

LUDWIG-MAXIMILIANS-UNIVERSITÄT  
MÜNCHEN

BACHELOR THESIS

---

**Analysis of Public Transport  
Connectivity and Different Places of  
Residence in Munich**

Described by Rental Prices and the Isar

---



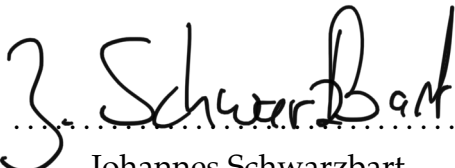
*Author:*  
Johannes SCHWARZBART

*Supervisors:*  
Cornelia FÜTTERER  
Prof. Dr. Thomas AUGUSTIN

May 27, 2021

# Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. This paper was not previously presented to another examination board and has not been published.

 .....

Johannes Schwarzbart

München, .....

May 27th, 2021

# Abstract

The thesis explores the relationship between public transport connectivity and their corresponding rental prices in different parts of the city. Creating a regression model with these factors tests whether an area's good connectivity by transit also means higher rental prices in the surrounding apartments. In order to accomplish this, we create the definition of a station's reachability as the percentage of stations one can reach within ten minutes. The resulting models indeed indicate a positive influence of apartments' rental prices on the area's reachability. Furthermore, we test if the river Isar, which runs through Munich, negatively influences the city's connectivity. This is tested by comparing transit speeds of connections crossing the Isar with the ones that stay on just one side, by creating a regression model containing Isar information, and by clustering the city. While the results indicate that the river does play a modest role in Munich's public transport system, it cannot be shown to be a strongly negative influence on transit connection durations. Finally, the transit network is transferred into graph form to use graph-theoretical notions of node importance, called *centrality*, to characterize the station's importance in the network. These centrality measures are then compared to rental price information and the previously defined reachability, showing generally positive correlations between these values.

# Contents

<b>Declaration of Authorship</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Data Description</b>	<b>2</b>
2.1 Public Transport Data . . . . .	2
2.2 Geographical Data . . . . .	4
2.3 Rental Data . . . . .	5
<b>3 Reachability Analysis</b>	<b>10</b>
3.1 Introduction . . . . .	10
3.2 Descriptive Analysis . . . . .	11
3.2.1 Reachability by Station . . . . .	11
3.2.2 Reachability by Borough & Sub-Borough . . . . .	13
3.3 Reachability & Rental Prices . . . . .	14
3.3.1 Generalized Linear Models (GLMs) . . . . .	15
3.3.2 Reachability Analysis by Apartment . . . . .	16
3.3.3 Reachability Analysis by Borough . . . . .	18
3.3.4 Reachability Analysis by Sub-Borough . . . . .	20
3.3.5 Rent to Reachability Ratio . . . . .	23
<b>4 Isar as a Potential Barrier for Public Transportation Reachability</b>	<b>25</b>
4.1 Speed of Transportation Methods . . . . .	25
4.2 Modeling Public Transportation Ride Times . . . . .	30
4.3 Cluster Analysis . . . . .	36
4.4 Clustering Evaluation . . . . .	36
4.4.1 Adjusted Rand Index . . . . .	36
4.5 Cluster Analysis Algorithms . . . . .	37
4.5.1 Hierarchical Clustering . . . . .	37
4.5.2 k-Means Clustering . . . . .	39

<b>5 Graph Analysis</b>	<b>42</b>
5.1 Introduction to Graph Theory . . . . .	42
5.2 Centrality Measures . . . . .	44
5.2.1 Degree Centrality . . . . .	44
5.2.2 Betweenness Centrality . . . . .	47
5.2.3 Closeness Centrality . . . . .	49
5.2.4 Eigenvector Centrality . . . . .	50
5.3 Relationship between Reachability, Centrality Measures & Rental Prices	54
<b>6 Summary &amp; Outlook</b>	<b>57</b>
<b>Bibliography</b>	<b>58</b>
<b>A Appendix</b>	<b>60</b>

# List of Figures

2.1	Public Transport Stations in Munich . . . . .	4
2.2	Boroughs . . . . .	5
2.3	Sub-Boroughs . . . . .	5
2.4	Rental Price vs. Rental Time . . . . .	6
2.5	Rental Time . . . . .	6
2.6	Net Rent Comparison . . . . .	6
2.7	Living Space Comparison . . . . .	6
2.8	Map of Mietspiegel Rental Prices . . . . .	7
2.9	Map of Immobilienscout24 Rental Prices . . . . .	7
2.10	Median Rental Price per Borough: Mietspiegel . . . . .	8
2.11	Median Rental Price per Borough: Immobilienscout24 . . . . .	8
2.12	Median Rental Price per Sub-Borough: Mietspiegel . . . . .	8
2.13	Median Rental Price per Sub-Borough: Immobilienscout24 . . . . .	8
3.1	Reachability Universität . . . . .	11
3.2	Station Reachability . . . . .	12
3.3	Reachability by Station: Histogram . . . . .	13
3.4	Reachability by Station: Boxplot . . . . .	13
3.5	Reachability Boroughs . . . . .	14
3.6	Reachability Sub-Boroughs . . . . .	14
3.7	Logistic Regression Mietspiegel: Full Data . . . . .	17
3.8	Logistic Regression Mietspiegel: New Contracts . . . . .	17
3.9	Regression Immobilienscout24 . . . . .	17
3.10	Regression Mietspiegel By Borough . . . . .	20
3.11	Regression Mietspiegel by Borough (New Contracts) . . . . .	20
3.12	Regression Immobilienscout24 by Borough . . . . .	20
3.13	Regression Mietspiegel By Sub-Borough . . . . .	22
3.14	Regression Mietspiegel by Sub-Borough (New Contracts) . . . . .	22
3.15	Regression Immobilienscout24 by Sub-Borough . . . . .	22
3.16	Rent to Reachability Ratio by Borough: Mietspiegel Data . . . . .	23
3.17	Rent to Reachability Ratio by Borough: Immobilienscout24 Data . . . . .	23
3.18	Rent to Reachability Ratio by Sub-Borough: Mietspiegel Data . . . . .	24

3.19	Rent to Reachability Ratio by Sub-Borough: Immobilienscout24 Data	24
4.1	Speed Comparison by Distance Calculation Method	26
4.2	Average Approximated Transportation Speed: Bicycle Distances	27
4.3	Share of Direct Connections: By Product and Isar Position	28
4.4	Average Public Transportation Speeds in Relation to Isar	28
4.5	Distance vs. MVG Ride Time	31
4.6	Distance vs. Time: Base Model	32
4.7	Hierarchical Clustering of MVG Network	38
4.8	k-Means Clustering Example	40
4.9	k-Means Clustering of MVG Network (2 Clusters)	40
4.10	k-Means Clustering of MVG Network (5 Clusters)	41
5.1	Simple Graph	42
5.2	Graph Adjacency Matrix	43
5.3	Weighted Digraph Adjacency Matrix	44
5.4	Public Transport Network as a Graph	45
5.5	Degree Example	45
5.6	Degree Centrality	46
5.7	Betweenness Centrality	48
5.8	Closeness Centrality	49
5.9	Eigenvector Centrality	51
5.10	Boxplot of Centrality Measures	52
5.11	Histogram of Centrality Measures	53
5.12	Relationship between Reachability and Centrality Measures (Log Values)	54
5.13	Reachability, Centrality Measures & Rental Prices	55
A.1	Map of Mietspiegel Rental Prices (New Contracts)	60
A.2	Median Rental Price per Borough: Mietspiegel (New Contracts)	60
A.3	Median Rental Price per Sub-Borough: Mietspiegel (New Contracts)	60
A.4	Rent to Reachability Ratio by Borough: Mietspiegel (New Contracts)	61
A.5	Rent to Reachability Ratio by Sub-Borough: Mietspiegel (New Contracts)	61
A.6	Residual Diagnostics for Distance vs. Transit Ride Time Base Model	61
A.7	Relationship between Reachability and Centrality Measures	62
A.8	Reachability, Centrality Measures & Rental Prices: Immobilienscout24 Data	62
A.9	Reachability, Centrality Measures & Rental Prices: Mietspiegel Data	63

A.10 Reachability, Centrality Measures & Rental Prices: Sub-Borough . . . 63



# List of Tables

3.1	Stations with the highest Reachability . . . . .	13
3.2	Regression by Apartment Summary . . . . .	16
3.3	Regression by Borough Summary . . . . .	19
3.4	Regression by Sub-Borough Summary . . . . .	21
4.1	t-test Transportation Speed and the Isar . . . . .	29
4.2	t-test Subway Speed and the Isar . . . . .	29
4.3	Linear Regression: Distance vs. MVG Duration . . . . .	32
4.4	ANOVA output for models 4.3 & 4.4 . . . . .	32
4.5	ANOVA output for models 4.4 & 4.5 . . . . .	33
4.6	ANOVA output for models 4.5 & 4.6 . . . . .	33
4.7	Comparison of MVG duration regression models . . . . .	35
5.1	Stations with the highest Degree Centrality . . . . .	47
5.2	Stations with the highest Betweenness Centrality . . . . .	48
5.3	Stations with the highest Closeness Centrality . . . . .	50
5.4	Stations with the highest Eigenvector Centrality . . . . .	52

# 1 Introduction

Public transportation networks allow us to travel around in big cities relatively quickly and inexpensively, thus creating an option for the masses to travel without needing a car. Consequently, many people use these networks regularly, as is the case in Munich, where the city's public transit conveyed more than 600 million passengers in 2019 [18].

As such, it becomes an essential factor for regular users, how well one is connected by public transport, which might also play an instrumental part in the search for an apartment. Consequently, apartments in well-connected areas might demand higher rents, thus creating a positive relationship between rent and reachability in a city. Testing this belief will be a major part of this thesis.

Another part of the thesis will explore the river Isar's role regarding transit connections. Since it runs from south to north through Munich, it might take more time to use transit connections crossing the Isar than the ones staying within one side.

Finally, we will explore Munich's public transportation network from a graph-theoretical perspective, where we explore different centrality measures of stations around the city.

## 2 Data Description

In this chapter, we will describe the data used in the thesis and explain their origins. Munich's public transport information, spatial information, and rental prices are the three different kinds of data.

### 2.1 Public Transport Data

The most important data for the analysis were accurate connection information for the public transport system in Munich. The *Münchener Verkehrsgesellschaft* (MVG; Munich Transport Company) is the company responsible for operating public transport in Munich [18] and can therefore be seen as the authoritative source for correct transport information in the city. Detailed transport information from each station to any other station within the MVG network can be queried on their website [mvg.de](http://mvg.de).

The required inputs are the start and destination station, the date and time of the desired connection, and options to exclude certain products (U-Bahn, S-Bahn, tram, bus) in the route planning. For the analysis, we queried the connections for March 15th, 2021, at 8:00 am without excluding any product. The output of each query consists of several route options the user could take to get to the destination. These options can differ regarding their starting time, the used products, and the route duration. Each route option contains information about the used products and their label, e.g., *U1* for the subway line 1, start and end times plus information about required changes and wait times. The transit duration is not given to us directly but is only conveyed by displaying start and end times by hour and minute. A connection leaving the station at 10:00:00 o'Clock and arriving at the next station at 10:00:59, a duration of 59 seconds, is presented in our data as leaving at 10:00 and arriving at 10:00; thus, the duration time is zero minutes. On the other hand, connections leaving at 10:00:59 and arriving at 10:02:00, a connection time of 61 seconds, are shown as leaving at 10:00 and arriving at 10:02, suggesting a two-minute ride. Throughout several stops, this naturally averages out, but since we are often dealing with one connection at a time, this potentially presents a problem, especially when dealing with graphs in Chapter 5, where an edge weight of zero

might indicate the absence of a connection. Consequently, all connections with the same start and end time are given the transit duration of half a minute instead of zero minutes.

Getting a complete picture of the MVG network within the city of Munich now requires querying all connection combinations. Since the number of stations is constantly changing and this analysis does not cover the whole MVG area, just the city of Munich, it was impossible to find a definitive, up-to-date list of the city's public transport stations. In order to get all or as many stations as possible, three different methods were combined. First, we scraped all station names listed on *mux.de*. For all these station names, we queried the MVG API to receive more information about the station, such as the offered products and coordinates per station. Each station has an MVG internal ID, and most stations queried from *mux.de* were well within the ID range from 1 to 5000. By querying the MVG API for all stations with IDs from 1 to 10,000, some more stations could be found, although most are either not in use anymore or used MVG internally only, such as depot facilities. Lastly, when querying all station combinations, we checked for each station on the way from station A to station B, whether that station is present in our database of known stations, and would add stations that were not. Using these methods, we were able to get information about 1095 public transport stations in Munich. In Figure 2.1 all these stations are mapped with the outline of the city of Munich. The popup over the *Universität* station shows the station name, its products, and the borough/sub-borough of the station.

The number of scraped connections is  $1095^2 = 1,199,025$ . Out of these connections, some station combinations did not return any results. Querying connections to the *Langwieder See* from *Marienplatz* for example, returned some results while the other way round did not return any. In this case, the reason is the special nature of the *Langwieder See* station, which is only served in summer on days with fair weather [16]. Such special cases occur quite rarely, which causes 1240 possible connections to be missing and leaves us with 1,197,785 connections. With the goal in mind to cause the least possible strain on the MVG website, we spread out the data scrape over several weeks and only queried for the lightweight *JSON* formatted version of the data. All acquired data was transferred to a PostgreSQL database for efficient storage and further analysis.

Since there are several routes for each connection, we picked the route per connection with the shortest total duration for further analysis. In case of a tie, the one with the earlier start time was used.

### Public Transport Stations in Munich

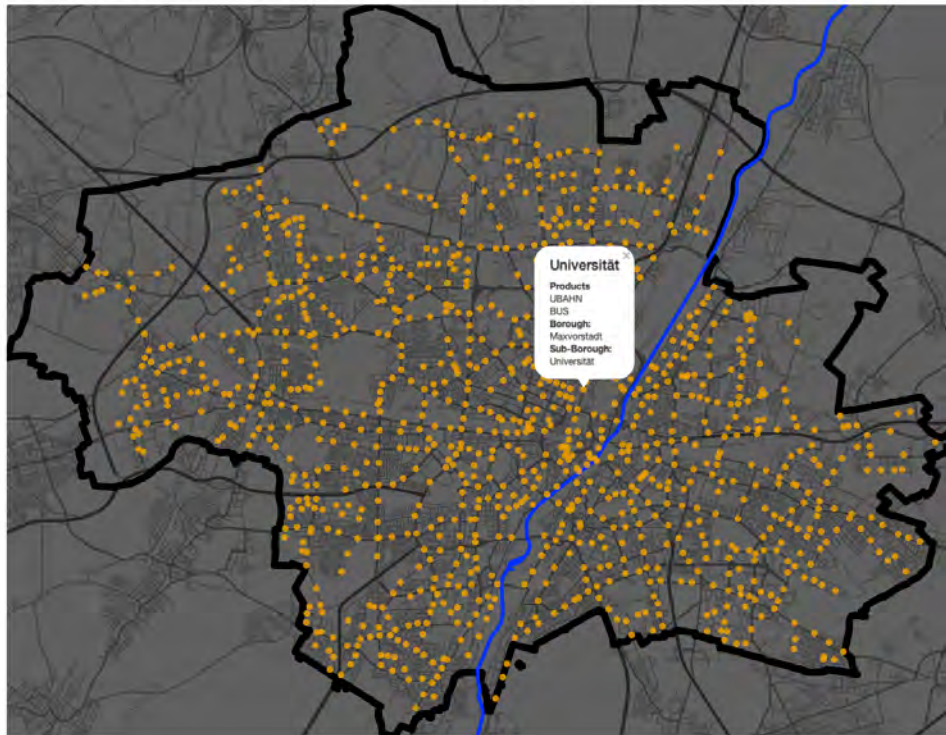


FIGURE 2.1: Public Transport Stations in Munich

## 2.2 Geographical Data

Further geographical data for the city of Munich was acquired from [openstreetmap.org](https://openstreetmap.org) [22]. Isar coordinates were used to assign each station to one side of the river, such that we gain information on whether connections cross the Isar. Furthermore, this allowed calculating the straight distance to the closest Isar location from each station. For mapping and calculating geographical data, the open-source PostgreSQL extension *PostGIS* was used.

The city of Munich is divided into 25 boroughs (Bezirke) and 108 sub-boroughs (Bezirksteile) [17]. The coordinates for these boroughs and the outline of Munich were also acquired from the same *OpenStreetMap* dataset, which allowed the assignment of each station to a borough and sub-borough.

The straight-line distances between all stations, calculated using their coordinates, are sometimes far off from the actually traveled distances between stations. Therefore the data was extended by retrieving the distances between stations using the car, bicycle, and footway. This data was provided by the [openrouteservice.org](https://openrouteservice.org) API [21].



FIGURE 2.2:  
25 Boroughs



FIGURE 2.3:  
108 Sub-Boroughs

## 2.3 Rental Data

Rental data for apartments in Munich consists of two different datasets, data acquired by scraping *immobilienscout24.de* in March and May 2021 and data from a rent index analysis (*Mietspiegel*) for the city of Munich conducted in 2018 [15], which was provided to us by LMU's department for statistics. For that analysis, the authors acquired a representative sample of rental object information in Munich by getting the information from actual tenants throughout the city. The most relevant information from both data sets includes net prices, square meters per apartment, and location information. For the *ImmobilienScout24* data, only apartments with precise locations were considered, resulting in 2606 rental objects. The *Mietspiegel* data assigns each apartment the coordinates of the center of their neighborhood (*Stadtviertel*), and all 3024 rental objects were used in the analysis.

One significant difference between the datasets is that the *Mietspiegel* data uses prices from existing rental contracts, whereas the *ImmobilienScout24* data uses proposed rental prices that may or may not result in actual rental contracts. Another drawback from the *ImmobilienScout24* data is the possibility that relatively cheap apartments are not listed on the website for a long time before the apartment is rented out, while expensive apartments may stay on the website for quite some time. These factors could lead to generally higher prices in the dataset than in reality.

On the other hand, the *Mietspiegel* rental contracts tend to be cheaper the older the rental contract is, as seen in Figure 2.4, where the rental time is plotted against the rental price per square meter. Figure 2.5 depicts a histogram of rental times, which shows that a significant number of apartments (1154) have a rental contract older than a decade, which contrasts starkly with the potentially new contracts from the *ImmobilienScout24* data.

When comparing the net rental price of the *Mietspiegel* and *ImmobilienScout24* datasets in Figure 2.6, it becomes clear how differently priced the apartments are. The median net rent of 21.59 Euros per square meter for the *ImmobilienScout24* apartments is almost twice the median net rent per square meter of 11.74 Euros from the *Mietspiegel* data and the highest

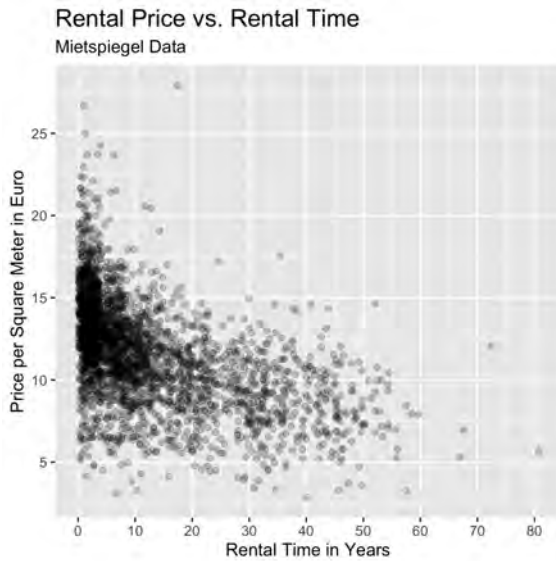


FIGURE 2.4:  
Rental Price vs. Time

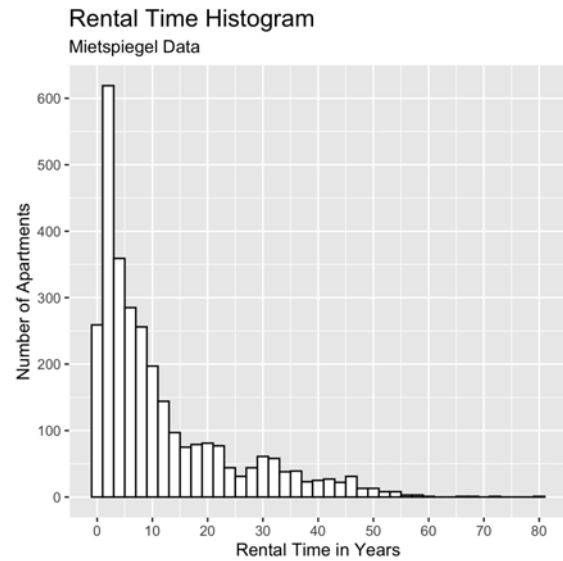


FIGURE 2.5:  
Rental Time

prices in the *ImmobilienScout24* data range up to almost 80 Euros, while not a single one in the *Mietspiegel* dataset reaches the 30 Euro mark. When comparing the living space of the apartments in the datasets, we notice the higher median value of 73.5 square meters for the *Mietspiegel* apartments compared to 58 square meters in the other dataset.

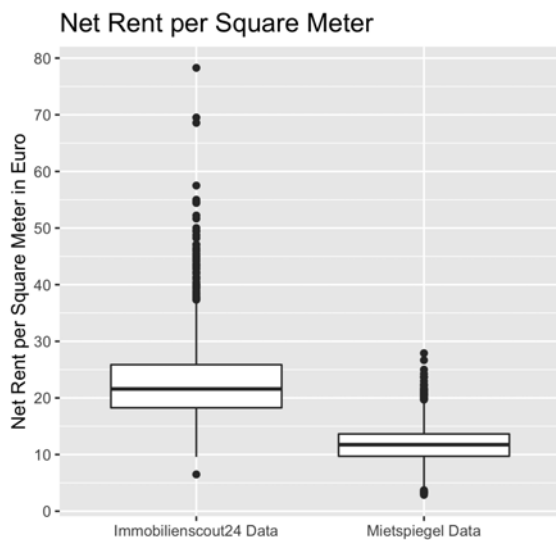


FIGURE 2.6:  
Net Rent Comparison

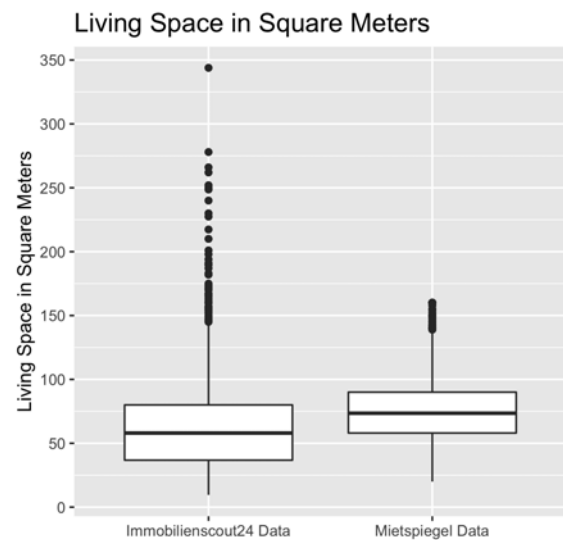


FIGURE 2.7: Living  
Space Comparison

Because of the fundamentally different characteristics of the two datasets, the data was never mixed and always used separately in the forthcoming analysis.

