown in Chapters 6 and 7. Once the theory has been established, the bike share trip data is presented alongside the data from other sources in Chapter 8, while the approach for modelling the data is shown in Chapter 9. Theory and data meet in practice in Chapters 10 to 12. Finally, a case study on the expansion of the bike share system in New York City can be found in Chapter 13

2. K-means

Out of the many algorithms that may be used to classify a set of data points, one of the most popular algorithms is the k-means algorithm. The k-means algorithm is a partitioning algorithm, i.e. the goal of the algorithm is to obtain a partition of a data set into k groups or clusters such that an objective function is extremised. While the k-means algorithm is somewhat elementary, its ease of implementation, low computational complexity and convergent property makes it a very attractive and frequently used clustering method.

The k-means algorithm is a partitional algorithm meaning that the goal of the algorithm is to partition the data set $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ of d-dimensional data points into k clusters $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ such that [KR05, p. 38]

1. $\mathcal{C}_j \neq \emptyset, \quad j = 1, \ldots, k,$
2. $\bigcup_{j=1}^{k} \mathcal{C}_j = \mathcal{D},$
3. $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \quad i \neq j.$

One simple way of partitioning $\mathcal{D}$ is by defining k representatives $\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_k$, one for each cluster, and then assigning each data point to the cluster with the nearest representative [Bis06, p. 424]. This in turn introduces the problem of how to determine these representatives. Randomly choosing data points to act as representatives will probably not yield good clustering. However, iteratively updating the representatives based on the data points in their cluster may yield better choices and better clustering over time. This is the idea behind the k-means algorithm.

The k-means algorithm is classically derived by minimising the Sum-of-Squares Error (SSE) objective function. Before minimising, we define the indicator variables r_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ as binary variables which indicate the cluster which $\boldsymbol{x}_i \in \mathcal{D}$ is assigned to. More formally, we define [Bis06, p. 424]

$$r_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in \mathcal{C}_j, \\ 0, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n \quad j = 1, \ldots, k. \tag{2.1}$$

The objective function to be minimised is then

$$E = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|\boldsymbol{x}_i - \boldsymbol{m}_j\|_2^2 \tag{2.2}$$

representing the sum of squared euclidean distances between each data point and the representative of the cluster which the data point is assigned to. The objective function needs to be minimised with respect to both the indicator variables r_{ij} and the representatives $\boldsymbol{m}_j$. In practice, some data points are chosen as the initial $\boldsymbol{m}_j$. Subsequently, E is minimised iteratively with each step consisting of two minimisations; one with respect to the r_{ij} keeping the $\boldsymbol{m}_j$ fixed, and one with respect to the $\boldsymbol{m}_j$ keeping the r_{ij} fixed [Bis06, p. 425].

The minimisation of E with respect to the r_{ij} is relatively straightforward. Note that the terms in Eq. (2.2) involving different i are independent, meaning we can minimise for each i separately. Thus, for $i = 1, \ldots, n$ the $\boldsymbol{m}_j$ which minimises the distance $\|\boldsymbol{x}_i - \boldsymbol{m}_j\|_2$ is simply the one closest to $\boldsymbol{x}_i$. The r_{ij} which minimise E are therefore the ones for which [Bis06, p. 425]

$$r_{ij} = \begin{cases} 1, & \text{if } j = \underset{1 \leq l \leq k}{\arg\min} \|\boldsymbol{x}_i - \boldsymbol{m}_l\|_2, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

Consider the minimisation of E with respect to the $\boldsymbol{m}_j$ keeping the r_{ij} fixed. Note again that the terms in Eq. (2.2) with different j are independent from each other. Thus, for a specific j we minimise the sum

$$\sum_{i=1}^{n} r_{ij} \|\boldsymbol{x}_i - \boldsymbol{m}_j\|_2^2 \tag{2.4}$$

which is a quadratic function of $\boldsymbol{m}_j$. Setting the derivative of Eq. (2.4) to zero yields [Bis06, p. 425]

$$2 \sum_{i=1}^{n} r_{ij} (\boldsymbol{x}_i - \boldsymbol{m}_j) = \boldsymbol{0} \tag{2.5}$$

which, when solving for $\boldsymbol{m}_j$, yields

$$\boldsymbol{m}_j = \frac{\sum_i r_{ij} \boldsymbol{x}_i}{\sum_i r_{ij}}. \tag{2.6}$$

Note that $\sum_i r_{ij} = |\mathcal{C}_j|$ by the definition of r_{ij}. Thus, Eq. (2.6) is simply calculated as taking the average over all $\boldsymbol{x}_i \in \mathcal{C}_j$.

The two steps of calculating Eq. (2.3) and Eq. (2.6) are repeated until some convergence criterion has been fulfilled. One possible criterion may be to define an $\varepsilon > 0$ and stop when $\|\boldsymbol{m}_j^{old} - \boldsymbol{m}_j^{new}\|_2 < \varepsilon$ for two consecutive iterations of $\boldsymbol{m}_j$. Another criterion may be to set a threshold on the number of reassignments of r_{ij} by Eq. (2.3). The algorithm has been well documented to converge, and we will also

provide a justification in Chapter 3. However, global convergence is not guaranteed. A pseudo-code of the k-means algorithm involving the discussed steps can be seen in Algorithm 1.

Algorithm 1 k-means

Input: dataset: $\mathcal{D} = \{\boldsymbol{x}_i\}$, number of clusters: k
Initialise: $\boldsymbol{m}_j, j = 1, \ldots, k$
while not converged **do**
  for $i = 1, \ldots, n$ **do**
    Find $J = \underset{1 \leq j \leq k}{\operatorname{argmin}} \|\boldsymbol{x_i} - \boldsymbol{m}_j\|_2$
    Assign $\boldsymbol{x}_i$ to $\mathcal{C}_J$
  end for
  Update $\boldsymbol{m}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\boldsymbol{x} \in \mathcal{C}_j} \boldsymbol{x}, \quad j = 1, \ldots, k$
end while
Output: k clusters

One critical aspect of the k-means algorithm is the initial choice of the representatives $\boldsymbol{m}_j$. Typically, the first representatives are chosen randomly as either some already existing points from the data set or as new randomly generated data points. However, a preliminary search for good initial representatives can result in fewer iterations being necessary for convergence as well as to avoid some local minima.

While the k-means algorithm is simple to implement and has a low complexity, its efficiency is limited by the type of data to be clustered. For instance, the algorithm performs well for data in well-separated and spherical clusters but if the data clusters are elongated rather than spherical, the performance of the algorithm drops significantly. This is shown in Fig. 2.1 where we have generated two types of data and performed the k-means algorithm with initial representatives drawn randomly from the data set. Both types of data contain two clusters of normally distributed data points, but one data type is more elongated than the other. Despite the clusters being clearly separated, the algorithm performs poorly for the data set containing elongated clusters.

Another limitation of the algorithm is its use of the l_2-norm as a measure of dissimilarity between points. This again limits the type of data which is appropriate to be used in the algorithm, and data where a different dissimilarity measure may be appropriate is not guaranteed to be clustered correctly. However, by introducing a more general dissimilarity measure $d(\cdot, \cdot)$ we can generalise the k-means algorithm by minimising the more general objective function [Bis06, p. 428]

$$\tilde{E} = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} d(\boldsymbol{x}_i, \boldsymbol{m}_j). \tag{2.7}$$

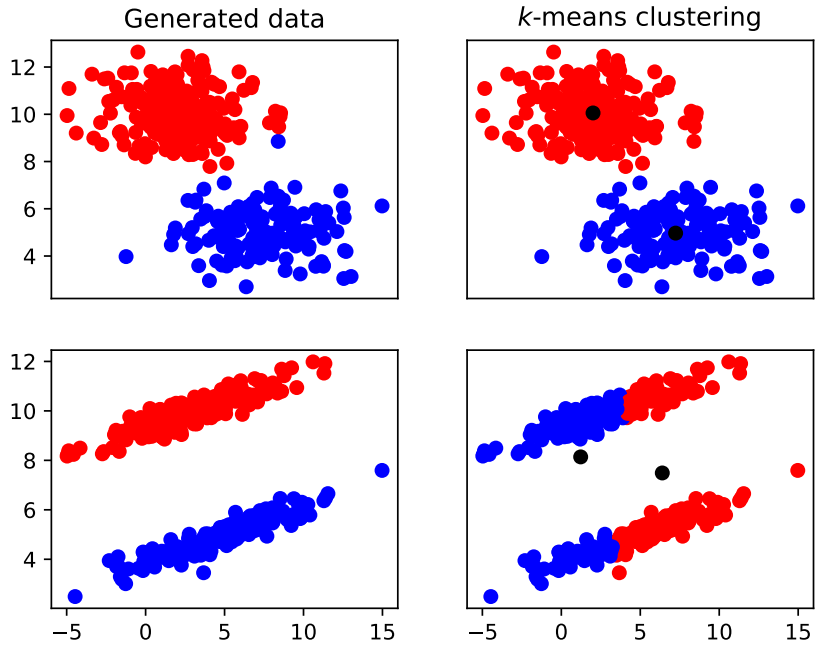


Figure 2.1: Test of the k-means algorithm using randomly generated data for $k = 2$. The representative vectors are shown in black.

The minimisation of Eq. (2.7) with respect to the r_{ij} is the same as before. Each data point should be assigned to the cluster with the most similar representative. However, the minimisation of Eq. (2.7) with respect to the $\boldsymbol{m}_j$ is more complex for a general choice of dissimilarity measure and closed form solutions may be difficult to determine [Bis06, p. 428]. In the absence of a closed form solution for a new representative $\boldsymbol{m}_j$, a discrete search over all points $\boldsymbol{x} \in \mathcal{C}_j$ may be made for this representative. Thus, instead of considering $\boldsymbol{m}_j \in \mathbb{R}^d$ like in the k-means algorithm, we consider only points from the data set as valid representatives i.e. $\boldsymbol{m}_j \in \mathcal{D}$ [Bis06, p. 428]. These representatives are commonly called medoids and thus the algorithm for finding them is called the k-medoids algorithm.

2.1 k-medoids

The k-medoids algorithm consists of two phases [KR05, pp. 102-104]. In the first phase, called the BUILD phase, the initial medoids are determined. The first medoid is chosen such that Eq. (2.7) is minimised for $k = 1$. Thus, the first medoid is the data point which is the most centered in the data set. The subsequent medoids are then chosen such that Eq. (2.7) is further minimised while keeping already found medoids fixed. This is continued until k initial medoids are found.

In the second phase, called the SWAP phase, a swap of each medoid with another data point is considered. For a medoid $\boldsymbol{m}_j$ representing the cluster $\mathcal{C}_j$, the cost function is evaluated for all possible swaps of $\boldsymbol{m}_j$ with the points $\boldsymbol{x}_i \in \mathcal{C}_j$. This is then done for every medoid and the pair $(\boldsymbol{x}_i, \boldsymbol{m}_j)$ which yields the greatest reduction in cost will be swapped such that $\boldsymbol{x}_i$ is the new medoid for $\mathcal{C}_j$. Further swaps are then done until some convergence criterion is fulfilled. A pseudo-code of the k-medoids algorithm can be seen below.

Algorithm 2 k-medoids

Input: data set: $\mathcal{D} = \{\boldsymbol{x}_i\}$, number of clusters: k
for $j = 1, \ldots, k$ **do**
  Choose $\boldsymbol{m}_j \in \mathcal{D}$ such that Eq. (2.7) is minimised
end for
for $i = 1, \ldots, n$ **do**
  Find $J = \underset{j=1,\ldots,k}{\operatorname{argmin}}\, d(\boldsymbol{x}_i, \boldsymbol{m}_j)$
  Assign $\boldsymbol{x}_i$ to $\mathcal{C}_J$
end for
Calculate current cost $\tilde{E}$
while $\tilde{E}$ decreases **do**
  for $j = 1, \ldots, k$ and all $\boldsymbol{x}_i \in \mathcal{C}_j$ with $\boldsymbol{x}_i \neq \boldsymbol{m}_j$ **do**
    Consider the pair $(\boldsymbol{x}_i, \boldsymbol{m}_j)$
    Calculate the cost E' if $\boldsymbol{m}_j \leftarrow \boldsymbol{x}_i$
  end for
  Choose the pair $(\boldsymbol{x}_i, \boldsymbol{m}_j)$ with the smallest E'
  $\boldsymbol{m}_j \leftarrow \boldsymbol{x}_i$
  $\tilde{E} \leftarrow E'$
  Reassign $\boldsymbol{x}_i, \quad i = 1, \ldots, n$
end while
Output: k clusters

As with the k-means algorithm, the k-medoids algorithm is best suited for spherical and well separated data. The SWAP phase also introduces a higher computational complexity since it evaluates the cost of every possible swap. However, the k-medoids algorithm does have advantages over the k-means algorithm. Other than being more applicable for general dissimilarity measures, the k-medoids algorithm is also more robust against outliers due to the objective function being a sum of dissimilarities rather than a sum of squares [KR05, p. 117].

3. Expectation Maximisation

One of the features of partitional algorithms is the fact that each data point is assigned to one and only one cluster. This hard classification of data points may be appropriate for points which are close to the centers of the clusters, but for points between clusters the assignment to one cluster over another may seem more arbitrary. Instead of assigning each data point to a single cluster, we can assign the point k values, one for each cluster, which indicate how much the data point belongs to each cluster in some sense. This type of classification is usually called fuzzy classification [DHS01, p. 1922]. In this chapter, we will provide a clustering algorithm which provides this kind of classification.

Recall that the Gaussian distribution over a vector $\boldsymbol{x} \in \mathbb{R}^d$ is defined as [Bis06, p. 25]

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{3.1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is the covariance matrix assumed to be positive semidefinite and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. Suppose that we are given a data set $\boldsymbol{X} = [\boldsymbol{x}_1 \quad \boldsymbol{x_2} \quad \cdots \quad \boldsymbol{x}_n]^T$ where each data point $\boldsymbol{x}_i$ is assumed to be drawn independently from a Gaussian distribution. Through maximum likelihood estimation, estimates of the parameters of the distribution can be found to be [Bis06, pp. 93-94]

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \tag{3.2}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T. \tag{3.3}$$

However, suppose that we are given a data set like in Fig. 3.1. By visual inspection, we see that the most of the data is concentrated in two clusters. If we assume that the data points are drawn from the same Gaussian distribution as before, modelling of this data set will fail to capture this clustered structure. Rather than assuming that each data point is drawn from a single Gaussian distribution, we can assume that the distribution is instead a mixture or superposition of multiple Gaussians.

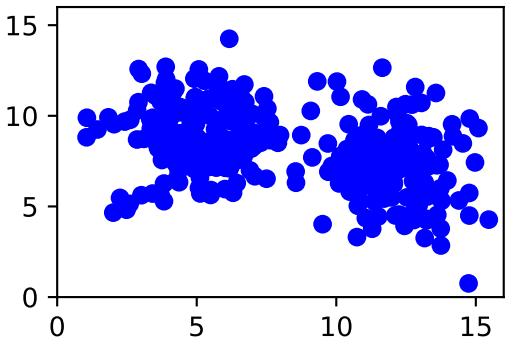


Figure 3.1: Clusterered data set.

Thus, we will assume that the data points are drawn independently from a Gaussian mixture model with its density given by [Bis06, p. 111]

$$p(\boldsymbol{x}) = \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{3.4}$$

where k is the given number of densities in the Gaussian mixture. In this context, the densities $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ are called the components of the distribution while the parameters π_j are called the mixing coefficients. Note that by integrating both sides of Eq. (3.4) we find that

$$\sum_{j=1}^{k} \pi_j = 1. \tag{3.5}$$

Since we also require that $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \geq 0$ and $p(\boldsymbol{x}) \geq 0$, we see from Eq. (3.4) that $\pi_j \geq 0$ for all j. Combining this with Eq. (3.5) yields that

$$0 \leq \pi_j \leq 1, \quad j = 1, \ldots, k. \tag{3.6}$$

Thus, the mixing coefficients can be seen as probabilities. Indeed, the marginal density of $\boldsymbol{x}$ is given by [Bis06, p. 112]

$$p(\boldsymbol{x}) = \sum_{j=1}^{k} p(j) p(\boldsymbol{x}|j) \tag{3.7}$$

which is equivalent to Eq. (3.4) with $\pi_j = p(j)$ and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sim p(\boldsymbol{x}|j)$. Therefore, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ can be interpreted as the density of $\boldsymbol{x}$ given the component j while π_j is the probability of picking that component.

We will introduce a k-dimensional binary random variable $\boldsymbol{z} = [z_1 \quad z_2 \quad \cdots \quad z_k]^T$ where a particular entry z_j is equal to 1 and all other entries are zero. Thus, there are k possible states of $\boldsymbol{z}$ according to the placement of the non-zero entry. We will

call these possible variables latent variables. We will define the marginal distribution of $\boldsymbol{z}$ using the mixing coefficients such that [Bis06, p. 430]

$$p(z_j = 1) = \pi_j = p(j), \quad j = 1, \ldots, k \tag{3.8}$$

which is equivalent to

$$p(\boldsymbol{z}) = \prod_{j=1}^{k} \pi_j^{z_j}. \tag{3.9}$$

The conditional density of $\boldsymbol{x}$ given a particular value for $\boldsymbol{z}$ is given by

$$p(\boldsymbol{x}|z_j = 1) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{3.10}$$

which can also be written as

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{j=1}^{k} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)^{z_j}. \tag{3.11}$$

We note that by using Eq. (3.9) and Eq. (3.11) we obtain [Bis06, p. 431]

$$\begin{aligned}
p(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) && (3.12)\\
&= \sum_{\boldsymbol{z}} p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z}) && (3.13)\\
&= \sum_{\boldsymbol{z}} \prod_{j=1}^{k} \pi_j^{z_j} \prod_{l=1}^{k} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)^{z_l} && (3.14)\\
&= \sum_{\boldsymbol{z}} \prod_{j=1}^{k} \left(\pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right)^{z_j} && (3.15)\\
&= \pi_1 \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \cdots + \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) && (3.16)\\
&= \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) && (3.17)
\end{aligned}$$

which is equivalent to Eq. (3.4). Thus, we can formulate the Gaussian mixture model using the latent variables. Since we have formulated the marginal probability $p(\boldsymbol{x})$ using the joint probability $p(\boldsymbol{x}, \boldsymbol{z})$, it follows that each data point $\boldsymbol{x}_i$ has a corresponding latent variable $\boldsymbol{z}_i$ [Bis06, p. 431]. This provides an intuitive interpretation. One of the Gaussians in the mixture must be responsible for the generation of $\boldsymbol{x}_i$ and the latent variable $\boldsymbol{z}_i$ provides that information. An obvious question would be: Given a data point $\boldsymbol{x}$ is it possible to determine its corresponding latent variable and thus which Gaussian it came from?

The key to answering this question lies in the conditional probability of $\boldsymbol{z}$ given $\boldsymbol{x}$. We will use $\gamma(z_j)$ to denote $p(z_j = 1|\boldsymbol{x})$ which can be determined using Bayes' theorem

$$\gamma(z_j) = p(z_j = 1|\boldsymbol{x}) \tag{3.18}$$
$$= \frac{p(z_j = 1)p(\boldsymbol{x}|z_j = 1)}{p(\boldsymbol{x})} \tag{3.19}$$
$$= \frac{p(z_j = 1)p(\boldsymbol{x}|z_j = 1)}{\sum_{l=1}^{k} p(z_l = 1)p(\boldsymbol{x}|z_l = 1)} \tag{3.20}$$
$$= \frac{\pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}. \tag{3.21}$$

We note that $\gamma(z_j)$ provides the posterior probability that $\boldsymbol{x}$ has been generated by the j'th component, while π_j is the prior probability. We will call $\gamma(z_j)$ the responsibility that the j'th component takes for explaining $\boldsymbol{x}$.

For a given data set $\boldsymbol{X}$, we will denote the corresponding set of latent variables as the $n \times k$ matrix $\boldsymbol{Z} = [\boldsymbol{z}_1 \quad \boldsymbol{z}_2 \quad \cdots \quad \boldsymbol{z}_n]^T$. Recall that the data points in $\boldsymbol{X}$ are i.i.d. In order to estimate the parameters of the mixed Gaussian distribution we aim to maximise the log-likelihood function given by [Bis06, p. 433]

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \Big\{ \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \Big\} \tag{3.22}$$
$$= \sum_{i=1}^{n} \ln \Big\{ \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \Big\}. \tag{3.23}$$

Before maximising the log-likelihood we note that the derivative of Eq. (3.1) with respect to $\boldsymbol{\mu}$ is

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \frac{\partial}{\partial \boldsymbol{\mu}} \exp\Big\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \Big\} \tag{3.24}$$
$$= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{\partial}{\partial \boldsymbol{\mu}} \Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \Big) \tag{3.25}$$
$$= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{\partial}{\partial \boldsymbol{\mu}} \Big(-\frac{1}{2}(\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - 2\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) \Big) \tag{3.26}$$
$$= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Big(-\frac{1}{2}(2\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - 2\boldsymbol{\Sigma}^{-1} \boldsymbol{x}) \Big) \tag{3.27}$$
$$= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{3.28}$$

Using Eq. (3.1), Eq. (3.21), and Eq. (3.28), the derivative of Eq. (3.23) with respect to $\boldsymbol{\mu}_j$ is

$$\frac{\partial}{\partial \boldsymbol{\mu}_j} \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\mu}_j} \ln \left\{ \sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \right\} \tag{3.29}$$

$$= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\mu}_j} \sum_{m=1}^{k} \pi_m \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{3.30}$$

$$= \sum_{i=1}^{n} \frac{\pi_j \frac{\partial}{\partial \boldsymbol{\mu}_j} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{3.31}$$

$$= \sum_{i=1}^{n} \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j) \tag{3.32}$$

$$= \sum_{i=1}^{n} \gamma(z_{ij}) \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j) \tag{3.33}$$

where $\gamma(z_{ij}) = p(z_j = 1|\boldsymbol{x}_i)$. Setting Eq. (3.33) to zero and multiplying by $\boldsymbol{\Sigma}_j$ yields

$$\boldsymbol{0} = \sum_{i=1}^{n} \gamma(z_{ij}) \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j) \tag{3.34}$$

$$\boldsymbol{0} = \sum_{i=1}^{n} \gamma(z_{ij})(\boldsymbol{x}_i - \boldsymbol{\mu}_j) \tag{3.35}$$

$$\boldsymbol{\mu}_j \sum_{i=1}^{n} \gamma(z_{ij}) = \sum_{i=1}^{n} \gamma(z_{ij}) \boldsymbol{x}_i \tag{3.36}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n_j} \sum_{i=1}^{n} \gamma(z_{ij}) \boldsymbol{x}_i \tag{3.37}$$

where $n_j = \sum_{i=1}^{n} \gamma(z_{ij})$ can be seen as the effective number of points in the j'th cluster. Thus, the mean of the j'th cluster is estimated by a weighted average of the data points with the responsibilities of each cluster being the weights. Through a similar process, the maximisation of Eq. (3.23) with respect to $\boldsymbol{\Sigma}_j$ can be found to yield [Bis06, p. 436]

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{n_j} \sum_{i=1}^{n} \gamma(z_{ij})(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \tag{3.38}$$

which has the same form as Eq. (3.3) except with each data point being weighted with the responsibility of its cluster.

When maximising Eq. (3.23) with respect to the mixing coefficients π_j, we note the constraint in Eq. (3.5) which needs to be taken into account. Thus, we aim to maximise

$$L = \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{j=1}^{k} \pi_j - 1 \right) \tag{3.39}$$

where λ is a Lagrange multiplier. The derivative of L with respect to π_j is

$$\frac{\partial}{\partial \pi_j} L = \frac{\partial}{\partial \pi_j} \left\{ \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{j=1}^{k} \pi_j - 1 \right) \right\} \tag{3.40}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \pi_j} \ln \left\{ \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right\} + \lambda \tag{3.41}$$

$$= \sum_{i=1}^{n} \left(\frac{\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \right) + \lambda. \tag{3.42}$$

Setting Eq. (3.42) to zero, multiplying by π_j and summing over j yields

$$0 = \sum_{i=1}^{n} \left(\frac{\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \right) + \lambda \tag{3.43}$$

$$= \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \right) + \lambda \sum_{j=1}^{k} \pi_j \tag{3.44}$$

$$= n + \lambda \tag{3.45}$$

which gives $\lambda = -n$. Substituting this into Eq. (3.43) and rearranging yields

$$n = \sum_{i=1}^{n} \frac{\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{3.46}$$

$$n\pi_j = \sum_{i=1}^{n} \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{3.47}$$

$$n\pi_j = \sum_{i=1}^{n} \gamma(z_{ij}) \tag{3.48}$$

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \gamma(z_{ij}) = \frac{n_j}{n}. \tag{3.49}$$

Thus, the j'th mixing coefficient is found by taking the average responsibility that the j'th cluster takes for each data point.

We note that the estimates of the parameters in Eqs. (3.37), (3.38), and (3.49) all require an estimate of the posterior probabilities which in turn is calculated using the parameters. While the estimates are not in a closed form, they do suggest an algorithm which iteratively estimates the responsibilities and parameters of the mixed Gaussian. By initialising the parameters of the mixed Gaussian, we can estimate the responsibilities by using Eq. (3.21) and then use these estimates to re-estimate the parameters of the density. This will then be repeated until some convergence criterion is fulfilled. Typically, the convergence criterion is either based on the change of the parameters between iterations or the change in log-likelihood given by Eq. (3.23).

The resulting algorithm is called the Expectation-Maximisation (EM) algorithm. The estimation of the responsibilities is called the E step while the estimation of the parameters is called the M step. A pseudo-code of the algorithm containing these steps is given in Algorithm 3.

Algorithm 3 Expectation-Maximisation

Input: data set: $\boldsymbol{X}$, number of clusters: k
Initialise: $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$ and π_j for $j = 1, \ldots, k$
while not converged **do**
    for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ **do**
        Evaluate the responsibilities:

$$\gamma(z_{ij}) \leftarrow \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{3.50}$$

    end for
    for $j = 1, \ldots, k$ **do**
        Re-estimate the parameters:

$$n_j \leftarrow \sum_{i=1}^{n} \gamma(z_{ij}) \tag{3.51}$$

$$\boldsymbol{\mu}_j \leftarrow \frac{1}{n_j} \sum_{i=1}^{n} \gamma(z_{ij}) \boldsymbol{x}_i \tag{3.52}$$

$$\boldsymbol{\Sigma}_j \leftarrow \frac{1}{n_j} \sum_{i=1}^{n} \gamma(z_{ij}) (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \tag{3.53}$$

$$\pi_j \leftarrow \frac{n_j}{n} \tag{3.54}$$

    end for
end while
Output: Responsibilities $\gamma(z_{ij})$

One of the disadvantages of the EM algorithm is the amount of calculations required in each iteration. The algorithm also requires more iterations before convergence compared to less complex methods like the k-means algorithm. However, this relatively slow convergence time can be alleviated by initialising the algorithm with good approximations of the parameters. For instance, it is typical in practice to initialise the $\boldsymbol{\mu}_j$ using means obtained from a prior application of the k-means algorithm [Bis06, p. 438].

An application of the EM algorithm on the data set in Fig. 3.1 can be seen in Fig. 3.2. In the figure, we have coloured each data point according to the responsibility of each of the two components. For instance, if a data point $\boldsymbol{x}_i$ has the corresponding responsibilities $\gamma(z_{i1}) = 0.5$ and $\gamma(z_{i2}) = 0.5$, i.e. the point is in the

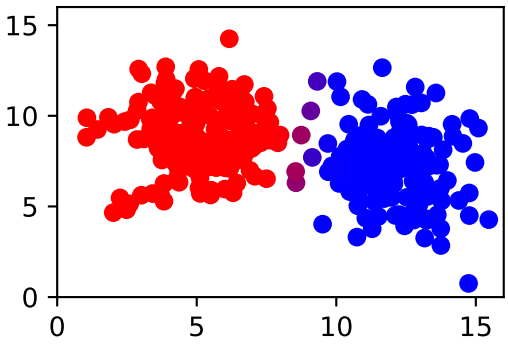


Figure 3.2: Classification of data set in Fig. 3.1 using the EM algorithm.

middle of the two components, the point is coloured with equal amounts of blue and red colour and thus appears as purple. This highlights one of the advantages of the EM algorithm over other partitional algorithms such as the k-means algorithm which provide a hard classification of each data point. Using a more fuzzy classification like the EM algorithm yields greater insights into the uncertainty of the classifications of edge cases.

One of the main advantages to the EM algorithm is its broad applicability. We have derived the algorithm assuming a mixed Gaussian density but the same derivation can also be done for other probabilistic models. For the rest of this chapter, we will discuss the EM algorithm in a more general setting and justify the convergence of the algorithm. While doing so, we will develop the primary steps of the EM algorithm for a general probabilistic model of the data set.

Recall that we denote the set of all observed data points as $\boldsymbol{X}$ and the set of corresponding latent variables as $\boldsymbol{Z}$. Suppose that we know the corresponding latent variable of each data point. We will then denote the set $\{\boldsymbol{X}, \boldsymbol{Z}\}$ as the complete data set. In practice, $\boldsymbol{Z}$ is not known and we will thus call $\boldsymbol{X}$ the incomplete data set.

Consider a probabilistic model which is parameterised by the parameter vector $\boldsymbol{\theta}$. The aim is to maximise the likelihood function

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}). \tag{3.55}$$

We will assume that the maximisation of $p(\boldsymbol{X}|\boldsymbol{\theta})$ is difficult while the maximisation of $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is significantly easier. We will also introduce a distribution over the latent variables and denote it as $q(\boldsymbol{Z})$.

We will turn our attention to the log-likelihood $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ and make the following decomposition

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) + \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) \tag{3.56}$$

$$- \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) + \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \tag{3.57}$$

$$= \sum_{\boldsymbol{Z}} \Big(q(\boldsymbol{Z}) \ln p(\boldsymbol{X}|\boldsymbol{\theta}) + q(\boldsymbol{Z}) \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) - q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) \Big) \tag{3.58}$$

$$- \sum_{\boldsymbol{Z}} \Big(q(\boldsymbol{Z}) \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) - q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) \Big) \tag{3.59}$$

$$= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \Big(\ln p(\boldsymbol{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) - \ln q(\boldsymbol{Z}) \Big) \tag{3.60}$$

$$- \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \Big(\ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) - \ln q(\boldsymbol{Z}) \Big) \tag{3.61}$$

$$= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}|\boldsymbol{\theta}) p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{3.62}$$

$$= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{3.63}$$

$$= \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p) \tag{3.64}$$

where [Bis06, p. 450]

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{3.65}$$

$$KL(q||p) = - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}. \tag{3.66}$$

Note that the decomposition in Eq. (3.64) holds for any choice of $q(\boldsymbol{Z})$. The decomposition in Eq. (3.64) can be used to define the EM algorithm in general and to show that the algorithm does indeed converge. We first note that Eq. (3.66) is the Kullback-Leibler distance between $q(\boldsymbol{Z})$ and the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ which satisfies that $KL(q||p) \geq 0$ with equality if and only if $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$. Thus, from Eq. (3.64) we have that $\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\boldsymbol{X}|\boldsymbol{\theta})$ i.e. $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound of the log-likelihood function.

Suppose that we have a current value of $\boldsymbol{\theta}$ which we denote $\boldsymbol{\theta}^{old}$. In the E step of the algorithm we aim to maximise $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ with respect to $q(\boldsymbol{Z})$. Since $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{old})$ does not depend on $q(\boldsymbol{Z})$ we see from Eq. (3.64) that $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ is maximised when $KL(q||p) = 0$ and will in fact be equal to $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{old})$. This is obtained by setting $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$. In the M step, the lower bound is maximised with respect to the parameters $\boldsymbol{\theta}$ while holding $q(\boldsymbol{Z})$ fixed yielding new parameter estimates $\boldsymbol{\theta}^{new}$. Unless $\mathcal{L}(q, \boldsymbol{\theta})$ is already maximised, its maximisation will lead to an increase in $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$. Substituting $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$ into Eq. (3.65) yields [Bis06, p. 452]

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})} \right\} \tag{3.67}$$

$$= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \tag{3.68}$$

$$= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) \tag{3.69}$$

$$= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) + \text{const.} \tag{3.70}$$

where the constant is the entropy of the q distribution and thus does not depend on $\boldsymbol{\theta}$ and

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = E_{\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}} [\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] \tag{3.71}$$

is the expectation of the complete-data log-likelihood $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ with respect to the latent variables conditioned on the data and the previous parameter estimations [Bis06, p. 452]. Thus, when we are maximising $\mathcal{L}(q, \boldsymbol{\theta})$ we are actually maximising the expectation $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ with respect to the parameters $\boldsymbol{\theta}$ which also justifies the name of the algorithm.

Since $q(\boldsymbol{Z})$ is held fixed and was determined using the old parameter values we have that $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \neq p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{new})$. Thus, $KL(q||p) > 0$ and by Eq. (3.64) we see that the increase in $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ is greater than the increase in the lower bound. However, setting $\boldsymbol{\theta}^{old} = \boldsymbol{\theta}^{new}$ and repeating the E and M step iteratively will lead to a repeating maximisation of the lower bound and thus also the log-likelihood until convergence. More formally, if $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^*)$ is the maximum of the log-likelihood then for any $\varepsilon > 0$ there exists an iteration number such that $|\ln p(\boldsymbol{X}|\boldsymbol{\theta}^*) - \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{new})| < \varepsilon$. However, it is important to note that the EM algorithm does not guarantee a global maximum of the log-likelihood. We will also note that the k-means algorithm is a particular case of the EM algorithm (see [Bis06, p. 441-444]), meaning that the previous discussion on convergence also applies to k-means.

We will end by compiling the steps discussed above into a more general EM algorithm.

Algorithm 4 General Expectation-Maximisation

Input: data set: $\boldsymbol{X}$

1: Initialise: initial parameters $\boldsymbol{\theta}^{old}$
2: **while** not converged **do**
3: Evaluate $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$
4: Choose $\boldsymbol{\theta}^{new}$ such that

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{3.72}$$

5: $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$
6: **end while**

Output: optimal parameters $\boldsymbol{\theta}^*$

4. Hierarchical Clustering

Clustering algorithms such as k-means aim to divide a set of data points into a predetermined number of subsets and then iteratively update these subsets. In contrast, hierarchical clustering works by iteratively merging points into subsets based on a distance criterion, thereby obtaining a hierarchy of clusters. This means that clusters obtained from the algorithm may contain one or more subclusters which in turn may also contain several subclusters themselves [DHS01, pp. 550-551].

The idea behind hierarchical clustering is relatively simple. In the first step, each of the points in the data set $\mathcal{D}$ is assigned to their own singleton cluster, yielding n clusters. We denote these clusters as the sets $\mathcal{C}_i = \{\boldsymbol{x}_i\}$ for $i = 1, \ldots, n$. Then, in the next step the two closest clusters are aggregated into one cluster. As an example, $\{\boldsymbol{x}_1\}$ and $\{\boldsymbol{x}_2\}$ may be combined into the cluster $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ yielding $n - 1$ remaining clusters. This procedure is then repeated iteratively until k clusters remain. This bottom-up approach is called agglomerative hierarchical clustering. A pseudo-code containing the most important steps can be seen below [DHS01, p. 552].

Algorithm 5 Agglomerative Hierarchical Clustering

Input: data set: $\mathcal{D}$, number of clusters: k
1: Initialise: $\hat{k} \leftarrow n$, $\mathcal{C}_i \leftarrow \{\boldsymbol{x}_i\}$ for $i = 1, \ldots, n$
2: **while** $k < \hat{k}$ **do**
3: Find the nearest clusters, eg. $\mathcal{C}_i$ and $\mathcal{C}_j$
4: Merge $\mathcal{C}_i$ and $\mathcal{C}_j$
5: $\hat{k} \leftarrow \hat{k} - 1$
6: **end while**
Output: k clusters

One critical aspect of Algorithm 5 is how to determine the distance between clusters and thus how to determine the two closest clusters. Some commonly used distances are [DHS01, p. 553]:

$$d_{min}(\mathcal{C}_i, \mathcal{C}_j) = \min_{\boldsymbol{x}\in\mathcal{C}_i, \boldsymbol{x}'\in\mathcal{C}_j} d(\boldsymbol{x}, \boldsymbol{x}') \tag{4.1}$$

$$d_{max}(\mathcal{C}_i, \mathcal{C}_j) = \max_{\boldsymbol{x}\in\mathcal{C}_i, \boldsymbol{x}'\in\mathcal{C}_j} d(\boldsymbol{x}, \boldsymbol{x}') \tag{4.2}$$

$$d_{avg}(\mathcal{C}_i, \mathcal{C}_j) = \frac{1}{n_i n_j} \sum_{\boldsymbol{x}\in\mathcal{C}_i} \sum_{\boldsymbol{x}'\in\mathcal{C}_j} d(\boldsymbol{x}, \boldsymbol{x}') \tag{4.3}$$

$$d_{mean}(\mathcal{C}_i, \mathcal{C}_j) = d(\boldsymbol{m}_i, \boldsymbol{m}_j) \tag{4.4}$$

where $n_i = |\mathcal{C}_i|$ and $\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x}\in\mathcal{C}_i} \boldsymbol{x}$.

Note how despite being different, all distances defined above require a measure of distance between individual data points, which we have denoted as $d(\boldsymbol{x}, \boldsymbol{x}')$. Typically, this distance measure is defined by the l_2-norm i.e. $d(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|_2$, but other definitions may be more applicable depending on the data. If the data points are in well separated spherical clusters, the choice of distance measure between clusters does not matter as much, since all of the distances will usually lead to a good clustering [DHS01, p. 553]. However, the choice of distance measure can have an important effect on the results if clusters are close to each other or not spherical.

If d_{min} is chosen as the distance between two clusters, the hierarchical clustering algorithm is sometimes called a nearest neighbour algorithm. A threshold for the distance between nearest clusters may also be introduced as a stop criterion for this algorithm. If that is the case, then the algorithm is called a single-linkage algorithm. One drawback of the nearest neighbour algorithm is its handling of outlying data points. Due to the distance measure, the algorithm may prefer joining two close but otherwise separated clusters than joining a cluster with an outlying singleton cluster. This is called the chaining effect [DHS01, pp.553-554]. An illustration of this can be seen in Fig. 4.1. Another problem with the nearest neighbour algorithm is its complexity. For a collection of n d-dimensional data points the full complexity of the algorithm is $\mathcal{O}(kn^2d)$ assuming that the inter-point distances are calculated using the l_2-norm which has a complexity of $\mathcal{O}(d)$.

When d_{max} is used, the distance between two clusters is determined by the most distant points in the clusters [DHS01, p. 554]. Due to this, the algorithm is often called the farthest neighbour algorithm. If a threshold on the distance between the two nearest clusters is introduced as a stopping criterion, the algorithm is called a complete-linkage algorithm. Unlike the nearest neighbour algorithm, the farthest neighbour algorithm discourages joining close clusters together and instead may favour adding outliers to clusters. Having a large threshold may result in few large clusters while having low threshold may yield many smaller clusters. This distance has the same complexity as for d_{min}.

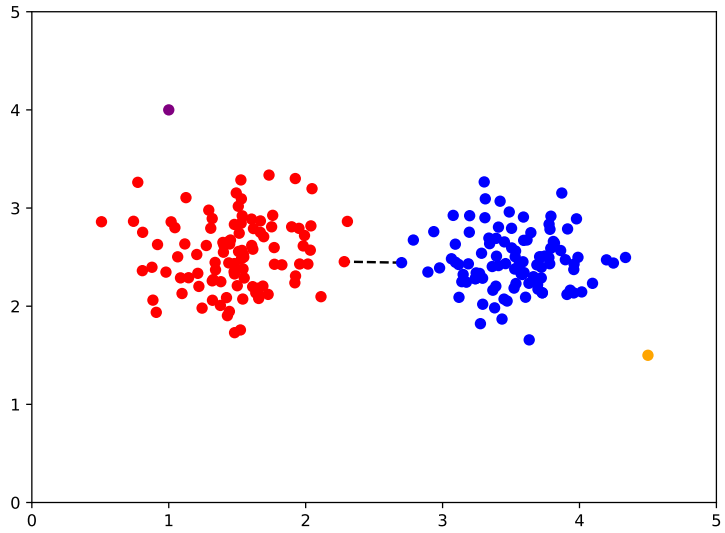


Figure 4.1: The chaining effect on randomly generated data clustered using the nearest neighbour algorithm. The dashed line represents the smallest distance and thus indicates the clusters to be merged leaving one large cluster and two outlying singleton clusters. In this case, using a farthest neighbour algorithm may yield better clustering.

Both d_{min} and d_{max} present two extreme ways of defining the distance between two clusters and both measures may be sensitive to outliers. The two measures d_{avg} and d_{mean} are both used as a compromise between d_{min} and d_{max} [DHS01, p. 555]. Using d_{mean} provides the least computational complexity. However, depending on the choice of dissimilarity measure as it may be difficult if not impossible to define the mean vector of the cluster and the dissimilarity between these means. In cases like this, d_{avg} may be more applicable [DHS01, p. 555].

5. Cluster Validation

One of the most important choices in the design of clustering algorithms is how to choose the number of clusters k. For up to 3-dimensional data, visual inspection is the most straight-forward approach to find a reasonable number of clusters. For data of slightly higher dimension a dimensionality-reduction algorithm may be used to project the data into a more visually friendly space. However, for high-dimensional data visual inspection may be unfeasible and a heuristic-based method may be used instead. Usually in practice, the clustering algorithm will have to be run several times with different choices of k and each clustering result will be evaluated using some predefined measure. The k which results in the best clustering in some predefined sense will then be chosen. An abundance of research has been done to find measures which assess the fitness of the clustering. Typically, the types of measures developed fall into one of the two categories below [JD88, p. 161]

External measures: The validity of the clustering is determined using a priori information.

Internal measures: The validity of the clustering is determined using only the data itself.

We will primarily focus on internal measures since a priori information about the data is rarely known in practice.

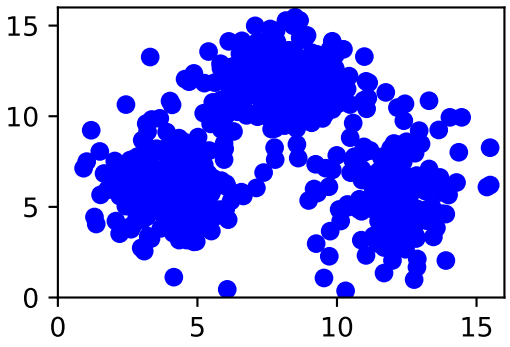


Figure 5.1: Generated clustered data.

For this chapter, we will consider the generated data set in Fig. 5.1. The data set was generated using a mixed Gaussian with three components, and by visual inspection it also looks like $k = 3$ will result in the best clustering. We will use this data set as an example to illustrate how the different measures respond for different k. The data will be clustered using the k-means algorithm.

5.1 The Elbow Criterion

One obvious way to check the validity of the choice of k is to compute the objective function of the clustering algorithm after convergence and see how this value behaves for different k. Recall that for the k-means algorithm, the objective function to be minimised is

$$E = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|\boldsymbol{x}_i - \boldsymbol{m}_j\|_2^2 \tag{5.1}$$

where $\boldsymbol{m}_j$ is the average of the data points in $\mathcal{C}_j$, and r_{ij} indicates the assignment of $\boldsymbol{x}_i$.

The objective function in Eq. (5.1) is the sum of squared distances between each data point and its nearest cluster representative $\boldsymbol{m}_j$. Thus, it is readily seen that an increase in k will always lead to a decrease in Eq. (5.1) since the number of representatives increases. Indeed, if $k = n$ then $E = 0$ by setting $\boldsymbol{m}_j = \boldsymbol{x}_i$ resulting in each data point having its own singleton cluster. However, having n clusters does not yield an interesting clustering result. A plot of the obtained minima of Eq. (5.1) for different k can be seen in Fig. 5.2.

From the figure, we see that for $k < 3$ an increase in k leads to a sharp decrease in the SSE. However, for $k \geq 3$ this decrease is more stagnant. The point where there is a significant change in the slope of the objective function is called the elbow of the objective function and is typically used to determine the best k for the data set. In

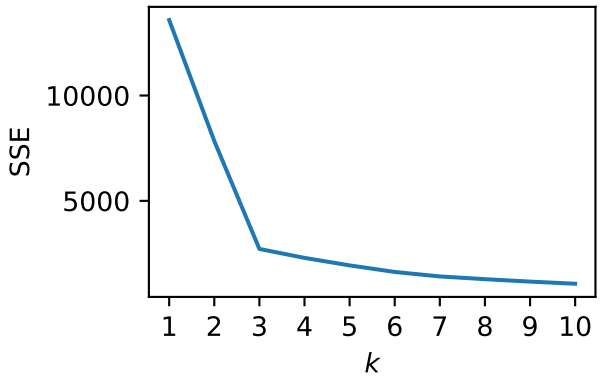


Figure 5.2: Sum of squares error for k-means for different k.

this case, $k = 3$ seems like a good choice since choices of larger k do not seem lead to a significant decrease in the objective function compared to that between $k = 2$ and $k = 3$.

5.2 Davies-Bouldin Index

When deciding on the best choice of k, one needs to set a criterion which signifies good clustering. Typically, the goal of a clustering algorithm is to obtain clusters which are well separated from each other, while the points within each cluster are close to each other. This intuitive idea leads to the formulation of the Davies-Bouldin index. The goal is to define a general cluster separation measure using the intuition from above. First, a dispersion measure $S_j = S(\mathcal{C}_j)$ of a cluster $\mathcal{C}_j$ is defined as a function such that [DB79]

a) $S(\mathcal{C}_j) \geq 0$

b) $S(\mathcal{C}_j) = 0$ if and only if $\boldsymbol{x}_i = \boldsymbol{x}_l$ for all $\boldsymbol{x}_i, \boldsymbol{x}_l \in \mathcal{C}_j$.

Additionally, $M_{i,j} = M(\mathcal{C}_i, \mathcal{C}_j)$ denotes the distance between the clusters $\mathcal{C}_i$ and $\mathcal{C}_j$ defined using an appropriate distance measure. We denote $R_{i,j} = R(S_i, S_j, M_{i,j})$ as the cluster similarity measure defined as a function which satisfies the criteria

a) $R(S_i, S_j, M_{i,j}) \geq 0$

b) $R(S_i, S_j, M_{i,j}) = R(S_j, S_i, M_{j,i})$

c) $R(S_i, S_j, M_{i,j}) = 0$ if and only if $S_i = S_j = 0$

d) $R(S_i, S_j, M_{i,j}) > R(S_i, S_l, M_{i,l})$ if $S_j = S_l$ and $M_{i,j} < M_{i,l}$

e) $R(S_i, S_j, M_{i,j}) > R(S_i, S_l, M_{i,l})$ if $M_{i,j} = M_{i,l}$ and $S_j > S_l$.

The above criteria are satisfied if [DB79]

$$R_{i,j} = R(S_i, S_j, M_{i,j}) = \frac{S_i + S_j}{M_{i,j}}. \tag{5.2}$$

From Eq. (5.2), we see that small dispersions within clusters and large distances between clusters lead to smaller values of $R_{i,j}$. Thus, good clusterings are indicated by small values of $R_{i,j}$. For each cluster, its worst case separation measure can be defined as

$$R_i = \max_{j \neq i} R_{i,j}. \tag{5.3}$$

Intuitively, this is the separation measure between $\mathcal{C}_i$ and the closest and most disperse other cluster. The Davies-Bouldin index is then obtained by taking the average of Eq. (5.3) over all the clusters [DB79]

$$DB = \frac{1}{k} \sum_{i=1}^{k} R_i. \tag{5.4}$$

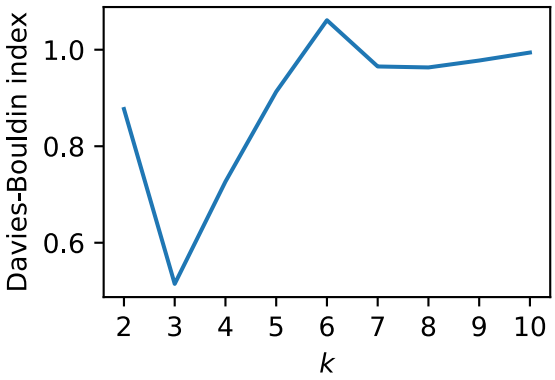


Figure 5.3: Davies-Bouldin index for different k.

We note that due to the definition in Eq. (5.3), the Davis-Bouldin index is only defined for $k \geq 2$. One of the advantages of the Davies-Bouldin index is its ability to apply a broad range of different dispersion and distance measures. In the original article by Davies and Bouldin, they chose the dispersion measure

$$S_j = \left(\frac{1}{|\mathcal{C}_j|} \sum_{\boldsymbol{x} \in \mathcal{C}_j} \|\boldsymbol{x} - \boldsymbol{m}_j\|_p^q \right)^{1/q} \tag{5.5}$$

which is the q'th root of the q'th moment of the l_p-norm between the data points in $\mathcal{C}_j$ and the representative $\boldsymbol{m}_j$, and the distance measure

$$M_{i,j} = \|\boldsymbol{m}_i - \boldsymbol{m}_j\|_p. \tag{5.6}$$

Fig. 5.3 shows the Davies-Bouldin index for different choices of k when clustering the data set in Fig. 5.1 using k-means. We have used the measures in Eqs. (5.5) and (5.6) with $p = 2$ and $q = 1$ which reduces Eq. (5.5) to the average euclidean distance between the data points and their representative. As before, a lower Davies-Bouldin index indicates better clustering which is seen for $k = 3$.

5.3 Dunn Index

For the Dunn index, the intuition is similar to the Davies-Bouldin index. The idea is that we want small compact clusters and we want the clusters to be far apart. To measure the compactness of a cluster $\mathcal{C}_j$, we can for some dissimilarity measure $d(\cdot, \cdot)$ define [Dun73]

$$\text{diam}(\mathcal{C}_j) = \max_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C}_j} d(\boldsymbol{x}, \boldsymbol{x}') \tag{5.7}$$

as the effective diameter of $\mathcal{C}_j$. We can also define the distance between clusters as

$$\text{dist}(\mathcal{C}_i, \mathcal{C}_j) = \min_{\boldsymbol{x} \in \mathcal{C}_i, \boldsymbol{x}' \in \mathcal{C}_j} d(\boldsymbol{x}, \boldsymbol{x}'). \tag{5.8}$$

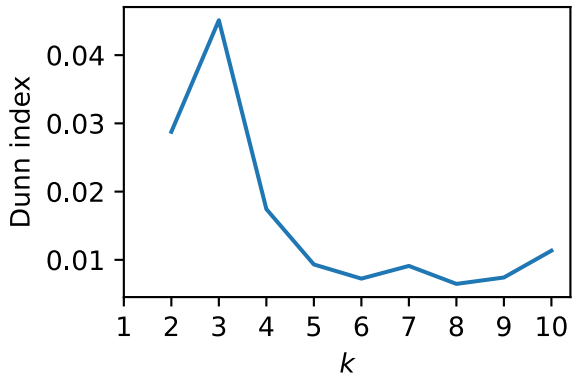


Figure 5.4: Dunn index for different k.

