Forecasting User Behavior to Optimize Fleet Distribution

Investigating the Viability of Predictive Models to Harmonize Supply and Demand in Berlin Carsharing

Florian König

Software Engineering

Bachelor of Science

Berlin, June 14th, 2025

Forecasting User Behavior to Optimize Fleet Distribution

Investigating the Viability of Predictive Models to Harmonize Supply and Demand in Berlin Carsharing

Florian König

[redacted]

Supervisor: Samuel Boguslawski
Professor of Software Engineering
CODE University of Applied Sciences

Co-supervisor: Dr. Kristian Rother
Data Science Trainer
Academis

Bachelor of Science Software Engineering

Spring Semester 2025

Bachelor's Thesis

Berlin, June 14th, 2025

Abstract

Free-floating carsharing systems face challenges in balancing vehicle supply and demand, often resulting in inefficient fleet distribution and reduced vehicle utilization. This thesis explores the potential of predictive models to forecast vehicle demand and user trip destinations within Berlin's carsharing. The central hypothesis states that both demand for and destination of trips can be accurately predicted, enabling targeted incentives that naturally align supply and demand.

To explore this, usage data was collected over 100 days, observing more than one million trips and 7,200 vehicles from Berlin's largest FFCS provider. This data was complemented with contextual information, including points of interest, weather conditions, and traffic data. To determine the possibility of predicting demand hotspots and trip destinations, a combination of spatiotemporal analysis and machine learning models was utilized.

The findings reveal that POIs are a significant factor influencing long-term demand patterns, while weather and traffic show little predictive power. Although the study evaluates multiple models for next-place prediction, it concludes that datasets lacking user-specific information and broader contextual information cannot provide the required accuracy. Despite these limitations, the research underscores the potential for future advancements by incorporating additional data sources.

Contents

List of Figures

List of Tables

Acronyms

FFCS	free-floating carsharing
SBCS	station-bound carsharing
POI	point of interest
O/D	origin/destination

AUC	area under the curve
GAM	generalized additive model
GRU	gated recurrent unit
GWR	geographically weighted regression
LASSO	least absolute shrinkage and selection operator
MC	Markov chain
OLS	ordinary least squares
PCA	principal component analysis
RNN	recurrent neural network
ROC	receiver operating characteristic

1
Introduction

Since the commercial introduction of free-floating carsharing (FFCS) by car2go in 2010, this mobility solution demonstrated environmental benefits that contribute to urban sustainability (Firnkorn et al., 2011). Unlike traditional station-bound and peer-to-peer carsharing, FFCS offers users the flexibility to pick up and drop off vehicles anywhere within a defined service area. This flexibility has fueled the consistent expansion of carsharing services (BCS, 2024a). In 2024, Berlin had one shared vehicle for every 126 privately owned vehicles (KBA, 2025; BCS, 2024b).

Despite its advantages, FFCS faces a challenge in predicting and balancing vehicle demand. Demand in different locations fluctuates throughout the day, creating areas with an oversupply of unused vehicles or unmet demand (Jorge et al., 2015; Lippoldt et al., 2019). Unlike bikes or kick scooters, cars are costly and logistically challenging to relocate, as they cannot be easily transported to high-demand areas by truck (Stokkink et al., 2021; Caggiani et al., 2013). Moreover, increasing the number of vehicles in the fleet does not necessarily resolve this imbalance (Weikl et al., 2015).

FFCS operators address this issue through various strategies. These include employing relocation drivers, commissioning crowdsourced relocation providers, and offering user incentives for long-time underutilized vehicles (Schulte et al., 2015; *Job Post* 2025; *Streetcrowd* 2025). However, these approaches often fall short of providing a sustainable, long-term solution. In search of a better, more organic solution, this thesis collects and analyzes anonymous real-world usage data (section 3.2 on page 11). Specifically, I will investigate whether it is possible to predict areas with higher demand (chapter 4 on page 16) and to determine potential trip destinations (chapter 5 on page 26), therefore understanding if a trip is likely to improve or worsen the balance of supply and demand.

Validating this hypothesis could enable more targeted incentives for users, encouraging service usage in ways that positively contribute to the system. This would enhance the competitiveness of FFCS and deliver broader benefits for urban mobility.

1.1 Hypothesis

I hypothesize that it is feasible to develop predictive models capable of accurately forecasting vehicle demand and user trip destinations within Berlin's carsharing ecosystem. These capabilities could enable targeted incentive mechanisms to encourage users to make trips if they are likely to go to areas experiencing negative supply-demand imbalances.

2 Related Work

Multiple studies emphasize the relevance of my research in an implicit user-based relocation strategy. Schiffer et al. (2021) demonstrated that implementing such strategies can increase fulfilled rental requests by 21% and operator revenue by 10%. Similarly, Stokkink et al. (2021) found that user incentives in a station-bound carsharing (SBCS) system are more profitable and sustainable than staff-based relocations and showed that a hybrid operator-user relocation system maximized profit and service level. Willing et al. (2017) further noted that user-based relocations contribute to the self-balancing of carsharing systems.

Over the past decades, carsharing has evolved significantly. Early research focused on station-bound carsharing, which is often divided into roundtrip and one-way modes of operation (Illgen et al., 2019). More recent studies have shifted to free-floating carsharing (FFCS), where vehicles can be left at any legal parking spot. This version of carsharing is now predominant in many cities, consistently experiencing higher growth than SBCS (Shaheen et al., 2015; BCS, 2024a; Jorge et al., 2013).

For the purposes of my research, this review of related work focuses on FFCS, however, also lists relevant SBCS findings. In many cases, free-floating demand can be treated analogously as areas of higher and lower demand instead of specific stations.

Ciari et al. (2014) used MATSim to simulate both SBCS and FFCS in Berlin, finding that the two systems are complementary rather than competitive. Each system serves distinct trip types and temporal patterns, highlighting the need for models that accommodate multiple carsharing modes.

Additionally, many current studies focus on either fully electric or combustion fleets. Wu et al. (2022) note that more research is needed for mixed vehicle fleets to reflect reality more closely. Even though user booking behavior does not differ significantly between the vehicle types (Niels et al., 2017), electric vehicles introduce additional complexity as relocation strategies have to be charge-aware.

2.1 Predicting Demand

Research into carsharing demand has been extensive, identifying various influencing factors. Statistical techniques like seasonal autoregressive integrated moving averages have been used to identify patterns (Müller et al., 2015). To understand flows within a city, Wang et al. (2018) and Le Vine et al. (2014) have leveraged the gravity and radiation models for urban mobility. Other studies applied machine learning approaches such as random forest regression and support vector regression (Cocca et al., 2020).

The literature identifies four key categories of factors influencing carsharing demand: environmental factors, temporal influences, demographics, and neighborhood characteristics.

2.1.1 Environmental Factors

Points of interest (POIs), such as malls, nightclubs, restaurants, or medical facilities, are significant drivers of carsharing demand. Schmöller et al. (2015) analyzed the spatial distribution of booking data and found that demand correlates with urban structures, concentrating around temporal peaks and spatial hotspots. This was further demonstrated by Wagner et al. (2015), who applied their model to the area surrounding the analyzed operating zone before expanding the latter and proving their predictions' accuracy.

In addition to POIs, Ménoire et al. (2020) and Celsor et al. (2007) found that both the walkability of the area around a carsharing station and regular bus service contribute positively to usage. For FFCS, however, Willing et al. (2017) observed that a lower density of bus stops increases carsharing activity. At the same time, train services seem to be used in direct combination with shared cars and were found to impact usage positively.

Interestingly, Kang et al. (2016) reveals that in Seoul, areas with more registered cars and fewer subway entrances are associated with higher carsharing demand. This suggests a difference in carsharing use between western regions and Asia.

Weather conditions further influence demand. Initially, Schmöller et al. (2014) did not observe a significant impact of temperature or rainfall on bookings on a daily scale. However, follow-up research by the same authors revealed that sudden weather changes, namely the start or end of rainfall, affect short-term demand dynamics (Schmöller et al., 2015). A case study in Basel highlighted that FFCS becomes more attractive compared to public transport at night, during cold weather, or in rainy conditions (Becker et al., 2017). This study also noted FFCS is primarily used for discretionary trips when public transport alternatives were substantially inferior.

2.1.2 Temporal Influences

Willing et al. (2017) expanded on the POI-based approaches by Schmöller et al. and Wagner et al., incorporating a temporal dimension to analyze FFCS demand in Amsterdam. Their findings revealed that different POI categories have varying impacts on demand at different times of the day. Restaurants, for example, showed a negative impact in the early morning but reversed this in the evening. They validated their price optimization and service area decision support system by applying it to Berlin.

Furthermore, Messa et al. (2021) identified a stronger demand correlation for work-related POIs during peak hours, while leisure destinations dominated off-peak hours, and recreational and social destinations were prevalent on weekends.

Studies consistently agree that temporal patterns in FFCS vary significantly between weekdays and weekends. Weekday demand is primarily driven by commuting trips, whereas weekends see a shift toward discretionary travel (Schmöller et al., 2015; Messa et al., 2021; Weikl et al., 2013).

2.1.3 Demographics

Numerous studies across different have identified age as one of the most reliable predictors of carsharing use, with young adults between 18 and 37 years old being the primary drivers of demand (Carrone et al., 2024; Schmöller et al., 2014). Jo et al. (2024) and Chun et al. (2019) show that the significance of this predictor is further increased in areas with lower rental activity. Similarly, areas with a higher household income and higher shares of educated people tend to experience shorter vehicle idle times (Carrone et al., 2024). A higher concentration of companies in an area also positively correlates with increased usage (Schmöller et al., 2015).

These findings are also evident for SBCS. For instance, Ménoire et al. (2020) used a growth model to show that areas with a higher share of unemployed people impacted usage negatively. However, they also identified people aged 35-44 as the most relevant user group for SBCS.

Demographic factors influence long-term demand patterns, remaining relatively stable throughout the day (Schmöller et al., 2015).

2.1.4 Neighborhood Characteristics

Lastly, neighborhood factors, such as population density or parking pressure, impact carsharing demand. Areas with limited parking make car ownership less convenient, thereby increasing the attractiveness of carsharing. These areas often also feature high population densities, which provide a large customer base with typically lower travel and vehicle ownership rates (Celsor et al., 2007).

2.1.5 Short-Term Demand

Short-term demand is shaped by dynamic factors, such as weather changes and the proximity of available vehicles. While demographics and neighborhood characteristics stay consistent for months or years, and temporal and geographical influences often follow predictable patterns, short-term trends are more variable. For example, Herrmann et al. (2014) developed a short-term demand model using real-time FFCS data and revealed evidence of the importance of relocation strategies for the FFCS business.

Weikl et al. (2013, 2015) differentiated between the different demand patterns by creating complementary models. One model utilized historical data to forecast long-term demand, while the other monitored real-time demand to adjust outputs and predict optimal future vehicle distributions. This approach reduced vehicle idle times by 4%.

To better understand short-term demand, some studies, such as Wang et al. (2019), additionally analyzed the use of the corresponding carsharing app. By observing actions like station selection and refreshes, they inferred demand for vehicles. Combined with temporal patterns, this allowed them to activate incentives for vehicles more quickly and increase utilization.

Notably, Wagner et al. (2015) highlighted that ”carsharing is a complex business and humans can anticipate certain effects, such as the buzz created by the opening of a new shopping center before they are reflected in the data.” While predictive systems can process vast amounts of data, designing them to match the adaptability of human intuition remains a challenge.

2.2 Predicting Trip Destinations

Understanding where users are likely to take a vehicle is critical, as vehicle supply can directly influence carsharing demand (Balac et al., 2016). While trip destinations are independent observations, origins depend on vehicle availability (Willing et al., 2017).

Archetti et al. (2023) evaluated the benefits of requiring users to input departure and arrival times and locations before booking a trip. This approach allowed the system to preemptively assign cars to reservations, doubling both reservation satisfaction and vehicle utilization. Their findings prove the value of accurate trip destination predictions, not only for user-based relocations but also for improving overall system efficiency.

Predicting trip destinations is closely related to studies of human mobility. Song et al. (2010) found that mobility is highly predictable, with a 93% predictability rate based on historical information, even though this value varies strongly between users.

2.2.1 Extrapolating Trips

Besse et al. (2016) clustered historical taxi trips geographically and applied Gaussian Mixture Models to predict destinations, achieving 85% accuracy in cluster identification. Their predictions improved as trips progressed, narrowing from one out of 10 clusters at 50% completion to one out of 100 at 70%. Similarly, Wang et al. (2018) applied these concepts to SBCS, correctly identifying areas of stations in 92% of cases for trips past 70% completion. Liu et al. (2021) further enhanced prediction accuracy by incorporating user data outside of trips, frequent user locations, app usage, and system-wide utilization metrics.

Recent advancements in deep learning have yielded significant improvements in destination prediction. For example, Casabianca et al. (2021) integrated attention layers into bidirectional LSTM models, enabling predictions of all destinations within 500 meters for Beijing vehicles after the halfway point of a trip. The attention mechanism dynamically weighted previous states, improving results and reducing training time.

Research on private car trips has shown that classifying users by entropy ranges — measuring the variability of their travel patterns — can enhance prediction accuracy, as driver predictability varies significantly (Jiang et al., 2021). This finding suggests that incorporating user trip history could enhance accuracy in destination predictions for carsharing systems.

However, to enable effective incentives for user-based relocation systems, destination predictions must be made before a trip begins.

2.2.2 Next-Place Prediction

A wide range of techniques has been explored for the next-place prediction problem, including (hidden) Markov models, support vector machines, recurrent neural networks (RNNs), and advanced algorithms like Active LeZi. A survey by Schreckenberger et al. (2018) identified RNNs as the best-performing method, with most studies incorporating both temporal and spatial features alongside additional context in 41% of cases.

Early research predominantly relied on Markovian models. Gambs et al. (2012) extended mobility Markov chains (Markov chains with location states) by incorporating users' previously visited locations. They achieved optimal results using two prior locations, predicting the next POI with 70–95% accuracy; adding additional history yielded negligible results. This demonstrated the predictability of human mobility when recent history is considered. In another approach, Mathew et al. (2012) employed hidden Markov models to predict future locations, treating movements as outputs from hidden contextual variables like activities and goals. However, their accuracy was limited, likely due to the application of planet-scale models to localized datasets.

Comito (2020) introduced decision trees for analyzing location-based social network data, differentiating between personal, cumulative, and mass mobility patterns. Their method achieved over 80% accuracy and revealed that crowd effects observed through social media could be leveraged to predict mobility. While the use of such data would introduce additional privacy considerations, similar effects may be detectable in carsharing platform metadata.

Recurrent neural networks have shown significant promise. For instance, Al-Molegi et al. (2016) proposed an RNN model using time and POI embeddings, outperforming traditional Markov chain and machine learning models. Their approach also reduced manual decisions for clustering parameters in spatio-temporal combinations. Supporting this, Chekol et al. (2022) found that merging multiple contextual features consistently improved performance, further validating the superiority of deep learning techniques over traditional methods.

Despite the extensive research on general destination and individual next-place prediction, a significant research gap exists for carsharing systems. Most studies, as reviewed by Schreckenberger et al. (2018), rely on large open datasets like Microsoft GeoLife, Nokia MDC, or MIT Reality Mining. To my knowledge, no open real-world dataset for carsharing trips exists, suggesting additional research opportunities if such data becomes available.

2.2.3 Systemic Factors

Contextual and environmental factors play a highly relevant role in next-place prediction. Consistent with the factors influencing demand, temporal and POI predictors are essential and can be grouped into three trip categories: systemic (repeating origin/destination pairs on weekdays), occasional (linked to entertainment and shopping on weekends), and trips influenced by public transport integration, particularly in areas with limited public transport options (Messa et al., 2021). Formentin et al. (2015) utilized similar patterns to decompose trip series into stationary stochastic, linear trend, and seasonal components.

Notably, the contextual factors influencing trip destinations differ between cities. While in Zurich, carsharing was predominantly used for discretionary trips with worse public transport alternatives (Becker et al., 2017), in Madrid, limited parking availability in certain areas made parking regulation and supply data more predictive than traditional activity-based features (Ampudia-Renuncio et al., 2020a).