Figure 2.8 depicts all apartments from the *Mietspiegel* dataset and Figure 2.9 all apartments from the *ImmobilienScout24* dataset, colored by their respective rental price per square meter, grouped into their respective four rental price quartiles. For the *Mietspiegel* map, we



### Rental Price per Square Meter Mietspiegel (Full Data)

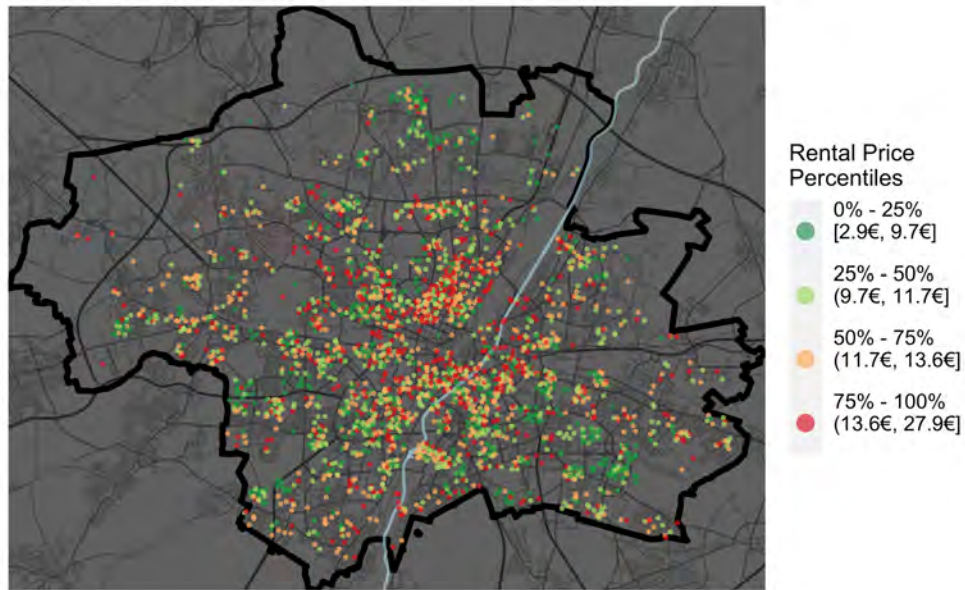


FIGURE 2.8: Map of Mietspiegel Rental Prices

### Rental Price per Square Meter ImmobilienScout24 Data

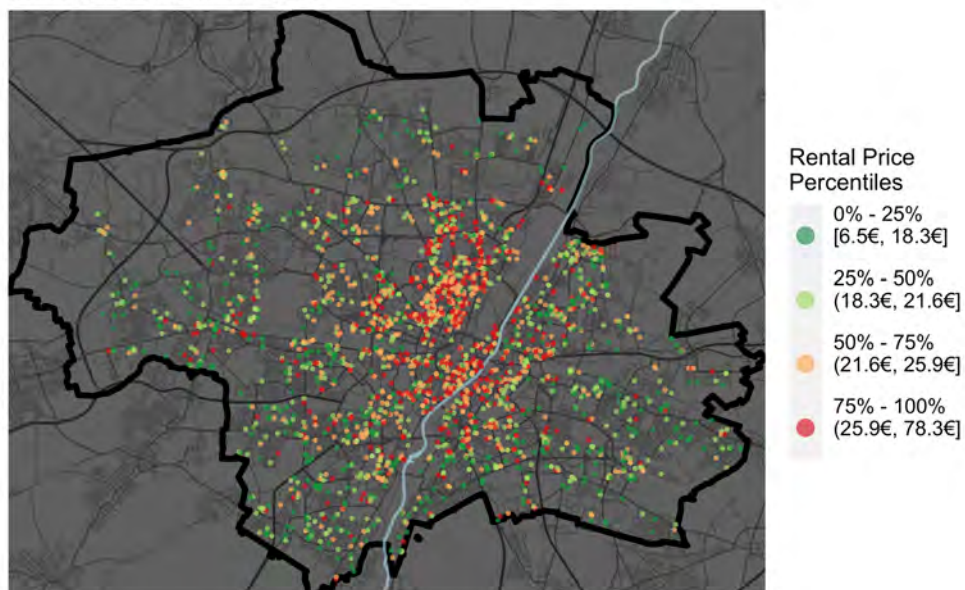


FIGURE 2.9: Map of ImmobilienScout24 Rental Prices

had to apply a small jitter for each observation, slightly changing the position of the data points to increase visibility since the coordinates in this dataset are not precise. Instead, several apartments in the dataset are clustered together into the same close-by location.

Both maps show how the number of apartments decreases the closer one gets to the city



boundaries and how a large number of apartments in the city center are mostly colored red and orange, suggesting rent higher than the median rental prices in their respective datasets. One structural difference between the maps is that there is a significant amount of apartments priced in the third and fourth quartile at the outskirts of Munich in the *Mietspiegel* dataset, while that trend is less noticeable in the *ImmobilienScout24* dataset. This fact can be noticed as well when aggregating the data on the sub-borough level, as depicted in Figures 2.12 and 2.13, where the median rental price per square meter is shown for both datasets. The white, missing sub-boroughs indicate that there are no apartments in the dataset located in these sub-boroughs.

**Median Rental Price per Borough**  
Mietspiegel (Full Data)

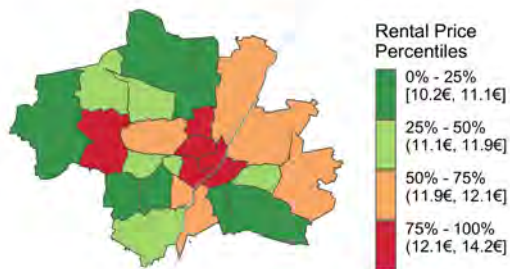


FIGURE 2.10:  
Median Rental Price per  
Borough:  
Mietspiegel

**Median Rental Price per Borough**  
ImmobilienScout24 Data

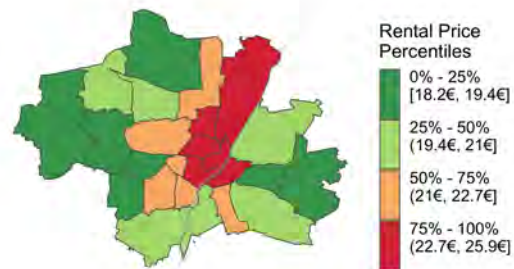


FIGURE 2.11:  
Median Rental Price per  
Borough:  
ImmobilienScout24

**Median Rental Price per Sub-Borough**  
Mietspiegel (Full Data)

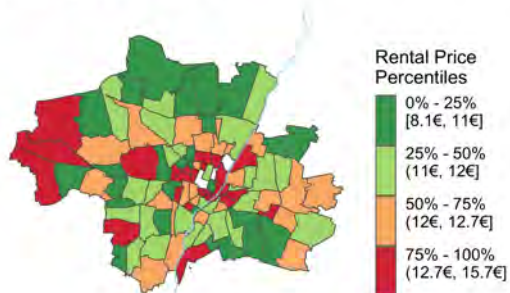


FIGURE 2.12:  
Median Rental Price per  
Sub-Borough:  
Mietspiegel

**Median Rental Price per Sub-Borough**  
ImmobilienScout24 Data



FIGURE 2.13:  
Median Rental Price per  
Sub-Borough:  
ImmobilienScout24

Since the *Mietspiegel* dataset contains a significant amount of old contracts, the question occurs if there are different trends regarding the rental price distribution throughout the

city when only considering more recent contracts. However, when only considering newer contracts of the *Mietspiegel* dataset, the ones with lengths of less than ten years, the same general trends can be seen as in the full dataset and the corresponding maps are depicted in Figures A.1, A.2 and A.3 in Appendix A.

## 3 Reachability Analysis

This chapter will illustrate the concept of a reachability analysis and apply it to the existing data. First, we will define reachability and descriptively apply it to the city's stations and aggregate them by borough and sub-borough. Afterward, regression models will describe the relationship between rental prices and reachability from the *Mietspiegel* and *ImmobilienScout24* datasets.

### 3.1 Introduction

When considering reachability, we are generally interested in the number of stations one can reach starting from any station in a network [2]. For a public transportation network, each station is ideally reachable from any other station. The interesting aspect becomes the number of stations one can travel to within a given time period.

A quick visual understanding of a station's connectedness can be given by considering Figure 3.1. Pictured is the outline of Munich in black, the Isar in blue, and all the MVG stations as circles in a color range from green to orange to red, indicating the time needed to get to them from the start station. The start station *Universität* is indicated by the big triangle.

Generally, the further away we get from the start station, the longer the travel time, until it takes more than forty minutes, where stations are colored red. However, some stations are quite far away but are still reachable quickly, such as the *Brudermühlstraße* station, the most southern dark green circle on the map. Because of the direct connection by subway, it is still a short ride between the stations. Other stations with a seemingly short geographical distance to the *Universität* station, on the other hand, are colored yellow or even orange, signaling relatively long rides. These examples amplify the notion of the analysis that travel times are considered in the form of public transportation durations instead of spatial distances. For defining reachability concretely, we need to set a specific threshold, and in this thesis, we will pick one of 10 minutes.

With an  $n \times n$  matrix,  $n$  as the number of stations in the public transportation network, and  $n_{ij}$  as the shortest travel time in minutes from station  $i$  to station  $j$  by public transport, we can formalize reachability of a station  $i$  as:

$$Reachability_i = \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{n_{ij} \leq 10\}} \quad (3.1)$$

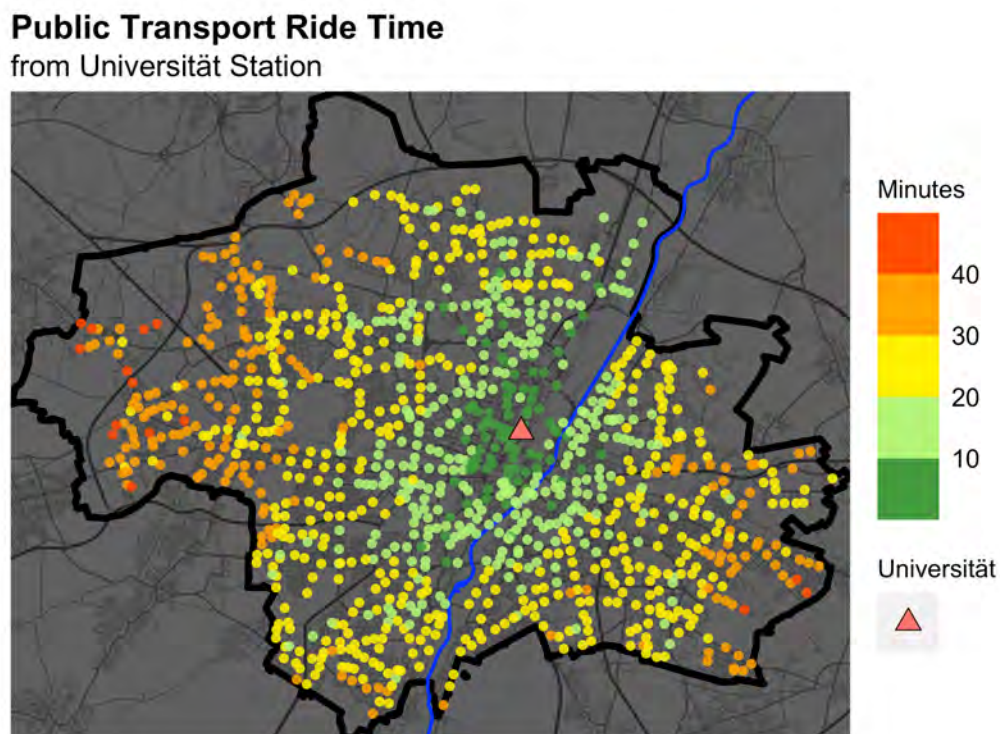


FIGURE 3.1: Reachability Universität

Informally we can say that the reachability of a station is defined as the percentage of all stations in the public transportation network one can travel to in ten minutes or less. Thus, the dark green circles represent the stations considered in our reachability definition, the ones reachable in up to ten minutes. In total, 93 stations are reachable within ten minutes starting from *Universität*, resulting in a reachability of 8.5% for this station.

A station's reachability can be seen as a proxy for its attractiveness as a location to live within the city under the assumption that high reachability is desirable and thus correlated with high rents and high demand for living around such stations. People would therefore pay more for the ability to travel to lots of stations quickly.

## 3.2 Descriptive Analysis

In order to get an overview of reachability for stations, boroughs, and sub-boroughs, we will descriptively analyze their reachability values, which will later allow us to analyze the data inferentially.

### 3.2.1 Reachability by Station

The reachability for all stations can be obtained in the same manner as the calculation for the *Universität* station we have just seen. The result is depicted in Figure 3.2, where each dot

represents a station. In order to get a better overview of the reachability distribution, the stations are colored according to groups of percentiles, where the red dots are stations with reachability in the first quartile, the blue dots with values in the second and third quartile, thus representing the middle 50% of the data. The fourth quartile is further divided in pink dots in the 99th percentile to show the biggest outliers in the data, and stations in green, representing the rest of the fourth quartile. Furthermore, the dots' sizes are proportional to their respective reachabilities. Generally, stations close to the city limits tend to have the worst reachability, although not exclusively, as one can notice some red stations very close to the city center. Stations in the fourth quartile are distributed throughout the city but are most commonly found close to the city center, where most of the stations in the 99th percentile are located as well. These ten stations in pink with the highest reachability scores are listed in Table 3.1. The central station is achieving the highest reachability of fourteen percent, while the other notable station is *Scheidplatz*, being located comparatively far away from the city center in the north, while still being one of the top ten stations at 9.3%.

### Reachability by Station

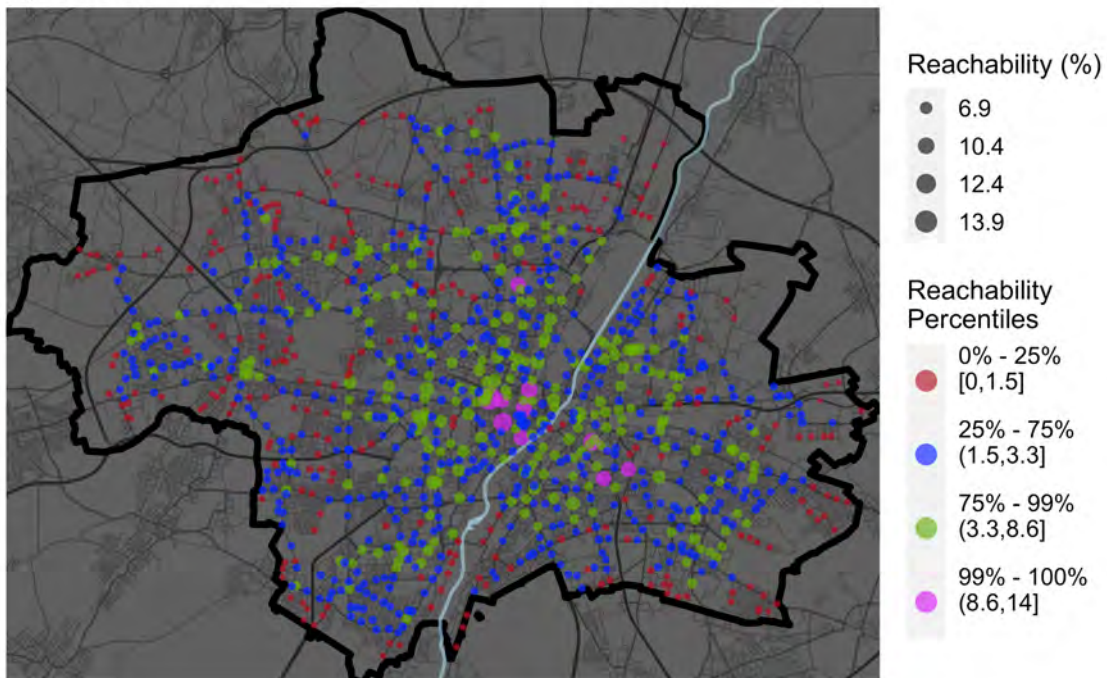
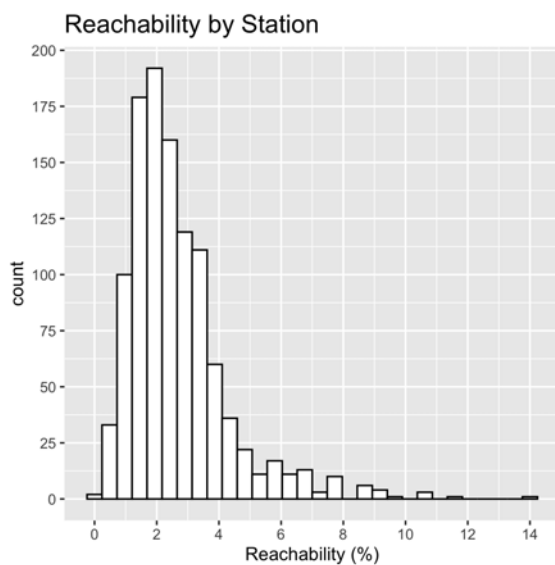
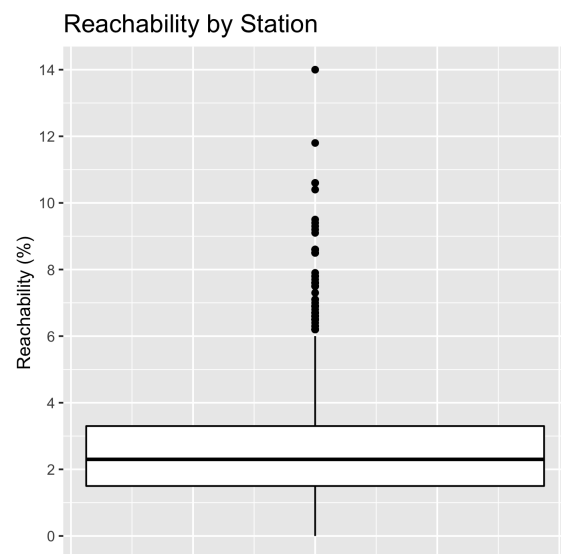


FIGURE 3.2: Station Reachability

In Figures 3.3 and 3.4 the reachability distribution for all stations is depicted in the form of a boxplot on the right and a histogram on the left. We can see a right-skewed distribution, where most stations have reachability scores of one to five percent, while the values range up to fourteen percent for the central station. The boxplot amplifies the notion that the top reachability stations are quite the outliers, achieving reachability scores of almost six times the median value of 2.3%.

Station	Reachability (%)
Hauptbahnhof (S, U, Bus, Tram)	14.0
Sendlinger Tor	11.8
Odeonsplatz	10.6
Ostbahnhof	10.6
Marienplatz	10.4
Fraunhoferstraße	9.5
Karlsplatz (Stachus)	9.4
Scheidplatz	9.3
Innsbrucker Ring	9.2
Karl-Preis-Platz	9.1

TABLE 3.1: Stations with the highest Reachability

FIGURE 3.3:  
Reachability by Station:  
HistogramFIGURE 3.4:  
Reachability by Station:  
Boxplot

### 3.2.2 Reachability by Borough & Sub-Borough

Reachability can also be aggregated on different levels. When considering the attractiveness of an apartment location in a city, it might not just be of interest how well the station closest to the apartment is connected with the rest of the city. When assuming that most people spend lots of time in their own neighborhoods, it might be of interest how well the boroughs are connected in general and how they compare with each other. Such an aggregation is depicted in Figure 3.5 on the borough level and in Figure 3.6 on the sub-borough level. The reachability is considered for every station within these boroughs or sub-boroughs, and the aggregated median value is then used as reachability for these subsections of Munich. The color-scales in the graphics are divided by quartile, where red boroughs and sub-boroughs are in their corresponding first quartile regarding all median reachability values of the (sub-)boroughs, the orange ones in the second quartile, lime-green in the third, and dark-green



in the fourth quartile. The results generally show worse reachability the further away the area is from the city center. There are some peculiarities, however. For instance, there is one white-colored sub-borough, *Schönfeldvorstadt*, where there are no stations. Furthermore, some sub-boroughs are quite close to the city center but are seemingly poorly connected by public transport, two of them right along the Isar, the *Maximilianeum* and *Dreimühlen* sub-boroughs. The opposite, where the area is far away but well connected, also occurs in sub-boroughs such as *Aubing-Süd* in the west and *Neuperlach* in the south-eastern part with median reachability measures of 2.7 and 3.1 percent respectively. Hence, the stations in *Aubing-Süd* have median reachability of 2.7%, which, per our reachability definition 3.1, is the percentage of stations reachable within 10 minutes.

The sub-borough and borough with the best reachability are the *Hackenviertel* sub-borough and the *Maxvorstadt* borough, both areas in the heart of Munich, with reachability values of 11.8% and 3.9%.

In contrast, the sub-boroughs and boroughs with the most limited reachability, *Fürstenried-West* and *Feldmoching-Hasenbergl*, have values of 0.9% and 1.5%, respectively.

**Median Reachability**  
per Borough

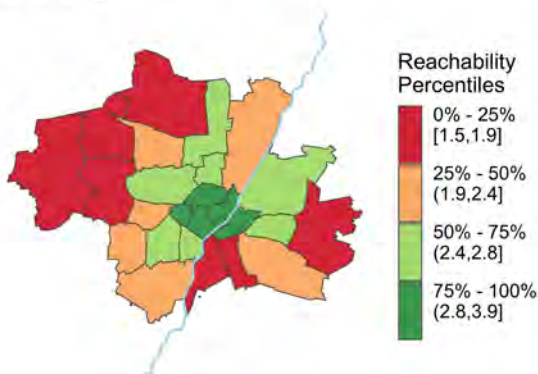


FIGURE 3.5:  
Reachability  
Boroughs

**Median Reachability**  
per Sub-Borough

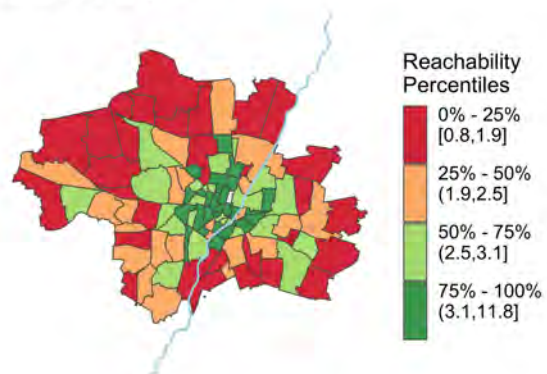


FIGURE 3.6:  
Reachability  
Sub-Boroughs

### 3.3 Reachability & Rental Prices

After exploring reachability properties around the city, it remains to be explored how rental prices relate to reachability. Generally, one would assume that high rental prices per square meter are associated with a good location within a city. Consequently, a good location in the city could be well-connected, allowing the resident to travel around the city quickly, making rental prices a proxy for reachability.