A measure of compact and separated clusters can then be defined as [Dun73]

$$D = \frac{\min\limits_{1\leq q\leq k}\ \min\limits_{1\leq r\leq 1,\ r\neq q} \operatorname{dist}(\mathcal{C}_q, \mathcal{C}_r)}{\max\limits_{1\leq j\leq k} \operatorname{diam}(\mathcal{C}_j)}. \tag{5.9}$$

The intuition is simple, the Dunn index is the ratio between the largest diameter of a cluster and the smallest distance between two clusters. Thus, large values of the Dunn index will indicate compact and separated clusters and thus a good choice for k. The Dunn index can also easily be modified by replacing Eqs. (5.7) and (5.8) with more applicable measures. Fig. 5.4 shows the Dunn index for different choices of k. As seen from the figure, choosing $k = 3$ leads to the highest Dunn index.

5.4 Silhouette Index

The silhouette method is another way to determine the validity of the clustering. The idea is that for each data point $\boldsymbol{x}_i$ assigned to some cluster $\mathcal{C}_j$ we will determine if $\mathcal{C}_j$ is the best "fit" for $\boldsymbol{x}_i$ relative to some other cluster. Due to this, the method is only applicable for $k \geq 2$. For some dissimilarity measure $d(\boldsymbol{x}, \boldsymbol{x}')$ we define the measure [Rou87]

$$a(\boldsymbol{x}_i) = \frac{1}{|\mathcal{C}_j| - 1} \sum_{\boldsymbol{x}'\in\mathcal{C}_j, \boldsymbol{x}'\neq\boldsymbol{x}_i} d(\boldsymbol{x}_i, \boldsymbol{x}'), \quad \boldsymbol{x}_i \in \mathcal{C}_j \tag{5.10}$$

as the average dissimilarity between a data point $\boldsymbol{x}_i$ and all the other data points in the same cluster. Note however that Eq. (5.10) is only defined if $|\mathcal{C}_j| > 1$ and thus is only applicable if the clustering does not contain singleton clusters. This will be discussed later. We will also define the dissimilarity between a data point $\boldsymbol{x_i} \in \mathcal{C}_j$ and a cluster $\mathcal{C}_l \neq \mathcal{C}_j$ as

$$d(\boldsymbol{x}_i, \mathcal{C}_l) = \frac{1}{|\mathcal{C}_l|} \sum_{\boldsymbol{x}'\in\mathcal{C}_l} d(\boldsymbol{x}_i, \boldsymbol{x}'), \quad \boldsymbol{x}_i \notin \mathcal{C}_l \tag{5.11}$$

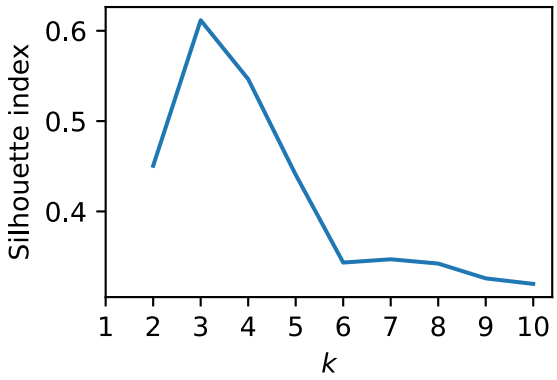


Figure 5.5: Silhouette index for different k.

which yields the average dissimilarity between $\boldsymbol{x}_i$ and all other data points in $\mathcal{C}_l$. This provides an intuitive measure of a data point's closeness to a cluster with nearer clusters having a smaller dissimilarity measure to $\boldsymbol{x}_i$. In this sense, the nearest cluster to $\boldsymbol{x}_i$ (apart from the point's own cluster) has the dissimilarity

$$b(\boldsymbol{x}_i) = \min_{\mathcal{C}_l \neq \mathcal{C}_j} d(\boldsymbol{x}_i, \mathcal{C}_l), \quad \boldsymbol{x}_i \in \mathcal{C}_j. \tag{5.12}$$

The measures in Eqs. (5.10) and (5.12) provide an intuitive measure of the classification of a data point $\boldsymbol{x}_i$. If $a(\boldsymbol{x}_i) < b(\boldsymbol{x}_i)$ then $\boldsymbol{x}_i$ is on average closer to the data points in its own cluster than the points in the second closest cluster indicating a proper classification. However, if the opposite is true then this indicates a poor classification. If $a(\boldsymbol{x}_i) = b(\boldsymbol{x}_i)$, then we can not say for certain if $\boldsymbol{x}_i$ belongs to its own cluster or the second closest one. This leads to the definition of the silhouette score for $\boldsymbol{x}_i$

$$s(\boldsymbol{x}_i) = \begin{cases} 1 - a(\boldsymbol{x}_i)/b(\boldsymbol{x}_i), & \text{if } a(\boldsymbol{x}_i) < b(\boldsymbol{x}_i) \\ 0, & \text{if } a(\boldsymbol{x}_i) = b(\boldsymbol{x}_i) \\ b(\boldsymbol{x}_i)/a(\boldsymbol{x}_i) - 1, & \text{if } a(\boldsymbol{x}_i) > b(\boldsymbol{x}_i) \end{cases} \tag{5.13}$$

From Eq. (5.13), we see that $-1 \leq s(\boldsymbol{x}_i) \leq 1$ with a higher score yielding better clustering. We noted earlier that $a(\boldsymbol{x}_i)$ is only defined if the cluster containing $\boldsymbol{x}_i$ is not a singleton cluster. If there is a singleton cluster, we will set $s(\boldsymbol{x}_i) = 0$ [Rou87, p. 56]. To assess the validity of the clustering, we simply take the average of the silhouette scores

$$S = \frac{1}{n} \sum_{i=1}^{n} s(\boldsymbol{x}_i) \tag{5.14}$$

which is called the silhouette index. A plot of the silhouette index for different k when clustering the data set in Fig. 5.1 can be seen in Fig. 5.5. Here, we have defined $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$. As seen from the figure, the best clustering can be found for $k = 3$.

6. Generalised Linear Models

Generalised linear models are types of statistical models which can be used to explain the relationship between statistical variables. They are able to model more complex interactions such as non-linear relationships between the variables.

6.1 The Exponential Family

In generalised linear models, the data is assumed to be distributed according to a probability distribution which belongs to a general family of distributions called the exponential dispersion family. In order to define this family, we will first present the natural exponential family.

Definition 6.1 (Natural Exponential Family)
[MT11, p. 90] Let Y be a random variable which is distributed according to a density $f_Y(y;\theta)$. The density $f_Y(y;\theta)$ is said to be a member of the natural exponential family of distributions if it can be written on the form

$$f_Y(y;\theta) = c(y)\exp\big(\theta y - \kappa(\theta)\big), \quad \theta \in \Omega \subseteq \mathbb{R} \tag{6.1}$$

with $\kappa(\theta)$ being the cumulant generator.

The representation in Eq. (6.1) is called the canonical parameterisation of the family, while θ is called the canonical parameter belonging to the parameter space Ω. The cumulant generator is important as together with the support of $f_Y(y,\theta)$ it characterises the distribution.

With the addition of another parameter, we can define the exponential dispersion family.

Definition 6.2 (Exponential Dispersion Family)
[MT11, p. 90] A family of probability densities which can be written on the form

$$f_Y(y;\theta,\lambda) = c(y,\lambda)\exp\Big(\lambda\big(\theta y - \kappa(\theta)\big)\Big) \tag{6.2}$$

is called an exponential dispersion family of distributions.

The parameter $\lambda > 0$ is called the precision parameter of the index parameter. It can be seen that by setting $\lambda = 1$, Eq. (6.2) is equivalent to Eq. (6.1). Thus, densities, which are part of the natural exponential family are also part of the exponential dispersion family with $\lambda = 1$.

Examples of distributions which are members of the exponential dispersion family are the Poisson distribution, the Gaussian distribution and the binomial distribution. For the parameters of these distributions we refer to the table in [MT11, p. 96].

An example of a distribution belonging to the natural exponential family is the Bernoulli distribution defined by the density

$$f_Y(y;q) = p^y(1-q)^{1-y} \tag{6.3}$$

where $p(Y=1) = q$ and $p(Y=0) = 1-q$. To see this, we rewrite Eq. (6.3) as

$$\begin{aligned} f_Y(y;q) &= q^y(1-q)^{1-y} && (6.4)\\ &= \exp\Big(\ln\left(q^y(1-q)^{1-y}\right)\Big) && (6.5)\\ &= \exp\left(y\ln q + (1-y)\ln(1-q)\right) && (6.6)\\ &= \exp\Big(y(\ln q - \ln(1-q)) + \ln(1-q)\Big) && (6.7)\\ &= \exp\left(y\ln\Big(\frac{q}{1-q}\Big) + \ln(1-q)\right) && (6.8)\\ &= \exp\left(\theta y - \ln\Big(\frac{1}{1-q}\Big)\right) && (6.9) \end{aligned}$$

where $\theta = \ln\left(q/(1-q)\right)$ and the cumulant generator is expressed in terms of q by $\ln\left(1/(1-q)\right)$. To express the cumulant generator in terms of θ, we note that

$$\begin{aligned} \theta &= \ln\Big(\frac{q}{1-q}\Big) && (6.10)\\ e^\theta &= \frac{q}{1-q} && (6.11)\\ e^\theta - e^\theta q &= q && (6.12)\\ e^\theta &= q(1+e^\theta) && (6.13)\\ q &= \frac{e^\theta}{1+e^\theta}. && (6.14) \end{aligned}$$

Substituting Eq. (6.14) into Eq. (6.9) yields

$$f_Y(y;\theta) = \exp\left(\theta y - \ln\left(\frac{1}{1-\frac{e^\theta}{1+e^\theta}}\right)\right) \tag{6.15}$$

$$= \exp\left(\theta y - \ln\left(\frac{1+e^\theta}{1+e^\theta-e^\theta}\right)\right) \tag{6.16}$$

$$= \exp\left(\theta y - \ln\Big(1+e^\theta\Big)\right) \tag{6.17}$$

$$= \exp\left(\theta y - \kappa(\theta)\right) \tag{6.18}$$

where

$$\kappa(\theta) = \ln\Big(1+e^\theta\Big). \tag{6.19}$$

Therefore, the Bernoulli distribution is a member of the natural exponential family and thus also the exponential dispersion family with $\theta = \ln\left(q/(1-q)\right)$, $\lambda = 1$, $c(y) = 1$ and the cumulant generator given by Eq. (6.19).

The cumulant generator is particularly interesting, as the properties of the exponential dispersion family depend on the cumulant generator.

$$\mathrm{E}[Y] = \kappa'(\theta) \tag{6.20}$$

$$\mathrm{Var}[Y] = \frac{\kappa''(\theta)}{\lambda} \tag{6.21}$$

However, the variance also depends on the precision parameter. Therefore, we introduce a concept which is related to the variance of the distribution but which does not depend on the precision parameter. In order to do so, we define the function

$$\tau(\theta) = \kappa'(\theta) = \mathrm{E}[Y] \tag{6.22}$$

which maps the canonical parameter θ from the parameter space Ω into the mean value space $\mathcal{M}$.

Using this function and the properties of the exponential dispersion family, we define a function which is related to the variance but only depends on the mean value of the distribution.

Definition 6.3 (Variance Function)
[MT11, p. 92] The variance function of the mean value of the distribution is defined as

$$V(\mu) = \kappa''\big(\tau^{-1}(\mu)\big) \tag{6.23}$$

where $\mu = \mathrm{E}[Y]$.

The variance function is also called the unit variance function as it will be used in the definition of the unit deviance. Note that the variance function of the distribution should not be confused with the variance operator which yields the variance of the distribution.

As seen in the definition of the variance function, the mean value parameter together with the variance function characterises the exponential dispersion family. Therefore, the variance function together with the mean value parameter can be seen as an alternative parameterisation of the exponential family, yielding two parameterisations each with their own advantages.

1. κ, θ, λ
2. V, μ, λ

The advantage of $\theta \in \Omega$ is that it resides on the real line, while $\mu \in \mathcal{M}$ can be directly measured as the mean and compared to measured values.

These two parameterisations are related by the so-called canonical link function

Definition 6.4 (Canonical Link Function)
[MT11, p. 95] The mapping between the parameterisations θ and μ which is the inverse of Eq. (6.22)

$$\theta = \tau^{-1}(\mu) \tag{6.24}$$

is called the canonical link function.

In general, the mapping $g(\cdot)$ from the mean to another parameter is called a link function.

6.2 Generalised Linear Models

Based on the definition of the exponential dispersion family, we can define the generalised linear model which is able to model random variables following an exponential dispersion model. When working with generalised linear models, we will use the link function as a mapping from the mean value to a linear predictor $\boldsymbol{\eta} = g(\boldsymbol{\mu})$.

Definition 6.5 (Generalised Linear Model)
[MT11, p. 99] Let $Y_1, Y_2, \ldots, Y_n$ be mutually independent random variables with

$$Y_i \sim ED(\mu_i, V(\mu_i)/\lambda_i), \quad i = 1, 2, \ldots, n \tag{6.25}$$

where $ED(\mu_i, V(\mu_i)/\lambda_i)$ is an exponential dispersion family distribution with mean parameter μ_i and precision parameter λ_i. In addition, let the variance

function $V(\cdot)$ be the same for all Y_i, and define the linear predictor $\boldsymbol{\eta} = [\eta_1 \quad \eta_2 \quad \cdots \quad \eta_n]^T$ such that

$$\eta_i = g(\mu_i), \quad i = 1, 2, \ldots, n \tag{6.26}$$

where g is a link function.

A generalised linear model for $Y_1, Y_2, \ldots, Y_n$ describes the hypothesis that $\boldsymbol{\eta}$ subtracted with known offset values $\boldsymbol{\eta}_0$ belongs to a linear subspace L, i.e.

$$\mathcal{H}_0 \ : \ \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L \tag{6.27}$$

where L is a linear subspace of $\mathbb{R}^n$ of dimension d.

There are thus multiple elements which need to be specified in order to obtain a generalised linear model for the random variables $Y_1, Y_2, \ldots, Y_n$. The type of probability distribution in the exponential family which it is hypothesised that the random variables follow, the link function, and the linear predictor $\boldsymbol{\eta}$.

The linear predictor is related to a d-dimensional coefficient vector $\boldsymbol{\beta}$ via vectors for concrete observations called model vectors. We let the model vector for the i'th observation be given as $\boldsymbol{x}_i = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{id}]^T$. The matrix obtained by collecting the model vectors is called the design matrix.

Definition 6.6 (Design Matrix for Generalised Linear Model)
[MT11, p. 99] Let the $n \times d$ matrix $\boldsymbol{X}$ have rows consisting of model vectors $\boldsymbol{X} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n]^T$ and let the subspace L be spanned by the columns of $\boldsymbol{X}$ such that the hypothesis in Eq. (6.27) can be written as

$$\boldsymbol{\eta} - \boldsymbol{\eta}_0 = \boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^d \tag{6.28}$$

where $\boldsymbol{X}$ has full rank. The matrix $\boldsymbol{X}$ is then called the design matrix.

In other words, a generalised linear model stipulates that the mean of the observations of the random variables Y can be modelled by a transformation of a linear relation between the design matrix and the coefficients in $\boldsymbol{\beta}$

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) \tag{6.29}$$

or equivalently

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} = g(\boldsymbol{\mu}) \tag{6.30}$$

with the offset $\boldsymbol{\eta}_0 = \boldsymbol{0}$.

6.3 Significance Tests

Upon determining the coefficients of the model, we wish to assess the significance of the determined coefficients. This is done by statistical tests following a general structure which is applicable to generalised linear models.

The process of determining the significance of the model and the coefficients therein follows a pattern as follows

1. Formulate a sufficient model by including all of the parameters as terms in the model.
2. Test whether the model can be reduced to the null model, i.e. whether we can reject the null hypothesis. If the null hypothesis can be rejected, we know that at least some of the parameters are necessary.
3. Test whether the model can be reduced by eliminating terms which are not significant.
4. Analyse the residuals in order to validate the model.

The test which is used for model reduction is the likelihood ratio test. We first define the likelihood ratio.

Definition 6.7 (Likelihood Function)
[MT11, p. 14] Given the parametric density $f_Y(\boldsymbol{y}; \boldsymbol{\phi})$ with parameter vector $\boldsymbol{\phi} \in \Phi^d$ for the observations $\boldsymbol{y} = [y_1 \quad y_2 \quad \cdots \quad y_n]^T$ the likelihood function for $\boldsymbol{\phi}$ is the function

$$L(\boldsymbol{\phi}; \boldsymbol{y}) = c(y_1, y_2, \ldots, y_n) f_Y(y_1, y_2, \ldots, y_n; \boldsymbol{\phi}) \tag{6.31}$$

where $c(y_1, y_2, \ldots, y_n)$ is a constant.

Definition 6.8 (Likelihood Ratio)
[MT11, p. 26] Consider the null hypothesis $\mathcal{H}_0 : \boldsymbol{\phi} \in \Omega_0$ against the alternative $\mathcal{H}_1 : \boldsymbol{\phi} \in \Omega \setminus \Omega_0$ with $\Omega_0 \subseteq \Omega$, where $\dim(\Omega_0) = m$ and $\dim(\Omega) = d$.

For given $y_1, y_2, \ldots, y_n$ the likelihood ratio is defined as

$$\lambda(\boldsymbol{y}) = \frac{\sup_{\boldsymbol{\phi} \in \Omega_0} L(\boldsymbol{\phi}; \boldsymbol{y})}{\sup_{\boldsymbol{\phi} \in \Omega} L(\boldsymbol{\phi}; \boldsymbol{y})}. \tag{6.32}$$

For each distribution in the exponential family, the unit deviance $d(y; \mu)$ is given based on its variance function.

Definition 6.9 (Unit Deviance)
[MT11, p. 93] Given an observation y and a parameter value μ, the unit deviance function is given as

$$D(y;\mu) = \int_{\mu}^{y} \frac{y-u}{V(u)} du \tag{6.33}$$

where $V(\cdot)$ is the variance function.

Using the unit deviance of the distribution and the weights if any, the residual deviance can be defined.

Definition 6.10 (Residual Deviance)
[MT11, p. 105] The deviance of a distribution in the exponential family with weight parameters w_i, means μ_i and observations y_i for $i = 1, 2, \ldots, n$ is given as

$$\mathcal{D}(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} w_i D(y_i;\hat{\mu}_i) \tag{6.34}$$

where $\hat{\mu}$ is the maximum likelihood estimate of μ_i.

These definitions are used in the likelihood ratio test.

Theorem 6.11 (Likelihood Ratio Test)
[MT11, p. 111] Let $\boldsymbol{\eta} \in \mathbb{R}^n$ be given as a transformation of the mean values $\boldsymbol{\mu}$ of mutually independent random variables $Y_1, Y_2, \ldots, Y_n$ following an exponential dispersion model with the dispersion parameter $\sigma^2 = \frac{1}{\lambda}$ such that

$$\eta_i = g(\mu_i), \quad i = 1, 2, \ldots n \tag{6.35}$$

and let L be a d-dimensional linear subspace of $\mathbb{R}^n$.

Assume that the generalised linear model

$$\mathcal{H}_1 \,:\, \boldsymbol{\eta} \in L \subset \mathbb{R}^d \tag{6.36}$$

holds with L parameterised as $\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1$.

Consider the hypothesis

$$\mathcal{H}_0 \,:\, \boldsymbol{\eta} \in L_0 \subset \mathbb{R}^m \tag{6.37}$$

where $\boldsymbol{\eta} = \boldsymbol{X}_0\boldsymbol{\beta}_0$ and $m < d$, with the alternative $\mathcal{H}_1 \,:\, \boldsymbol{\eta} \in L \setminus L_0$.

Then the likelihood ratio test for $\mathcal{H}_0$ has the test statistic

$$-2\log\lambda(\boldsymbol{Y}) = \mathcal{D}(\boldsymbol{\mu}_0(\boldsymbol{\beta}_0);\boldsymbol{\mu}_1(\boldsymbol{\beta}_1)) \tag{6.38}$$

where $\mathcal{D}$ is the deviance statistic, and $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are the mean function under $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively.

When $\mathcal{H}_0$ is true, the test statistic $\mathcal{D}(\boldsymbol{\mu}_0(\boldsymbol{\beta}_0); \boldsymbol{\mu}_1(\boldsymbol{\beta}_1))$ will asymptotically follow a $\sigma^2\chi^2(d-m)$ distribution.

When performing a likelihood ratio test in practice, we can thus compute the deviance statistic $\mathcal{D}(\boldsymbol{\mu}_0(\boldsymbol{\beta}_0); \boldsymbol{\mu}_1(\boldsymbol{\beta}_1))$ and compare it to the $\sigma^2\chi^2(k-m)$ distribution. Typically we are looking for a 95% significance level, which means that if the deviance is in the tails with probability 2.5% on either side of the distribution, we can reject $\mathcal{H}_0$ with 95% confidence.

The deviance statistic $\mathcal{D}(\boldsymbol{\mu}_0(\boldsymbol{\beta}_0); \boldsymbol{\mu}_1(\boldsymbol{\beta}_1))$ can be determined via the Pythagorean relation

$$\mathcal{D}(\boldsymbol{\mu}_0(\boldsymbol{\beta}_0); \boldsymbol{\mu}_1(\boldsymbol{\beta}_1)) = \mathcal{D}(\boldsymbol{y}; \boldsymbol{\mu}_0(\boldsymbol{\beta}_0)) + \mathcal{D}(\boldsymbol{y}; \boldsymbol{\mu}_1(\boldsymbol{\beta}_1)) \tag{6.39}$$

relating the deviance statistic with the residual deviances, i.e. the deviances of the means relative to the observations. [MT11, p. 112]

6.3.1 Goodness of fit

Recall that the first step in determining the significance of a generalised linear model is to formulate a comprehensive model termed the sufficient model by including all the available parameters as terms in the model. After formulating a sufficient model, we can test the goodness of fit in order to determine whether it is possible to reject the null hypothesis. [MT11, p. 112]

The goodness of fit test is a Likelihood Ratio Test where the full model, allowing each observation to have its own mean, is compared with the null hypothesis. That is, let $\mathcal{H}_{\text{full}} : \boldsymbol{\mu} \in \mathbb{R}^n$ and $\mathcal{H}_0 : \eta \in L \subset \mathbb{R}^d$ with L parameterised as $\boldsymbol{\eta} = \boldsymbol{X}_0\boldsymbol{\beta}$.

Since we let each observation have its own freely varying mean value, the residual deviance $\mathcal{D}(\boldsymbol{y}, \boldsymbol{\mu}_{\text{full}}))$ is 0. Therefore, by Eq. (6.39), the deviance statistic is equal to the residual deviance $\mathcal{D}(\boldsymbol{y}, \boldsymbol{\mu}_0(\boldsymbol{\beta}_0)$. By Theorem 6.11, when $\mathcal{H}_0$ is true the deviance statistic is distributed as $\chi^2(n-d)$ assuming that the dispersion $\sigma^2 = 1$.

The residual deviances are sometimes alternatively referred to as the goodness of fit statistics. In that case, the goodness of fit statistic of the model under hypothesis $\mathcal{H}_0$ is denoted $G^2(\mathcal{H}_0)$. The partitioning of the deviance in Eq. (6.39) is then denoted as

$$G^2(\mathcal{H}_0|\mathcal{H}_1) = G^2(\mathcal{H}_0) - G^2(\mathcal{H}_1) \tag{6.40}$$

with $G^2(\mathcal{H}_0|\mathcal{H}_1) = \mathcal{D}(\boldsymbol{\mu}_0(\boldsymbol{\beta}_0); \boldsymbol{\mu}_1(\boldsymbol{\beta}_1))$ being interpreted as the goodness of fit statistic of $\mathcal{H}_0$ conditioned on $\mathcal{H}_1$ being true. [MT11, p. 113]

The goodness of fit statistics can be represented in an analysis of deviance table as seen in Table 6.1. In the table, the deviances of the model and the residual are laid out alongside their degrees of freedom and mean deviance. In addition, we show their interpretation as goodness of fit of the hypotheses.

Source	f	Deviance	Mean Deviance	Interpretation as goodness of fit
Model $\mathcal{H}_{\text{null}}$	$k-1$	$\mathcal{D}(\boldsymbol{\mu}_1(\boldsymbol{\beta}_1);\boldsymbol{\mu}_{\text{null}})$	$\frac{\mathcal{D}(\boldsymbol{\mu}_1(\boldsymbol{\beta}_1);\boldsymbol{\mu}_{\text{null}})}{k-1}$	$G^2(\mathcal{H}_{\text{null}}\|\mathcal{H}_1)$
Residual (Error)	$n-k$	$\mathcal{D}(\boldsymbol{y};\boldsymbol{\mu}_1(\boldsymbol{\beta}_1))$	$\frac{\mathcal{D}(\boldsymbol{y};\boldsymbol{\mu}_1(\boldsymbol{\beta}_1))}{n-k}$	$G^2(\mathcal{H}_1)$
Corrected total	$n-1$	$\mathcal{D}(\boldsymbol{y};\boldsymbol{\mu}_{\text{null}})$		$G^2(\mathcal{H}_{\text{null}})$

Table 6.1: Analysis of deviance table for the model $\mathcal{H}_1$ compared with the minimal model H_{null}. f indicates the degrees of freedom.

From the goodness of fit interpretation it is seen that the deviance of the residual indicates whether the model can be maintained at all by comparing it with the percentiles of the $\chi^2(n-k)$ distribution. Upon determining that the model can be maintained, it can be examined whether the null hypothesis can be rejected by comparing $G^2(\mathcal{H}_{\text{null}}|\mathcal{H}_1)$ with $\chi^2(n-k)$. By rejecting the null hypothesis of a model, the model $\mathcal{H}_1$ has been established as a sufficient model. Subsequently, we can investigate whether the sufficient model can be reduced to a model with fewer parameters, thereby determining whether the terms in the sufficient model are necessary. [MT11, p. 115]

When the model has been determined to be sufficient, the parameters can be investigated to determine whether they have a statistically significant effect on the model. This is called model reduction. Model reduction can be done by performing a likelihood ratio test between two models at a time. If the aim is to see whether one specific parameter should be included in the model, a likelihood ratio test can be performed between the model including this parameter and the model without.

6.3.2 Wald Test

For generalised linear models, it is also possible to perform inference on the individual parameters of the model directly without the need to perform successive likelihood ratio tests. The method which is commonly used is termed the Wald test

Theorem 6.12 (Wald Test)
[MT11, p. 116] Let the hypothesis be given such that the parameter β_j has a specific value, that is, $\mathcal{H} : \beta_j = \beta_{0,j}$. The hypothesis $\mathcal{H}$ is tested by the test statistic

$$u_j = \frac{\hat{\beta}_j - \beta_{0,j}}{\sqrt{\hat{\sigma}^2 \hat{\sigma}_{jj}}} \tag{6.41}$$

where $\hat{\sigma}^2$ indicates the estimated dispersion parameter, while σ_{jj} is the j'th diagonal element in the dispersion matrix $\hat{\Sigma}$. If the hypothesis $\mathcal{H}$ is true, u_j is approximately distributed as a standard normal distribution.

Proof.
Omitted. The proof involves determining that $\hat{\beta}_j$ is a ML estimator, which by Theorem 2.4 in [MT11] under some regularity conditions is normally distributed. ■

The test statistic u_j can be compared to the quantiles of a standard normal distribution and rejected for large values of $|u_j|$.

Specifically, the hypothesis $\mathcal{H}_0 : \beta_j = 0$ for whether a parameter can be omitted gives the test statistic

$$u_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \hat{\sigma}_{jj}}}. \tag{6.42}$$

The p-value can then be found as $p = 2(1 - \Phi(|u_j|)$ with Φ being the probit function. The p-value is a mapping of the test statistic into the range 0 to 1, signifying the probability that the observations follow the null hypothesis. Thus, with a low p-value, we can reject the null hypothesis.

Equivalently, the test can be done using the test statistic $z_j = u_j^2$ which can be rejected for $z_j > \chi^2_{1-\alpha}$.

7. Logistic Regression

The logistic regression model has a rich history in statistics. From its first development in the mid 20th century [Ber44] many refinements have been made to the model [Cox58] and its usefulness in classification has been well demonstrated over time. In its simplest form the logistic regression model is used as a binary classifier, but several extensions have since been made to the model such as the extension to data with multiple classes [The69].

In this chapter, we will present the method of logistic regression and its uses in classification. We will start by considering classification of data consisting of two classes and afterwards extend the problem to multiple classes.

7.1 Binary Logistic Regression

For a d-dimensional data point $\boldsymbol{x}$ belonging to one of two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ we aim to classify the point to the correct class. The idea in logistic regression is to compute the posterior probability $p(\mathcal{C}_1|\boldsymbol{x})$ based on the class conditional probability $p(\boldsymbol{x}|\mathcal{C}_1)$ and the prior $p(\mathcal{C}_1)$ using Bayes' Theorem and then make a decision based on these probabilities. This probabilistic view of classification is also used in fuzzy classification algorithms such as the EM-algorithm discussed in Chapter 3. The difference however is that in fuzzy classification $p(\boldsymbol{x}|\mathcal{C}_1)$ and $p(\mathcal{C}_1)$ are estimated iteratively by improving the classification of each data point, while in logistic regression the relation between the posterior probabilities and the features of the data points will be modeled using prior data with known classifications. Assume that we have a d-dimensional data point $\boldsymbol{x}$ with an unknown classification. We define the label corresponding to $\boldsymbol{x}$ as

$$l = \begin{cases} 1, & \text{if } \boldsymbol{x} \in \mathcal{C}_1, \\ 0, & \text{if } \boldsymbol{x} \in \mathcal{C}_2. \end{cases} \tag{7.1}$$

Since the label is unknown, we can model it as a random binary variable following a Bernoulli distribution

$$p(l) = p(\mathcal{C}_1|\boldsymbol{x})^l p(\mathcal{C}_2|\boldsymbol{x})^{1-l} \tag{7.2}$$

$$= p(\mathcal{C}_1|\boldsymbol{x})^l \Big(1 - p(\mathcal{C}_1|\boldsymbol{x})\Big)^{1-l} \tag{7.3}$$

with the posterior probability $p(\mathcal{C}_1|\boldsymbol{x})$ as the parameter. By the Bernoulli distribution, the expected value of l will also be $p(\mathcal{C}_1|\boldsymbol{x})$. From Bayes' Theorem, we can write the posterior probability of $\boldsymbol{x}$ belonging to $\mathcal{C}_1$ as [Bis06, p. 197]

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{7.4}$$

$$= \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)/p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{1 + p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)/p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{7.5}$$

$$= \frac{e^a}{1 + e^a} \tag{7.6}$$

where

$$a = \ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{7.7}$$

is called the log-odds.

The right hand side of Eq. (7.6) is an important function and deserves its own definition.

Definition 7.1 (Logistic Sigmoid)
The function

$$\sigma(a) = \frac{e^a}{1 + e^a}, \quad a \in \mathbb{R} \tag{7.8}$$

is called the logistic sigmoid function. [Bis06, p.197]

Some properties of the logistic sigmoid function are shown in the following proposition.

Proposition 7.2 (Properties of the Logistic Sigmoid)
For the logistic sigmoid function $\sigma(a)$, it follows that for all $a \in \mathbb{R}$

a) $0 < \sigma(a) < 1$

b) $\sigma(-a) = 1 - \sigma(a)$

c) $\sigma^{-1}(a) = \ln\left(\frac{a}{1-a}\right)$

d) $\frac{\partial \sigma(a)}{\partial a} = \sigma(a)(1 - \sigma(a))$

Proof.
Property a)

For $a \to \infty$ we have that $e^a \to \infty$. Thus, $\sigma(a) \to 1$. For $a \to -\infty$ we have that $e^a \to 0$, yielding that $\sigma(a) \to 0$.

Property b)

$$\sigma(-a) = \frac{e^{-a}}{1+e^{-a}} = \frac{1+e^{-a}-1}{1+e^{-a}} = 1 - \frac{1}{1+e^{-a}} = 1 - \frac{e^a}{1+e^a} = 1 - \sigma(a). \quad (7.9)$$

Property c)

$$\sigma^{-1}(\sigma(a)) = \ln\left(\frac{e^a/(1+e^a)}{1-e^a/(1+e^a)}\right) = \ln\left(\frac{e^a}{1+e^a-e^a}\right) = \ln e^a = a. \quad (7.10)$$

Property d)

Using the product rule, we have that

$$\frac{\partial\sigma(a)}{\partial a} = \frac{1}{1+e^a}\frac{\partial}{\partial a}e^a + e^a\frac{\partial}{\partial a}\frac{1}{1+e^a} \quad (7.11)$$

$$= \frac{e^a}{1+e^a} - \frac{(e^a)^2}{(1+e^a)^2} = \frac{e^a(1+e^a)-(e^a)^2}{(1+e^a)^2} \quad (7.12)$$

$$= \frac{e^a}{(1+e^a)^2} = \sigma(a)\frac{1}{1+e^a} = \sigma(a)\frac{e^{-a}}{1+e^{-a}} \quad (7.13)$$

$$= \sigma(a)\sigma(-a) = \sigma(a)(1-\sigma(a)) \quad (7.14)$$

where the last equation holds due to Property b). ■

The inverse of the logistic sigmoid function is called the logit function. The introduction of these functions may seem arbitrary but the form of the logit function is important for the model. By Eq. (7.7), we can rewrite the logit function as

$$a = \ln\{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)\} - \ln\{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)\} \quad (7.15)$$

$$= \ln\left\{\frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x})}p(\boldsymbol{x})\right\} - \ln\left\{\frac{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\boldsymbol{x})}p(\boldsymbol{x})\right\} \quad (7.16)$$

$$= \ln p(\mathcal{C}_1|\boldsymbol{x}) + \ln p(\boldsymbol{x}) - \ln p(\mathcal{C}_2|\boldsymbol{x}) - \ln p(\boldsymbol{x}) \quad (7.17)$$

$$= \ln p(\mathcal{C}_1|\boldsymbol{x}) + \ln p(\mathcal{C}_2|\boldsymbol{x}) \quad (7.18)$$

showing that the logit function is log-linear with respect to the posterior probabilities. A great deal of simplification can be obtained by having a be a linear function of $\boldsymbol{x}$

which is the case if the posterior densities are part of the exponential family [Bis06, p. 203]. Under this assumption, we can write the logit function as [Bis06, p. 198]

$$a(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 \tag{7.19}$$

where $\boldsymbol{w}$ is a vector of weights and w_0 is a constant bias parameter. Typically, Eq. (7.19) is written more compactly as [Bis06, pp. 204-205]

$$a(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}} \tag{7.20}$$

where $\tilde{\boldsymbol{w}} = [w_0 \quad \boldsymbol{w}^T]^T$ and $\tilde{\boldsymbol{x}} = [1 \quad \boldsymbol{x}^T]^T$. We will make use of the same convention but keeping with the previous notation such that $\boldsymbol{w}$ and $\boldsymbol{x}$ refer to $\tilde{\boldsymbol{w}}$ and $\tilde{\boldsymbol{x}}$.

Using the expression in Eq. (7.20), we can define the logistic regression model.

Definition 7.3 (Binary Logistic Regression Model)
[Bis06, p.205] Let $\boldsymbol{x}$ be a d-dimensional data point belonging to one of the classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Let l be the corresponding label of $\boldsymbol{x}$ such that $l = 1$ if $\boldsymbol{x} \in \mathcal{C}_1$ and $l = 0$ otherwise. The relation

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x}) \tag{7.21}$$

with $p(\mathcal{C}_2|\boldsymbol{x}) = 1 - p(\mathcal{C}_1|\boldsymbol{x})$ and σ being the logistic sigmoid function, is called a logistic regression model with the d-dimensional vector $\boldsymbol{w}$ containing its parameters.

The model is equivalently formulated as

$$\boldsymbol{w}^T\boldsymbol{x} = \sigma^{-1}\big(p(\mathcal{C}_1|\boldsymbol{x})\big) = \sigma^{-1}\big(E[l]\big). \tag{7.22}$$

Eq. (7.22) shows that the logistic regression model is a generalised linear model with the logit function as the canonical link function and l and $\boldsymbol{x}$ as the response and explanatory variable respectively [MT11, p. 99-100]. The equation also provides an interpretation of the parameters of the model. For any $i = 1, 2, \ldots, d$, a unit increase in x_i will result in a w_i increase or decrease in the log-odds $\sigma^{-1}\big(p(\mathcal{C}_1|\boldsymbol{x})\big)$, depending on the sign of w_i.

7.1.1 Parameter Estimation

The parameters of the model are estimated based on data points with known classifications which makes logistic regression a supervised learning method for classification. Assume that we have n data points with each point being in either $\mathcal{C}_1$ or $\mathcal{C}_2$. For a data point $\boldsymbol{x}_i$ we will assign the point a label $l_i \in \{0, 1\}$ such that $l_i = 1$ if $\boldsymbol{x}_i \in \mathcal{C}_1$ and $l_i = 0$ if $\boldsymbol{x}_i \in \mathcal{C}_2$ and let $\boldsymbol{l} = [l_1 \quad l_2 \quad \cdots \quad l_n]^T$. Assuming that the data points are mutually independent, we can write the likelihood [Bis06, p.206]

$$p(\boldsymbol{l}|\boldsymbol{w}, \boldsymbol{x}) = \prod_{i=1}^{n} p(\mathcal{C}_1|\boldsymbol{x_i})^{l_i}\Big(1 - p(\mathcal{C}_1|\boldsymbol{x_i})\Big)^{1-l_i}. \tag{7.23}$$

As usual, the estimation of $\boldsymbol{w}$ is done by maximising the likelihood. Taking the negative log of the likelihood and using Eq. (7.6) with Eq. (7.20) yields

$$E(\boldsymbol{w}) = -\ln p(\boldsymbol{l}|\boldsymbol{w}, \boldsymbol{x}) \tag{7.24}$$

$$= -\sum_{i=1}^{n} \ln \left\{ p(\mathcal{C}_1|\boldsymbol{x_i})^{l_i} \Big(1 - p(\mathcal{C}_1|\boldsymbol{x_i})\Big)^{1-l_i} \right\} \tag{7.25}$$

$$= -\sum_{i=1}^{n} \Big(l_i \ln \sigma(a_i) + (1 - l_i) \ln \Big\{ 1 - \sigma(a_i) \Big\} \Big). \tag{7.26}$$

where $a_i = \boldsymbol{w}^T \boldsymbol{x}_i$. Thus, $\boldsymbol{w}$ is found by minimising $E(\boldsymbol{w})$. The error function in Eq. (7.26) is also called the cross-entropy error function [Bis06, p.206]. It can be shown that the error function is convex and has a unique minimum. In order to show this, we will first derive the Hessian matrix of Eq. (7.26).

Lemma 7.4 (Hessian of Cross-entropy Error Function)
The Hessian matrix of the cross-entropy error function in Eq. (7.26) is

$$\boldsymbol{H} = \nabla\nabla E(\boldsymbol{w}) = \sum_{i=1}^{n} \sigma(a_i)(1 - \sigma(a_i)) \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{7.27}$$

Proof.
Using Property d) of Proposition 7.2 and the chain rule, the gradient of Eq. (7.26) with respect to $\boldsymbol{w}$ is

$$\nabla E(\boldsymbol{w}) = -\sum_{i=1}^{n} \Big\{ l_i \nabla \ln \sigma(a_i) + (1 - l_i) \nabla \ln \big(1 - \sigma(a_i)\big) \Big\} \tag{7.28}$$

$$= -\sum_{i=1}^{n} \Bigg\{ l_i \frac{\partial \ln \sigma(a_i)}{\partial \sigma(a_i)} \frac{\partial \sigma(a_i)}{\partial a_i} \nabla a_i + (1 - l_i) \frac{\partial \ln \big(1 - \sigma(a_i)\big)}{\partial \big(1 - \sigma(a_i)\big)} \frac{\partial \big(1 - \sigma(a_i)\big)}{\partial a_i} \nabla a_i \Bigg\} \tag{7.29}$$

$$= -\sum_{i=1}^{n} \left\{ \frac{l_i}{\sigma(a_i)} \sigma(a_i)(1 - \sigma(a_i)) \boldsymbol{x}_i - \frac{1 - l_i}{1 - \sigma(a_i)} \sigma(a_i)(1 - \sigma(a_i)) \boldsymbol{x}_i \right\} \tag{7.30}$$

$$= -\sum_{i=1}^{n} \Big\{ l_i (1 - \sigma(a_i)) \boldsymbol{x}_i - (1 - l_i) \sigma(a_i) \boldsymbol{x}_i \Big\} \tag{7.31}$$

$$= -\sum_{i=1}^{n} \Big\{ l_i \boldsymbol{x}_i - l_i \sigma(a_i) \boldsymbol{x}_i - \sigma(a_i) \boldsymbol{x}_i + l_i \sigma(a_i) \boldsymbol{x}_i \Big\} \tag{7.32}$$

$$= \sum_{i=1}^{n} (\sigma(a_i) - l_i) \boldsymbol{x}_i. \tag{7.33}$$

The Hessian matrix of Eq. (7.26) is found by computing an additional gradient of Eq. (7.33) with respect to $\boldsymbol{w}^T$. Doing this yields

$$\boldsymbol{H} = \nabla\nabla E(\boldsymbol{w}) = \sum_{i=1}^{n} \frac{\partial \sigma(a_i)}{\partial a_i} \boldsymbol{x}_i \nabla a_i = \sum_{i=1}^{n} \sigma(a_i)(1-\sigma(a_i))\boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{7.34}$$

■

Having obtained the Hessian matrix in Lemma 7.4, we can use it to show that the error function in Eq. (7.26) is convex.

Theorem 7.5 (Convexity of Cross-entropy Error Function)
The cross entropy error function defined in Eq. (7.26) is a convex function and has a unique minimum.

Proof.
The function is strictly convex if and only if its Hessian matrix is positive definite. From Lemma 7.4, the Hessian matrix can be written as [Bis06, pp. 207-208]

$$\boldsymbol{H} = \sum_{i=1}^{n} \sigma(a_i)(1-\sigma(a_i))\boldsymbol{x}_i \boldsymbol{x}_i^T = \boldsymbol{X}^T \boldsymbol{R} \boldsymbol{X} \tag{7.35}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n]^T$ and $\boldsymbol{R}$ is an $n \times n$ diagonal matrix with

$$\boldsymbol{R}_{ii} = \sigma(a_i)(1-\sigma(a_i)). \tag{7.36}$$

Recall that $\boldsymbol{H}$ is positive definite if and only if $\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u} > 0$ for all $\boldsymbol{u} \in \mathbb{R}^d \backslash \{\boldsymbol{0}\}$. Using Eq. (7.27), we can write

$$\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u} = \boldsymbol{u}^T \boldsymbol{X}^T \boldsymbol{R} \boldsymbol{X} \boldsymbol{u} = (\boldsymbol{X}\boldsymbol{u})^T \boldsymbol{R} \boldsymbol{X} \boldsymbol{u} = \boldsymbol{v}^T \boldsymbol{R} \boldsymbol{v} \tag{7.37}$$

where $\boldsymbol{v} = \boldsymbol{X}\boldsymbol{u}$. Recall from property a) in Proposition 7.2 that $0 < \sigma(a) < 1$ for any a. Thus, all entries in $\boldsymbol{R}$ are strictly positive, yielding that $\boldsymbol{v}^T \boldsymbol{R} \boldsymbol{v} > 0$ for all $\boldsymbol{v} \neq \boldsymbol{0}$. By Eq. (7.37), it follows that $\boldsymbol{H}$ is positive definite. Thus, the error function in Eq. (7.26) is a convex function of $\boldsymbol{w}$ and has a unique minimum. ■

Theorem 7.5 is an important result since it allows estimation of $\boldsymbol{w}$ using conventional algorithms for convex optimisation problems such as gradient-descent algorithms.

7.2 Multinomial Logistic Regression

The multinomial logistic regression model provides an extension of the logistic regression model to classify data when there are $k > 2$ classes. In this case, we will turn to representing the labels of each data point as latent variables. For a data point $\boldsymbol{x}$, we define its corresponding label as a k-dimensional vector $\boldsymbol{l} = [l_1 \quad l_2 \quad \cdots \quad l_k]^T$ such that

$$l_j = \begin{cases} 1, & \text{if } \boldsymbol{x} \in \mathcal{C}_j, \\ 0, & \text{otherwise.} \end{cases} \tag{7.38}$$

Akin to the binomial case, the distribution of the label can be modelled using a Multinoulli distribution. The distribution is again governed by the posterior probabilities $p(\mathcal{C}_j|\boldsymbol{x})$ for $j = 1, \ldots, k$ and the expected value of $\boldsymbol{l}$ is

$$E[\boldsymbol{l}] = \begin{bmatrix} p(\mathcal{C}_1|\boldsymbol{x}) & p(\mathcal{C}_2|\boldsymbol{x}) & \cdots & p(\mathcal{C}_k|\boldsymbol{x}) \end{bmatrix}^T. \tag{7.39}$$

By Bayes' Theorem, the j'th posterior probability can be written as [Bis06, p. 209]

$$p(\mathcal{C}_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_j)p(\mathcal{C}_j)}{\sum_{m=1}^{k} p(\boldsymbol{x}|\mathcal{C}_m)p(\mathcal{C}_m)} \tag{7.40}$$

$$= \frac{p(\boldsymbol{x}|\mathcal{C}_j)p(\mathcal{C}_j)/p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{1 + \sum_{m=1}^{k-1} p(\boldsymbol{x}|\mathcal{C}_m)p(\mathcal{C}_m)/p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)} \tag{7.41}$$

$$= \frac{e^{a_j}}{1 + \sum_{m=1}^{k-1} e^{a_m}} \tag{7.42}$$

with

$$a_j = \ln \frac{p(\boldsymbol{x}|\mathcal{C}_j)p(\mathcal{C}_j)}{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}. \tag{7.43}$$

The a_j defined above are called the activations of the logistic regression model and give the log-odds of being in the j'th cluster relative to the k'th cluster [Bis06, p. 209]. The k'th cluster is sometimes called the reference cluster since all log-odds are in reference to this cluster. The right hand side of Eq. (7.42) is a generalisation of the sigmoid function called the softmax function.

Definition 7.6 (Softmax Function)
Let $\boldsymbol{a} = [a_1 \quad a_2 \quad \cdots \quad a_k]^T$. The function

$$s(a_j) = s(a_j; \boldsymbol{a}) = \frac{e^{a_j}}{\sum_{m=1}^{k} e^{a_m}} \tag{7.44}$$

is called the softmax function.

We note that if the a_j are defined as in Eq. (7.43) then the softmax function is given by Eq. (7.42) since $a_k = 0$.

Proposition 7.7 (Derivative of Softmax Function)
For a chosen $h = 1, 2, \ldots, k$, the derivative of the softmax function is given as

$$\frac{\partial s(a_j)}{\partial a_h} = s(a_j)(I_{jh} - s(a_h)) \tag{7.45}$$

where I_{jh} is the (j, h)'th entry in the $k \times k$ dimensional identity matrix.

Proof.
In the case where $j \neq h$, we have that

$$\frac{\partial s(a_j)}{\partial a_h} = \frac{\partial}{\partial a_h} \frac{e^{a_j}}{\sum_{m=1}^{k} e^{a_m}} = e^{a_j} \frac{-1}{(\sum_{m=1}^{k} e^{a_m})^2} e^{a_h} = -s(a_j)s(a_h) \tag{7.46}$$

For the case where $j = h$, using the product rule yields

$$\frac{\partial s(a_h)}{\partial a_h} = \frac{e^{a_h}}{\sum_{m=1}^{k} e^{a_m}} + e^{a_h} \frac{-1}{(\sum_{m=1}^{k} e^{a_m})^2} e^{a_h} = s(a_h)(1 - s(a_h)). \tag{7.47}$$

The derivatives in Eqs. (7.46) and (7.47) can be combined such that [Bis06, p. 209]

$$\frac{\partial s(a_j)}{\partial a_h} = s(a_j)(I_{hj} - s(a_h)). \tag{7.48}$$

■

In addition, we define the vector-valued function

$$\boldsymbol{s}(\boldsymbol{a}) = \begin{bmatrix} s(a_1) & s(a_2) & \cdots & s(a_k) \end{bmatrix}^T. \tag{7.49}$$

We will assume that a_j can be written as a linear function of $\boldsymbol{x}$ i.e.

$$a_j(\boldsymbol{x}) = \boldsymbol{w}_j^T \boldsymbol{x} \tag{7.50}$$

where $\boldsymbol{w}_j$ contains the weights for the j'th label. As in the binary case, this assumption is valid when the posterior densities are part of the exponential family. Under this assumption, we can write

$$\boldsymbol{a} = \boldsymbol{W}\boldsymbol{x} \tag{7.51}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \cdots \quad \boldsymbol{w}_k]^T$ is a $k \times d$ matrix containing all weights of the model. Using this, we can define the multinomial logistic regression model.

Definition 7.8 (Multinomial Logistic Regression Model)
Let $\boldsymbol{x}$ be a d-dimensional data point belonging to one of the classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$. Let $\boldsymbol{l} = [l_1 \quad l_2 \quad \cdots \quad l_k]^T$ be a latent variable of $\boldsymbol{x}$ such that $l_j = 1$ if $\boldsymbol{x} \in \mathcal{C}_j$ and 0 otherwise. The relation

$$E[\boldsymbol{l}] = \boldsymbol{s}(\boldsymbol{W}\boldsymbol{x}) \tag{7.52}$$

with $\boldsymbol{s}(\boldsymbol{a})$ being the softmax function taken entry-wise on a is called a multinomial logistic regression model with $\boldsymbol{W} = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \cdots \quad \boldsymbol{w}_k]^T$ containing its parameters. The model is equivalently formulated as

$$p(\mathcal{C}_j|\boldsymbol{x}) = E[l_j] = s(\boldsymbol{w}_j^T\boldsymbol{x}), \quad j = 1, \ldots, k \tag{7.53}$$

with s being the softmax function.