2.3 Incentive Mechanisms

Various incentive policies have been studied to address vehicle imbalances in carsharing systems. These include trip merging and splitting (Barth et al., 2004); suggesting destination changes (Di Febbraro et al., 2012; Wagner et al., 2015; Di Febbraro et al., 2019; Clemente et al., 2018); offering incentives for unused vehicles or trips ending in specific areas (Lippoldt et al., 2018, 2019); and implementing zone-based pricing (Willing et al., 2017). Jorge et al. (2015) proposed adjusting prices depending on a trip's impact on fleet distribution, and Wang et al. (2021) and Waserhole et al. (2016) used both monetary incentives and surcharges to address vehicle imbalances.

Herrmann et al. (2014) surveyed carsharing users and found that 85% would choose a more distant car if offered at a 10ct/km discount, with an additional 13% willing at higher discounts. Furthermore, a similar proportion of users were open to specifying their trip destination at the start, indicating strong user acceptance of incentive-based interventions.

2.3.1 Incentive Performance

Studies in Milan and Cologne examined the effects of surcharges for low-demand areas, reduced rates, and bonus minutes. Bonus minutes were found to increase take rates and lead to quicker rentals, while reduced rates resulted in trips ending in attractive zones more often. However, the success of these incentives depended on their timing and spatial granularity. Poorly targeted or delayed incentives were less effective, and overly large incentive zones failed to align with actual demand hotspots (Lippoldt et al., 2018, 2019).

Jorge et al., 2015; Wang et al., 2021 developed models for dynamic pricing and combinatorial incentives and surcharges, demonstrating that such policies can improve fleet balance and increase operator revenue by up to 22.5%. Similarly, Waserhole et al. (2016) provided analytical insights into pricing optimization, showing that well-designed incentives significantly improve system throughput.

2.3.2 Behavioral Impacts

User acceptance of incentives is generally high. However, while Herrmann et al. (2014) reported that 85% of users would choose a more distant car for a small discount, user sensitivity varies by demographic factors. Younger, male, and frequent users are more responsive to incentives (Wang et al., 2021).

Lippoldt et al. (2018, 2019) observed that zone-based bonus minutes often led to unintended behaviors, such as short trips to the border of incentivized areas, as users sought to maximize rewards. In contrast, reduced rates were more likely to encourage trips ending in attractive zones. Implicit incentives can influence rental destinations. However, they cannot fully replace operator-based relocations or incentives for explicit destinations.

2.3.3 Considerations

While incentive mechanisms can significantly reduce the reliance on costly operator-based relocations and enhance system profitability, their design and implementation present challenges. Complex pricing schemes or poorly communicated incentives confuse users or lead to unintended behaviors, such as short trips solely to claim bonuses. Moreover, from a societal perspective, directing vehicles away from low-demand areas may reduce mobility options for some residents (Willing et al., 2017). Ultimately, successful incentive policies must be carefully calibrated and continuously evaluated to achieve a balance between operator goals and user experience.

3 Data

The primary dataset for this analysis comprises carsharing activity collected from Berlin's largest FFCS provider, MILES. MILES currently offers more than 7,200 shared vehicles across Berlin to their more than 2.3 million customers (*Miles Sustainability Report* 2023).

The data were gathered over a duration of 100 days between March and June 2025, during which more than 7,200 vehicles and over 1 million trips were tracked. The dataset includes vehicle id and status, trip start and end points, reservations, and discounts, with a typical spatial accuracy within 50 meters. The delay between actual trip start and end events and their recording was observed to be less than 2 minutes. Additional data points are detailed in section 3.3 on page 13. Furthermore, traffic and weather data were obtained periodically, and POI data was collected at the end of the observation period to analyze their relationship with trip patterns.

Compared to the most commonly used open geolocation datasets, such as GeoLife or the Mobile Data Challenge, this dataset contains substantially more trips, although over a shorter observation period and without linkage to individual users. While much larger publicly available and continuously updated datasets exist for other modes of transport, such as the New York City Taxi and Limousine Commission (TLC) dataset with over 3 billion taxi and for-hire trips and the Citi Bike System dataset with more than 100 million bikesharing records, large-scale FFCS data are not commonly made publicly available (*TLC Trip Record Data* 2025; *Citi Bike System Data* 2025).

Most recent carsharing studies rely on data obtained through sharing agreements with operators, often involving predecessors of Free2Move (Kortum et al., 2016). A notable exception is the InnoZ project, a collaboration between the city of Berlin, Deutsche Bahn, and T-Systems, which scraped over 50 million trips over five years across multiple cities and providers. However, this dataset is not publicly available, and InnoZ itself no longer exists.

To my knowledge, the dataset used in this thesis, with its focus on Berlin's FFCS, is more comprehensive than any currently available public dataset in this context.

Figure 3.1: *Geographical distribution of trip start points*

Figure 3.2: *Estimate of shared fleet sizes in Berlin by carsharing operator, based on observations during April 20—23, 2025.*

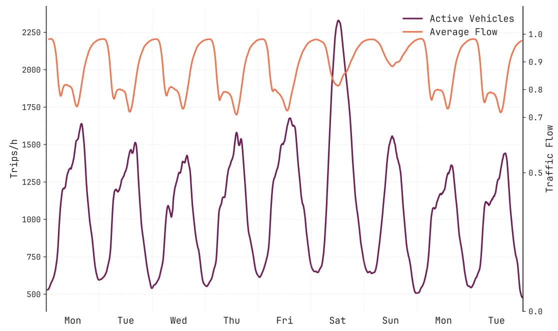


Figure 3.3: *Count of trips per hour and average city-wide traffic flow (March 31st—April 8th, 2025)*

3.1 Privacy Implications

The use of real-world carsharing data raises privacy considerations, even in the absence of direct personal identifiers. The dataset used in this thesis was obtained from the MILES API, which, at the time of collection, allowed retrieval of vehicle status and location using randomized authentication. This included both resting and active states of vehicles. The service implements security measures that are inadequate to mitigate potentially malicious attacks, even lacking automated rate limiting or consistent route authorization.

Notably, while the dataset includes data for thousands of vehicles and more than one million trips, it does not contain user data or associate trips with specific individuals. Fortunately, MILES does not appear to expose user data publicly. Despite this, the possibility of tracking real-time trip locations could potentially be exploited. If a malicious actor could, for example, associate a license plate and time of use with a specific individual, this would enable them to reconstruct this individual's entire trip — including origin, destination, and intermediate points. Moreover, users with regular travel patterns, such as daily commutes or visits to loved ones could have their routines extracted, potentially leading to their identification.

This is clearly a complicated issue to resolve. Not only is it necessary to make resting vehicle states available publicly for users to use the service, but Shared mobility is often distributed through a multitude of partners, such as Jelbi, Bolt, Sixt, and Freenow, each implementing a different security concept.

3.2 Data Collection

The data collection process primarily relied on automated scraping of carsharing activity from the MILES API, complemented by periodic retrieval of weather and traffic data. This method aligns with approaches used in other studies that have utilized scraping to gather carsharing data, such as Ampudia-Renuncio et al. (2020b) and Kortum et al. (2016). The system architecture is outlined in figure 3.5 on page 13.

3.2.1 Carsharing Data

Vehicle data were continuously monitored over a 100-day period between March 23 and July 5, 2025, with a three-day gap caused by a server issue. When a vehicle became unavailable, its status was checked to determine whether it was reserved, a trip had started, or it had entered a service mode. Once a vehicle became available again, the current reservation or trip was considered complete. To minimize the impact on the operator's systems, the frequency of API requests was deliberately limited.

Tracking active vehicles posed additional challenges due to the API's design. Higher FFCS usage throughout the day led to faster trip detection but resulted in fewer updates during trips. For reservations, the vehicle's status was monitored frequently until it expired or a trip began. Vehicles in service states were excluded from tracking to ensure data accuracy and efficient monitoring.

3.2.2 Complementary Data

Weather and traffic data collection began on March 30, 2025, and thus covers one week less than the carsharing data. Weather information was retrieved from WorldWeatherOnline through wttr.in, while traffic data were obtained from TomTom. Both datasets are indexed by EPSG:3857 tiles, allowing for efficient usage with vehicle locations. For traffic data, the average traffic flow (ranging from 0% for full congestion to 100% for free flow) and a street density score were calculated per tile and observation. The traffic flow calculation process is shown in figure 3.4.

POIs were requested from OpenStreetMap on June 22nd, 2025. For this, 14 categories (described in table B.1 on page 42) were defined and exported within the service area's bounds.

3.2.3 Data Cleaning

Data cleaning was focused primarily on the carsharing dataset to ensure the quality and relevance of the analysis. Trips that were either very short (less than 1 km) or long-time rentals (more than 12 hours) were removed, as these do not represent the typical ad-hoc use I am analyzing and, in some cases, represent scraping issues. This removed 24.7% of trips. To better capture typical usage patterns, trips occurring on holidays were also filtered out, though school holidays are included (5.3% of trips). Only public trips were retained, with internal trips excluded from the dataset. Finally, the dataset was limited to trips in passenger vehicles (which excluded sizes L and XL) that both started and ended within the Berlin city boundaries. An additional 0.4% of trips were removed for being incomplete or being otherwise irregular. After cleaning, 71% of trips remained, and only associated vehicles, waypoints, and reservations were kept. The exact counts of records after cleaning are summarized in table 3.1 on page 14.

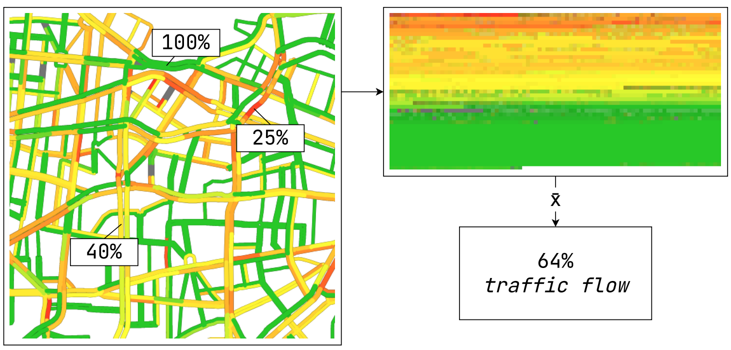


Figure 3.4: *Illustration of traffic flow calculation from a traffic map tile*

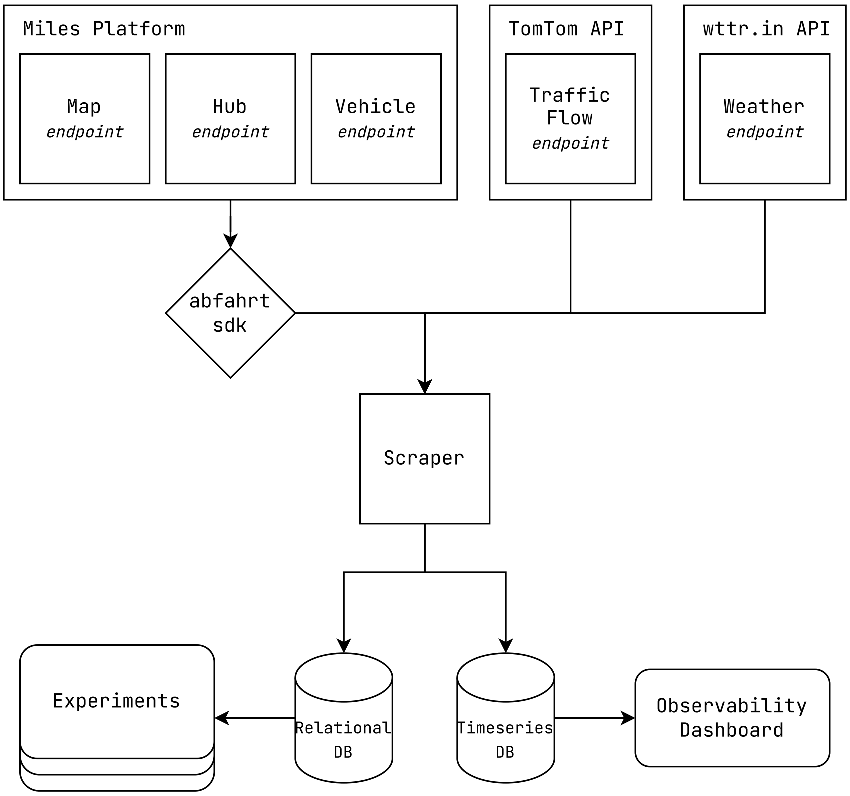


Figure 3.5: *System architecture of the data scraper*

3.3 Data Description

Each vehicle in the dataset includes detailed information, such as license plates and vehicle models. For every trip, the dataset records start and end points, intermediate points with timestamps and locations, the vehicle's current status, and any discounts applied. Reservations are also tracked, including their start and end times, along with associated locations. Additionally, changes to discounts for each vehicle are logged, offering insights into dynamic pricing and incentive mechanisms. Vehicle counts are aggregated every 15 minutes.

Weather data, such as temperature and precipitation, is available at 60-minute intervals and is spatially organized into tiles of approximately 2.4 km. Traffic flow and street density data are provided every 20 minutes for smaller tiles of 1.2 km, offering a high-resolution view of urban mobility conditions. These contextual variables enable the analysis of how external factors influence carsharing usage.

Due to the aforementioned API request restriction, there is a delay of up to 2 minutes between the actual start or end of a trip and its observation in the dataset. During active trips, vehicle status updates were typically recorded every 10–20 minutes. However, during peak times, the update interval for intermediate points could extend to around 60 minutes.

The dataset reflects high rental activity, with an average of 150 trips per vehicle and a median of 170 trips. The data model is illustrated in figure 3.6 on the following page, and detailed value counts are provided in table 3.1 on the next page.

Table 3.1: *Overview of cleaned data*

Data Source	Attributes	Volume/Frequency
Miles	*Trips:* Full trips with start and end points; optionally intermediate points; data on discounts	1,180,173 trips
	Waypoints: Point along a trip route with time; location; and vehicle status	5,204,771 waypoints
	Vehicles: Every vehicle in Berlin, including metadata	6,486 cars, 877 vans
	Reservations: Start and end times; location; discount at reservation start	1,393,230 reservations
	Discount Change: Placed, changed, or removed incentive; time	1,892,705 changes
	Density: Vehicle counts by size group	15-minute per postcode
TomTom	*Flow:* Average area-wide traffic flow and street density	20-minute per 1.2km tile
WorldWeather-Online	*Weather:* Current conditions including temperature; rain; visibility; and other factors	Hourly per 2.4km tile
OpenStreetMap	*POIs:* Categorized in 14 types; name and location	42,021 POIs