### 3.3.1 Generalized Linear Models (GLMs)

Linear models are appropriate for regression analysis with continuous and approximately normally distributed response variables [8]. However, in many applications, the response is not a continuous variable but rather binary, categorical, or a count variable. When thinking about the reachability of a station, we can imagine it as a binary variable, where every other station can either be reached within ten minutes or not. In GLMs, the distribution of the response is assumed to belong to a single family of distributions known as the exponential family, which includes the normal, Bernoulli, binomial, and Poisson distributions. A transformation of the mean response is then linearly related to the covariates via an appropriate *link function*.

In logistic regression, the predictor could be a continuous variable that can take on values over the entire real line, whereas the response is a probability and is therefore constrained to fall between 0 and 1. Here, the link function is called logit function  $\eta_i = \log\left(\frac{p}{1-p}\right)$ . We can see that the logit function transforms a variable constrained to  $[0, 1]$  to a variable that can take values over the entire real line. The link function makes the response compatible with the predictor variables, and hence it is possible to make it a linear function of the predictors plus a random component.

Thus, our model will generally look as follows:

$$g(\text{reachability}) = \log\left(\frac{\text{reachability}}{1 - \text{reachability}}\right) = \eta = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k. \quad (3.2)$$

Overall, for distributions of the exponential family with  $b$  as an arbitrary function,  $\phi$  as an arbitrary scale parameter and  $\theta$  as the canonical parameter of the distribution, we have [8]:

- a)  $E(Y_i) = b'(\theta_i) = \mu_i = h(\eta_i) = h(x_i' \beta)$ , with  $h$  as invertable response function,
- b)  $x_i' \beta = \eta_i = g(\mu_i)$ , with  $g = h^{-1}$  as link function,
- c)  $\text{Var}(Y_i) = \phi b''(\theta_i)$

As a criterium for model fit, we can use the deviance, which compares the maximum of the log-likelihood of the estimated model with the log-likelihood of the perfect model. With  $l_i(\text{fit})$  as the fitted model's log-likelihood of group  $i$  and  $l_i(\text{full})$  as the saturated model's log-likelihood, we get the deviance as [8]

$$D = -2 \sum_{i=1}^G \{l_i(\text{fit}) - l_i(\text{full})\}$$

For a more comparable measure, we can use *Deviance Explained*, calculated as

$$D_{\text{explained}} = 1 - \frac{D_{\text{pred}}}{D_{\text{null}}},$$

with  $D_{\text{pred}}$  as the estimated deviance of the model and  $D_{\text{null}}$  as the model's null-deviance.

Concretely, with

$$\eta = \beta_0 + \beta_1 \cdot \text{rentsqm},$$



we can now express the function in an exponential form by transforming it with the exponential function, resulting in

$$\frac{\text{reachability}}{1 - \text{reachability}} = \exp(\beta_0) \cdot \exp(\beta_1 \cdot \text{rentsqm})$$

An increase of rent per square meter by one unit, one Euro, in this case, results, *ceteris paribus*, in an increase of the reachability odds by the factor  $\exp(\beta_1)$ .

### 3.3.2 Reachability Analysis by Apartment

Now, we will conduct the analysis based on the *Mietspiegel* and *ImmobilienScout24* datasets. For the moment, however, we still have disjunct datasets of reachability per station and rental prices per apartment. We have several options to connect them, such as aggregation per borough and per sub-borough, but first, we will assign each apartment the station with the maximum reachability within a 500-meter radius. The purpose behind this approach is that the station with the maximum reachability within an apartment's walking distance is probably one of the most frequented ones to get around the city. Out of the 3024 rental objects in the *Mietspiegel* dataset, sixteen did not have stations within that radius and were assigned reachability 0, whereas each of the 2606 *ImmobilienScout24* apartments had a station in the 500-meter radius.

With  $g$  as log-likelihood function, our simple model then looks like this:

$$g(\widehat{\text{reachability}}) = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \text{rentsqm} \quad (3.3)$$

	<i>Dependent variable:</i>		
	<b>reachability</b>		
	Mietspiegel	Mietspiegel	ImmobilienScout24
	Full Data	Contr. Len. < 10 yrs	
<b>rentsqm</b>	0.019*** (0.001)	0.031*** (0.001)	0.013*** (0.001)
<b>Intercept</b>	-3.281*** (0.010)	-3.487*** (0.015)	-3.764*** (0.012)
Observations	3,024	1,865	2,606
Deviance Explained	1.37%	3.19%	2.46%

Note:

\* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$

TABLE 3.2: Regression by Apartment Summary

Table 3.2 shows the regression results in the left column with  $\widehat{\beta}_0 = -3.281$  and  $\widehat{\beta}_1 = 0.019$ , meaning that c.p. an increase of rental price per square meter by one Euro increases

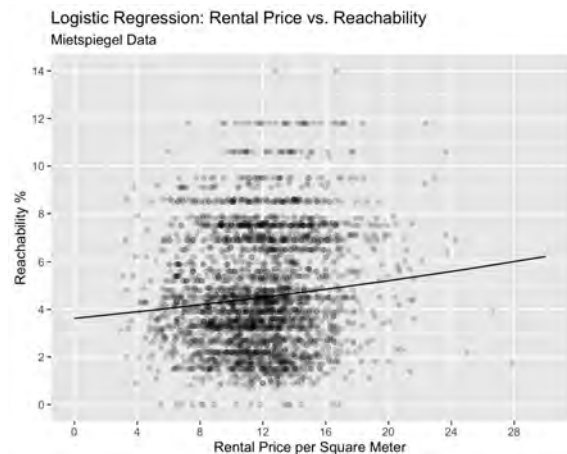


FIGURE 3.7:  
Regression Mietspiegel:  
Full Data

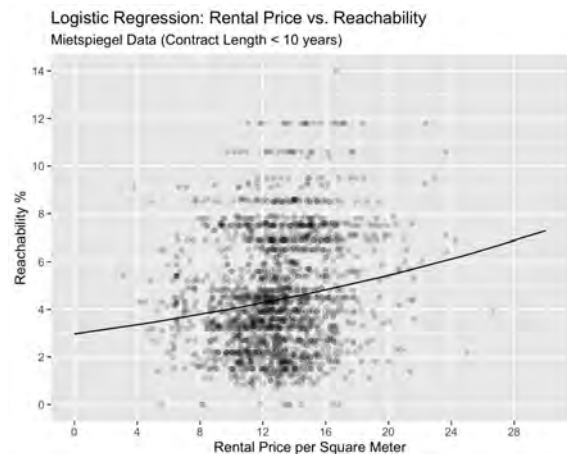


FIGURE 3.8:  
Regression Mietspiegel:  
New Contracts

the reachability odds by factor  $\exp(0.019) = 1.02$ . While this suggests a positive relationship between rent and reachability and both intercept and the *rentsqm* variable are significant at the 5% level, the explained deviance lies at just 1.37%. In Figure 3.7 the scatter plot of rental price and reachability is depicted with the corresponding regression line, showing that the data is very noisy and does not seem to follow a clear linear trend, with the regression line consequently not being a great fit on the data.

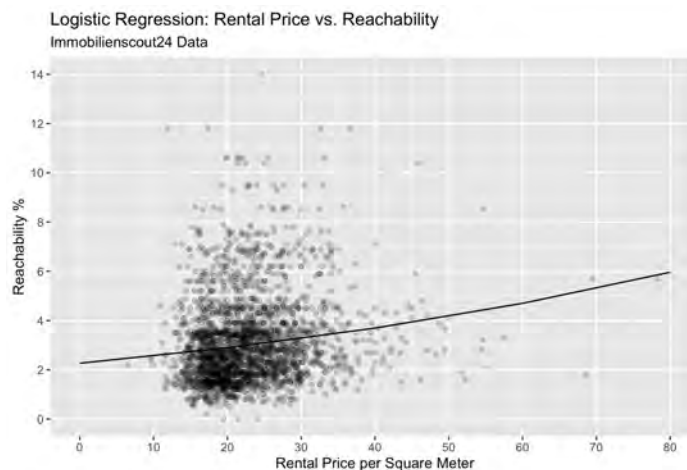


FIGURE 3.9: Regression Immobilienscout24

We have established before how the *Mietspiegel* dataset contains many apartments with old contracts and have seen lower prices with older contracts, introducing heterogeneity into the data. When conducting the same regression as before on apartments with contracts that are less than ten years old, the explanatory power of the model can be slightly improved, increasing the explained deviance to 3.19%, while keeping the covariables statistically significant. The exact output of the corresponding results can be seen in the middle column

in Table 3.2 and the scatter plot with the regression line is shown in Figure 3.8. For the dataset with these new contracts, the number of observations is reduced to 1864, and the median price per square meter increases to 12.81 Euros, while the median living space stays the same at 73 square meters. The price increase can be noticed when comparing the scatter plots again, where the cloud of observations seems slightly shifted to the right for the new contracts, amplifying how many cheaper apartments are not present in the reduced dataset anymore.

As before with the *Mietspiegel* dataset, we can conduct the corresponding analysis with the *ImmobilienScout24* data.

Table 3.2 shows the regression results, where we can see a positive relationship between rent and reachability, although the model's explanatory power is relatively weak, as indicated by the explained deviance value of 2.46%. These values are in line with the results from the *Mietspiegel* dataset, seen in Table 3.2. The scatter plot depicted in Figure 3.9 with the corresponding regression line further indicates very noisy data with no clear linear trend.

### 3.3.3 Reachability Analysis by Borough

When aggregating the data by borough, we get 25 median rental prices per square meter and median reachability scores in percent, based on the observations throughout Munich. The highest median price for the full *Mietspiegel* dataset is in the *Altstadt-Lehel* borough and the lowest in *Feldmoching-Hasenbergl* at 14.2 and 10.2 Euros respectively. The number of apartments per borough in the dataset ranges from 29 in *Allach-Untermenzing* to 235 in *Neuhausen-Nymphenburg*. For the data where only apartments with rental durations of less than ten years are considered, the number of rental objects per borough varies from 22 apartments in the *Allach-Untermenzing* borough to 156 in *Neuhausen-Nymphenburg*. The lowest median rent per square meter can also be found in *Allach-Untermenzing* at 11.5 Euros, and the highest once again in *Altstadt-Lehel* at 15.1 Euros.

In the *ImmobilienScout24* dataset, the number of apartments per borough ranges from 39 in *Schwanthalerhöhe* to 231 in *Bogenhausen*.

The lowest median reachability is assigned to *Feldmoching-Hasenbergl* at 1.5% stations reached within 10 minutes, whereas 3.9% can be reached in the same time starting from *Maxvorstadt* and the borough with highest median rent at 25.9 Euros is *Altstadt-Lehel*, as in the *Mietspiegel* data, and the one with the lowest is *Trudering-Riem* at 18.2 Euros per square meter. The borough with the lowest median rent in the *ImmobilienScout24* dataset is consequently higher than the lowest one in the *Mietspiegel* data, highlighting again how different the nominal prices are between them.

Since we now deal with reachability percentages of grouped data, we have to account for possible overdispersion. We can incorporate it using the *quasibinomial* family instead of the binomial one and keeping the logit as the link function, weighted by the number of apartments per borough. This way, the additional dispersion parameter is estimated to scale the standard errors.

Now we get the following model for the median reachability per borough.

$$g(\widehat{\text{median\_reachability\_borough}}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{median\_rentsqm\_borough} \quad (3.4)$$

In Table 3.3 the regression results are shown, with the full *Mietspiegel* data in the left column, the results for observations with new contracts in the middle one and for *Immobilien Scout24* data in the right one. Both variables, *rentsqm* and the intercept are statistically significant at the 5% level. For the full data, a rent per square meter increase of one Euro means that the proportion of reachable stations to the ones not reachable increases by the factor  $\exp(0.190) = 1.21$ , while the proportion increases by the factor  $\exp(0.221) = 1.25$  for the apartments with a contract length of fewer than ten years and by the factor  $\exp(0.079) = 1.08$  for the *Immobilien Scout24* data. In Figure 3.10 and Figure 3.11, we can see the higher price for apartments with rental lengths of less than ten years compared to the entire data set on the left. The scatter plot on the left looks more noisy, reflecting the lower deviance explained from the model, than the one on the right, where the predicted values, indicated by the regression line, seem to be a better fit for the data compared to the model with the entire dataset. Figure 3.12 depicts the scatter plot and the corresponding regression line for *Immobilien Scout24* data, showing a clear positive trend between rent and reachability, which is also expressed by the highest deviance explained value of 52.2%.

Overall, there seems to be a clear trend between higher rental prices and higher median reachability per borough.

	<i>Dependent variable:</i>		
	<b>reachability</b>		
	Mietspiegel	Mietspiegel	Immobilien Scout24
	Full Data	Contr. Len. < 10 yrs	
<b>rentsqm</b>	0.190** (0.070)	0.221*** (0.054)	0.079*** (0.016)
<b>Intercept</b>	-5.880*** (0.828)	-6.511*** (0.699)	-5.378*** (0.353)
Observations	25	25	25
Deviance Explained	27.9%	44.3%	52.2%

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 3.3: Regression by Borough Summary

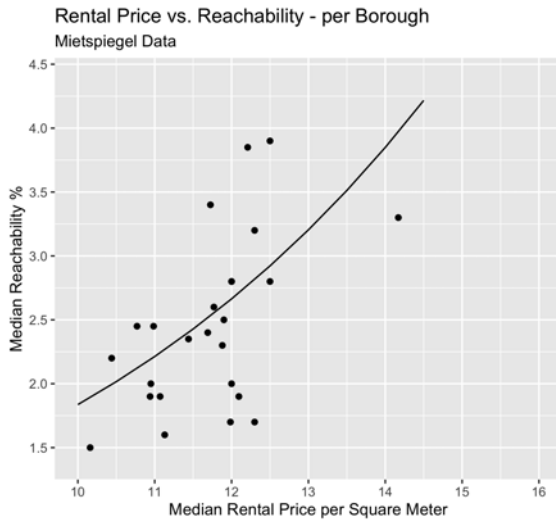


FIGURE 3.10:  
Regression Mietspiegel  
By Borough  
(Full Data)

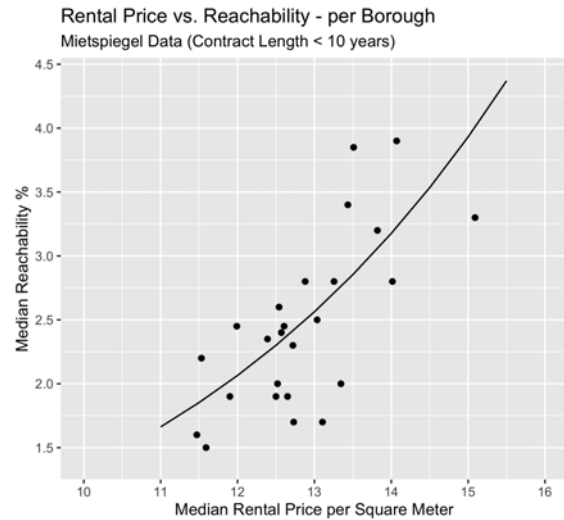


FIGURE 3.11:  
Regression Mietspiegel  
By Borough  
(New Contracts)

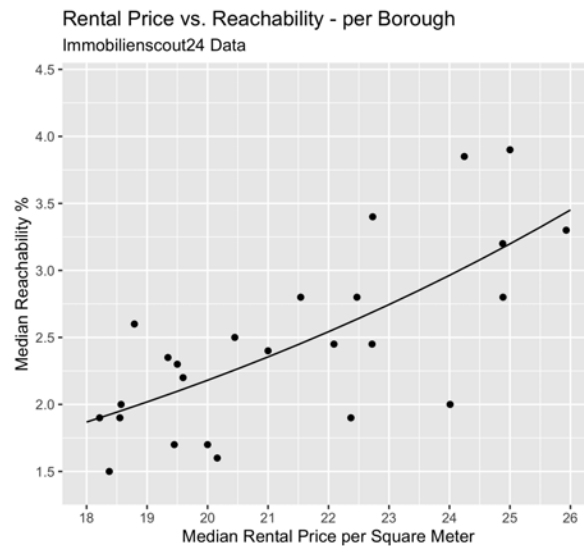


FIGURE 3.12: Regression Immobilienscout24 by Borough

### 3.3.4 Reachability Analysis by Sub-Borough

After aggregating the data by borough, we can do the same on the sub-borough level, resulting in median rental prices and median reachability scores for the 105 sub-boroughs in which the *Mietspiegel* dataset has rental objects. Hence, three of the sub-boroughs do not have any apartments within their borders. The number of apartments per borough now ranges from just one in *Biederstein*, *Freiham*, *Graggenau* and *Ludwigsfeld* to 106 in *Nymphenburg*. The median rental prices per square meter now range from 8.1 Euros in *Blumenau* to 15.7 Euros in *Hackenviertel* for the whole dataset. When limiting the observations to apartments with

rental duration fewer than ten years, the most expensive sub-borough stays the *Hackenviertel*, now at 17.8 Euros, while the cheapest one is *Am Hart* at 8.1 Euros. Five of the sub-boroughs now have only one apartment within their borders, while the most populous one in regards to the number of apartments is now *Obergiesing* at 72. Additionally, *Biederstein* has no apartments in the reduced dataset and is therefore excluded in the analysis of new contracts.

The *ImmobilienScout24* dataset has apartments in 106 of the 108 sub-boroughs, with the highest median rent of 29.8 Euros of 17 apartments in *Englischer Garten Süd* and the lowest price of 15.3 Euros in *Oberwiesenfeld* for just one apartment.

There is one sub-borough in the city with no public transport stations, *Schönfeldvorstadt*, where the median reachability was consequently set to 0%. The one with the lowest median reachability apart from this area is *Fürstenried-West* with a value of 0.85%, while the highest median reachability of 11.8% can be found in the most expensive sub-borough *Hackenviertel*. As before, we will use a GLM with the quasibinomial family and the logit as the link function, weighted by the number of apartments per sub-borough, resulting in the following model.

$$g(\widehat{median\_reachability\_sub\_borough}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot median\_rentsqm\_sub\_borough \quad (3.5)$$

	Dependent variable:		
	reachability		
	Mietspiegel Full Data	Mietspiegel Contr. Len. < 10 yrs	ImmobilienScout24 (3)
<b>rentsqm</b>	0.124*** (0.031)	0.160*** (0.025)	0.073*** (0.013)
<b>Intercept</b>	-5.060*** (0.367)	-5.679*** (0.336)	-5.193*** (0.287)
Observations	105	104	106
Deviance Explained	15.4%	28.5%	26.8%

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 3.4: Regression by Sub-Borough Summary

We can see the regression results in Table 3.4, with the full *Mietspiegel* data in the left column, the results for observations with new contracts in the middle and the *ImmobilienScout24* results in the right one. Both variables, *rentsqm* and the intercept are statistically significant at the 5% level. For the full dataset with any contract length, a rent per square meter increase of one Euro means that, on average, the proportion of reachable stations to the ones not reachable is increased by the factor  $\exp(0.124) = 1.13$ , while the proportion increases by the factor  $\exp(0.160) = 1.17$  for the apartments with a contract length fewer than ten years and by factor  $\exp(0.073) = 1.08$  for *ImmobilienScout24* data. Figures 3.13 and 3.14 show once

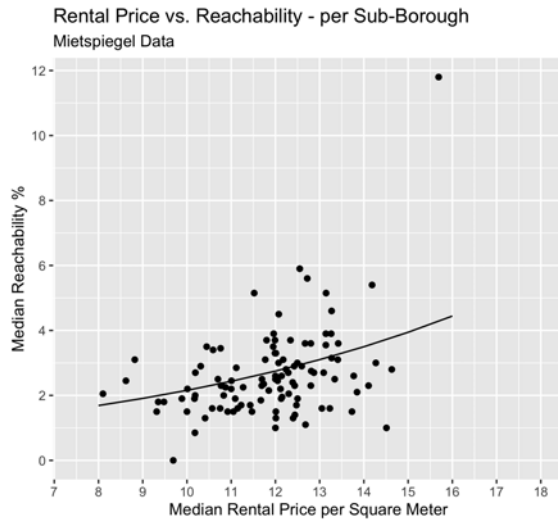


FIGURE 3.13:  
Regression Mietspiegel  
By Sub-Borough  
(Full Data)

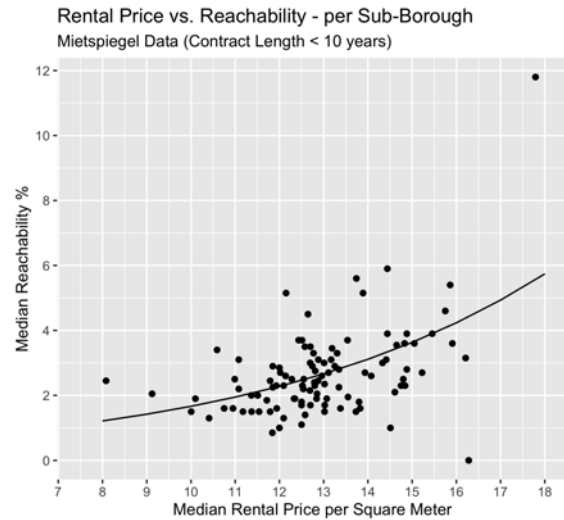


FIGURE 3.14:  
Regression Mietspiegel  
By Sub-Borough  
(New Contracts)

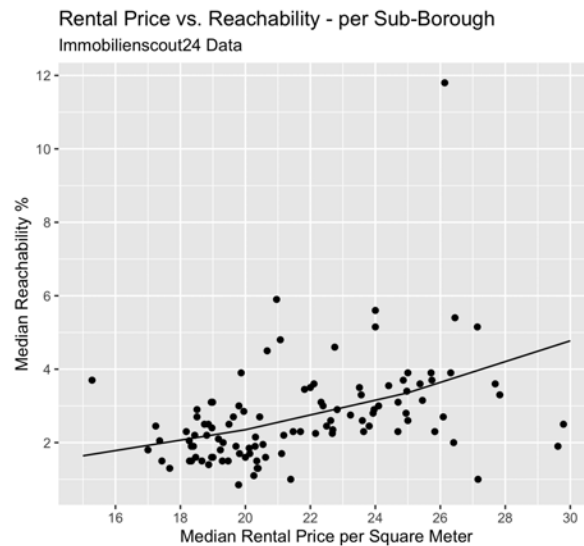


FIGURE 3.15: Regression Immobilienscout24 by Sub-Borough

again the higher price for *Mietspiegel* apartments with rental lengths of less than ten years compared to the full data set on the left. These Figures and Figure 3.15 show the positive trends between rent and reachability in all datasets. As before, when grouping the data by borough, the explained deviance is higher for the model with the reduced dataset at 28.5% compared to the 15.4% explained deviance for the *Mietspiegel* model fitted on the entire dataset, whereas the *Immobilienscout24* has the second-highest deviance explained at 26.8%.