In the binary case, we saw that the logistic regression model is a generalised linear model under the assumption of linearity. This is also the case for multinomial logistic regression, however it requires an extension of generalised linear models to the multivariate case. For more information on multivariate generalised linear models, we refer to [FT01]. For the case of multinomial logistic regression, we can alternatively formulate the model as

$$\boldsymbol{W}\boldsymbol{x} = \boldsymbol{g}(E[\boldsymbol{l}]) \tag{7.54}$$

where the link function $\boldsymbol{g}(\cdot)$ is the inverse of $\boldsymbol{s}(\cdot)$. It can be shown that Eq. (7.54) holds if we define the link function as

$$\boldsymbol{g}(E[\boldsymbol{l}]) = \begin{bmatrix} g_1(E[\boldsymbol{l}]) & g_2(E[\boldsymbol{l}]) & \cdots & g_k(E[\boldsymbol{l}]) \end{bmatrix}^T \tag{7.55}$$

where [FT01, p. 73]

$$g_j(E[\boldsymbol{l}]) = \ln \frac{p(\mathcal{C}_j|\boldsymbol{x})}{1 - \sum_m^{k-1} p(\mathcal{C}_m|\boldsymbol{x})}. \tag{7.56}$$

To see this, we use Eq. (7.52), the definition of $\boldsymbol{s}(\boldsymbol{a})$, and Bayes' theorem to rewrite Eq. (7.56) as

$$
\begin{aligned}
g_j(E[\boldsymbol{l}]) &= g_j(\boldsymbol{s}(\boldsymbol{a})) && (7.57)\\
&= \ln \frac{s(a_j)}{1 - \sum_m^{k-1} s(a_m)} && (7.58)\\
&= \ln \frac{p(\mathcal{C}_j|\boldsymbol{x})}{1 - \sum_m^{k-1} p(\mathcal{C}_m|\boldsymbol{x})} && (7.59)\\
&= \ln \frac{p(\boldsymbol{x}|\mathcal{C}_j)p(\mathcal{C}_j)/p(\boldsymbol{x})}{1 - \sum_m^{k-1} p(\boldsymbol{x}|\mathcal{C}_m)p(\mathcal{C}_m)/p(\boldsymbol{x})} && (7.60)\\
&= \ln \frac{p(\boldsymbol{x}|\mathcal{C}_j)p(\mathcal{C}_j)}{p(\boldsymbol{x}) - \sum_m^{k-1} p(\boldsymbol{x}|\mathcal{C}_m)p(\mathcal{C}_m)} && (7.61)\\
&= \ln \frac{p(\boldsymbol{x}|\mathcal{C}_j)p(\mathcal{C}_j)}{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)} = a_j && (7.62)
\end{aligned}
$$

which shows that $\boldsymbol{g}(\boldsymbol{s}(\boldsymbol{a})) = \boldsymbol{a}$. Recall that the label $\boldsymbol{l}$ was assumed to be multinomial Bernoulli distributed. As was done when discussing latent variables in the derivation of the EM algorithm in Chapter 3, the distribution of $\boldsymbol{l}$ given the weight matrix $\boldsymbol{W}$ can be expressed as

$$
p(\boldsymbol{l}|\boldsymbol{W}, \boldsymbol{x}) = \prod_{j=1}^{k} p(\mathcal{C}_j|\boldsymbol{x})^{l_j}. \tag{7.63}
$$

Let $\boldsymbol{l}_i$ be the label for the i'th data point $\boldsymbol{x}_i$ for $i = 1, \ldots, n$. In order to fit the multinomial logistic regression model, we define the label matrix as the $n \times k$ matrix $\boldsymbol{L} = [\boldsymbol{l}_1 \quad \boldsymbol{l}_2 \quad \cdots \quad \boldsymbol{l}_n]^T$ with elements l_{ij}. Assuming that the data points are mutually independent, we can write the likelihood function as

$$
p(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{X}) = \prod_{i=1}^{n} \prod_{j=1}^{k} p(\mathcal{C}_j|\boldsymbol{x}_i)^{l_{ij}}. \tag{7.64}
$$

with $\boldsymbol{X} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n]^T$. Taking the negative logarithm of Eq. (7.64) yields [Bis06, p. 209]

$$
E(\boldsymbol{W}) = -\ln p(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{X}) = -\sum_{i=1}^{n} \sum_{j=1}^{k} l_{ij} \ln p(\mathcal{C}_j|\boldsymbol{x}_i) = -\sum_{i=1}^{n} \sum_{j=1}^{k} l_{ij} \ln s(\boldsymbol{w}_j^T \boldsymbol{x}_i) \tag{7.65}
$$

which is the cross-entropy function for multinomial classification. The weight matrix $\boldsymbol{W}$ is thus found by minimisation of Eq. (7.65).

We will end by showing that similarly to the binary case, the error function in Eq. (7.65) is convex and thus it can be minimised using convex optimisaion algorithms.

Lemma 7.9 (Hessian of Multinomial Cross-entropy Error)
[Bis06, p. 210] Let $s_j(\boldsymbol{x}) = s(a_j) = s(\boldsymbol{w}_j^T \boldsymbol{x})$. The Hessian matrix of the cross entropy error function for multinomial logistic regression is a $kd \times kd$ block matrix where the (h, j)'th block is a $d \times d$ matrix given by

$$\boldsymbol{H}_{h,j} = \nabla_{\boldsymbol{w}_j} \nabla_{\boldsymbol{w}_h} E(\boldsymbol{W}) = \sum_{i=1}^{n} s_h(\boldsymbol{x}_i)(I_{hj} - s_j(\boldsymbol{x}_i))\boldsymbol{x}_i \boldsymbol{x}_i^T \tag{7.66}$$

Proof.
Using Proposition 7.7, the gradient of the cross entropy error function with respect to a chosen weight vector $\boldsymbol{w}_h$ is

$$\nabla_{\boldsymbol{w}_h} E(\boldsymbol{W}) = -\sum_{i=1}^{n} \left\{ \nabla_{\boldsymbol{w}_h} \sum_{j=1}^{k} l_{ij} \ln s_j(\boldsymbol{x}_i) \right\} \tag{7.67}$$

$$= -\sum_{i=1}^{n} \left\{ l_{ih} \nabla_{\boldsymbol{w}_h} \ln s_h(\boldsymbol{x}_i) + \sum_{j \neq h} l_{ij} \nabla_{\boldsymbol{w}_h} \ln s_j(\boldsymbol{x}_i) \right\} \tag{7.68}$$

$$= -\sum_{i=1}^{n} \left\{ l_{ih} \frac{\partial \ln s_h(\boldsymbol{x}_i)}{\partial s_h(\boldsymbol{x}_i)} \frac{\partial s_h(\boldsymbol{x}_i)}{\partial a_h} \nabla_{\boldsymbol{w}_h} a_h + \sum_{j \neq h} l_{ij} \frac{\partial \ln s_j(\boldsymbol{x}_i)}{\partial s_j(\boldsymbol{x}_i)} \frac{\partial s_j(\boldsymbol{x}_i)}{\partial a_h} \nabla_{\boldsymbol{w}_h} a_h \right\} \tag{7.69}$$

$$= -\sum_{i=1}^{n} \left\{ \frac{l_{ih}}{s_h(\boldsymbol{x}_i)} s_h(\boldsymbol{x}_i)(1 - s_h(\boldsymbol{x}_i))\boldsymbol{x}_i - \sum_{j \neq h} \frac{l_{ij}}{s_j(\boldsymbol{x}_i)} s_j(\boldsymbol{x}_i) s_h(\boldsymbol{x}_i) \boldsymbol{x}_i \right\} \tag{7.70}$$

$$= \sum_{i=1}^{n} \left\{ -l_{ih}(1 - s_h(\boldsymbol{x}_i))\boldsymbol{x}_i + \sum_{j \neq h} l_{ij} s_h(\boldsymbol{x}_i) \boldsymbol{x}_i \right\} \tag{7.71}$$

$$= \sum_{i=1}^{n} \left\{ -l_{ih}\boldsymbol{x}_i + l_{ih} s_h(\boldsymbol{x}_i)\boldsymbol{x}_i + \sum_{j \neq h} l_{ij} s_h(\boldsymbol{x}_i) \boldsymbol{x}_i \right\} \tag{7.72}$$

$$= \sum_{i=1}^{n} \left\{ -l_{ih}\boldsymbol{x}_i + \sum_{j=1}^{k} l_{ij} s_h(\boldsymbol{x}_i)\boldsymbol{x}_i \right\} = \sum_{i=1}^{n} \left\{ -l_{ih}\boldsymbol{x}_i + s_h(\boldsymbol{x}_i)\boldsymbol{x}_i \sum_{j=1}^{k} l_{ij} \right\} \tag{7.73}$$

$$= \sum_{i=1}^{n} \left\{ s_h(\boldsymbol{x}_i)\boldsymbol{x}_i - l_{ih}\boldsymbol{x}_i \right\} = \sum_{i=1}^{n} (s_h(\boldsymbol{x}_i) - l_{ih})\boldsymbol{x}_i \tag{7.74}$$

where we have used that $\sum_{j=1}^{k} l_{ij} = 1$. Taking a second derivative of Eq. (7.74)

with respect to a chosen weight vector $\boldsymbol{w}_j$ for $j = 1, 2, \ldots, k$ yields

$$\begin{aligned}
\boldsymbol{H}_{h,j} &= \nabla_{\boldsymbol{w}_j} \nabla_{\boldsymbol{w}_h} E(\boldsymbol{W}) && (7.75)\\
&= \sum_{i=1}^{n} \nabla_{\boldsymbol{w}_j} s_h(\boldsymbol{x}_i) \boldsymbol{x}_i && (7.76)\\
&= \sum_{i=1}^{n} \frac{\partial s_h(\boldsymbol{x}_i)}{\partial a_j} \boldsymbol{x}_i \nabla_{\boldsymbol{w}_j} a_j && (7.77)\\
&= \sum_{i=1}^{n} s_j(\boldsymbol{x}_i)\big(I_{jh} - s_h(\boldsymbol{x}_i)\big)\boldsymbol{x}_i \boldsymbol{x}_i^T && (7.78)
\end{aligned}$$

The Hessian matrix of the error function in Eq. (7.65) is thus a $kd \times kd$ block matrix where the (h, j)'th block is a $d \times d$ matrix given by Eq. (7.78). ■

Theorem 7.10 (Convexity of Multinomial Cross-entropy Error)
The cross-entropy error function for the multinomial case defined in Eq. (7.65) is convex.

Proof.
Similar to the proof for the binary case, we will prove that the Hessian matrix in Lemma 7.9 is positive-semidefinite. Let $\boldsymbol{H}$ be the $kd \times kd$ Hessian block matrix with its (h, j)'th block defined as in Eq. (7.78) and let $\boldsymbol{u} = [\boldsymbol{u}_1^T \quad \boldsymbol{u}_2^T \quad \cdots \quad \boldsymbol{u}_k^T]^T \in \mathbb{R}^{kd} \setminus \{\boldsymbol{0}\}$ where $\boldsymbol{u}_j \in \mathbb{R}^d$ for $j = 1, 2, \ldots, k$. Then the Hessian matrix $\boldsymbol{H}$ is positive-semidefinite if and only if

$$\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u} = \sum_{j=1}^{k} \sum_{h=1}^{k} \boldsymbol{u}_h^T \boldsymbol{H}_{h,j} \boldsymbol{u}_j \geq 0. \tag{7.79}$$

By insertion of Eq. (7.78), we can write

$$\begin{aligned}
\sum_{j=1}^{k} \sum_{h=1}^{k} \boldsymbol{u}_h^T \boldsymbol{H}_{h,j} \boldsymbol{u}_j &= \sum_{j=1}^{k} \sum_{h=1}^{k} \boldsymbol{u}_h^T \Bigg(\sum_{i=1}^{n} s_h(\boldsymbol{x}_i)\big(I_{hj} - s_j(\boldsymbol{x}_i)\big)\boldsymbol{x}_i \boldsymbol{x}_i^T \Bigg) \boldsymbol{u}_j && (7.80)\\
&= \sum_{j=1}^{k} \sum_{h=1}^{k} \sum_{i=1}^{n} s_h(\boldsymbol{x}_i)\big(I_{hj} - s_j(\boldsymbol{x}_i)\big) \boldsymbol{u}_h^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j && (7.81)\\
&= \sum_{i=1}^{n} \sum_{h=1}^{k} s_h(\boldsymbol{x}_i) \boldsymbol{u}_h^T \boldsymbol{x}_i \boldsymbol{x}_i^T \sum_{j=1}^{k} \big(I_{hj} - s_j(\boldsymbol{x}_i)\big) \boldsymbol{u}_j. && (7.82)
\end{aligned}$$

We wish to show that every term in the outer sum of Eq. (7.82) is non-negative. We note that

$$\sum_{j=1}^{k} (I_{hj} - s_j(\boldsymbol{x}_i))\boldsymbol{u}_j = \boldsymbol{u}_h - \sum_{j=1}^{k} s_j(\boldsymbol{x}_i)\boldsymbol{u}_j \tag{7.83}$$

since $\sum_{j=1}^{k} I_{hj}\boldsymbol{u}_j = \boldsymbol{u}_h$. Using Eq. (7.83), we have that for every i

$$\sum_{h=1}^{k} s_h(\boldsymbol{x}_i)\boldsymbol{u}_h^T\boldsymbol{x}_i\boldsymbol{x}_i^T \sum_{j=1}^{k} (I_{hj} - s_j(\boldsymbol{x}_i))\boldsymbol{u}_j \tag{7.84}$$

$$= \sum_{h=1}^{k} s_h(\boldsymbol{x}_i)\boldsymbol{u}_h^T\boldsymbol{x}_i\boldsymbol{x}_i^T \left(\boldsymbol{u}_h - \sum_{j=1}^{k} s_j(\boldsymbol{x}_i)\boldsymbol{u}_j\right) \tag{7.85}$$

$$= \sum_{h=1}^{k} s_h(\boldsymbol{x}_i)\boldsymbol{u}_h^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{u}_h - \sum_{h=1}^{k} s_h(\boldsymbol{x}_i)\boldsymbol{u}_h^T\boldsymbol{x}_i\boldsymbol{x}_i^T \sum_{j=1}^{k} s_j(\boldsymbol{x}_i)\boldsymbol{u}_j \tag{7.86}$$

$$= \sum_{h=1}^{k} s_h(\boldsymbol{x}_i) f_i(\boldsymbol{u}_h) - f_i\left(\sum_{h=1}^{k} s_h(\boldsymbol{x}_i)\boldsymbol{u}_h\right) \tag{7.87}$$

where we have defined the function $f_i(\boldsymbol{u}) = \boldsymbol{u}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{u}$. For every i, we have that

$$f_i(\boldsymbol{u}) = \boldsymbol{u}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{u} = (\boldsymbol{u}^T\boldsymbol{x}_i)(\boldsymbol{u}^T\boldsymbol{x}_i) = (\boldsymbol{u}^T\boldsymbol{x}_i)^2 \geq 0. \tag{7.88}$$

Thus, the function $f_i(\boldsymbol{u})$ is convex. Using that $\sum_{h=1}^{k} s_h(\boldsymbol{x}_i) = \sum_{h=1}^{k} p(\mathcal{C}_h|\boldsymbol{x}_i) = 1$, we have from Jensen's inequality that [Bis06, p. 56]

$$f_i\left(\sum_{h=1}^{k} s_h(\boldsymbol{x}_i)\boldsymbol{u}_h\right) \leq \sum_{h=1}^{k} s_h(\boldsymbol{x}_i) f_i(\boldsymbol{u}_h). \tag{7.89}$$

Using this, we have that Eq. (7.87) is non-negative for every i, meaning that Eq. (7.82) is non-negative. Thus, $\boldsymbol{H}$ is positive-semidefinite showing that the cross-entropy error function in Eq. (7.65) is convex. ■

8. Data and Pre-processing

In order to obtain parameters for a statistical model on bike share traffic patterns, it is necessary to obtain data on bike share trips.

Through an extensive search we have found 19 cities which have bike share trip data openly available. These cities are listed in Table 8.1. As seen in the table, the bike share systems have varying sizes and amounts of traffic. The number of stations in these systems lies between 56 and 938 in 2019, while the number of trips ranges from just 123 thousand trips among 163 stations to 26 million among 399 stations. The systems also vary both in density and extent. In order to narrow the scope of the project, we chose 8 cities which were used in the modelling: 4 from the United States and 4 from Europe. From the US, these cities are in order of system size New York City, Chicago, and Washington DC, and Boston while the cities from Europe are London, Helsinki, Oslo, and Madrid. The cities were chosen both based on the sizes and maturity of their bike share systems and availability of other data used in the analysis. For the purpose of illustration, examples from the analysis will be based on the bike share system in New York City, as this is a large system with many stations and trips which has been the subject of many prior studies [Chi+20; FIE16; OS15].

In addition to bike share trip data, we have compiled other types of data which do not pertain to the bike share systems directly. These types of data include land use data of the cities, population data, and location data of transit systems in the cities as well as city centers. As these types of data are external to the bike sharing systems we will refer to them as external data. The sources of the data are listed in Table 8.2. The different types of data are described in the subsequent sections.

8.1 Bike Share Trip Data

The bike share trip data is obtained directly from the websites of the individual bike sharing providers or from city data portals. In this project, we use data sets from 2019 as this is the most recent year with normal operation prior to the COVID-19 pandemic. All of the data sets used contain data on every individual trip made in the system including trip duration, time of departure from the start station, start station ID, start station name, time of arrival at 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 8.2.

For cities in the US, the data sets also include the type of user which used the bicycle, primarily split between subscribers who pay an annual subscription fee and casual users who pay for individual trips or for a limited time of use.

City	System name	# stations	# trips 2019	Reference
Bergen	Bergen Bysykkel	90	898 276	[Ber]
Boston	Bluebikes	341	2 522 771	[Blu]
Buenos Aires	EcoBici	417	5 238 643	[Ecoa]
Chicago	Divvy	593	3 614 078	[Div]
Edinburgh	Just Eat Cycles	163	123 684	[Jus]
Guadalajara	MiBici	275	4 625 130	[MiB]
Helsinki	Helsinki City Bikes	348	3 784 877	[Hel]
London	Santander Cycles	753	8 829 104	[Trab]
Los Angeles	Metro Bike Share	234	276 943	[Met]
Madrid	BiciMad	214	3 956 099	[EMT]
Mexico City	EcoBici	480	8 349 075	[Ecob]
Minneapolis	Nice Ride	179	263 169	[Nic]
Montreal	Bixi	619	5 442 288	[BIX]
New York City	Citi Bike	938	20 551 396	[Cita]
Oslo	Oslo Bysykkel	254	2 237 092	[Osl]
San Francisco	Bay Wheels	351	2 296 199	[Lyf]
Taipei	YouBike	399	26 484 903	[You]
Trondheim	Trondheim Bysykkel	56	356 189	[Tro]
Washington, D.C.	Capital Bikeshare	429	3 281 231	[Cap]

Table 8.1: Bike share systems around the world, including number of stations and number of trips in 2019. The systems chosen for this project are highlighted.

Data set	Area	Provider	Reference
Trip Data	New York City	Citi Bike	[Cita]
Trip Data	Chicago	Divvy Bikes	[Div]
Trip Data	Washington D.C.	Capital Bikeshare	[Cap]
Trip Data	Boston	Bluebikes	[Blu]
Trip Data	London	Transport for London	[Trab]
Trip Data	Helsinki	Helsinki Region Transport	[Hel]
Trip Data	Oslo	Oslo City Bike	[Osl]
Trip Data	Madrid	BiciMad	[EMT]
Station Data	Chicago	City of Chicago	[Citc]
Station Data	Washingon D.C.	Department of Real Estate Services	[Citd]
Station Data	London	Transport for London	[Traa]
Station Data	Madrid	BiciMad	[EMT]
Land Use Data	New York City	New York City Department of City Planning	[New]
Land Use Data	Chicago	City of Chicago	[Citb]
Land Use Data	Washington D.C.	Government of the District of Columbia	[Cite]
Land Use Data	Boston	Boston Planning and Development Agency	[Bos]
Land Use Data	Europe	European Enviroment Agency	[Eurc]
Population Data	US	US Census Bureau	[USCB]
Population Data	Europe	European Enviroment Agency	[Eurc], [Eurb]
Transit Data	All cities	OpenStreetMap	[Opea]
City Centers	All cities	OpenStreetMaps	[Opea]

Table 8.2: Data sources.

8.2 Station Service Area

Each station in the system is assigned a designated service area. These service areas are determined using a Voronoi tessellation which defines boundaries such that land is assigned to the closest station. The areas are then truncated such that there is a maximum of 500 meters from the station to the furthest point in its service area in Euclidean distance. This distance was measured using an azimuthal equidistant projection centered at a predetermined center point of the city. We choose 500 meters as this is assumed to be the maximum distance people are willing to walk to get to a station. This distance also conforms to general consensus in the US of having transit-oriented development planning areas extend between a quarter-mile and a half-mile from a transit station [FTA02, p. 78]. Euclidean distance is used as the bike sharing systems are in urban areas with a high density of roads and intersections. Having a highly interconnected road network makes the Euclidean distance more likely to be a valid metric since it will not be significantly shorter than the conventional network distance following the road network, as suggested by [OCB14]. This simplification is further justified if we assume that a user will walk the most direct route to a station without taking detours. The service areas are further truncated such that they do not span over bodies of water such as seas, rivers, and lakes when possible. This is done by intersection with polygons from the land use data described below.

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 system during the year of 2019. However, at no point in time has the system had 938 active stations simultaneously, since some stations have been created, relocated and/or removed entirely. Thus, calculating 938 service areas will not give a representative view of the system and how it was used at any given time. To account for this, we calculate the service areas of the system in each day of the year. An example of a map of the stations in New York City and their service areas for one day can be seen in Fig. 8.1.

The relocation and removal of stations also affects other variables which are derived from the service areas and the location of the stations. These variables include the population density around the station, land use, and distance to nearest transit points. To alleviate this, for each station all variables used in the modelling are calculated for each day the station has been used and then averaged over those days.

8.3 Land Use Data

For US cities, land use is obtained from zoning data provided by the cities. The data contains polygons delineating 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

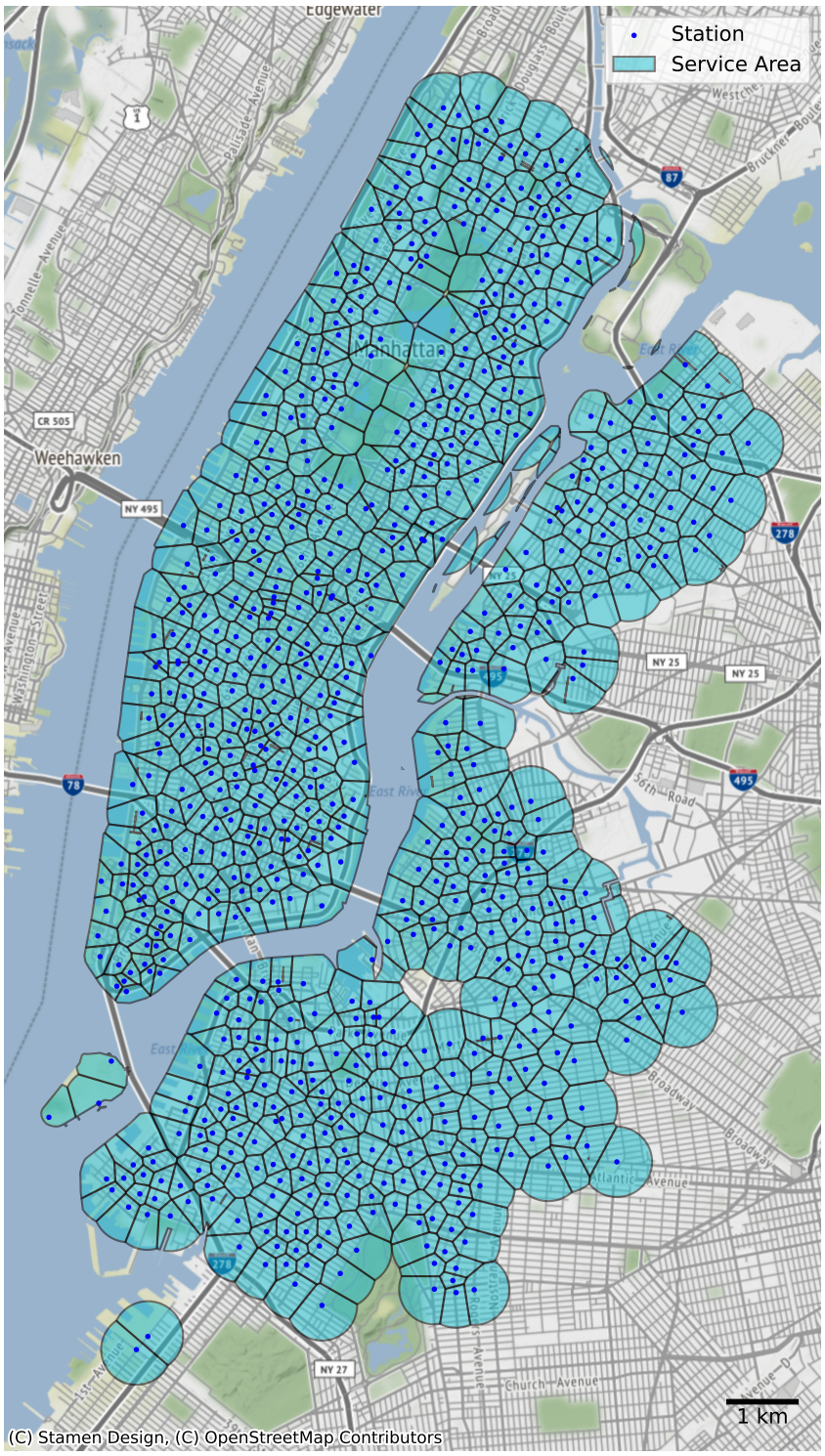


Figure 8.1: Service areas for New York City on October 23rd 2019.

historical zoning data was found, we use the most recent data as of October 2021 provided by the cities. It is probable that the zoning has changed from 2019 till 2021. However, we operate under the assumption that the changes in this time-frame were relatively minor with regard to the general ridership of the bike sharing systems.

For European cities, zoning data is not available in a standardised form as land use regulations differ between countries and regions. Instead, we use land use data from Urban Atlas 2018 in the Copernicus Land Monitoring Service provided by the European Environment Agency. This data includes polygons representing different land areas and a general description of their use. These areas are then classified into the same categories as the US cities. We again assume no significant land use changes from 2018 to 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. We deem this as appropriate since the zoning data in the US does not separate roads from zoning areas.

8.4 Population Data

The United States Census Bureau provides historical census data for 2019 on census tract level and polygons of the census tracts. Using this data, we calculate the population density of each census tract in number of people per 100 square meters. For Helsinki, Oslo and Madrid, population estimates are provided for each polygon in the land use data from Urban Atlas 2018. For London, population estimates from Urban Atlas 2012 were used instead due to discrepancies found in the population data from Urban Atlas 2018. We note that land use polygons from 2012 may differ from those in 2018 but they will only be used to estimate the population density in the station service area and not for land use.

We calculate the population density of each station's service area as a weighted average of the population densities of the census tracts or land use polygons within the service area, where the weights are the polygons' share of the station service area.

8.5 Transit Data

Transit data is obtained using the Overpass API from OpenStreetMap. The data contains locations of metro stations and railway stations.

8.6 City Centers

The center of each city is obtained using the Overpass API from OpenStreetMap. According to the OpenStreetMap wiki, city centers are located at places like cen-

tral squares, central administrative or religious buildings and central road junctions [Opeb]. As OpenStreetMap data is created by user contributions, the city center locations represent the consensus among the contributors. The centers have also been checked and deemed sensible by the authors.

8.7 Pre-processing of Data

Several considerations have been made in terms of the trip data and stations used in the modelling. We are primarily concerned with trips taken on business days since the traffic patterns generally are more predictable due to commuting. This is also done to ease comparison between cities since we expect the traffic patterns in the business days to be similar for all cities. We also aim to remove as many trips as possible which are considered recreational trips since these trips act as noise in the commuter traffic patterns and the models. Thus, trips which do not start on a business day are removed from the trip data along with trips which were taken on holidays where people might use the bike sharing system more leisurely. Users who are not subscribed to the bike sharing system are also more likely to use the bikes leisurely rather than for commuting purposes, as shown by [NSG19]. Leisure trips are also responsible for the majority of loop trips i.e. trips starting and ending at the same station, as was observed by [ZWD15] and [NSG19]. Therefore, we remove trips taken by casual users in cities where the user type is present in the data as well as loop trips.

Typically, the providers of the bike sharing data remove trips taken by staff for maintenance as well as trips taken to and from test stations. Trips which are up to 60 seconds long are also removed by some providers since these trips are assumed to be false starts or users ensuring that the bike is locked. However, the pre-processing of the data done by the providers is not consistent for all systems. Thus, to ensure consistency in the trip data we remove all trips with a duration of up to 60 seconds. We also remove stations which appear to be test stations or otherwise used for maintenance purposes.

Some bike share stations have a very low amount of traffic. If a station has only 1 or 2 trips per day, then the traffic pattern of that station may be more erratic than a typically used station. This can pose as a problem in the clustering since it can create outlier clusters consisting of low traffic stations. In addition, calculations of external variables derived from the stations location and service area are also prone to errors. The bike share trip data only provides the timestamps when a trip has taken place and not the days in which each station was active but unused. Thus, the external variables are calculated for each day the station was used and not necessarily each day the station was in service, meaning that days were there are no trips to or from the station are not counted when averaging and may give a misrepresentation of the external variables. This issue is only significant for low-traffic stations which only receive few trips a day since they are more probable to receive no trips on a day they were active. Therefore, in the modelling we exclude stations which have less than 8

daily trips on average. This number is chosen based on preliminary testing to provide a balance between clear traffic patterns while not excluding too many stations. The number of trips and stations removed after excluding low-traffic stations can be seen in Table 8.3.

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 DC	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 8.3: Number of trips and stations retained after removing low-traffic stations.

9. Modelling Approach

In this chapter, we describe our approach to modelling both the shape and volume of the average daily traffic of the bike share stations. The modelling can be done in two parallel stages: The first stage models the shape of the traffic by clustering the normalised traffic of each station and then relating this clustering to the external data variables discussed in the previous chapter using a Logistic Regression (LR) model. The second stage models the average amount of daily trips to and from each station also using the external data as predictor variables using a Generalised Linear Model (GLM). An overview of our modelling approach can be seen in Fig. 9.1.

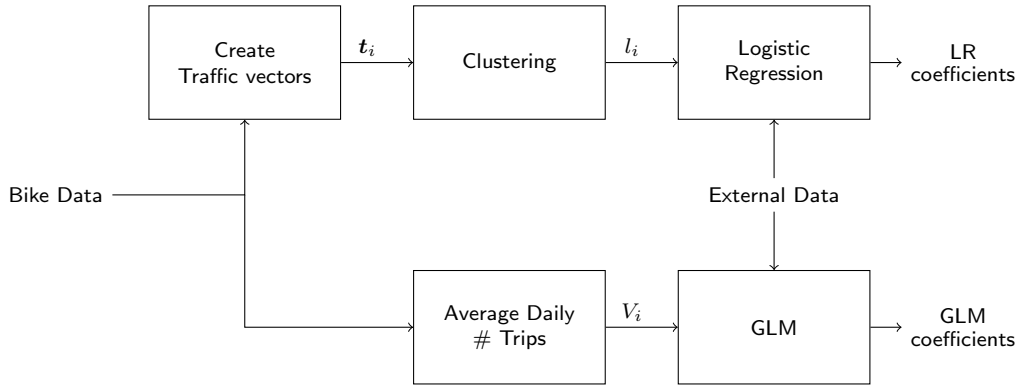


Figure 9.1: Flowchart of the modelling approach.

Using the bike share trip data, we calculate the hourly number of arrivals and departures for each station every day in which the station was used. The number of arrivals and departures for a specific hour are counted from the start of the hour to the end of the hour e.g. for hour 16 the arrivals and departures are counted from 16:00:00 to 16:59:59. Previous studies have established that a resolution of one hour yields a good trade-off between temporal resolution and fluctuations [Bor+11]. If the temporal resolution is higher, the patterns will be more different day-to-day, while a lower temporal resolution will smooth the traffic pattern and obscure the peaks in the pattern.

Let D_i be the set of days where station i has been used and let $\boldsymbol{d}_{d,i}$ and $\boldsymbol{a}_{d,i}$ be two 24-dimensional vectors representing the hourly number of departures and arrivals respectively for the station at day $d \in D_i$ starting at hour 0. Both traffic vectors are then averaged over all days such that

$$\boldsymbol{d}_i = \frac{1}{|D_i|} \sum_{d \in D_i} \boldsymbol{d}_{d,i} \qquad \text{and} \qquad \boldsymbol{a}_i = \frac{1}{|D_i|} \sum_{d \in D_i} \boldsymbol{a}_{d,i}$$

represent the average daily number of departures and arrivals for station i respectively. We define the average traffic volume (or demand) of a station as

$$V_i = \|\boldsymbol{d}_i\|_1 + \|\boldsymbol{a}_i\|_1 \tag{9.1}$$

i.e. the number of departures and arrivals of the station on an average day. When clustering, we normalise the average daily traffic of each station such that

$$\boldsymbol{d}'_i = \frac{\boldsymbol{d}_i}{V_i} \qquad \text{and} \qquad \boldsymbol{a}'_i = \frac{\boldsymbol{a}_i}{V_i} \tag{9.2}$$

yield the share of traffic for each hour on an average day. This is done since we are primarily concerned with clustering the stations based on the shape of their average daily traffic rather than the absolute amount of traffic. For instance, a highly trafficked station may be clustered differently from another station due to the higher amount of traffic despite them both being of the same type.

In order to mitigate the effect of the concentration of trips in the rush hours on the traffic pattern, we also redefine our traffic vectors as the hourly differences between the normalised number of arrivals and departures. We define the final traffic vector for station i as the 24-dimensional vector containing the relative difference in departures and arrivals for each hour

$$\boldsymbol{t}_i = \boldsymbol{d}'_i - \boldsymbol{a}'_i. \tag{9.3}$$

The vectors obtained from Eq. (9.3) are assigned to k clusters using one of the clustering algorithms discussed in Chapters 2 to 4. The labels obtained from the clustering, l_i, for each station are then used as response variables to train a LR model described in Chapter 7 while using the external data discussed in Chapter 8 as predictor variables. These variables are also used in the GLM where the V_i are used as response variables. The summary statistics of the external data variables on service area level can be seen in Table 9.1.

Combining the two models is straightforward. From the LR model, the shape of the traffic can be predicted, while the volume is predicted by the GLM. Multiplying the mean vector of the cluster which the station is predicted to be in with the predicted volume will result in a prediction in how many arrivals and departures the station will receive in each hour.

	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 [per 100 sq. m]	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 [per 100 sq. m]	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 [per 100 sq. m]	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 [per 100 sq. m]	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 9.1: Summary statistics of the variables used in the model.

9.1 Preliminary Clustering Analysis

The choice of clustering method and parameters is an important part of our modelling approach and will have a significant impact on future results. When doing clustering of data, three key specifications need to be set beforehand. These are:

- The clustering algorithm
- The distance measure between points (and possibly clusters)
- The number of clusters

The specifications above are evaluated using the cluster validation measures presented in Chapter 5. The choice of algorithm, distance measure and number of clusters all affect these measures which makes finding an optimal combination of these choices particularly difficult and finding a combination which is optimal for all cities close to impossible. The goal is to identify a clustering method which is considered a good fit for all cities and useful for further analysis while not necessarily optimal for all cities.

In order to test the choice of clustering algorithm, we clustered the stations in each city for different choices of the number of clusters. In both the k-means and EM algorithm, the distance measure was defined as the l_2-norm as in their definitions. For k-medoids and hierarchical clustering, we used an implementation of the Dynamic Time Warping (DTW) algorithm which was also used in [SLM15]. The DTW algorithm is typically used to compare time-series data while also taking temporal displacement into account. In the case of bike sharing stations, having rush hours at different times can cause this displacement. An introduction to the DTW algorithm can be found in [Sen08]. The reason behind using different distance measures is based on the design of the algorithms. The k-means algorithm is defined by minimising an l_2-norm and thus using this distance gives this algorithm a particular advantage. However, the strength of the k-medoids and hierarchical clustering algorithms is their flexibility with respect to the distance measure. Therefore, while k-means always minimises the l_2-norm, the performance of k-medoids and hierarchical clustering can be improved by using a different distance measure such as DTW, and possibly exceed the performance of the k-means algorithm.

For the hierarchical clustering algorithm we define the distance between clusters as the average distance between points in each cluster as described by Eq. (4.3).

A comparison between the clustering algorithms for all cities can be seen in Fig. 9.2. Both the silhouette and Davies-Bouldin index indicate a clear ordering between the algorithms in terms of their performance with the hierarchical clustering algorithm having the best performance while the EM algorithm has the worst. The hierarchical clustering algorithm also scores significantly better in the Dunn index for low choices of k. The EM algorithm seems to be either on par or significantly worse

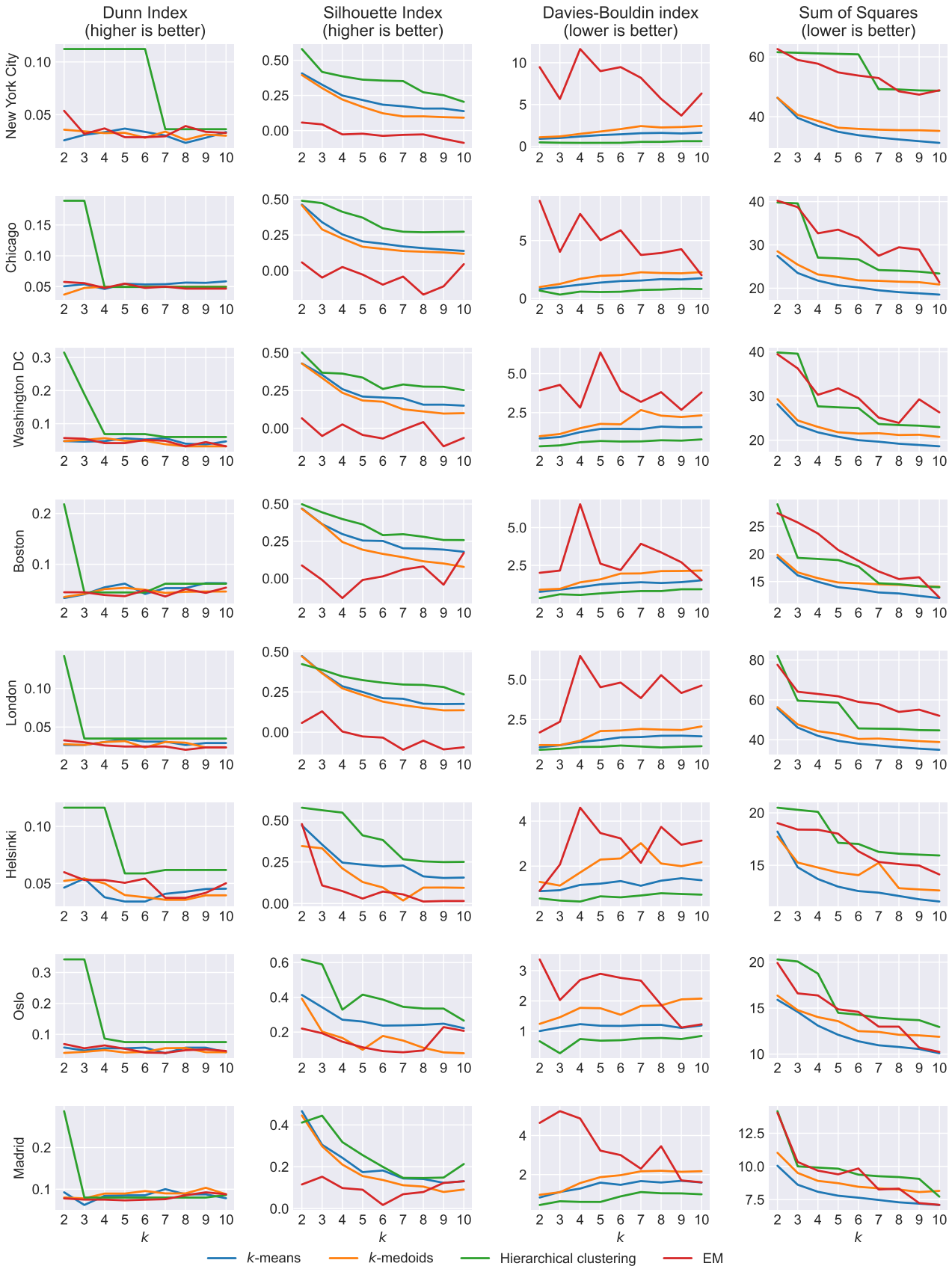


Figure 9.2: Comparison of the clustering algorithms for different choices of k for each city.

than the other algorithms. This is curious as the EM algorithm can be seen as a generalisation of the k-means algorithm under the assumption of normally distributed data points. The algorithm was initialised with cluster centers obtained from the k-means algorithm and was executed 100 times, picking the result with the highest likelihood by Eq. (3.23). The reason for the low performance may be due to the normality assumptions not being valid and/or error in estimation of the covariance matrices. Constraining the algorithm to only estimate the covariance matrices as scaled identities, and thus assuming spherical clusters of data points, leads to similar performance to the k-means algorithm. This does indicate that the data points are approximately spherically clustered although it is unclear why this was not picked up on by the EM algorithm. Another explanation could be that the measures used to evaluate the clustering all prefer spherical clusters and thus constraining the EM algorithm to provide spherical clusters gives better results in the eyes of these measures. The clustering from the EM algorithm may be deemed better when using a measure which is more relaxed on the shape of the clusters.

From Fig. 9.2, it seems that the hierarchical clustering algorithm fits best to the data. However, observing the clusters resulting from the algorithm reveals a large number of singleton clusters as well as clusters only containing a couple of stations. This may be attributed to the chaining effect discussed in Chapter 4 although these small sized clusters were also found using the complete linkage distance measure defined in Eq. (4.2). In the analysis in this project, we are not interested in clusters consisting of a low number of outlier traffic pattern stations, but would rather obtain clusters which better represent the general traffic patterns and are more balanced in number of stations. Therefore, we deem that the hierarchical clustering algorithm is not suitable for this matter. The k-means algorithm does not exhibit this behaviour and we will thus use this in the clustering experiments.

Having chosen the clustering algorithm, the number of clusters remains to be decided. Fig. 9.3 shows the score of the validation indices for different k when using the k-means algorithm. Both the Davies-Bouldin and silhouette index favor few clusters. For the sum of squares, we also observe elbows for $3 \leq k \leq 5$ in most cities. For values of k in this interval, the Dunn index peaks at $k = 5$ for all cities except Helsinki where it peaks at $k = 3$. We note that in terms of bike sharing there is a trade-off between the number of clusters and the ability to compare clusters between cities. Fewer clusters will yield a lower resolution of the clustering but the clustering will likely be similar between cities yielding easy comparisons. Conversely, having many clusters may result in different types of clusters across the cities due to local factors having a higher impact on the clustering. An argument can also be made for preferring an uneven amount of clusters if we assume that at a given time, some clusters will have a high amount of departures while other clusters will have a high amount of arrivals. With an uneven amount of clusters this leaves space for a neutral reference cluster with similar amounts of arrivals and departures. With these considerations we deem that $k = 5$ is reasonable for the clustering.

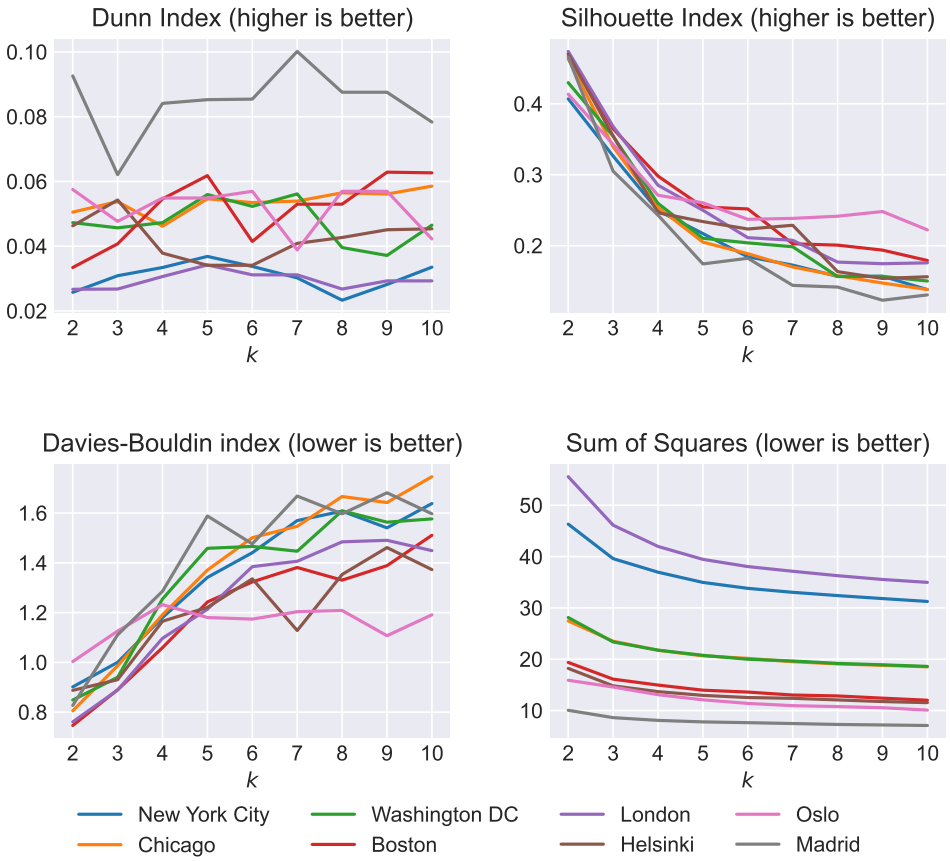


Figure 9.3: Davies-Bouldin, Dunn, and silhouette indices, as well as SSE for all cities and different choices of k using the k-means algorithm.

10. Clustering Results

In this chapter, we present results obtained from the k-means clustering of the traffic patterns of stations in the bike share systems with $k = 5$ as specified in Chapter 9 and section 9.1. The cluster centers for each city are seen in Fig. 10.1 and the number of stations in each cluster is shown in Table 10.1. The centers generally follow 5 types of clusters:

Low morning source: Traffic mostly concentrated around rush hours with a clear separation between the amount of departures and the amount of arrivals. The morning rush hours are dominated by departures while the evening rush hours are dominated by arrivals.

Low morning sink: Traffic mostly concentrated around rush hours with a clear separation between the amount of departures and the amount of arrivals. The morning rush hours are dominated by arrivals while the evening rush hours are dominated by departures.

High morning source: Very high concentration of traffic around rush hours with a large separation between the amount of departures and the amount of arrivals. The morning rush hours are dominated by departures while the evening rush hours are dominated by arrivals.

High morning sink: Very high concentration of traffic around rush hours with a large separation between the amount of departures and the amount of arrivals. The morning rush hours are dominated by arrivals while the evening rush hours are dominated by departures.

Reference: Cluster which does not follow the pattern of the previous types. Used as reference in LR modelling.

For most cities, the reference cluster contains stations which have approximately the same number of departures and arrivals for any given time. However, in Oslo there is a notable absence of this type of cluster. Instead, Oslo has a cluster where departures and arrivals are balanced in the morning while in the afternoon and throughout the evening most of the trips are departures. 23 stations or 9.2% of Oslo's system is comprised of these types of stations and a large majority of these

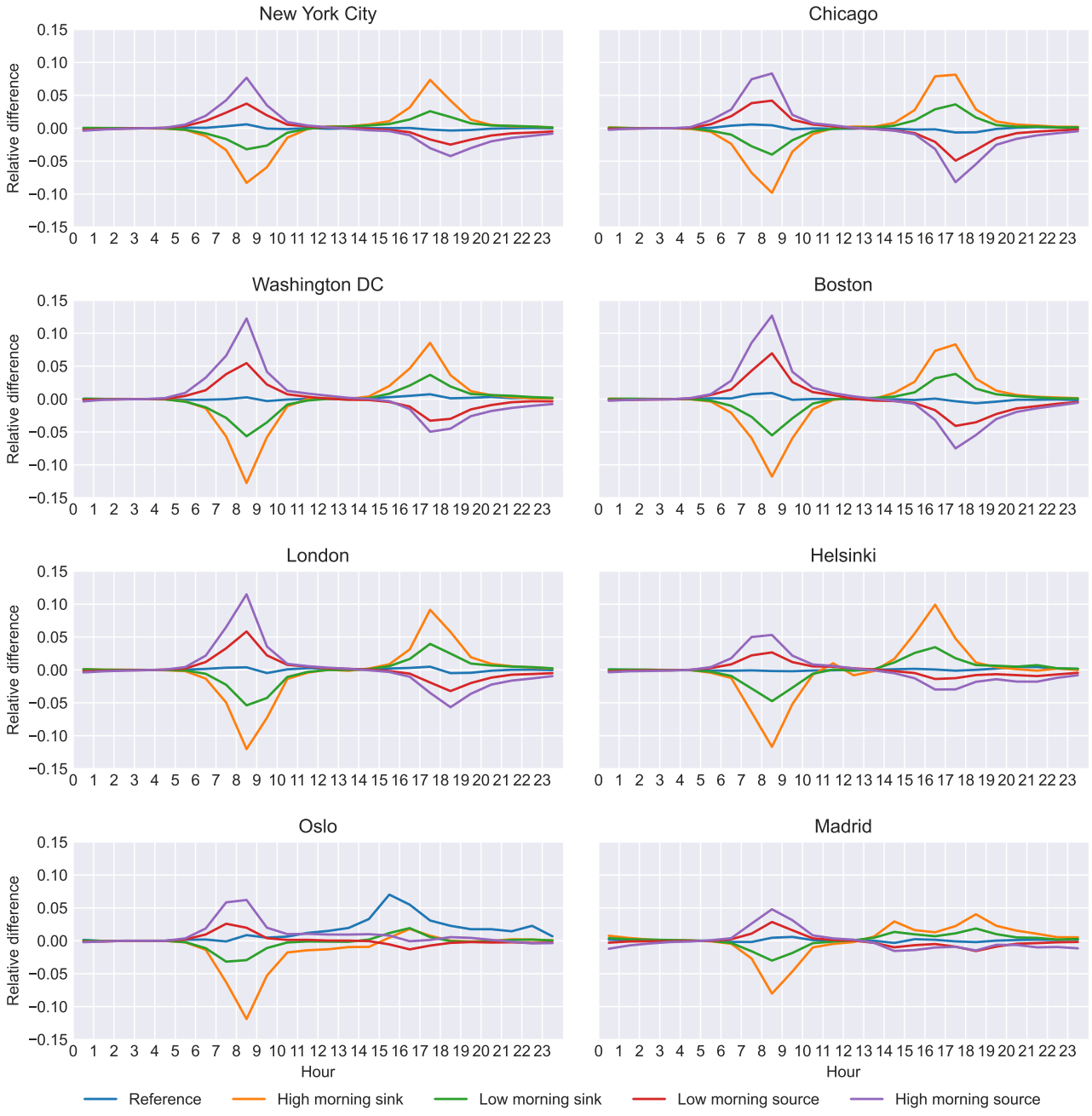


Figure 10.1: Cluster centers for all cities.

City	Reference	High morning sink	Low morning sink	Low morning source	High morning source
NYC	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%)
Wash. 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 10.1: Size of the 5 clusters obtained from the clustering. The size is represented as a percentage of the total number of stations below.

stations are located in or near the Ullevål district which contains Oslo University, see Fig. 10.2. 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. If that is the case, it shows how public transit and the bike sharing system in Oslo complement each other for the use-case of the university students. With the limits of the available open data it is not possible to confirm or deny this hypothesis. Therefore, it might be relevant to either do surveys among Oslo University bike share users or obtain other types of data such as bike share data with user ids from the provider or trip data from other forms of public transport in order to further investigate this imbalance.