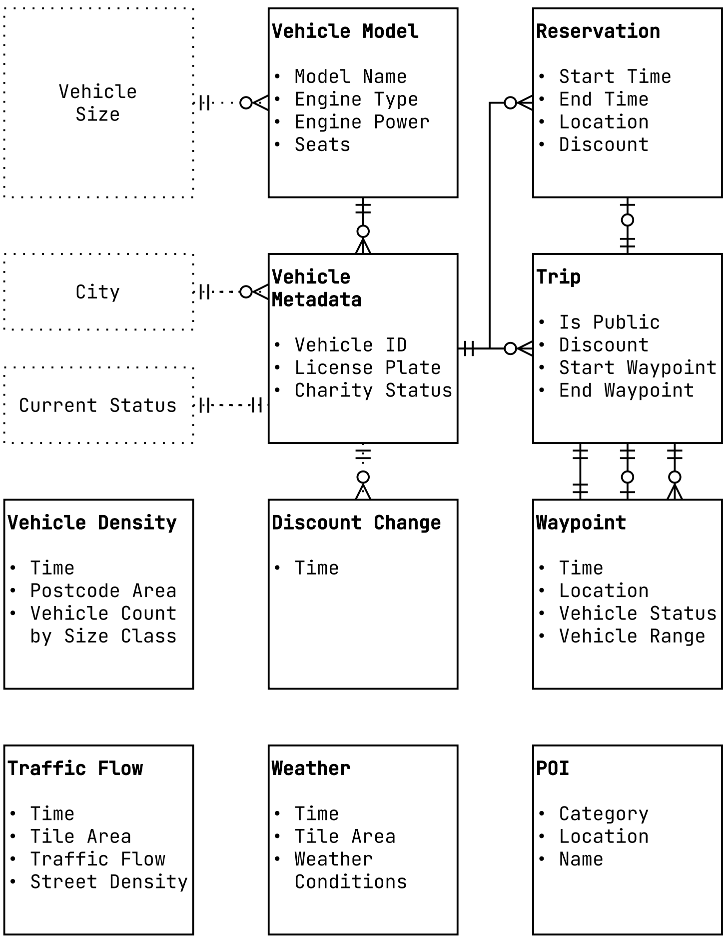


Figure 3.6: *Entity-relationship diagram of scraped data*

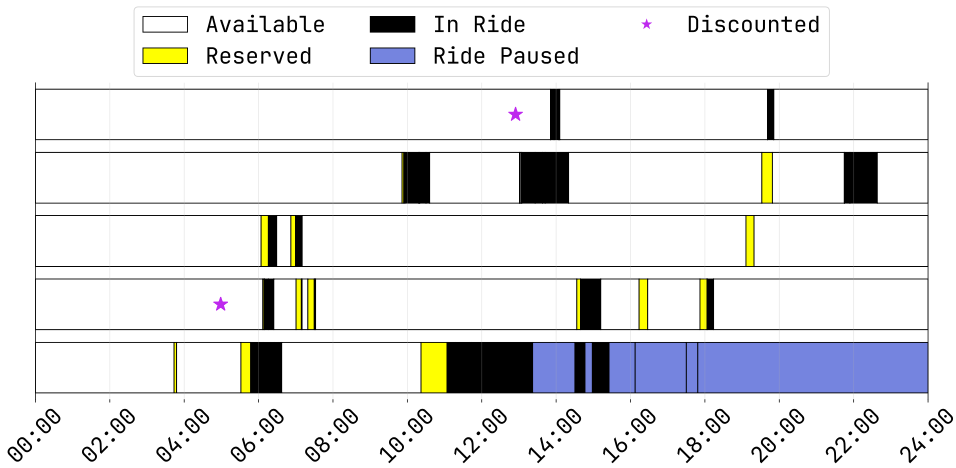


Figure 3.7: *Timeline of randomly sampled vehicles on June 20th, 2025*

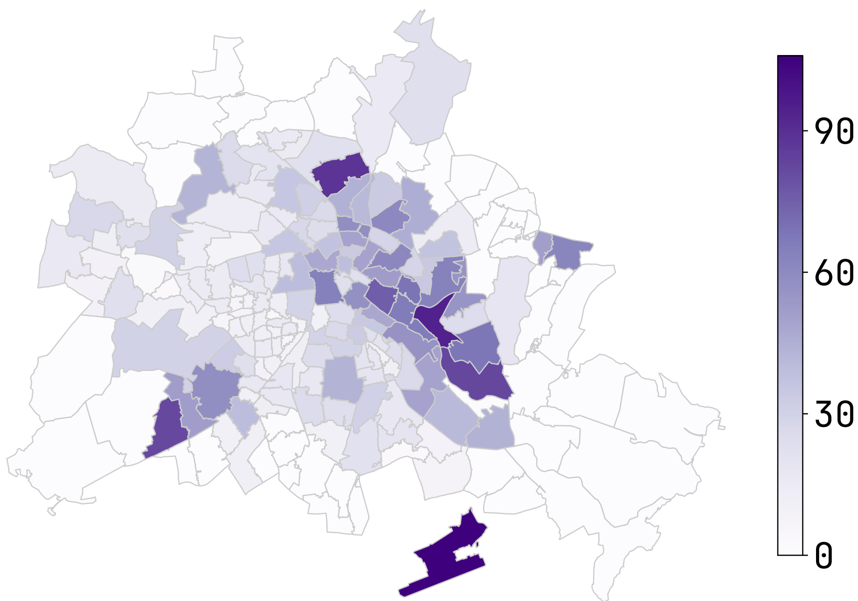


Figure 3.8: *Average vehicle count per Berlin postcode and BER airport*

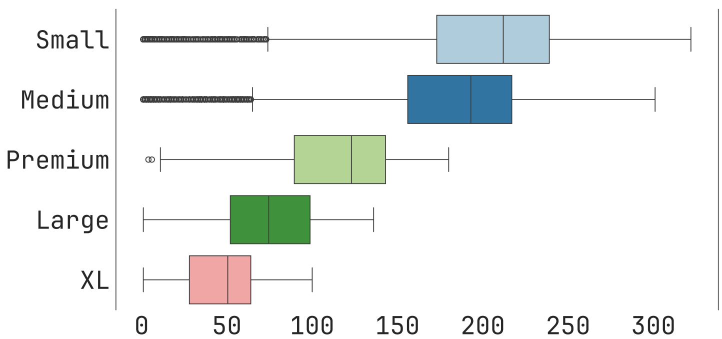


Figure 3.9: *Trips recorded per vehicle by size class*

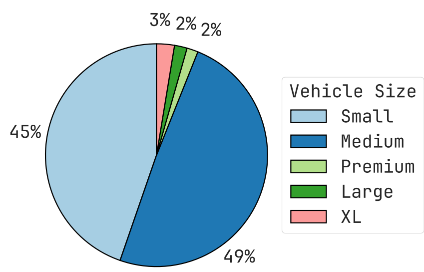


Figure 3.10: *Trips recorded per size class*

Figures 3.9 and 3.10 include large and XL size classes for reference

4
Predicting Demand

4.1 Methods

This experiment investigates whether the available data supports short- and long-term demand predictions for FFCS vehicles. Trip starts are interpreted as observed demand, and the influence of various contextual factors on this demand is analyzed.

4.1.1 Aggregation and Detrending

Previous studies on urban mobility consistently report distinct mobility patterns on workdays and weekends (Messa et al., 2021; Weikl et al., 2013). Reflecting this and the temporal structure illustrated in figure 4.2 on page 18, all trips were partitioned into 5 time bins for either day type: early morning (00:00—05:59), morning (06:00—09:59), midday (10:00—15:59), evening (16:00—19:59), and late evening (20:00—23:59). Traffic and weather data were aggregated into hourly intervals to align with this temporal division.

For spatial aggregation, I applied Uber's H3 hexagon grid, assigning all values to a standard spatial index. At resolution 8, each cell represents an area of approximately 0.65km^2, offering neighborhood scale and matching previous research (Casabianca et al., 2021). H3 is emerging as an industry standard for geospatial analytics and has successfully been applied in similar studies, such as Jo et al. (2024).

Both trip and POI counts are heavily influenced by the underlying urban density. To isolate deviations attributable to explanatory features, a one-predictor linear regression was applied to trips per cell and time bin, controlling for total daily trips by day type, before mean-centering for further analysis.

Similarly, POI counts were processed by regressing each category's count against the total POI count. The residuals represent irregular over- or underrepresentation of specific categories.

4.1.2 Spatial Clusters and Colocation

Local hot and cold spots for both trip starts and POI density were identified using the Getis-Ord G_i^* statistic (Getis et al., 1992), with inverse-distance weights applied to a three-ring neighborhood. The G_i^* statistic evaluates how much a cell's and its neighborhood's feature values deviate from the expected value to cluster high and low values. Focussing on clusters instead of isolated extreme cells mitigates the multiple comparisons problem and highlights areas of actual relevance.

POI categories are often strongly collinear, even if general urban activity has been accounted for. Redundant categories were identified using an l_1-penalised OLS regression (LASSO). Additionally, Local Moran's I was calculated to identify common colocations of unusual densities across POI categories.

4.1.3 Global and Local Effects

To evaluate city-wide associations between demand and potential predictors, Spearman correlations were calculated for residualized variables. Ordinary Least Squares (OLS) models were fitted to analyze the relationships between trips in different time bins and predictors such as traffic, weather, and POI categories. To account for potential non-linearities and better capture primary effects, the OLS results were compared against those from a generalized additive model (GAM).

While global coefficients are easy to interpret, they can hide spatial heterogeneity. To address this, geographically weighted regression (GWR) was applied to both the POI and traffic/weather feature sets. Bandwidths were selected using the corrected Akaike Information Criterion (AICc). The standard deviation of the GWR coefficients was used to determine whether an observed effect was global (city-wide) or local (neighborhood-specific). Additionally, the GWR's R^2 served as a meaningful measure of a predictor's explanatory relevance for the collected data.

4.2 Results

4.2.1 Clusters of Trip Starts

After accounting for general urban activity, the Getis-Ord G_i^* statistic reveals distinct spatiotemporal patterns in trip start counts, representing local demand for vehicles. In figure 4.3 on page 19, cells with G_i^*-scores above the 99% confidence level are highlighted as hot or cold spots, with those above the 95% confidence level are marked as warm or cool.

During early mornings on weekends, demand falls almost entirely below the residualized weekend baseline, aligning with intuitive expectations. On weekdays, this time bin shows hotspots in the outer region of the service area, while the central district exhibits below-expected demand.

As the day progresses, demand gradually migrates inward. By midday, the highest intensities are concentrated in and around the city center, possibly reflecting work- or lunch-related travel. This westward shift continues throughout the evening. Late evenings on weekdays display a more even surface overall, interrupted by a few hotspots in the city center and farther west.

Weekend dynamics differ substantially. Central hotspots reappear during the morning and evening bins, expanding into northern and western cells, yet the same areas flip to cold at midday. Overall, trip starts on weekends are more evenly distributed than on weekdays.

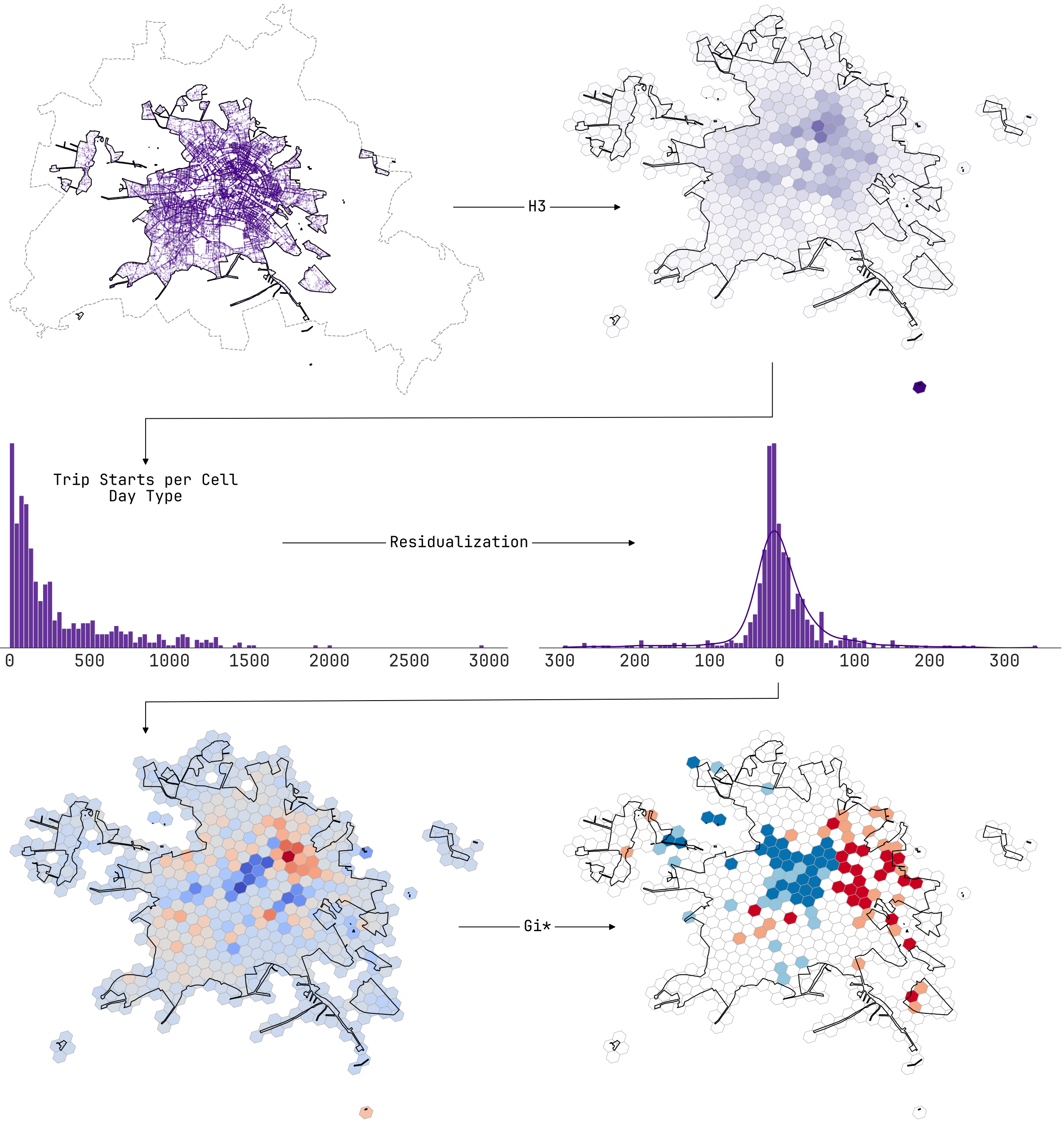


Figure 4.1: *POI analysis pipeline*

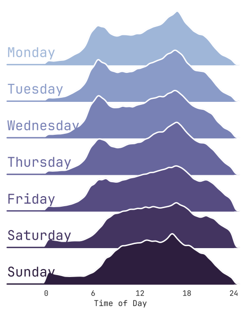


Figure 4.2: *Temporal distribution of trip starts by day of the week*

4.2.2 Clusters of Points of Interest

Applying the same G_i^* procedure to residualised POI counts highlights the diverse land use across Berlin (figure 4.4 on the following page). Accommodation and restaurant venues form the most prominent hot cluster in the central district. This area also concentrates transit POIs, stretching from the central station through Alexanderplatz and north to Gesundbrunnen.

Shops, in contrast, expand much farther beyond the central district, appearing as hotspots across nearly half the service area. Their only notable cold spot is in the southwest, where bars, entertainment, and health institutions form hot clusters. While not immediately intuitive, this reflects how G_i^* identifies unusual values within neighborhoods, not city-wide absolutes.

Education, religion, and public service facilities are evenly distributed and show up as largely neutral. Culture and finance, in contrast, show the opposite extreme. In contrast, culture and finance-related POIs are highly concentrated. Cultural POIs cluster tightly around the city center and Gesundbrunnen, while finance-related POIs form a corridor between the center, Neukölln, and adjacent districts. These elevated local means push most other areas into the cold category.

Finally, the nature category appears nearly homogeneous. Nonetheless, the algorithm marks a cold spot close to the Grunewald forest and a hot patch in Lichtenberg, highlighting a discrepancy in how the POI dataset represents reality.

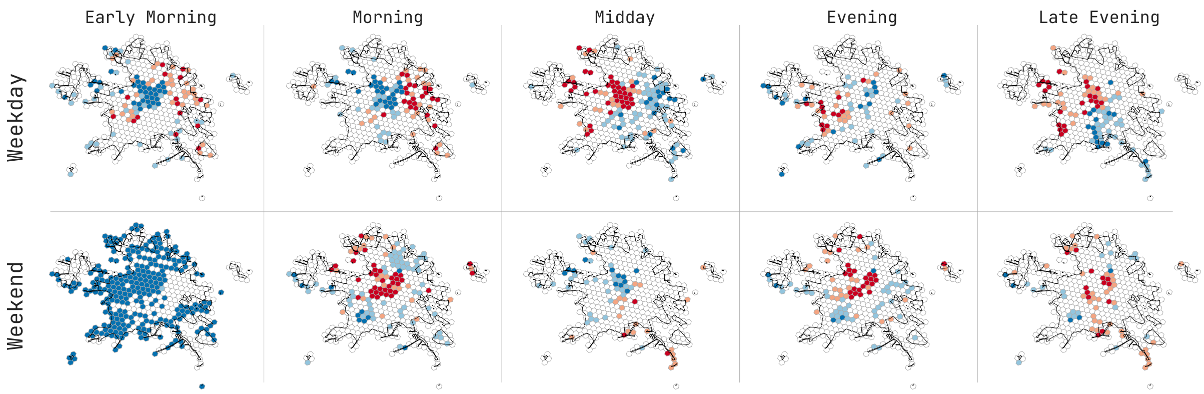


Figure 4.3: *G_i^* hotspots for trip starts in each time bin*

4.2.3 Point of Interest Categories

A city-wide LASSO regression reveals significant collinearity between some POI categories. The strongest correlation is observed between health and shopping, which is notable, given that the hotspot analysis did not identify overlapping clusters for these categories. This does not explain, however, whether this reflects pharmacies and medical practices clustering near retail corridors or simply a shared prevalence of both categories in lower-density neighborhoods.