We have now seen the same general trend for both datasets for different aggregations. The explanatory power of the full *Mietspiegel* dataset could be improved by reducing the data to more recent rental contracts and was afterward comparable to the one from the

*ImmobilienScout24* data. Thus, we can conclude there to be a significant trend between the rental price per square meter and the reachability of stations near the corresponding apartments, confirming our initial hypothesis of increased reachability for areas with higher rental prices.

### 3.3.5 Rent to Reachability Ratio

After exploring the relationship between reachability and (sub-)boroughs, it remains to be explored which parts of Munich have good reachability but relatively low rent and vice versa, the expensive ones that do not allow for good reachability. With that goal in mind, we will look at the median rental price per square meter per (sub-)borough and divide it by its median reachability score in percent, giving us a price per reachability percentage. A small ratio means consequently that the reachability can be bought relatively cheaply, while a high ratio means that the borough is expensive for the reachability it offers.

**Rent to Reachability Ratio per Borough**  
Mietspiegel Data

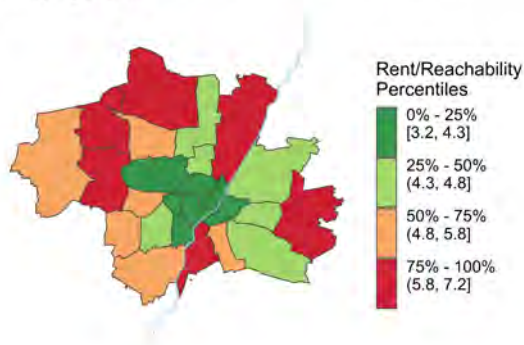


FIGURE 3.16:  
Rent-Reach. Ratio  
By Borough:  
Mietspiegel (Full)

**Rent to Reachability Ratio per Borough**  
ImmobilienScout24 Data

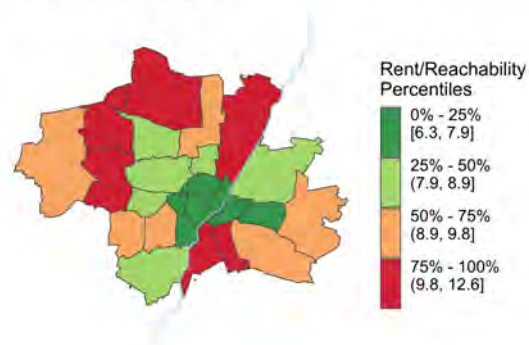


FIGURE 3.17:  
Rent-Reach. Ratio  
By Borough:  
ImmobilienScout24

Figures 3.16 and 3.17 depict these rent to reachability ratios for boroughs and from the full *Mietspiegel* data and *ImmobilienScout24*. We can see that the ratio is generally the smallest in the city center, indicating relatively low rental prices for the reachability one gets. While the boroughs for both datasets are not always in the same quartiles regarding the rent to reachability ratio, there are no stark differences between them.

Figures 3.18 and 3.19 show the ratios for the sub-boroughs and the same datasets as before. Once again, the smallest ratios are found around the city center, indicating that the high price is made up for by good reachability. Noticeably different from that trend in both datasets are the green colored sub-boroughs *Aubing-Süd* in the east, *Obersendling* in the south, and *Gartenstadt Trudering* in the west; all sub-boroughs far away from the city center but having rent to reachability scores greater than the median. On the other hand,



**Rent to Reachability Ratio per Sub-Borough**  
Mietspiegel Data

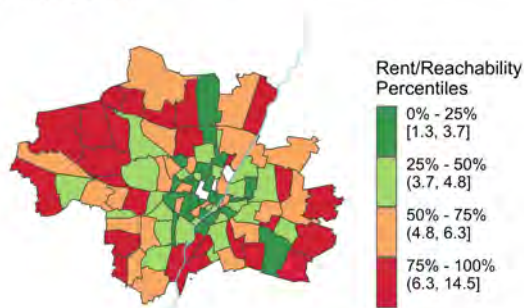


FIGURE 3.18:  
Rent-Reach. Ratio  
By Sub-Borough:  
Mietspiegel (Full)

**Rent to Reachability Ratio per Sub-Borough**  
ImmobilienScout24 Data

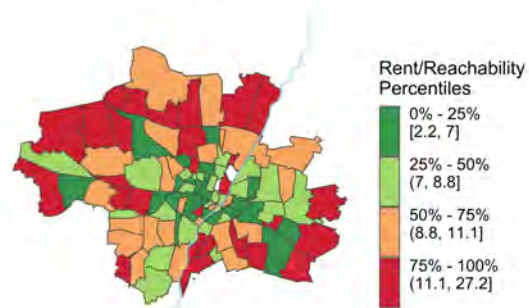


FIGURE 3.19:  
Rent-Reach. Ratio  
By Sub-Borough:  
ImmobilienScout24

the centrally located *Englischer Garten* sub-borough directly west of the Isar has high rents relative to the offered connectedness, being in the third and fourth quartile in the datasets.

Overall, the datasets are in general agreement regarding the rent to reachability quartile for most sub-boroughs, with some exceptions of greater differences, such as the *Freimann* sub-borough, which is dark-green colored for the *Mietspiegel* data, indicating a ratio in the first quartile. At the same time, it is colored orange for *ImmobilienScout24*, signaling the sub-borough to have a ratio above the median for the data.

The same general trends of the *Mietspiegel* data are visible as well when reducing the data to the new contracts, which can be seen in the Appendix A in figures A.4 and A.5.

# 4 Isar as a Potential Barrier for Public Transportation Reachability

In this chapter, we will explore the role the river Isar plays regarding public transport reachability. As it runs straight through the city, our initial hypothesis states, that the river could act as a barrier for public transport reachability, resulting in slower connections when crossing the river compared to connections that run on just one side of it. For that purpose, we will first explore the average public transport speeds and compare the ones running across the river to the ones that are not. Afterward, a regression model for the duration of public transport connections is constructed, with the goal of improving the model by using Isar related information. Finally, we will cluster the stations in the city and explore the role the Isar could play in the resulting clusterings.

## 4.1 Speed of Transportation Methods

Depending on the method of transportation, one moves at different speeds through the city. Naturally, the speed is slower when walking from station to station than taking the subway connection. Between the other transportation products, the speed difference is less clear and therefore worth exploring. The data for the exact distances traveled between stations is not available here, so we need to approximate the average speeds of the different products. The data we do have are distances between all stations traveled by foot, bike, car, and straight-line distances. In this case, only connections without changes are considered. Connections with wait times or changing transportation methods are therefore excluded. In cases with several direct connections between two stations, only the fastest option will be considered, which leaves us with a total of 33,922 direct connections. Naturally, the straight-line distances are the shortest, which translates to the slowest approximated average speeds. Car distances are generally the longest distances and lead to the fastest approximated speeds overall.

The boxplots comparing the four different distances are displayed in Figure 4.1. Calculating the speed based on the bicycle distances leads to a middle ground between the fast speeds of the car distances and the slow ones of the footway distances. Hence, it is a good compromise between the two, and we will conduct our analysis based on the speeds that we get when calculating the distances between stations based on the way one would take by bicycle. For consistency reasons, this distance will be applied to all the public transportation

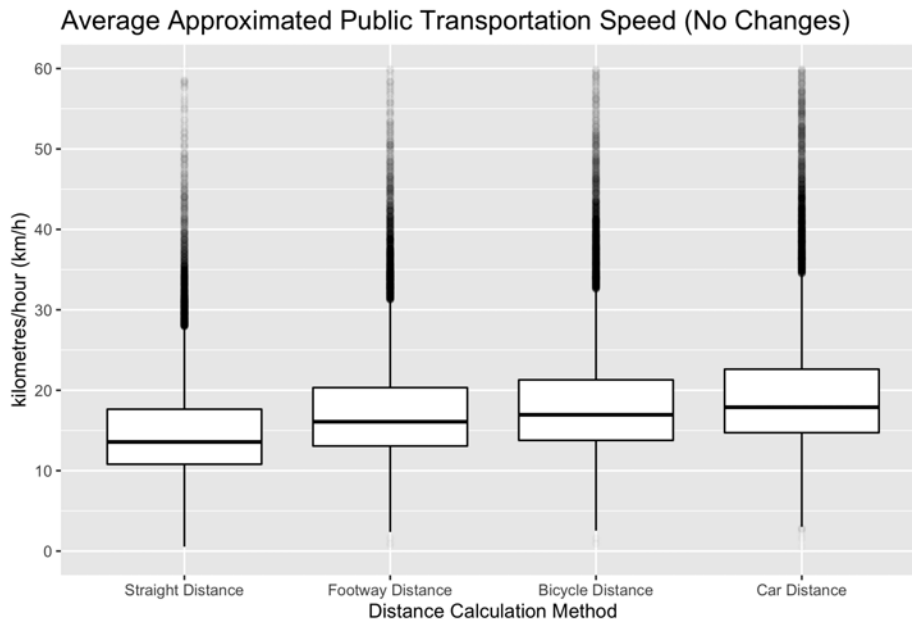


FIGURE 4.1: Speed Comparison by Distance Calculation Method

methods, even though the straight-line distances might be a better approximation for the subway and the car distances might be the best approximation for the bus routes.

Displayed in Figure 4.2 are the average speeds of the different public transportation methods based on the bicycle distances, ordered by their median speed, generally resulting in three different speed categories. The slowest one is walking, followed by the *bus* and *tram* as second speed category and the transportation methods *U-Bahn*, *Bahn* and *S-Bahn* as the fastest methods. These three categories in the same ordering result independent of the distance method used. We have official information regarding their average speeds about three of the transport options. The *U-Bahn's* average speed is 34.8 km/h officially, compared to the calculated average speed of 39.8 km/h; the *tram's* actual speed is 18.5 km/h and 20.2 km/h calculated with the bike distances, while the average official bus speed is 17.9 km/h, while the speed we calculated is 17.5 km/h [18]. Although we have these official speeds, we still do not have information about speeds for connections crossing the river compared to the ones that do not and we will therefore continue to use the calculated speeds. While there is some deviation from the official speeds, the estimates seem to be in the right ballpark, enabling us to continue this analysis.

Around 75% of all stations in Munich serve busses, and consequently, around the same percentage of all direct connections are bus rides. This result is not surprising since it is the most flexible option, and bus stops and connections can be added or removed with relative ease, while the other transportation methods will have to invest into new infrastructure quite heavily before creating new connections and stations. However, this flexibility is limited when crossing a river since crossing one of only a few bridges is required to do so. In this case, the ability of the *U-Bahn* and *S-Bahn* to ride underground seems like the natural

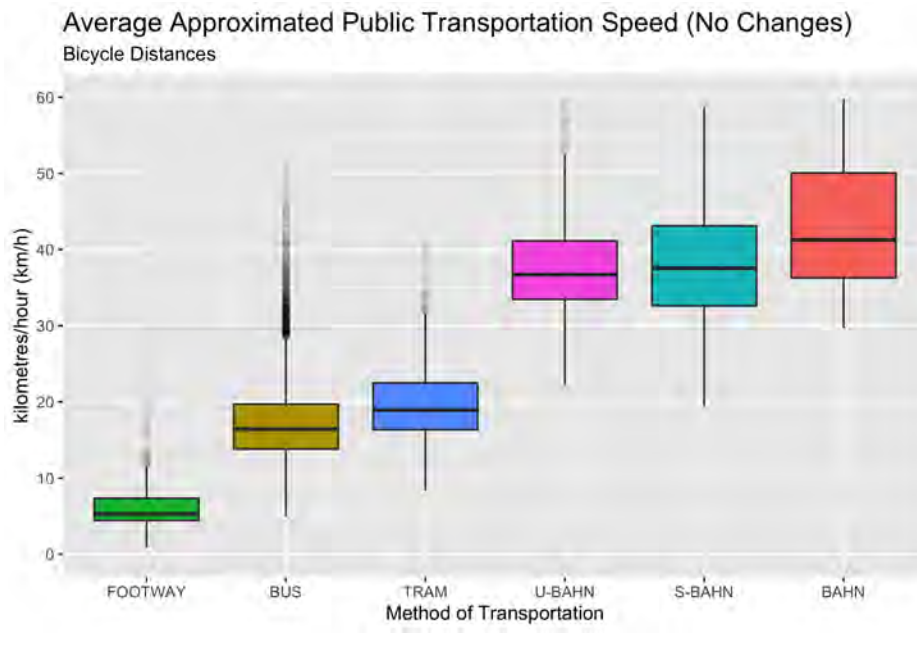


FIGURE 4.2:  
Average Transportation Speed:  
Bicycle Distances

way to efficiently cross the river, and the question arises if there are, in fact, fewer bus rides crossing the river relative to the bus rides that stay on just one side. The result is captured in Figure 4.3 via a bar plot with two groups, where green bars represent direct connections that do not cross the Isar, whereas the blue bars signal that the direct connections have crossed the Isar. There are 3567 such direct connections that cross the river and 30,355 direct connections staying on the same side relative to the Isar.

The y-axis signals the percentage of connections for each product within their group that has been made with the product. For example, only around one percent of all connections that cross the Isar are made by foot, whereas 6.5% of connections that stay on the same side relative to the Isar are footway connections. As we can see, there are noticeable differences for all the products. 78% of connections that do not cross the Isar are busses, but only 52% of those that do cross it are bus connections, which seems to go in line with our theory that the bus flexibility could be better suited for connections staying within the same side. The inverse trend for footway and bus connections can be observed for the other products. Their share increases drastically for Isar crossing connections and especially the *U-Bahn* (5% vs. 22%), and *S-Bahn* (1% vs. 10%) products are used vastly more for the direct connections which lead across the river.

When comparing the average speeds of connections within an Isar side and the ones crossing the Isar in Figure 4.4, we can see that the opposite of our initial hypothesis seems to be true. Average speeds for Isar crossing connections are actually higher than the ones staying within one side.

In order to statistically test the rather vague Isar hypothesis, we need to formulate a

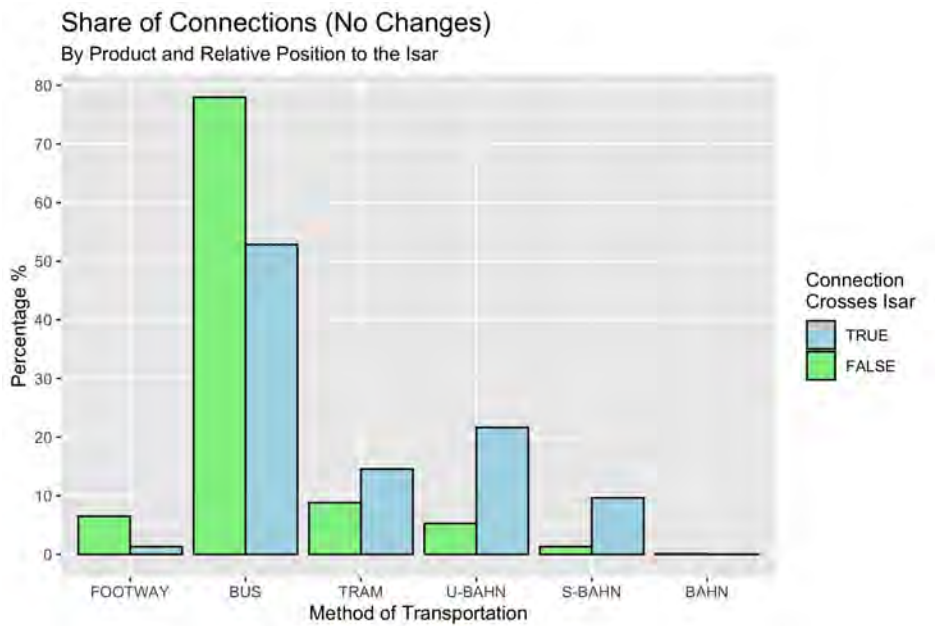


FIGURE 4.3:  
Share of Direct Connections:  
By Product and Isar Position

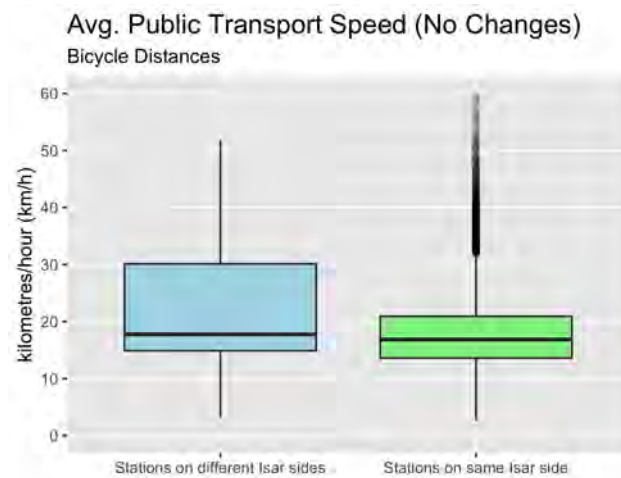


FIGURE 4.4:  
Average Transportation Speeds  
in Relation to Isar

concrete null hypothesis. First, it is defined as the thesis stating that the average speed of all connections crossing the Isar is greater or equal to the average speed of all connections not crossing the Isar.

$$\begin{aligned}
 H_0 : \bar{v}_{isar\_crossed} - \bar{v}_{isar\_not\_crossed} &\geq 0 \\
 &vs. \\
 H_1 : \bar{v}_{isar\_crossed} - \bar{v}_{isar\_not\_crossed} &< 0
 \end{aligned}
 \tag{4.1}$$

When conducting the corresponding Welch two sample t-test, as shown in Table 4.1, the null hypothesis cannot be rejected at the 5% level.

**Welch Two Sample t-test**

x:  $\bar{v}_{isar\_crossed}$   
 y:  $\bar{v}_{isar\_not\_crossed}$   
 alt. hypothesis: true difference in means is less than 0

t	df	p-value	95 percent conf. interval	mean of x	mean of y
17.168	4550.3	1	[-Inf, 3.41]	22.22	19.11

TABLE 4.1: t-test Transportation Speed and the Isar

However, this was to be expected, since the box plots in 4.4 already indicated a *higher* average speed for the Isar crossing connections. One main reason for the increased speeds is the predominant usage of the fast subways instead of busses when crossing the river, as seen before in Figure 4.2.

Accordingly, we can conduct another test with a slightly modified null hypothesis where the connections are filtered to the ones using only the subway (*U-Bahn*).

$$\begin{aligned}
 H_0 : \bar{v}_{isar\_crossed\_subway} - \bar{v}_{isar\_not\_crossed\_subway} &\geq 0 \\
 &vs. \\
 H_1 : \bar{v}_{isar\_crossed\_subway} - \bar{v}_{isar\_not\_crossed\_subway} &< 0
 \end{aligned}
 \tag{4.2}$$

When conducting the Welch t-test with the filtered data and the new null hypothesis, it is now possible to reject the null hypothesis at the  $\alpha = 0.05$  level with  $p - value < 2.2e - 16$ . The exact results are shown in 4.2.

**Welch Two Sample t-test**

x:  $\bar{v}_{isar\_crossed\_subway}$   
 y:  $\bar{v}_{isar\_not\_crossed\_subway}$   
 alt. hypothesis: true difference in means is less than 0

t	df	p-value	95 percent conf. interval	mean of x	mean of y
-16.112	2369.9	< 2.2e-16	[-Inf, -5.78]	35.49	41.93

TABLE 4.2: t-test Subway Speed and the Isar

Interestingly, these test results do not only occur when filtering on the subway product, but also when filtering on any single product - *S-Bahn*, *Tram*, *Bus*, *Bahn* and also *Footway*. Hence, the seemingly clear trend from before where public transportation seemed faster when crossing the Isar reverses completely when looking at the individual product level. What we have here is an example of *Simpson's paradox*, in which a trend for several combined groups of data disappears or reverses when split into their respective groups [26].

Since all public transportation methods in Munich are faster on average when not crossing the Isar, it can therefore be stated, that the Isar is actually a barrier for public transportation reachability. By aptly using the fast subway more frequently in transport planning, this fact is almost made irrelevant however.