There are also apparent differences in the non-reference clusters across cities. Madrid stands out in this regard, due to having three peaks in the daily traffic: one from 7:00 to 10:00, another from 14:00 to 15:00 and a third from 17:00 to 20:00. This is likely a consequence of the work culture in Madrid where it is customary to take a break from work around midday and resume the workday at a later time. In New York City, Chicago, Washington DC, Boston, and London the peaks of the morning sources and the valleys of the morning sinks are mostly aligned at 8:00, meaning that most of the morning commute in these cities is done from 8:00 to 9:00. However, in

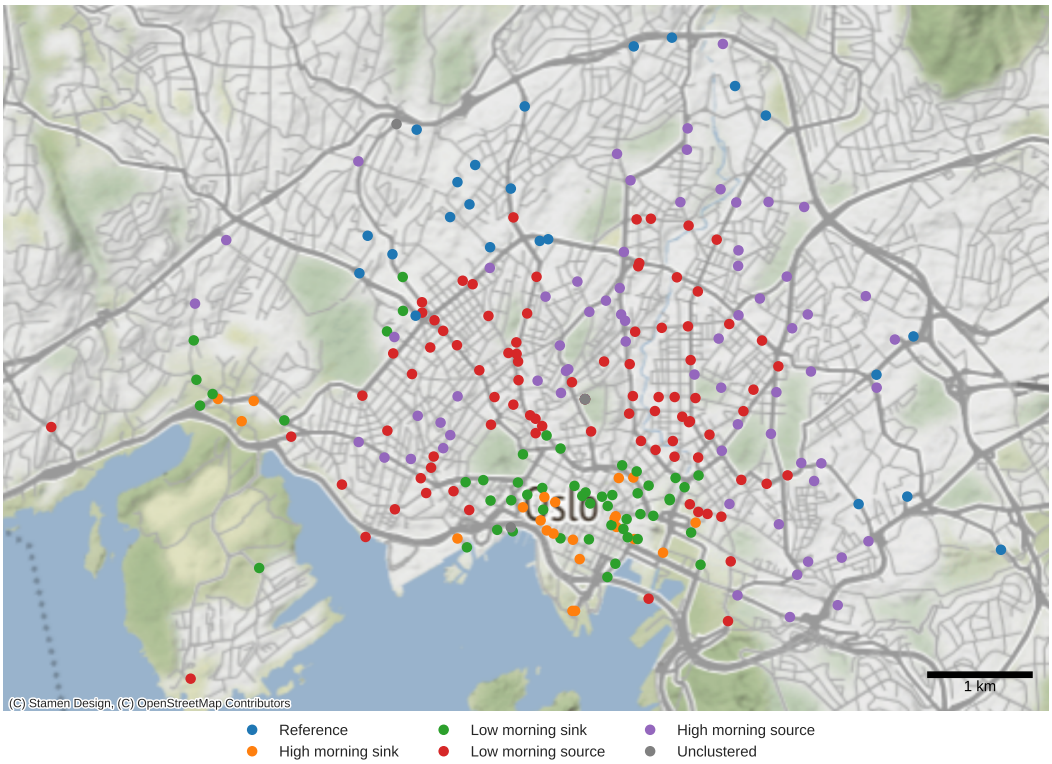


Figure 10.2: Clustered stations in Oslo. Grey stations are unclustered due to low traffic.

the afternoon commute there is a noticeable one hour shift between the peaks of the morning sinks and the valleys of the morning sources. This may be an effect of a more relaxed commute in the afternoon hours where punctuality is less of a concern but it could also be an artifact of our 1-hour bins coupled with peoples work schedules. For instance, it is possible that people are expected to arrive at work before 9:00 so people use the bike sharing system in the interval 8:00-9:00. In the afternoon however, people might leave work later in the hour e.g. at 17:50 and then finish their commute after 18:00 meaning that the arrival is counted in the 18:00-19:00 bin resulting in the misalignment. The shape of the peaks may also show variations in peoples work schedules across cities. In New York City, people usually depart to work from 8:00 to 9:00 and then depart from work from 17:00 to 18:00. In Chicago, the wideness of the peaks indicate that people depart to work from 7:00 to 9:00 and then depart from work from 15:00 to 17:00 indicating that some people arrive at and depart from work earlier than others.

In New York City, Washington DC, London, and Helsinki, at first sight, it seems like in the afternoon rush hours the peaks of the morning sinks are considerably larger than the valleys of the morning sources, i.e. there appears to be an imbalance between arrivals and departures. In Helsinki, Oslo and Madrid, an opposite imbalance is also

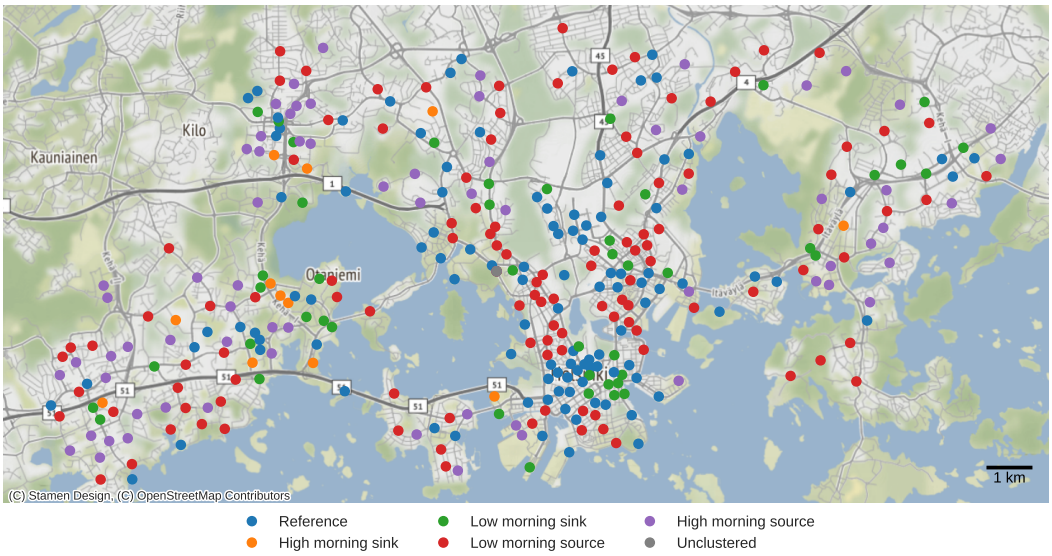


Figure 10.3: Clustered stations in Helsinki. Grey stations are unclustered due to low traffic.

seen in the morning rush hours. However as there is always one arrival for each departure, these imbalances are likely caused by a disparity between the number of morning sink and morning source stations. In Helsinki, where these imbalances are most prominent, 16.3% of the stations are of morning sink types while 52.6% are morning sources, a difference of 36.3 percentage points. This indicates that the trips emanating 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. When looking at the placement of the morning sink stations, they are typically concentrated in commercial and educational areas such as in the Helsinki city center around Kaartinkaupunki, Otaniemi which hosts a high number of schools and universities, and the Pitäjänmäki district which contains many IT and manufacturing companies, see Fig. 10.3. On the other hand, morning source stations in Helsinki are mostly placed in residential areas surrounding the commercial areas. The city which seems to be most balanced is Chicago. Here, there is only a 18.6 percentage point difference between the amount of morning sink and morning source stations. However, using these differences to solely explain the imbalances is not adequate. In Washington DC, there is only a 14.1 percentage point difference meaning that other local factors may contribute to the imbalance observed before.

Another explanation for the imbalance in the arrivals and departures in the afternoon hours could be that arrivals are more spread out in the afternoon. This is for instance the case in Washington DC where departures from morning sinks are mostly concentrated from 15:00 to 19:00 while arrivals to morning sources are primarily from 17:00 and throughout the evening. This can be an effect of what was discussed previously with bike share users having a more relaxed commute in the

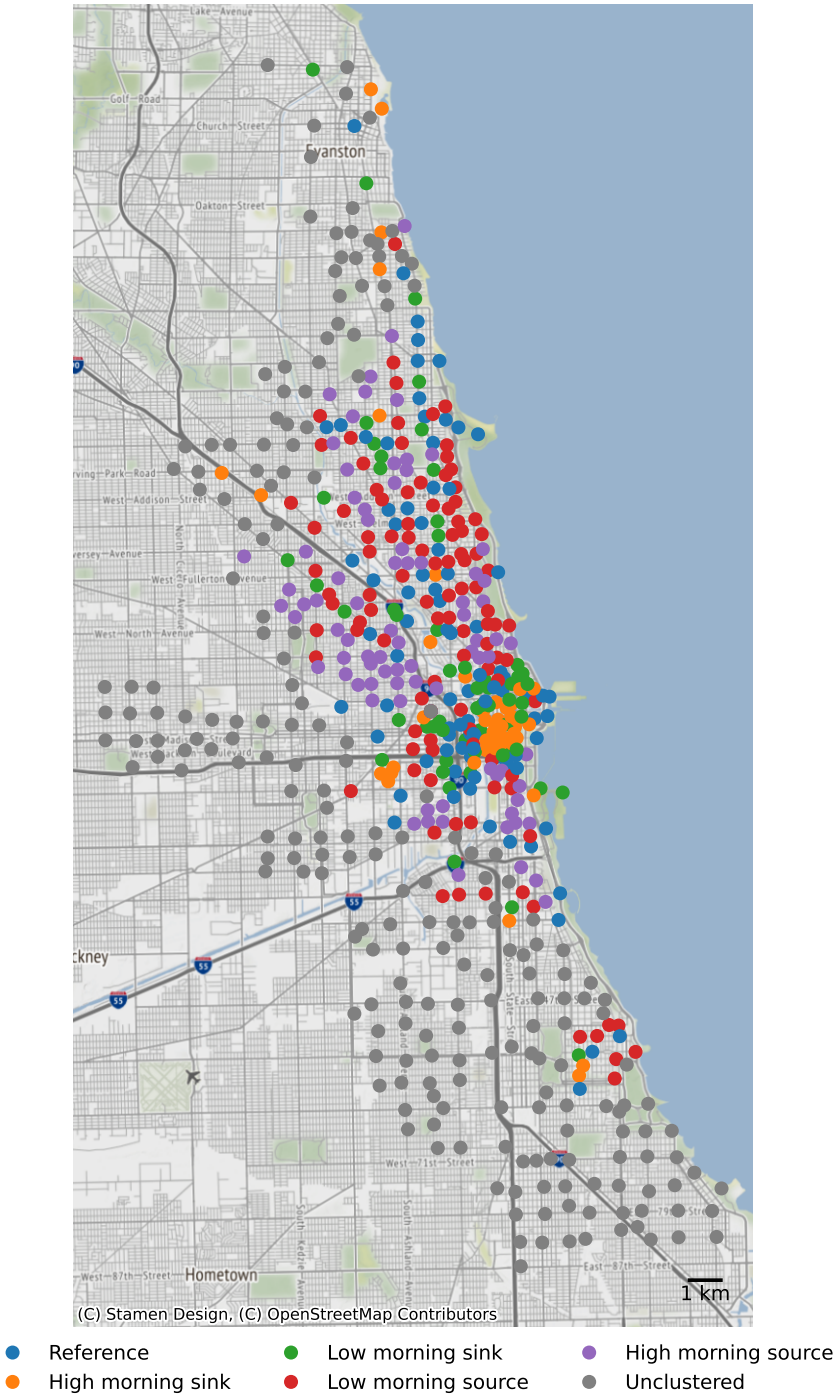


Figure 10.4: Clustered stations in Chicago. Grey stations are unclustered due to low traffic.

afternoon or people using bike share for recreational purposes or to get to leisure activities.

Interesting results can also be obtained by considering the stations which have been left out of the clustering due to having less than 8 trips a day on average. As seen in Table 8.3, there is generally a higher percentage of stations which are retained in the European cities which indicates that nearly all parts of the systems are frequently used. However, in US cities many stations, typically at the edge of the bike sharing system, have been removed due to low traffic. The worst case is in Chicago where about 37.77% of the stations are excluded which account for 19.15% of the trips indicating that a large part of the system is infrequently used. This can be attributed to many factors. One important factor is the income disparity in the city which has increased over time in these areas [Nat]. Over time, poor neighbourhoods like the ones containing these low traffic stations have become poorer, while rich neighbourhoods became richer. These low-income neighbourhoods are also mainly residential areas which in accordance with the zoning ordinance are restrictive towards business developments. Coupled with the fact that these areas are not well connected to public transit systems (see Fig. C.2) and are generally far away from the core of Chicago, the people living in these neighbourhoods have a higher incentive to use a car instead of bike sharing. Similar results where the outer residential areas are geographically disincentivised from using the bike sharing system have also been observed for the other US cities.

10.1 Logistic Regression

The coefficients obtained from the LR models trained on each city are presented in Table 10.2. To better visualise the relationship between the external variables of a city and the output from the LR model, heat map plots can be constructed for each city showing the external variables and the corresponding output. This is done by dividing the city area into 200 m $\times$200 m cells and then calculating the values of the external variables for each cell. These variables are then used as input for an LR model trained on the city in order to obtain the probability of a station placed at the center point of each cell belonging to a particular cluster. The resulting heat maps for New York City can be seen in Fig. 10.5. Similar heat maps for other cities can be found in Appendix C.

As seen in Table 10.2, morning sources are usually highly associated with residential land use. This is also easily observed in New York City as seen in Fig. 10.5, where morning source stations are more likely to be in residential areas which is a pattern that is shared between all US cities. In regards to commercial land use, stations in New York City in or near commercial areas such as the Diamond District and the Financial District in Manhattan, as well as along Hudson River are more probable to be morning sink stations. This can also be seen to be the case for Washington DC. Thus for these two cities, looking at the residential and commercial land use together can say a lot about the type of the station. The same is valid for Boston albeit to a lesser degree, as the coefficients for Boston are not all statistically significant.

Cluster	Coef. name	NYC	Chicago	Wash. DC	Boston	London	Helsinki	Oslo	Madrid
High Morning Sink	Const.	−0.247	**2.193**	**2.210**	**3.623**	**−11.171**	−3.792	**19.973**	−2.938
	Share of residential use	**−4.200**	**−3.260**	**−6.538**	−0.185	**15.293**	−2.039	−10.165	**13.670**
	Share of commercial use	**2.016**	−0.511	−0.559	−0.241	**13.703**	1.636	−11.843	4.654
	Share of recreational use	−1.279	−7.586	**−3.661**	**−3.909**	6.408	2.102	−11.076	3.946
	Population density [per 100 m²]	**−1.809**	**−1.886**	−1.797	**−6.079**	**−3.521**	−4.598	−1.424	**−2.813**
	Distance to nearest subway [km]	−0.502	**−2.349**	0.332	0.014	0.083	−0.345	**6.194**	1.198
	Distance to nearest railway [km]	−0.205	−0.383	**0.493**	0.427	0.664	−0.024	**−6.149**	**−2.133**
	Distance to city center [km]	0.152	0.080	**−0.577**	**−0.609**	**−0.321**	**0.415**	**−3.707**	0.034
Low Morning Sink	Const.	**1.785**	**2.234**	1.238	**2.787**	**−3.956**	−0.515	**18.862**	−2.188
	Share of residential use	−0.875	−1.261	**−3.797**	−1.147	**6.567**	0.333	−9.182	**9.133**
	Share of commercial use	0.301	−1.798	−0.400	−0.661	**5.758**	1.266	−10.123	3.983
	Share of recreational use	−0.093	**−2.628**	**−1.747**	−2.090	1.863	−2.341	−11.704	1.379
	Population density [per 100 m²]	**−0.769**	**−1.675**	−0.713	**−1.902**	**−1.726**	**−1.625**	−1.760	**−2.052**
	Distance to nearest subway [km]	**−1.972**	**−2.140**	−0.341	−0.323	−0.046	−0.172	**2.543**	2.100
	Distance to nearest railway [km]	−0.143	0.138	**0.351**	−0.306	**1.001**	0.034	**−3.175**	−0.687
	Distance to city center [km]	−0.065	0.007	**−0.233**	−0.162	**−0.356**	0.102	**−2.093**	0.040
Low Morning Source	Const.	**−2.281**	**−1.507**	**−2.333**	−0.726	−0.004	**−4.019**	15.057	−6.033
	Share of residential use	**1.708**	**3.955**	**2.904**	0.134	−1.814	2.028	−10.586	−1.215
	Share of commercial use	−1.031	0.434	−2.536	−1.062	−2.432	−1.359	−12.577	4.138
	Share of recreational use	**1.544**	−0.320	0.231	**−3.259**	−1.298	2.262	−8.750	6.801
	Population density [per 100 m²]	0.271	−0.243	**1.712**	−0.379	**1.111**	**1.261**	1.739	**1.908**
	Distance to nearest subway [km]	**1.243**	**1.013**	**0.825**	**0.599**	**0.987**	**0.340**	**3.012**	0.732
	Distance to nearest railway [km]	**0.283**	**0.654**	**−0.426**	**0.704**	**−0.924**	**0.385**	−1.313	**−0.868**
	Distance to city center [km]	0.011	**−0.226**	0.150	0.017	**0.177**	0.075	**−1.880**	0.242
High Morning Source	Const.	**−5.030**	−1.263	**−6.701**	−0.657	−1.461	**−2.356**	10.181	−9.216
	Share of residential use	**3.002**	**6.045**	**5.584**	0.805	−0.633	−0.564	−6.766	−1.206
	Share of commercial use	**−3.195**	−2.686	**−8.969**	−3.322	−2.532	**−7.125**	−9.546	2.369
	Share of recreational use	**3.623**	**−3.690**	0.668	**−5.676**	−1.905	−1.580	−4.786	5.945
	Population density [per 100 m²]	**0.693**	−0.101	**2.268**	−1.100	**0.913**	0.867	1.314	**1.941**
	Distance to nearest subway [km]	**2.950**	**1.441**	**1.230**	0.469	**2.241**	0.222	**2.585**	**4.715**
	Distance to nearest railway [km]	**0.413**	0.353	**−0.523**	**0.751**	**−0.777**	0.256	**−1.507**	**−1.167**
	Distance to city center [km]	−0.061	**−0.376**	**0.686**	0.045	0.112	**0.359**	−0.805	**1.025**

Table 10.2: Coefficients of LR models trained on different cities. Bold coefficients are statistically significant ($p < 0.05$).

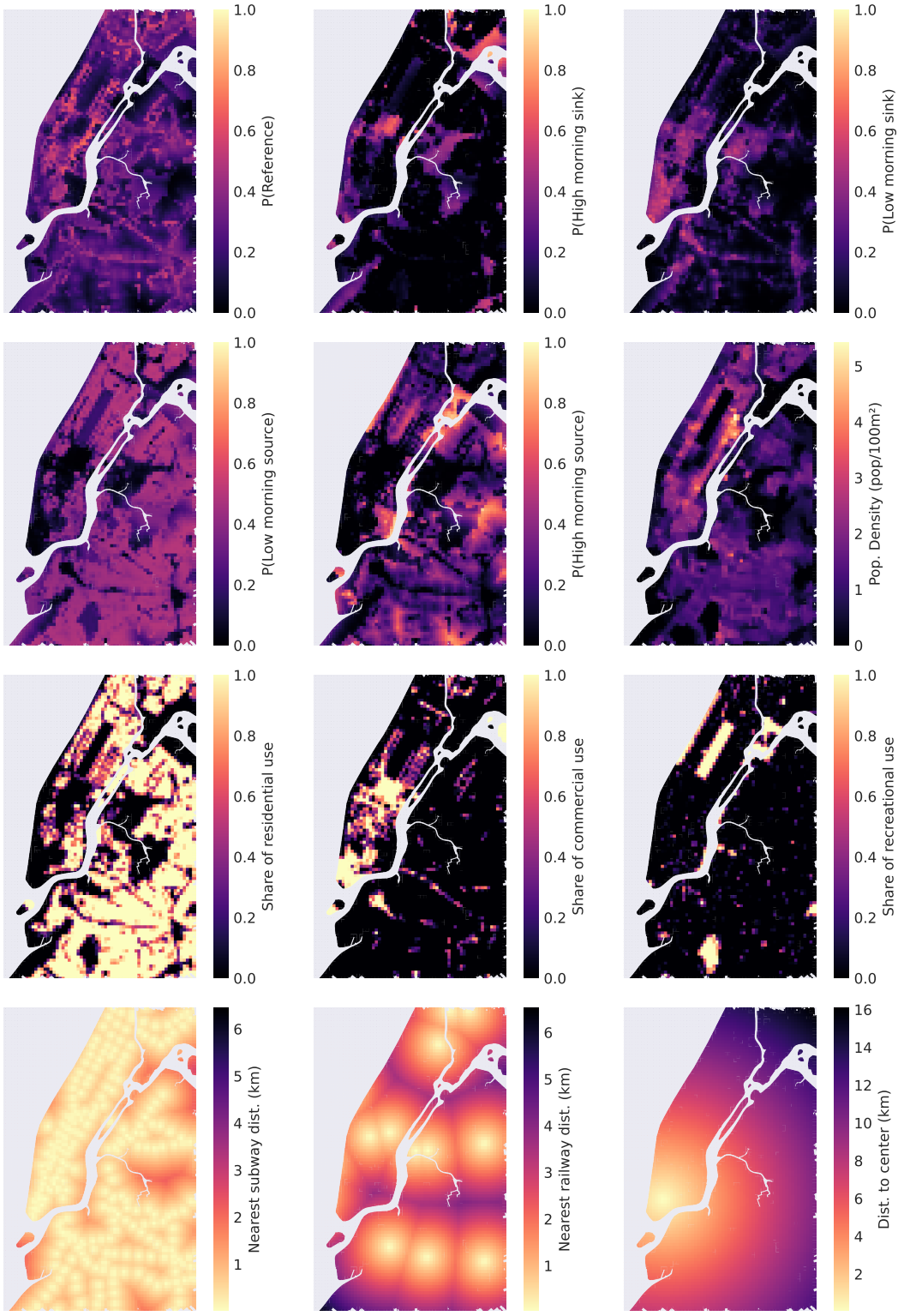


Figure 10.5: Heat maps of probabilities of belonging to different clusters alongside external variables for New York City.

It is important to note that there are two equally valid interpretations of this result depending on the interpretation of morning sources and morning sinks. One interpretation is that morning source stations have an abundance of departures in the morning while morning sink stations have an abundance of arrivals. In this framework, morning source stations are imbalanced due to the relatively high number of people living near these stations yielding a large amount of potential cyclists who use these stations during their commute. Likewise, morning sink stations are in commercial areas which have a large amount of work spaces, shops and other venues providing many reasons for cyclists to arrive at nearby stations. Another interpretation is that the imbalance in morning source stations is caused by an absence of arrivals while for morning sink stations it is due to an absence of departures. Residential areas have little to no commercial purposes so there is little reason for people to go to these areas in the middle of a business day. Likewise, commercial areas do not have as many residents and therefore a low supply of cyclists. Both of these interpretations are supported by how zoning in US cities is typically regulated. Each type of zone is usually defined with one specific purpose in mind such as commercial and residential use, and mixed-use zones are generally few and far between. This leads to cities having large areas in which only commercial use is permitted and likewise for residential use. Thus, when people are commuting in the morning they are most likely departing from residential zones and arriving in commercial zones to work.

A city which was expected to show a similar behaviour to this is Chicago. However, it was found that stations around commercial areas are more likely to be low morning sources. When examining the zoning data from Chicago, it was found that the buildings along the main roads of the city are zoned as commercial areas which is likely the cause of the different pattern in the coefficient. However, these commercial areas are also, despite their intended use, relatively lowly developed and provide fewer work opportunities compared to areas closer to the city center. Thus, stations which are close to these areas beside main roads may still act as morning sources since the low amount of work opportunities provides an absence of arrivals. If this is the case, investing more in the development of these areas could change the stations to be more balanced throughout the day.

The above discussion can also shed some light on the predicting power of land use in European cities where counter intuitive behaviour can be found. In London, morning sinks are associated with both residential and commercial areas while morning sources are more associated with residential than commercial areas. Helsinki has a similar behaviour to what was observed in US cities while Madrid exhibits the opposite, with stations in residential areas being more likely to be morning sinks. Oslo also exhibits very counter intuitive behaviour with regards to the share of land use with morning sinks being disassociated with both residential and commercial land use although this may be an effect of the reference cluster. One reason that the European cities do not share a specific pattern can be a lack of standardisation in how

land use is regulated in the countries. European cities also generally do not separate different zoning types to the same degree as cities in the US, yielding a higher blend of different types of land use throughout the cities.

It is difficult to say something general about recreational use. As the primary type of residential land use is parks, the relationship between recreational use and cluster type might say something about the location and usage of the parks in each city. In New York City there is a clear relationship between recreational use and cluster type, likely due to the presence of Central Park in the middle of Manhattan which is surrounded by residential areas. On the other hand, in Boston the model has negative coefficients for both high morning sink and high morning source clusters. In London, a relationship opposite of that to New York is observed, although the coefficients are not statistically significant. The major parks in London are Hyde Park and Regent's Park which while being close to residential areas are more associated with balanced stations, see Fig. C.5.

It was found that the station type is highly associated with the population density of the surrounding area. For all cities, morning sink stations are generally in areas with a low population density while morning sources are in areas with a high population density. This aligns with the expectation that people travel from their homes in the morning and arrive later in the afternoon. One should also note the relation between an area's land use and its population density. Residential areas generally have a higher population density than commercial areas and thus one would expect that the coefficients in the model with respect to share of residential use and population density will behave similarly. While this is the case for cities in the US, it is not the case for cities in Europe, possibly due to what was previously discussed regarding land use differences between US cities and European cities. Thus for European cities, population density might be a more important predictor of station type than how the surrounding land is used.

Another pattern shared between almost all cities is how the distance to the nearest subway relates to the station type. In all cities apart from Oslo and Madrid, morning source stations are generally far from the nearest subway while morning sinks generally are closer. For most cities, it is also possible to distinguish between the two types of morning stations with high morning source stations being further away from the nearest subway than low morning source stations. How the station type is related to the distance of the nearest subway can be attributed to many factors. The most straightforward interpretation is that people use bike sharing as a first-mile solution. At the start of a user's trip, they may use bike sharing to cover the distance between their origin and the nearest subway station which will cover the remaining distance making the subway station a morning sink. It should be noted that stations close to subways are not necessarily always morning sinks since people can also use a station after the subway in order to cover the remaining trip distance, making the station a morning source. However, in most cities the subway network

is more dense in downtown areas meaning the subway will likely take you within walking distance to your place of work, lowering the need for bike sharing. This also inspires another interpretation. Commercial areas tend to have a higher concentration of subway stations than residential areas and thus it makes sense that morning sinks are closer to subways since they tend to be in commercial areas. Therefore, the coefficients related to the distance to the nearest subway may be a result of a correlation between land use and the density of the subway network.

In regards to railway stations, they share the same pattern as for subway stations except for in Washington DC, London, and Oslo where morning sources tend to be closer to railway stations than morning sinks. As an example, the two bike sharing stations in London which have the highest amount of traffic are next to Waterloo Station and King's Cross Station, and both of these bike sharing stations are very high morning sources. A likely explanation of this is that many people working in the mentioned cities are living outside of the city and use railways in their commute. Due to the low number of railway stations, an additional transportation method may be required to cover the last mile, one of which is bike sharing. In Madrid, high morning sinks and sources tend to be nearer to railway stations than low morning sinks and sources indicating a combination of source and destination train stations.

Having people from outside the city using the bike sharing system can also affect the predicting power of population density, since a large share of people going into the bike sharing system in the morning are coming from outside the city through railways instead of coming from residential areas inside the city.

10.1.1 Generalisation test

In order to see how a model trained on one city generalises to another city, we tested each of the 8 models on the same data used to train the other models. The rate at which the models predicted the cluster types correctly can be seen in Fig. 10.6. When training and testing on the same city, we split the stations randomly into a training set and test set with the training set having about 80% of the stations. Due to this random split, we found that the success rates on the diagonal changed a lot depending on the chosen seed in the random number generator. To counteract this, we computed these success rates 50 times at different seeds and then took an average.

US cities were observed to be similar to each other and this is also shown in the success rates. Training an LR model on a US city and then testing it on another US city yields an average success rate of 31.17% while training on a US city and testing on a European city yields a success rate of 28.38% on average. However, the success rates in the quadrant of Fig. 10.6 where we train on the US and test in Europe also have a large variance with one success rate being lower than random guessing and one as high as 38%, the highest success rate of a model tested on a foreign city. The poor success rate came as a result of training the model on Boston and then testing it Oslo. These low rates are likely an effect of the different reference cluster. How-

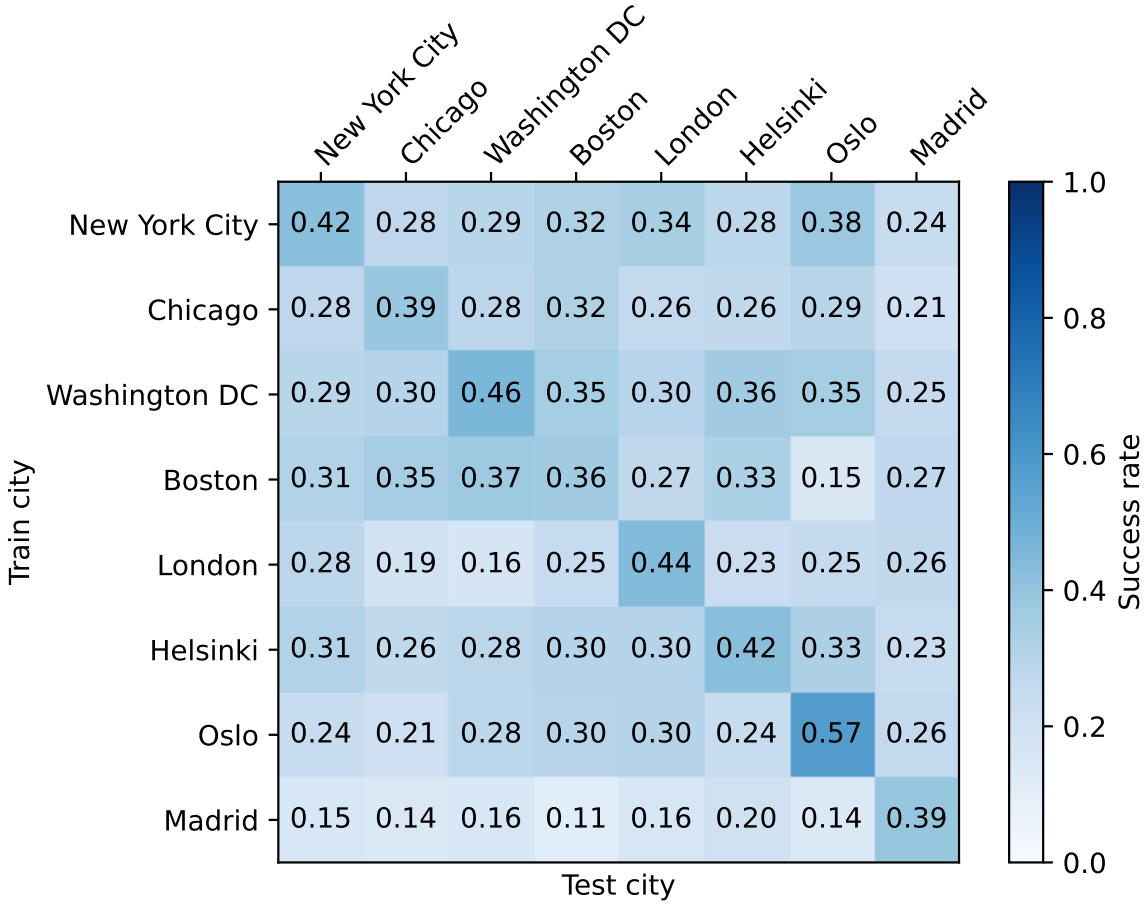


Figure 10.6: Success rates of models trained and tested on different cities.

ever, models trained on other US cities perform better on Oslo despite the different reference cluster. It is seen that the model trained on Washington DC is generally good at predicting the station type of European cities with success rates between 30% and 36% except for Madrid where the rate is 25%. In this sense, it can be said that Washington DC is the most European-like city out of the four cities in the US.

When training on the European cities and testing on US cities, the models score a success rate of 22.63% on average. When training on a European city and testing on a foreign European city this average score drops to 18.13% indicating that when predicting the type of the stations in a European city it may actually be better to train the model on a city in the US than another European city. Most of the low success rates are due to poor performances on models trained on Madrid as well as other models tested on Madrid. In fact, the model trained on Madrid did not manage to get a success rate better then randomly guessing when tested on other cities.

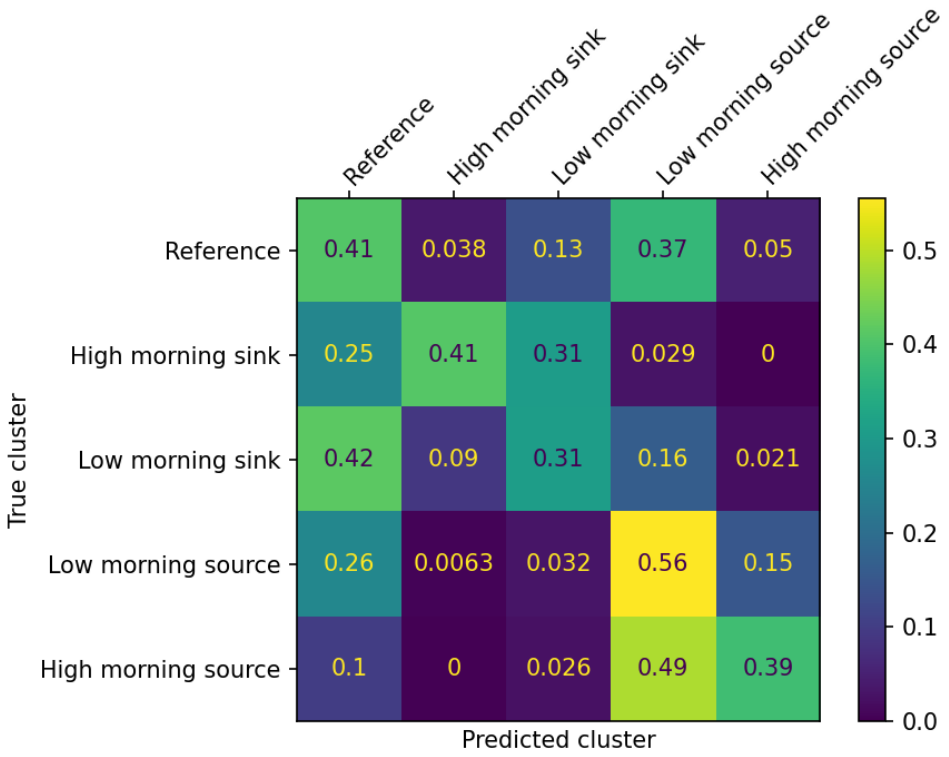


Figure 10.7: Confusion matrix of LR model trained and tested on New York City. The values are normalised with respect to the true labels.

The cluster centers in Fig. 10.1 also suggest how well a model trained on some cities will perform on others. Despite the minor differences between the cluster centers in different cities mentioned before, there is still a high degree of similarity between the traffic patterns of the cluster centers in US cities indicating that a model trained on one US city will likely perform well on another US city as was also observed. For the European cities it was found that the cluster centers are more different from each other which may have an influence on the generalisation of the models.

When it comes to predicting the type of a station, some prediction errors are more severe than others. For instance, having a high morning source station being predicted as a low morning source station is not as detrimental when predicting the daily traffic than if it was predicted as a high morning sink station. Thus, to better understand the predictions of the LR model it can be useful to look a the confusion matrix of the model. A confusion matrix of a model trained and tested on New York City can be seen in Fig. 10.7. Note that the values in the matrix are normalised with respect to the true labels. For instance, in Fig. 10.7 low morning source sta-

tions are being predicted to balanced stations 26% of the time, low morning source stations 56% of the time and high morning source stations 15% of the time. We also note that the confusion matrix has been averaged over multiple iterations and as a result of this, success rates on each row may not sum up to 1. Crucially however, low morning source stations are only predicted to be sink stations under 4% of the time. Morning sink stations are sometimes predicted to be morning source stations but only a relatively low number of times. Thus, the model does not tend to mix up morning source stations and morning sink stations, which was also observed on the models trained on other cities, see Appendix B. Many stations are also incorrectly predicted to be balanced stations, but the severity of this error is not as big as mixing up morning sources and morning sinks. Thus, even though the success rates in Fig. 10.6 can be seen as relatively low, the impact of the classifications may actually be relatively minor in the prediction of the daily traffic.

We conclude this chapter by listing the key findings from the clustering analysis. These are:

- Most of the cities (all of US and 3 from Europe) share the same types of stations albeit with differences in the peaks and valleys in the rush hours which may be attributed to local factors.
- Oslo lacks a cluster of balanced stations and instead has a cluster containing stations which are balanced in the morning and sources in the afternoon. This may be related to students in the Ullevål district.
- Madrid has three periods in the day with high traffic. This has been attributed to the working culture in Madrid and their midday break.
- US cities have more low-traffic stations (under 8 trips a day on average). These stations are primarily located at the edge of the bike sharing system in low income areas. This was not observed in European cities.
- For the US cities, stations in residential areas are more likely to be morning sources while stations in commercial areas are more likely to be morning sinks. This pattern may be attributed to the zoning regulations of US cities.
- Population density is an important predictor of the station type, although this may be due to a correlation between population density and residential use.
- In the US cities, London and Helsinki, stations close to subways are more likely to be morning sinks while stations further away are more likely morning sources. This may be due to people using bike sharing to cover the first mile.
- The relation between railway stations and the station type depends on how people use the railway system in their commute. In some cities, the impact from railway stations is similar to subway stations while in other cities morning

sources are generally close to railway stations possibly due to working people not living within the city.

- The model trained on Washington DC performs best on average on European cities compared to models trained on other US cities.
- Models trained on US cities perform better on average when tested on European cities than models trained on European cities. This is mostly because of the poor performance of the model trained on Madrid.

11. Demand Prediction

In this chapter, we present results gathered by predicting the traffic volume of stations in the bike sharing systems as described in Chapter 9. The model uses a Gaussian exponential family distribution with the natural logarithm as link function as this was found to provide the best results in preliminary analyses. The resulting coefficients are presented in Table 11.1. A heat map of predicted demand for 200 m ×200 m cells in New York City can be seen in Fig. 11.1. Similar heat maps for other cities can be found in Appendix D.

Coef. name	NYC	Chicago	Wash. DC	Boston
Const.	**5.851**	**4.349**	**4.687**	**4.586**
Share of residential use	**−0.691**	**−0.553**	0.084	**0.716**
Share of commercial use	**0.298**	**0.497**	0.082	0.304
Share of recreational use	0.147	−0.381	0.035	−0.157
Population density [per 100 m^2]	**0.264**	**0.728**	**0.638**	−0.200
Distance to nearest subway [km]	**−0.340**	**−0.401**	**−0.381**	−0.154
Distance to nearest railway [km]	**−0.277**	**−0.286**	**−0.080**	−0.176
Distance to city center [km]	**−0.113**	**−0.085**	**−0.237**	**−0.205**
	London	Helsinki	Oslo	Madrid
Const.	**7.244**	**6.204**	**5.082**	**5.055**
Share of residential use	**−2.327**	**−0.599**	**−0.553**	−0.261
Share of commercial use	**−2.135**	0.015	0.019	−0.130
Share of recreational use	**−2.11**	−0.230	0.251	0.211
Population density [per 100 m^2]	**−0.180**	0.049	**0.220**	**0.109**
Distance to nearest subway [km]	**−0.459**	**−0.323**	0.191	−0.296
Distance to nearest railway [km]	0.080	**−0.175**	−0.139	**−0.116**
Distance to city center [km]	**−0.127**	**−0.084**	**−0.366**	**−0.076**

Table 11.1: Coefficients of demand regression model on different cities. Bold coefficients are statistically significant ($p < 0.05$).

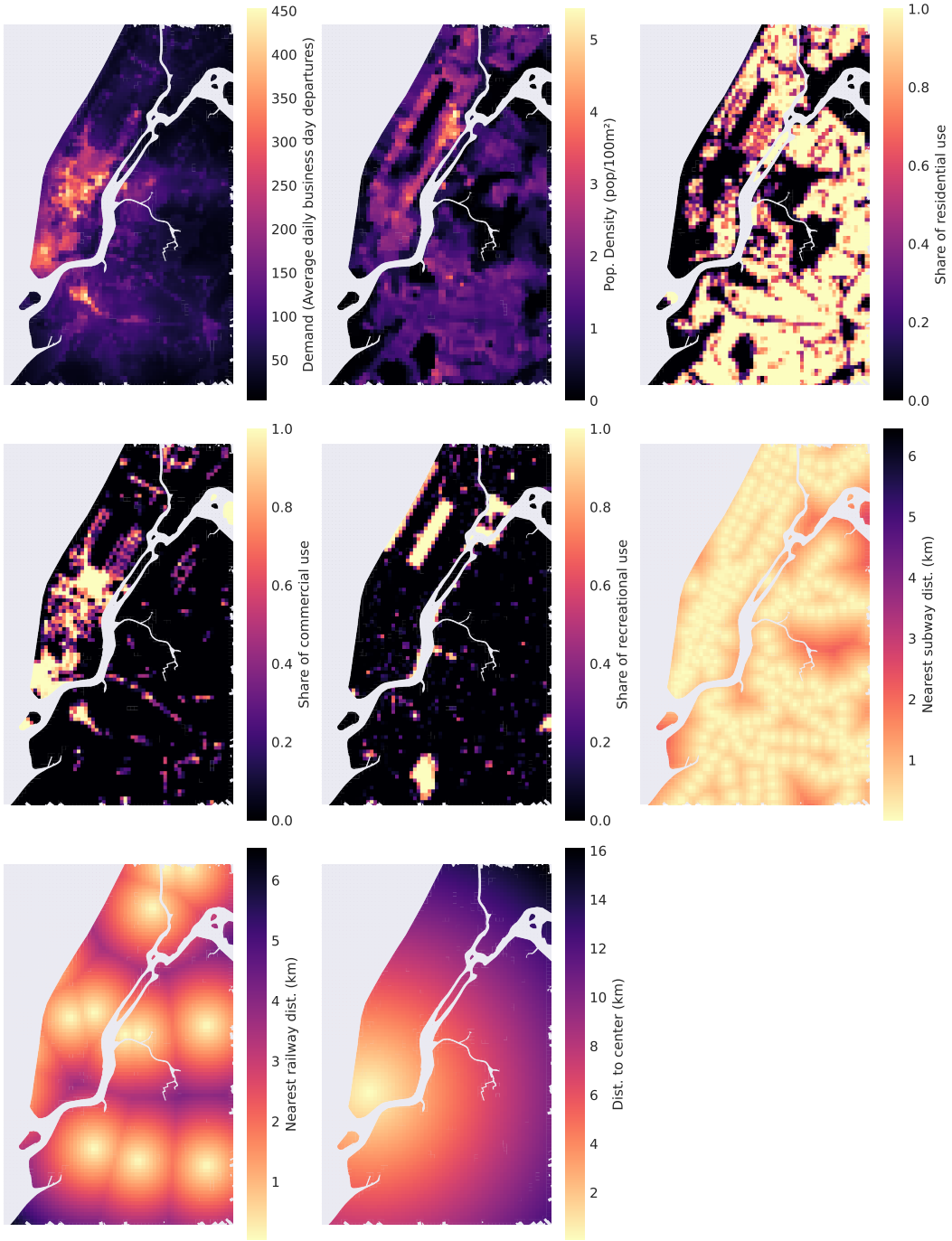


Figure 11.1: Heat map of predicted demand and external variables for New York City.

For most cities, stations in residential areas generally have less demand than stations in commercial areas. However, this is with the exception of Boston which shows the opposite behaviour and also Washington DC where land use is not as important when predicting demand. However, population density is an important predictor of demand, something which is shared between all cities although with Boston and London as exceptions where a higher population density negatively affects demand. This can be considered strange since residential areas usually have a high population density. However, when comparing land use, commercial areas have much more traffic than residential areas since they typically act as morning sinks i.e. many trips end up at and depart from these stations. Having a higher population density regardless of the type of area will always increase the demand, with the exception of Boston and London. Thus, even though residential areas tend to have a higher population density, the stations in these areas still typically have a lower demand than stations in commercial areas by virtue of being in residential areas.

Distance to transportation hubs is also a good predictor of demand with stations closer to subway and railway stations having more demand. This is also true for the distance to the city center which was found to be significant in predicting demand for all cities.

Residual plots for models trained on all cities can be seen in Fig. 11.2. The residual was calculated as the actual number of average daily trips minus the predicted number. It can be observed for all cities, that the variance of the residuals increases with respect to the predicted number of trips for each station. This is typically an indication that a variance stabilising transformation of the response variable is needed. However, we were unable to find such a transformation. This may be taken as evidence that a more advanced model may be needed to predict the demand of the stations more accurately. An important thing to note however is that the residuals do not indicate that a linear model is inadequate at modelling the demand. One sign which would indicate that would be if the residuals are concentrated in a non-linear pattern such as an arch. However, this was not observed for our models.

A good way of gauging how the model acts depending on the actual traffic of the station is by plotting the predicted number of trips for each station against the actual number. This is shown in Fig. 11.3. On the dashed lines, the predicted demand is equal to the actual demand and points far from the lines indicate larger errors. From the plots, it was found that the stations in which the model has the largest error are the ones which have a high amount of traffic compared to other stations and that the model always underestimates the traffic volume of these stations. By inspection, it was found that these outlying stations also produce the outliers seen in the residual plots.

Some noteworthy outlying stations are stations next to Waterloo Station in London in which the model underestimates the traffic volume by about 200 trips, stations outside of Union Station in Washington DC where the model underestimates the traf-

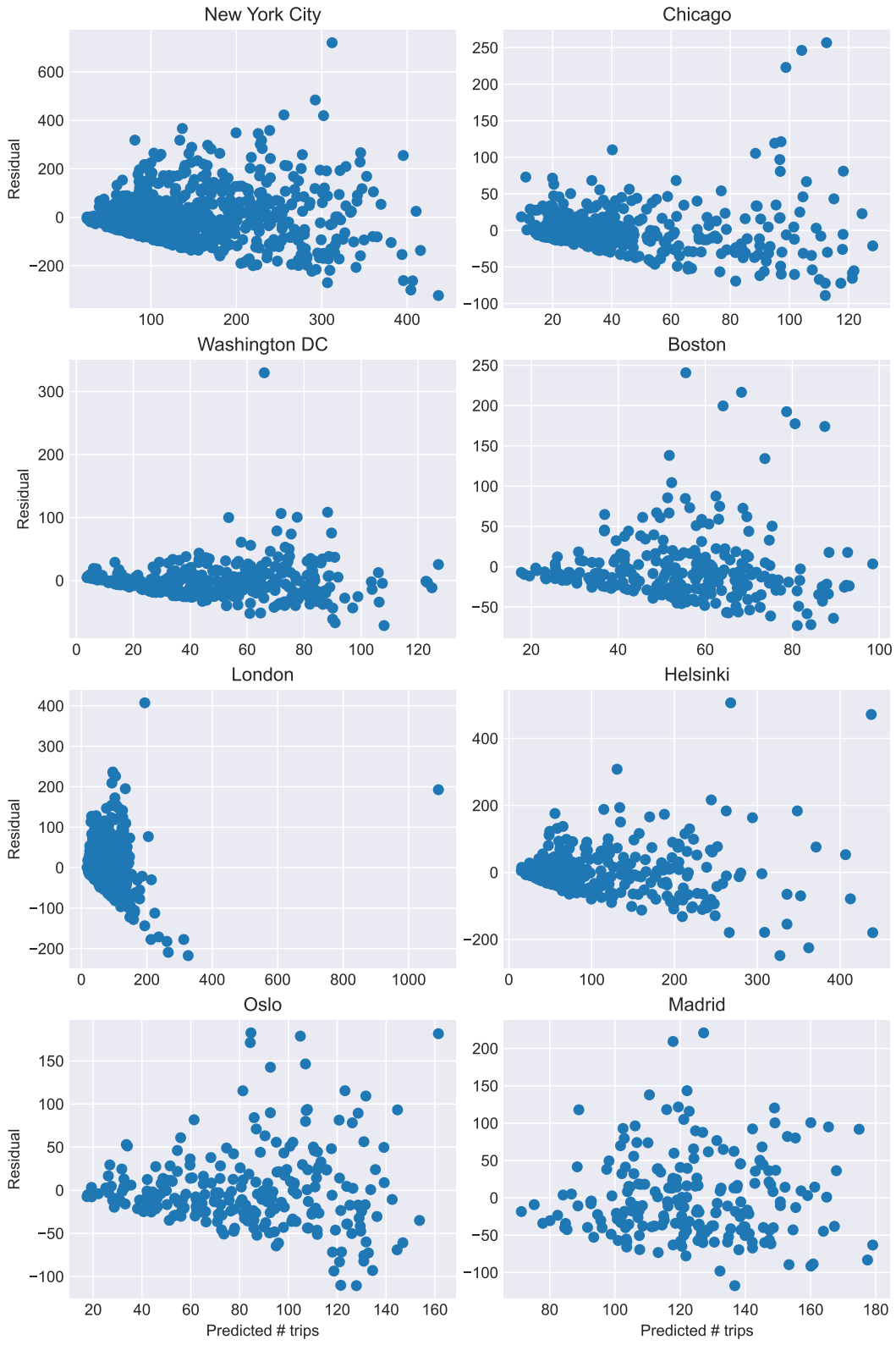


Figure 11.2: Residual plots of models trained on all cities.

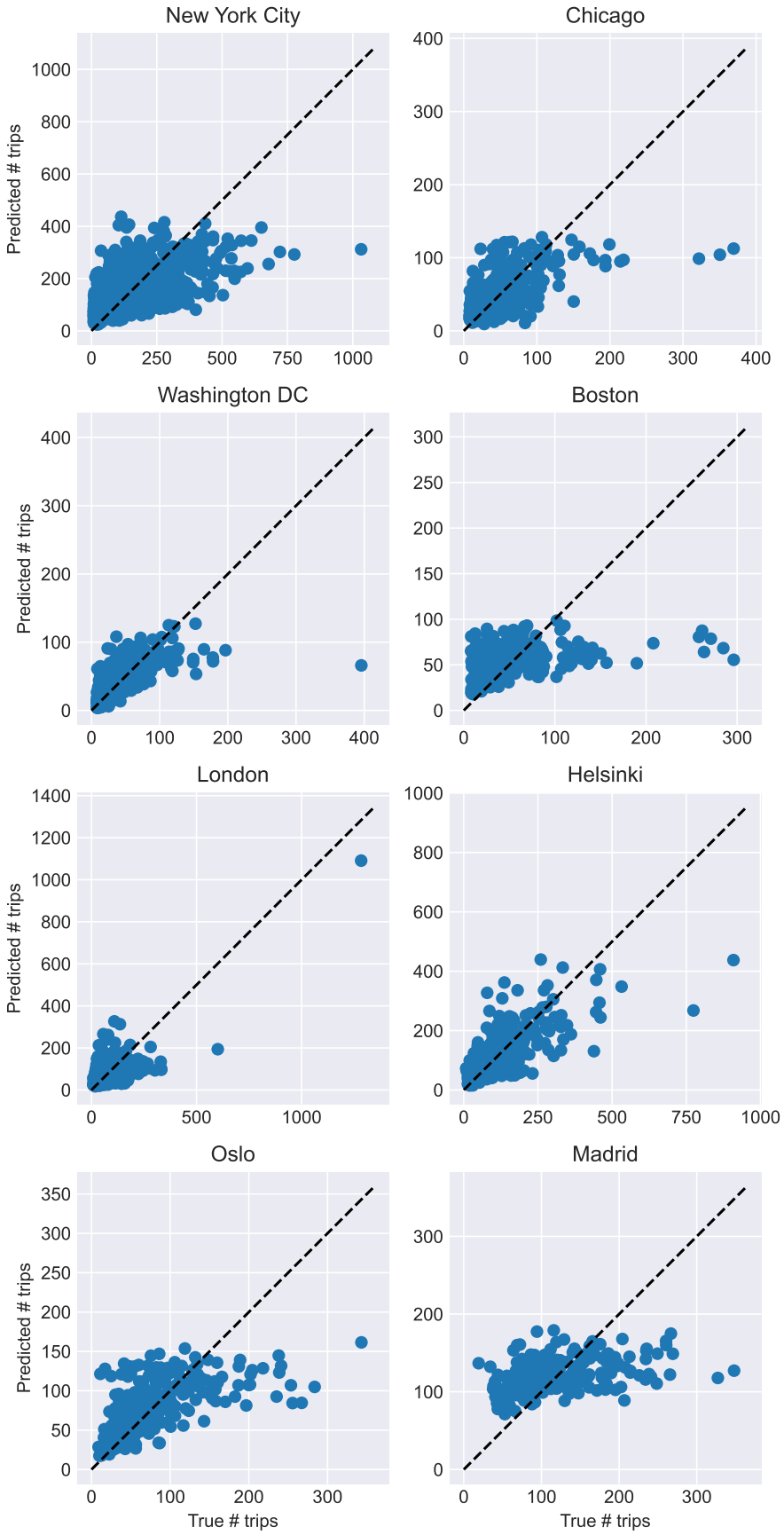


Figure 11.3: True number of trips vs. predicted number of trips on models trained on each city. Points closer to the dashed line have lower error.

