# Geospatial Analysis of Property Values Using Spatial & Social Ratings Data

Master Thesis Submitted to

**Prof. Dr. Wolfgang K. Härdle**
**Prof. Dr. Stefan Lessmann**

School of Business and Economics
Ladislaus von Bortkiewicz Chair of Statistics

**Humboldt-Universität zu Berlin**

by

**Alex Walker Truesdale**

598854

in partial fulfillment of the requirements for the degree of
**Master of Science Wirtschaftsinformatik**

November 11, 2020

Berlin, Germany

# Acknowledgement

I would like to express a heartfelt gratitude to Prof. Dr. Wolfgang K. Härdle for providing opportunity and guidance in both the writing and preparation of this master's thesis as well as in the numerous courses I have taken at the statistics chair under his supervision. The close communities of curious, intellectual minds in these environments have shaped my academic, professional, and personal achievements profoundly.

Additionally, I would like to extend a thank you to Dr. Prof. Stefan Lessmann, who is responsible for the exceptional curriculum and culture at the Wirtschaftswissenschaftliche Fakultät. The numerous WiWi courses taken under his purview have provided me with invaluable technical and business skills, which have been of immediate use in industry.

I am sincerely thankful for the instruction and, on occasion, the academic and/or professional counseling from Dr. Alisa Kim and Dr. Johannes Haupt. Along with Dr. Prof. Lessmann, they provided the sound structural foundation for the Wirtschaftsinformatik degree programme, which has propelled myself and others forward with state-of-the-art training in both the practical and academic realms.

**Alex Truesdale**

# Abstract

Housing in urban centres is a sensitive issue both politically and in a market context. This is particularly true for the housing market in Berlin, Germany, where widespread speculation and international investment have resulted in sharp increases in property and rental costs in a short period of time. For all stakeholders in this equation (politicians, residents, property investors/developers), it is important to understand what drives price and demand in this unique environment.

To model these dynamics, this thesis employs a geographically weighted regression (GWR) and extends the hedonic pricing method (HPM) to include features involving neighbourhood popularity, access to nature (parks, rivers), access to public transportation, and proximity to shopping and schools, among other kinds of locations. Google place data & review counts represent general popularity of locations. Public data from the Berlin Senate and the VBB provide geospatial information on natural resources and access to public transportation, respectively. Kernel density estimation (KDE) is employed to analyse spatial patterns and distribution of places of interest (POIs), with hotspot analysis performed using the Getis-Ord statistic. Property data comes from a dataset of bookings made on a local platform for furnished apartment rentals. GWR results show a high correlation between POI clustering and price as well as significant improvements in model performance. The GWR notably succeeds in capturing spatial heterogeneity in feature effects that global OLS regression overlooks. This study provides a unique marriage of public and private data sources to arrive at a novel analysis of the Berlin property market.

**Keywords**: Housing Price Analysis, Spatial Modelling, Neighbourhood Effects, Social Network Data, Berlin Housing Market

# Contents

# List of Figures

# List of Tables

# 1 Introduction

The real estate market represents an economic cornerstone for both developing and developed nations alike. It often serves as a proxy for important indicators such as rate of business activity, economic growth, development of consumer purchasing power, amongst others. According to OECD et al. (2013), there are numerous arenas in society in which individuals or organisations strive to develop new and advanced mechanisms for asset appraisal and decision making in this sensitive market. For example, residential property price indices (RPPIs) are commonly used to inform the formulation and implementation of economic policy. Given their macroeconomic impact and degree of incorporation into economic structures, it is unsurprising that the estimation of real estate prices is a long-standing and vibrant area of research interest. Zooming in to the city of Berlin, Germany, this interest is intensified, particularly so in recent years. Since the fall of the Berlin Wall, private and public development efforts in the city have increased substantially along with the city's popularity as a destination for tourists and immigrants alike. Precipitous increases in rental prices and property value in this time have raised alarm in many areas from social justice and cohesion interests to business and political concerns. Regardless of where one falls on the politics of such development, however, the fact remains that Berlin as a real estate market has many interesting research facets to explore.

The aim of this thesis is to apply tried and tested methods in real estate pricing analysis to the Berlin market with the use of a unique mix of otherwise disparate data sources. While the statistical methodology of this study is a classic approach, the careful selection of novel data sources seeks to provide improved predictive power that remains, to date, relatively untapped. This allows, in quantitative terms, a better analysis and understanding of the dynamics that drive the Berlin property and rental markets. Additionally, this work provides a baseline for new methodologies which seek to leverage the ever-growing availability of high-fidelity and/or high-volume data from public or private providers. The remainder of this thesis is broken down into the following structure:

Section 2 explores the existing body of work and foundations for this thesis' academic and practical approaches, respectively. Section 3 introduces the study area and data sources which supply the regression analysis (properties dataset, ratings & reviews data, and municipal data). Section 4 provides detailed descriptions of the statistical methods employed: Kernel Density Estimate (for identifying high concentrations of POIs), The Getis-Ord statistic (used to detect hot spots using ratings & reviews data), and the framework of GWR modelling, which is used to perform the core analysis. Section 5 presents a discussion of the hot- and cold-spots for different types of POIs, GWR coefficients, and how the two relate to one another. The thesis is concluded in Section 6 with a discussion of the study's achievements and a summary of the value it offers in furthering the research in the areas of real estate pricing & dynamics and urban studies in general. Data analysis tools used in this article are Python3 and numerous Python modules for GIS mapping and data analysis, along with statistical methods libraries.

# 2 Literature Review

Residential property is a durable commodity with a versatile value profile. Not only is this profile made up of physical features of an apartment like size, age, and composition of a property, but locational, social, and environmental features play important roles in determining true property value as well. These additional components of an object's value are made up numerous sub-features, which are often intricately connected and combine to form latent characteristics. Examples of this are a given area's "vibrancy", as introduced by Barreca et al. (2020), or its access to nature or "environmental goods" (Ridker and Henning, 1967; Irwin, 2002; Wen et al., 2015).