Classic commercial correlations also emerge, such as strong associations between food and drink, as well as shopping and accommodation with food, suggesting mixed shopping-nightlife districts where hotels are likewise overrepresented. Recreation facilities frequently appear near schools and universities, while finance POIs draw higher concentrations of nearby transit stops. Conversely, transit clusters are less likely to overlap with bar districts, and health POIs are uncommon in restaurant-dense cells.

Local Moran's I paints a similar picture at the neighborhood scale. High-high clusters of restaurants and bars dominate the center and Prenzlauer Berg, although areas exist where only one of the two categories is unusually common. The recreation-education colocation suggested by the LASSO results is also evident on smaller scales.

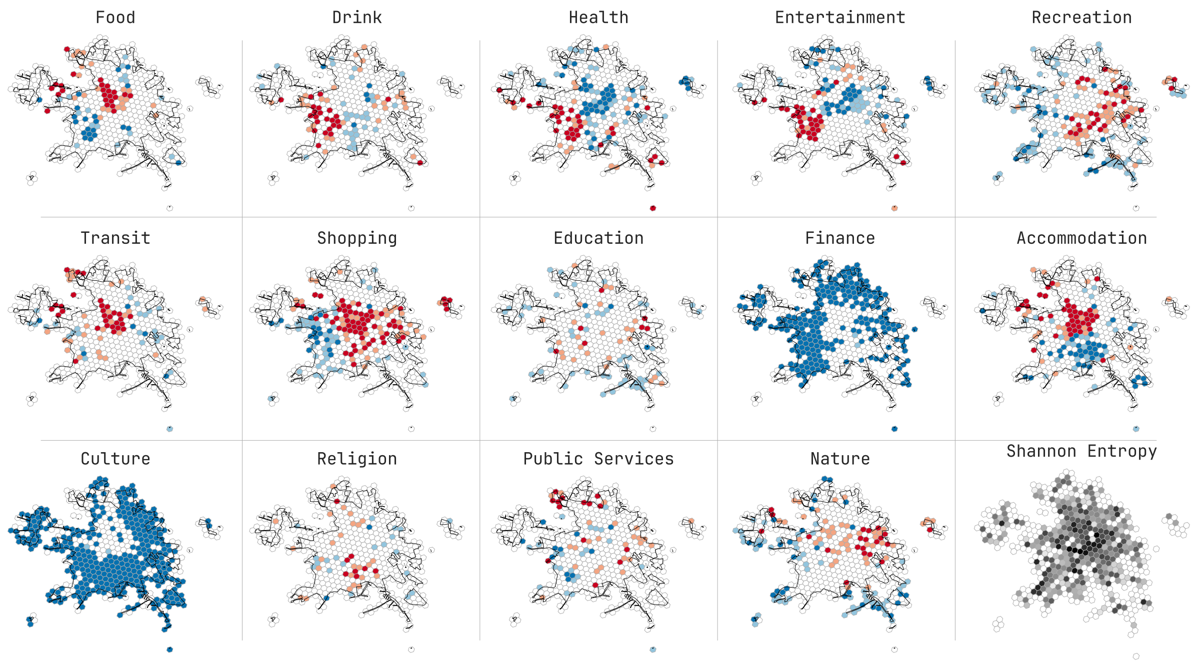


Figure 4.4: *G_i^* hotspots for POIs by category*

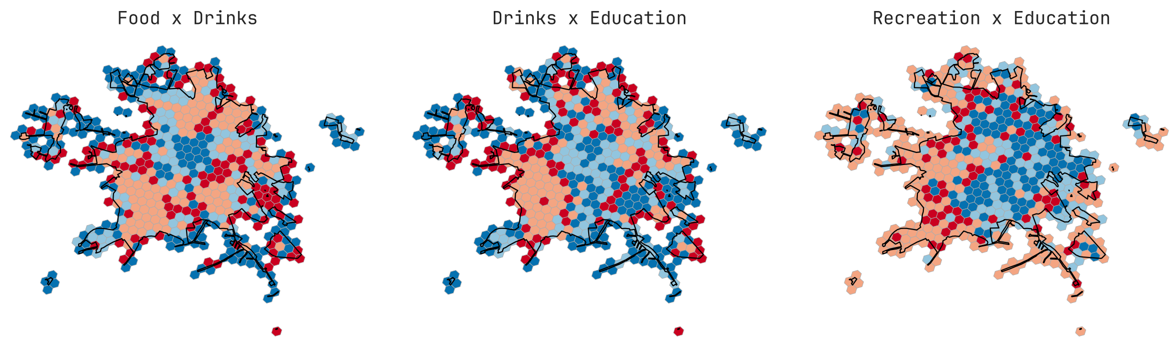


Figure 4.5: *Locan Moran's I for select POI category combinations*

4.2.4 Long-Term Demand

Basic Pearson correlations show that the strength and sign of POI-demand relationships vary throughout the day.

In the early-morning weekday bin, education exhibits the strongest positive correlation, whereas accommodation, culture, and food show adverse effects. This could represent people leaving residential districts with more schools and fewer hotels or museums. This pattern persists into the morning rush hour, but by midday, accommodation and culture turn positive. At this point, transit stops and restaurants also contribute positively, while education shifts to a negative influence on demand.

In the evening, when the G_i^* analysis shows that trip starts are more spatially balanced, POI influence fades. After 20:00, however, bars emerge as the dominant factor every weekday, showing slightly weaker effects on Thursdays but remaining consistent on Fridays.

On weekends, in the early mornings, bars contribute positively, with similar patterns to weekdays by morning. In the late evening, weekend trips increasingly start in areas with leisure-oriented POIs, such as restaurants, bars, and entertainment. While these relationships are not as strong as some others, they counteract the negative impacts of categories like health, recreation, and education during the same period.

Applying GWR reveals the spatial texture of these effects. Most categories exhibit uniform behavior in the core of the service area and vary only in the outskirts. However, a few notable exceptions emerge.

The late-evening bar effect, while positive city-wide, is much stronger in eastern Berlin. On weekdays, accommodation has a nearly global positive impact at midday but reverses to a negative influence in the early morning. During this time, areas near the airport signal a slight positive effect, contrary to their negative impact at noon. Interestingly, this area often opposes general city-wide effects, such as food and entertainment.

While some categories, like restaurants, are much more homogenous, transit locations stand out for the commuting patterns they reveal. On weekday mornings, transit suppresses demand in the center and east but boosts it in the west, suggesting that commuters in these areas use FFCS vehicles, while those in the inner city walk or cycle to public transport. By midday, this pattern inverts, showing a strong positive influence throughout the city, except in areas with positive early-morning correlations. By late evening, the positive influence shifts further west, beginning to transition north and back to the early-morning pattern. On weekends, transit effects are weaker and less structured, though the east retains its midday positive influence.

The standard deviation of local GWR coefficients confirms the localized nature of transit effects, whereas categories like food, drink, health, and shopping behave more globally. Category-wise R^2 values confirm the Pearson correlation findings, reiterating the varying relevance of different POI categories throughout the day.

Shopping locations present a high relevance but highly local effects. Conversely, education has a strong and relatively global impact during early weekday mornings. High individual R^2 scores, exceeding 40%, underscore the collinearity observed earlier. In a multivariate GWR, all POI categories combined explain more than 50% of weekday demand deviance (except during the evening bin) and 23-34% on weekends. This illustrates both the power and the limitations of this static context for modeling demand.

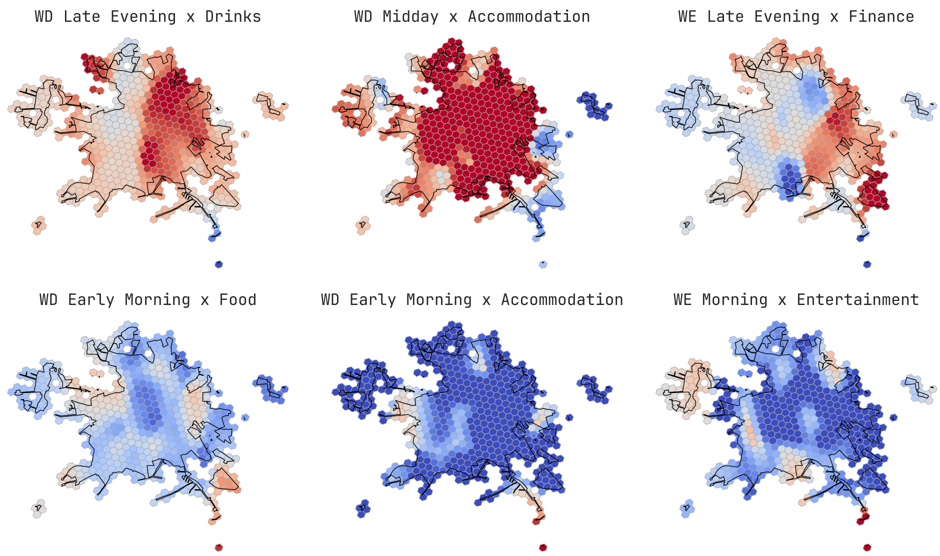


Figure 4.6: *GWR local effects for select POI and time bin combinations*

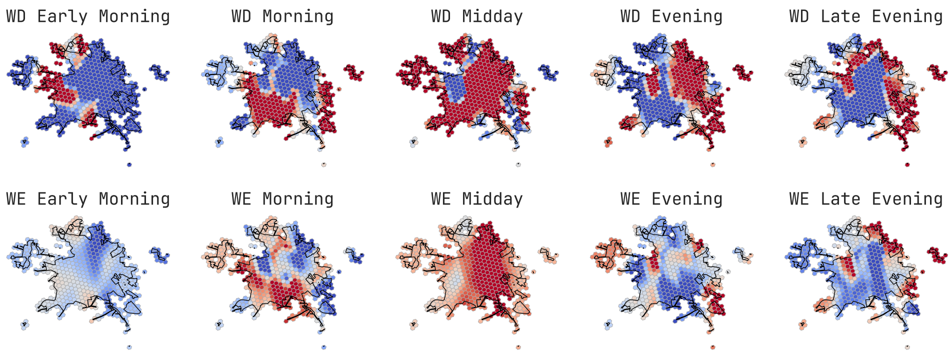


Figure 4.7: *GWR local effects for transit POI and time bin combinations*

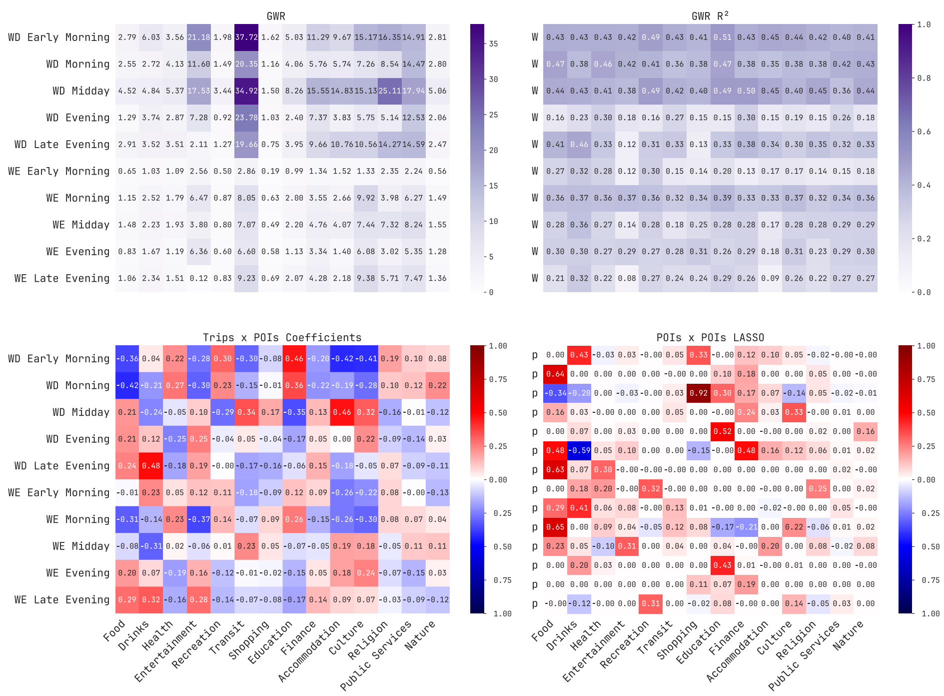


Figure 4.8: *Trip and POI category correlation results*

4.2.5 Short-Term Demand

For traffic and weather data, simple correlations did not reveal meaningful signals, suggesting that potential effects may be obscured by spatial heterogeneity or non-linearity. The city-wide OLS (table C.1 on page 43), however, confirms their very low practical relevance, explaining only 0.7% of the deviation in trip start counts, with weather factors alone contributing to less than half of this. The GAM results further eliminate non-linearity as a significant factor, as they show minimal improvement over the OLS. Similarly, the GWR analysis eliminates spatial heterogeneity, as effects are globally negligible.

The slight yet statistically significant effects, amplified by the large sample size, largely conform to intuition. Traffic flow follows an inverted U-curve, where less congestion correlates with more demand, but no congestion has a negative effect, possibly reflecting reduced activity in unusually less busy areas. Temperature is mildly U-shaped, with middling values suppressing demand, while heat increases it slightly more than low temperatures, perhaps due to mode shifts away from cycling or public transport.

Light rain shows a marginally positive correlation with demand, whereas heavy rain reverses this effect. However, the sample size for heavy rainfall is too small to draw definitive conclusions. For changes in precipitation, the two models diverge. The GAM suggests a drop in demand when rain begins and a rebound when it stops, while the OLS finds a weak overall decline that is only significant at large deltas.

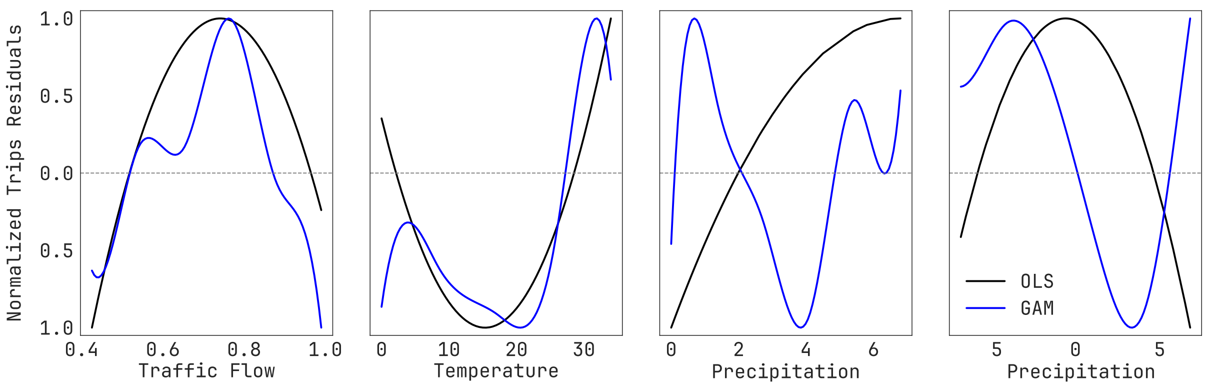


Figure 4.9: *Comparison of OLS and GAM models for traffic, temperature, and precipitation*

4.3 Discussion

In the first experiment, I analyzed the relationships between the trip start counts, different POI categories, and traffic and weather influences, identifying recurring patterns. While I did not attempt to predict future demand, the results demonstrated that POIs play a significant role in where demand exists at different times. These findings align with much of the existing literature, including (Schmöller et al., 2015) and (Wagner et al., 2015). However, contrary to Schmöller et al., I was unable to identify a practically relevant influence of precipitation changes on FFCS demand.