Train city \ Test city	NYC	Chicago	Wash. DC	Boston	London	Helsinki	Oslo	Madrid
NYC	73.9	72.0	47.7	92.0	84.4	47.8	97.7	132.6
Chicago	77.0	21.4	21.8	29.1	44.7	80.0	45.0	152.1
Wash. DC	74.0	19.8	17.7	30.7	55.9	74.0	75.8	378.0
Boston	95.2	22.7	22.1	30.8	34.8	75.9	41.0	70.4
London	135.1	170.2	139.6	121.0	36.4	73.8	72.6	65.8
Helsinki	92.5	124.4	91.2	140.9	119.2	48.8	137.0	116.9
Oslo	96.6	24.1	23.2	34.2	42.2	70.2	35.7	52.0
Madrid	82.8	38.1	31.3	46.9	40.2	61.1	43.2	46.3

Table 11.2: MAE when training and testing the demand model across cities. Lower is better.

fic volume by over 300 trips, Grand Central Terminal in New York City where the estimate is about 650 trips lower than the actual volume as well as three stations in Chicago where the model greatly underestimates the traffic volume. Conversely, the model seems to generally overestimate the traffic volume of station which have a low amount of traffic.

Table 11.2 shows the Mean Absolute Error (MAE) when training and testing the demand model across cities. The test has been done in the same manner as in the clustering analysis, i.e. when training and testing on the same city, a $80\% - 20\%$ random split of the complete data set was done first and when comparing between two different cities, the model was trained on the complete data set of the train city and tested on the complete data set of the test city. When training and testing on the same city, we also average the error over multiple iterations. The MAE is intuitive to use since it tells how many trips the model is off on average.

In the results from the LR model, we found that there was a similarity between the US cities and the European cities. This can also be found in these results with the notable exception of New York City where the demand model performs significantly worse than other models trained on US cities. This is most likely caused by the large amount of traffic within the system of New York City relative to the other US cities. From Fig. 11.3, it can be seen that New York City has many stations with over 500 trips daily on average while other US cities only have stations with up to 400 daily trips on average. This can skew the model trained on New York City to expect more daily trips. The amount of traffic in the systems of Chicago, Washington DC, and Boston is more comparable and thus demand models trained and tested across these cities perform significantly better. This result indicates that if one were to plan a bike sharing system in a new city using a demand model, then the model should be trained on a city with an existing system which has a similar amount of traffic to what is expected in the new city.

London also acts as an outlier in Table 11.2 and produces very large errors for the US cities. This may be due to both the demand model being skewed by Waterloo station and also that the demand model for London treats higher population density as a negative effect on demand.

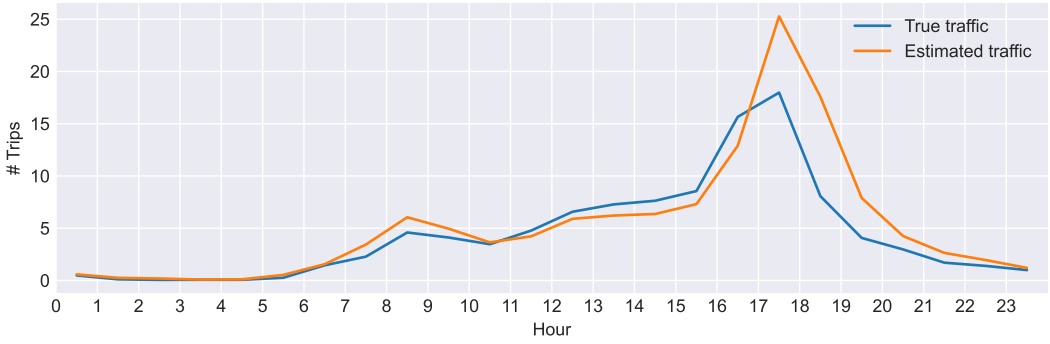
We previously saw that LR models trained on US cities are generally better at predicting the station types in European cities than models trained on European cities, likely due to cultural differences between European cities. However, this is not observed in the case of demand. In fact, models trained on Helsinki, Oslo, and Madrid perform moderately well when tested on Chicago, Washington DC and Boston. As seen in Table 11.1, there are not many differences in how the models act. This indicates that while cultural differences between the cities can have a large impact on the shape of the overall traffic of the stations, they have a relatively low impact on the actual demand of the stations. Another example of this is the demand models trained and tested across Oslo and Madrid, two cities with different cycling culture as evidenced by the bike share traffic patterns. The demand model with the best performance in Madrid was trained in Oslo and the model trained in Madrid also performs relatively well in Oslo.

The main takeaways from the demand model analysis are thus as follows:

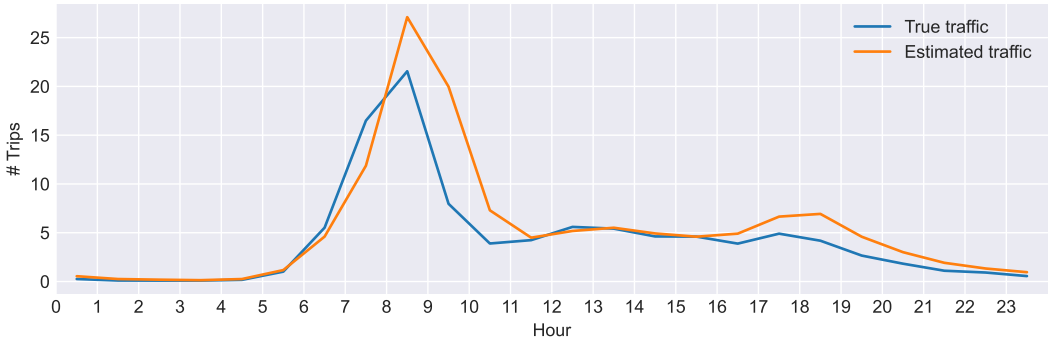
- Share of residential use negatively affects demand while share of commercial use positively affects demand. This is with the exception of Boston where the opposite is true and Washington DC where land use is less significant.
- Higher population density generally increases demand except for in Boston and London.
- Stations closer to subways, railway stations and city centers generally have more demand.
- The variance of the residuals grows with the predicted traffic volume indicating that a more advanced statistical model may be better able to model the demand.
- The demand models tend to overestimate the traffic volume of low-traffic stations while underestimating the traffic volume of high-traffic stations.
- Performance of demand models can be heavily affected by differences in the amount of traffic between systems. Demand models used on a new city should be trained on a city with a similar expected amount of traffic.

12. Traffic Prediction

So far we have seen how both the type and traffic volume of a station can be modelled based on external data. In order to combine these two results, recall from Chapter 9 that multiplying the predicted shape of the traffic with the predicted volume will give a prediction on the hourly traffic of an average day. An example of such a predicted traffic pattern is seen in Fig. 12.1, which shows a high morning sink station in New York City where the combination of the models for the traffic pattern and volume



(a) Estimated number of departures.



(b) Estimated number of arrivals.

Figure 12.1: Predicted number of departures and arrivals for a station on W. 52nd St. & 5th Ave. in New York City.

has estimated the average daily traffic. In this case, the model performs very well both in predicting the correct shape of the traffic and an appropriate traffic volume. The predicted traffic volume was 253.45 trips while the true traffic volume was 206.26 daily trips on average, an overestimate of around 47 trips.

The example illustrates that even though the traffic volume was overestimated by close to 23% of the actual amount, the predicted traffic pattern is reasonably close to the actual traffic since the 47 falsely predicted trips are spread out over departures and arrivals throughout the day.

To see how well the model generalises to new stations in the same bike sharing system, we split the stations in each system with 80% being in the training set and the rest in the test set. In Chapter 11, it was found that the demand prediction behaved differently depending on the true traffic volume of the stations. Thus, to make a more representative test set, we classified the 20% lowest traffic stations as low-traffic, the top 20% most trafficked stations as high-traffic, and the rest as mid-traffic stations. The test set was then constructed by picking random stations while conserving the 20/60/20 relationship between low-, mid- and high-traffic stations. The 20/60/20 split is of course arbitrary and other methods can be used to classify the stations into these categories. In a preliminary experiment, we attempted to use a k-means classification but this resulted in few high-traffic stations due to outliers. Therefore, this more rudimentary method provided more reliable results.

12.1 Results

After constructing the test set for each city, the traffic patterns of these stations were predicted. The MAE of the prediction for each hour can be seen in Fig. 12.2. As can be seen in the figure, the MAE is largest during times were there is a high amount of traffic in the system. The city were the traffic predictions have the highest MAE is New York City, were the MAE goes up to 4 in the rush hours for both departures and arrivals, meaning that the prediction is off by 8 trips on average. The standard deviations also show that the error varies a lot in the rush hours. In London, the prediction error in number of departures and arrivals is the same on average but during the morning rush hours, the prediction error of departures has a larger standard deviation than arrivals. This can likely be attributed to a particularly high-traffic station near King's Cross Station, where the number of departures was underestimated by 200 in the morning rush hours and similarly for the number of arrivals, see Fig. 12.3. This station was also wrongly predicted to be a low morning source while it actually was an extreme case of a high morning source type station.

By making a distinction between low-, mid-, and high-traffic stations, we are able to gauge how the traffic prediction performs for each of these three types of stations. Figs. 12.4 and 12.5 show the Mean Error (ME) of the traffic prediction when tested on the three types of stations. The tested stations are all from the same test set

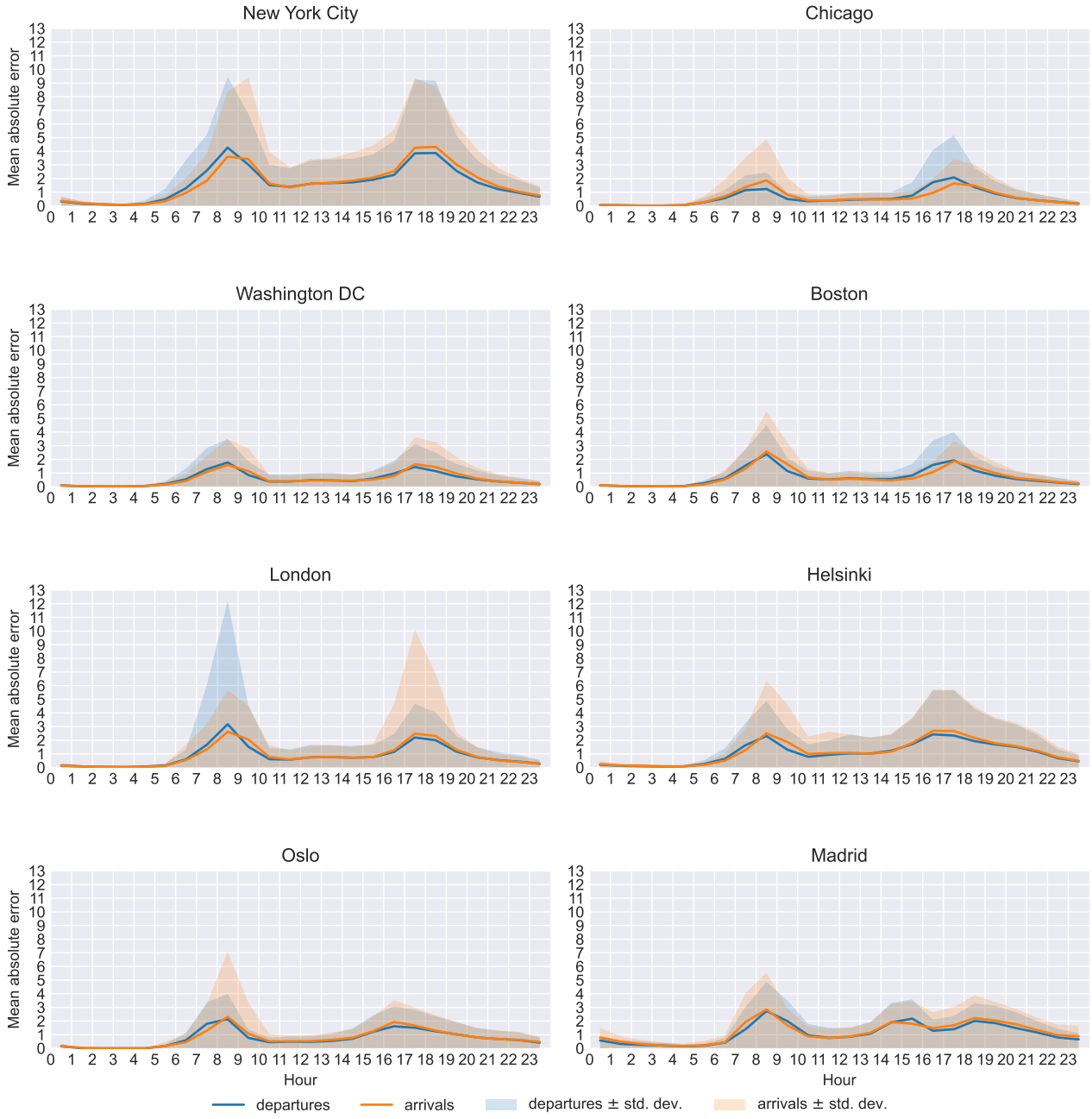
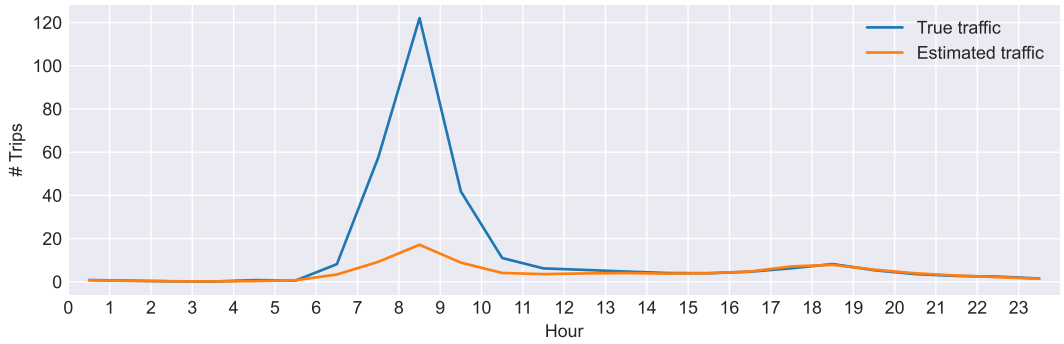
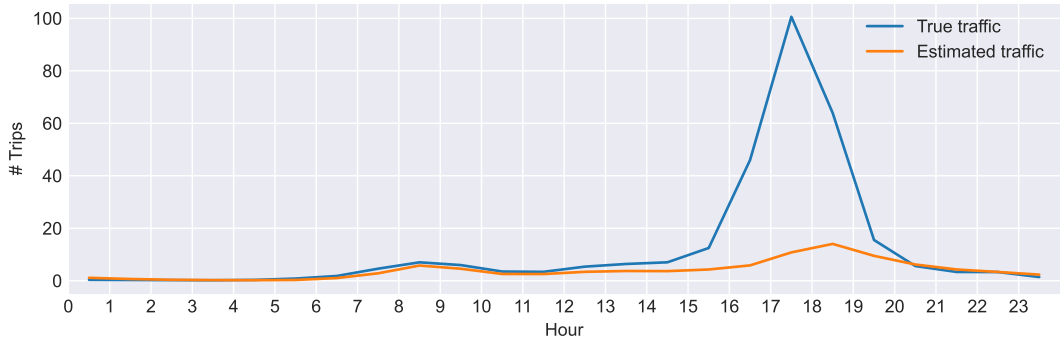


Figure 12.2: MAE of predicted number of departures and arrivals for each hour.



(a) Estimated number of departures.



(b) Estimated number of arrivals.

Figure 12.3: Predicted number of departures and arrivals for a station on Belgrove St. near King's Cross Station in London.

discussed previously and the models have been trained using the same training set. Keep in mind that the test sets have varying sizes. We opt for the ME since it also tells if the models under- or overestimates the traffic.

From the figures, we observe that the models generally underestimate the traffic pattern of high-traffic stations which matches what was found in Chapter 11. In Chicago, the number of departures was only underestimated in the afternoon, while the number of arrivals were only underestimated in the morning. However, the underestimation in the morning is likely due to the test set having many high morning sink stations which were predicted to be low morning sink or balanced stations. An example can be seen in Fig. 12.6 were a high morning sink station was predicted to be a balanced station. Regarding mid-traffic and high-traffic stations, for most cities the performance of the models slightly overestimates the traffic for low-traffic stations which is also consistent with the analysis of the demand model. In Washington DC and Helsinki, the performance of the traffic prediction is very similar between low-traffic station and mid-traffic stations. This also makes sense when observing Fig. 11.3 where both of these types of stations do not stray off from the diagonal line too far compared to other cities.

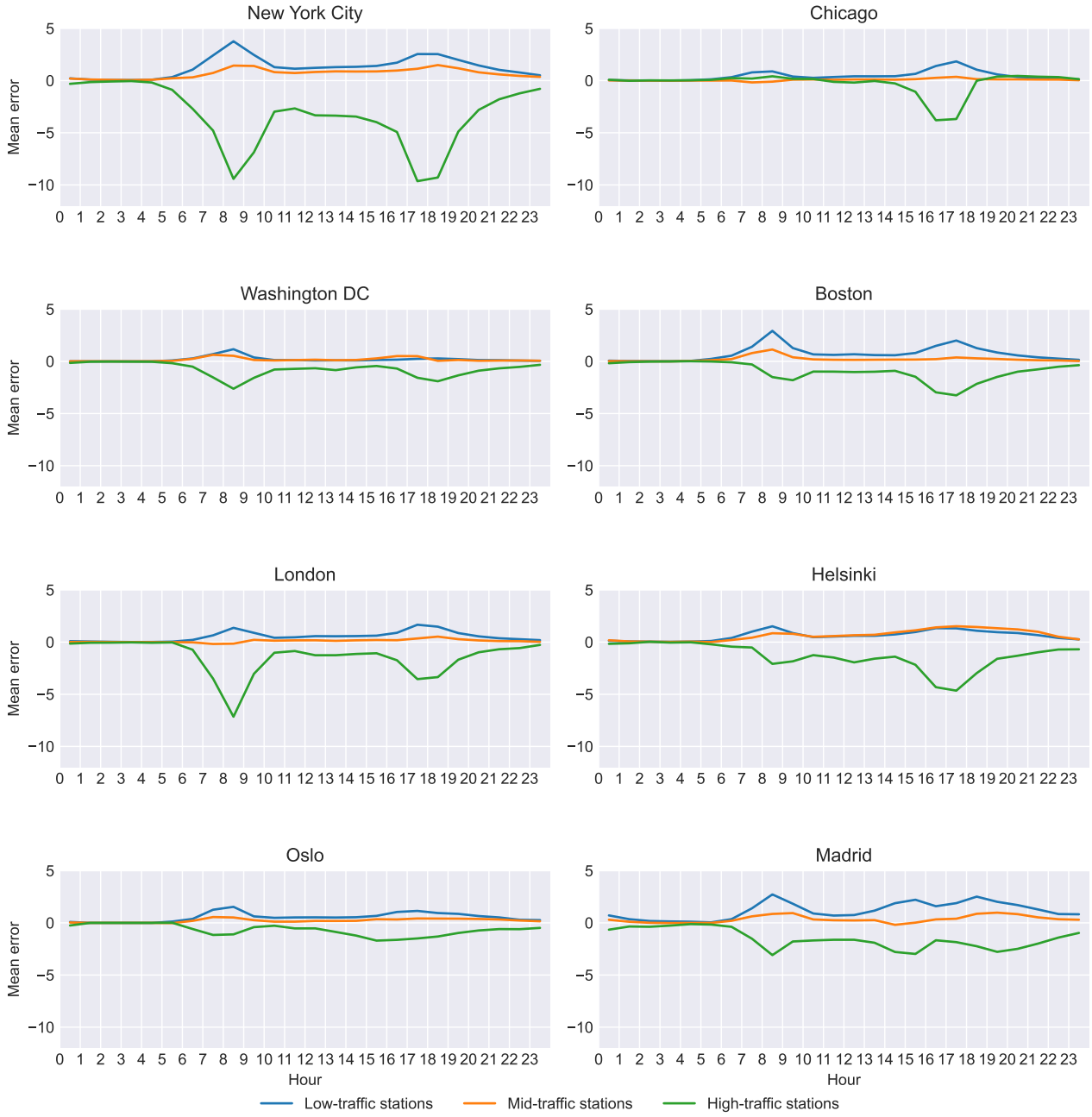


Figure 12.4: Mean error of predicted number of departures each hour when tested on low-, mid-, and high-traffic stations.

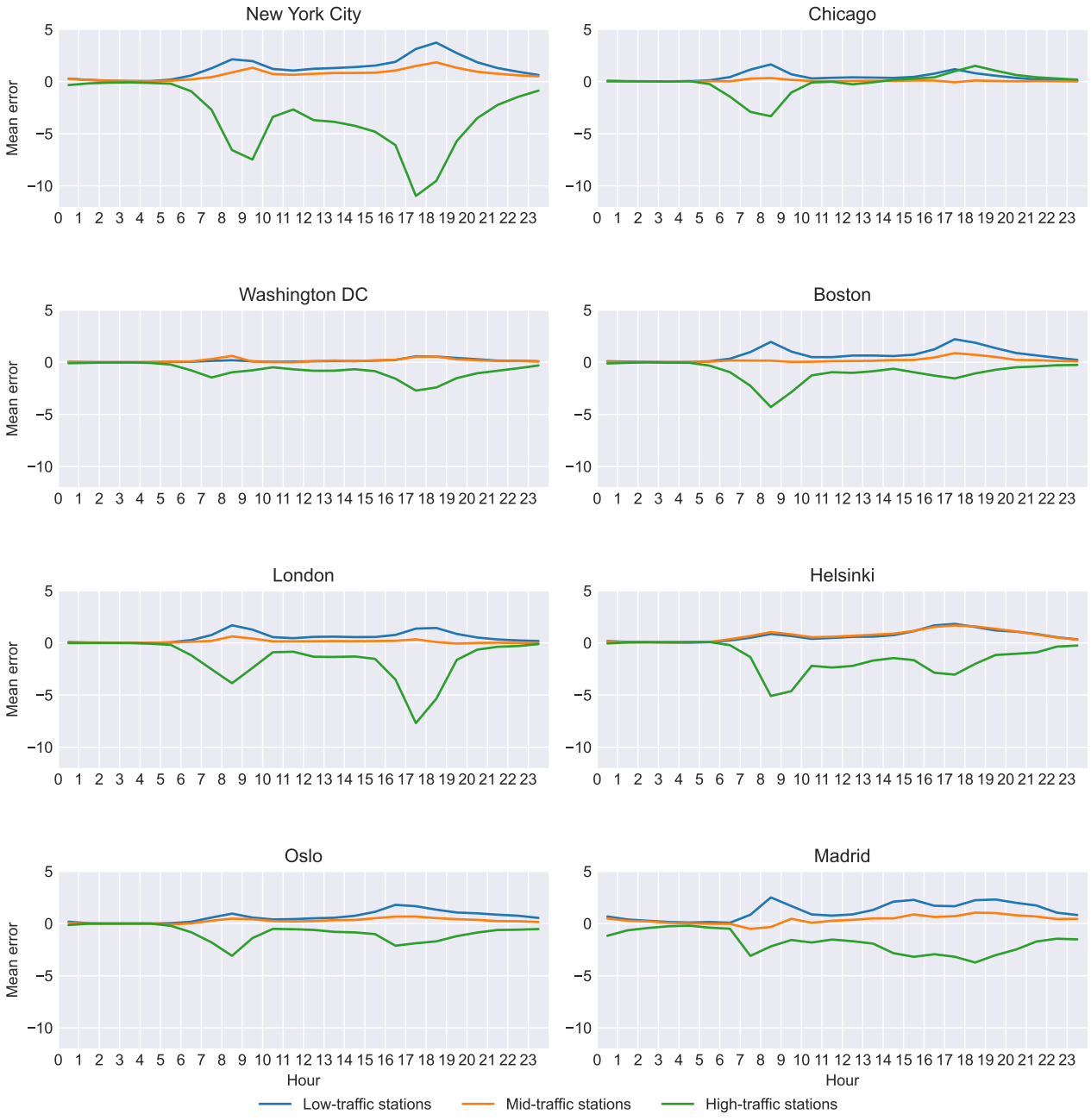
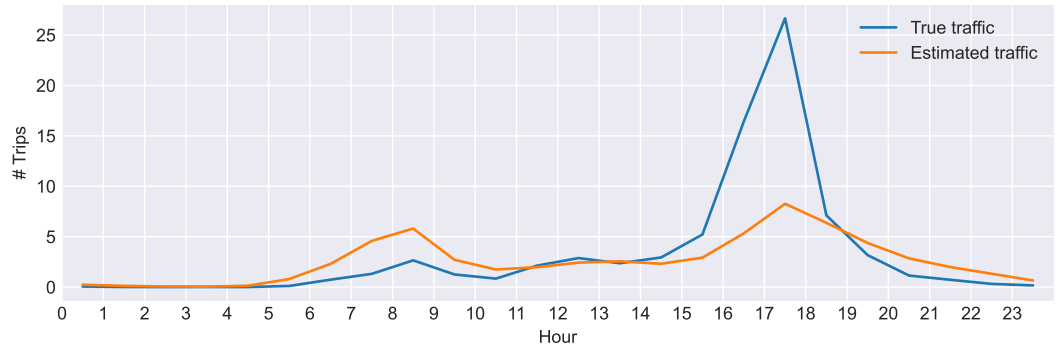
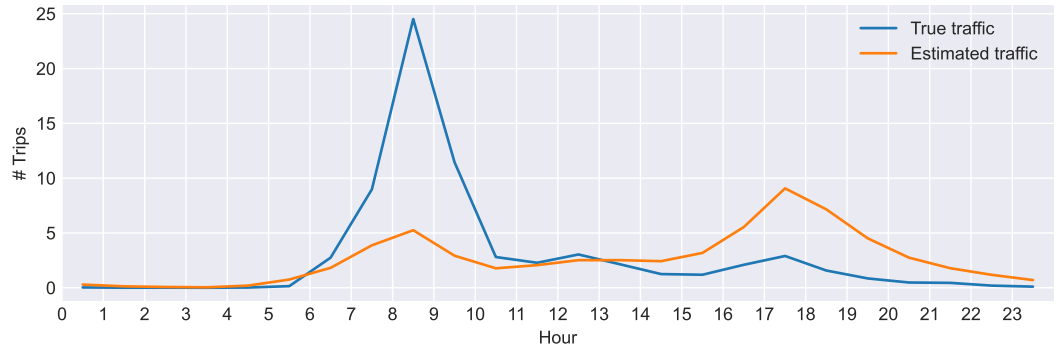


Figure 12.5: Mean error of predicted number of arrivals each hour when tested on low-, mid-, and high-traffic stations.



(a) Estimated number of departures.



(b) Estimated number of arrivals.

Figure 12.6: Predicted number of departures and arrivals for a station on Orleans St. & Merchandise Mart Plaza in Chicago.

13. Case: The New York City 2019 System Expansion

In autumn of 2019, the Citi Bike system in New York City was expanded into a new area straddling the boundary between Queens and Brooklyn. The area consisting of parts of the neighbourhoods East Williamsburg, Bushwick and Ridgewood had been identified as the next area for expansion as part of phase 3 of the New York Citi Bike system development.

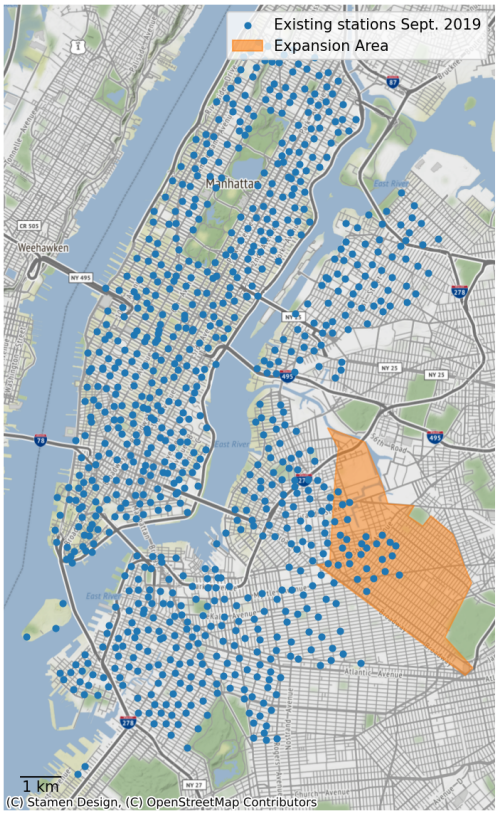
When the Citi bike system was first planned in 2009, three phases for the roll-out of the system were laid out. The area identified for phase 1 was built in 2013, while the area of phase 2 was split up as phase 2 in 2015-2017 and phase 3 which started construction in 2019 and is expected to be completed in 2023. [NYC09]

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. 13.1a.

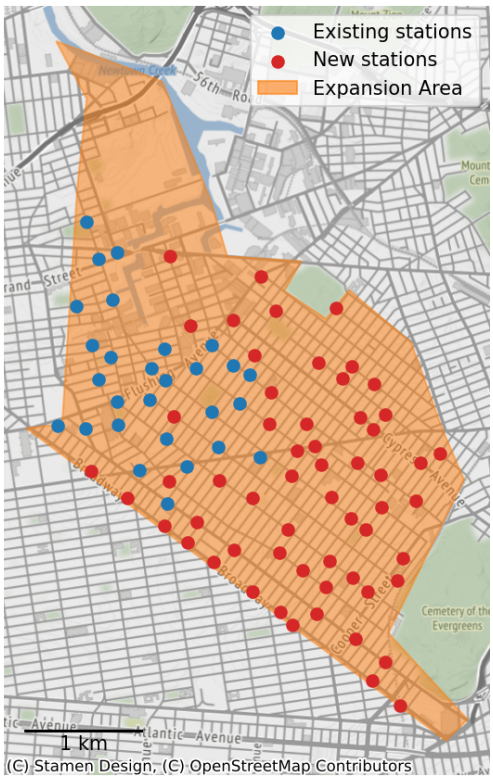
The selection of the areas for the different phases was primarily based on estimated demand for a bike share system. Factors such as the amount of residents cycling or walking to work, the percentage of workers living within 2.5 miles ($\approx$ 4 km) or 5 miles ($\approx$ 8 km) of their workplace as well as population density influenced the decision. This was done in order to optimise the popularity and thereby the profitability of the system, as the system was designed to operate without government subsidies. [NYC09]

The area for the 2019 expansion scores high in all the aforementioned factors compared to the other areas included in the phase 3 expansion, [NYC09, p.71-73] which likely contributes to the decision by the NYC Department of Transportation of this area as the first in phase 3.

At the start of the 2019 expansion in October, 28 stations existed in the expansion area. At the end of November, 58 new stations had been placed all over the expansion area, expanding the service area of the bike share system as well as filling in gaps and increasing density between existing stations in the expansion area. The existing stations as well as the stations which were placed in October and November 2019 can be seen in Fig. 13.1b. From the Department of City Planning materials, it is not obvious which factors have influenced the placement of the stations within the



(a) New York City station map September 2019 prior to expansion.



(b) Expansion area with existing stations in September 2019 in blue as well as stations placed in October and November in red.

Figure 13.1: New York City 2019 expansion area.

expansion area. However, similarly to how the expected demand for the bike share system has been instrumental to the selection of service areas for the different expansion phases, the demand can also be used to determine locations for new stations within the expansion area.

13.1 Application of Demand Model

The aim of this case study is to apply the demand model established in Chapter 11 to predict the demand in the expansion area in order to determine a station placement configuration which maximises the demand that is served by the bike share system.

In order to do so, the model is fitted to average station data for the months of 2019 prior to October. This is to fit the model only to data which existed prior to the expansion, as well as to make sure that the actual realised station placement does not influence the model coefficients. The demand prediction resulting from such a model alongside the external variables in the expansion area are shown in Fig. 13.2.

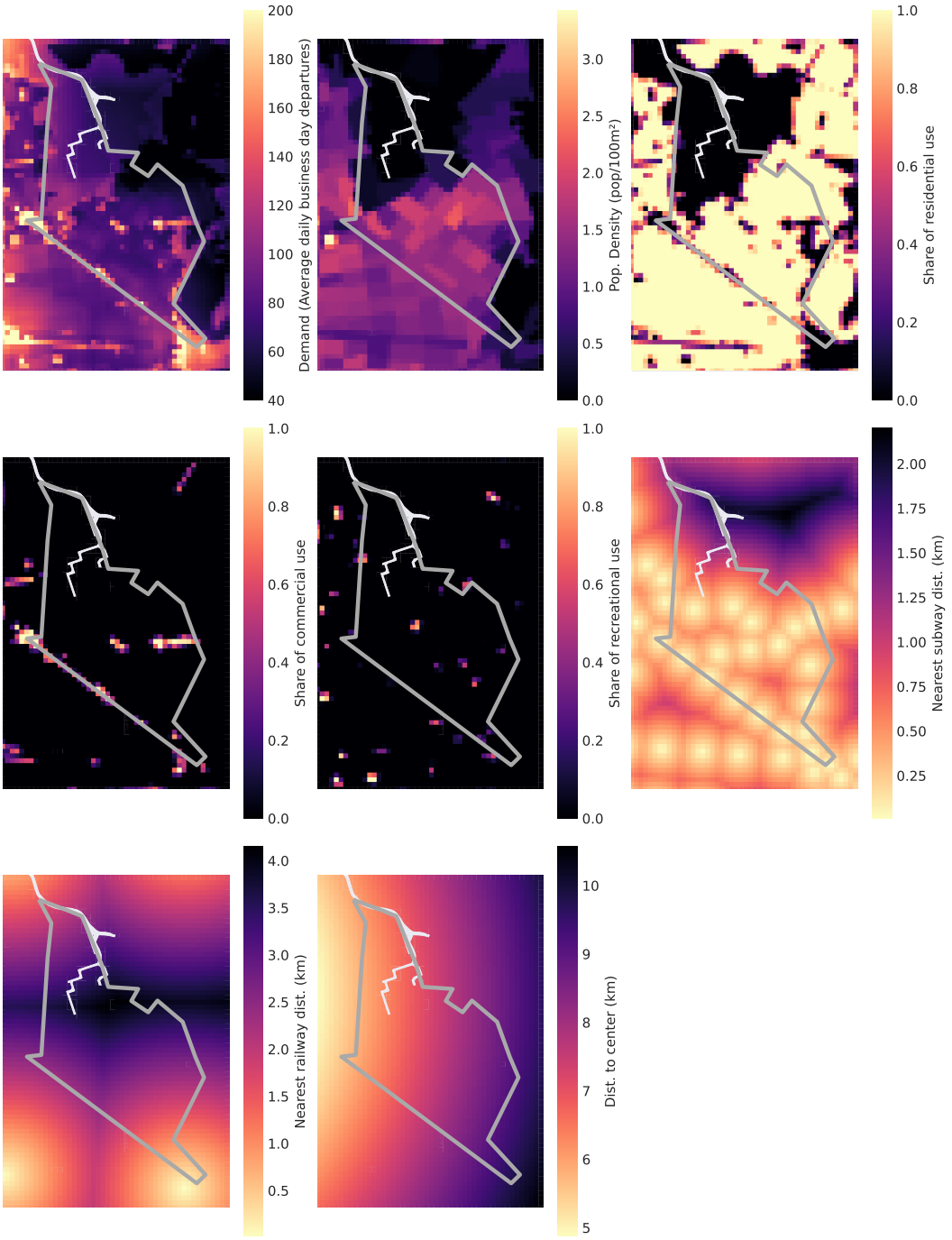


Figure 13.2: Heatmaps showing a demand model of the expansion area. The grey outline indicates the expansion area.

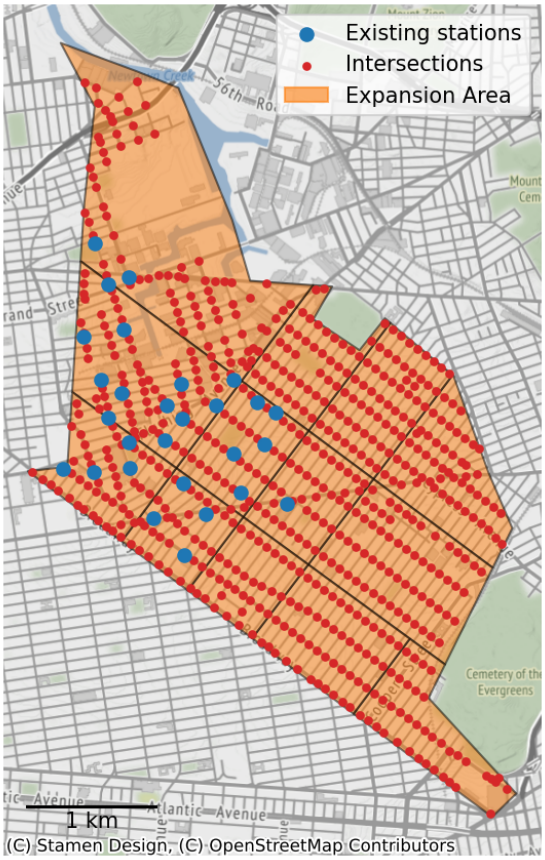


Figure 13.3: New York City 2019 expansion area intersections in subdivided polygons.

There are in theory an infinite number of possible locations for new stations. However, in order to limit the investigation to a finite subset of points, the set of all road intersections in the expansion area is used as candidate points. The intersections are 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 each road lane. Therefore, intersections which are nearer than 20m to each other are merged to one intersection at the mean point. In the expansion area, this yields $n = 643$ candidate locations as shown in Fig. 13.3.

After calculating the predicted demand for each of these candidate locations, it seems to be a simple optimisation problem to determine m stations which serve the highest demand in their service area. We can define the optimisation problem as follows.

Let $\mathcal{I} = \{i\}$ with $|\mathcal{I}| = n$ be the set of candidate locations for the placement of a station and compute the expected demand e_i for each $i \in \mathcal{I}$. Let $\mathcal{S} \subseteq \mathcal{I}$ with $|\mathcal{S}| = m$

be the set of chosen candidate locations, and define the indicator variables

$$s_i = \begin{cases} 1 & \text{if } i \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases} \tag{13.1}$$

The optimisation problem can then be formulated as

$$\underset{\mathcal{S} \subseteq \mathcal{I}}{\text{maximise}} \quad \sum_{i \in \mathcal{I}} e_i s_i, \tag{13.2}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}} s_i = m. \tag{13.3}$$

However, if the demand in the service area of the stations is the sole criterion for the selection of stations, the selected stations will be grouped together in the locations where the expected demand is the highest without regard for factors such as spacing between the stations, meaning that some demand is served by multiple stations simultaneously. Therefore, it is necessary to introduce a constraint to the optimisation problem to ensure a wider coverage. One such constraint could be to define a minimum distance d_{min} between selected stations. Doing this, the more constrained optimisation problem is

$$\underset{\mathcal{S} \subseteq \mathcal{I}}{\text{maximise}} \quad \sum_{i \in \mathcal{I}} e_i s_i, \tag{13.4}$$

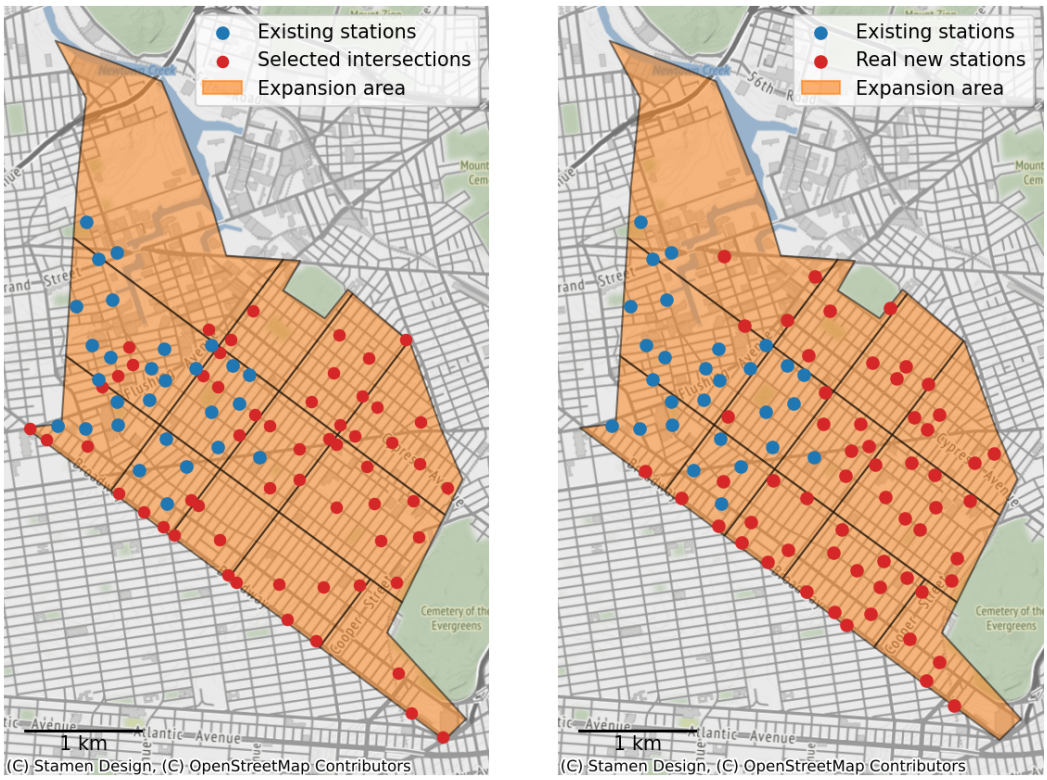
$$\text{s.t.} \quad \sum_{i \in \mathcal{I}} s_i = m, \tag{13.5}$$

$$\min_{i,j \in \mathcal{S},\ i \neq j} d(i,j) > d_{min}, \tag{13.6}$$

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

The problem with this kind of nonlinear constraint is that this combinatorial optimisation problem becomes difficult to impossible to solve using typical optimisation methods. At the same time, selecting $m = 58$ stations out of $n = 643$ possible locations yields more than $2 \cdot 10^{83}$ combinations, so it is impossible to compute the suitability of all combinations within any reasonable amount of time.

One solution to this problem is to divide the area into subsections. In Fig. 13.3, 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 problem much more tractable. 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 similarly to how the NYC department of City Planning determined the expansion areas. This yields from 1 to 8 stations to be placed per sub-area in addition to the existing stations in each area if any. The number of different combinations of station selections in each sub-area then varies from 83 to 2.5 billion which is much more suitable for optimisation by exhaustion.



(a) Selected intersections.

(b) Real implemented locations.

Figure 13.4: New York City 2019 solutions.

13.2 Comparison of Solutions

With a minimum distance between the stations of $d_{min} = 250\,\text{m}$ and demand predictions obtained from the model illustrated in Fig. 13.4a, the optimal solution is seen in Fig. 13.4a. For comparison, the realised solution as implemented by the Citi Bike system is seen in Fig. 13.4b.

Comparing the determined optimal solution to the implemented solution using the objective function for the optimisation in Eq. (13.2) shows a clear advantage to the solution determined by the optimisation procedure, with a score of 6749.2 for the implemented solution vs. a score of 8513.6 for the solution obtained by the optimisation. It is thus clear that the solution determined serves more expected demand than what was really implemented. However, this conclusion comes with a number of caveats which can be divided into three categories.

Problems with the solution obtained by optimisation: The way of determining the optimal solution by subdividing the expansion area has the major flaw of border effects. As the optimisation of station placement is done separately in each subdivision in isolation, 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. 13.4a.

Selection of objective function: The objective function is designed purely to optimise the demand met by the bike share stations. As the optimisation is performed with regard to this objective function. At the same time, 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 Fig. 13.4b. They may for example put more weight on the 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 optimise a certain objective function performs better than a solution which is designed to optimise a different unknown objective.

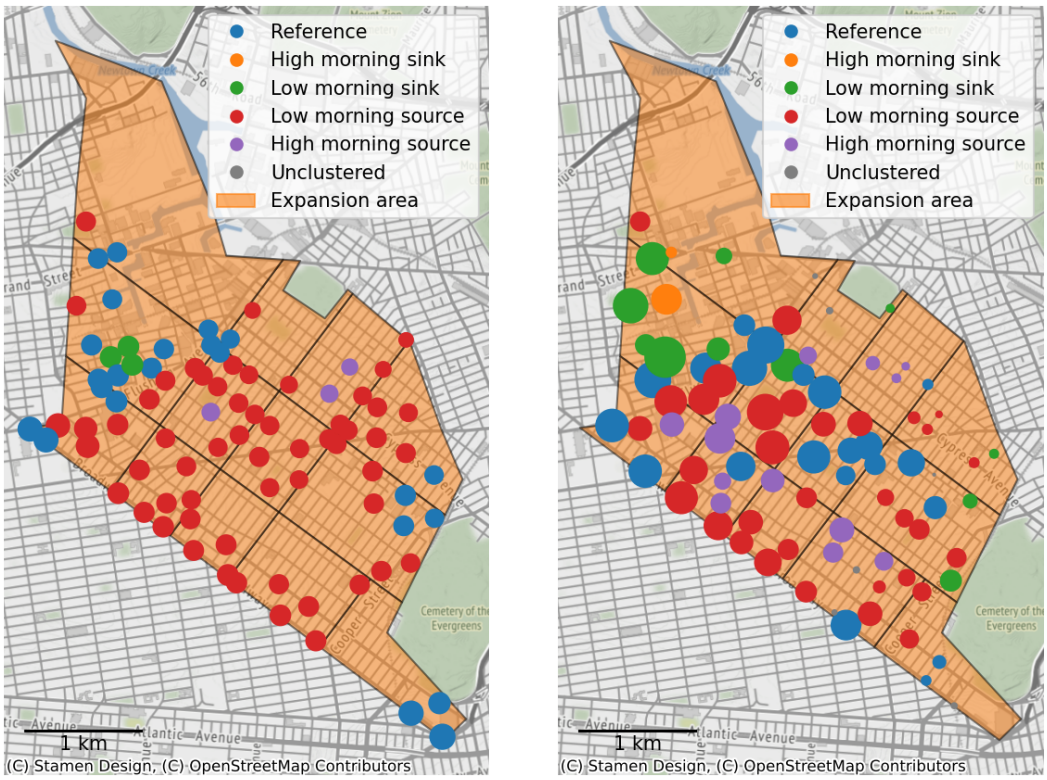
Real life considerations: When looking purely theoretically at station placement, we do not have to concern ourselves with factors such as whether there is space on the street or sidewalk for a bike share station, or potential complaints from residents. The selection of station locations by the Department of Transportation may be based on opaque criteria, but it is through this non-rigid selection process that they can accommodate practical concerns about fitting in to the urban environment. With that being said, an optimisation approach with intersections as candidate locations can be a suitable first step, which can then be further refined by subsequent consideration of practical considerations followed by feedback rounds to gather input from local residents.

13.3 Comparison of Predicted Traffic

In order to get an idea of how well the model predictions of the traffic at the stations matches with the real world, we use the models to predict the traffic patterns including both the predicted cluster type and volume of traffic for each selected station in the expansion area. This can be seen in Fig. 13.5a. The model is trained on data from September 2019 and thus has no prior knowledge about the new stations. For comparison, the real traffic data from November 2019 is shown in Fig. 13.5b.

From the figures, it seems like 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 mainly 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.

For the volume, the prediction consistently underestimates the amount of traffic with only a handful of exceptions at the edge of the expansion area where the model overestimates the traffic. In short, while the models are not completely inaccurate, they fail to predict the large concentration of trips in the middle of the expansion



(a) Selected intersections with predicted cluster and predicted amount of traffic. Model trained on data from September 2019.

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

Figure 13.5: New York City 2019 stations with colours corresponding to traffic pattern cluster type and sizes corresponding to number of trips.

area. This may be related to the fact that most of these stations were still new in November 2019, and might attract more trips than it would otherwise do. However, since the models predict the average traffic pattern and do not take sudden surges in traffic into account, the behaviour of the real stations may get closer to the predicted behaviour over time.

14. Conclusion

The aim of this project was to develop a way to predict average daily traffic patterns of stations using a combination of clustering analysis and statistical modelling. In accordance with this, several clustering methods and statistical tools have been presented and used to inform decisions in the design of the models.

A preliminary clustering analysis found that using k-means with five clusters yielded good separation between stations based on the shape of the traffic patterns. These traffic patterns are shared between the cities, albeit with some differences which can be attributed to local factors in each city. The clusters were classified as low morning source, high morning source, low morning sink, and high morning sink as well as a balanced reference cluster for all cities except for Oslo. In the further analysis, it was found that the cluster type of a station is strongly associated with how the surrounding area of that station is used. For instance in the cities in the US, land use such as residential and commercial use were important predictors of station type, possibly due to zoning regulation standards in the country. Common between all cities is that population density is a powerful predictor of station type with morning sources being in highly populated areas. City infrastructure in terms of location of transit systems can also be a good predictor of station type, however this is highly dependent on how people tend to use these systems in their commute, particularly if the transit trips are contained within the city boundaries or if they stem from workers commuting into the city.

Findings from the clustering were also reflected in the demand models, with morning sinks having a higher increase in demand due large concentrations of trips in these stations, while population density is also a very important predictor of demand. Distance to city center was also found to be very important for demand which is in line with observations from previous studies.

When testing the models between cities, it was found that local differences can play a large role when predicting the cluster type of stations in the bike sharing system. Models trained and tested between US cities perform better on average than models trained and tested between European cities when predicting station types, which may be due to more cultural homogeneity between cities in the US. This indicates that if one were to predict the cluster types of potential stations in a new city based on a model trained on a city with an existing bike sharing system, then

similarity in culture can be very important for the accuracy of the model. However, cultural differences were not found to be important when predicting the demand of the demand of the stations. Instead, differences in overall traffic of the systems can be a bigger contributor to prediction error, meaning that the expected overall amount of traffic in a city has to be taken to account in the planning process of a new bike sharing system when using demand models from other cities.