In order to account for these relationships in this thesis, traditional statistical methods are used in combination with modern data sources only recently available to researchers. A popular method employed in the study of housing prices is the hedonic pricing method (HPM), explained by Adair et al. (1996) & Zilisteanu et al. (2019), which is further extended by the geographically weighted regression (GWR) method, theorised by Brunsdon et al. (1996) and applied in the works of Hiebert and Allen (2019) & Cellmer et al. (2020). These are both extensions of ordinary least squares (OLS) regression in which non-structural features (specifics of a property's surroundings rather than the property itself) and the spatial variability of these features are taken into account. Just as in the case of OLS, HPM & GWR models seek to explicitly determine drivers of market prices in the way of feature coefficients. A further extension of the GWR method is the mixed geographically weighted regression (MGWR) proposed by Fotheringham et al. (2017) and applied by Icon et al. (2019); Zhang et al. (2019). Here, the scale at which particular features variate is also taken into account. Employing these methods, prior research, for example that of Zhang (2019); Long (2020) & Zhang et al. (2020), has successfully identified that proximity of property to specific amenities bears a clear relationship to property value. This can include but is not limited to centres of commerce, natural spaces (parks, leafy areas or bodies of water), as studied by Daams et al. (2019) & Kim et al. (2019), or other important locations like schools and hospitals, effects of which are shown in the works of Kain and Quigley (2012) & Rivas et al. (2019). Proximity to these non-structural or 'environmental' features can be considered as a means by which to capture added value of a property's neighbourhood, which is often how individuals consider real estate value in day-to-day life. Typically, access to such amenities is studied by measuring distance from a given property to certain points of interest (POIs) (Tamara Sliskovic, 2019). In similar consideration, distance from a property to public transit connections and hubs is also a well-studied factor, as explored by So et al. (1997) & Zhang et al. (2019). This thesis takes into account green spaces, bodies of moving or standing water (rivers, lakes, etc.), commercial and business locations, and public transport access as primary POI types in testing whether these amenities fulfill specific functions in urban geography and to what degree their presence influences housing value.

A key detail of this analysis which provides a bridge between static structural / environmental features and the dynamicity of real-world human activity is the use of publicly-generated ratings & reviews data to study housing prices from a social media & big data perspective. While traditional methods of assessing environmental assets of a property are useful (distance to POIs), they ultimately fall short in achieving their task of properly representing the milieu that shapes an object's contextual value. This failure comes in the way of accounting only for distance-based access to such localities as discussed in the paragraph prior. What lacks is the socio-psychological component of human preference, that is what drives an individual to visit specific POIs. Simply put, is a given location or area popular, a dynamic explored by Chen et al. (2019), and how can this be represented with data? According to Chen et al. (2017) & Ghania et al. (2019), since the onset of Web 2.0 technologies, namely social media platforms and services, the data generated by Internet users has increased dramatically in both volume and diversity. This has direct implications for the study of urban commerce, preference dynamics, and geospatial analyses in general, say Marti et al. (2019) & Owuor and Hochmair (2020). This takes shape foremost in significant increases data points which are invaluable for understanding the embedded, underlying drivers of market behaviour in real estate. Social media has already been successfully utilised to study this area by way of, for example, sentiment analysis of Twitter data, explored by Hannum et al. (2019), analysis of geotagged photos and other media by Qi et al. (2019), or the consideration of housing platforms themselves, in which the market is analysed based on the nature in which properties and property information are shared (Hua et al., 2019). This basis of prior research highlights the value of these data sources and is the basis upon which this thesis' approach is founded.

The specific method through which social media data is put to use is inspired by a research article by Wu et al. (2016), which studies housing prices in Shenzhen, China by way of hot spot analysis of check-in data at POIs like bars, restaurants, parks, etc. The main idea here is that, while distance proximity to such amenities is valuable in modelling price, being able to classify them as particularly 'vibrant' or not rounds out the analysis further still. To this end, environmental features can be supplemented with hot- cold-spot analysis powered by big-data / social-media data sources. Hot spots in the context of this thesis refer to type-specific, spatial clusterings of POIs. These concentrations of POIs are important to take into account, as this clustering behaviour often follows particular dynamics by which popular and high-quality establishments (restaurants, for example) operate in near proximity to one another and form agglomerations, a phenomenon outlined by Mossay et al. (2020). Commercial, recreational, or green spaces which come together to form hot spots can be considered more capable of drawing higher numbers of guests and, by extension, also increase the so-called vibrancy of the area. This preponderance of activity, in turn, translates to a higher value reflected in property prices, as shown in the study by Wu et al. (2016). The inverse is also possible in that over-concentrations of said POIs can present negative effects on value – take, for example, the case of noise or air pollution in urban centres.

This research contributes to the existing body of research by providing an analysis of the Berlin real estate market with a focus on employment of unique data sources. Housing value is analysed with traditional methods and supplemented by data points from novel, big-data sources (ratings & review data) and highly granular data from public institutions (data on green space, bodies of water, and building age). This marriage of organically generated big data and curated municipal data helps to account for both static and more dynamic variables in the real estate value equation. The quality of the municipal data is particularly encouraging in that accurate and complete accounts of public, natural resources (parks, lakes, ect.) and information on building ages are particularly difficult to come by. One example of GWR housing price analysis by Tomal (2020) in Kraków, Poland highlights the importance of building age, which is especially noteworthy in European cities where pre-war housing stock plays a unique role in the urban housing market.

Furthermore, this study benefits from the use of well-governed property data that represents actual marketplace transactions to supply a robust target variable for the various analyses. In this case, the transactions are for furnished rental apartments in Berlin. While this is a particular sub-market of the real estate market as a whole, effects from spatial features are expected to generalise to larger markets like property sales or unfurnished rentals (the other two primary sub-markets in the city). The degree to which dynamics can translate between sub-markets is addressed in part by Clark and Lomax (2020), who explore the price-to-rent ratio (the ratio of home prices to annualised rent) for sub-markets in England. This article notes the importance of this ratio in being able to compare rental and purchase market dynamics. Talk of housing bubbles in Germany (Cholodilin and Michelsen, 2019) and in Berlin, specifically, might point to a divergence of purchase and rental dynamics in the real estate market. The recent introduction of the so-called Mietendeckel legislation in Berlin, explained by Dolls et al. (2020), brings this relationship further into question. For the scope of this analysis, however, these finer considerations will remain points for further research in future work.

# 3 Data

The data itself is a core component of this thesis and its contribution to the existing body of research. This section will present the structure and sources for each of the following datasets, respectively: property objects, social reviews data, and municipal data gathered from the Berlin Senate and VBB. First, however, a brief overview of the recent history of Berlin in how it informs the current state of the housing market provides valuable background for later unpacking the data exploration and analysis results.