Changes in precipitation, traffic, and additional weather variables proved to be statistically

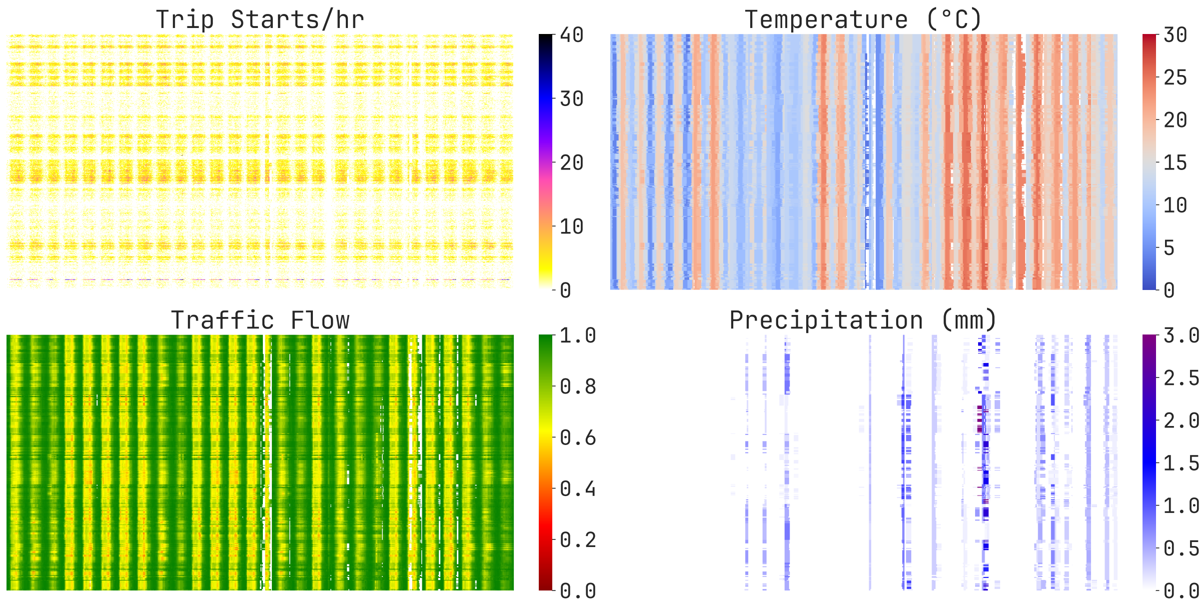


Figure 4.10: *Dynamic factors between May 9th and June 9th, 2025*

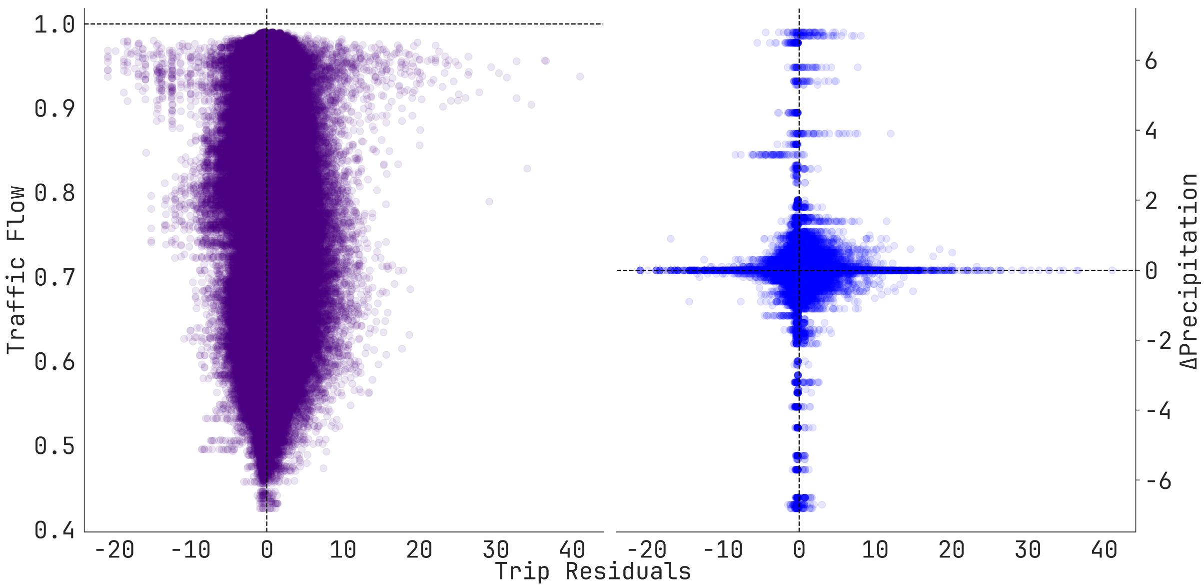


Figure 4.11: *Traffic and precipitation against trip starts*

significant yet practically irrelevant. With a combined explanatory power of less than 1%, these factors offer little value in predicting dynamic demand. Additional, potentially relevant predictors such as events, transit outages, school holidays, or real-time vehicle reservation rates exist but were not included in the thesis dataset. Furthermore, live user application data, as utilized by (Wang et al., 2019) and (Liu et al., 2021), has been shown to be highly valuable in similar studies but was unavailable for this analysis.

Based on the dataset collected for this thesis, the first part of my hypothesis is rejected. It is not possible to forecast vehicle demand more meaningfully than static patterns using the available data. Short-term demand predictions would require additional predictors, although the literature suggests that meaningful forecasting is feasible with the right data.

There are also methodological limitations to my approach. POI reflect only the count of locations, not their relevance. For example, a local pharmacy is treated equivalently to a clinic, and Tempelhof Field is considered the same as Viktoria-Luise-Platz. This clearly does not reflect reality. While I identified significant collinearities between POI categories, I did not address them using techniques such as principal component analysis or SHAP profiles for the LASSO regression. My results show that POIs significantly impact observed demand, but could have been more consequential for land-use types like those determined by (Klemmer et al., 2016).

Data limitations also extend to weather factors. The dataset contained only a small number of days with heavy rain and none with extreme heat. Furthermore, while I detrended trip counts, it might have been valuable to use alternative detrending bases, such as population or employment census data, which were not part of the thesis dataset.

Finally, trip start points are not entirely independent, as they depend on the destination of the previous trip (Willing et al., 2017). My approach treated demand as the number of trip starts without accounting for unfulfilled demand caused by vehicle unavailability. (Schmöller et al., 2015) addressed this by clustering days and regions into different demand categories, which then served as a baseline for required supply. While the observed fleet's size might render this issue negligible, it could still have influenced my results.

5

Predicting Trip Destinations

5.1 Methods

The second experiment attempts to predict trip destinations using common models to determine if the dataset allows for accurate forecasting. Three models were compared against a baseline and theoretical maximum: a statistical multinomial logit, a gradient-boosted decision tree, and a neural network.

5.1.1 Data Exploration and Preparation

The service area contains 535 resolution-8 cells, but initial tests showed that some models became computationally expensive or unstable due to sparse data in the outer cells. To address this, the dataset was downsampled to 109 classes at resolution 7, where each cell has an average edge length of approximately 1320 meters. Additionally, trips to and from Berlin's airport were excluded from the analysis after determining that these trips differed significantly from other system-wide trips.

To identify macro-patterns before model fitting, origin-destination flows, median trip distances, and prevailing travel directions were mapped for each time bin. Residualized destination counts were analyzed using the Getis-Ord G_i^* statistic (analogous to section 4.1.2 on page 16) to compare hotspots in trip starts and ends.

The dataset was split into a twelve-week training window and a two-week hold-out window, representing approximately 13% of the records. For the baseline and logit model, context combinations that did not occur in the training set were removed to mitigate inaccuracies caused by an insufficient observation duration. Laplace smoothing was applied to probabilities to reduce noise in rare contexts and improve model robustness.

5.1.2 Baselines

To assess the determinability of the classification problem and contextualize the prediction results, I established a range for potential improvement. The lower bound was defined by a majority-vote classifier, which always predicts the most frequent destination for a given context.

The upper bound Π_{max} was estimated using the Bayes error rate, representing the highest theoretically achievable prediction accuracy for a model trained on the available context set in the dataset (equation (D) on page 45).

For additional insight, I also calculated the theoretical maximum predictability for each context when considering human mobility (Π^*). This approach, introduced by Song et al. (2010), was based on studying phone users' mobility patterns. While this upper bound is only valid for trips associated with individual users and when user history is known, it offers a perspective on the potential impact of user-specific correlations on prediction accuracy.

Each improvement range was defined using a subset of the available features: starting cell, hour of the week, applied discounts, reservations, traffic, temperature, precipitation, POI-based land use, and dominant POI category. Traffic and weather variables were discretized into three categories, and six land-use clusters were derived from POI category residuals using PCA.

5.1.3 Models

While trajectory-based approaches have proven to be highly accurate, they are not applicable in this experiment, as the trip's destination must be predicted before the trip begins. Therefore, the problem was framed as a classification task, with three common approaches applied to calculate accuracies for the top 1 through top 3 predictions.

In theory, the majority-vote baseline would converge into Π_{max} if applied to an infinitely large and perfectly representative dataset. However, real-world datasets, including the one used in this study, are finite and inherently skewed. Observations are more frequent in the core of the service area and less common during extreme weather, periods of congestion, or relevant events.

To close this gap, I fitted three models to infer relationships between context variables. Categorical features were encoded as integers, and numerical features were standardized based on the training set. Model performance was evaluated using top-1 accuracy and the ROC AUC for each class.

The first model applied was a multinomial logit for its transparency in revealing cross-context relationships. This model provides straightforward insights into the linear dependencies between predictors and outcomes, making it useful for comparison.

To capture more complex relationships between context variables, I used a gradient-boosted decision tree (CatBoost) optimized through grid search-tuned hyperparameters. This algorithm improved the modeling of potentially complex context relationships between predictors and outcomes.

Finally, I implemented a recurrent neural network (RNN) to model the spatiotemporal data. Specifically, I used a gated recurrent unit (GRU) with a simple attention layer to focus on the most relevant features during prediction. RNNs have been identified as the best-performing method for demand prediction by Schreckenberger et al. (2018) and have shown significant potential for next-place predictions in the work of Al-Molegi et al. (2016).

5.2 Results

The initial exploration revealed that destination points are more concentrated than start points, as determined by local residuals. However, the overall demand patterns for pick-ups and drop-offs are similar across most time bins. Trip distances are homogeneous throughout the city, with a slight decrease observed in Friedrichshain, and show no meaningful variation over time. Trips started on the outskirts of the service area cover longer distances. The average trip direction exhibits a clear west-east orientation with a slight southward trend. This pattern persists for shorter trips, albeit less prominently, with a localized version of the same trend emerging in Spandau.

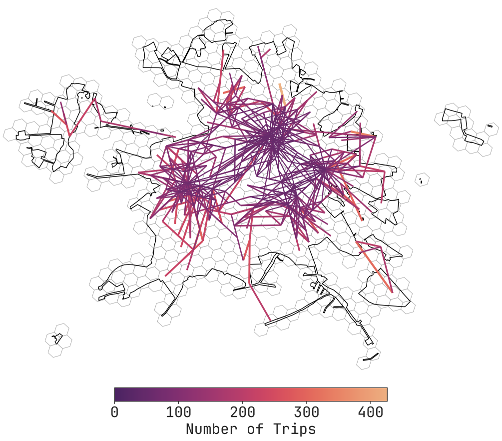


Figure 5.1: *3 most common trips per starting cell with 50+ observations at resolution 8*

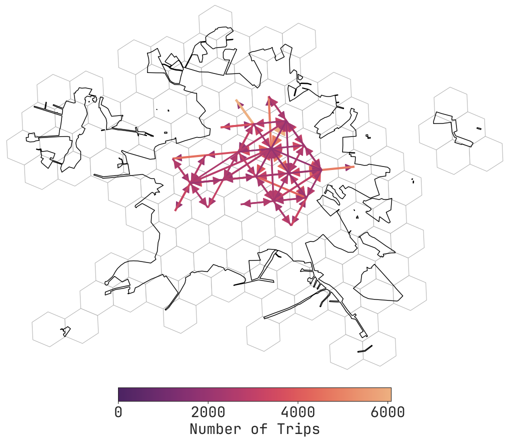


Figure 5.2: *100 most common trips in Berlin at resolution 7*

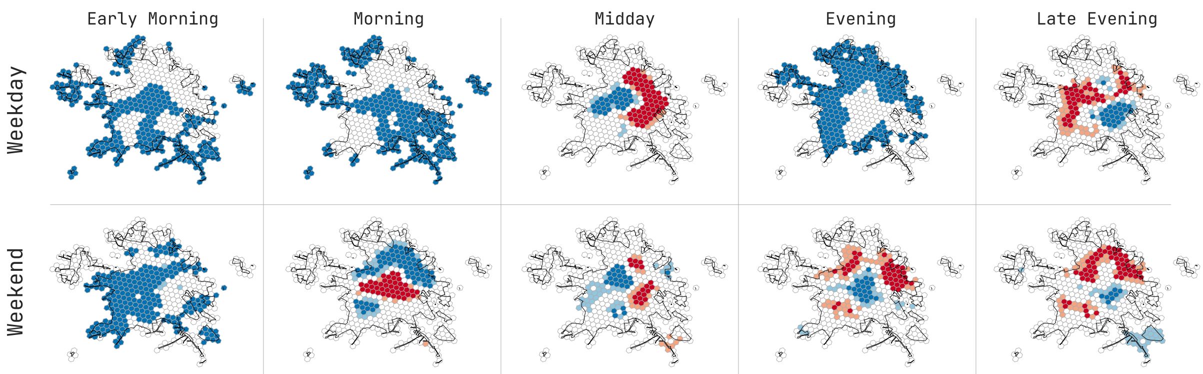


Figure 5.3: *Gi* hotspots for trip destinations in each time bin*

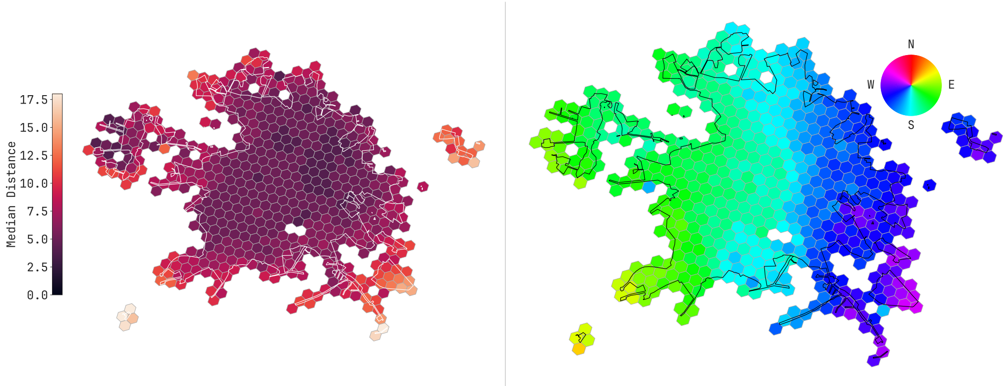


Figure 5.4: *Median distance and average direction of trips*

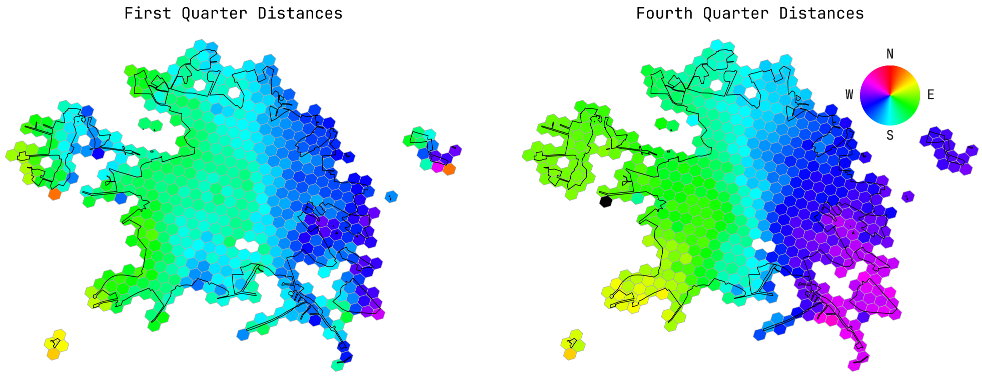


Figure 5.5: *Average direction of trips with different lengths*

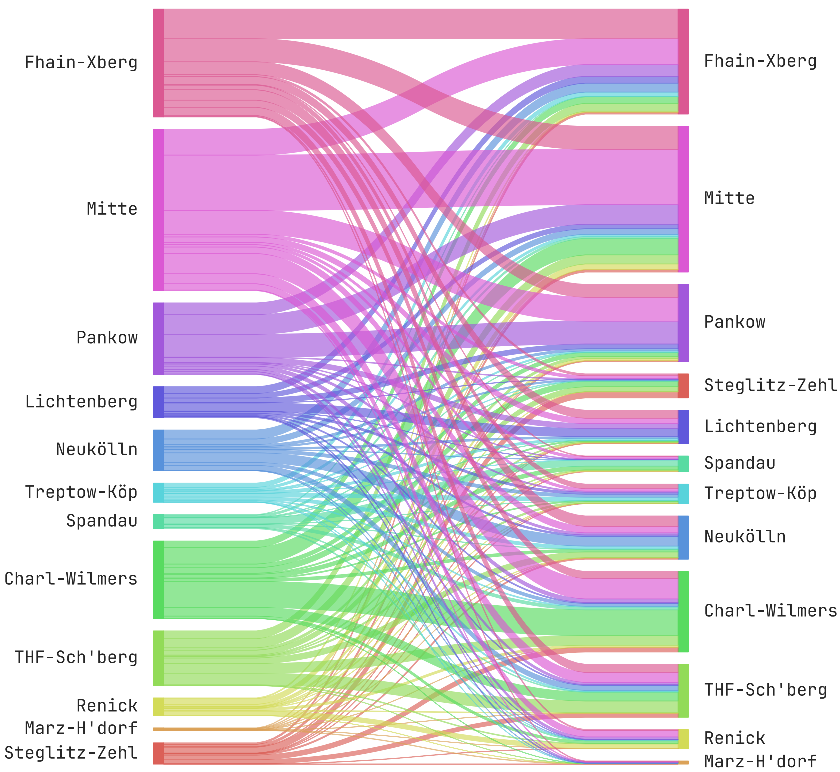


Figure 5.6: *Trip flows between districts*

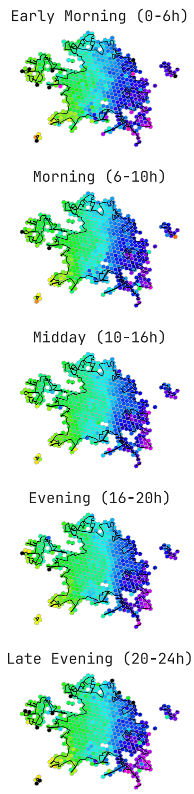


Figure 5.7: *Average direction of trips on weekdays*

5.2.1 Baseline

Using only the starting location and hour of the week as predictors establishes a theoretical maximum accuracy of 11.7% and a top-1 baseline of 7.9%. While this baseline is already far better than random guessing, $\Pi_{\max}$ does not leave much room for improvement.