Combining the prediction of traffic shape and demand resulted in the prediction of average daily traffic which had a reasonable precision when tested on stations in the bike sharing systems, with the highest prediction error being in New York City where the prediction was off by 8 trips on average in the rush hours. The error of the traffic estimation is also very dependent on the actual demand of the station, where the models tend to underestimate the traffic for high-traffic stations while also overestimating the traffic for low-traffic stations. The demand model was applied to the case of Citi Bike's system expansion in the autumn of 2019 with a simple optimisation problem concerning the placement of new stations. While the obtained solution was predicted to cover more demand, it did have a lower coverage of the area. When testing the traffic prediction on the obtained stations, it was found that the traffic of the real implemented stations was more concentrated in the middle of the expansion area than expected from the model, meaning that more can still be learned with regards to how urban features can affect bike sharing as a whole.

Bibliography

[AB21] T. Askarzadeh and R. Bridgelall, "Micromobility Station Placement Optimization for a Rural Setting," *Journal of Advanced Transportation*, vol. 2021, e9808922, Sep. 2021. DOI: 10.1155/2021/9808922.

[BB12] D. Buck and R. Buehler, "Bike Lanes and Other Determinants of Capital Bikeshare Trips," 2012.

[Ber44] J. Berkson, "Application of the Logistic Function to Bio-Assay," *Journal of the American Statistical Association*, vol. 39, no. 227, pp. 357–365, Sep. 1944. DOI: 10.1080/01621459.1944.10500699.

[Bis06] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York: Springer, 2006.

[Bor+11] P. Borgnat, C. Robardet, J.-B. Rouquier, P. Abry, E. Fleury, and P. Flandrin, "Shared Bicycles in a City: A Signal Processing and Data Analysis Perspective," *Advances in Complex Systems*, vol. 14, no. 3, pp. 415–438, Jun. 2011. DOI: 10.1142/S0219525911002950.

[Chi+20] F. Chiariotti, C. Pielli, A. Zanella, and M. Zorzi, "A Bike-sharing Optimization Framework Combining Dynamic Rebalancing and User Incentives," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 14, no. 3, 11:1–11:30, Feb. 2020. DOI: 10.1145/3376923.

[CMF18] L. Conrow, A. T. Murray, and H. A. Fischer, "An optimization approach for equitable bicycle share station siting," *Journal of Transport Geography*, vol. 69, pp. 163–170, May 2018. DOI: 10.1016/j.jtrangeo.2018.04.023.

[CO14] E. Côme and L. Oukhellou, "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*, vol. 5, no. 3, 39:1–39:21, Jul. 2014. DOI: 10.1145/2560188.

[Cox58] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958.

[CYI18] D. Çelebi, A. Yörüsün, and H. Işık, "Bicycle sharing system design with capacity allocations," *Transportation Research Part B: Methodological*, vol. 114, pp. 86–98, Aug. 2018. DOI: 10.1016/j.trb.2018.05.018.

[Cô+14] E. Côme, A. Randriamanamihaga, L. Oukhellou, and P. Aknin, "Spatio-temporal Analysis of Dynamic Origin-Destination Data Using Latent Dirichlet Allocation. Application to Vélib' Bikesharing System of Paris.," Jul. 2014.

[Dad12] D. W. Daddio, "Maximizing Bicycle Sharing: An Empirical Analysis of Capital Bikeshare Usage," M.S. thesis, University of North Carolina, Chapel Hill, 2012.

[DB79] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979. DOI: 10.1109/TPAMI.1979.4766909.

[DeM+21] P. DeMaio, C. Yu, O. O'Brien, R. Rabello, S. Chou, and T. Benicchio, "The Meddin Bike-sharing World Map - Mid-2021 Report," Tech. Rep., Oct. 2021, p. 26.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. New York: Wiley, 2001.

[Dun73] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, Jan. 1973. DOI: 10.1080/01969727308546046.

[EU20] E. Eren and V. E. Uz, "A review on bike-sharing: The factors affecting bike-sharing demand," *Sustainable Cities and Society*, vol. 54, p. 101 882, Mar. 2020. DOI: 10.1016/j.scs.2019.101882.

[FAZ17] Y. Feng, R. C. Affonso, and M. Zolghadri, "Analysis of bike sharing system by clustering: The Vélib' case," *IFAC-PapersOnLine*, 20th IFAC World Congress, vol. 50, no. 1, pp. 12 422–12 427, Jul. 2017. DOI: 10.1016/j.ifacol.2017.08.2430.

[FIE16] A. Faghih-Imani and N. Eluru, "Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of New York CitiBike system," *Journal of Transport Geography*, vol. 54, pp. 218–227, Jun. 2016. DOI: 10.1016/j.jtrangeo.2016.06.008.

[FR15] I. Frade and A. Ribeiro, "Bike-sharing stations: A maximal covering location approach," *Transportation Research Part A: Policy and Practice*, vol. 82, pp. 216–227, Dec. 2015. DOI: 10.1016/j.tra.2015.09.014.

[FT01] L. Fahrmeir and G. Tutz, *Multivariate statistical modelling based on generalized linear models*, 2nd ed. / with contributions from Wolfgang Hennevogl, ser. Springer series in statistics. New York: Springer, 2001.

[FTA02] FTA, "Transit-Oriented Development and Joint Development in the United States: A Literature Review," Washington, DC, USA, Tech. Rep., 2002, p. 144.

[Hyl+18] M. Hyland, Z. Hong, H. K. R. d. F. Pinto, and Y. Chen, "Hybrid cluster-regression approach to model bikeshare station usage," *Transportation Research Part A: Policy and Practice*, Smart urban mobility, vol. 115, pp. 71–89, Sep. 2018. DOI: 10.1016/j.tra.2017.11.009.

[JD88] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, 1st ed. Prentice Hall, 1988.

[KR05] L. Kaufmann and P. J. Rousseeuw, *Finding Groups in Data*, 2nd ed. John Wiley & Sons, Inc, 2005.

[Li+19] Y. Li, Z. Zhu, D. Kong, M. Xu, and Y. Zhao, "Learning Heterogeneous Spatial-Temporal Representation for Bike-Sharing Demand Prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 1004–1011, Jul. 2019. DOI: 10.1609/aaai.v33i01.33011004.

[Ma+22] X. Ma, Y. Yin, Y. Jin, M. He, and M. Zhu, "Short-Term Prediction of Bike-Sharing Demand Using Multi-Source Data: A Spatial-Temporal Graph Attentional LSTM Approach," *Applied Sciences*, vol. 12, no. 3, p. 1161, Jan. 2022. DOI: 10.3390/app12031161.

[MS14] E. W. Martin and S. A. Shaheen, "Evaluating public transit modal shift dynamics in response to bikesharing: A tale of two U.S. cities," *Journal of Transport Geography*, vol. 41, pp. 315–324, Dec. 2014. DOI: 10.1016/j.jtrangeo.2014.06.026.

[MT11] H. Madsen and P. Thyregod, *Introduction to general and generalized linear models*, ser. Chapman & Hall/CRC texts in statistical science series. Boca Raton: CRC Press, 2011.

[NSG16] R. B. Noland, M. J. Smart, and Z. Guo, "Bikeshare trip generation in New York City," *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 164–181, Dec. 2016. DOI: 10.1016/j.tra.2016.08.030.

[NSG19] ——, "Bikesharing Trip Patterns in New York City: Associations with Land Use, Subways, and Bicycle Lanes," *International Journal of Sustainable Transportation*, vol. 13, no. 9, pp. 664–674, Oct. 2019. DOI: 10.1080/15568318.2018.1501520.

[NYC09] NYC Department of City Planning, "Bike-Share Opportunities in New York City," Tech. Rep., 2009, p. 142.

[OCB14] O. O'Brien, J. Cheshire, and M. Batty, "Mining bicycle sharing data for generating insights into sustainable transport systems," *Journal of Transport Geography*, vol. 34, pp. 262–273, Jan. 2014. DOI: 10.1016/j.jtrangeo.2013.06.007.

[OS15] E. O'Mahony and D. B. Shmoys, "Data Analysis and Optimization for (Citi)Bike Sharing," p. 8, 2015.

[PS17] C. Park and S. Y. Sohn, "An optimization approach for the placement of bicycle-sharing stations to reduce short car trips: An application to the city of Seoul," *Transportation Research Part A: Policy and Practice*, vol. 105, pp. 154–166, Nov. 2017. DOI: 10.1016/j.tra.2017.08.019.

[PZ17] A. Pal and Y. Zhang, "Free-floating bike sharing: Solving real-life large-scale static rebalancing problems," *Transportation Research Part C: Emerging Technologies*, vol. 80, pp. 92–116, Jul. 2017. DOI: 10.1016/j.trc.2017.03.016.

[Rom+12] J. P. Romero, A. Ibeas, J. L. Moura, J. Benavente, and B. Alonso, "A Simulation-optimization Approach to Design Efficient Systems of Bike-sharing," *Procedia - Social and Behavioral Sciences*, vol. 54, pp. 646–655, Oct. 2012. DOI: 10.1016/j.sbspro.2012.09.782.

[Rou87] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987. DOI: 10.1016/0377-0427(87)90125-7.

[Sen08] P. Senin, "Dynamic Time Warping Algorithm Review," Tech. Rep., 2008.

[SLM15] A. Sarkar, N. Lathia, and C. Mascolo, "Comparing cities' cycling patterns using online shared bicycle maps," *Transportation*, vol. 42, no. 4, pp. 541–559, Jul. 2015. DOI: 10.1007/s11116-015-9599-9.

[The69] H. Theil, "A Multinomial Extension of the Linear Logit Model," p. 10, 1969.

[VGM11] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns," *Procedia - Social and Behavioral Sciences*, vol. 20, pp. 514–523, 2011. DOI: 10.1016/j.sbspro.2011.08.058.

[VM11] P. Vogel and D. C. Mattfeld, "Strategic and Operational Planning of Bike-Sharing Systems by Data Mining – A Case Study," in *Computational Logistics*, J. W. Böse, H. Hu, C. Jahn, X. Shi, R. Stahlbock, and S. Voß, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2011, pp. 127–141. DOI: 10.1007/978-3-642-24264-9_10.

[XC20] S. J. Xu and J. Y. J. Chow, "A longitudinal study of bike infrastructure impact on bikesharing system performance in New York City," *International Journal of Sustainable Transportation*, vol. 14, no. 11, pp. 886–902, Sep. 2020. DOI: 10.1080/15568318.2019.1645921.

[Yan+16] Z. Yang, J. Hu, Y. Shu, P. Cheng, J. Chen, and T. Moscibroda, "Mobility Modeling and Prediction in Bike-Sharing Systems," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, Singapore Singapore: ACM, Jun. 2016, pp. 165–178. DOI: 10.1145/2906388.2906408.

[Yan+20a] H. Yang, Y. Zhang, L. Zhong, X. Zhang, and Z. Ling, "Exploring spatial variation of bike sharing trip production and attraction: A study based on Chicago's Divvy system," *Applied Geography*, vol. 115, p. 102 130, Feb. 2020. DOI: 10.1016/j.apgeog.2019.102130.

[Yan+20b] Y. Yang, A. Heppenstall, A. Turner, and A. Comber, "Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems," *Computers, Environment and Urban Systems*, vol. 83, p. 101 521, Sep. 2020. DOI: 10.1016/j.compenvurbsys.2020.101521.

[Zha+18] C. Zhang, L. Zhang, Y. Liu, and X. Yang, "Short-term Prediction of Bike-sharing Usage Considering Public Transport: A LSTM Approach," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, HI: IEEE, Nov. 2018, pp. 1564–1571. DOI: 10.1109/ITSC.2018.8569726.

[Zho15] X. Zhou, "Understanding Spatiotemporal Patterns of Biking Behavior by Analyzing Massive Bike Sharing Data in Chicago," *PLOS ONE*, vol. 10, no. 10, e0137922, Oct. 2015. DOI: 10.1371/journal.pone.0137922.

[ZM18] Y. Zhang and Z. Mi, "Environmental benefits of bike sharing: A big data-based analysis," *Applied Energy*, vol. 220, pp. 296–301, Jun. 2018. DOI: 10.1016/j.apenergy.2018.03.101.

[ZQB19] Z. Zhang, C. Qian, and Y. Bian, "Bicycle–metro integration for the 'last mile': Visualizing cycling in Shanghai," *Environment and Planning A: Economy and Space*, vol. 51, no. 7, pp. 1420–1423, Oct. 2019. DOI: 10.1177/0308518X18816695.

[ZWD15] J. Zhao, J. Wang, and W. Deng, "Exploring bikesharing travel time and trip chain by gender and day of the week," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 251–264, Sep. 2015. DOI: 10.1016/j.trc.2015.01.030.

Data References

[Ber] Bergen City Bike, *Historical data - Bergen City Bike*. URL: https://bergenbysykkel.no/en/open-data/historical (visited on 11/03/2021).

[BIX] BIXI Montréal, *Open Data*. URL: https://bixi.com/en/open-data (visited on 11/03/2021).

[Blu] Bluebikes Boston, *Bluebikes System Data*. URL: http://www.bluebikes.com/system-data (visited on 11/03/2021).

[Bos] Boston Maps, *Boston Zoning Subdistricts - Analyze Boston*. URL: https://data.boston.gov/dataset/boston-zoning-subdistricts1 (visited on 04/25/2022).

[Cap] Capital Bikeshare, *System Data*. URL: http://www.capitalbikeshare.com/system-data (visited on 11/03/2021).

[Cita] Citi Bike NYC, *Citi Bike System Data*. URL: http://ride.citibikenyc.com/system-data (visited on 11/03/2021).

[Citb] City of Chicago, *Boundaries - Zoning Districts (current)*. URL: https://data.cityofchicago.org/Community-Economic-Development/Boundaries-Zoning-Districts-current-/7cve-jgbp (visited on 04/25/2022).

[Citc] ——, *Divvy Bicycle Stations*. URL: https://data.cityofchicago.org/Transportation/Divvy-Bicycle-Stations-All-Map/bk89-9dk7 (visited on 04/22/2022).

[Citd] City of Washington, DC, *Capital Bike Share Locations*. URL: https://opendata.dc.gov/datasets/DCGIS::capital-bike-share-locations/about (visited on 04/22/2022).

[Cite] ——, *Zoning Regulations of 2016*. URL: https://opendata.dc.gov/datasets/DCGIS::zoning-regulations-of-2016/about (visited on 04/25/2022).

[Div] Divvy Chicago, *Divvy System Data*. URL: http://www.divvybikes.com/system-data (visited on 11/03/2021).

[Ecoa] EcoBici Buenos Aires, *Buenos Aires Data*. URL: https://data.buenosaires.gob.ar/dataset/bicicletas-publicas (visited on 11/03/2021).

[Ecob] EcoBici Ciudad de México, *Open data | ECOBICI*. URL: https://www.ecobici.cdmx.gob.mx/en/informacion-del-servicio/open-data (visited on 11/03/2021).

[EMT] EMT Madrid, *Datos estáticos BICIMAD*. URL: https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1) (visited on 11/03/2021).

[Eura] European Commission, *Cycling Policy and Background*. URL: https://transport.ec.europa.eu/transport-themes/clean-transport-urban-transport/cycling/guidance-cycling-projects-eu/cycling-policy-and-background_en (visited on 06/02/2022).

[Eurb] European Environment Agency (EEA), *Urban Atlas 2012 — Copernicus Land Monitoring Service*, Land item. URL: https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012 (visited on 05/25/2022).

[Eurc] ——, *Urban Atlas 2018 — Copernicus Land Monitoring Service*, Land item. URL: https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018 (visited on 04/25/2022).

[Hel] Helsinki City Bike, *City bike stations' Origin-Destination (OD) data - Helsinki Region Infoshare*. URL: https://hri.fi/data/en_GB/dataset/helsingin-ja-espoon-kaupunkipyorilla-ajatut-matkat (visited on 11/03/2021).

[Jus] Just Eat Cycles Edinburgh, *Historical data - Just Eat Cycles*. URL: https://edinburghcyclehire.com/open-data/historical (visited on 11/03/2021).

[Lyf] Lyft San Francisco, *System Data | Bay Wheels | Lyft*. URL: https://lyft.com/bikes/bay-wheels/system-data (visited on 11/03/2021).

[Met] Metro Bike Share, *Data*. URL: https://bikeshare.metro.net/about/data/ (visited on 11/03/2021).

[MiB] MiBici Guadalajara, *MIBICI | Open data*. URL: https://www.mibici.net/en/open-data/ (visited on 11/03/2021).

[Nat] Nathalie P. Voorhess Center for neighborhood and Community Improvement, *Who Can Live in Chicago?* URL: https://voorheescenter.uic.edu/who-can-live-in-chicago/ (visited on 05/05/2022).

[New] New York City Department of City Planning, *NYC GIS Zoning Features*. URL: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-gis-zoning.page (visited on 04/25/2022).

[Nic] Nice Ride Minnesota, *System Data*. URL: http://www.niceridemn.com/system-data (visited on 11/03/2021).

[Opea] OpenStreetMap, *Overpass API*. URL: http://overpass-api.de/ (visited on 04/25/2022).

[Opeb] OpenStreetMap Contributors, *Tag:place=city - OpenStreetMap Wiki*. URL: https://wiki.openstreetmap.org/wiki/Tag:place%3Dcity (visited on 05/31/2022).

[Osl] Oslo City Bike, *Historical data - Oslo City Bike*. URL: https://oslobysykkel.no/en/open-data/historical (visited on 11/03/2021).

[Traa] Tranport for London, *Transport for London Unified API*. URL: http://api.tfl.gov.uk (visited on 04/22/2022).

[Trab] Transport for London, *Cycling.data.tfl.gov.uk*. URL: https://cycling.data.tfl.gov.uk/ (visited on 11/03/2021).

[Tro] Trondheim City Bike, *Historical data - Trondheim City Bike*. URL: https://trondheimbysykkel.no/en/open-data/historical (visited on 11/03/2021).

[UN] UN DESA, *68% of the world population projected to live in urban areas by 2050, says UN*. URL: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html (visited on 05/26/2022).

[Uni] United Nations, *Goal 11 | Department of Economic and Social Affairs*. URL: https://sdgs.un.org/goals/goal11 (visited on 06/02/2022).

[USCB] U. C. United States Census Bureau, *Data*. URL: https://www.census.gov/data (visited on 04/25/2022).

[You] YouBike Taipei, *YouBike*票證刷卡資料. URL: https://drive.google.com/drive/folders/1QsROgp8AcER6qkTJDxpuV8Mt1Dy6lGQO (visited on 11/03/2021).

A. Clustering Results

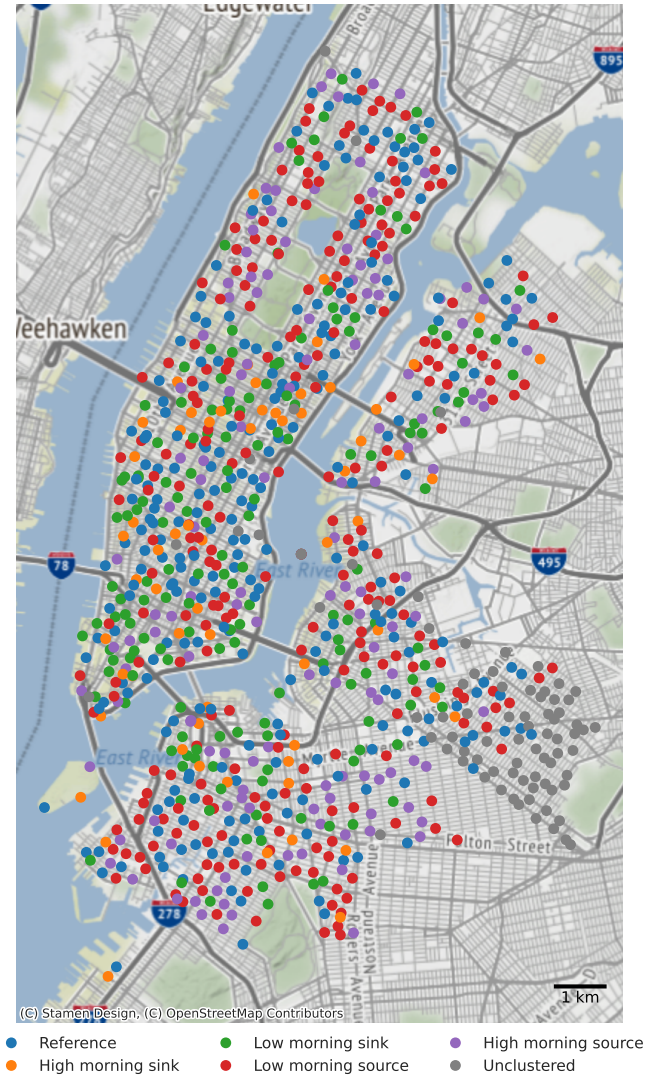


Figure A.1: Results of clustering for New York City.

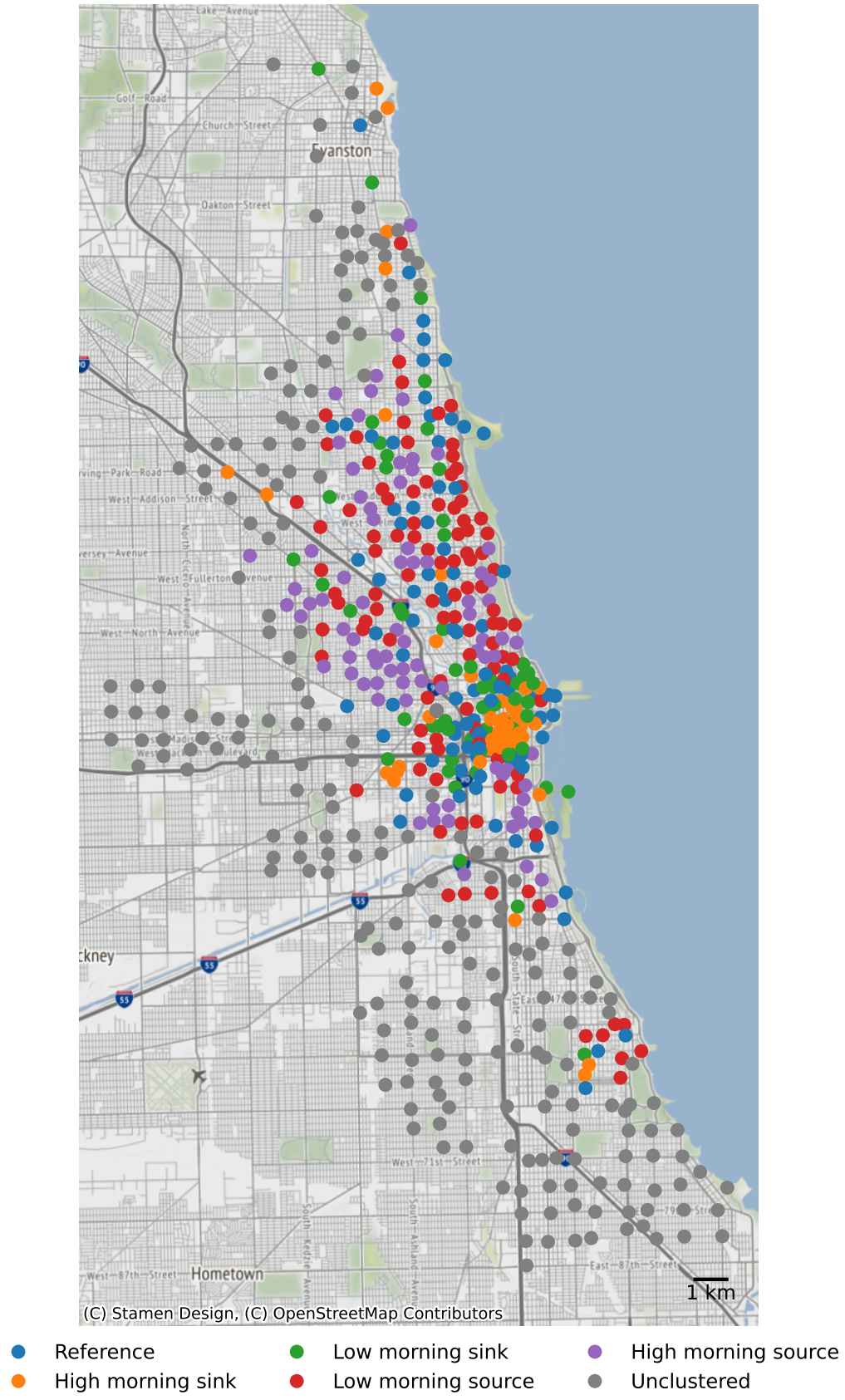


Figure A.2: Results of clustering for Chicago.

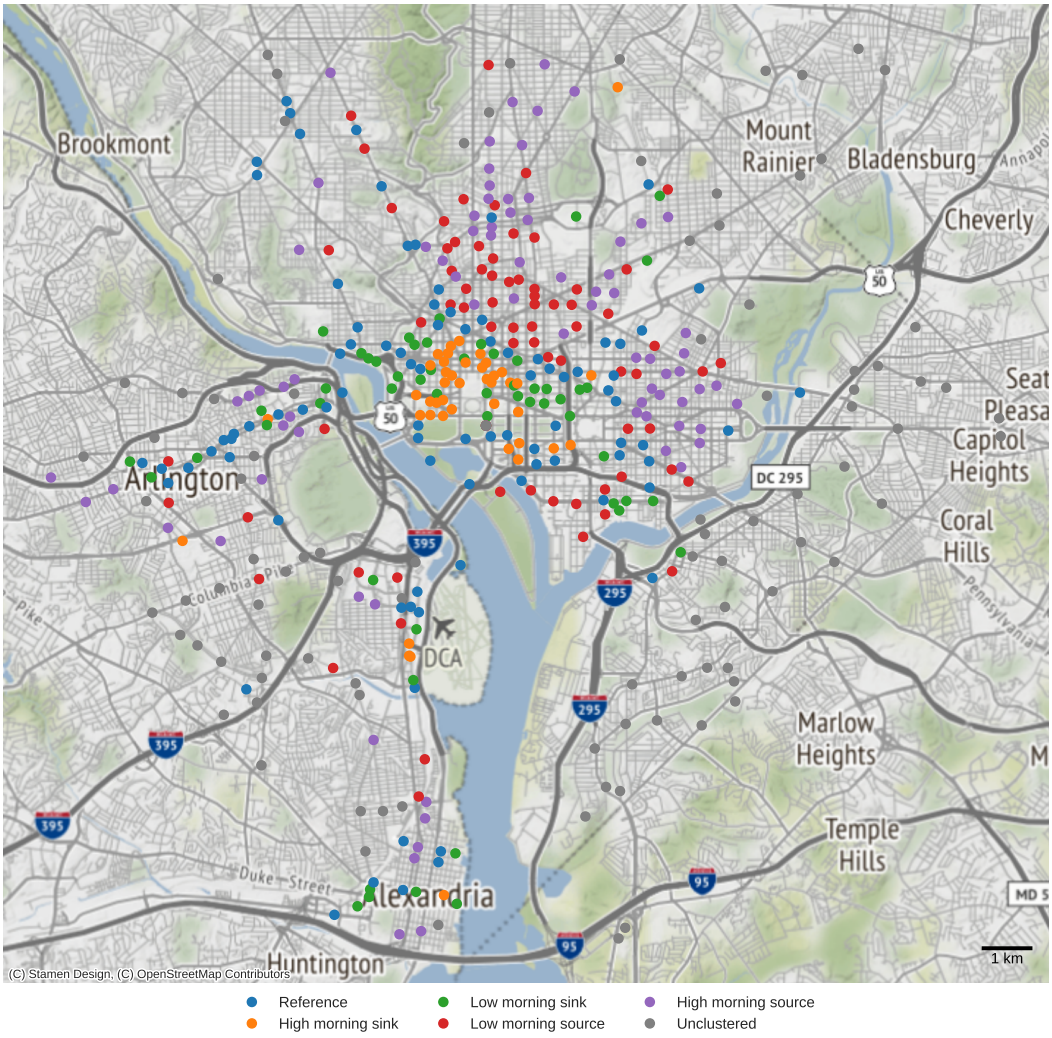


Figure A.3: Results of clustering for Washington DC.

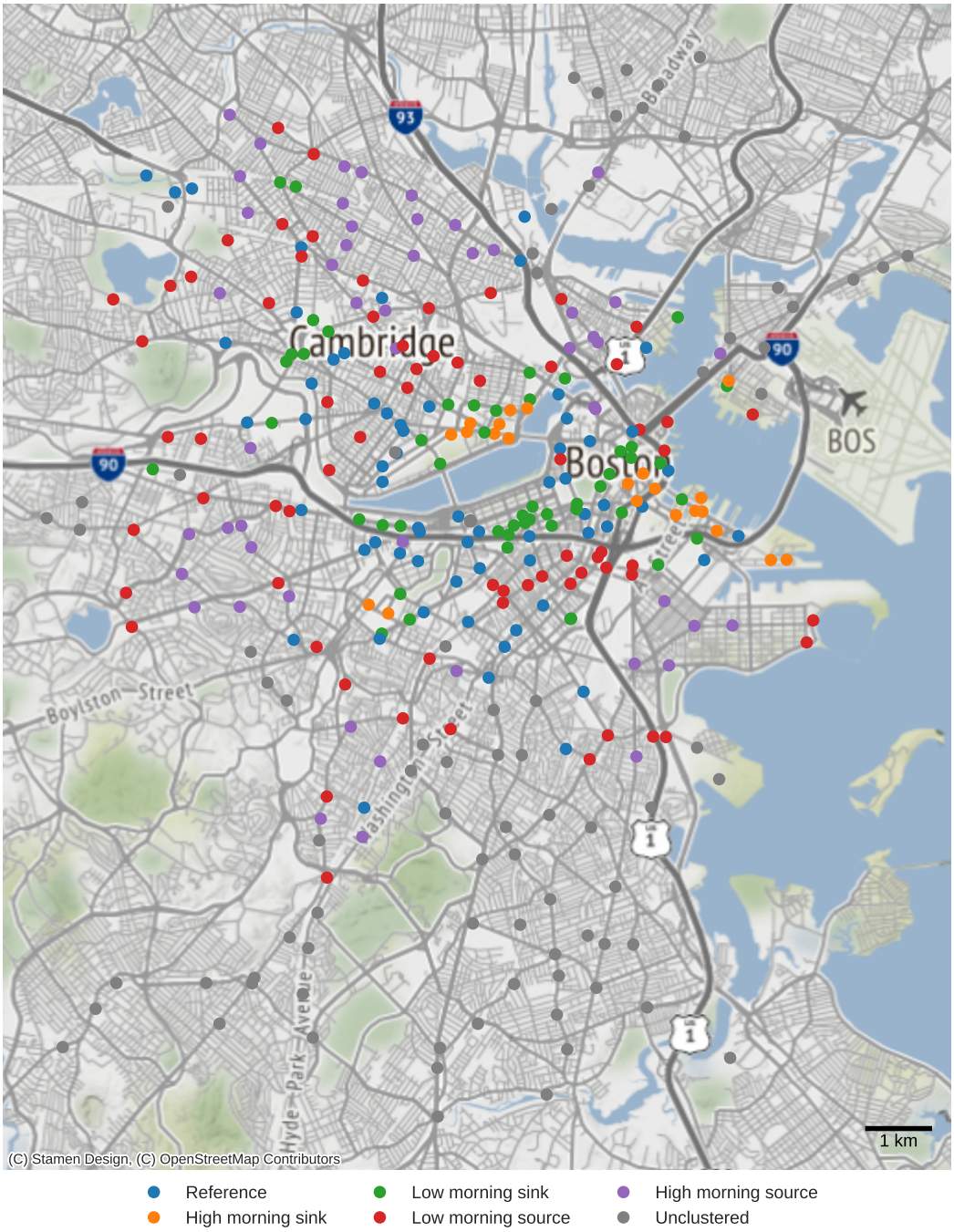


Figure A.4: Results of clustering for Boston.

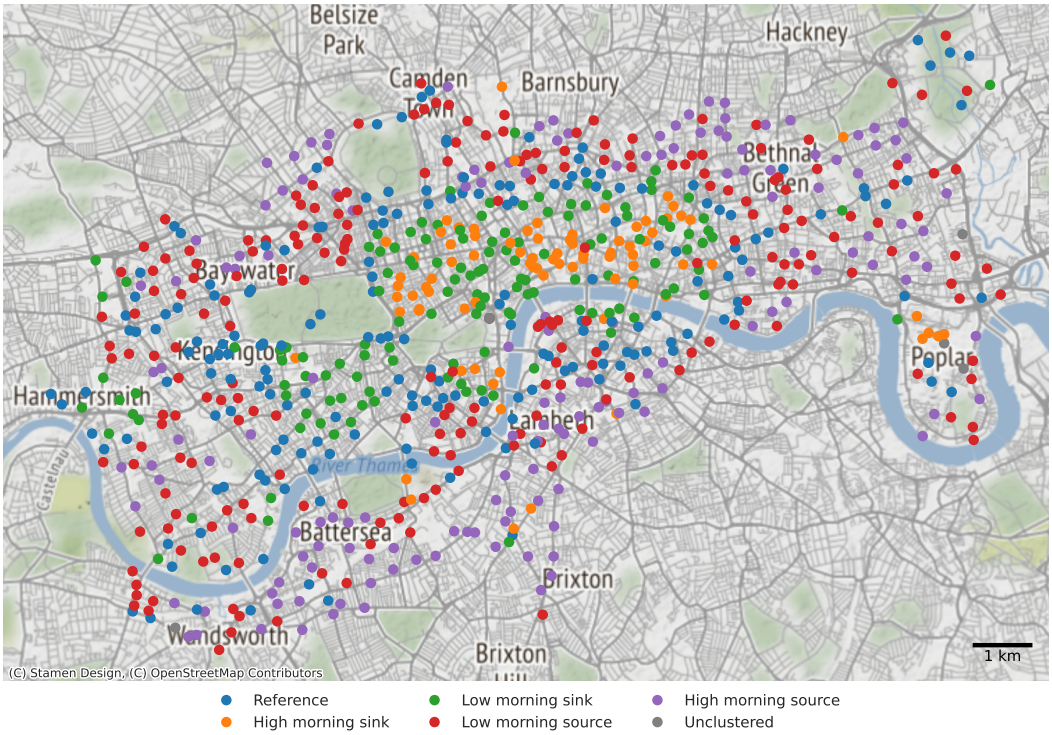


Figure A.5: Results of clustering for London.

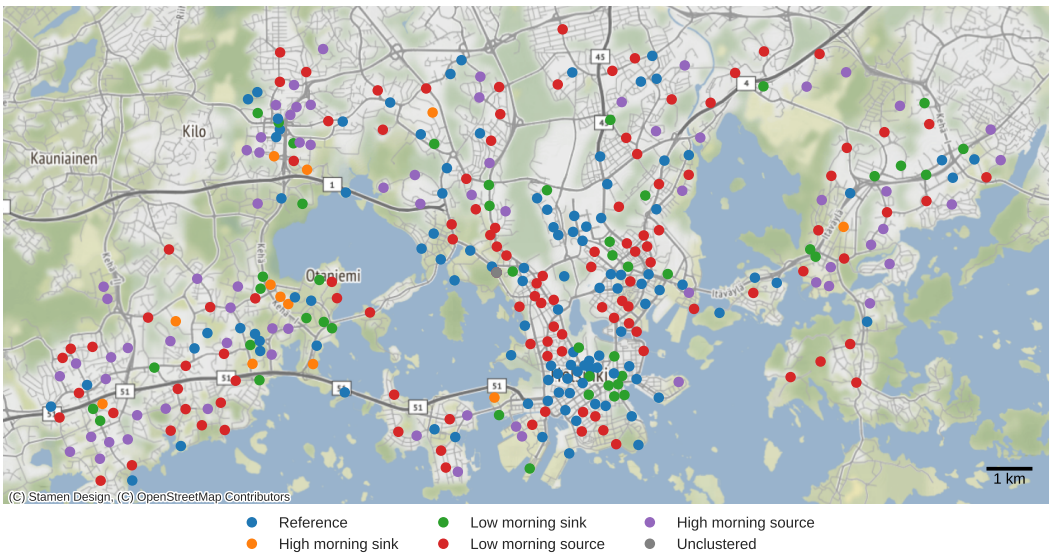


Figure A.6: Results of clustering for Helsinki.

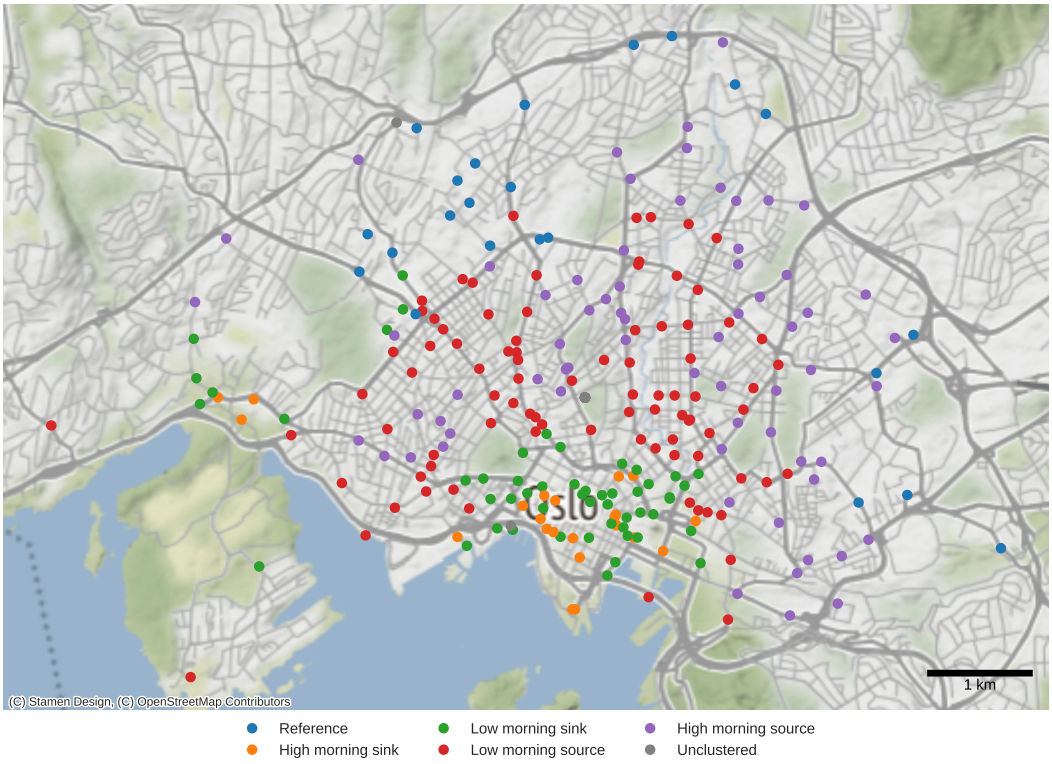


Figure A.7: Results of clustering for Oslo.

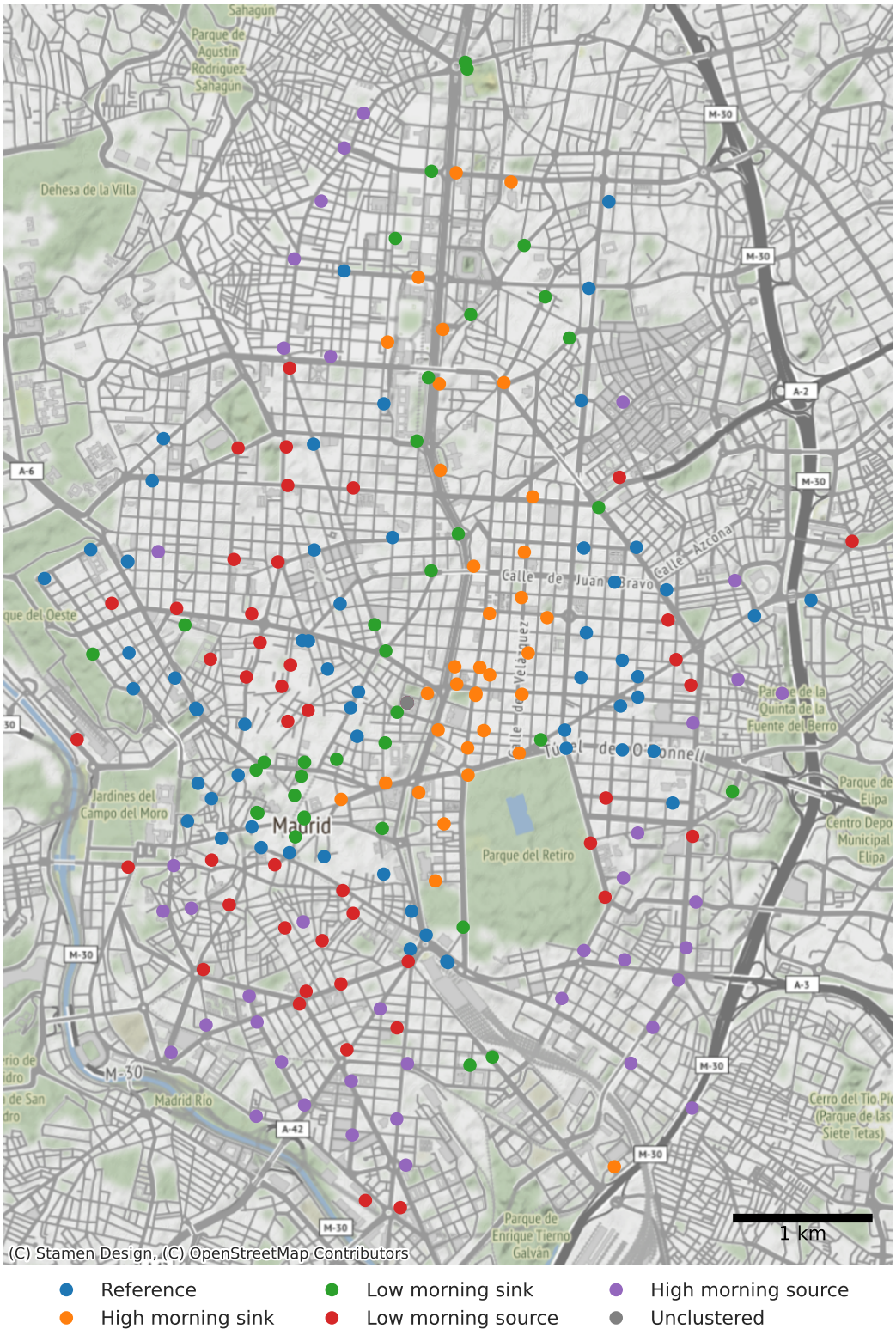


Figure A.8: Results of clustering for Madrid.

B. Confusion Matrices For All Cities

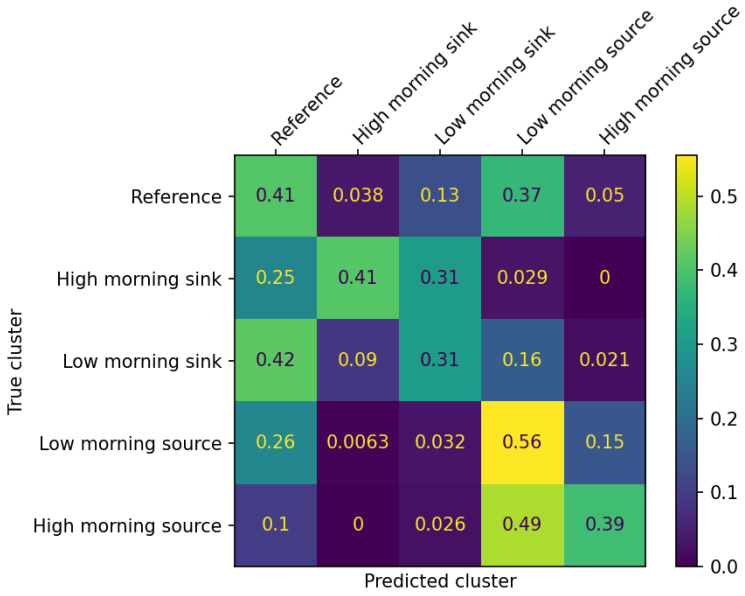


Figure B.1: Confusion matrix of LR model trained and tested on New York City. The values are normalised with respect to the true labels.

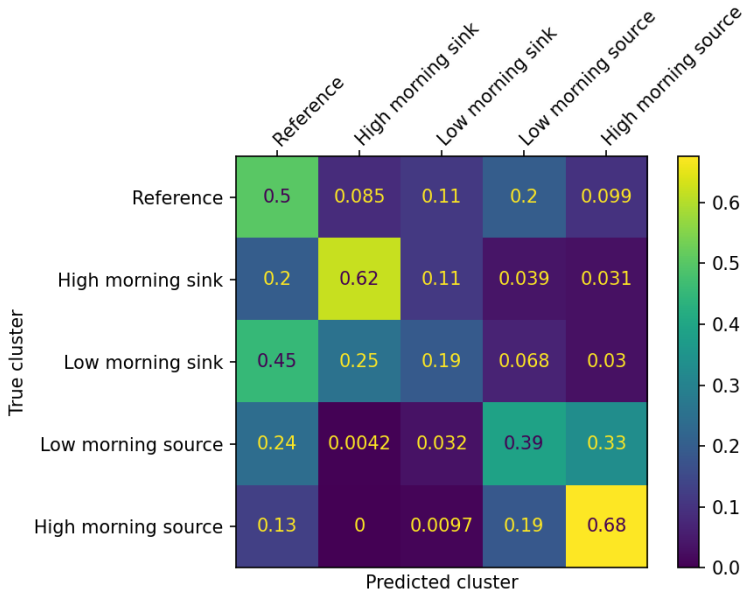


Figure B.3: Confusion matrix of LR model trained and tested on Washington DC. The values are normalised with respect to the true labels.

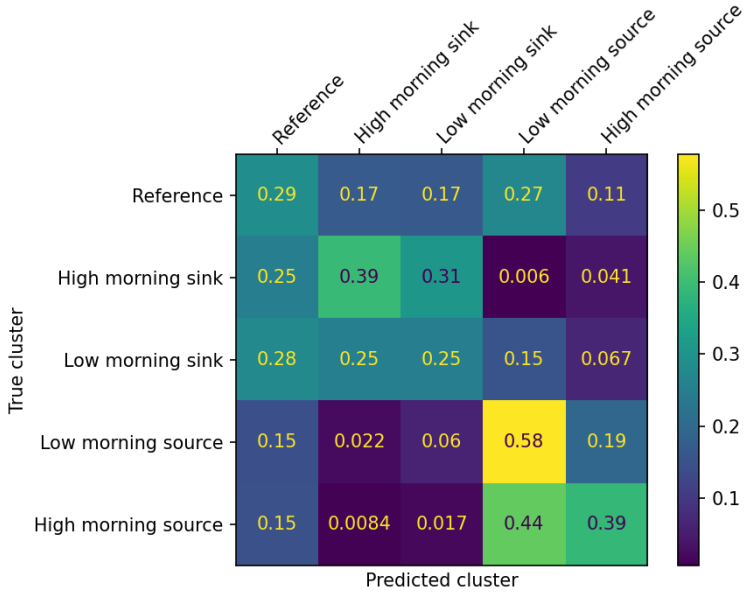


Figure B.2: Confusion matrix of LR model trained and tested on Chicago. The values are normalised with respect to the true labels.

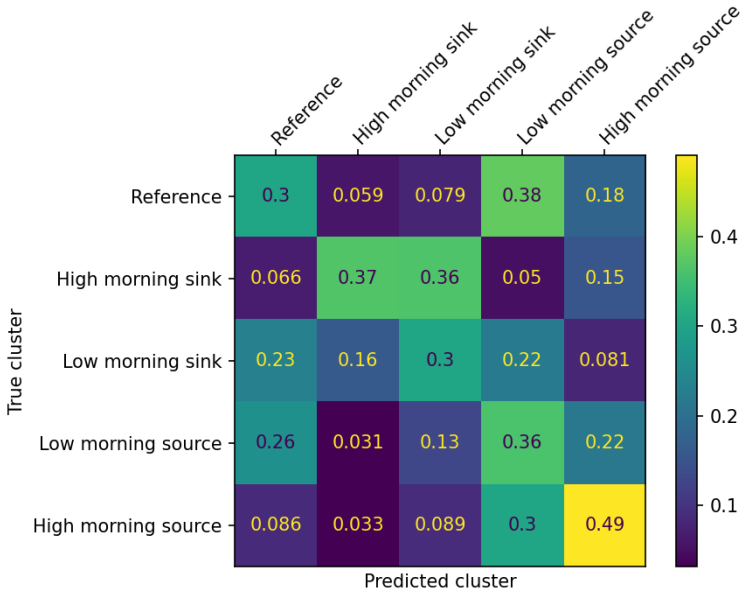


Figure B.4: Confusion matrix of LR model trained and tested on Boston. The values are normalised with respect to the true labels.

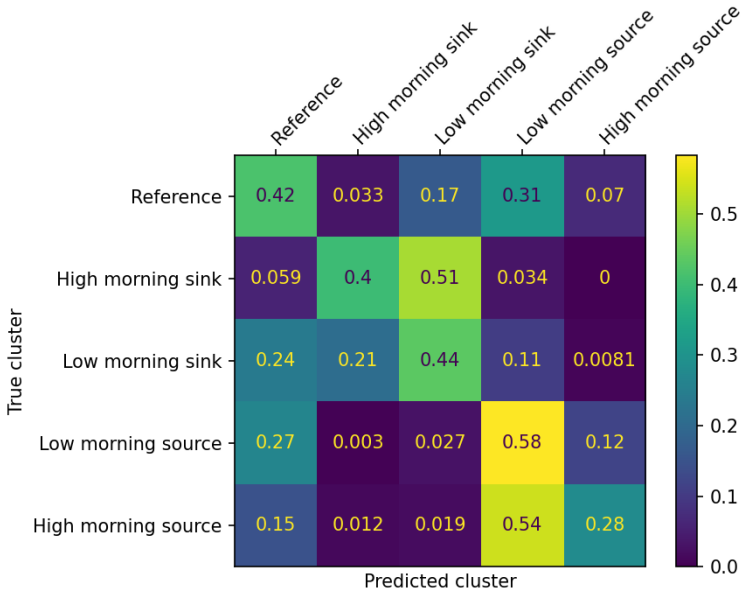


Figure B.5: Confusion matrix of LR model trained and tested on London. The values are normalised with respect to the true labels.

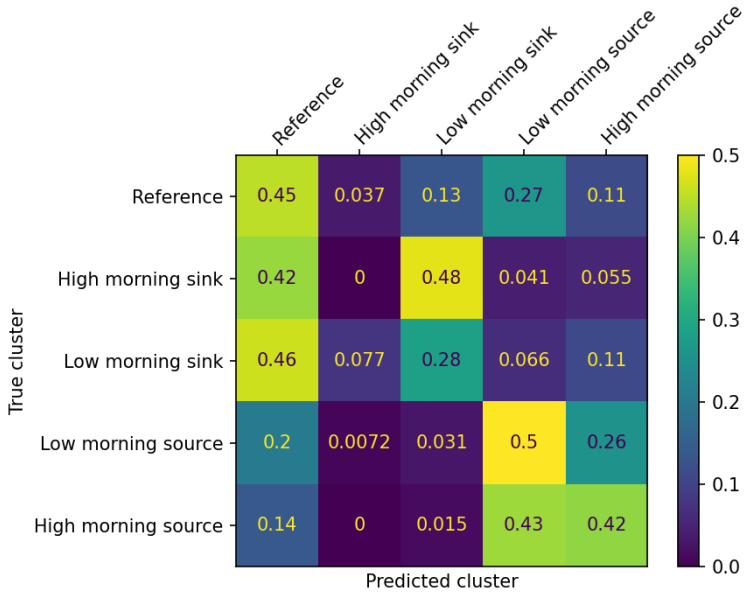


Figure B.6: Confusion matrix of LR model trained and tested on Helsinki. The values are normalised with respect to the true labels.