## 3.1 Case Study: Berlin

Berlin is the most populous city in the European Union and the capital & largest city of Germany, both by area and population. Germany, in turn, is the 4th largest world economy in GDP. The city, located in central-northern Europe, covers a landmass of 891,1 km$^2$ with a city and metro population of 3.769.495 and 6.144.600 inhabitants, respectively. Berlin is composed of 12 districts or Bezirke, which are listed as follows (ordered by population density): Friedrichshain-Kreuzberg (1), Mitte (2), Neukölln (3), Tempelhof-Schöneberg (4), Lichtenberg (5), Charlottenburg-Wilmersdorf (6), Marzahn-Hellersdorf (7), Pankow (8), Steglitz-Zehlendorf (9), Reinickendorf (10), Spandau (11), and Treptow-Köpenick (12) (Visualised in Figure 3.1).

The city has a storied history, undergoing particular turbulence in the 20$^{\text{th}}$ century as the central point of the German Empire in both WW1 and WW2, as well as the crucible in which significant episodes played out between the United States and the Soviet Union during the Cold War. From 1961 to 1989, Berlin was divided in two, separated by the Berlin Wall, upon either side of which dramatically different economic and political structures shaped their respective portions of the city. Prior to this time, the boroughs of Berlin had indeed already developed many of their defining and distinguishing characteristics, but the postwar reconstruction of the city and the split due to the Berlin Wall left long-term residual effects upon the communities & (sub-)cultures within the city, as well as on the economic structures and architectural developments that rose out of the period. During postwar reconstruction, guest workers primarily from struggling economies in Southern Europe and Turkey came to Berlin to help rebuild the city and earn a more stable living than was available in their home countries. Many of these guest workers started families in Berlin and stayed in the city. These communities formed by sizeable immigrant populations became strongly rooted in their respective regions of the city and remain today as integral components of the Berlin's overall cultural makeup. Another notable occurrence in the post-war time is the development of not a single nucleus around which outer districts of the city orbit, but rather a multitude of urban nuclei, as noted by Broadbent (1990). Berlin famously has no true center and is instead often appears to be an agglomeration of sub-cities, each with a unique character and identity of its own.

Berlin's current role is again re-imagined, reuniting in 1990 as a single city and, in the past 30 years, rapidly developing once again into a global metropolis with a diverse, burgeoning cultural scene as well as a highly-developed economy, operating primarily in the high-tech, media, and service industries. The complex identity of the city continues to draw many immigrants and transplants from within Germany and from around the world, seeking either to take part in the growing technology sector or in a corner of the largely-alternative creative and cultural scene, among other motivators. This regular influx of new residents and continually growing industry has lead to high volumes of property investment in the recent decades, producing a very active market in which purchase and rental prices are rising at some of the fastest speeds in the world of late, as reported by DW (2017). For this reason, the housing market in Berlin continues to receive a high degree of scrutiny from diverse political, social, and economic interests. The city thereby offers an fascinating case study by which to explore the various dynamics at play in shaping property demand and market dynamics. Specifically, the decentralised and sporadic nature of Berlin's recent development are of particular interest in researching the geo-spatial non-stationarity of certain market drivers.

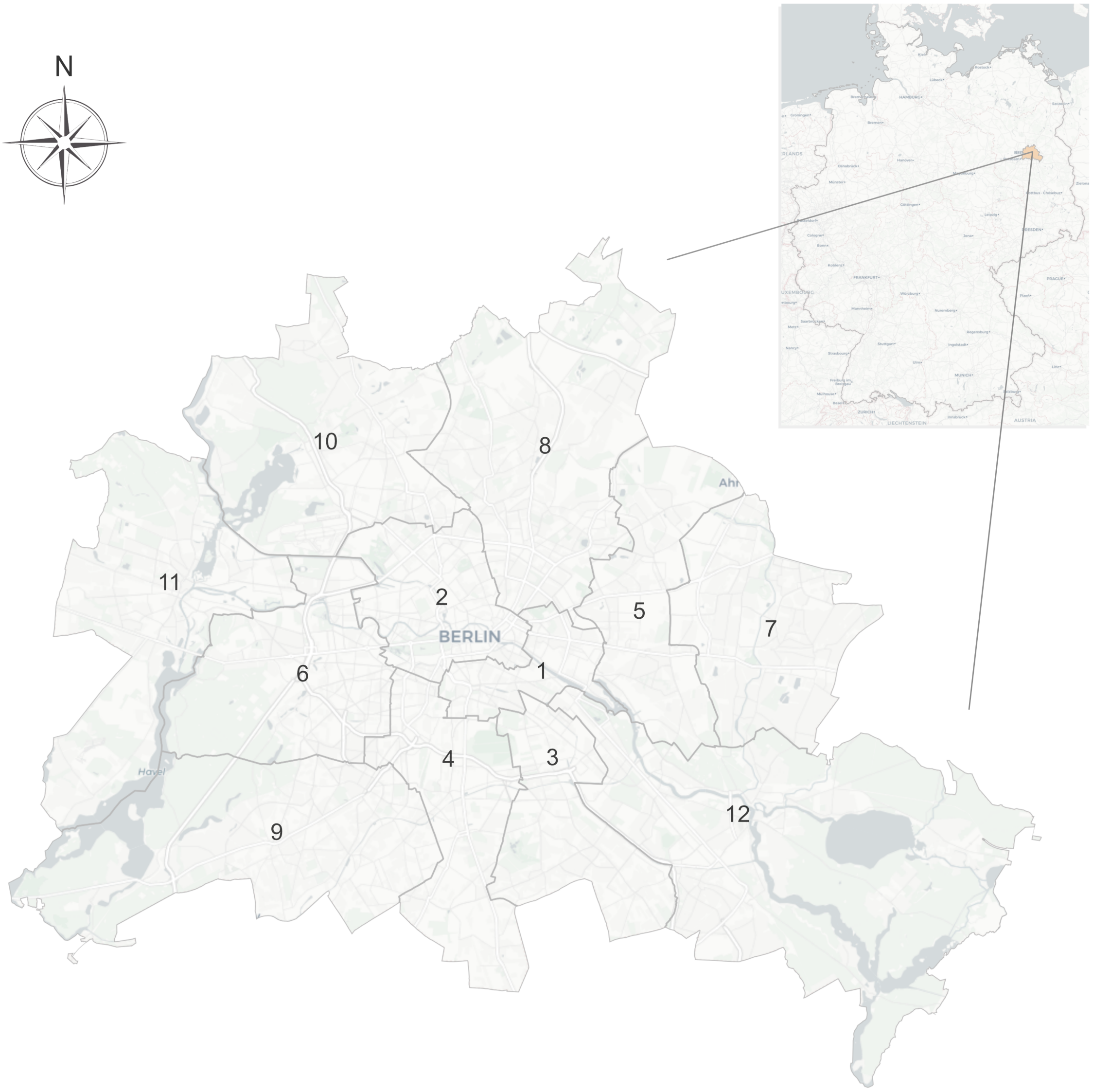

Figure 3.1: Berlin Districts (Bezirke) Bezirke

The study of this thesis will take into account the entirety of Berlin; however, the economic activity as well as the strength of the study's data is better represented closer to the geographical center of the city. The de facto delineator between the inner and outer districts of Berlin is the Berliner Ringbahn (shown in Figure 3.2), a continuous railway loop, which is part of the Berlin S-Bahn or Stadt-Bahn (City Rail) network and belts the innermost localities of the city. Observations within the Ringbahn will receive special attention, as the more desirable residential localities generally fall very near to or within this city boundary. The districts that make up this kernel of Berlin are the innermost sub-districts of Friedrichshain-Kreuzberg (1), Mitte (2), the northernmost sections of