Unfortunately, and consistent with the findings of the first experiment, adding additional context only increases the theoretical maximum to 15.2% top-1 accuracy. This demonstrates that, even with the proposed predictors, the classification problem is not deterministic. Moreover, the majority baseline accuracy decreases when additional context is introduced. This is due to smaller observation sizes per context, which flatten the smoothed counts toward uniformity.

Removing either the start location or time from the context confirms what the entropy analysis implies. These two predictors are by far the most relevant. In contrast, neither of the POI-based predictors improves the majority baseline or the theoretical maximum. This suggests that any contextual information they provide is already encoded in the starting location cell.

5.2.2 Logit

Fitting the multinomial logit model using the starting cell and time, as well as traffic and weather context, improves upon the baseline by roughly 35%, or three percentage points. This improvement highlights the model's ability to infer relationships in the sparser dataset as more variables are added. However, it only marginally improves upon the most effective baseline variant, suggesting the model is unable to benefit from the additional context effectively.

The results confirm that, without the starting cell, the remaining predictors provide little to no value. Even with generous regularization, the model drives most non-spatial coefficients to zero. Weather and traffic data are even less informative for predicting trip destinations than they were for predicting trip starts in the first experiment.

5.2.3 Gradient-Boosting

An initial, uncalibrated run of the CatBoost gradient-boosting algorithm yielded lower results than the logit model. Even after carefully tuning hyperparameters, the resulting accuracy gains were minimal. Encoding historical k-fold origin/destination (O/D) probabilities reasonably improved the model's performance, but ROC curve analysis revealed that this process inadvertently leaked future knowledge into the test set.

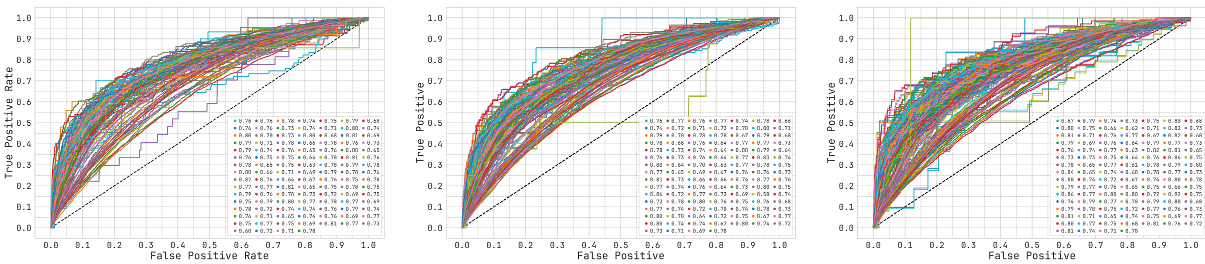
Without access to data that is unattainable in real-world scenarios, the gradient-boosting algorithm does not outperform the logit regression. This suggests that complex interactions between the proposed predictors do not drive the trips in the dataset. Analyzing the model's feature importance reveals that starting cell and time account for more than 90% of the prediction performance.

5.2.4 Recurrent Neural Network

Similarly, the GRU model failed to outperform the previous approaches when trained on 15-minute snapshots over 4-hour windows. Encoding additional spatial information, such as the start point's distance from the center of Berlin's carsharing activity or its parent cell at a lower resolution, did not yield significant improvements. Adding an attention layer to capture better the progression of system state over 24-hour windows slightly improved accuracy, but the GRU still barely surpassed the previous approaches.

After careful configuration, the results show no evidence of significant overfitting. However, the GRU does not extract more value from traffic and weather data than the minimal insights provided by the logit or CatBoost models. This suggests either that the dataset lacks sufficient information to leverage advanced algorithms or that the proposed predictors offer no additional value.

These findings align with the initial exploration. However, possible destinations are almost purely predicted based on starting location and time, which is why all models barely improve upon the baseline. However, unlike the majority-vote baseline, models incorporating cross-influences perform better in previously unseen contexts. This is evident in the consistent performance of both CatBoost and GRU when applied to all test contexts (chapter D on page 45).



Logit CatBoost GRU + Attention

Figure 5.8: *ROC curves per tested model at resolution 7.*

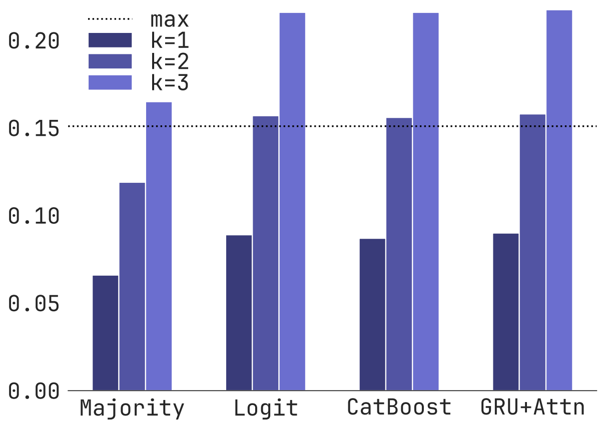


Figure 5.9: *Comparison of destination prediction model accuracies*

5.3 Discussion

I trained three classification models of increasing complexity to address the next-place prediction problem: a multinomial logit, gradient-boosted decision trees, and a compact neural network. While these models improved measurably upon the baseline majority-vote, their performance was nearly indistinguishable from one another. None of the models achieved a level of accuracy that would justify deployment.

Even after excluding trips to and from the airport and removing holiday data (section 3.2.3 on page 12), the theoretical ceiling set by context entropy remains at 15% — a score much too low to base discounts for implicit user relocations on effectively.

These results are consistent with the findings from chapter 4 on page 16. Dynamic influences, such as traffic, temperature, and precipitation, provide minimal predictive value. Consequently, the models rely almost exclusively on starting location and time, both of which are already well explained by the baseline prediction.

Unlike the first experiment, the spatial dimension of the destination prediction only aimed to understand if different locations are popular rather than why they are popular. Therefore, adding POI information did not improve predictions, as this information was already encoded in the starting cell context. Understanding the reasons for cell popularity would be relevant when applying such models to a different city (Willing et al., 2017). However, the proposed land-use and dominant category indicators were insufficient to replace the starting cell context.

Several opportunities for improvement emerge from these findings. Analogous to chapter 4 on page 16, incorporating additional context data, such as major events, transit disruptions, or unplanned construction, could benefit prediction accuracy. The dataset may also not be representative of factors such as extreme weather and too noisy in such instances. The deep learning approach, in particular, might further benefit from a more extensive observation period to extract meaningful patterns. Additionally, the dataset contains features that were not included in this analysis, such as the difference between *premium* and *standard* vehicles, which could be investigated further.

One context that could dramatically improve predictions is user identification and history. My calculations, following the approach by Song et al. (2010), suggest that accuracy could increase to as much as 45% if anonymous user data were added to the analyzed context. Similarly, Gambs et al. (2012) and Al-Molegi et al. (2016) have demonstrated the high predictability of human mobility patterns. Liu et al. (2021) further showed that knowledge of frequent user locations can improve trip extrapolation accuracy, which could apply to next-place predictions as well.

Based on the results of this experiment, however, the second part of the hypothesis is rejected. With the dataset collected for this thesis and the proposed predictors, it is not possible to accurately predict a trip's destination before it begins. While user history could significantly improve predictions, it is unlikely to achieve an accuracy level suitable for practical deployment unless more relevant predictors are provided.

6
Conclusion

The two experiments conducted in this thesis approached the hypothesis from different sides. The first explained short- and long-term demand patterns using exploratory methods, while the second applied three different classification strategies to predict likely carsharing trip destinations. Both experiments yielded the same result: trip metadata, POIs, traffic, and weather data are insufficient to model human behavior within a carsharing system.

The thesis demonstrated that traffic, temperature, and precipitation have minimal influence on both demand and usage for FFCS. While the results limit the feasibility of short-term demand estimation based on the analyzed dataset, they also reveal significant patterns in long-term demand. These patterns were shown to be largely explained by differences in nearby points of interest (POIs). This finding, despite limited by the unresolved collinearity of categories, can inform strategic decisions such as service area design.

When predicting individual trip destinations, the limitations become more pronounced. None of the three tested models (multinomial logit, CatBoost, and GRU with attention) achieved accuracy levels sufficient for operational deployment. Without user data or richer dynamic context, the theoretical upper bound for prediction accuracy tops out at just 15%, far below what would be needed for implicit user-based relocation incentives.

The hypothesis that predictive models could accurately forecast demand and trip destinations within Berlin's FFCS was thoroughly tested but ultimately not supported. Instead, this thesis highlights the need for richer datasets to enable more accurate real-time predictions and provides insights into long-term demand patterns.

6.1 Future Work

The challenges explored in this thesis remain relevant for optimizing carsharing systems. The literature consistently emphasizes that user-based relocations are more effective than operator-based strategies and that users can be incentivized to choose alternative vehicles or parking locations (Herrmann et al., 2014; Lippoldt et al., 2019). However, realizing the full potential of these strategies requires further research.

Future studies should incorporate additional data, particularly user-specific information, such as personal location history and app interaction rates. System metadata, including changes in reservation behavior and external factors, such as transit disruptions, major events, or city-wide activities, could also enhance the predictive power of demand and destination models.

Furthermore, exploring these methods in different urban contexts could provide additional insights. The literature indicates that carsharing systems are used differently across cities, and this research suggests that Berlin's patterns may be more complex than those reported elsewhere. Understanding these differences could improve the generalizability of predictive models.

By leveraging richer datasets, future research may achieve prediction accuracies suitable for operational deployment. Such advancements would not only improve the efficiency of carsharing systems but also contribute to more sustainable urban mobility.

References

Al-Molegi, Abdulrahman et al. (2016). "STF-RNN: Space Time Features-based Recurrent Neural Network for Predicting People next Location". In: *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2016 IEEE Symposium Series on Computational Intelligence (SSCI). Athens, Greece: IEEE, pp. 1–7. ISBN: 978-1-5090-4240-1. DOI: 10.1109/SSCI.2016.7849919. URL: http://ieeexplore.ieee.org/document/7849919/ (visited on 06/19/2025).

Ampudia-Renuncio, María et al. (2020a). "Electric Free-Floating Carsharing for Sustainable Cities: Characterization of Frequent Trip Profiles Using Acquired Rental Data". In: *Sustainability* 12.3, p. 1248. ISSN: 2071-1050. DOI: 10.3390/su12031248.

Ampudia-Renuncio, María et al. (2020b). "Understanding the Spatial Distribution of Free-Floating Carsharing in Cities: Analysis of the New Madrid Experience through a Web-Based Platform". In: *Cities* 98, p. 102593. ISSN: 02642751. DOI: 10.1016/j.cities.2019.102593.

Archetti, Claudia et al. (2023). "The Benefit of Complete Trip Information in Free-Floating Carsharing Systems". In: *EURO Journal on Transportation and Logistics* 12, p. 100110. ISSN: 2192-4376. DOI: 10.1016/j.ejtl.2023.100110.

Balac, Milos et al. (2016). "Carsharing Demand Estimation: Zurich, Switzerland, Area Case Study". In: *Transportation Research Record: Journal of the Transportation Research Board* 2563.1, pp. 10–18. ISSN: 0361-1981, 2169-4052. DOI: 10.3141/2536-02.

Barth, Matthew et al. (2004). "User-Based Vehicle Relocation Techniques for Multiple-Station Shared-Use Vehicle Systems". In: *Transportation Research Record: Journal of the Transportation Research Board* 1887.1, pp. 137–144. ISSN: 0361-1981. DOI: 10.3141/1887-16.

BCS (2024a). *Carsharing Users by Variant*. Bundesverband Carsharing, provided by Statista. URL: https://de.statista.com/statistik/daten/studie/202416/umfrage/entwicklung-der-carsharing-nutzer-in-deutschland/ (visited on 06/17/2025).

BCS (2024b). *Number of car sharing vehicles in selected cities in Germany in 2024*. Bundesverband Carsharing, provided by Statista. URL: https://www.statista.com/statistics/1092513/car-sharing-number-of-vehicles-in-german-cities/ (visited on 06/22/2025).

Becker, Henrik et al. (2017). "Modeling Free-Floating Car-Sharing Use in Switzerland: A Spatial Regression and Conditional Logit Approach". In: *Transportation Research Part C: Emerging Technologies* 81, pp. 286–299. ISSN: 0968090X. DOI: 10.1016/j.trc.2017.06.008.

Besse, Philippe C. et al. (2016). *Destination Prediction by Trajectory Distribution Based Model*. DOI: 10.48550/arXiv.1605.03027. arXiv: 1605.03027 [stat]. URL: http://arxiv.org/abs/1605.03027 (visited on 06/11/2025). Pre-published.

Caggiani, Leonardo and Michele Ottomanelli (2013). "A Dynamic Simulation Based Model for Optimal Fleet Repositioning in Bike-sharing Systems". In: *Procedia - Social and Behavioral Sciences* 87, pp. 203–210. ISSN: 18770428. DOI: 10.1016/j.sbspro.2013.10.604.

Carrone, Andrea Papu et al. (2024). "Analysis of Car Sharing Operation Area Performance: An Idle Time Prediction Approach". In: *Transportation Research Part A: Policy and Practice* 190, p. 104241. ISSN: 0965-8564. DOI: 10.1016/j.tra.2024.104241.

Casabianca, Pietro et al. (2021). "Vehicle Destination Prediction Using Bidirectional LSTM with Attention Mechanism". In: *Sensors (Basel, Switzerland)* 21.24, p. 8443. ISSN: 1424-8220. DOI: 10.3390/s21248443. PMID: 34960532.

Celsor, Christine and Adam Millard-Ball (2007). "Where Does Carsharing Work?: Using Geographic Information Systems to Assess Market Potential". In: *Transportation Research Record: Journal of the Transportation Research Board* 1992.1, pp. 61–69. ISSN: 0361-1981, 2169-4052. DOI: 10.3141/1992-08.

Chekol, Ayele Gobezie and Marta Sintayehu Fufa (2022). "A Survey on next Location Prediction Techniques, Applications, and Challenges". In: *EURASIP Journal on Wireless Communications and Networking* 2022.1, p. 29. ISSN: 1687-1499. DOI: 10.1186/s13638-022-02114-6.