## 4.2 Modeling Public Transportation Ride Times

After confirming differences in the average product speeds related to the Isar, we can now model the duration of the shortest connection between all stations. Hence, we now also consider connections again where changes are necessary, resulting in 1,197,785 observations. First, we will construct a base model with the distance between stations as the only independent variable. This is where the relationship between the variables seems to be the most clear - longer distances between stations generally mean longer ride times. We will use bicycle distances between stations again, along with the ride time in minutes of the shortest connections between all stations. Figure 4.5 shows the relationship between these variables in a scatter plot with hexagonal binning, which divides the space into hexagons, counts the number of observations in each hexagon and plots these hexagons color coded by the number of observations in each one. From these hexagons and the density plots on the top and on the right, we can see how most MVG connections lie in the light blue area from about zero to 20,000 meters with a ride time of less than an hour. The light blue area seems to appear close to a logarithmic or square root function and allows us to linearize the relationship by log- or square root transformation. As we can see in Figure 4.5, the distributions are not heavily skewed and the less aggressive transformation variant square root might be the better fit. While in this case, log transforming both variables would actually achieve a slightly better model fit than the square root as measured by the  $R^2$  metric, residual analysis shows, that the homoscedasticity assumption is hurt severely and that the residuals deviate from the normal distribution. Since the linear regression assumptions are fulfilled (see Figure A.6 in Appendix A ) for the model with the square root transformed independent variable and its fit is almost as good as the log-log transformed one, we can go along with it from here on. Consequently, our linear regression model takes on the following form.

$$\widehat{duration\_mvg} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \sqrt{distance\_bike}. \quad (4.3)$$

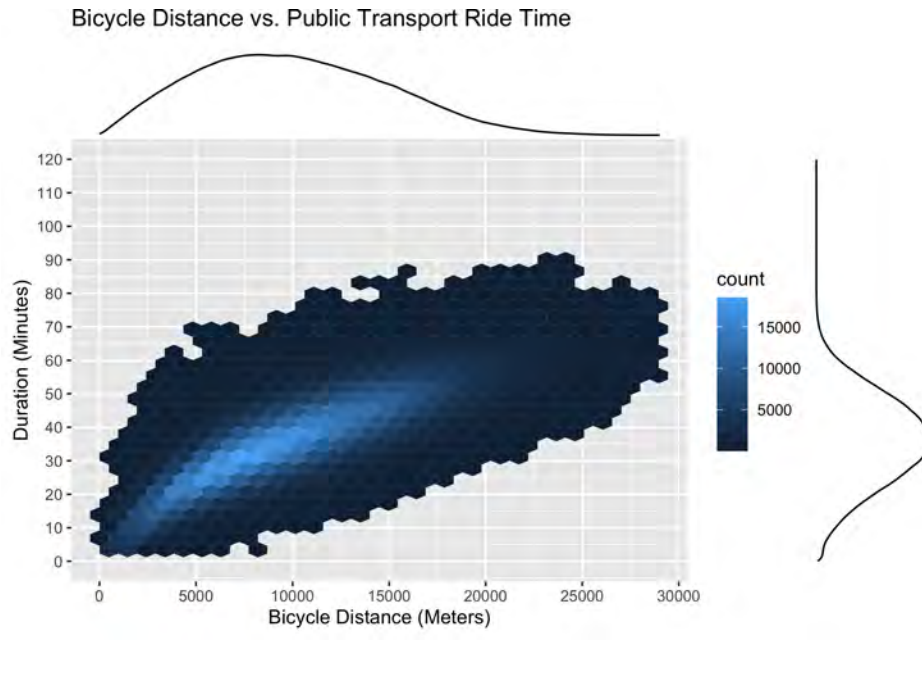


FIGURE 4.5: Distance vs. MVG Ride Time

Table 4.3 shows the regression results where  $\beta_1$  is estimated as 0.435, which in the context of the square root transformed independent variable means, that the duration is c.p. increased by 0.435 for an increase of the bike distance's square root by one meter. Note however, that the intercept is negative and connections with a bike distance of less than 214 meters will be predicted as having negative ride times. As only 0.01% of all connections' start and end stations are separated by less than 214 meters, we will be going on with the model regardless. Both, the intercept and the distance are statistically significant at the 5% level and the model shows an  $R^2$  value of 0.726, indicating that 72.6% of the variance in the data can be explained by the model. When plotting the regression line onto the untransformed data in Figure 4.6, we can see how the model follows the data nicely.

After we have established a solid base model, we can now add variables related to the Isar, to check if they contain information that can improve it. One of the simplest ones we can add is the binary variable *same\_isar\_side*, indicating whether the start and end stations of a connection are on the same side of the Isar or not.

$$\widehat{duration\_mvg} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \sqrt{distance\_bike} + \hat{\beta}_2 \cdot same\_isar\_side, \quad (4.4)$$

In order to test whether the new variable is improving the model in a meaningful way, we conduct the analysis of variance (ANOVA) for the models to test if the more complex model has significantly reduced residual sum of squares (RSS) compared to the simpler model. For this purpose, the *F-test* is used, where the null hypothesis states, that the models do not differ significantly in the amount of variance they explain [8]. In Table 4.4 we can see the result of the ANOVA for the base model and the one with the added Isar variable. As the p-value for the F-test is smaller than 0.05 we can reject the null hypothesis at this



Dependent variable:	
duration_mvg	
distance_bike	0.435*** (0.0002)
Intercept	-6.284*** (0.025)
Observations	1,197,785
R <sup>2</sup>	0.726
Adjusted R <sup>2</sup>	0.726
Residual Std. Error	6.845 (df = 1197783)
F Statistic	3,171,729.000*** (df = 1; 1197783)
Note:	*p<0.1; **p<0.05; ***p<0.01

TABLE 4.3: Linear Regression: Distance vs. MVG Duration

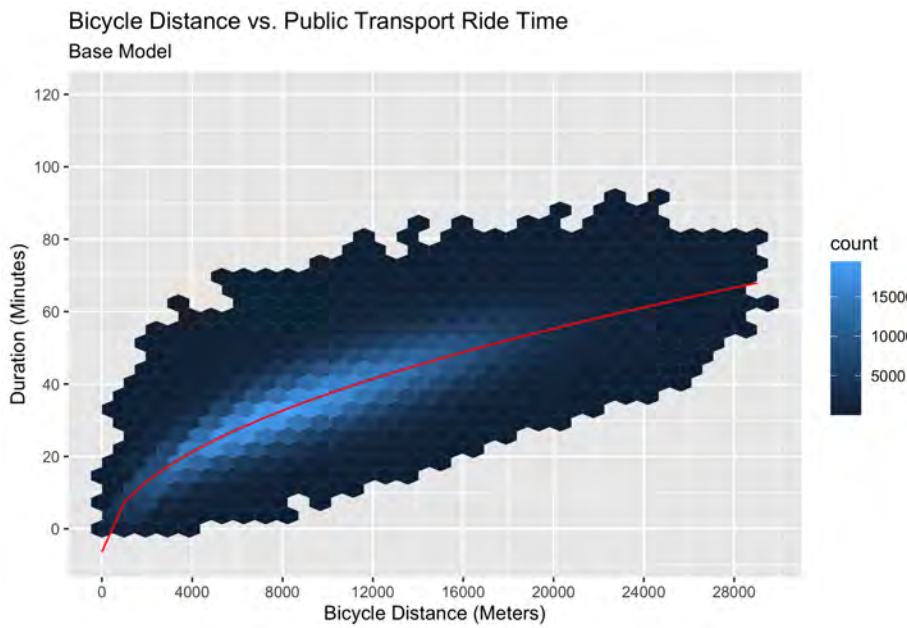


FIGURE 4.6: Distance vs. Time: Base Model

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1197783	56127391				
2	1197782	56023607	1	103785	2219	0.0000

TABLE 4.4: ANOVA output for models 4.3 & 4.4

level and conclude, that the *same\_isar\_side* variable significantly improves the model. In Table 4.7 the exact regression outputs for all models are shown. As with the the average

speed of all products before, the model suggests that the ride times are generally longer, if the connection stays within one side of the Isar, although this difference is *ceteris paribus* only 0.668 minutes. While the added variable is statistically significant at the 5% level, the adjusted  $R^2$  value increases only slightly from 0.726 to 0.7264.

$$\begin{aligned} \widehat{duration\_mvg}_i = & \hat{\beta}_0 + \hat{\beta}_1 \cdot \sqrt{distance\_bike_i} + \\ & \hat{\beta}_2 \cdot same\_isar\_side + \\ & \hat{\beta}_3 \cdot isar\_dist\_start + \hat{\beta}_4 \cdot isar\_dist\_end, \end{aligned} \tag{4.5}$$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1197782	56023607				
2	1197780	54673872	2	1349735	14785	0.0000

TABLE 4.5: ANOVA output for models 4.4 & 4.5

The next variables we can add to the model are the distance in meters to the Isar from the start and end station respectively. The idea here is that being far away from the Isar at the beginning or end of a connection signals that the station is either far east or far west, where reachability is generally worse at the city limits, and consequently, the ride times would be longer. The ANOVA 4.5 shows how those variables improve the explanatory power of the model; in Table 4.7 the coefficients are positive and indicate increased average ride times the further away from the Isar a connection starts and ends. However, the *same\_side\_isar* variable is now negative, but reduced to being close to zero.

$$\begin{aligned} \widehat{duration\_mvg} = & \hat{\beta}_0 + \hat{\beta}_1 \cdot \sqrt{distance\_bike_i} + \\ & \hat{\beta}_2 \cdot same\_isar\_side + \hat{\beta}_3 \cdot isar\_dist\_start + \hat{\beta}_4 \cdot isar\_dist\_end + \\ & \hat{\beta}_5 \cdot (same\_side\_isar : isar\_dist\_start) + \\ & \hat{\beta}_6 \cdot (same\_side\_isar : isar\_dist\_end) \end{aligned} \tag{4.6}$$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1197780	54673872				
2	1197778	54083846	2	590026	6534	0.0000

TABLE 4.6: ANOVA output for models 4.5 & 4.6

When being far away from the Isar when starting and ending the connection, it matters a lot whether both stations are on the same side of the Isar or not. If they are, it is further indication that both stations are near each other while being on different sides of the river would indicate higher ride times. That is why we add interaction terms between the binary *isar side* variable and the distances to the isar for start and end stations, resulting in our

final model 4.6. The ANOVA once again confirms model improvement by adding these variables and the regression output Table 4.7 shows an increased  $R^2$  value of 0.736 and all variables being statistically significant at the 5% level. For  $same\_side\_isar = 0$ , the model is reduced to

$$\begin{aligned} \widehat{duration\_mvg} = & -6.397 + 0.434 \cdot \sqrt{dist\_bike} + \\ & 0.00006 \cdot isar\_dist\_start - 0.00007 \cdot isar\_dist\_end. \end{aligned} \quad (4.7)$$

While these Isar distance variables are statistically significant, their coefficients are quite small, even a distance of 15,000 meters from the start station to the Isar, the maximum value in the dataset, would increase the ride time in the model by merely 0.9 minutes, while the same distance between the end station and the Isar would decrease the ride time by only 1.05 minutes. Hence, the distances variables can only influence the ride times by about a minute in each direction. Consequently, when crossing the Isar the model is basically reduced back to the base model, where the square root transformed distance between stations is the most influential variable.

When  $same\_side\_isar = 1$  on the other hand, we can simplify the full model as

$$\begin{aligned} \widehat{duration\_mvg} = & -6.397 - 2.792 + 0.434 \cdot \sqrt{dist\_bike} + \\ & (0.00006 + 0.0004) \cdot isar\_dist\_start + (-0.00007 + 0.0003) \cdot isar\_dist\_end = \\ & -9.189 + 0.437 \cdot \sqrt{dist\_bike} + 0.00064 \cdot isar\_dist\_start + 0.00027 \cdot isar\_dist\_end \end{aligned} \quad (4.8)$$

Compared to the Isar crossing connections, the average ride times are now generally reduced by 2.792 minutes, while the distance is increased by 0.0004 minutes for every meter the start station is away from the Isar and is further increased by 0.0003 minutes for every meter the end station is away from the Isar. This results in the fact, that ride times seem to be reduced compared to Isar crossing connections, as long as start and end station are not far away from the Isar. However, one explanation for these reduced ride times when not too far from the Isar could be the generally better connected stations near the city center.

Overall, we have seen how adding information regarding the Isar could help improve modeling the public transport ride times and that connections that do not cross the Isar have reduced ride times, as long as their start and end station distances to the Isar are not too big.

<i>Dependent variable:</i>				
<b>duration_mvg</b>				
	Model 4.3	Model 4.4	Model 4.5	Model 4.6
sqrt(dist_bike)	0.435*** (0.0002)	0.441*** (0.0003)	0.420*** (0.0003)	0.434*** (0.0003)
same_side_isar		0.668*** (0.014)	-0.098*** (0.015)	-2.792*** (0.028)
isar_dist_start			0.0003*** (0.00000)	0.00006*** (0.00000)
isar_dist_end			0.0002*** (0.00000)	-0.00007*** (0.00000)
same_side_isar: isar_dist_start				0.0004*** (0.00000)
same_side_isar: isar_dist_end				0.0003*** (0.00000)
Intercept	-6.283*** (0.025)	-7.237*** (0.032)	-6.942*** (0.031)	-6.397*** (0.032)
Observations	1,197,785	1,197,785	1,197,785	1,197,785
Adjusted R <sup>2</sup>	0.726	0.726	0.733	0.736
F Statistic	3,171,729***	1,589,911***	821,971***	556,136***

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 4.7: Comparison of MVG duration regression models

### 4.3 Cluster Analysis

After conducting statistical tests and regression analysis, another method for figuring out the Isar's influence is using unsupervised methods. More specifically, clustering methods can be used to create station partitions within the city based on the ride times between them.

The general goal of clustering is identifying structures or clusters solely based on the unlabeled observations  $x_i \in \mathbb{R}^n$ . Observations within a cluster are then supposed to be as similar as possible to each other, while elements in different clusters should be as dissimilar as possible. As such, intra-cluster similarity should be high and inter-cluster similarity low for the resulting clusters. The resulting partitions are also called a *clustering* [14].

### 4.4 Clustering Evaluation

When such clusterings are created, one needs to be able to compare different results to decide which ones are the best fitting outcomes for the current scenario and there are many methods trying to solve this problem. Depending on the problem, it is possible to create a single quality score resulting purely from the outcome data, called *internal* evaluation. In other scenarios there exists a ground truth, where one can compare the resulting cluster assignments to the actual classifications of the objects, which is called *external* evaluation [9]. Other situations demand expert intervention, calling on a person familiar with the subject matter to judge the situation. In our case, we will look at one external method, the *adjusted Rand Index* in order to compare the resulting clusterings to the western and eastern Isar sides in Munich.

#### 4.4.1 Adjusted Rand Index

The *Rand Index* measures similarity of different clusterings by comparing all possible observation pairs between them, where the number of observations  $n$  needs to be the same in both clusterings. Hence,  $\binom{n}{2}$  comparisons are made and for each of these comparisons of clusterings  $C_1$  and  $C_2$ , one of four results is possible:

- **a:** same cluster in  $C_1$ , same cluster in  $C_2$
- **b:** same cluster in  $C_1$ , different cluster in  $C_2$
- **c:** different cluster in  $C_1$ , same cluster in  $C_2$
- **d:** different cluster in  $C_1$ , different cluster in  $C_2$

Therefore, the clusterings agree in **a** and **d**, and when dividing the number of matching assignments with all comparisons, we get the Rand Index (RI) as [23]:

$$RI = \frac{a + d}{a + b + c + d} = \frac{a + d}{\binom{n}{2}} \quad (4.9)$$

As such,  $RI \in [0, 1]$ , where  $RI = 0$  means there are no agreements and  $RI = 1$  signals perfect agreement between the clusterings. However, one would expect some agreements by chance alone, which is why the *adjusted* Rand Index (ARI) can be used as an alternative. It adjusts the Rand Index by considering the expected number of agreements by chance, making the expected value of the ARI 0 and giving it a range of  $ARI \in [-1, 1]$  [13]. For clustering similarities worse than the expected number of agreements, the ARI can consequently now become negative as well.

The comparisons are done the same way as before for the Rand Index. Formally, the adjusted Rand Index is then calculated as follows [24]:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{n}{2} - [(a+b)(a+c) + (c+d)(b+d)]} \quad (4.10)$$

It is common to use the (Adjusted) Rand Index to compare a clustering to an existing classification, a ground truth, in order to judge the clustering algorithm. In our case, we are interested in finding out, whether the stations are clustered into a partition east and west of the Isar. Consequently, we will be using the classification of east and west of the river as our ground truth when using the ARI in this section.

## 4.5 Cluster Analysis Algorithms

A multitude of different cluster algorithms exist, and two popular approaches are *hierarchical* and *k-Means* clustering which we will explore in detail now.

### 4.5.1 Hierarchical Clustering

Generally, there are two types of hierarchical clustering, the agglomerative and the divisive variant. The divisive one starts in the situation, where all observations are clustered together and this single cluster is split up step by step until each observation is its own single cluster. The agglomerative variant, the one we will be using, starts with each observation being its own cluster and clusters are merged together until all observations are assigned to the same cluster.

Hence, in each step we want to merge the observations that are the most similar. This similarity is quantified by a metric and in our case we choose the *Euclidean distance*, defined as

$$d(a, b) = \|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2},$$

where  $a$  and  $b$  are single observations from the dataset [14].

When comparing the distance between sets of observations, we need to use a *linkage function*, which defines pairwise comparisons between the observations from the two clusters being compared. Once again, several options are possible, such as taking the minimum

or maximum distance between observations of the clusters. We will be using the average distance between all observations. Therefore, the distance between two clusters  $A$  and  $B$  is the average of all Euclidean distances  $d(a, b)$  between pairs of observations  $a \in A$  and  $b \in B$ . Formally, we can describe the average linkage function as [11]

$$d_{AverageLinkage}(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

---

**Algorithm 1** Hierarchical Clustering [14]
 

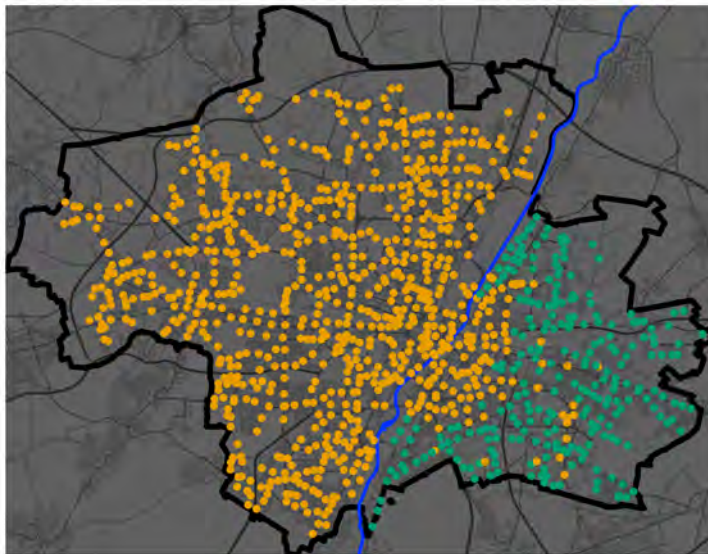
---

1. Begin with  $n$  observations and a distance measure of all the  $\binom{n}{2} = n(n-1)/2$  pairwise distances and treat each observation as its own cluster.
  2. for  $i = n, n-1, \dots, 2$ :
    - (a) Examine all pairwise inter-cluster distances among the  $i$  clusters and identify the pair of clusters that is the most similar, hence, the one with the smallest inter-cluster distance as calculated by the chosen linkage function. Merge these two clusters.
    - (b) Compute the new pairwise inter-cluster distances among the  $i-1$  remaining clusters.
- 

As we can see, the number of clusters ranges from the number of observations  $n$  to just one cluster and it is up to the practitioner to choose the final number of clusters one deems to be right for the situation.

**Agglomerative Hierarchical Clustering Result**

2 Clusters




---

 FIGURE 4.7: Hierarchical Clustering of MVG Network

Applied to our MVG dataset, we use the  $1095 \times 1095$  distance matrix as input, with  $n_{ij}$  as shortest distance between stations  $i$  and  $j$ . As we are mainly interested in noticing any effects the Isar might have, we will choose the number of clusters as 2, testing whether the city is divided into two parts along the riverside. The result is shown in Figure 4.7, where all stations are colored by the cluster they were assigned. Overall, there seems to occur an east-west separation and the Isar line is the boundary between clusters in the north and the south. The orange colored cluster traverses across the river through the city center however and there are many stations being clustered in the eastern part of the city. When assuming the separation of the city along the Isar line as ground truth, we get an adjusted Rand Index of 0.655, while a score of 1 would indicate a perfect match to the ground truth.

## 4.5.2 k-Means Clustering

The basic idea of k-Means clustering is finding  $k$  different clusters, where the observations within the clusters should be as *similar* as possible and the observations between clusters should be as *dissimilar* as possible. Each cluster has a cluster representative, usually calculated as the center of the cluster, and the Euclidean distances between these representatives and the observations are calculated to determine similarity to each cluster. The formal goal then becomes finding a clustering  $C$  with  $k$  clusters and cluster representatives  $\bar{x}_r$ , that minimizes the overall distance from the observations to the cluster representatives [12]:

$$\min_C \sum_{r=1}^k \sum_{x_i \in C_r} \|x_i - \bar{x}_r\|^2 \quad (4.11)$$

Several algorithms exist to find possible solutions, one of them is the following.

---

### Algorithm 2 k-Means Algorithm [10]

---

1. Initialization: Choose  $k$  arbitrary representatives
  2. Repeat
    - (a) Assign each observation to the cluster with the nearest representative as measured by Euclidean distances
    - (b) For each cluster  $r$ , calculate the new cluster centroids (representatives) as 
$$x_r = \frac{1}{|C_r|} \sum_{x_i \in C_r} x_i$$
    - (c) **STOP**, if cluster representatives do not change or a previously defined number of iterations  $n$  is achieved
- 

Figure 4.8 shows an example of k-Means with  $k = 2$ . The final result depends on the randomly picked initial cluster representatives, while the procedure generally only finds local optima [14]. For that reason it is customary to repeat the algorithm many times with changing initializations, picking the result that minimizes equation 4.11.



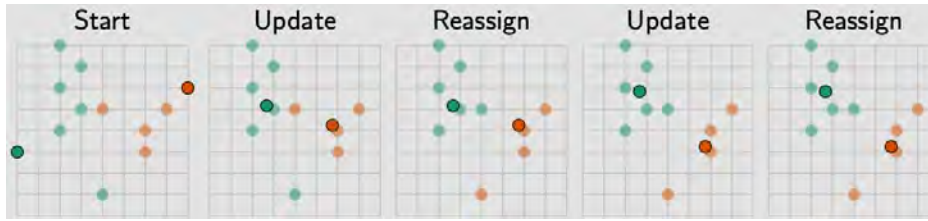


FIGURE 4.8: k-Means Clustering Example [25]

When applying the procedure to our MVG dataset and its  $1095 \times 1095$  distance matrix as before, and choosing  $k = 2$ , we get the result shown in Figure 4.9.

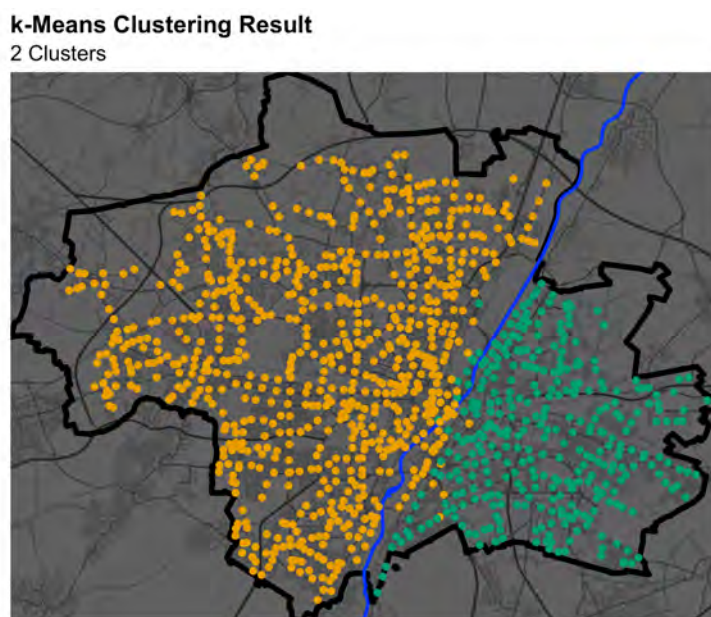


FIGURE 4.9: k-Means Clustering of MVG Network (2 Clusters)

Now, we notice how the city is cut into two along the Isar, although several stations close to the Isar still belong to the cluster on the other side of the river. However, the adjusted Rand Index is 0.888 now when considering the perfect separation of stations along the Isar as ground truth, indicating a very good match between the clusterings.