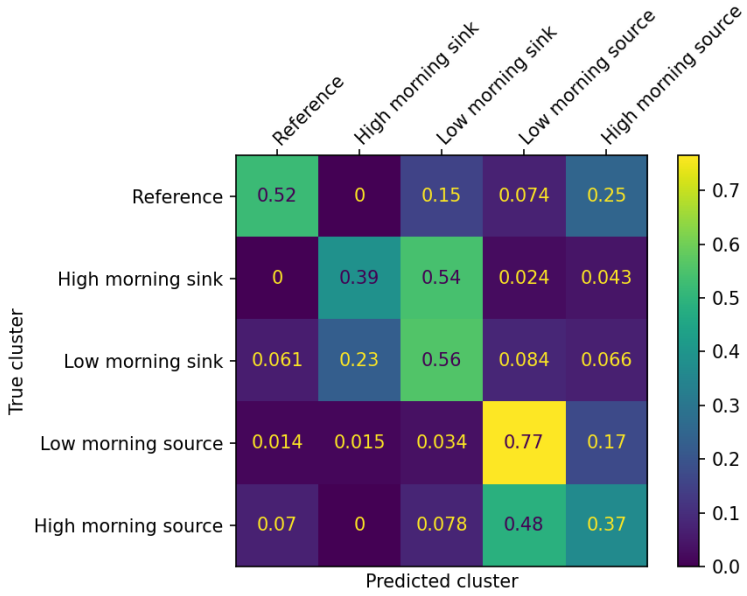


Figure B.7: Confusion matrix of LR model trained and tested on Oslo. The values are normalised with respect to the true labels.

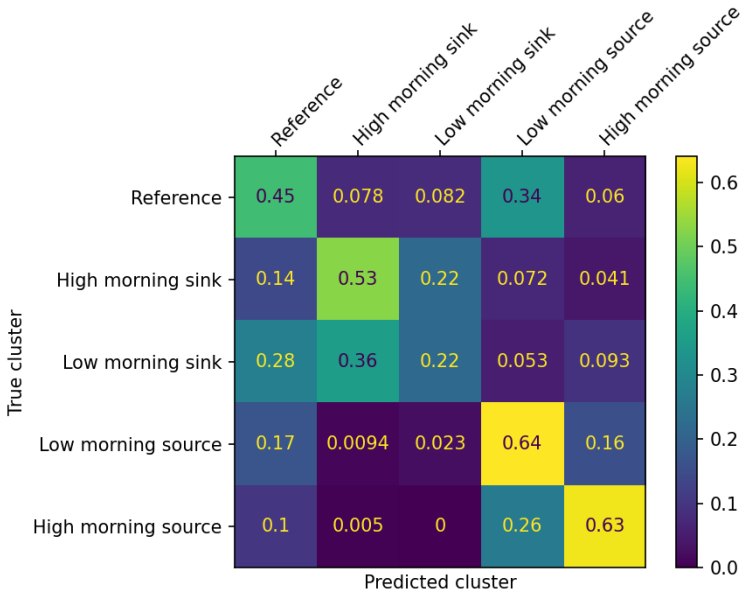


Figure B.8: Confusion matrix of LR model trained and tested on Madrid. The values are normalised with respect to the true labels.

C. Logistic Regression Heatmaps

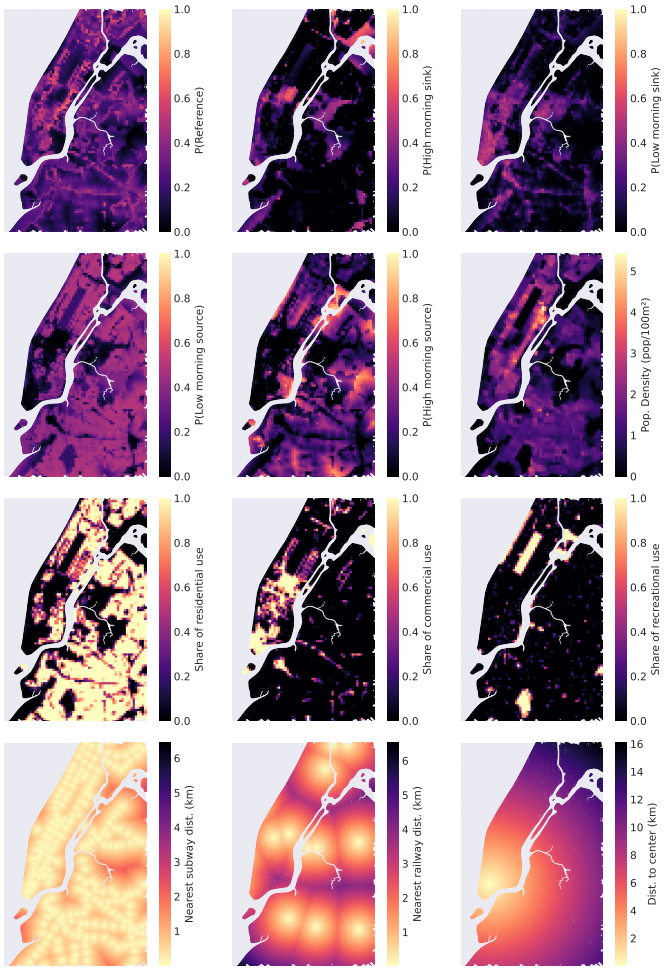


Figure C.1: Heat maps of probabilities of belonging to different clusters alongside external variables for New York City.

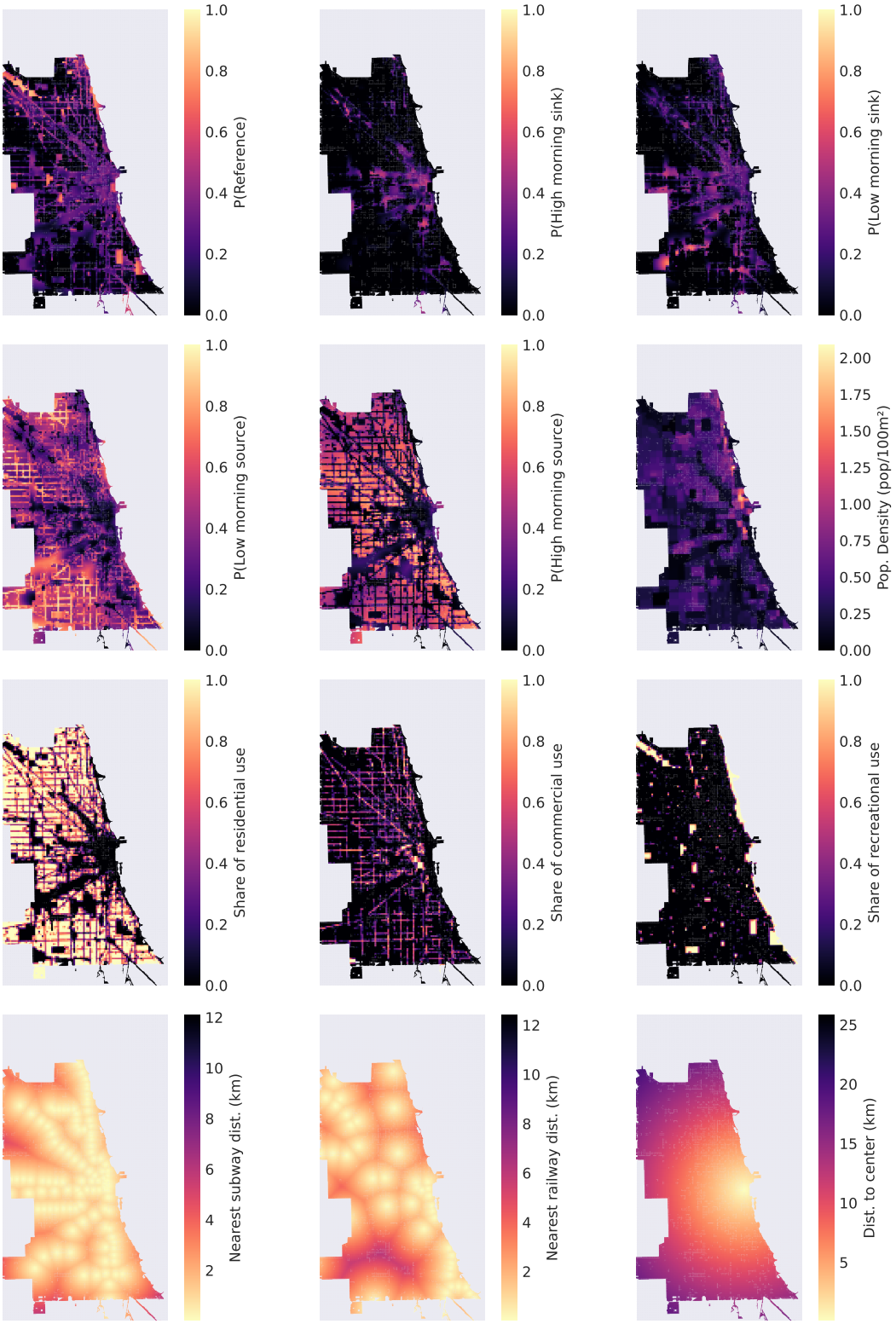


Figure C.2: Heat maps of probabilities of belonging to different clusters alongside external variables for Chicago.

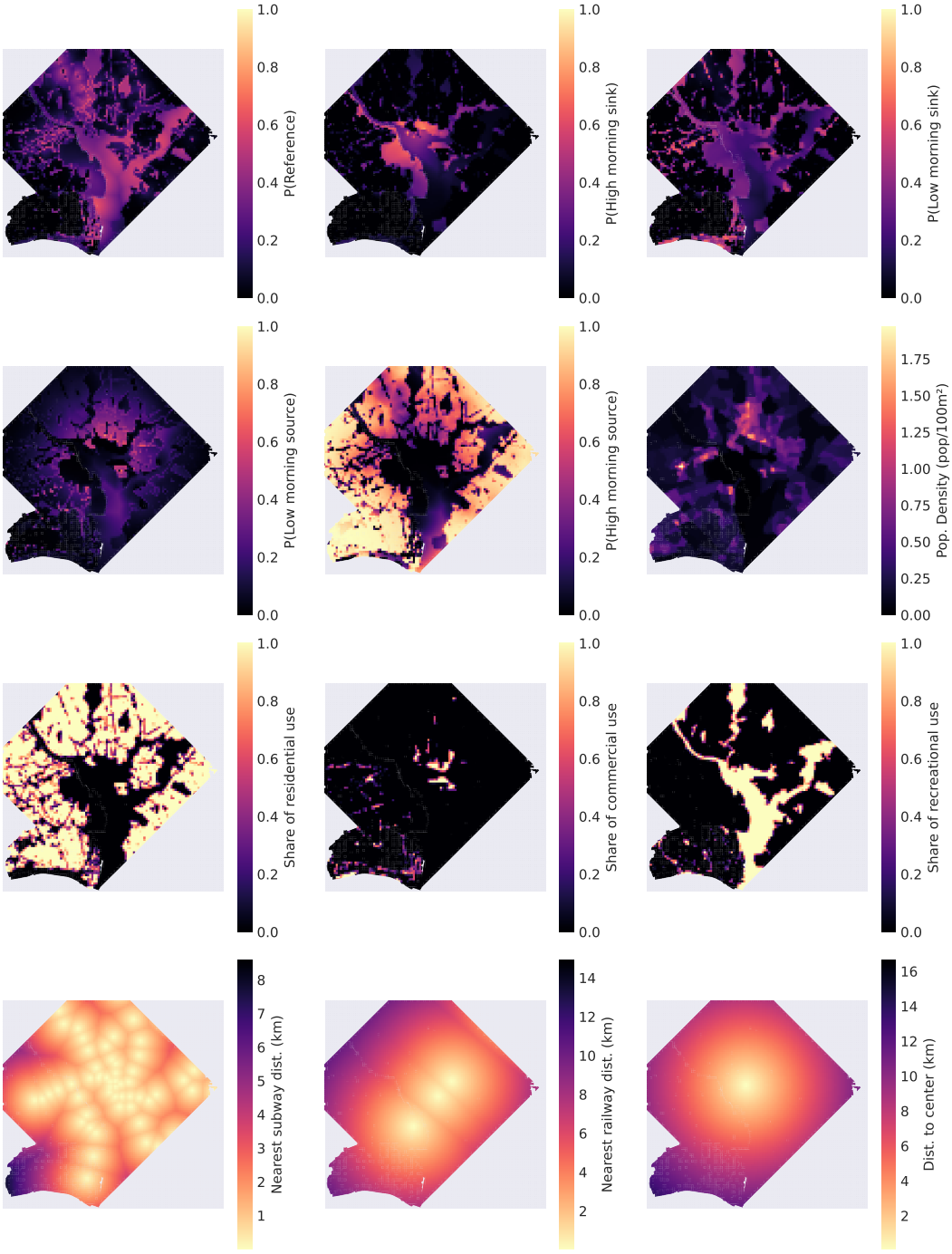


Figure C.3: Heat maps of probabilities of belonging to different clusters alongside external variables for Washington DC.

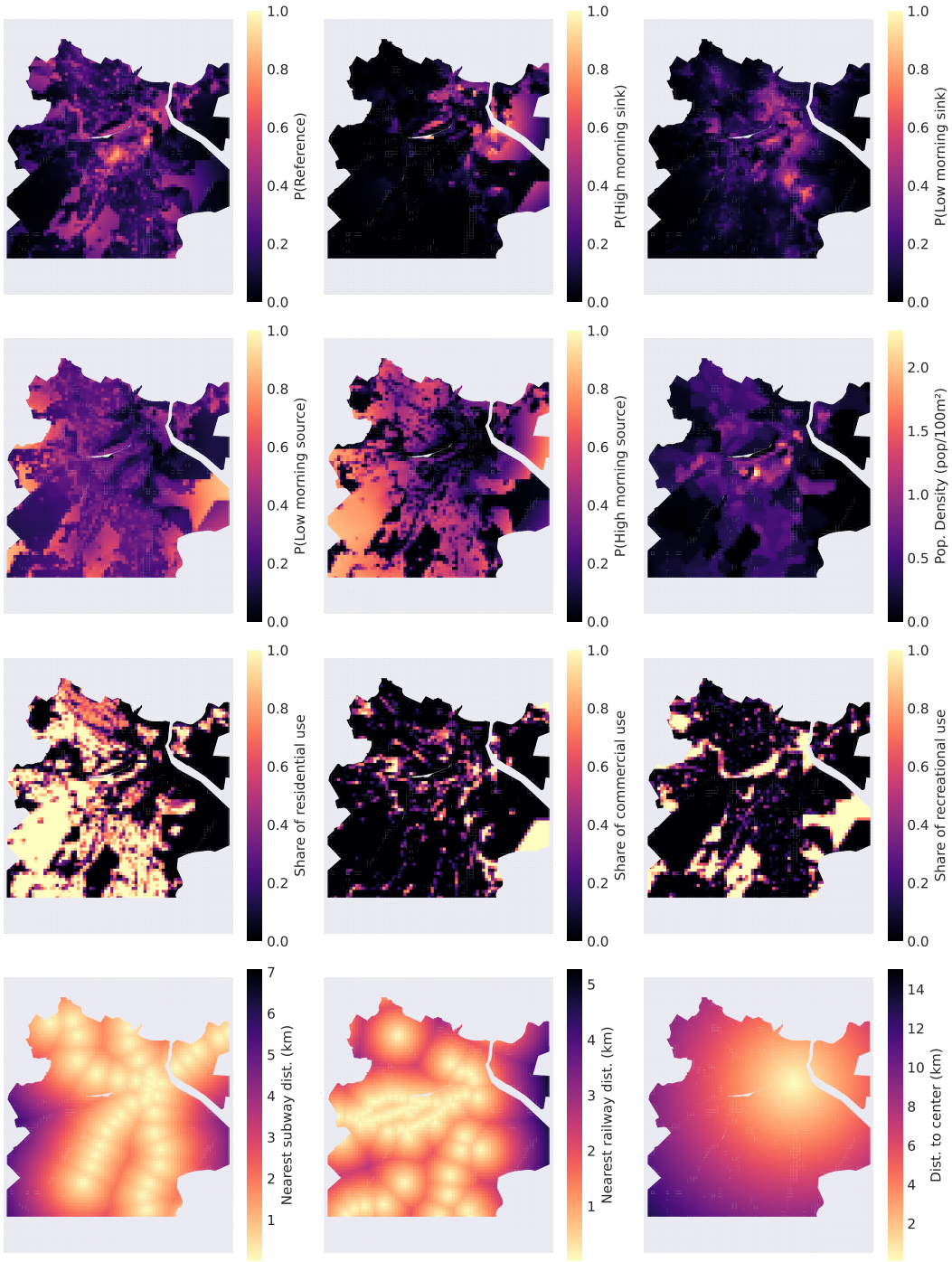


Figure C.4: Heat maps of probabilities of belonging to different clusters alongside external variables for Boston.

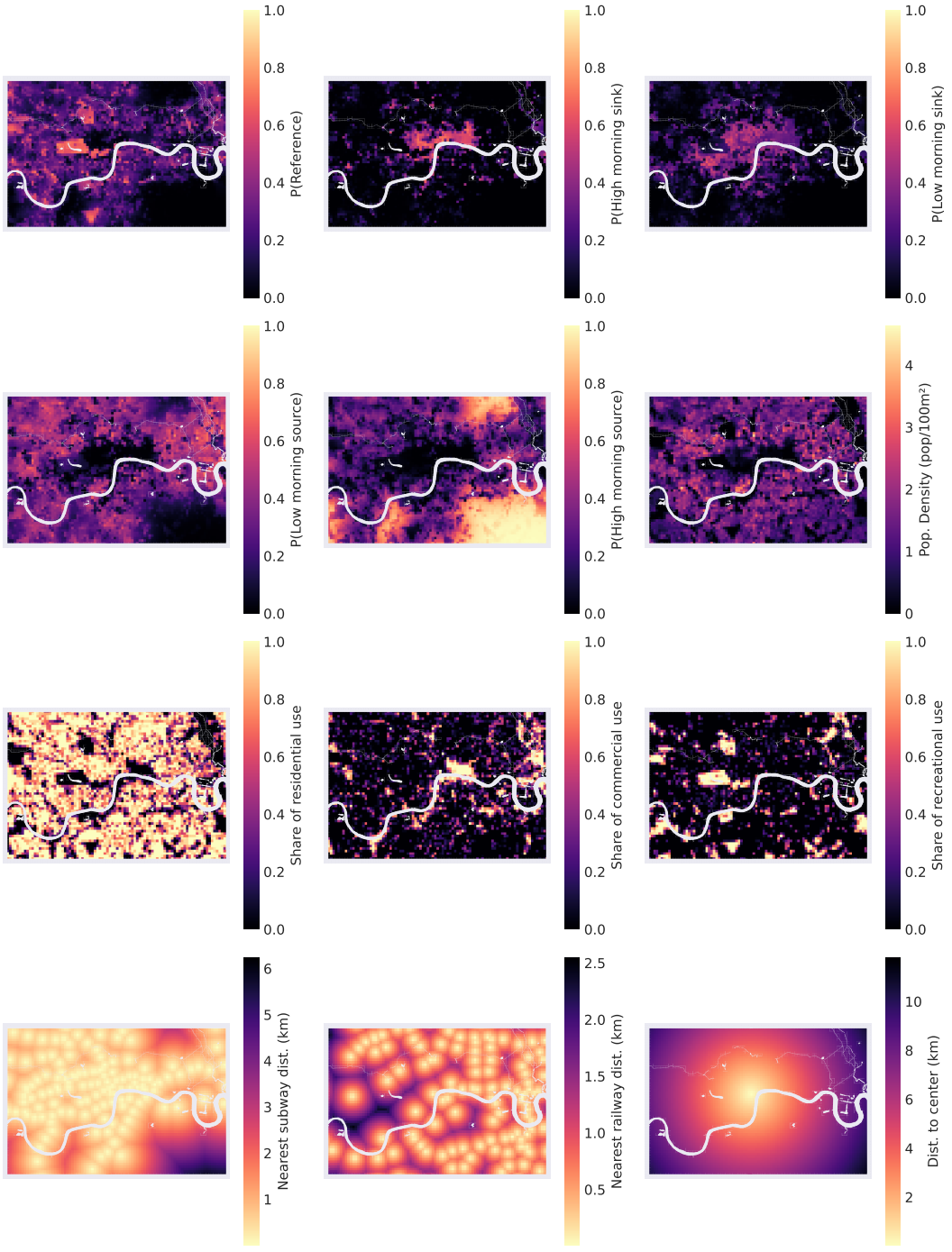


Figure C.5: Heat maps of probabilities of belonging to different clusters alongside external variables for London.

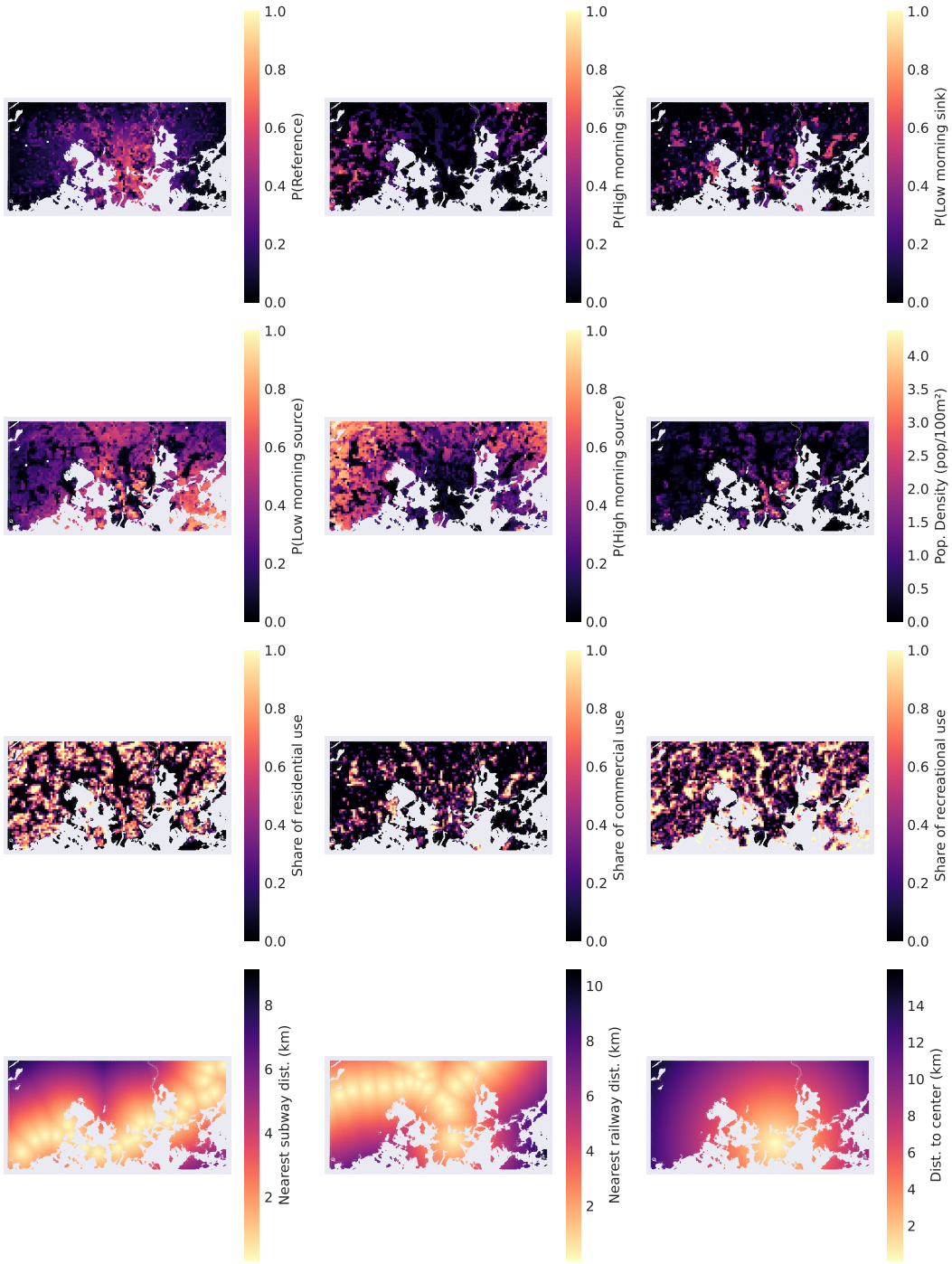


Figure C.6: Heat maps of probabilities of belonging to different clusters alongside external variables for Helsinki.

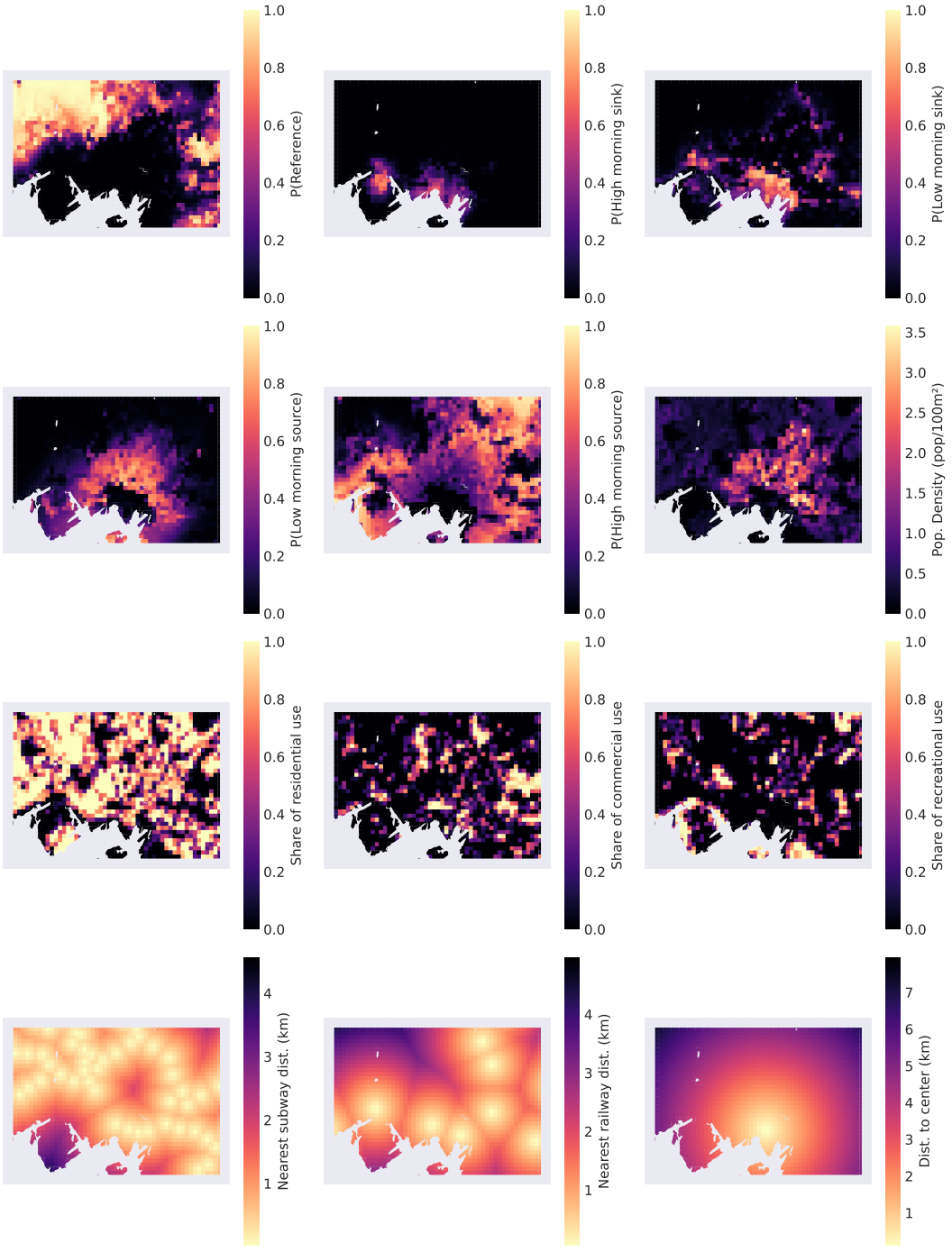


Figure C.7: Heat maps of probabilities of belonging to different clusters alongside external variables for Oslo.

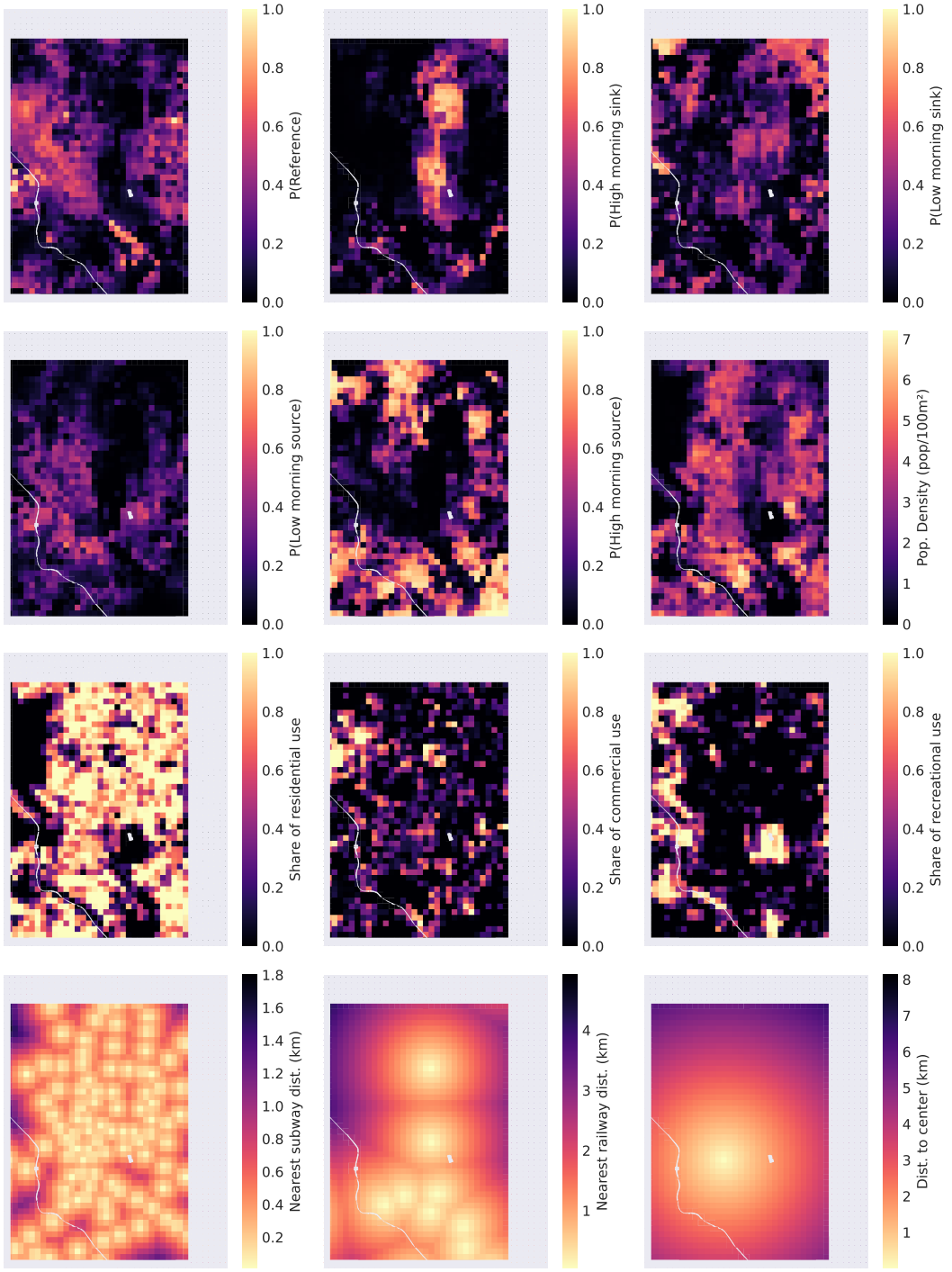


Figure C.8: Heat maps of probabilities of belonging to different clusters alongside external variables for Madrid.

D. Demand Model Heatmaps

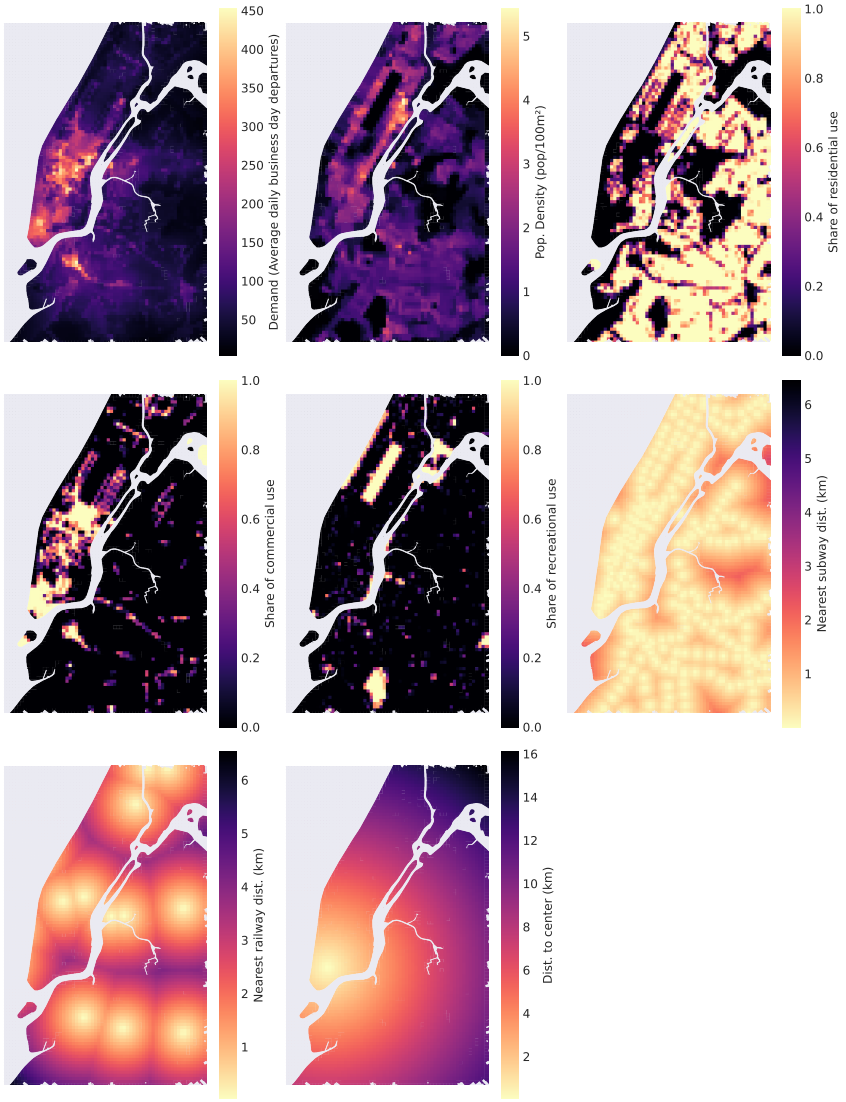


Figure D.1: Heat map of predicted demand and external variables for New York City.

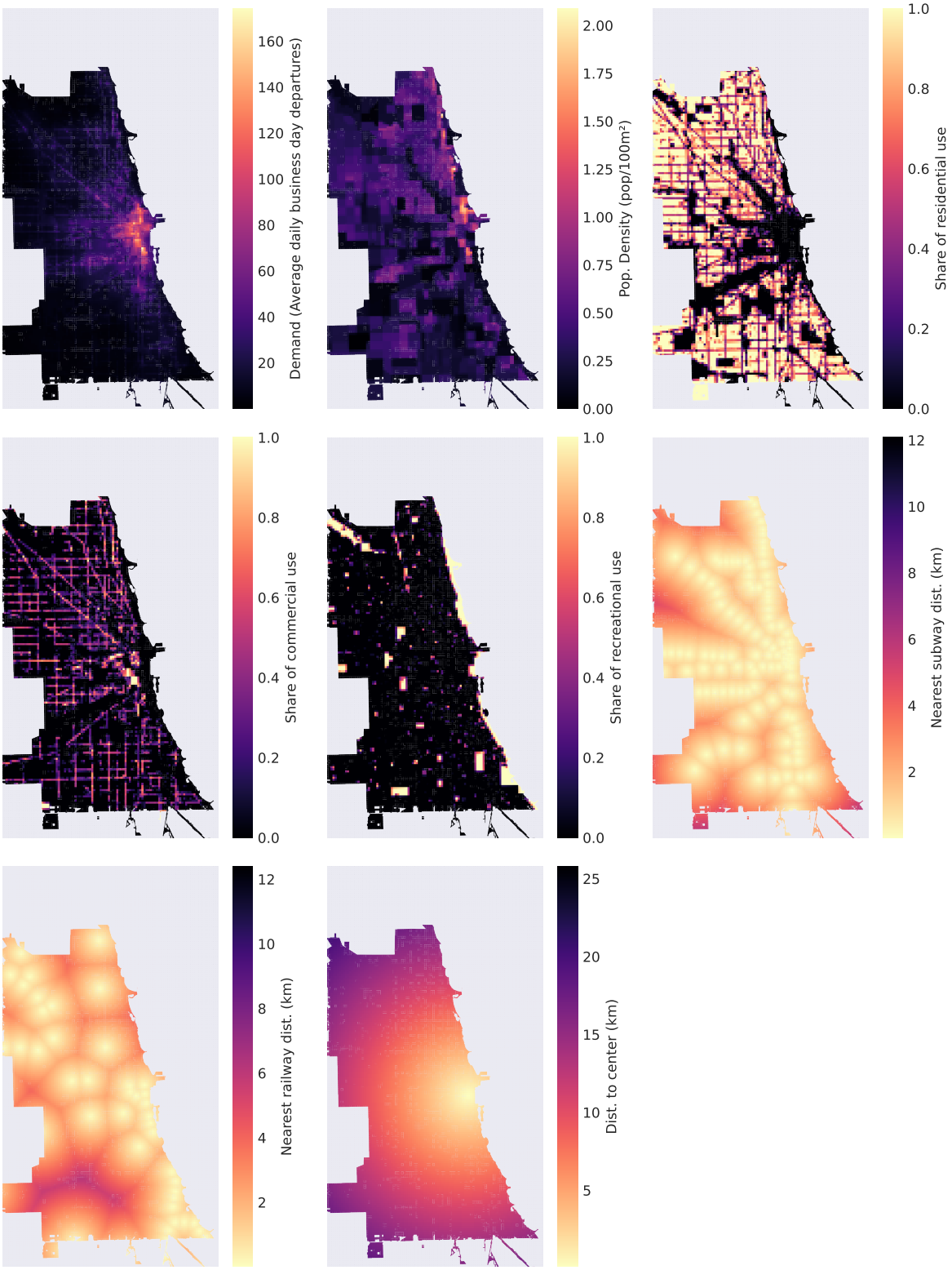


Figure D.2: Heat map of predicted demand and external variables for Chicago.

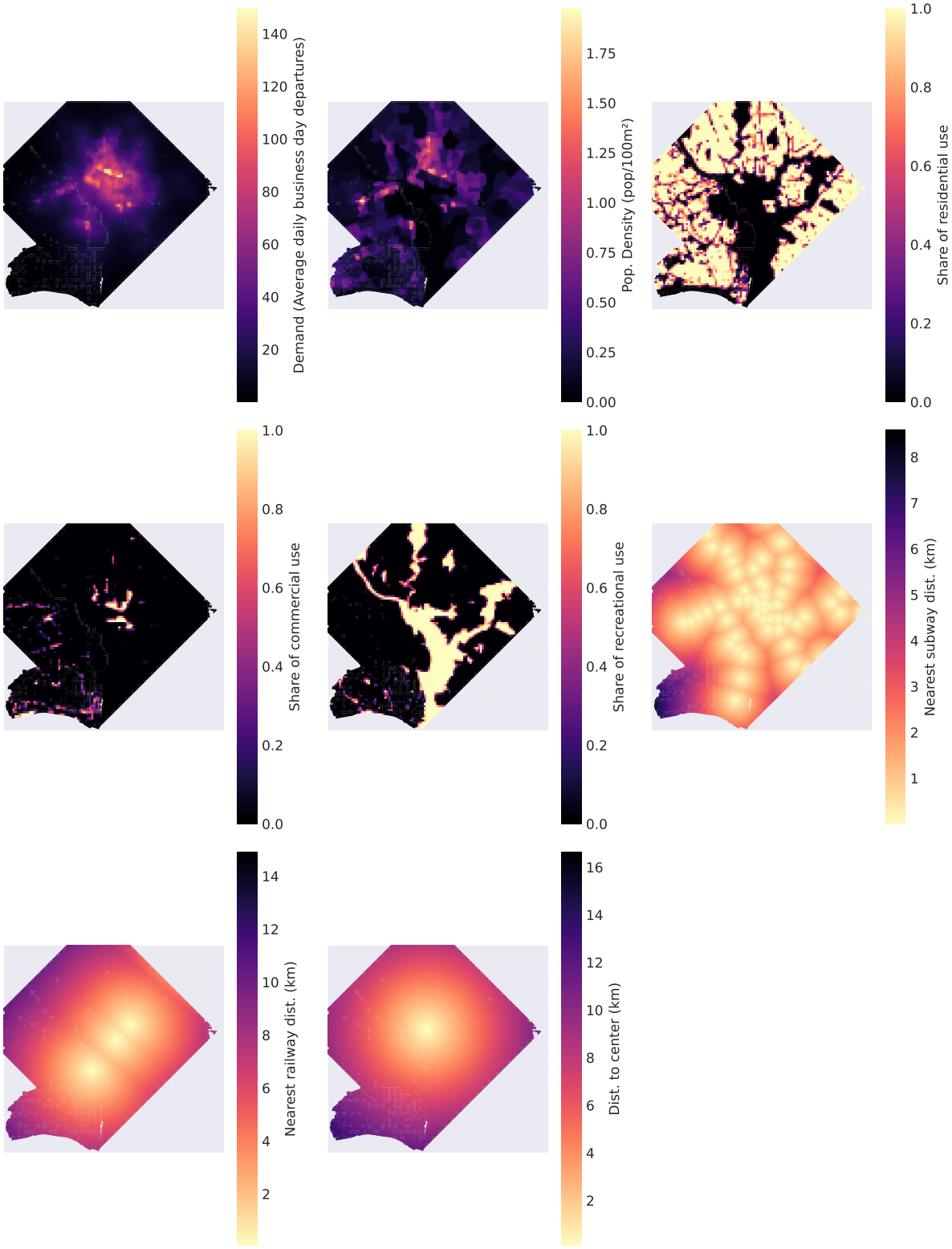


Figure D.3: Heat map of predicted demand and external variables for Washington DC.

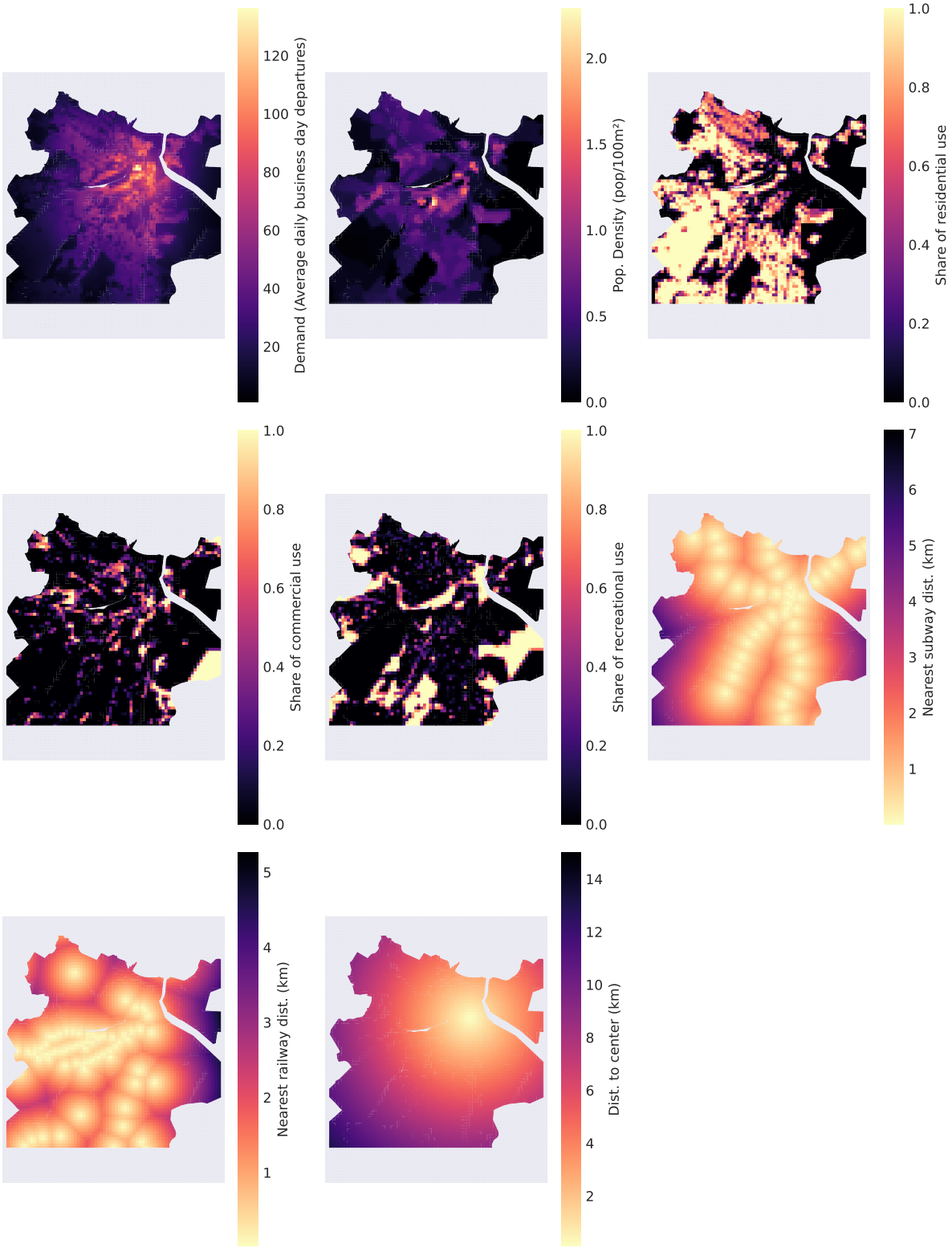


Figure D.4: Heat map of predicted demand and external variables for Boston.

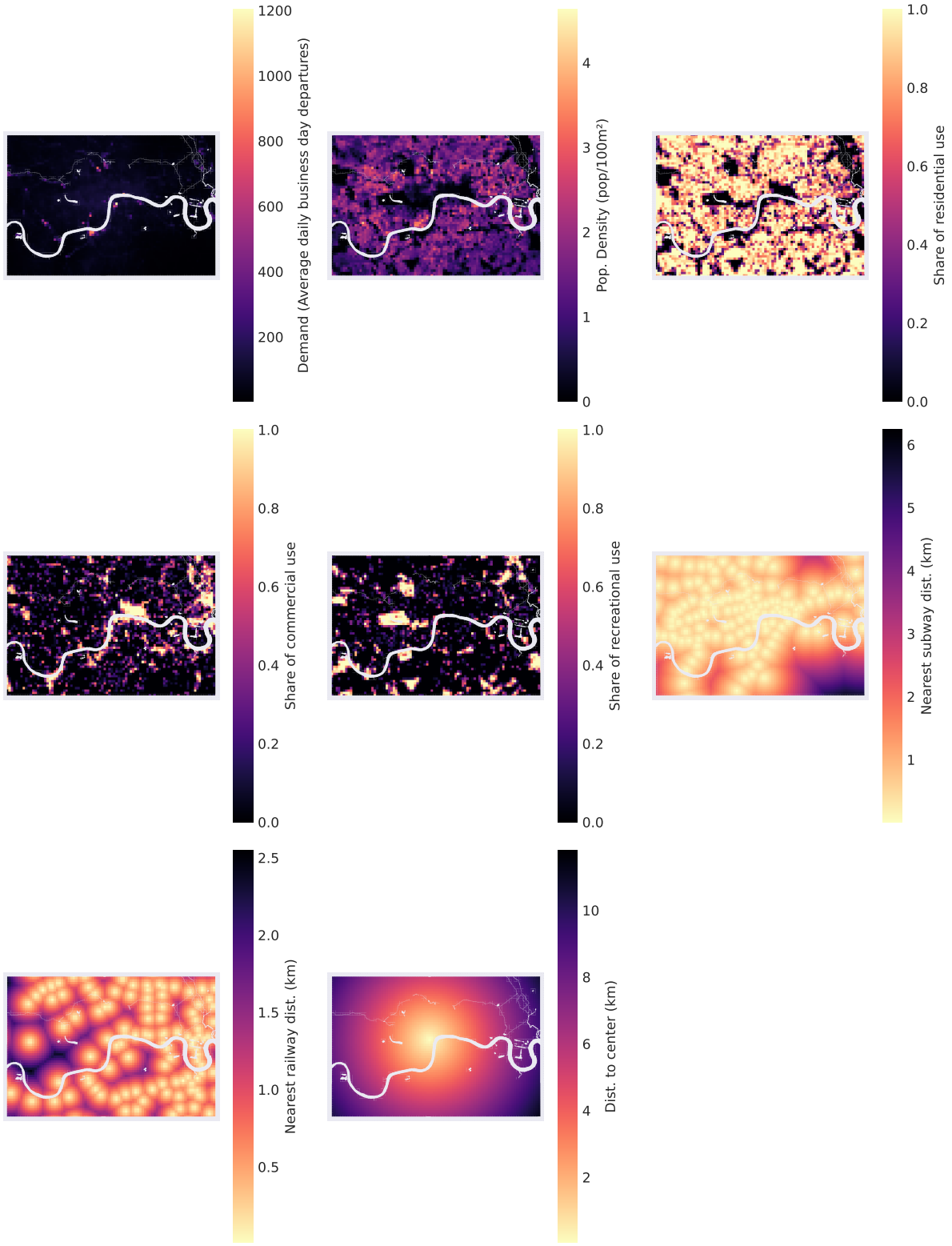


Figure D.5: Heat map of predicted demand and external variables for London.