Chun, Yoon-Young et al. (2019). "Exploring Factors Affecting Car Sharing Use Intention in the Southeast-Asia Region: A Case Study in Java, Indonesia". In: *Sustainability* 11.18 (18), p. 5103. ISSN: 2071-1050. DOI: 10.3390/su11185103.

Ciari, Francesco et al. (2014). "Modeling Station-Based and Free-Floating Carsharing Demand: Test Case Study for Berlin". In: *Transportation Research Record: Journal of the Transportation Research Board* 2416.1, pp. 37–47. ISSN: 0361-1981, 2169-4052. DOI: 10.3141/2416-05.

Citi Bike System Data (2025). City Bike NYC. URL: https://citibikenyc.com/system-data (visited on 06/20/2025).

Clemente, Monica et al. (2018). "A Decision Support System for User-Based Vehicle Relocation in Car Sharing Systems". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48.8, pp. 1283–1296. ISSN: 2168-2216, 2168-2232. DOI: 10.1109/TSMC.2017.2665881.

Cocca, Michele et al. (2020). "On Car-Sharing Usage Prediction with Open Socio-Demographic Data". In: *Electronics* 9.1 (1), p. 72. ISSN: 2079-9292. DOI: 10.3390/electronics9010072.

Comito, Carmela (2020). "NexT: A Framework for next-Place Prediction on Location Based Social Networks". In: *Knowledge-Based Systems* 204, p. 106205. ISSN: 09507051. DOI: 10.1016/j.knosys.2020.106205.

Di Febbraro, Angela et al. (2012). "One-Way Carsharing: Solving the Relocation Problem". In: *Transportation Research Record: Journal of the Transportation Research Board* 2319.1, pp. 113–120. ISSN: 0361-1981. DOI: 10.3141/2319-13.

Di Febbraro, Angela et al. (2019). "One-Way Car-Sharing Profit Maximization by Means of User-Based Vehicle Relocation". In: *IEEE Transactions on Intelligent Transportation Systems* 20.2, pp. 628–641. ISSN: 1524-9050, 1558-0016. DOI: 10.1109/TITS.2018.2824119.

Firnkorn, Jörg and Martin Müller (2011). "What Will Be the Environmental Effects of New Free-Floating Car-Sharing Systems? The Case of Car2go in Ulm". In: *Ecological Economics* 70.8, pp. 1519–1528. ISSN: 09218009. DOI: 10.1016/j.ecolecon.2011.03.014.

Formentin, Simone et al. (2015). "On the Prediction of Future Vehicle Locations in Free-Floating Car Sharing Systems". In: *2015 IEEE Intelligent Vehicles Symposium (IV)*. 2015 IEEE Intelligent Vehicles Symposium (IV). Seoul, South Korea: IEEE, pp. 1006–1011. ISBN: 978-1-4673-7266-4. DOI: 10.1109/IVS.2015.7225816. URL: http://ieeexplore.ieee.org/document/7225816/ (visited on 06/18/2025).

Gambs, Sébastien et al. (2012). "Next Place Prediction Using Mobility Markov Chains". In: *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*. MPM '12. New York, NY, USA: Association for Computing Machinery, pp. 1–6. ISBN: 978-1-4503-1163-2. DOI: 10.1145/2181196.2181199. URL: https://dl.acm.org/doi/10.1145/2181196.2181199 (visited on 06/18/2025).

Getis, Arthur and J. K. Ord (1992). "The Analysis of Spatial Association by Use of Distance Statistics". In: *Geographical Analysis* 24.3, pp. 189–206. ISSN: 1538-4632. DOI: 10.1111/j.1538-4632.1992.tb00261.x.

Herrmann, Sascha et al. (2014). "Increasing Acceptance of Free-Floating Car Sharing Systems Using Smart Relocation Strategies: A Survey Based Study of Car2go Hamburg". In: *Computational Logistics*. Ed. by Rosa G. González-Ramírez et al. Vol. 8760. Cham: Springer International Publishing, pp. 151–162. ISBN: 978-3-319-11420-0. DOI: 10.1007/978-3-319-11421-7_10. URL: http://link.springer.com/10.1007/978-3-319-11421-7_10 (visited on 06/11/2025).

Illgen, Stefan and Michael Höck (2019). “Literature Review of the Vehicle Relocation Problem in One-Way Car Sharing Networks”. In: *Transportation Research Part B: Methodological* 120, pp. 193–204. ISSN: 01912615. DOI: 10.1016/j.trb.2018.12.006.

Jiang, Hongbo et al. (2021). “An Empirical Study of Travel Behavior Using Private Car Trajectory Data”. In: *IEEE Transactions on Network Science and Engineering* 8.1, pp. 53–64. ISSN: 2327-4697, 2334-329X. DOI: 10.1109/TNSE.2020.3025529.

Jo, Hanbyul et al. (2024). *Vehicle Allocation Modeling: Optimizing the Logistics Behind a Free-Flow Carsharing in Milan, Italy*. Zenodo. DOI: 10.5281/zenodo.14232091. URL: https://zenodo.org/records/14232091 (visited on 06/01/2025).

Job Post (2025). *Job Post MILES Driver and Cleaner*. URL: https://web.archive.org/web/20240623203213/https://miles-mobility.jobs.personio.de/job/1148455?language=de&display=de (visited on 03/26/2025).

Jorge, Diana and Gonçalo Correia (2013). “Carsharing Systems Demand Estimation and Defined Operations: A Literature Review”. In: *European Journal of Transport and Infrastructure Research* 13.3 (3). ISSN: 1567-7141. DOI: 10.18757/ejtir.2013.13.3.2999.

Jorge, Diana et al. (2015). “Trip Pricing of One-Way Station-Based Carsharing Networks with Zone and Time of Day Price Variations”. In: *Transportation Research Part B: Methodological* 81, pp. 461–482. ISSN: 01912615. DOI: 10.1016/j.trb.2015.06.003.

Kang, Junhee et al. (2016). “Finding Factors That Influence Carsharing Usage: Case Study in Seoul”. In: *Sustainability* 8.8, p. 709. ISSN: 2071-1050. DOI: 10.3390/su8080709.

KBA (2025). *Pkw-Bestand in Berlin 2025*. Kraftfahrt-Bundesamt, provided by Statista. URL: https://de.statista.com/statistik/daten/studie/255179/umfrage/bestand-an-pkw-in-berlin/ (visited on 06/22/2025).

Klemmer, Konstantin et al. (2016). “Explaining Spatio-Temporal Dynamics in Carsharing: A Case Study of Amsterdam”. In: *Proceedings of the 22nd Americas Conference on Information Systems (AMCIS 2016)*. Americas Conference on Information Systems. San Diego, CA, USA: Association for Information Systems. URL: https://aisel.aisnet.org/amcis2016/ITProj/Presentations/35/.

Kortum, Katherine et al. (2016). “Free-Floating Carsharing: City-Specific Growth Rates and Success Factors”. In: *Transportation Research Procedia* 19, pp. 328–340. ISSN: 23521465. DOI: 10.1016/j.trpro.2016.12.092.

Le Vine, Scott et al. (2014). “A New Approach to Predict the Market and Impacts of Round-Trip and Point-to-Point Carsharing Systems: Case Study of London”. In: *Transportation Research Part D: Transport and Environment* 32, pp. 218–229. ISSN: 13619209. DOI: 10.1016/j.trd.2014.07.005.

Lippoldt, Katrin et al. (2018). “Effectiveness of Different Incentive Models in Free-Floating Carsharing Systems: A Case Study in Milan”. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 2018 21st International Conference on Intelligent Transportation Systems (ITSC). Maui, HI: IEEE, pp. 1179–1185. ISBN: 978-1-7281-0321-1. DOI: 10.1109/ITSC.2018.8569242. URL: https://ieeexplore.ieee.org/document/8569242/ (visited on 06/11/2025).

Lippoldt, Katrin et al. (2019). “Analyzing the Potential of User-Based Relocations on a Free-Floating Carsharing System in Cologne”. In: *Transportation Research Procedia* 37, pp. 147–154. ISSN: 23521465. DOI: 10.1016/j.trpro.2018.12.177.

Liu, Dongbo et al. (2021). “Real-Time Return Demand Prediction Based on Multisource Data of One-Way Carsharing Systems”. In: *Journal of Advanced Transportation* 2021.1, p. 6654909. ISSN: 2042-3195. DOI: 10.1155/2021/6654909.

Mathew, Wesley et al. (2012). “Predicting Future Locations with Hidden Markov Models”. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. Ubicomp ’12: The 2012 ACM Conference on Ubiquitous Computing. Pittsburgh Pennsylvania: ACM, pp. 911–918. ISBN: 978-1-4503-1224-0. DOI:

10.1145/2370216.2370421. URL: https://dl.acm.org/doi/10.1145/2370216.2370421 (visited on 06/18/2025).

Ménoire, Marlène et al. (2020). "Predicting Carsharing Station-Based Trip Generation Using a Growth Model". In: *Transportation Research Procedia* 48, pp. 1466–1477. ISSN: 23521465. DOI: 10.1016/j.trpro.2020.08.192.

Messa, Federico and Diego Deponte (2021). "Free-Flow Carsharing Systems in a Spatio-Temporal Urban Ecosystem: An Urban Informatics Approach". In: European Transport Conference 2021. Online: Association for European Transport. DOI: 10.5281/zenodo.6493824. URL: https://zenodo.org/records/6493824 (visited on 06/13/2025).

Miles Sustainability Report (2023). *MILES Sustainability Report 2023*. MILES. URL: https://archive.org/details/miles-sustainability-report-2023-en (visited on 06/20/2025).

Müller, Johannes and Klaus Bogenberger (2015). "Time Series Analysis of Booking Data of a Free-Floating Carsharing System in Berlin". In: *Transportation Research Procedia*. 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands 10, pp. 345–354. ISSN: 2352-1465. DOI: 10.1016/j.trpro.2015.09.084.

Niels, Tanja and Klaus Bogenberger (2017). "Booking Behavior of Free-Floating Carsharing Users: Empirical Analysis of Mobile Phone App and Booking Data Focusing on Battery Electric Vehicles". In: *Transportation Research Record* 2650.1, pp. 123–132. ISSN: 0361-1981. DOI: 10.3141/2650-15.

Schiffer, Maximilian et al. (2021). "A Polynomial-Time Algorithm for User-Based Relocation in Free-Floating Car Sharing Systems". In: *Transportation Research Part B: Methodological* 143, pp. 65–85. ISSN: 0191-2615. DOI: 10.1016/j.trb.2020.11.001.

Schmöller, Stefan and Klaus Bogenberger (2014). "Analyzing External Factors on the Spatial and Temporal Demand of Car Sharing Systems". In: *Procedia - Social and Behavioral Sciences* 111, pp. 8–17. ISSN: 18770428. DOI: 10.1016/j.sbspro.2014.01.033.

Schmöller, Stefan et al. (2015). "Empirical Analysis of Free-Floating Carsharing Usage: The Munich and Berlin Case". In: *Transportation Research Part C: Emerging Technologies* 56, pp. 34–51. ISSN: 0968090X. DOI: 10.1016/j.trc.2015.03.008.

Schreckenberger, Christian et al. (2018). "Next Place Prediction: A Systematic Literature Review". In: *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Prediction of Human Mobility*. PredictGIS 2018. New York, NY, USA: Association for Computing Machinery, pp. 37–45. ISBN: 978-1-4503-6042-5. DOI: 10.1145/3283590.3283596. URL: https://dl.acm.org/doi/10.1145/3283590.3283596 (visited on 06/18/2025).

Schulte, Frederik and Stefan Voß (2015). "Decision Support for Environmental-friendly Vehicle Relocations in Free- Floating Car Sharing Systems: The Case of Car2go". In: *Procedia CIRP* 30, pp. 275–280. ISSN: 22128271. DOI: 10.1016/j.procir.2015.02.090.

Shaheen, Susan A. et al. (2015). "One-Way Carsharing's Evolution and Operator Perspectives from the Americas". In: *Transportation* 42.3, pp. 519–536. ISSN: 0049-4488, 1572-9435. DOI: 10.1007/s11116-015-9607-0.

Song, Chaoming et al. (2010). "Limits of Predictability in Human Mobility". In: *Science* 327.5968, pp. 1018–1021. DOI: 10.1126/science.1177170.

Stokkink, Patrick and Nikolas Geroliminis (2021). "Predictive User-Based Relocation through Incentives in One-Way Car-Sharing Systems". In: *Transportation Research Part B: Methodological* 149, pp. 230–249. ISSN: 01912615. DOI: 10.1016/j.trb.2021.05.008.

Streetcrowd (2025). StreetCrowd. URL: https://www.streetcrowd.com/how-it-works (visited on 03/26/2025).

TLC Trip Record Data (2025). NYC Taxi & Limousine Commission. URL: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page (visited on 06/20/2025).

Wagner, Sebastian et al. (2015). "Data Analytics in Free-Floating Carsharing: Evidence from the City of Berlin". In: *2015 48th Hawaii International Conference on System Sciences*. 2015 48th Hawaii International Conference on System Sciences, pp. 897–907. DOI: 10.1109/HICSS.2015.112. URL: https://ieeexplore.ieee.org/abstract/document/7069916 (visited on 04/06/2025).

Wang, Lei et al. (2018). "GPS-data-driven Dynamic Destination Prediction for on-Demand One-Way Carsharing System". In: *IET Intelligent Transport Systems* 12.10, pp. 1291–1299. ISSN: 1751-9578. DOI: 10.1049/iet-its.2018.5250.

Wang, Lei et al. (2019). "Incentive-Based Approach to Control Demands for Operation of One-Way Carsharing System". In: *Transportation Research Record: Journal of the Transportation Research Board* 2673.4, pp. 427–438. ISSN: 0361-1981, 2169-4052. DOI: 10.1177/0361198119837497.

Wang, Lei et al. (2021). "Demand Control Model with Combinatorial Incentives and Surcharges for One-Way Carsharing Operation". In: *Transportation Research Part C: Emerging Technologies* 125, p. 102999. ISSN: 0968090X. DOI: 10.1016/j.trc.2021.102999.

Waserhole, Ariel and Vincent Jost (2016). "Pricing in Vehicle Sharing Systems: Optimization in Queuing Networks with Product Forms". In: *EURO Journal on Transportation and Logistics* 5.3, pp. 293–320. ISSN: 21924376. DOI: 10.1007/s13676-014-0054-4.

Weikl, Simone and Klaus Bogenberger (2013). "Relocation Strategies and Algorithms for Free-Floating Car Sharing Systems". In: *IEEE Intelligent Transportation Systems Magazine* 5.4, pp. 100–111. ISSN: 1941-1197. DOI: 10.1109/MITS.2013.2267810.

Weikl, Simone and Klaus Bogenberger (2015). "A Practice-Ready Relocation Model for Free-Floating Carsharing Systems with Electric Vehicles – Mesoscopic Approach and Field Trial Results". In: *Transportation Research Part C: Emerging Technologies* 57, pp. 206–223. ISSN: 0968-090X. DOI: 10.1016/j.trc.2015.06.024.

Willing, Christoph et al. (2017). "Moving in Time and Space – Location Intelligence for Carsharing Decision Support". In: *Decision Support Systems* 99, pp. 75–85. ISSN: 01679236. DOI: 10.1016/j.dss.2017.05.005.

Wu, Ting and Min Xu (2022). "Modeling and Optimization for Carsharing Services: A Literature Review". In: *Multimodal Transportation* 1.3, p. 100028. ISSN: 2772-5863. DOI: 10.1016/j.multra.2022.100028.