Neukölln (3) and Tempelhof-Schöneberg (4), respectively, Charlottenburg-Wilmersdorf (6), and Prenzlauer Berg in Pankow (8).

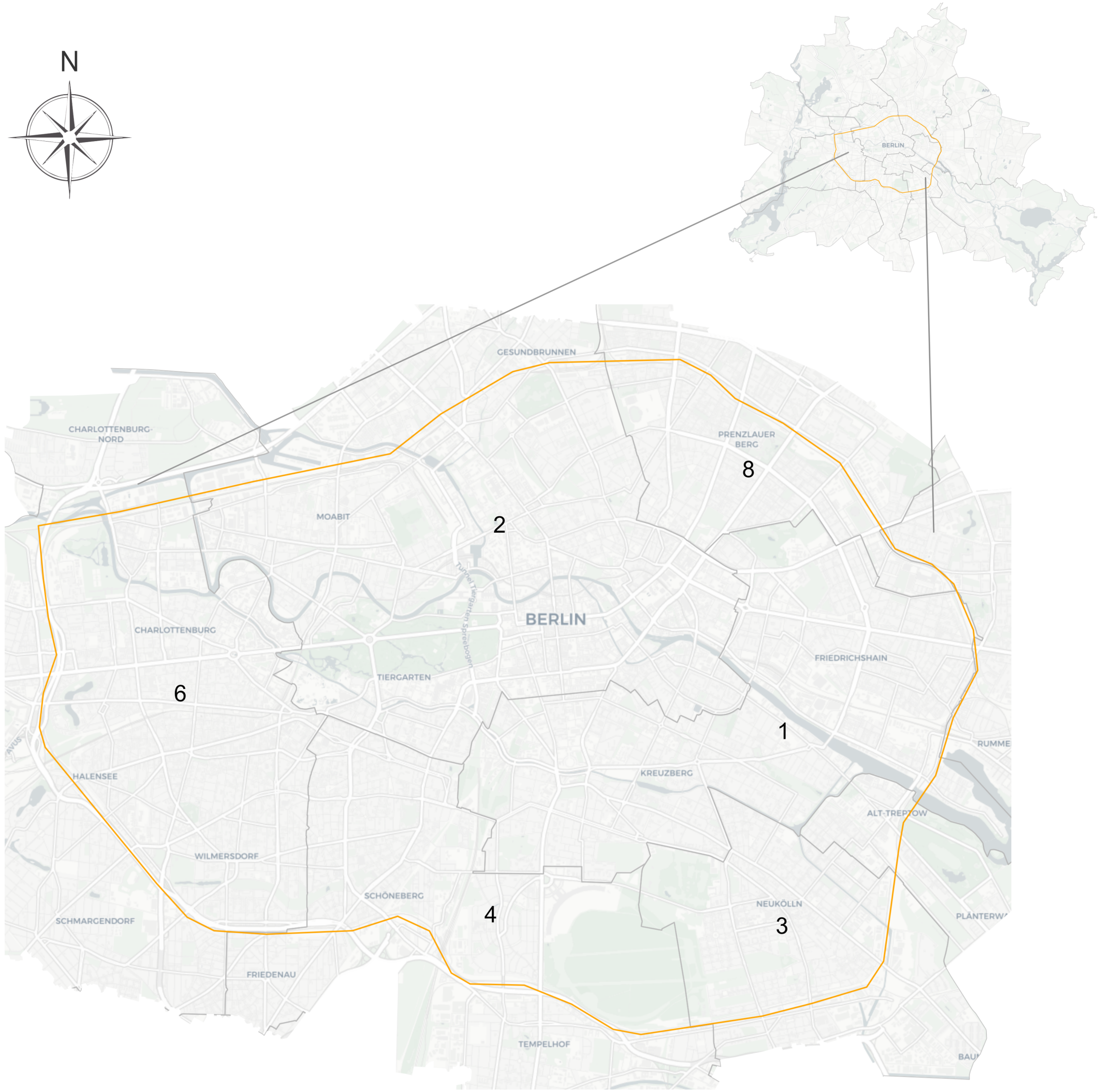

Figure 3.2: Berlin (within Ringbahn)

## 3.2 Properties Data

Structural data is the central component of any housing price analysis. While geospatial features and novel data sources add valuable nuance, the property-specific data points remain critical for establishing the foundation of the study.

This thesis draws the core property data from a local Berlin dataset supplied by Wunderflats GmbH, the largest marketplace provider for furnished and serviced mid- to long-term rentals in Germany. The dataset is comprised of 4.021 events representing a sample of confirmed apartment bookings between November, 2018 and February of 2019. Each event contains the structural details of the property, accounting for the basis of this study's independent variables: object floor, apartment area, number of rooms, how many people the unit accommodates, and whether or not the apartment has an elevator. Events also contain the price at which the apartment was booked. This is the dependent variable and feature of interest in this study. As is common in regression analyses of price, a normally distributed target leads to better model performance in that it stabilizes the variance of the underlying series, explained by Lütkepohl and Xu (2012). Price analyses like this one often sport a skewed price distribution in which supply on the lower end of the price spectrum is more heavily represented. In Figure 3.3, this stabilisation in the booking prices can be seen with the log-price distribution on the right:

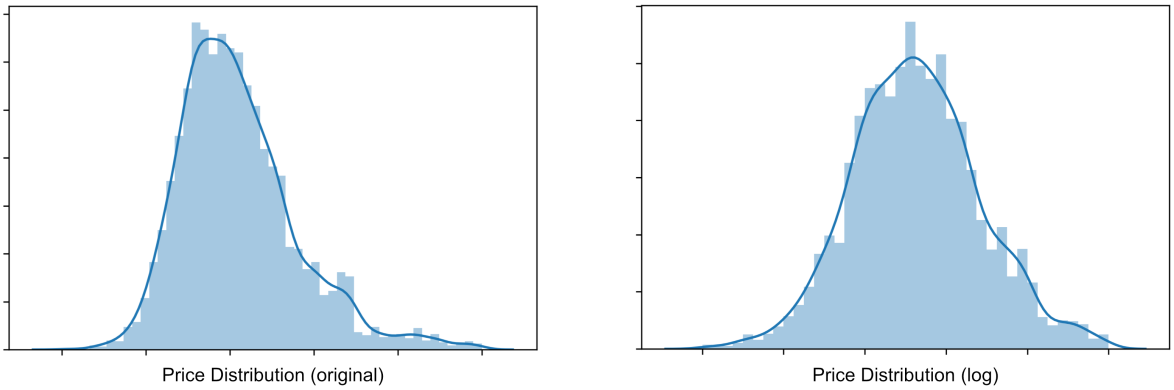

Figure 3.3: Price Distributions

In respecting the economic and political interests of Wunderflats, prices are scale-obfuscated, such that true booking prices remain unknown. This does not, however, have any material effect on the ability of this thesis to achieve its purpose in that concrete coefficients are secondary rather to the relative performance (signal magnitude and significance) of features derived from the public and big data sources of interest.

Additionally, the obfuscated location of apartments provide the basis by which access to environmental / spatial amenities is calculated. Specific locations of apartments are abstracted, again, in the interest of data security & privacy. Geospatial obfuscation can take a complex form, as seen, for example, in work from Zurbarán et al. (2018). In the scope of this thesis, however, a simple method is taken, in which a representative point is taken in the immediate area of the unit as a surrogate location. Through this approach, requisite feature measurements can still be made for engineered features without disclosing the precise object location itself. Finally, the log price per metre squared is is considered. Particularly luxurious apartments and other outliers which would skew regression results (apartments with a significantly deviating price (EUR/$m^2$) or area ($m^2$)) are removed from the dataset using the common IQR method (Seo, 2006).

In considering object floor, the architectural history of Berlin is taken into account. According to Emenlauer et al. (2018), the majority of pre-war city planning ordinances restricted house heights to between 5 and 7 floors, depending on the time period. In the post-war time, however, more liberal approaches were taken by both of the ruling bodies in Germany, resulting in numerous high-rise housing projects being constructed in both East and West Berlin. Furthermore, recent, modern policies on new developments in city also allow for high-rise apartment buildings. The instances in which the 'floor' feature exceeds a value of 7 are therefore often correlated with apartments being either in post-war, high-rise buildings or new-build developments. The distinction of year-of-build is addressed with building-age features from the Berlin Senate dataset, described in Figure 3.4. Unfortunately, the property data itself does not provide specific information here.

Descriptive statistics, variable descriptions and expected effect signs are presented in Figures 3.1 and 3.2.

| Variable | Description | Mean | Stdev. | Exp. Sign |
|---:|:---:|:---:|:---:|:---:|
| Floor | Floor in Building | 2,63 | 2,14 | + |
| Area | Area in $m^2$ | 55,86 | 22,95 | + |
| Rooms | Count of Rooms | 1,88 | 1,18 | + |
| Accommodates | Object Capacity | 2,34 | 1,04 | + |

Table 3.1: Property Dataset Features (Numeric)

| Variable | Description | True | False | Exp. Sign |
|---|---|---|---|---|
| Elevator | Presence of Elevator | 2.540 | 1.947 | + |

Table 3.2: Property Dataset Features (Boolean)

## 3.3 Ratings & Reviews Data

Data that represent human movement and preference in the urban environment provide unique insight for the study of housing price dynamics. Estimations or generalisations of social behaviour within a city may seek to offer value in this area but fail in capturing the fine-grain dynamics that make this kind of information so valuable. As discussed in the literature review (Section 2) and summarised by Marti et al. (2019), the volume and social nature of the data generated by Web 2.0 technologies supply urban studies new means by which to approach this information trove. Collection and successful utilisation of this data, however, is a significant challenge in and of itself. Working from methods utilised by Wu et al. (2016), this study takes advantage of the scale and detail of publicly generated social network data in the form of Google reviews data to measure the degree of visitor traffic to specific POIs of different types.

Substantial research has been done on how Internet reviews play a role in modern business success and in discoverability. Tang (2017) and Banerjee et al. (2017), as two examples, explore the role of "online word-of-mouth" in a business' bottom-line and the importance of agent's or participants' trust within the review network. Dou et al. (2012) addresses the question of whether the source of a review matters in how it affects a user's perception of a business. This touches on the core importance of utilising this sort of data – its representation of online networks which influence individual preferences and decision making, both in where to eat dinner as well as, presumably, where to rent an apartment. Ultimately, this thesis reserves questions of review content and challenging topics such as trust for the realm of further study, though it is important to note the additional depth of research available. In this thesis, the summed count of reviews for a given POI are the unit of measurement by which price effects of location-based popularity are examined. In their paper, Wu et al. (2016) focus on check-in data – digital declarations that an individual is visiting a business or location. Google reviews data differs slightly in that it does not explicitly capture check-ins but represents rather an asynchronous review of a POI. This can be regarded, however, as a sort of check-in with a looser consideration of the importance of time precision in when the event itself occurs. This loss of the time dimension is accounted for in the consistency and coverage of the data source itself; Google is likely the most comprehensive collector of data on diverse POIs.

In total, there are 18.691 data points for POIs in Berlin distributed amongst POI types as shown in Table 3.3. Food & Drink is the primary point of interest of the POI collections, making up the centrepiece of the ratings & reviews dataset and hosting the reviews data upon which local activity features are built. Similarly, review counts for parks and green spaces are collected. These reviews are used in conjunction with the Berlin Senate data, described in further detail in Section 3.4. The remaining POI types serve as POIs against which to measure distance from a given property. Groceries POIs simply cover grocery stores. Education consists of school locations at all levels of the German education system. Medical Centres refer to hospitals or collections of private practices. Shopping Centres refer to malls or large single-structure shopping centres.

| **POI Type** | **Count of Data Points** |
|---:|:---:|
| Groceries | 643 |
| Education | 2.257 |
| Food & Drink | 13.826 |
| Parks & Green Space | 1.349 |
| Medical Centre | 481 |
| Shopping Centre | 135 |