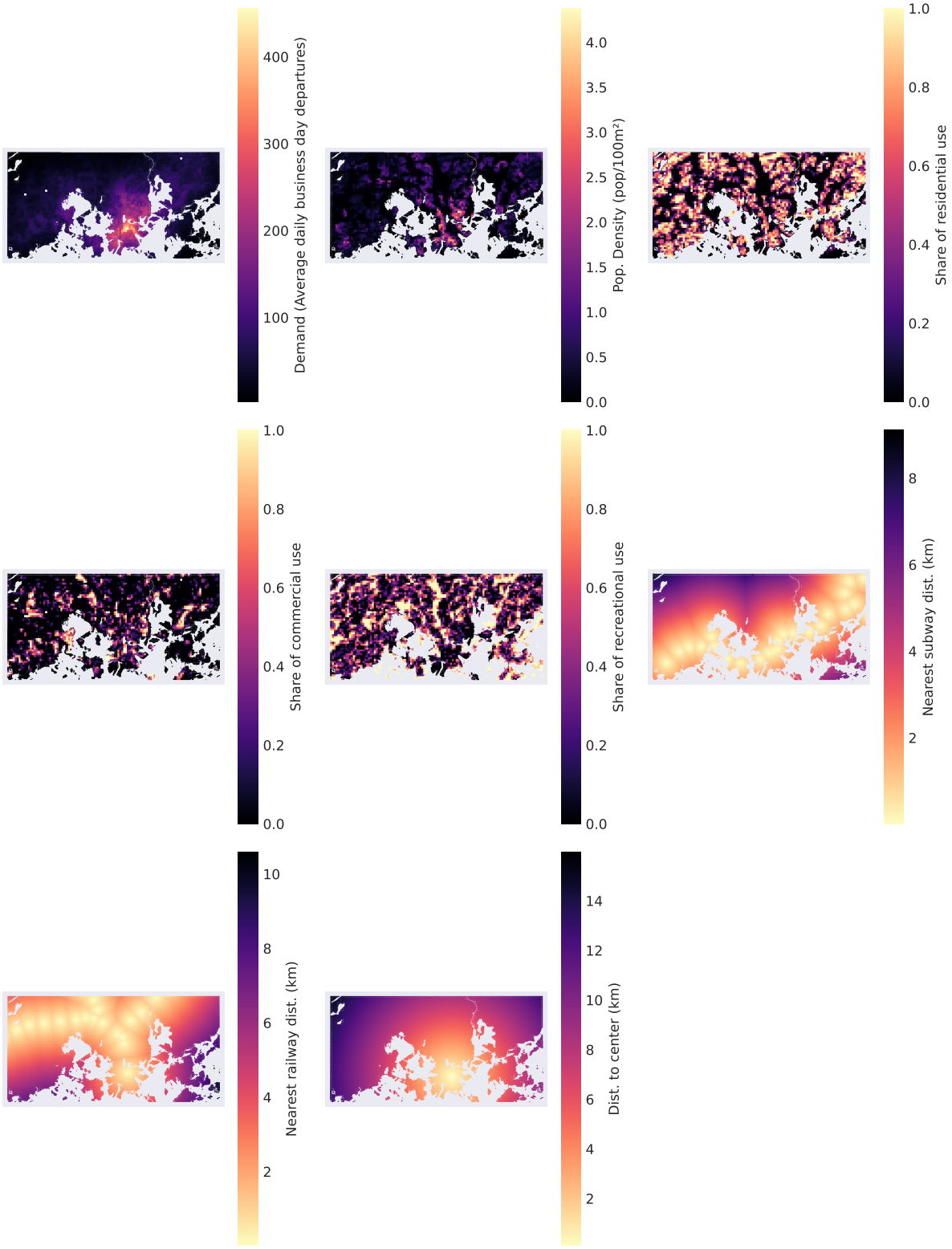


Figure D.6: Heat map of predicted demand and external variables for Helsinki.

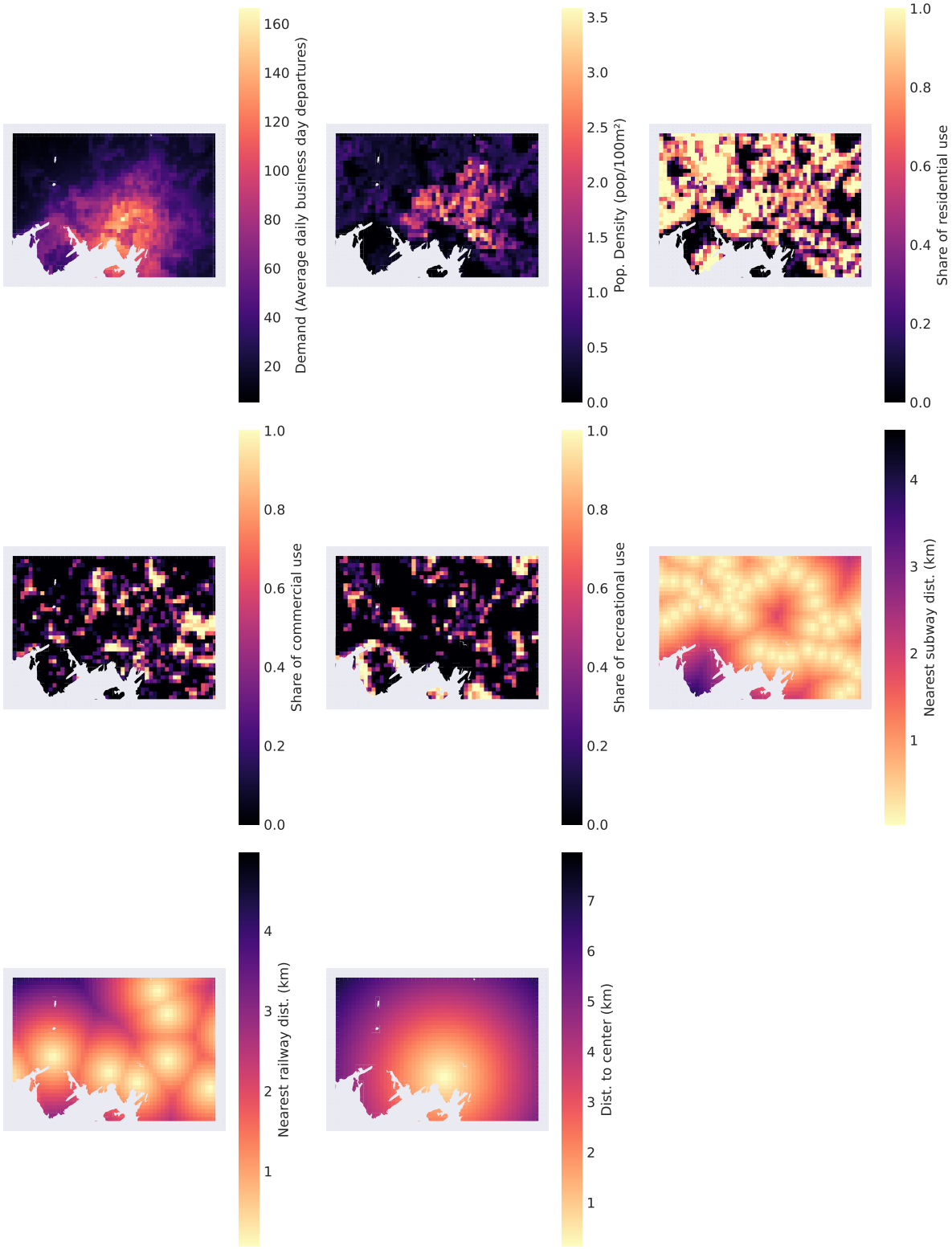


Figure D.7: Heat map of predicted demand and external variables for Oslo.

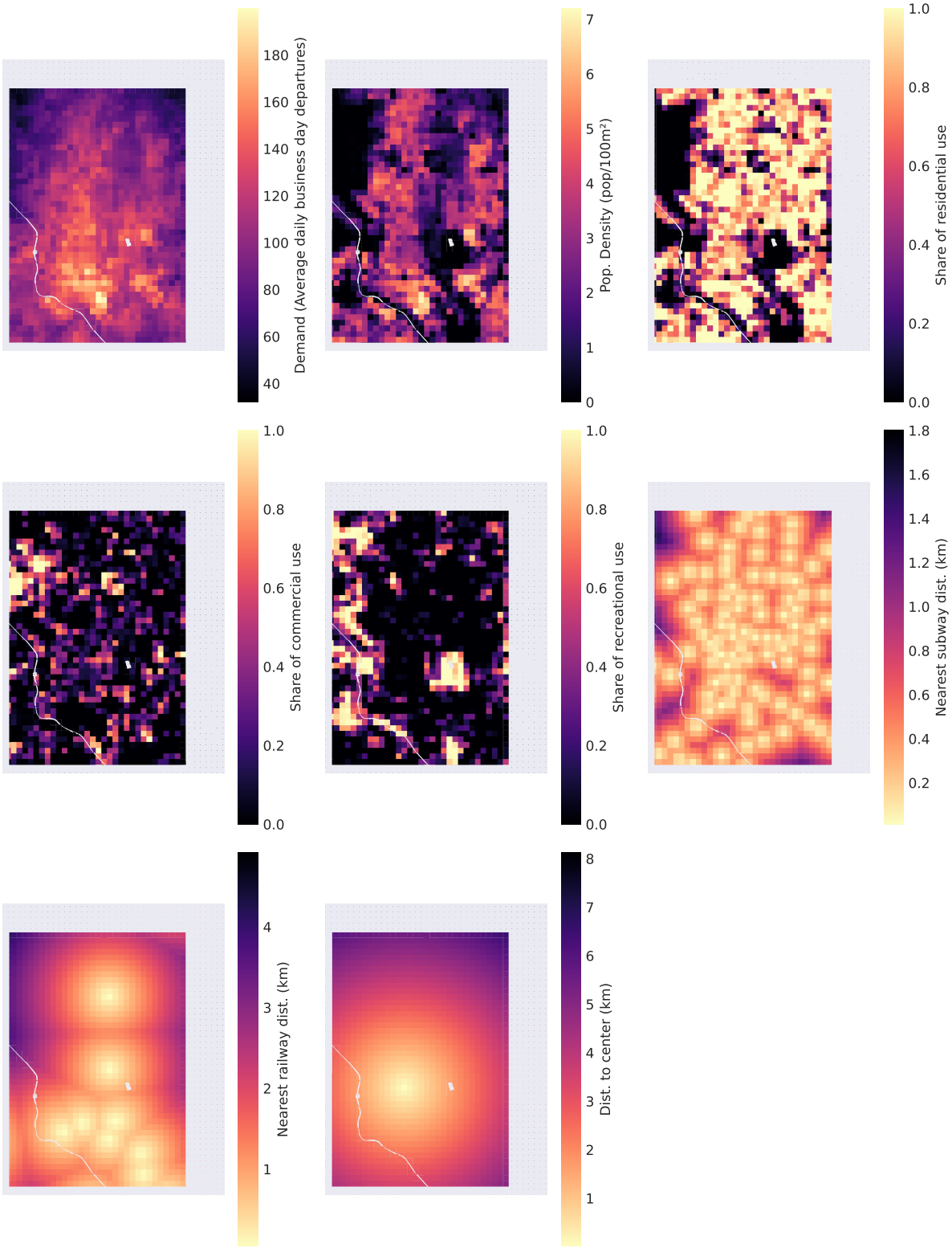


Figure D.8: Heat map of predicted demand and external variables for Madrid.

Bike Sharing Traffic Pattern Prediction from Urban Features and Automated Station Planning

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

[a]*Department of Mathematical Sciences, Aalborg University, Denmark*
[b]*Department of Electronic Systems, Aalborg University, Denmark*

Abstract

This is an abstract.

Keywords: Bicycles, Bikes, Bikers
2010 MSC: 00-01, 99-00

1. Introduction

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 the prediction of demand patterns in bike sharing systems and its relation to the built environment and to other variables such as transit stations. We will therefore divide existing studies in three categories, based on their main methodology: spatial and temporal clustering of stations and communities, short- and long-term traffic prediction, and analyses of the effect of the built environment on bike sharing usage. For a more complete review of the large body of work on bike sharing system, we refer the reader to Eren and Uz (2020); Albuquerque et al. (2021).

*Corresponding author
Email address: dvandi17@student.aau.dk (Daniel B. van Diepen)

2.1. Clustering

Clustering is an operation that divides 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 that have a particularly 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. (2020c). 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.

Spatial patterns can also be used to determine the impact of rare or anomalous events: one example is the disruption in urban mobility caused by public transit strikes, whose effect on bike sharing demand and geographic patterns was shown to be significant by Yang et al. (2022). These analyses of anomalous events can also help shed light on normal usage by contrast. During the Covid-19 pandemic, a spatial analysis by Pase et al. (2020) of the New York City bike sharing system showed that users favored longer trips, often between distant neighborhoods, with respect to the pre-pandemic baseline: this is consistent with the reported reluctance to use mass transit services in 2020, as they were seen as high-risk environment for potential contagion.

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. An analysis of the fraction of subscriber traffic, as compared to daily pass users (i.e., most likely tourists or temporary visitors, while subscribers are often daily commuters) and an analysis of the imbalance in traffic patterns between the mornings and evenings allowed the authors to determine the set of stations most likely used by commuters, relating the results to land use type and the directions of travel. The presence of well-connected and protected bike lanes is also noted as a potential factor

increasing long-distance trips, as users are encouraged to cycle if the route is safe and quick.

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.

2.2. Traffic prediction

Predicting traffic patterns is crucial in bike sharing systems, as the short-term trends are crucial for effective rebalancing, while longer-term trends can be used for proper planning, and this topic has been the subject of intense study in the past decade. The spatial and temporal aspects that affect future traffic are the same that need to be considered for the clustering, and several works exploit different combinations of input parameters, along with historical and immediate trends, to perform the prediction.

The two main approaches to demand prediction are to either model the traffic as a stochastic process, using statistical knowledge to determine the most likely future behavior, or to use a purely learning-based approach, trading the explainability of the model for the generalization and pattern-matching capabilities of deep neural networks.

We first consider the model-based approach: the first simple applications of models to traffic prediction consider individual stations Yang et al. (2016), creating simple heuristics to predict future demand based on historical behavior Sathishkumar et al. (2020). More complex models consider clusters of stations based on geographical and past trip information Li et al. (2015), allowing a coarser-grained prediction on the cluster level and a finer-grained one for individual stations Li and Zheng (2019). The correlation between close-by clusters is another significant piece of information, which can help predict spikes in demand Chen et al. (2016), and graph information in general can be a powerful tool to predict future behavior Yang et al. (2020c,a). The stochastic nature of traffic demand can also be taken into account, using Markov modeling and birth-death process theory to include uncertainty in the model Zhou et al. (2018). This type of models allows for a more precise estimation of future risk, which is particularly important when planning and managing the system, as average behavior might not be enough to provide full service availability even in worst-case scenarios of high and unbalanced demand Hulot et al. (2018). In order to further improve worst-case performance, risk and extreme value theory can be used effectively Sohrabi et al. (2020).

The use of Long Short-Term Memory (LSTM) networks has proven to be effective to capture temporal relations, both on the individual station level and for the network as a whole Wang and Kim (2018), and these networks can accommodate information such as public transit stops and schedules Zhang et al.

(2018), which can be extremely important to gauge the last-mile effectiveness of the service. The combination of recurrent and convolutional networks can exploit both spatial and temporal information at the same time, leading to a more complete learning model Xiao et al. (2021) which can perform better. The new attention-based neural network architecture can further improve prediction performance by considering data at different timescales, which is often problematic for standard LSTM. Learning models can also benefit from more complex representation, which can either rely on graph models Yang et al. (2020b) or be learned directly as a spatio-temporal graph along with external information such as the hour and weather Li et al. (2019).

2.3. Effect of the urban environment

Noland et al. (2016) estimated Bayesian regression models of trips at stations in order to examine the effects of bicycle infrastructure, population and employment, land use and transit access. They found that their model could show the effects of the factors, but that the models were not well suited for predicting the traffic the next year.

Daddio (2012) used an adjusted regression model on data from Washington, D.C. and found five statistically significant factors affecting the number of trips, including population, density of retail stores, and locations of metro rail stations. They used the model to make a heatmap of the expected number of trips in Washington D.C.

Hyland et al. (2018) is one paper where clustering and modelling are combined, using the cluster membership as a term in a mixed model to increase the accuracy of the model. They focus on modelling the number of trips to each station using mixed effect models. In order to improve the model, they proposed clustering stations according to the percentage of arrivals in the morning, trips by members, weekend trips and afternoon trips.

3. Data and Methods

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 following sections.

3.1. Bike Sharing Trip Data

The bike sharing trip data is obtained directly from the websites of the individual bike sharing providers or from city data portals. In this paper, we use datasets from 2019, as this is the most recent year with normal operation prior to the COVID-19 pandemic. 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

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.

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).

3.2. Station Service Area Determination

In order to match each station in the network to the land use features of its catchment area, we assign a designated service area to each station in the network. These service areas are determined using a Voronoi tessellation, which defines boundaries that assign each point in the city to the closest station. The areas are then truncated, so that there is a maximum Euclidean distance of 500 meters from the station to the furthest point in its service area. The 500 m limit is a conservative approach, which assumes that users are not willing to walk any further to reach a bike sharing station, and 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). This simplification is further justified if we assume that a user will walk the most direct route to a station without taking detours. The service areas are further truncated such that they do not span over bodies of water such as seas, rivers, and lakes. This is done by using polygons in the land use data described below.

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. An example of a map of the stations in New York City and their service areas for October 23rd can be seen in Fig. 1.

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.

3.3. Land Use, Census, and Transit Data

For U.S. cities, land use is 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 probable 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. This data includes polygons of different land areas, along with a general description of its use. These areas are then classified in the same way the US cities. 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. Since the zoning data in the US does not separate roads from

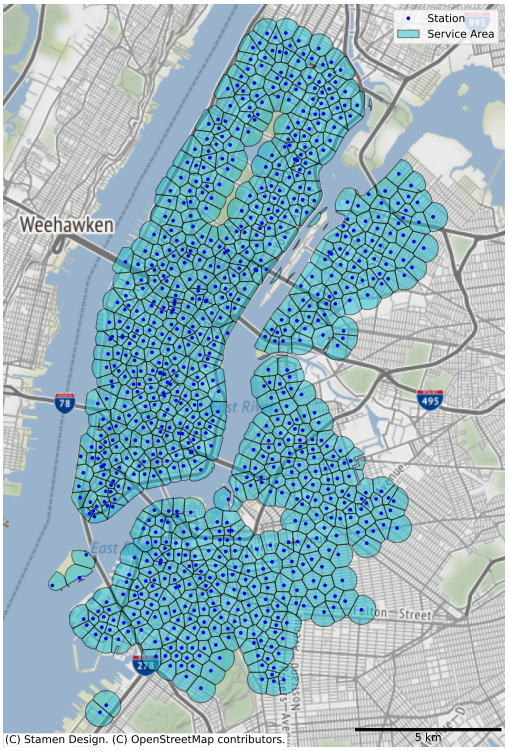


Figure 1: Service areas for New York City on October 23rd 2019.

zoning areas, this leads to a fairer comparison and a more uniform model of the zoning.

Historical census data for U.S. cities in 2019 is provided by the United States Census Bureau on the census tract level, along 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 European cities, population counts are provided for each polygon in the land use data from the 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, transit data was obtained using the Overpass API from OpenStreetMap. The data contains locations of subway and railway stations.

3.4. Data Processing

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

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

intense, with a correspondingly stronger impact on planning and management considerations, we only consider 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. Furthermore, we removed all trips that do not start on business days from the dataset, excluding both weekends and public holidays. Furthermore, we excluded two more kinds of trips: loop trips, i.e., trips that have the same departure and arrival point, which are often recreational Zhao et al. (2015), and trips taken by temporary users (in cities which have this distinction in the dataset), who are most likely tourists visting the city for a short period Noland et al. (2019). Finally, trips shorter than 60 seconds are 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. If a station has only 1 or 2 trips per day, individual users can have a significant effect on traffic patterns. This can bias the analysis, introducing outliers with limited value to the overall system; therefore, we remove any stations with 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. 2.

3.5. Modeling Approach

Our modeling approach follows other spatio-temporal clustering works, dividing bike sharing stations into groups based on their daily arrivals and departures. However, we perform a further step and attempt to connect traffic patterns with other features of the urban environment, building a predictive model that can be generalized to other neighborhoods and cities. An overview of our modelling approach can be seen in Fig. 2. 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. Finally, logistic regression is 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.

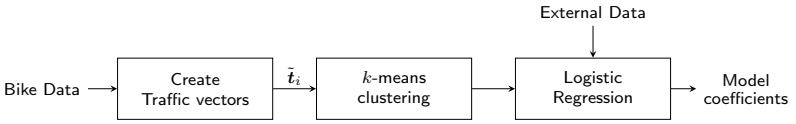


Figure 2: Flowchart of the modeling approach.

Using the bike sharing trip data, we calculate 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 start of the hour to the end of the hour, e.g., the arrivals and departures are counted from 16:00:00 to 16:59:59. Let $\mathcal{M}_i$ be the set of days in which station i has been used. We then define the two 24-element vectors $\mathbf{d}_{m,i}$ and $\mathbf{a}_{m,i}$, representing the departures and arrivals from and to station i in each hour of day m, respectively. In order to mitigate the effect of the concentration of trips in the rush hours on the traffic pattern, we consider the flow to the station, defined as the difference between the number of arrivals and departures:

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

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

$$\tilde{\mathbf{f}}_i = \frac{\sum_{m \in M_i} \mathbf{f}_{m,i}}{\sum_{m \in M_i} \sum_{h=0}^{23} \mathbf{a}_{m,i}(h) + \mathbf{d}_{m,i}(h)}. \tag{2}$$

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.

We then used the classical k-means algorithm to divide the stations into classes based on their traffic patterns. The algorithm is a partitional algorithm which divides the data points into k clusters, minimizing the distance between data points in a shared cluster while maximizing the distance between data points in different clusters. To initialize the algorithm, the number of clusters and an appropriate distance measure have to be chosen beforehand. 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 definition of the number of clusters k is more complex, as several clustering measures exist. In this paper, four of such measures have been used. The Davies-Bouldin, Dunn, and Silhouette indices, which were first presented in Davies and Bouldin (1979), Dunn (1973) and Rousseeuw (1987), respectively, all measure the cohesion and separation of clusters based on distances between data points and between clusters. These metrics are calculated for different choices of k. Clustering is better if it leads to high values of the Dunn and Silhouette indices and low values

of the Davies-Bouldin index. The fourth clustering measure used is the Sum of Squared Errors (SSE) between data points and their closest cluster centres, which is also the error minimized by the k-means algorithm.

The labels obtained from the clustering are used as dependent variables in a multinomial Logistic Regression (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. As independent variables, we use the urban and transit data for each station described in Sec. 3.3. Since the service areas of the stations can vary over time, so can these variables. Thus, each variable is averaged over the days in which the station was in use. The summary statistics on the independent variables on the service area level are listed in Table 3.

4. Results

4.1. Clustering

The number of clusters has been determined by observing the clustering measures and the resulting cluster centres for different choices of k while also considering the future use of the clustering in the model and in the comparison between cities. For instance, when relating the clustering to other geographical features and comparing between cities, it is preferable to have the same number and types of clusters for each city in order to ensure that the differences in the models are due to the inherent differences in the cities and not due to a different clustering. There is also a trade-off between resolution of the clustering and the ability to compare clusters between cities. A lower amount of clusters will yield a lower resolution of the clustering but the clustering will be similar between cities yielding easy comparisons. Conversely, having a high amount of clusters may result in different types of clusters across the cities due to local factors impacting the clustering.

The clustering measures for different k can be seen in Fig. 3. The Davies-Bouldin and Silhouette index generally prefer fewer clusters. The opposite is true for the sum of squares metric which decreases as k increases. In fact, the sum of squares will always decrease as a function of k and thus an elbow criterion is typically used such that for a chosen k' the decrease in SSE is significantly smaller for $k > k'$ than for $k < k'$. From the sum of squares, we thus find that 3 to 5 clusters seem to be reasonable choices. For most of the cities the Dunn index seems to prefer 5 clusters over 3 or 4 clusters.

With the above considerations, we determined that five clusters provide an adequate clustering for each city that also lends itself to comparisons between cities. The resulting cluster centres are shown in Fig. 4 and the size of each cluster is shown in Table 4. This clustering generally leads to 5 distinct station types: high morning sources which receive a large intake of trips in the morning, low morning sources which does the same albeit to a lesser degree, high morning sinks and low morning sinks which act opposite to high morning sources and low morning sources, and a fifth cluster which will act as a reference. There are

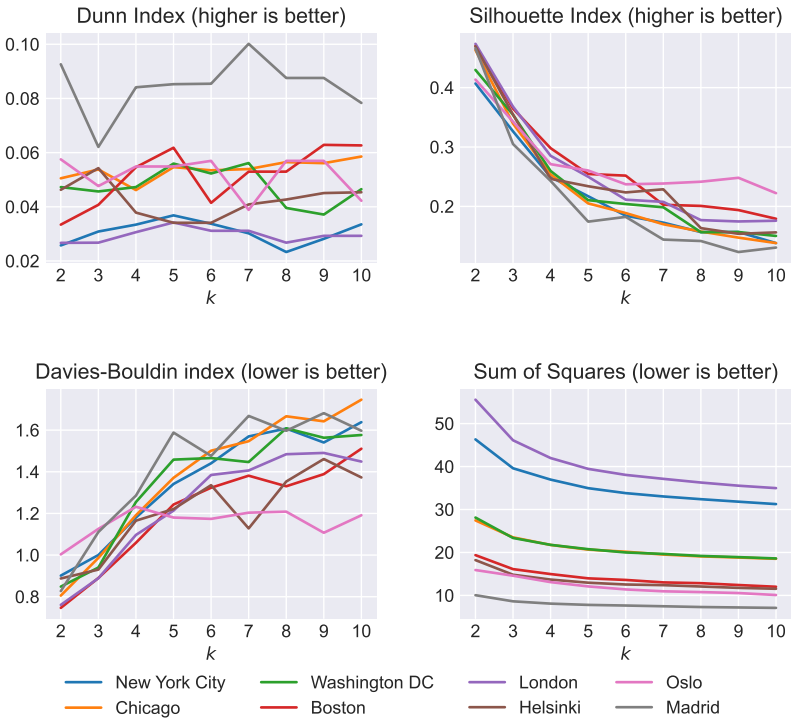


Figure 3: Davies-Bouldin, Dunn, and Silhouette index and SSE for all cities and different choices of k.

still interesting differences between the cities. In almost all cities, the reference cluster has a very low relative difference, meaning that the number of departures and arrivals at a station within the cluster is highly balanced at any given time. However, in Oslo there is a notable absence of this type of cluster. Instead, Oslo has a cluster where departures and arrivals are balanced in the morning while in the afternoon and throughout the evening most of the trips are departures. 23 stations or 9.2% of Oslo's network is comprised of these types of stations and a large majority of these stations are located in or near 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 bikesharing to depart from the university in the afternoon. Whether this is because the students are using the bicycles for recreational purposes or for other reasons, this may show how public transit and the bikesharing system in Oslo complement each other for the use-case of the university students.

Differences also lie in the non-reference clusters across cities. Madrid stands out in this regard, due to having three peaks in the daily traffic: one from 6 to 10, another from 13 to 15 and a third from 17 to 19. This is likely a consequence of the work culture in Madrid where it is customary to take a midday break from work and resume the workday at a later time. In New York City, Chicago, Washington DC, Boston, and London the peaks of the morning sources and the valleys of the morning sinks are mostly aligned at 8:00, meaning that most of the morning commute in these cities is done from 8:00 to 9:00. However, in the afternoon commute there is a noticeable one hour shift between the peaks of the morning sinks and the valleys of the morning sources. This may be an effect of a more relaxed commute in the afternoon hours where punctuality is less of a concern but it could also be an artifact of our 1-hour bins coupled with peoples work schedules. For instance, it is possible that people are expected to arrive at work before 9:00 so people use the bikesharing system in the interval 8:00-9:00. In the afternoon however people might leave work later in the hour e.g. at 17:50 and then finish their commute after 18:00 meaning that the arrival is counted in the 18:00-19:00 bin resulting in the misalignment. The shape of the peaks may also show variations in peoples work schedules across cities. In New York City, people usually depart to work from 8:00 to 9:00 and then depart from work at 17:00 to 18:00. In Chicago, the wideness of the peaks indicate that people depart to work from 7:00 to 9:00 and then depart from work from 15:00 to 17:00 indicating that some people arrive at and depart from work earlier than others.

In New York City, Washington DC, London, and Helsinki, 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 of morning sink types while 55.7% are morning sources, a difference of 39.4%. This indicates that the trips eminating 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. When looking at the placement of the morning sink stations, they are typically placed in few work-related areas such as in the Helsinki city centre around Kaartinkaupunki, Otaniemi which hosts a high number of schools and universities, and the Pitäjänmäki district which contains many IT and manufacturing companies. On the contrary, morning source stations in Helsinki are mostly placed in residential areas around the work-related areas. The city which seems to be most balanced is Chicago. Here, there is only a 18.9% difference between the amount of morning sink and morning source stations. However, using these differences to solely explain the imbalances is not adequate. In Washington DC, there is only a 16.5% difference but the system is less balanced than Chicago meaning that other local factors may contribute to this imbalance.

The clustering also suggest how well a model trained on some cities will perform on others. Despite the minor differences mentioned before, there is still a high degree of sameness between the US cities indicating that a model trained on one US city will likely perform well on another US city. The same can not be said for the European cities where the clusters are more varied. Whether this is due to cultural differences or other factors it may prove problematic for the models and their accuracy .

4.2. Modelling

The coefficients obtained from the LR models trained on each city are presented in Table 5. To better visualise the relationship between the external variables of a city and the output from the LR model, different heat maps were constructed for the city. This was done by dividing the city area into 200m $\times$ 200m cells and then calculating the share of different land uses, population density, and the distance to the nearest subway and railway from the centre of the cell. These variables were then used as input for a LR model trained on the city to obtain the probability of the centre of each cell being in each cluster. The resulting heat maps for New York City can be seen in Fig. 5. It is seen that stations in commercial and industrial areas such as the Diamond District, the Financial District, and along Hudson River are more probable to be morning sink stations while morning source stations are more likely to be in residential areas and areas with a high population density. This is also readily seen in the coefficients for the model where high morning sink stations have a significantly higher coefficient for share of commercial use while the opposite is true for residential use. Looking at the coefficient for share of residential use and comparing between cluster types, it is seen that there is 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 equally valid interpretations of his result depending on the understanding of a morning source station. Morning source stations are generally unbalanced with respect to the number of arrivals and departures. One interpretation is that this imbalance is caused by an abundance of departures in the morning. This is easily explained by the relatively large number of people in residential zones compared to other zones

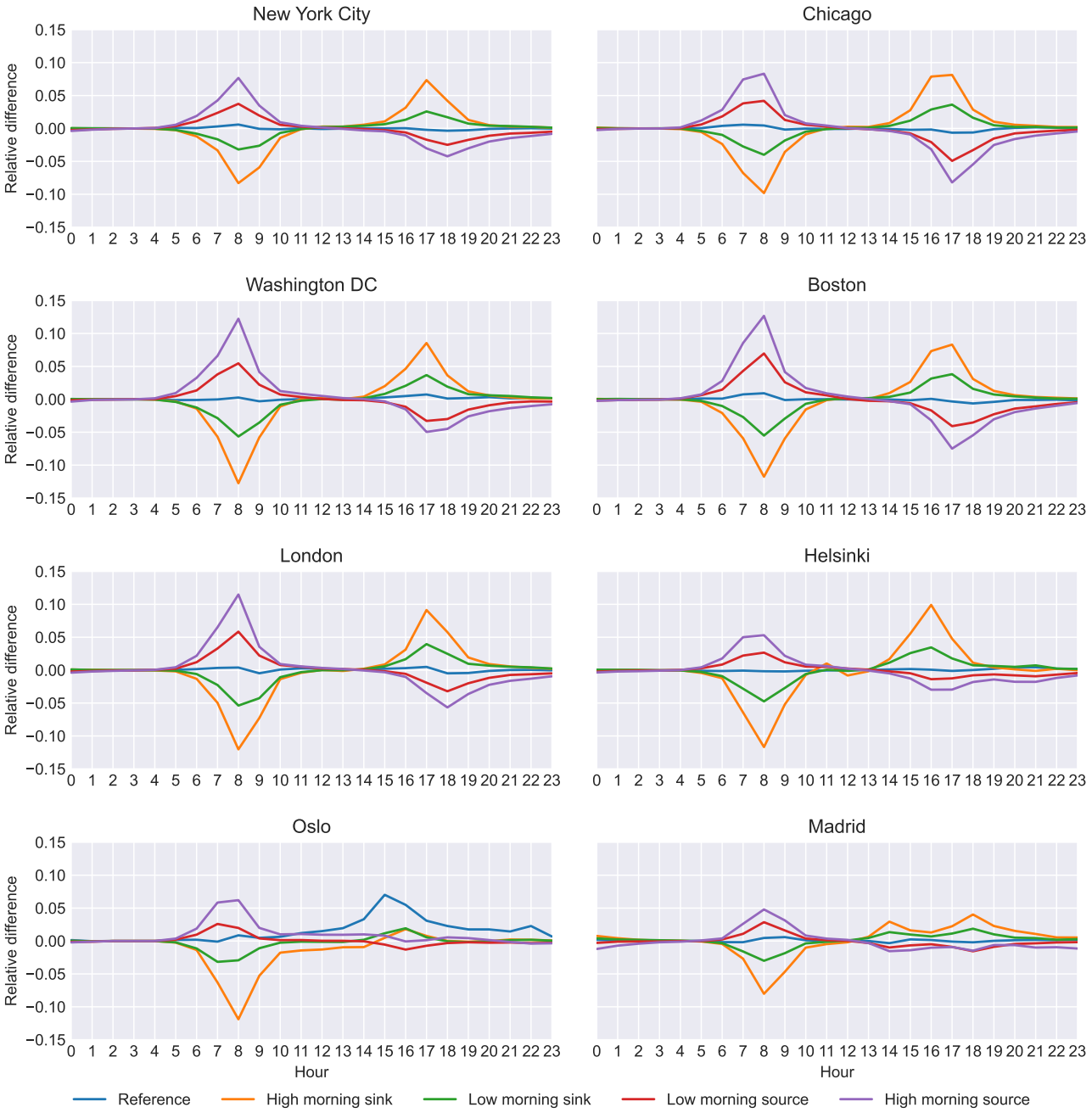


Figure 4: Cluster centres for all cities.

yielding a large amount of potential bikers who commute to commercial areas. Another interpretation is that the imbalance is caused by an absence of arrivals. Residential areas have little to no commercial purposes so there is little reason for people to be in these areas in the middle of a business day. Both of these interpretations are supported by how zoning in US cities are usually regulated. There is a high degree of separation between different types of areas and their uses. This leads to cities having large areas in which only commercial use is permitted and likewise for residential use. Thus, when people are commuting in the morning they are most likely departing from residential zones and arriving in commercial zones to work. By the discussion above, it stands to reason that the coefficient for the share of commercial use will have a monotone decrease from high morning sinks to high morning sources, an opposite pattern to the share of residential use. This is mostly the case for the US cities except for Chicago where the low morning source cluster has the highest coefficient.

In the European cities, is it difficult to detect a general pattern which is shared between the cities and with the US cities. In Helsinki and Madrid, there is a monotone decrease in the coefficient for share of commercial use, as was seen in the US cities. When looking at the share of residential use, in Helsinki the coefficients are higher in morning sources than morning sinks although without a monotone increase as was seen in the US cities. In Madrid, there is a monotone decrease in these coefficients, an opposite behaviour to what is seen in US cities. In London both the share of residential use and commercial use have high coefficients for morning sinks and low coefficients for morning sources. In the US cities, zoning regulations lead to a high degree of separation between residential and commercial areas, which led to general patterns in the coefficients. However, in European cities this separation is much less prominent. In London, residential areas are much more scattered around the city except for core business districts. This was also observed in the other European cities. This difference in zoning standards between the US cities and the European cities may affect the predicting power of models trained on a US city and tested on European cities and vice versa. Thus, the share of land use can be deemed more important when the models are only trained and tested on US cities.

Another good indicator of the station type is the distance to the nearest subway with morning sink stations being closer to subways than morning source stations. This also gives some insight into how people are using bike-sharing in conjunction with subways. At the start of a user's trip, they may use bike-sharing to cover the distance between their origin and the nearest subway which will cover the remaining distance. It should be noted that stations close to subways are not always morning sinks since people can also use a station after the subway in order to cover the remaining trip distance, making the station a morning source. However, in most US cities the subway network is more dense in downtown areas meaning the subway will likely take you within walking distance to your place of work, lowering the need for bike sharing. The proximity to a railway station is significant for most cities when the station is a morning source type. In most cities, morning source stations are further away from railway stations with the notable exceptions of Washington DC and London

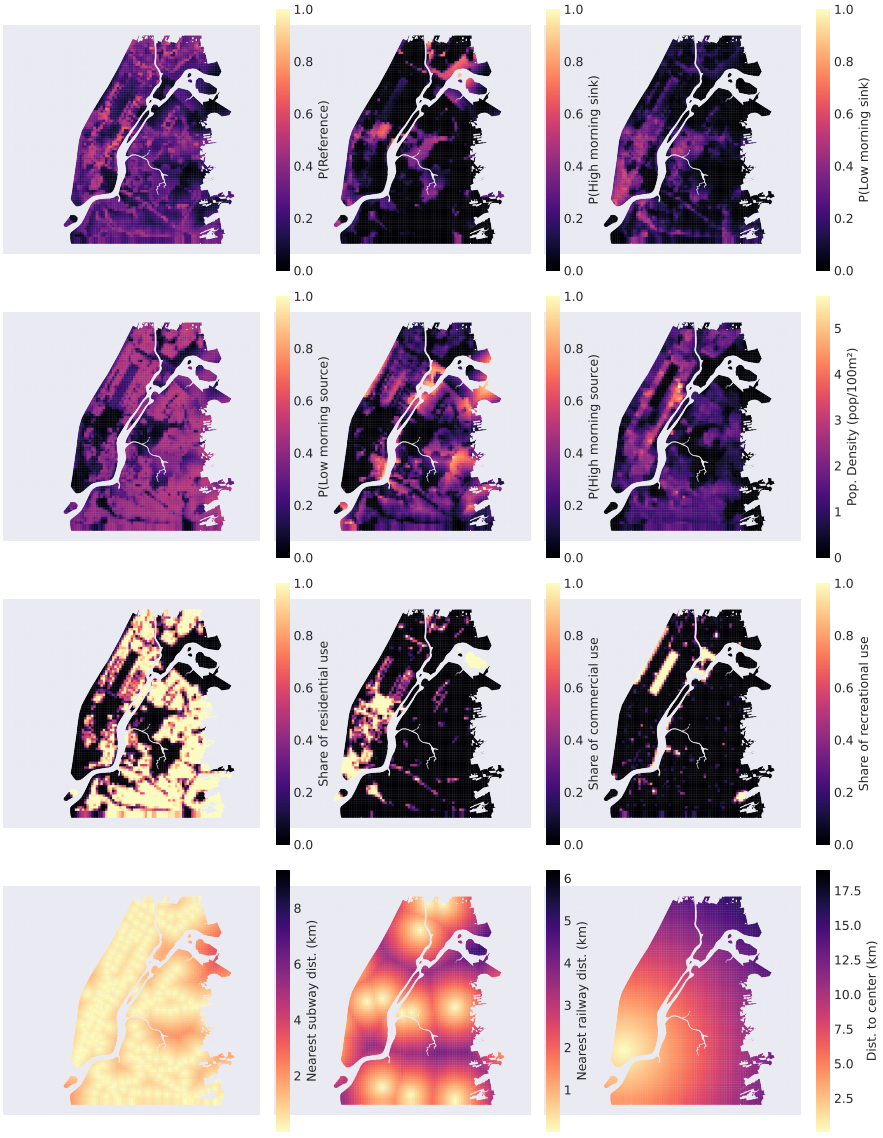


Figure 5: Heat maps of probabilities of being in different clusters.

where morning source station are closer to railway stations. The most trafficked bike sharing stations in Washington DC and London are located outside of Union Station and Waterloo Station which are the main railway stations in their respective cities. This may be related to people who are living outside of these cities commuting by railway into the cities and then use bike sharing for the rest of the commute.

In order to see how a model trained on one city generalises to another city, we tested each of the 8 models on the same data used to train the other models. The rate at which the models predicted the cluster types correctly can be seen in Fig. 6. When training and testing on the same city, we split the stations randomly into a training set and test set with the training set having about 80% of the stations. The results show that models trained on US cities perform well on other US cities with the exception of Washington DC. Excluding success rates of models trained and tested on the same city, success rates when testing on Washington DC ranged from 26% to 30% while for the other US cities, success rates ranged from 32% to 38% when training on US cities. This difference in the success rates is most likely due to the abnormal reference cluster in Washington DC which is more morning source-like than in the other clusters. When looking at the European cities separately, the model trained on Oslo is substantially more accurate when tested on Helsinki rather than London or Madrid. Likewise, the model trained on Helsinki performs particularly well on Oslo with almost twice the accuracy of random guessing. This suggest that Helsinki and Oslo are the most similar European cities. The model trained in London also performs best on Helsinki however the same performance is not seen when training the model in Helsinki and testing it on London. Madrid also serves as an outlier since the success rate of the model trained in Madrid range from 9% to 14% when testing on other cities, a significantly worse result than expected from randomly guessing. However, models trained on other cities seem to perform substantially better when tested on Madrid with success rates ranging from 23% to 33%.

When comparing between US cities and European cities, both New York City and Helsinki stand out as cities where other models perform particularly well. When testing on New York City using models trained all other cities except for Washington DC and Madrid, the success rates range from 32% to 36%. In Helsinki, the range changes to 30%-39%.

5. Conclusion

We conclude that this is a paper

References

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.

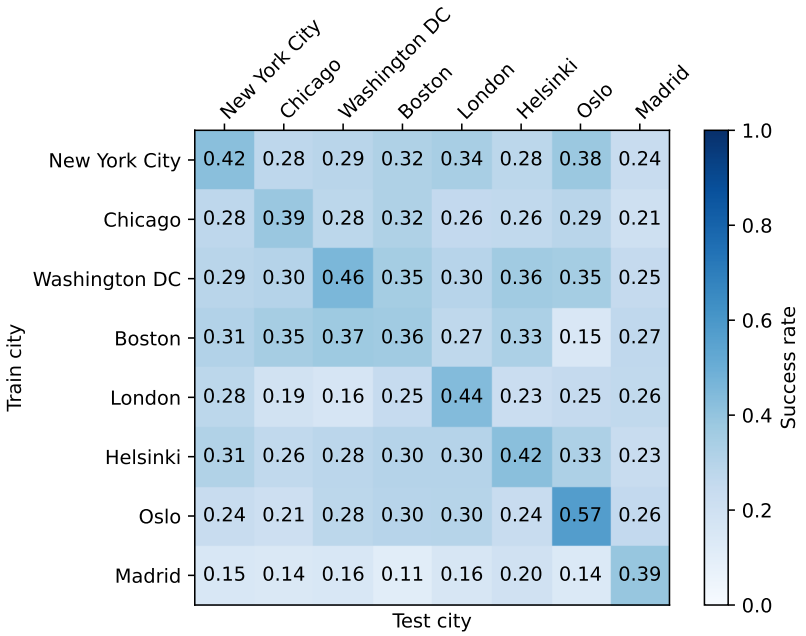


Figure 6: Success rates of models trained and tested on different cities.

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.

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, 39:1–39:21. URL: https://doi.org/10.1145/2560188, doi:10.1145/2560188. no Data (logprob: -177.683) ECC.

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 PAMI-1, 224–227. URL: http://ieeexplore.ieee.org/document/4766909/, doi:10.1109/TPAMI.1979.4766909.

Dunn, J.C., 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cyber-

netics 3, 32–57. URL: http://www.tandfonline.com/doi/abs/10.1080/01969727308546046, doi:10.1080/01969727308546046.

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

Hensher, D.A., 2020. What might Covid-19 mean for mobility as a service (MaaS)?

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. URL: https://www.sciencedirect.com/science/article/pii/S0965856417304676, doi:10.1016/j.tra.2017.11.009. 29 ECC.

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, 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.

Li, Y., Zhu, Z., Kong, D., Xu, M., Zhao, Y., 2019. Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction, in: Conference on Artificial Intelligence, AAAI. pp. 1004–1011.

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. URL: https://linkinghub.elsevier.com/retrieve/pii/S0965856416307716, doi:10.1016/j.tra.2016.08.030. 144 ECC.

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. URL: https://www.tandfonline.com/doi/full/10.1080/15568318.2018.1501520, doi:10.1080/15568318.2018.1501520. 11 ECC.

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. URL: https://linkinghub.elsevier.com/retrieve/pii/S0966692313001178, doi:10.1016/j.jtrangeo.2013.06.007.

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

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. URL: https://linkinghub.elsevier.com/retrieve/pii/0377042787901257, doi:10.1016/0377-0427(87)90125-7.

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. URL: http://link.springer.com/10.1007/s11116-015-9599-9, doi:10.1007/s11116-015-9599-9. 58 ECC.

Sathishkumar, V., Park, J., Cho, Y., 2020. Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications 153, 353–366.

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.

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. doi:10.1007/978-3-642-24264-9_10. 106 ECC.

Wang, B., Kim, I., 2018. Short-term prediction for bike-sharing service using machine learning. Transportation Research Procedia 34, 171–178.

Xiao, G., Wang, R., Zhang, C., Ni, A., 2021. Demand prediction for a public bike sharing program based on spatio-temporal graph convolutional networks. Multimedia Tools and Applications 80, 22907–22925.

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. URL: https://www.sciencedirect.com/science/article/pii/S0143622819305636, doi:10.1016/j.apgeog.2019.102130.

Yang, J., Guo, B., Wang, Z., Ma, Y., 2020b. Hierarchical prediction based on network-representation-learning-enhanced clustering for bike-sharing system in smart city. IEEE Internet of Things Journal 8, 6416–6424.

Yang, Y., Beecham, R., Heppenstall, A., Turner, A., Comber, A., 2022. Understanding the impacts of public transit disruptions on bikeshare schemes and cycling behaviours using spatiotemporal and graph-based analysis: A case study of four London Tube strikes. Journal of Transport Geography 98, 103255.

Yang, Y., Heppenstall, A., Turner, A., Comber, A., 2020c. 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, C., Zhang, L., Liu, Y., Yang, X., 2018. Short-term prediction of bike-sharing usage considering public transport: A LSTM approach, in: 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 1564–1571.

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. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968090X15000388, doi:10.1016/j.trc.2015.01.030.

Zhou, X., 2015. Understanding Spatiotemporal Patterns of Biking Behavior by Analyzing Massive Bike Sharing Data in Chicago. PLOS ONE 10, e0137922. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0137922, doi:10.1371/journal.pone.0137922. 125 ECC Publisher: Public Library of Science.

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 .

	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.

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 4: Size of the 5 clusters obtained from the clustering. The size is represented as a percentage of the total number of stations below.

Cluster	Coef. name	New York City	Chicago	Washington DC	Boston	London	Helsinki	Oslo	Madrid
High Morning Sink	Const.	−0.247	**2.193**	**2.210**	**3.623**	**−11.171**	−3.792	**19.973**	−2.938
	Share of residential use	**−4.200**	**−3.260**	**−6.538**	−0.185	**15.293**	−2.039	−10.165	**13.670**
	Share of commercial use	**2.016**	−0.511	−0.559	−0.241	**13.703**	1.636	−11.843	4.654
	Share of recreational use	−1.279	−7.586	**−3.661**	**−3.909**	6.408	2.102	−11.076	3.946
	Population density [per 100 sq. m]	**−1.809**	**−1.886**	−1.797	**−6.079**	**−3.521**	−4.598	−1.424	**−2.813**
	Distance to nearest subway [km]	−0.502	**−2.349**	0.332	0.014	0.083	−0.345	**6.194**	1.198
	Distance to nearest railway [km]	−0.205	−0.383	**0.493**	0.427	0.664	−0.024	**−6.149**	**−2.133**
	Distance to city center [km]	0.152	0.080	**−0.577**	**−0.609**	**−0.321**	**0.415**	**−3.707**	0.034
Low Morning Sink	Const.	**1.785**	**2.234**	1.238	**2.787**	**−3.956**	−0.515	**18.862**	−2.188
	Share of residential use	−0.875	−1.261	**−3.797**	−1.147	**6.567**	0.333	−9.182	**9.133**
	Share of commercial use	0.301	−1.798	−0.400	−0.661	**5.758**	1.266	−10.123	3.983
	Share of recreational use	−0.093	**−2.628**	**−1.747**	−2.090	1.863	−2.341	−11.704	1.379
	Population density [per 100 sq. m]	**−0.769**	**−1.675**	−0.713	**−1.902**	**−1.726**	**−1.625**	−1.760	**−2.052**
	Distance to nearest subway [km]	**−1.972**	**−2.140**	−0.341	−0.323	−0.046	−0.172	**2.543**	2.100
	Distance to nearest railway [km]	−0.143	0.138	**0.351**	−0.306	**1.001**	0.034	**−3.175**	−0.687
	Distance to city center [km]	−0.065	0.007	**−0.233**	−0.162	**−0.356**	0.102	**−2.093**	0.040
Low Morning Source	Const.	**−2.281**	**−1.507**	**−2.333**	−0.726	−0.004	**−4.019**	15.057	−6.033
	Share of residential use	**1.708**	**3.955**	**2.904**	0.134	−1.814	2.028	−10.586	−1.215
	Share of commercial use	−1.031	0.434	−2.536	−1.062	−2.432	−1.359	−12.577	4.138
	Share of recreational use	**1.544**	−0.320	0.231	**−3.259**	−1.298	2.262	−8.750	6.801
	Population density [per 100 sq. m]	0.271	−0.243	**1.712**	−0.379	**1.111**	**1.261**	1.739	**1.908**
	Distance to nearest subway [km]	**1.243**	**1.013**	**0.825**	**0.599**	**0.987**	**0.340**	**3.012**	0.732
	Distance to nearest railway [km]	**0.283**	**0.654**	**−0.426**	**0.704**	**−0.924**	**0.385**	−1.313	**−0.868**
	Distance to city center [km]	0.011	**−0.226**	0.150	0.017	**0.177**	0.075	**−1.880**	0.242
High Morning Source	Const.	**−5.030**	−1.263	**−6.701**	−0.657	−1.461	**−2.356**	10.181	−9.216
	Share of residential use	**3.002**	**6.045**	**5.584**	0.805	−0.633	−0.564	−6.766	−1.206
	Share of commercial use	**−3.195**	−2.686	**−8.969**	−3.322	−2.532	**−7.125**	−9.546	2.369
	Share of recreational use	**3.623**	**−3.690**	0.668	**−5.676**	−1.905	−1.580	−4.786	5.945
	Population density [per 100 sq. m]	**0.693**	−0.101	**2.268**	−1.100	**0.913**	0.867	1.314	**1.941**
	Distance to nearest subway [km]	**2.950**	**1.441**	**1.230**	0.469	**2.241**	0.222	**2.585**	**4.715**
	Distance to nearest railway [km]	**0.413**	0.353	**−0.523**	**0.751**	**−0.777**	0.256	**−1.507**	**−1.167**
	Distance to city center [km]	−0.061	**−0.376**	**0.686**	0.045	0.112	**0.359**	−0.805	**1.025**

Table 5: Coefficients of LR models trained on different cities. Bold coefficients are statistically significant ($p < 0.05$).