Appendices

A
Service Area

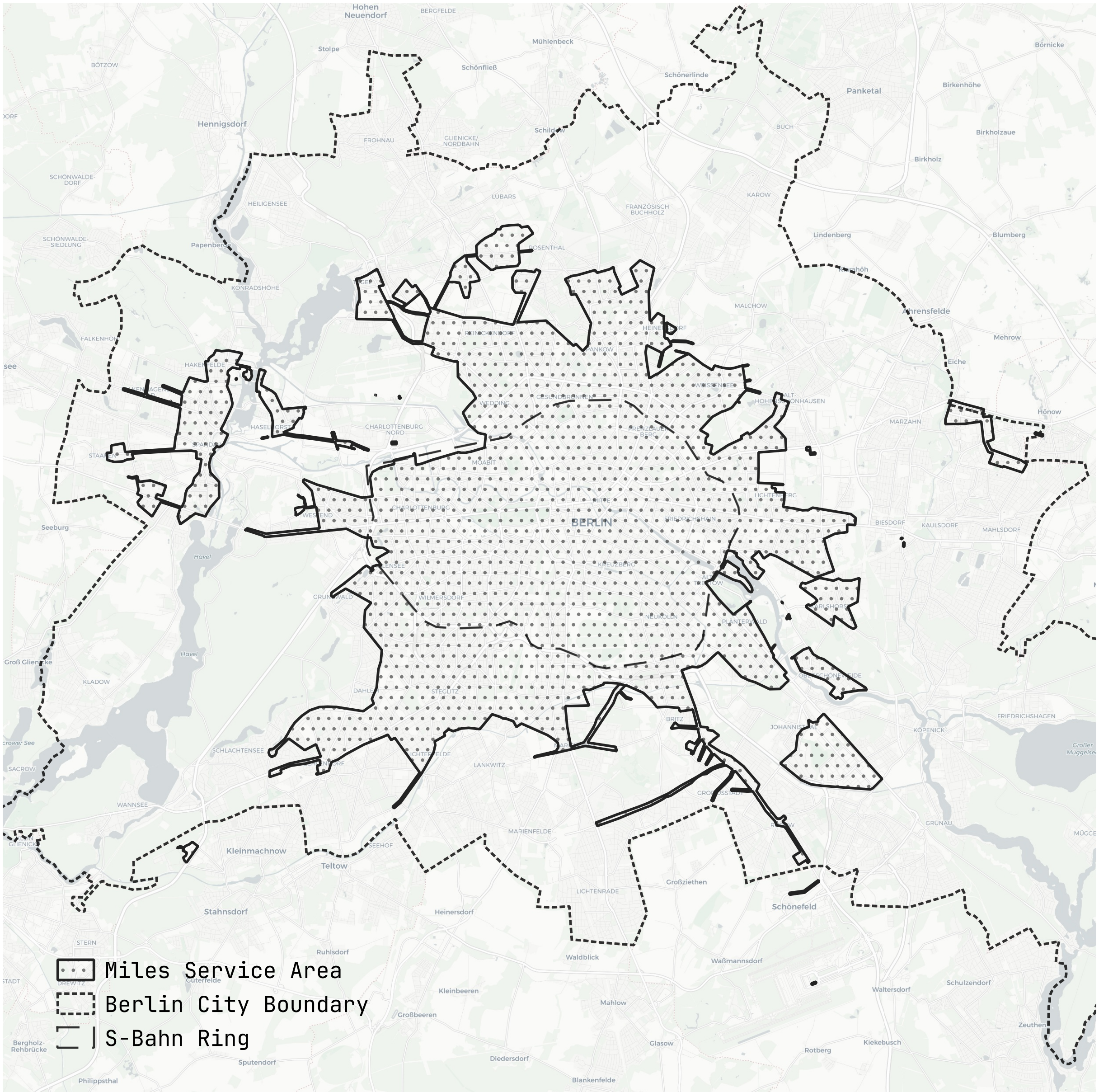


Figure A.1: *Service Area Context*
Basemap via CartoDB

B
Point of Interest Categories

Table B.1: *Point of interest categories*

Category Name	Included Tags	Number of POIs
Food	Restaurant	6078
Drink	Bar; Pub	3901
Health	Doctors; Dentist; Pharmacy; Clinic; Hospital	2478
Entertainment	Cinema; Theatre; Nightclub; Amusement Arcade	388
Recreation	Fitness Centre; Sports Centre; Swimming Pool; Pitch; Climbing; Track	4314
Transit	Railway Station; Bus Station; Ferry Terminal, Aerodrome	222
Shopping	Shop (all types)	16797
Education	School; University; Kindergarten; College	2488
Finance	Bank	1205
Accommodation	Hotel; Hostel; Guest House; Motel	712
Culture	Museum; Gallery; Library; Arts Centre	700
Religion	Place of Worship	497
Public Services	Post Office; Police; Fire Station; Townhall	237
Nature	Park; Wood	2004

C
Demand OLS Results

Table C.1: *OLS regression results for trip residuals with traffic and weather variables*

Dep. Variable:	Trip Residuals	**R-squared:**	0.007
Model:	OLS	**Adj. R-squared:**	0.007
Method:	Least Squares	**F-statistic:**	1057
No. Observations:	1,002,635	**Prob (F-statistic):**	0.000
Df Residuals:	1,002,627	**Log-Likelihood:**	-1.691e+06
Df Model:	7	**AIC:**	3.384e+06
Covariance Type:	Nonrobust	**BIC:**	3.384e+06

Predictor	**Coef.**	**Std. Err.**	**t**	**P>|t|**	**[0.025, 0.975]**
Constant (Intercept)	-1.9626	0.058	-33.863	<0.01	[-2.076, -1.849]
Traffic Flow	6.3016	0.148	42.681	<0.01	[6.012, 6.591]
Traffic Flow^2	-4.2506	0.092	-46.096	<0.01	[-4.431, -4.070]
Temperature	-0.0399	0.001	-47.632	<0.01	[-0.042, -0.038]
Temperature^2	0.0013	2.57e-05	49.227	<0.01	[0.001, 0.001]
Precipitation	0.1927	0.010	20.032	<0.01	[0.174, 0.212]
Precipitation^2	-0.0142	0.001	-16.806	<0.01	[-0.016, -0.013]
Precipitation Delta	-0.0167	0.007	-2.410	0.016	[-0.030, -0.003]
Precipitation Delta^2	-0.0142	0.001	-16.806	<0.01	[-0.016, -0.013]

Omnibus:	486702.589	Skew:	1.562
Durbin-Watson:	1.444	Prob(JB):	0.000
Prob(Omnibus):	0.000	Kurtosis:	28.421
Jarque-Bera (JB):	27405070.149	Cond. No.:	2.30e+15

Table C.2: *OLS regression results for trip residuals with POI categories during weekday middays*

Dep. Variable:	Trip Starts Midday Weekday	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.437
Method:	Least Squares	**F-statistic:**	30.79
No. Observations:	538	**Prob (F-statistic):**	2.28e-59
Df Residuals:	523	**Log-Likelihood:**	-3122.9
Df Model:	14	**AIC:**	6276.0
Covariance Type:	Nonrobust	**BIC:**	6340.0

Predictor	**Coef.**	**Std. Err.**	**t**	**P>|t|**	**[0.025, 0.975]**
Constant (Intercept)	17.9680	5.294	3.394	0.001	[7.568, 28.368]
Points of Interest: Food	0.9559	0.714	1.338	0.181	[-0.448, 2.360]
Points of Interest: Drink	-7.7722	0.871	-8.922	0.000	[-9.484, -6.061]
Points of Interest: Health	-0.2275	0.759	-0.300	0.765	[-1.719, 1.264]
Points of Interest: Entertainment	-10.4473	3.353	-3.116	0.002	[-17.035, -3.860]
Points of Interest: Recreation	-1.0442	0.515	-2.029	0.043	[-2.055, -0.033]
Points of Interest: Transit	21.2825	5.374	3.960	0.000	[10.725, 31.840]
Points of Interest: Shopping	0.1270	0.200	0.636	0.525	[-0.265, 0.519]
Points of Interest: Education	-3.3162	1.237	-2.680	0.008	[-5.747, -0.885]
Points of Interest: Finance	8.6880	1.906	4.558	0.000	[4.944, 12.432]
Points of Interest: Accommodation	13.1082	1.679	7.806	0.000	[9.809, 16.407]
Points of Interest: Culture	13.9205	1.894	7.349	0.000	[10.199, 17.642]
Points of Interest: Religion	1.3671	3.562	0.384	0.701	[-5.630, 8.365]
Points of Interest: Public Services	1.6461	4.560	0.361	0.718	[-7.313, 10.605]
Points of Interest: Nature	-2.6074	0.764	-3.414	0.001	[-4.108, -1.107]

Omnibus:	229.651	Skew:	1.595
Durbin-Watson:	1.906	Prob(JB):	0.000
Prob(Omnibus):	0.000	Kurtosis:	12.562
Jarque-Bera (JB):	2277.840	Cond. No.:	107.0

D
Destination Baseline & Results

For the majority baseline and the logit model, only contexts present in the training set were kept for the test set. Maximum set sizes: 836,023 training rows, 125,296 test rows. Resolution 7 corresponds to an average hexagon edge length of approximately 1320 maters, resolution 8 to approximately 500 meters.

The theoretical upper bound $\Pi_{\max}$ is the is the maximum accuracy a model can achieve given the context variables. It is implemented as the Monte-Carlo estimate of the Bayes-optimal prediction accuracy, or $\Pi_{max} = \max_{d \in D} P(D = d|C)$.

Π^* represents the theoretical maximum accuracy based on human mobility, if trips were user-bound and user history was available, using the context variables, following Song et al. (2010).

The top-k majority distribution shows the percentage of test rows that can be predicted correctly by selecting the top-k most frequent destinations in the training set.

Table D.1: *Upper bound, Π* and top-k majority distribution at H3 resolution 7*

Context Variables	$\Pi_{\max}$	Π^*	Top-1	Top-2	Top-3	Test Kept
C H	11.7%	37.6%	7.9%	14.25%	19.67%	99.9%
C H F T R	15.1%	44.5 %	6.63%	11.93%	16.51%	70.6%
C H L P	11.7%	37.6%	7.82%	13.94%	19.33%	99.9%
C H F T R L P	15.1%	44.5%	6.63%	11.93%	16.51%	70.6%
C H D B F T R L P	15.2%	45.4%	5.94%	10.60%	14.57%	69.1%
H F T R L P	6.4%	30.6%	5.16%	9.65%	13.79%	90.5%
C F T R:	13.1%	36.9%	8.13%	14.39%	20.12%	99.6%

Table D.2: *Top-k logit model predictions at H3 resolution 7*

Context Variables	Top-1	Top-2	Top-3	Test Kept
C H	8.73%	15.55%	21.56%	99.9%
C H F T R	8.87%	15.72%	21.65%	70.6%
C H L P	8.71%	15.52%	21.43%	99.9%
C H F T R L P	8.88%	15.73%	21.65%	70.6%
C H D B F T R L P	8.85%	15.70%	21.75%	69.1%
H F T R L P	5.64%	10.55%	15.08%	90.5%
C F T R:	8.42%	14.95%	20.70%	99.6%

Table D.3: *Top-k CatBoost model predictions at H3 resolution 7*

Context Variables	Top-1	Top-2	Top-3	Test Kept
C H F T R	8.69%	15.57%	21.58%	100%

Table D.4: *Top-k GRU model predictions at H3 resolution 7*

Context Variables	Top-1	Top-2	Top-3	Test Kept
C H F T R	8.97%	15.81%	21.75%	100%

Table D.5: *top-k majority distribution at H3 resolution 8*

Context Variables	Top-1	Top-2	Top-3	Test Kept
C H	1.92%	3.13%	4.16%	98%
C H F T R	1.33%	2.25%	3%	65%
C H L P	1.92%	3.13%	4.16%	98%
C H F T R L P	1.33%	2.25%	3%	65%
C H D F T R L P	1.33%	2.24%	3%	65%
H F T R L P	1.16%	2.18%	3.12%	90%
C F T R:	2.06%	3.74%	5.16%	99%

Legend:

- **C**: Start H3 Cell
- **H**: Hour of Week
- **D**: Discount Applied
- **B**: Was Reserved
- **F**: Traffic Flow
- **T**: Temperature
- **R**: Precipitation
- **L**: POI cluster
- **P**: Dominant POI category

Declaration

Florian König, CODE University of Applied Sciences.

This dissertation is original work, written solely for this purpose, and all the authors whose studies and publications contributed to it have been duly cited.

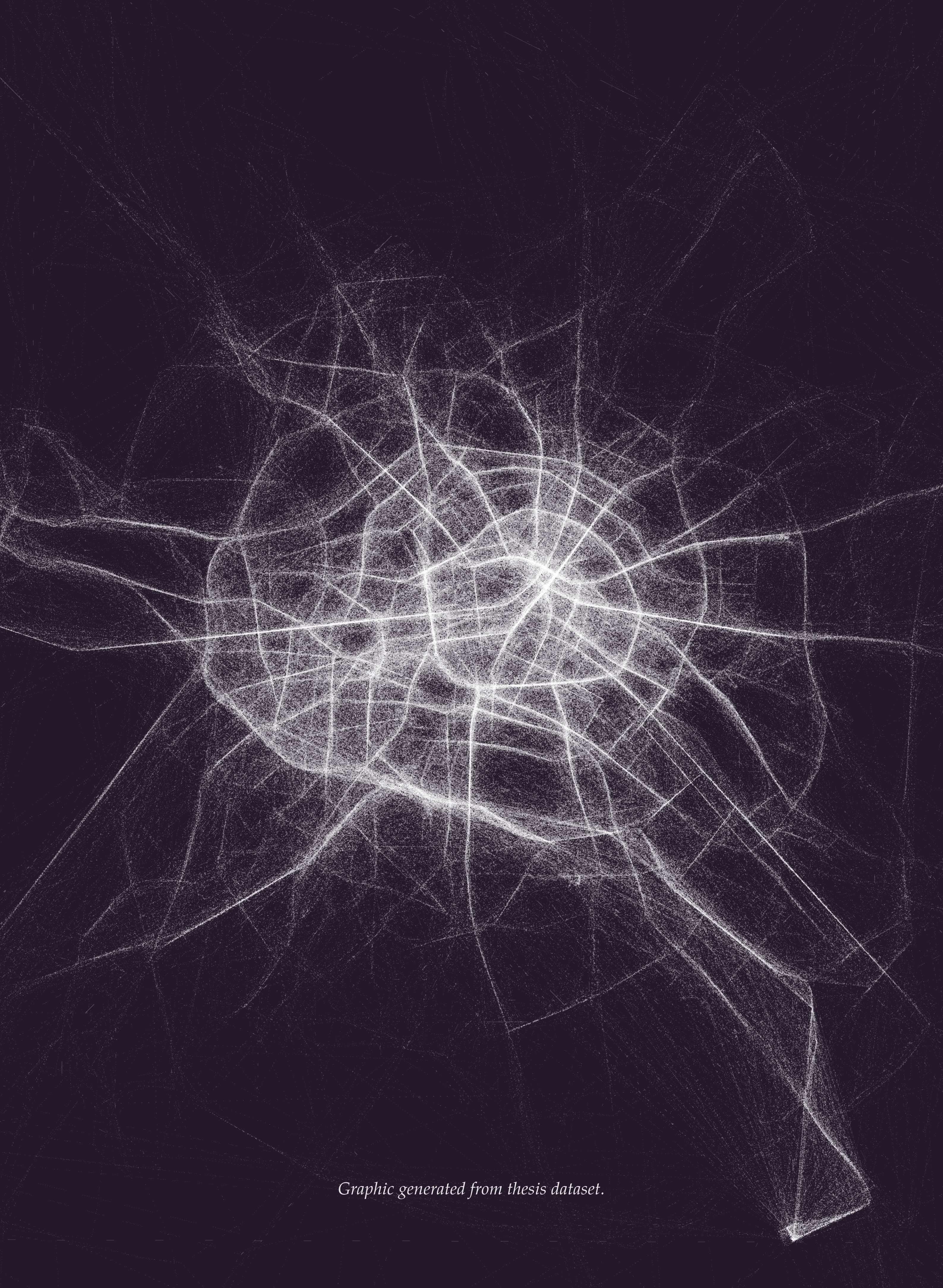
I hereby confirm that I have written the thesis titled *Forecasting User Behavior to Optimize Fleet Distribution* by myself, without contributions from any sources other than those cited in the text and bibliography. All graphics, drawings and images not created by me but included in this thesis, along with any uses of generative AI have been fully and accurately referenced.

Furthermore, I confirm that neither this work nor parts of it have been previously or concurrently used as an assessment submission in other courses or in other examination proceedings.

In the table below, I list the generative AI tools used in creating my thesis, as well as a description of the purpose and extent of the use of each tool.

AI Tool Used	**Purpose**
Anthropic Claude	Latex syntax generation
Consensus	Literature discovery
GitHub Copilot (OpenAI GPT)	Code completion in data analysis
Grammarly	Grammar checking
OpenAI GPT	First-loop feedback Editing suggestions
NotebookLM (Google Gemini)	Research notes structuring
Stanford STORM	Literature discovery

[redacted]



Graphic generated from thesis dataset.