Overall, we now have a result for hierarchical clustering, where the Isar's role is rather minor and a k-Means result with near perfect separation along the Isar, indicating how the river might play a role regarding public transport, but its influence is not strongly noticeable for every clustering algorithm.

When choosing other cluster numbers  $k$ , we often do not achieve clusterings where the Isar has a noticeable influence. For  $k = 5$ , for instance, shown in Figure 4.10, the result is closer to the hierarchical cluster result from before regarding the Isar separation. Now it seems like the city seems to be clustered in a northern, eastern, southern, western and

central part, although the central cluster stretches out deeply into each of the four other clusters.

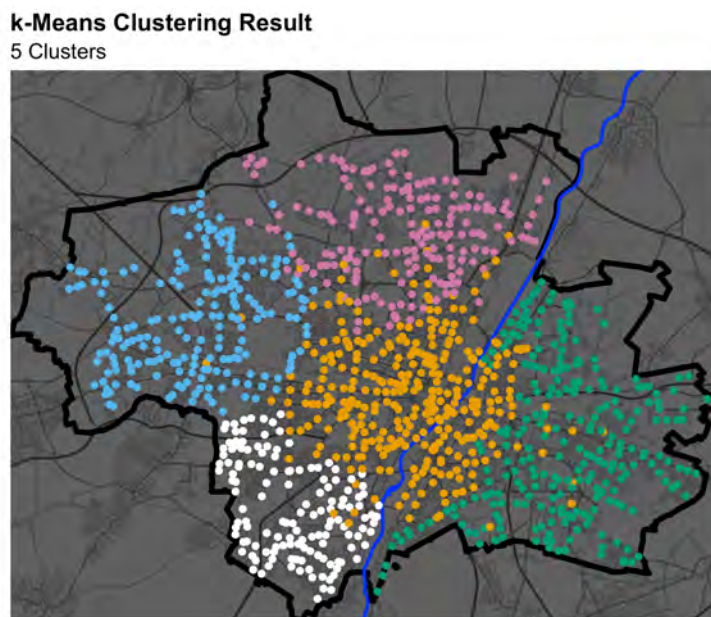


FIGURE 4.10: k-Means Clustering of MVG Network  
(5 Clusters)

In summary, we have now seen differing speeds of public transit when crossing the Isar, could achieve an improved regression model for the transit connections and were able to show a clustering separating the city in two across the Isar. While the overall speeds of each single product is slower when crossing the Isar, this is compensated by using faster products overall when crossing it. The final regression model showed an increased connection duration when being close to the Isar at the beginning or end of the connection, when crossing it, but this trend reversed when being rather far away at the end or the beginning of the journey. Overall this leads to the conclusion, that the Isar does play a minor role in Munich's public transportation system. It cannot be clearly shown, that it acts as a strongly negative influence or barrier for public transport connectivity, however.

## 5 Graph Analysis

It is possible to describe a public transportation network as a graph, allowing the usage of several established algorithms to determine how central the stations are within the transportation network. For this purpose, we will first describe the basic notions of graphs and how the public transit network can be seen as one. Afterward, several centrality measures will be introduced to quantify the location quality of stations within the network.

### 5.1 Introduction to Graph Theory

In its simplest form, a network can be structurally described by a graph where we write  $G = (V, E)$ , with  $V$  as a set of vertices or nodes and  $E \subseteq \{\{x, y\} | x, y \in V, x \neq y\}$  as the set of edges connecting the vertices. The order of these vertices is irrelevant in this case, and we get an *undirected simple graph*. The number of nodes  $|V|$  is also called *order* and the number of edges  $|E|$  the *size* of a graph. The *distance*  $d(u, v)$  between two nodes  $u$  and  $v$  is the number of edges in a shortest path connecting them [27]. Not allowed in this simple graph are, per definition, several distinct vertices between edges. Figure 5.1 shows an example of such a simple graph with an order of four and the size of three.

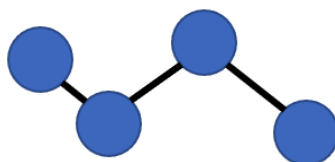


FIGURE 5.1: Simple Graph

When the order of nodes is important in a connection, one can use the directed simple graph, also called *digraph*. The set of edges is then defined as  $E \subseteq \{(x, y) | (x, y) \in V^2, x \neq y\}$ , where the set of edges now consists of an ordered pair of vertices instead of the unordered pair before.

Notice that now the distance  $d(u, v)$  between two vertices might be defined in one direction, while the other direction  $d(v, u)$  would describe the distance of a path that does not exist and is therefore not defined. Additionally, it is possible to assign weights to edges, called *edge weights*  $\omega_e$ , resulting in a *weighted graph*. This weight is typically given by a weight function  $\omega : E \rightarrow \mathbb{R}$  and the distance between two vertices is now defined as the sum of these edge weights in the shortest path connecting them [27].

Apart from representing a graph visually by drawing nodes as circles and the edges between them as lines, it is also common to store them in an **adjacency matrix** [7]. For this purpose, we assume that the nodes of a graph are numbered  $1, 2, \dots, |V|$  in some arbitrary manner. The resulting adjacency matrix representation of a simple graph then consists of a  $|V| \times |V|$  matrix  $A = (a_{ij})$ , such that

$$a_{ij} = \begin{cases} 1, & (i, j) \in E \\ 0, & \text{otherwise} \end{cases}$$

Figure 5.2 depicts such a transformation from the visual graph to the adjacency matrix. Each node now has an arbitrarily assigned number, and the matrix displays the presence of edges between these vertices with binary values.

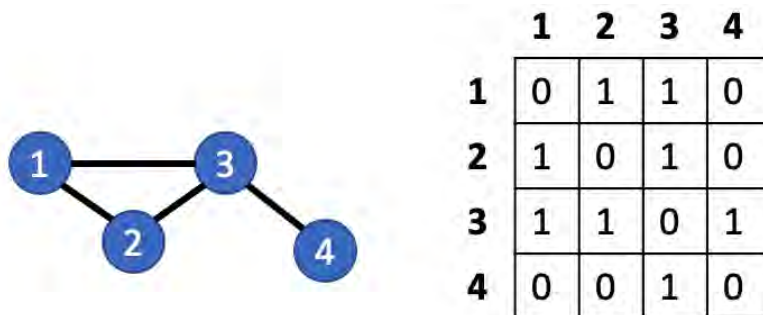


FIGURE 5.2: Graph Adjacency Matrix

This definition can also be adapted to weighted graphs. For the weighted graph  $G = (V, E)$  with the edge function  $\omega$ , instead of storing binary values in the adjacency matrix, we can use the weight  $\omega(u, v)$  of the edge  $(u, v) \in E$  as the entry in row  $u$  and column  $v$ . For non-existing edges *NULL* values or 0 and  $\infty$  are typically used as matrix entries.

In Figure 5.3, a weighted digraph is shown in visual form on the left and matrix form on the right. The edge weights are noted next to their corresponding edge and emphasize that the edge length does not necessarily correlate with the edge weights but can be arbitrarily chosen for a fitting visual representation. While the simple graph's adjacency matrix is symmetric, this is not necessarily the case anymore for digraphs.

The concepts above can now be directly applied to Munich's public transportation network by using the definition of a weighted digraph. Each station is a node, direct connections are the edges in the graph, and the ride time between the stations constitutes the edge weight. Notice how it is common that several connections are going from station A to station B in a public transportation timetable and how that violates the definition of a weighted digraph where only one edge with the same tail and head is allowed. Since slower alternatives are usually not an option when route planning, we will only consider the fastest option between all stations going forward, and in case of a tie between two connections, a random one will be picked.

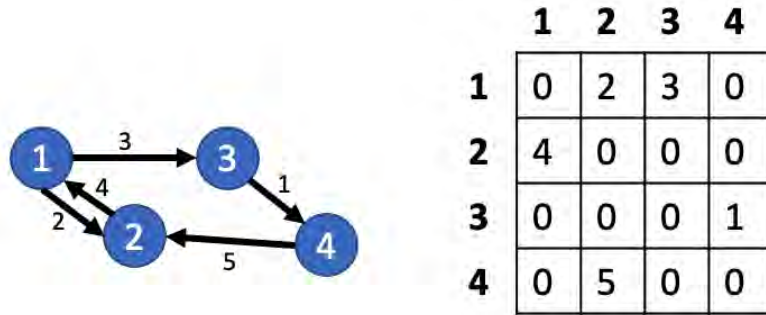


FIGURE 5.3: Weighted Digraph Adjacency Matrix

Almost all direct connections in the data are symmetric in the network. When we go from station A to station B, where no further stop lies between these two stations, it is usually possible to go back the same way in about the same amount of time.

However, some graph algorithms and centrality measures are not clearly defined for undirected graphs [19]. For that reason, the transportation network will be transformed into a graph in a simplified form. An undirected, weighted graph will be used, where the weights between the nodes are the average ride durations between the stations. Footway connections were only considered if they were five minutes or less, and no other connection was faster. The resulting graph is of order 1095, size 1768, and is displayed in Figure 5.4.

The stations as nodes are represented as circles, and the connections as edges are the lines between these nodes. The edges are color-coded by product, and it can be observed that most connections are served by busses. We can also see how the footway option, colored in green, is widespread in the city center and how the western part of Munich is more frequently connected by (*S*-)Bahn connections instead of the subway.

## 5.2 Centrality Measures

Centrality Measures try to quantify the importance of nodes in a graph. Like the reachability concept, high centrality measures of nodes could express a good location of the surrounding area. However, there are different approaches to determining this importance, and depending on the graph and the researcher's current interests, nodes can be of drastically different relevance. Hence, there are several different centrality measures, and we want to introduce some of them and apply them to our public transportation network.

### 5.2.1 Degree Centrality

Probably the most intuitive centrality measure is the *degree centrality*, which is based on the concept of the graph-theoretical *degree*. The degree  $deg(v)$  of a node  $v$  is the number of edges that are incident to the node  $v$  [5]. Applied to the example in Table 5.5, nodes 1, and 4



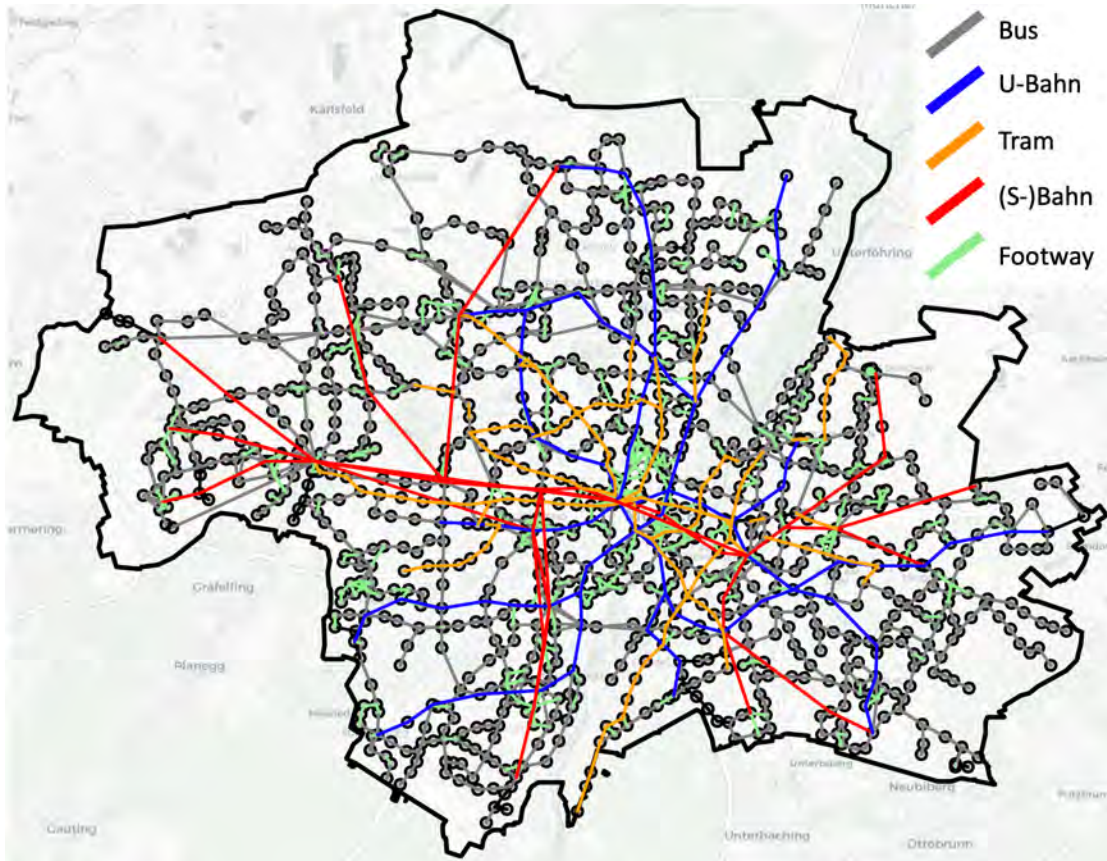


FIGURE 5.4: Public Transport Network as a Graph

have a degree of  $\deg(1) = \deg(4) = 1$ , while node 3 has two edges and therefore  $\deg(3) = 2$  and node 2 has no edges and  $\deg(2) = 0$ .

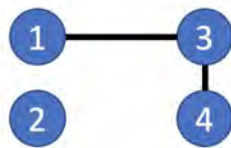


FIGURE 5.5: Degree Example

Now, a node's degree *centrality* is the same as its degree,  $C_{Deg(v)} = \deg(v)$ . Since the degree is an absolute number that can vary drastically depending on the graph order and its structure, it can be more intuitive analyzing the normalized degree centrality by dividing a node's degree by the maximum possible degree in the graph,  $n - 1$ , where  $n$  is the number of nodes in the graph.

$$C_{NormDeg(v)} = \frac{\deg(v)}{n - 1} \quad (5.1)$$

Going forward, we will be using both versions intermittently. Applied to Munich's transportation network, stations with direct connections to several different stations will

have high degree centrality measures while other stations, such as the last station of a subway connection line, will have low centrality. However, duration times are not considered, and each edge is considered to be equally important when calculating the degree centrality.

### Degree Centrality by Station

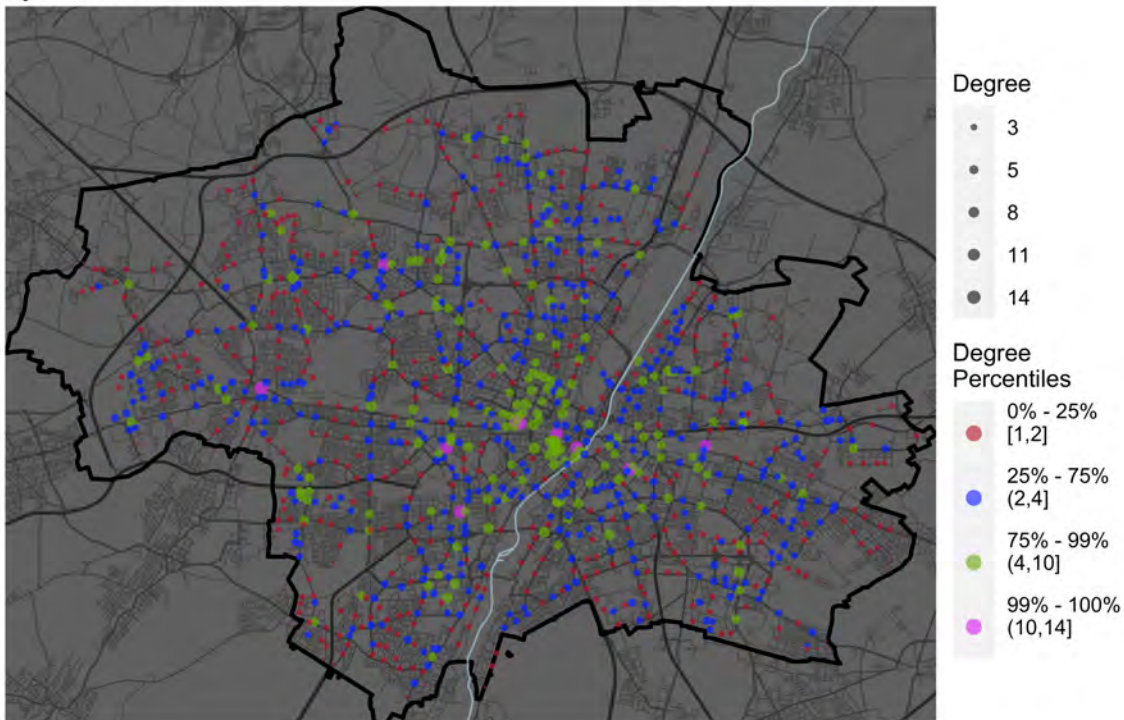


FIGURE 5.6: Degree Centrality

Nonetheless, when considering the stations with the highest degree centrality values, shown in Table 5.1, one can see how these are one of the main hubs in Munich, such as the central station with a degree centrality of 14, while the the average degree of all stations is 3.2. Concretely, the table displays the stations with a degree in the top percentile in the dataset. These stations are colored pink in Figure 5.6. Note that the central station is split into four distinct stations in the MVG network, *Hauptbahnhof (S, U, Bus, Tram)*, *München Hbf*, *Hauptbahnhof Süd* and *Hauptbahnhof Nord* where each of these is located in or around the central station. In Table 5.1, we can see how two of these four stations are in the list of stations with the highest degree values.

On the other end of the spectrum, we can see some red, tiny points in Figure 5.6 especially at the city boundaries, where stations only have one direct connection to another station and therefore a degree of 1.

Station	Degree (Normalized)
Hauptbahnhof (S, U, Bus, Tram)	14 (0.0128)
Pasing	14 (0.0128)
Ostbahnhof	13 (0.0119)
Harras	13 (0.0119)
Heimeranplatz	13 (0.0119)
München Hbf	12 (0.011)
Moosach	12 (0.011)
Marienplatz	11 (0.0101)
Isartor	11 (0.0101)
Berg am Laim	11 (0.0101)

TABLE 5.1: Stations with the highest Degree Centrality

## 5.2.2 Betweenness Centrality

Another intuitive measure is the *betweenness centrality*. It quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Formally, betweenness is defined as [4]

$$C_{\text{Betweenness}}(v) = \sum_{i \neq n \neq j} \frac{\sigma_{ij}(v)}{\sigma_{ij}}, \quad (5.2)$$

$\sigma_{ij}$  as number of shortest paths from node  $i$  to node  $j$ ,

$\sigma_{ij}(v)$  as number of those paths which pass through  $v$ .

These shortest paths are found by using algorithms such as **Breadth-First Search** or **Dijkstra's algorithm** [7]. The edge weights are thus incorporated into the betweenness measure by applying these algorithms on the weighted instead of the unweighted graph.

For easier comparisons, it is once again convenient to normalize the centrality measure by multiplying the factor  $\frac{2}{(n-1)(n-2)}$  with the betweenness measures.

In Figure 5.7 the normalized betweenness centrality for all stations in the MVG network is displayed. Each circle represents a station's betweenness, and the circle's size is proportional to the station's betweenness. It becomes apparent quickly how only a couple of stations have a considerable betweenness measure. In fact, only 187 out of the 1095 stations have a measure above 0.01 and only 12 have one above 0.1. Hence, big hubs are given disproportionately much importance in this measure.

In Table 5.3 all stations with a betweenness measure in the top percentile are displayed, which are the 11 pink-colored circles in Figure 5.7. We can immediately see how these stations resemble the stations with the highest degrees. All of these stations have in common that they serve several different products, such as *U-Bahn*, *Bahn*, *S-Bahn*, trams, and busses at just one station. These stations are big hubs within the network. Consequently, they get a considerable share of the betweenness centrality measure. The two stations with the highest betweenness values, the *München Hbf* (Munich central station) and the *Ostbahnhof* (Munich East station) are the connections west and east of the Isar river, where lots of connections are arriving before or after crossing the river.



**Betweenness Centrality**

by Station

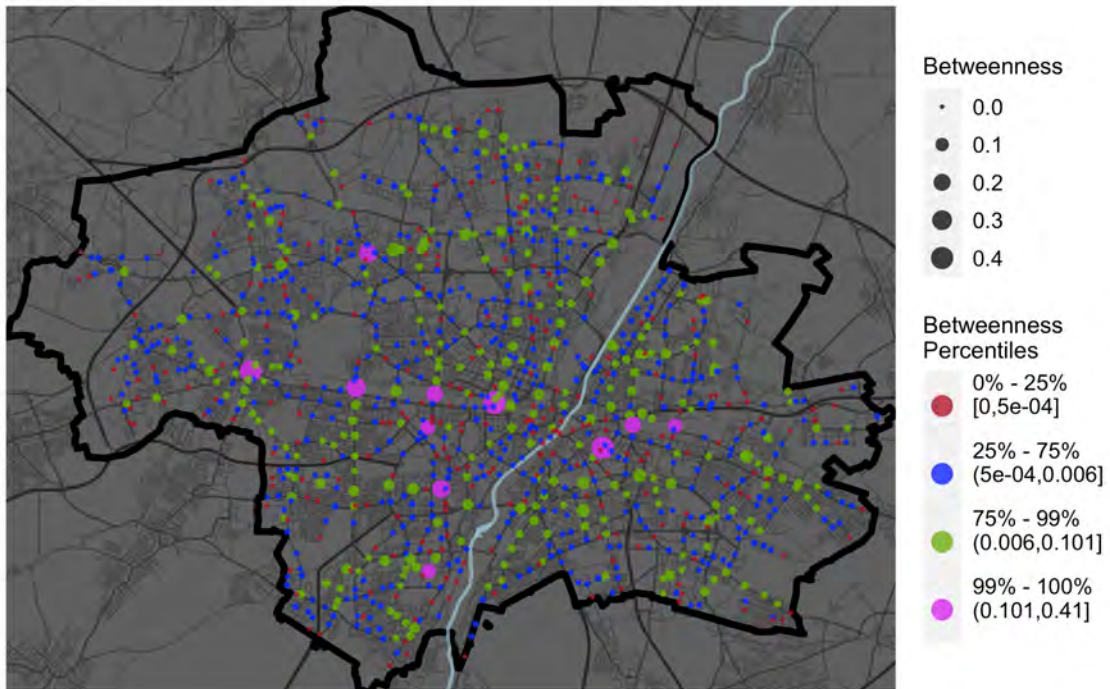


FIGURE 5.7: Betweenness Centrality

Many red-colored stations, the ones with betweenness below 0.1, are single-purpose bus stations, where no other products are used. Generally, the high betweenness stations seem clustered around the city center and especially along the *S-Bahn* line called **Stammstrecke** [1] going from *Pasing* in the west to *Ostbahnhof* in the east. Additionally, several other stations with relatively high betweenness are spread around the city, such as *Feldmoching* up north, *Siemenswerke* in the south, and *Trudering* in the east.