Table 3.3: POI Data Counts

For each data point from the properties dataset, proximity to the nearest grocery, respective school type, medical centre, and shopping centre is calculated in kilometres. For the recreation ratings & reviews data (food & drink) a search radius of 650 meters representing walking distance is prepared. Within this radius, counts of establishments including bars, cafés, and restaurants are summed as well as their respective counts of reviews. These sums are from here forward referred to as "local activity" features. Varying accounts define walking distance at different intervals and often note the fluidity of the value from study to study. Daniels and Mulley (2013) note in their inspection of urban public transport accessibility that walking distance can sometimes refer to distances of 400 to 500 metres, depending on the city. Colabianchi et al. (2007), on the other hand, identifies a range of buffer values in their research from 400 to 1000 metres. Wu et al. (2016) lands on a value of 1000 metres after testing a number of values in subsequent regressions. In this thesis, the selection of walking distance at 650 metres is based on a study of the average size of the Berlin "Kiez", a concept representing a small sub-community typically confined to a single postcode. In the case of the Berlin housing market, attractiveness of an apartment is often considered in reference to which Kiez the property is (or is not) in. The complete set of predictors derived from the ratings & reviews data can be seen in Table 3.4. An in-depth exploration of the most salient features here is carried out in Section 5.

| Variable | Description | Mean | Stdev. | Exp. Sign |
|---:|:---:|:---:|:---:|:---:|
| Nearest Grocery | Distance (km) | 0,33 | 0,24 | - |
| Nearest Grundschule | Distance (km) | 0,39 | 0,24 | - |
| Nearest Medical | Distance (km) | 0,34 | 0,23 | - |
| Nearest Kita | Distance (km) | 0,22 | 0,14 | - |
| Nearest Oberschule | Distance (km) | 0,53 | 0,35 | - |
| Nearest Shopping | Distance (km) | 0,85 | 0,58 | - |
| Recreation Access | Count Establishments | 135,53 | 98,50 | + |
| Recreation Popularity | Sum Review Count | 19.917 | 20.689 | + |
| Green Space Popularity | Sum Review Count | 14.201 | 12.418 | + |

Table 3.4: Ratings & Reviews Dataset Features

## 3.4 Berlin Senate & VBB Data

Municipal data is unique in that cities often allocate significant resources to the study of their infrastructure and general urban composition in a way that private actors do not. Berlin is no exception to this and offers a trove of highly-detailed data on a wide range of subjects (social, ecological, political, etc.) and their respective interplay with the city and its inhabitants. Of primary interest are the datasets on green space (parks, boulevards, etc.), waterways and standing bodies of water, and the registry on building ages throughout the city.

Daams et al. (2019) and Kim et al. (2019) highlight and quantify the value-added effect of green space in urban environments on housing prices in Amsterdam, Netherlands and Busan, S. Korea, respectively. The primary source for this thesis, Wu et al. (2016), displays furthermore the impact of natural greenery in the Shenzhen real estate market. With the Senate green space data, this analysis can be performed on the Berlin market as well. The green space dataset (Grünanlagenbestand Berlin) consists of 2.525 parks or similar green areas. Each observation has an object name, area in m$^2$, and an array of coordinates making up its shape and identifying its location. Similar to how proximities and activity scores are generated using the food & drink data from the ratings & reviews dataset, so too are values for green space proximity and activity values. Once more, a 650 metre search radius is drawn from each unique apartment from the properties dataset. In this case a green space is considered nearby if any part of its outer, bounding polygon falls within this search radius. Instead of creating the *green space access* feature from the simple count of nearby observations, however, the entire area in m$^2$ of the space is summed with that of the other nearby objects. Of further interest is the popularity of these green spaces. In joining the green space dataset from the Berlin Senate with the ratings & reviews dataset, review count data can be brought into the analysis. This way, following the same methodology by which local activity can be measured via bar, café, and restaurant reviews, so too can an impression for the value of a given park or green space be determined beyond simply noting its size.

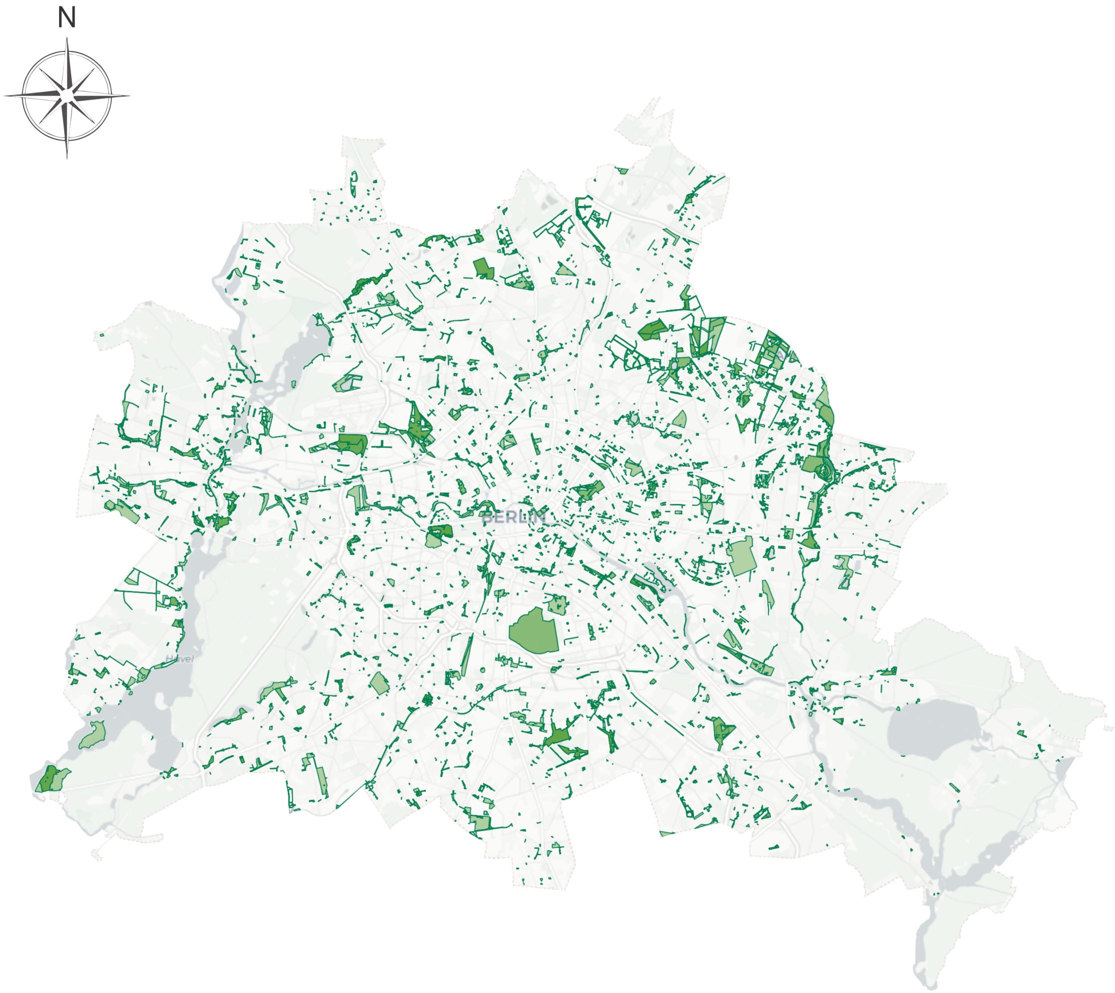

Figure 3.4: Berlin Parks Data (Development and Housing, 2020) 🔍 Green Space

Following the same logic, data on water proximity is also assessed to determine how an apartment's access to running or standing bodies of water affects the property's value. Similar to the green space dataset, the Berlin city datasets on bodies of water (Gewässerkarte & Gewässerverzeichnis) provide geospatial polygon objects to represent these water bodies in addition to values on their size in m$^2$. A similar search is again performed to generate the features *Running Water Access* & *Standing Water Access*. Popularity or activity relating to water access is not supported by the ratings & reviews data. The final green space and water-related features are as follows in Table 3.7.

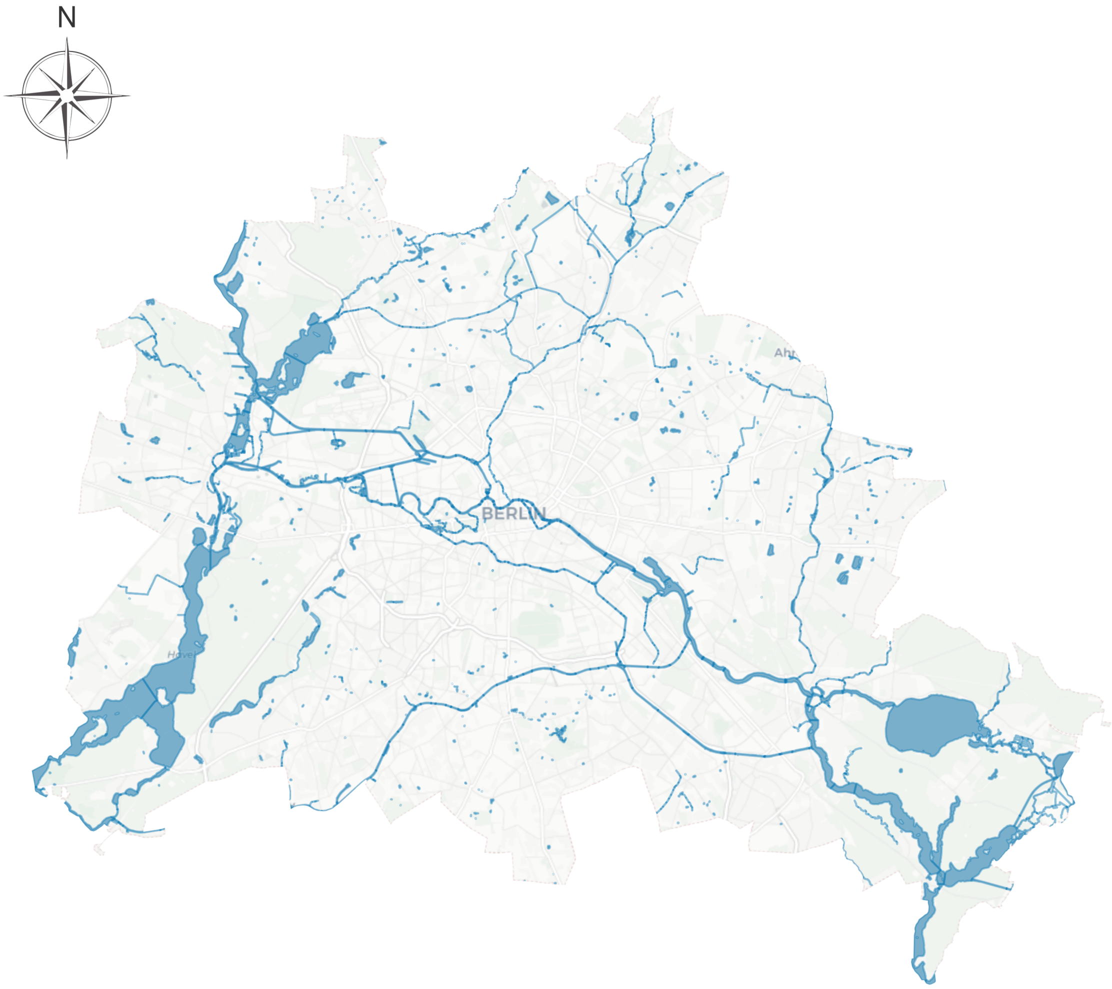

Figure 3.5: Berlin Water Data (Development and Housing, 2020) Water

| Variable | Description | Mean | Stdev. | Exp. Sign |
|---|---|---|---|---|
| Running Water Access | Sum Polygon Areas (km) | 60,48 | 375,71 | + |
| Standing Water Access | Sum Polygon Areas (km) | 2,51 | 15,59 | + |
| Green Space Access | Sum Polygon Areas (km) | 717,53 | 637,72 | + |

Table 3.5: Green Space & Water Features

The last of the three data subsets from the Berlin Senate is the dataset on building ages. While the structural variables available in the properties dataset provide a foundation upon which to estimate a property's value, these variables are not one in the same when considering a pre-war, Altbau house or a brand-new, modern project. The importance of building age is not so straightforward, however, that it can simply be stated that old houses are old and therefore less valuable and that the opposite is true for new builds. While this may often be case, Berlin, like other European cities, and especially like those that witnessed destruction at the hands of war, has a complicated architectural history. In many inner districts, for example, old houses which have been renovated now demand premiums based on their locations and mix of classic architecture with modern fittings. Post-war housing, on the other hand, is often characterised by less-refined architectural design and use of lower-quality and less-sustainable construction materials. This is mostly due to the post-war haste to rebuild in addition to the design school of thought at the time. Brand new projects run the gamut in terms of relative value but often correlate to higher energy efficiency and premiums for modern fittings and comforts. With the geographically weighted regression model's sensitivity to spatial clusterings, the expectation is that, for example, in the quarters of Berlin where older houses are in higher demand, the coefficient for old buildings will be greater than in those areas where these houses have not yet seen renovations and modernisation.