Station	Normalized Betweenness
München Hbf	0.410
Ostbahnhof	0.397
Pasing	0.370
Laim	0.276
Moosach	0.236
Harras	0.222
Donnersbergerbrücke	0.183
Leuchtenbergring	0.177
Berg am Laim	0.127
Heimeranplatz	0.118
Siemenswerke	0.115

TABLE 5.2: Stations with the highest Betweenness Centrality

Overall, the betweenness centrality measure seems to portray the most important stations in the network as accurately the junctions people will have to use to travel across the efficiently.

### 5.2.3 Closeness Centrality

Another centrality measure is **closeness**, indicating how close a node is to all other nodes in the network. Concretely, a node's closeness is defined as its average distance to all other nodes [20]:

$$C_{Closeness}(v) = \frac{1}{\sum_{u=1}^{n-1} d(v, u)}, \quad (5.3)$$

where  $d(v, u)$  is the distance between nodes  $v$  and  $u$ . The *Dijkstra Algorithm* is used to calculate the shortest distances with the edge weights as the distances between nodes. The closeness is multiplied with the factor  $(N - 1)$  to get the normalized closeness.

**Closeness Centrality**  
by Station

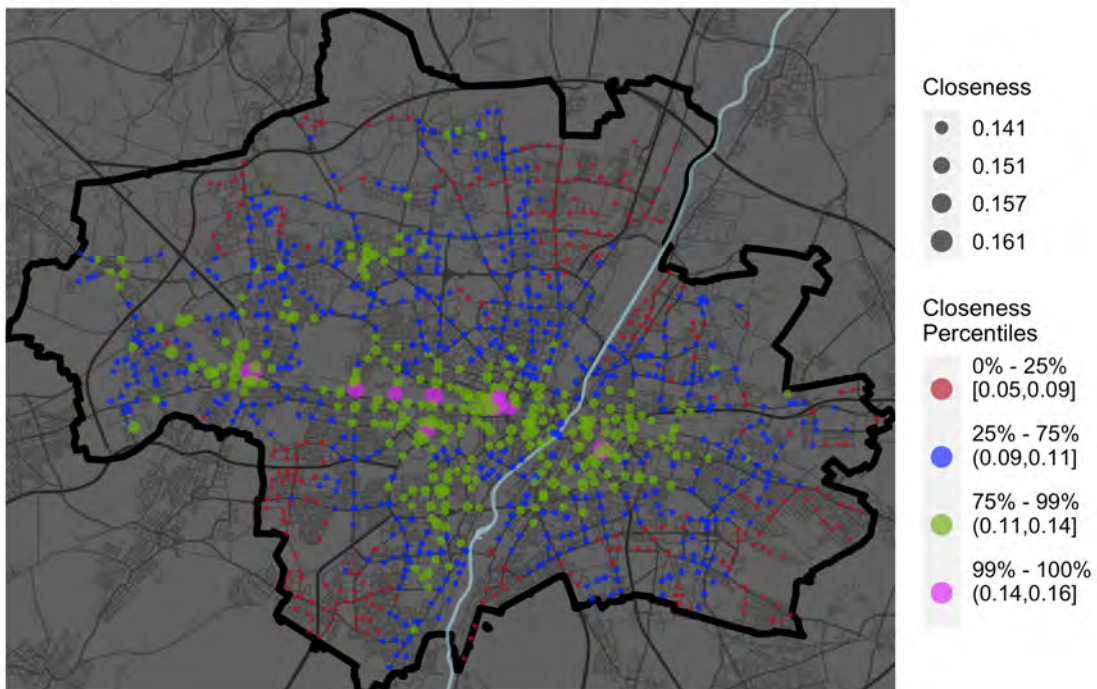


FIGURE 5.8: Closeness Centrality

Consequently, a closeness value of 1 would mean that all other nodes in the network are direct neighbors of a node. Figure 5.8 depicts the closeness measures for all stations and is color-coded like the betweenness centrality, where the first quartile is red, the second and third quartiles are orange, the fourth quartile up until the 99th percentile is green, and from

Station	Normalized Closeness
München Hbf	0.162
Pasing	0.160
Ostbahnhof	0.155
Donnersbergerbrücke	0.155
Laim	0.150
Heimeranplatz	0.148
Hauptbahnhof (S, U, Bus, Tram)	0.147
Karlsplatz (Stachus)	0.147
Hirschgarten	0.146
Elisenstraße	0.145
Hauptbahnhof Nord	0.144

TABLE 5.3: Stations with the highest Closeness Centrality

there, the stations are pink. Since the circle sizes are proportional to the stations' closeness, we can further differentiate their closeness.

As we can see by these circle sizes, the absolute difference with this centrality measure is smaller between the highest and lowest values than the range we have seen before in the betweenness measures. The closeness range goes from about 0.05 to 0.16. The stations with the highest values, in the fourth quartile, are mainly clustered around the city center, and other stations with high closeness measures mostly surround these points. This exemplifies how different the ideas of betweenness and closeness actually are. For instance, being a station close to the one with the highest centrality measure in the network is inevitably going to lead to a high centrality measure as well in regards to closeness, while this is not the case for betweenness. Consequently, centrality gives us more of an indicator for the connectedness of whole neighborhoods while betweenness highlights the most important stations in neighborhoods and their connectedness. An example of this difference can be seen when considering the central station. Three of four stations considered as variants of the central station (*München Hbf*, *Hauptbahnhof (S, U, Bus, Tram)*, *Hauptbahnhof Nord*) are within the top percentile of closeness and the fourth station, *Hauptbahnhof Süd*, lies within the top two percentiles. In contrast, only one of the four stations, *München Hbf*, is in the top betweenness percentile, *Hauptbahnhof (S, U, Bus, Tram)* and *Hauptbahnhof Nord* are in the upper fifteen percentiles while *Hauptbahnhof Süd* is just slightly above the median betweenness value.

## 5.2.4 Eigenvector Centrality

**Eigenvector centrality** computes node importance by considering the centrality of the node's neighbors. The eigenvector for a node  $v$  is the  $v$ -th element of the vector  $x$  defined by the eigenvector equation:

$$Ax = \lambda x,$$

with  $A$  as the graph's adjacency matrix and the eigenvalue  $\lambda$  [3]. Since  $A$  is a square matrix and only contains non-negative values, the matrix has a unique largest real eigenvalue  $\lambda$  and the corresponding eigenvector has strictly positive components per the *Perron-Frobenius Theorem* [6].

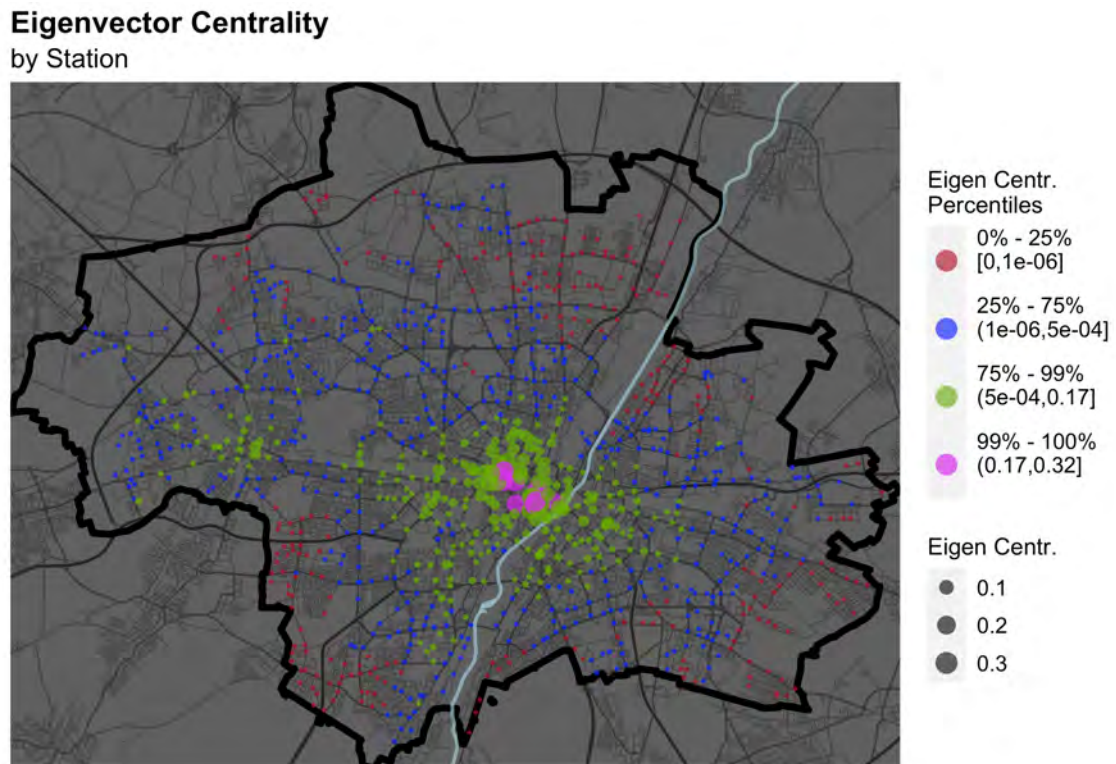


FIGURE 5.9: Eigenvector Centrality

While the top stations with regards to their centrality measure were spread around several locations in the city for betweenness and closeness values, the top percentile of stations is now located in a small radius around the central station. This highlights how stations need several direct neighbors with high centrality values to achieve high centrality themselves. Hence, most stations with eigenvector centrality measures in the fourth quartile are directly spread around the city center. The green cluster in the west around the *Pasing* station is connected to the central station via train and *S-Bahn* as well, such that most of the high centrality stations are within a short ride of the central station. While these stations look rather far away from the city center, from a graph perspective they are closer than many other points with a smaller spatial distance, highlighting once more that we are considering ride times in minutes between stations as edge weights instead of distances in meters.

The top percentile of stations, displayed in Table 5.4, shows stations that have not yet appeared as top centrality stations in the other measures, such as *Schrammshalle*, *Viktualienmarkt* and *Blumenstraße*, all stations within a couple of minutes to *Marienplatz*, one of the most famous and central places in Munich, which did not make an appearance in the top stations with the other centrality measures.



Station	Eigenvector Centrality
Hauptbahnhof (S, U, Bus, Tram)	0.318
München Hbf	0.25
Hauptbahnhof Nord	0.249
Marienplatz	0.238
Schrannenhalle	0.236
Karlsplatz (Stachus)	0.225
Elisenstraße	0.209
Sendlinger Tor	0.207
Viktualienmarkt	0.203
Blumenstraße	0.2
Isartor	0.18

TABLE 5.4: Stations with the highest Eigenvector Centrality

The fact that most stations are assigned a centrality measure of close to zero can also be seen in its strongly right-skewed histogram in Figure 5.11 and its boxplot in Figure 5.10, where the median centrality value is as low as  $0.00002$ , while the maximum eigenvector centrality value is bigger than  $0.3$ . With its interquartile range of just  $0.0005$  and its difference between the maximum and minimum value of  $0.317$ , 244 of the 1095 stations are considered outliers in the boxplot.

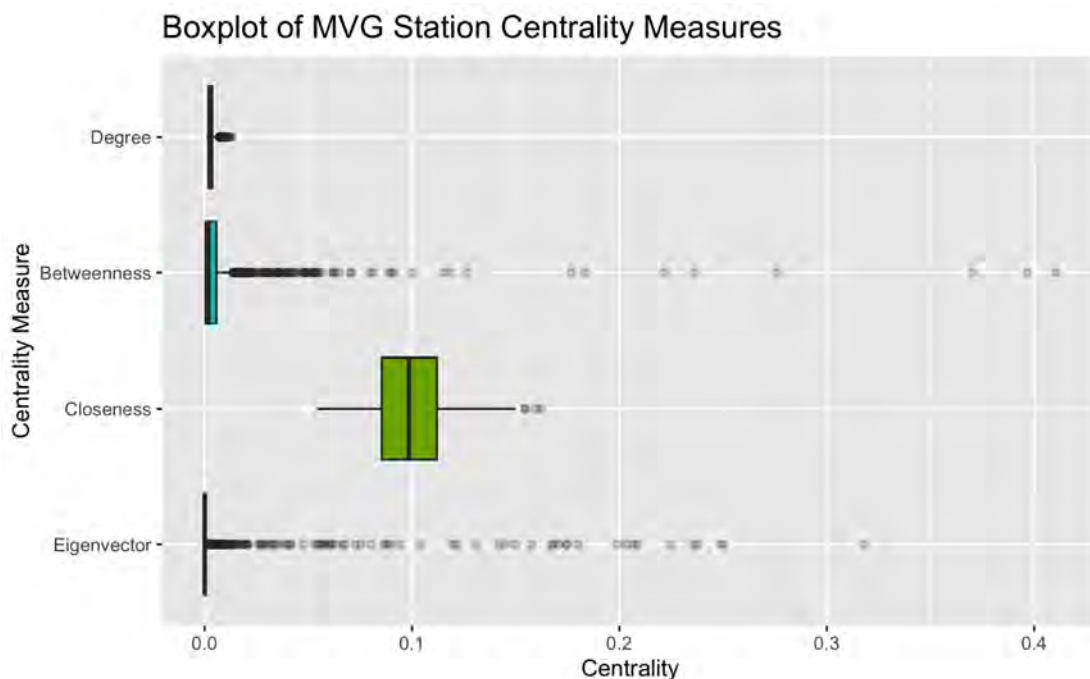


FIGURE 5.10: Boxplot of Centrality Measures

We have now seen how four different methods of assigning node importance do so in quite different ways and with differing results. From a real-life perspective, when deciding which stations one would like to live around based on the described methods, the degree of

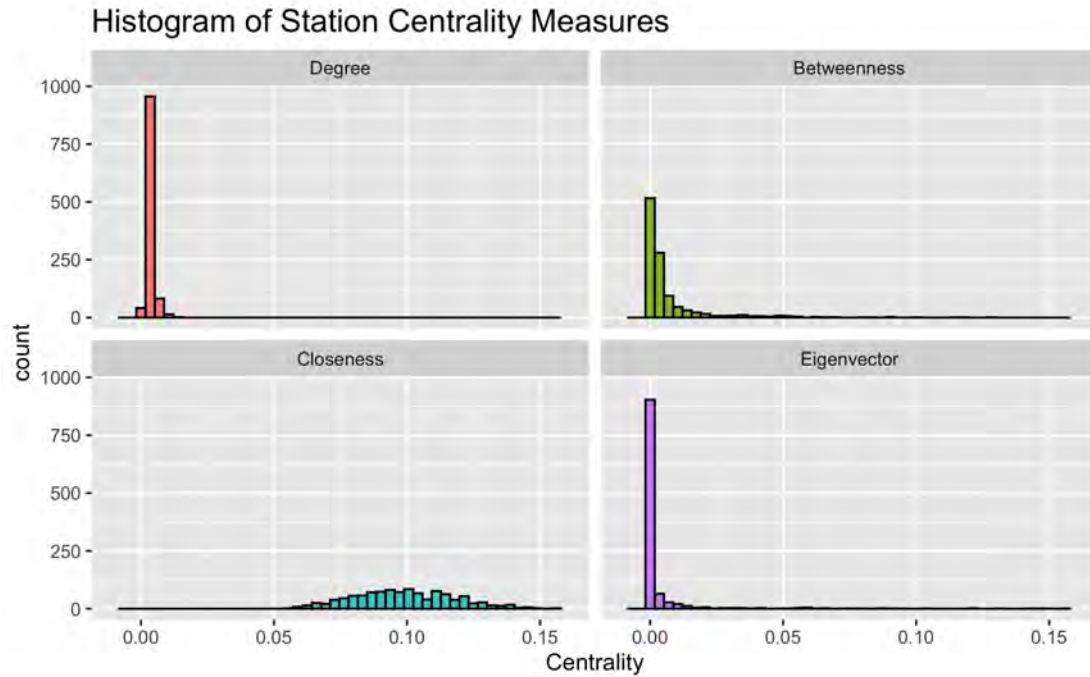


FIGURE 5.11: Histogram of Centrality Measures

a station by itself is not an immediate strong indicator for a well-connected station. While stations in the city center tend to have higher degree values, such a station would also be possible far away from the city center and other attractive city locations, only connecting several other stations at the city's outskirts.

On the other hand, high betweenness measures directly imply a well-connected station since many connections between other stations have their shortest ride time via that station. Hence, the high betweenness of a station gives it the status of a hub in the network and is therefore attractive when considering living locations. Since there are only a couple of stations with extremely high betweenness compared to others, it is also a great way to identify the major stations in a city.

This is not the case for eigenvector centrality, where the importance of the neighbors of a node matter. The *Blumenstraße* station for instance, is in the top percentile of node importance regarding eigenvector centrality by being close to two crucial stations, *Sendlinger Tor* and *Marienplatz*, while not being a major hub itself. Like closeness, which considers the distance to all other stations, eigenvector centrality is, therefore, more of an indicator for being considered central within a city, with less importance being laid on the single station itself but rather its surroundings.

### 5.3 Relationship between Reachability, Centrality Measures & Rental Prices

After looking into different centrality measures, we can now compare them to the reachability measure from chapter 3. While the reachability is not trying to describe the importance of stations in the network per se, it is doing so indirectly by assigning high values to well-connected stations and is thus comparable to the graph centrality measures.

Figure 5.12 depicts a pair plot of the centrality scores and reachability, the share of all stations, a station can cover with its connections within ten minutes. All measures generally range from 0 to 1, when considering the normalized centrality measures and when considering reachability as a score between 0 and 1. For improved readability and interpretability in the plot, however, the not-normalized degree centrality was used. We can then see the highest positive correlation of 0.761 between reachability and the degree, while there are also positive correlations between all other measures. In order to mitigate the issue of the strongly right-skewed betweenness and eigenvector values, we use the log values of these measures. For betweenness, 0-values were replaced by 10% of the second-smallest value.

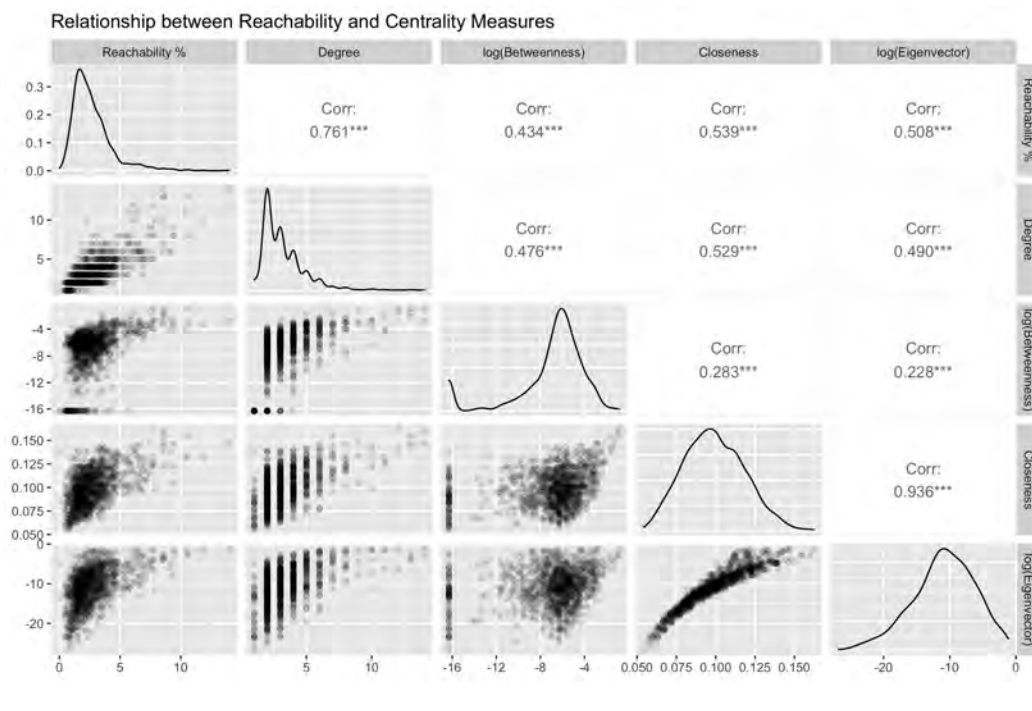


FIGURE 5.12: Relationship between Reachability and Centrality Measures  
Log values of Betweenness and Eig. Centrality

We notice in 5.12 how closeness and the log of the eigenvector centrality are now strongly correlated with a correlation coefficient of 0.936, while the correlation of the non-log transformed measures was significantly smaller at 0.373, which is depicted in Figure A.7

in Appendix A. This high correlation could be a consequence of the previously mentioned nature of these measures, where closeness and eigenvector centrality both value being close to many other nodes in the network highly. As such, both measures contain similar information about the network and are very strongly correlated. On the other hand, we can see how the correlation between the log-betweenness and all other measures is reduced now, apart from the log-eigenvector centrality. Generally, betweenness is the measure that seems to be the least related to the others, as its focus is rewarding hubs instead of rewarding being generally well connected.

Finally, we want to get a quick insight into how these graph measures relate to rental prices. Hence, we will aggregate the data by borough and consider the median values of rent, reachability, and centrality by borough. As rental dataset, we will use the *Mietspiegel* dataset restricted on the contracts newer than ten years, but similar results occur for the full data and the *ImmobilienScout24* data, as seen in the Appendix A in figures A.8 and A.9. Note, that for these aggregated values only the eigenvector centrality was log-transformed.

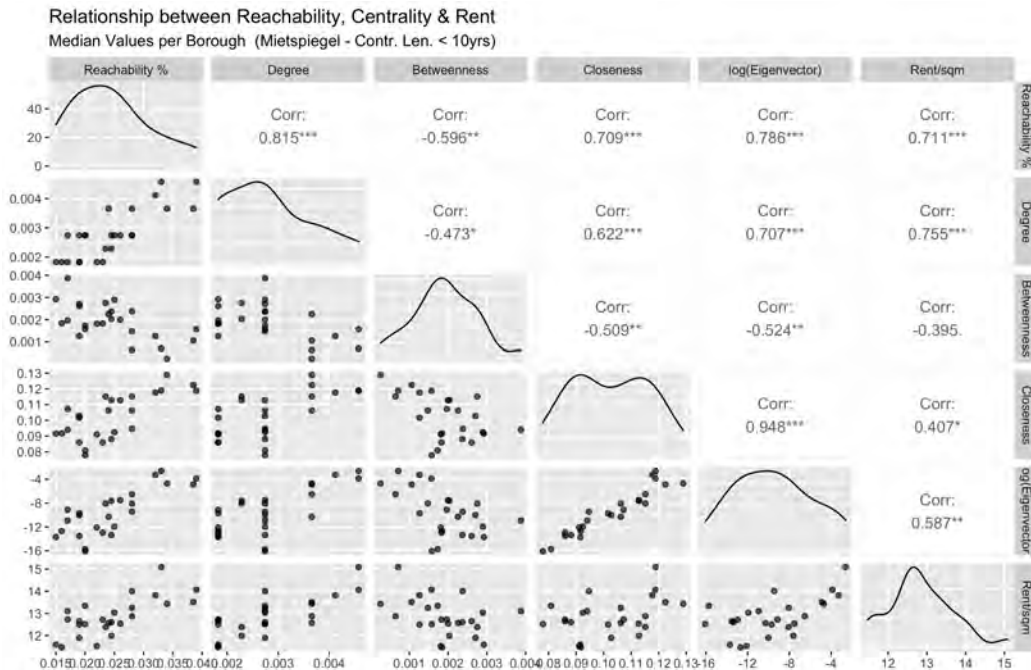


FIGURE 5.13: Reachability, Centrality Measures & Rental Prices

Figure 5.13 then shows the pair plot of the median values of reachability, the centrality measures, and the rental prices aggregated by borough. While we have seen strictly positive correlations before, we now notice negative correlations of betweenness and all other values.

However, all other measures are still positively correlated, with the strongest relationship still being the one between the log-transformed eigenvector centrality and closeness with a correlation of 0.948. Regarding rental prices, the strongest relationships we see here is to the degree centrality, with a value of 0.755, while the correlation with reachability is slightly



less at 0.711. Similar trends can be observed when aggregating the data by sub-borough, depicted in Figure A.10 in the Appendix A.

We have now seen how traditional graph centrality measures can describe the importance of public transportation stations within a city and how these measures are related to existing rental prices. Hence, a similar regression analysis between centrality measures and rental prices is possible. Additionally, the centrality scores help to quickly identify essential stations in cities, giving potential new residents a good overview of central hubs and well-connected neighborhoods.

## 6 Summary & Outlook

We have now seen how rental price information for the same city can vary drastically depending on the data gathering method. Nonetheless, we could use both datasets to model reachability by public transport in Munich on the apartment level and aggregated by borough and sub-borough. We have then seen Isar's influence on Munich's transit in the form of different transport method speeds, by using Isar related information to improve the connection duration regression model and by clustering the city into distinct parts. Finally, we explored how the transit network can be displayed as a graph and how to quantify station importance by using graph centrality measures.

Going forward, it will potentially be possible to use the presented reachability information and the graph centrality measures to improve existing rental price models within cities. Furthermore, the intuitive reachability could help new residents decide which locations in a city are suitable for them when looking for apartments that are not only close to work or school but generally well-connected. From the company's perspective, one might be interested in creating offices near stations that are big hubs, as indicated by the betweenness centrality, or just ones with generally high reachability.

# Bibliography

- [1] 2. *Stammstrecke München*. 2021. URL: <https://2.stammstrecke-muenchen.de>.
- [2] Matthias Althoff. “Reachability analysis and its application to the safety assessment of autonomous cars”. PhD thesis. Technische Universität München, 2010.
- [3] Phillip Bonacich. “Some unique properties of eigenvector centrality”. In: *Social networks* 29.4 (2007), pp. 555–564.
- [4] Ulrik Brandes. “A faster algorithm for betweenness centrality”. In: *Journal of mathematical sociology* 25.2 (2001), pp. 163–177.
- [5] Piotr Bródka et al. “A degree centrality in multi-layered social network”. In: *2011 International Conference on Computational Aspects of Social Networks (CASoN)*. IEEE. 2011, pp. 237–242.
- [6] Kung-Ching Chang, Kelly Pearson, and Tan Zhang. “Perron-Frobenius theorem for nonnegative tensors”. In: *Communications in Mathematical Sciences* 6.2 (2008), pp. 507–520.
- [7] Thomas H Cormen et al. *Introduction to algorithms*. MIT press, 2009.
- [8] Ludwig Fahrmeir et al. *Regression*. Springer, 2007.
- [9] Ronen Feldman, James Sanger, et al. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.
- [10] John A Hartigan and Manchek A Wong. “Algorithm AS 136: A k-means clustering algorithm”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 100–108.
- [11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- [12] Annika Hoyer. “Multivariate Verfahren”. Lecture Notes. 2020.
- [13] Lawrence Hubert and Phipps Arabie. “Comparing partitions”. In: *Journal of classification* 2.1 (1985), pp. 193–218.
- [14] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

- [15] Lehrstuhl für Statistik (Prof. Dr. Göran Kauermann) der Ludwig-Maximilians-Universität München Sozialreferat der Landeshauptstadt München Kantar TNS. *Mietspiegel für München 2019*. 2019.
- [16] muenchen.de. *Badebus Langwieder See*. URL: [www.muenchen.de/freizeit/baden/badebus.html](http://www.muenchen.de/freizeit/baden/badebus.html).
- [17] muenchen.de. *Historische Grunddaten zu den Münchner Stadtbezirken, Stadtteilen und Vororten*. URL: [www.muenchen.de/rathaus/Stadtverwaltung/Direktorium/Stadtarchiv/Publikationen/Von-Allach-bis-Zamilapark.html](http://www.muenchen.de/rathaus/Stadtverwaltung/Direktorium/Stadtarchiv/Publikationen/Von-Allach-bis-Zamilapark.html).
- [18] MVG. *About us, by us*. URL: [www.mvg.de/en/about.html](http://www.mvg.de/en/about.html).
- [19] UJ Nieminen. "On the centrality in a directed graph". In: *Social Science Research* 2.4 (1973), pp. 371–378.
- [20] Kazuya Okamoto, Wei Chen, and Xiang-Yang Li. "Ranking of closeness centrality for large-scale social networks". In: *International workshop on frontiers in algorithmics*. Springer. 2008, pp. 186–195.
- [21] openrouteservice.org by HeiGIT. *Data retrieved from openrouteservice.org API*. openrouteservice.org. 2021.
- [22] OpenStreetMap contributors. *Map Export retrieved from www.openstreetmap.org*. www.openstreetmap.org. 2021.
- [23] William M Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.
- [24] Jorge M Santos and Mark Embrechts. "On the use of the adjusted rand index as a metric for evaluating supervised classification". In: *International conference on artificial neural networks*. Springer. 2009, pp. 175–184.
- [25] Thomas Seidl. "Knowledge Discovery in Databases". Lecture Notes. 2019.
- [26] Clifford H Wagner. "Simpson's paradox in real life". In: *The American Statistician* 36.1 (1982), pp. 46–48.
- [27] Douglas Brent West et al. *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River, 2001.

# A Appendix

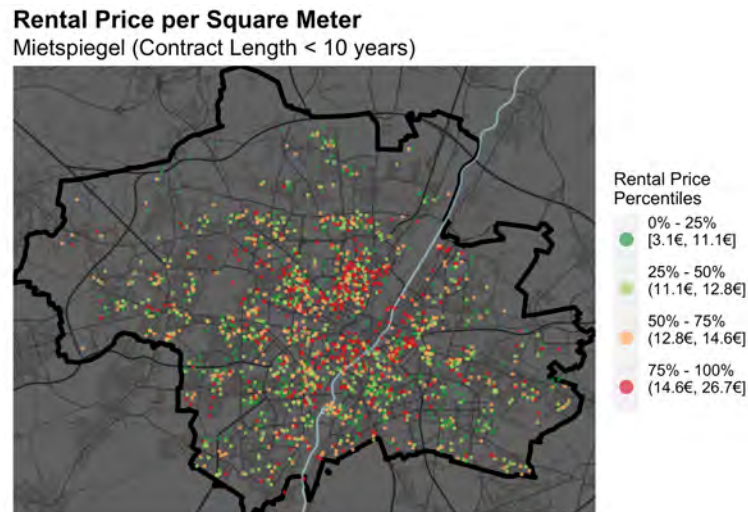


FIGURE A.1: Map of Mietspiegel Rental Prices (New Contracts)

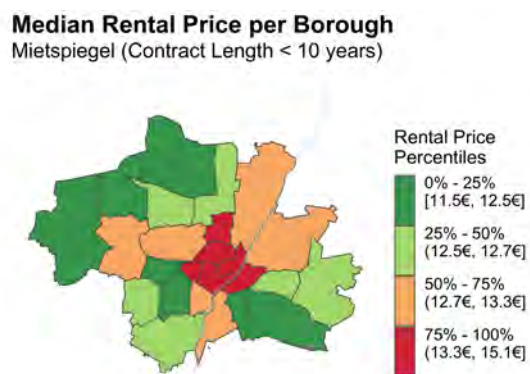


FIGURE A.2:  
Median Rental Price per  
Borough:  
Mietspiegel (New Con-  
tracts)

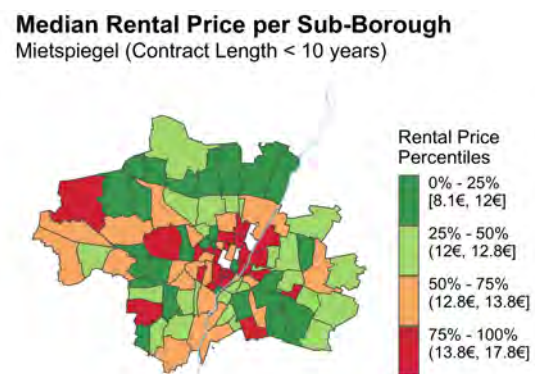
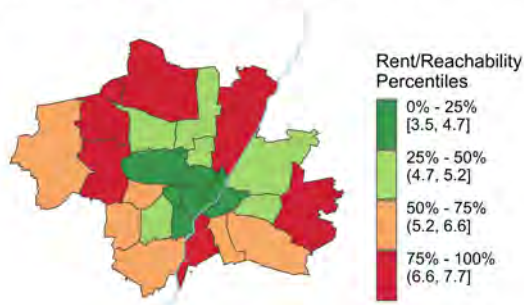


FIGURE A.3:  
Median Rental Price per  
Sub-Borough:  
Mietspiegel (New Con-  
tracts)

**Rent to Reachability Ratio per Borough**  
Mietspiegel Data (Contract Length < 10 years)



**Rent to Reachability Ratio per Sub-Borough**  
Mietspiegel Data (Contract Length < 10 years)

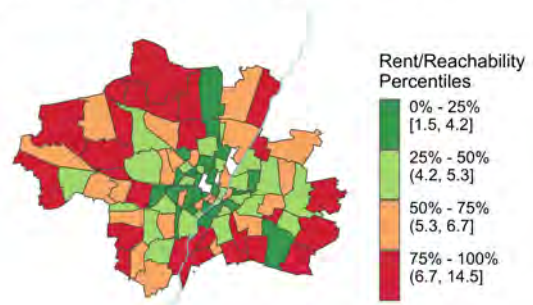


FIGURE A.4:  
Rent vs. Reachability  
Ratio  
By Borough:  
Mietspiegel (New Con-  
tracts)

FIGURE A.5:  
Rent vs. Reachability  
Ratio  
By Sub-Borough:  
Mietspiegel (New Con-  
tracts)

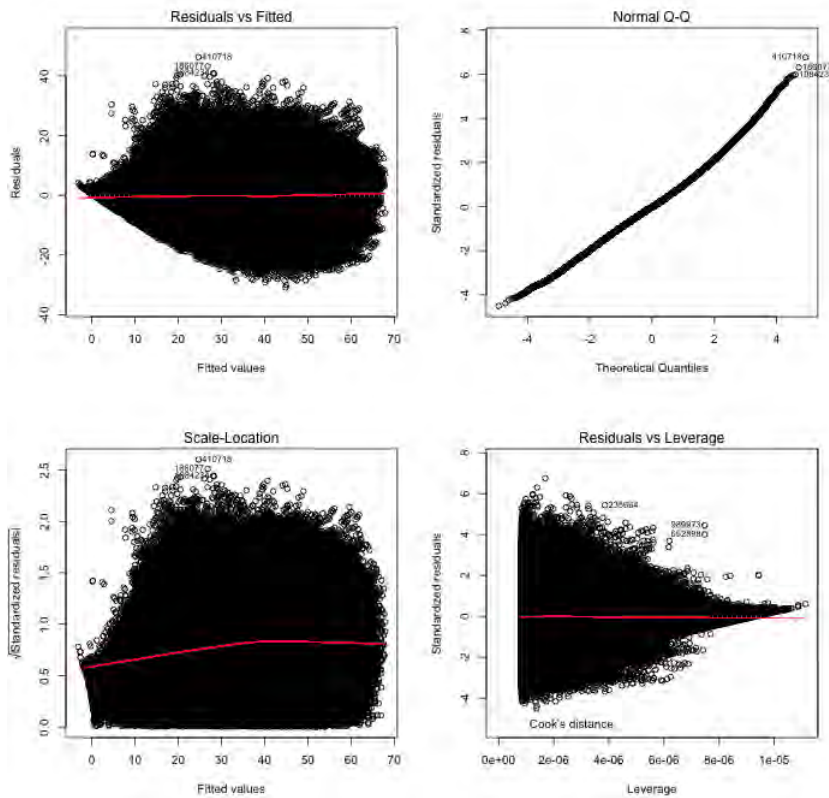


FIGURE A.6: Residual Diagnostics for Distance vs. Transit Ride Time  
Base Model

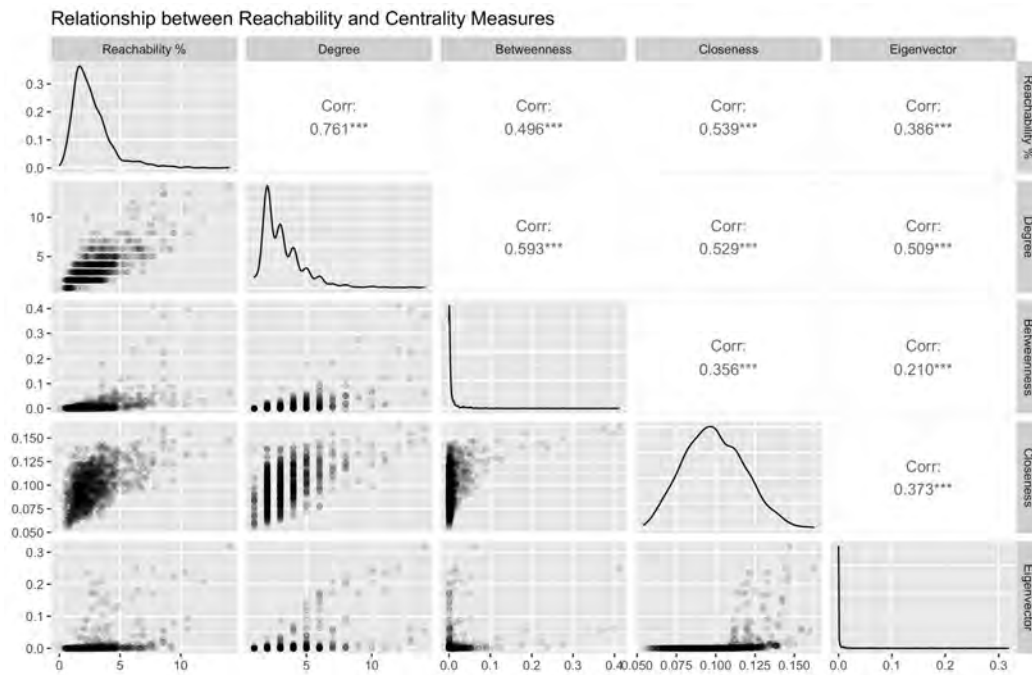


FIGURE A.7: Relationship between Reachability and Centrality Measures

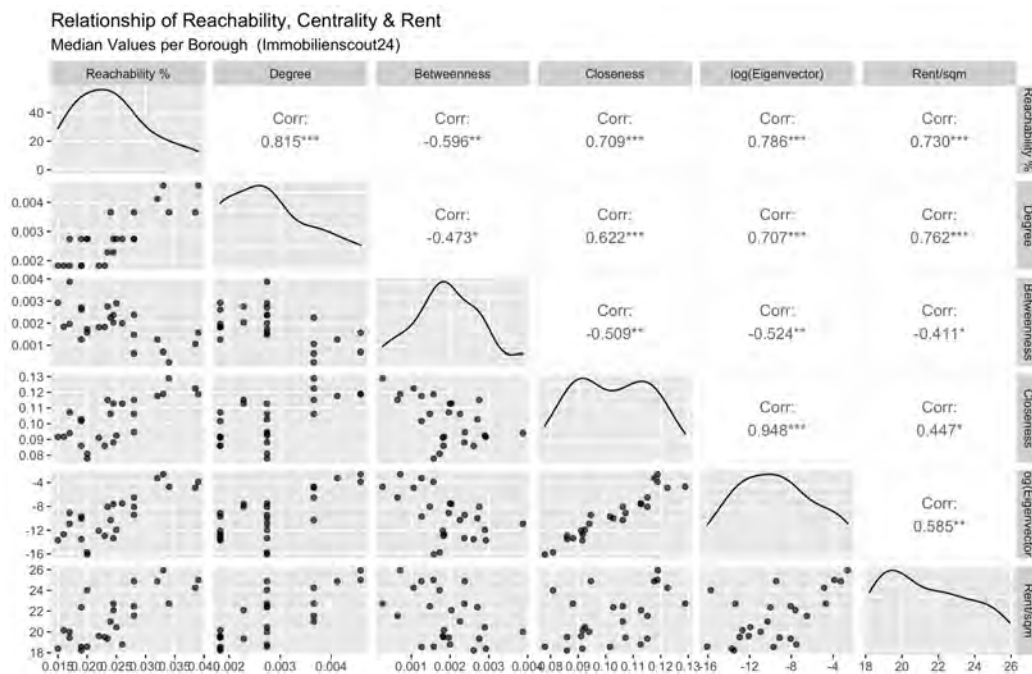


FIGURE A.8: Reachability, Centrality Measures & Rental Prices Immobilienscout24 Data

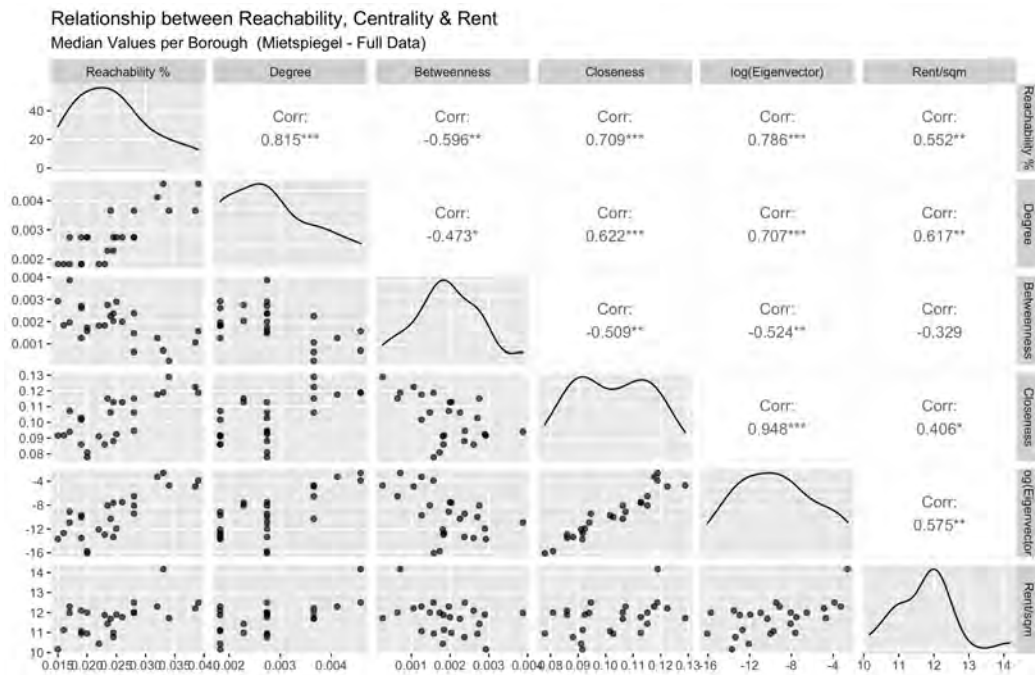


FIGURE A.9: Reachability, Centrality Measures & Rental Prices  
Mietspiegel Data

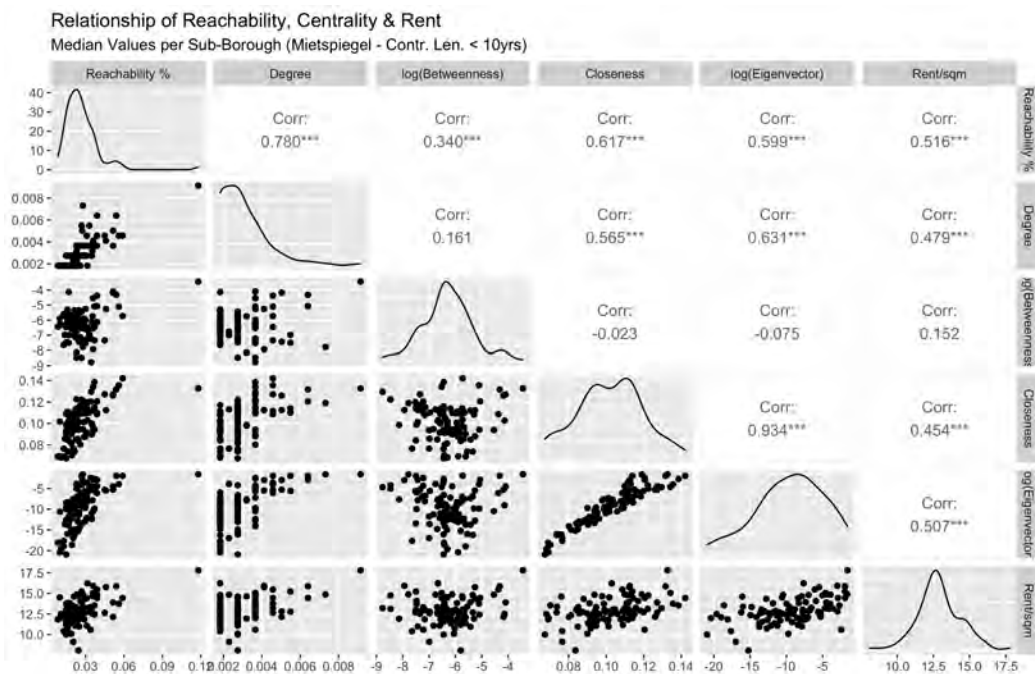


FIGURE A.10: Reachability, Centrality Measures & Rental Prices  
Sub-Borough