To capture this kind of variability in the relationship between age in housing stock and the value of a home, comprehensive, granular data on building ages is required. Thankfully, the Berlin Senate dataset on building ages (Berliner Gebäude Atlas) provides a nearly block-by-city-block account of the ratio of buildings in a given area classified by the decade in which they were built. In total, there are 13.091 polygons or areas in this dataset. In Figure 3.6 below, a map from Berliner Morgenpost's interactive publication *Gebäudealter Alt- Oder Neubau? So Wohnt Berlin* presents this data in striking colour, highlighting the expansion of the city over time and hinting at the destruction seen at the end of WW2.

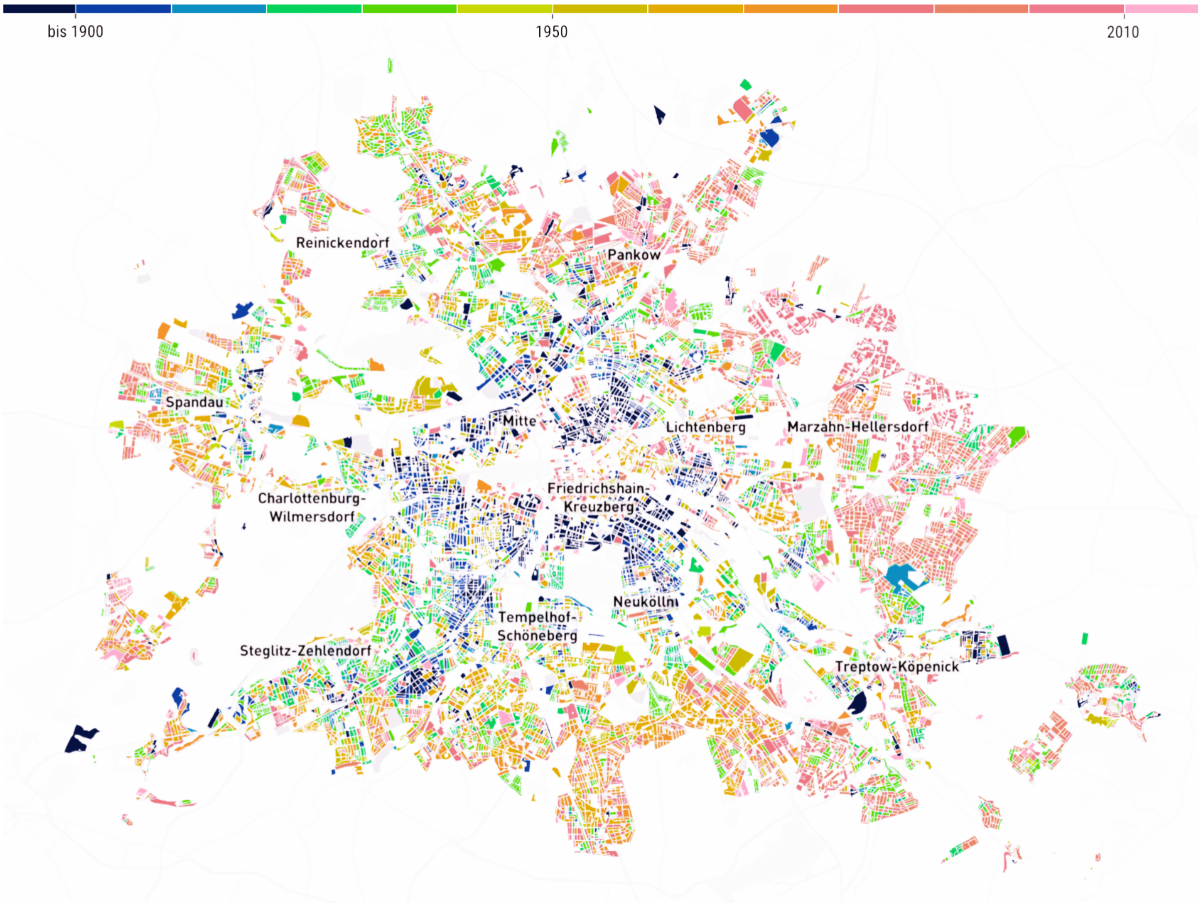

Figure 3.6: Berlin Building Ages (Timcke et al., 2018)

For this thesis, the available data values for ages of buildings are grouped into four time blocks: Altbau or Gründerzeit (pre 1900-1920), Post-Gründerzeit or Pre-War (1920-1950), Post-War (1950-1990), and Modern (1990-present). For each of the listings in the properties dataset, the area out of the 13.091 available polygons in which the apartment lies is determined. The grouped values for the ratios of building types are then assigned to that listing. These features are laid out in Table 3.6 below:

| Variable | Description | Mean | Stdev. | Exp. Sign |
|---|---|---|---|---|
| Altbau Ratio | Ratio of Total | 0,53 | 0,33 | + |
| Pre-War Ratio | Ratio of Total | 0,08 | 0,17 | - |
| Post-War Ratio | Ratio of Total | 0,31 | 0,32 | - |
| Modern Ratio | Ratio of Total | 0,08 | 0,18 | + |

Table 3.6: Building Age Ratio Features

In addition to the data made available by the Berlin Senate, the Verkehrsbund Berlin-Brandenburg (VBB) publishes geospatial information on public transit lines in Berlin. This data is useful in assessing the general mobility access of an individual property. In addition to the routes data shown in Figure 3.7, individual stop locations are also available. This allows precise distance-to-nearest-transport features to be calculated. In consideration are S-Bahn stops, U-Bahn stops, and Metrotram stops, as well as a mode-agnostic, nearest-transport feature. As a split city post-war, Berlin's public transport network developed under two divergent systems; in the East, street-level tram networks were prioritised (3.7 right). In the West, on the other hand, development focused on underground transport lines (U-Bahn; 3.7 centre). The S-Bahn infrastructure is more or less free from this phenomenon, geographically speaking (3.7 left). The split between the U-Bahn and Tram lines is still very much apparent today and is the reason for the mode-agnostic, nearest-transport feature, as distance to U-Bahn or Metrotram station is likely heavily correlated with geospatial location. This feature has the following characteristics:

| Variable | Description | Mean | Stdev. | Exp. Sign |
|---|---|---|---|---|
| Nearest Transport | Direct Distance (km) | 0,36 | 0,28 | + |

Table 3.7: Transport Feature

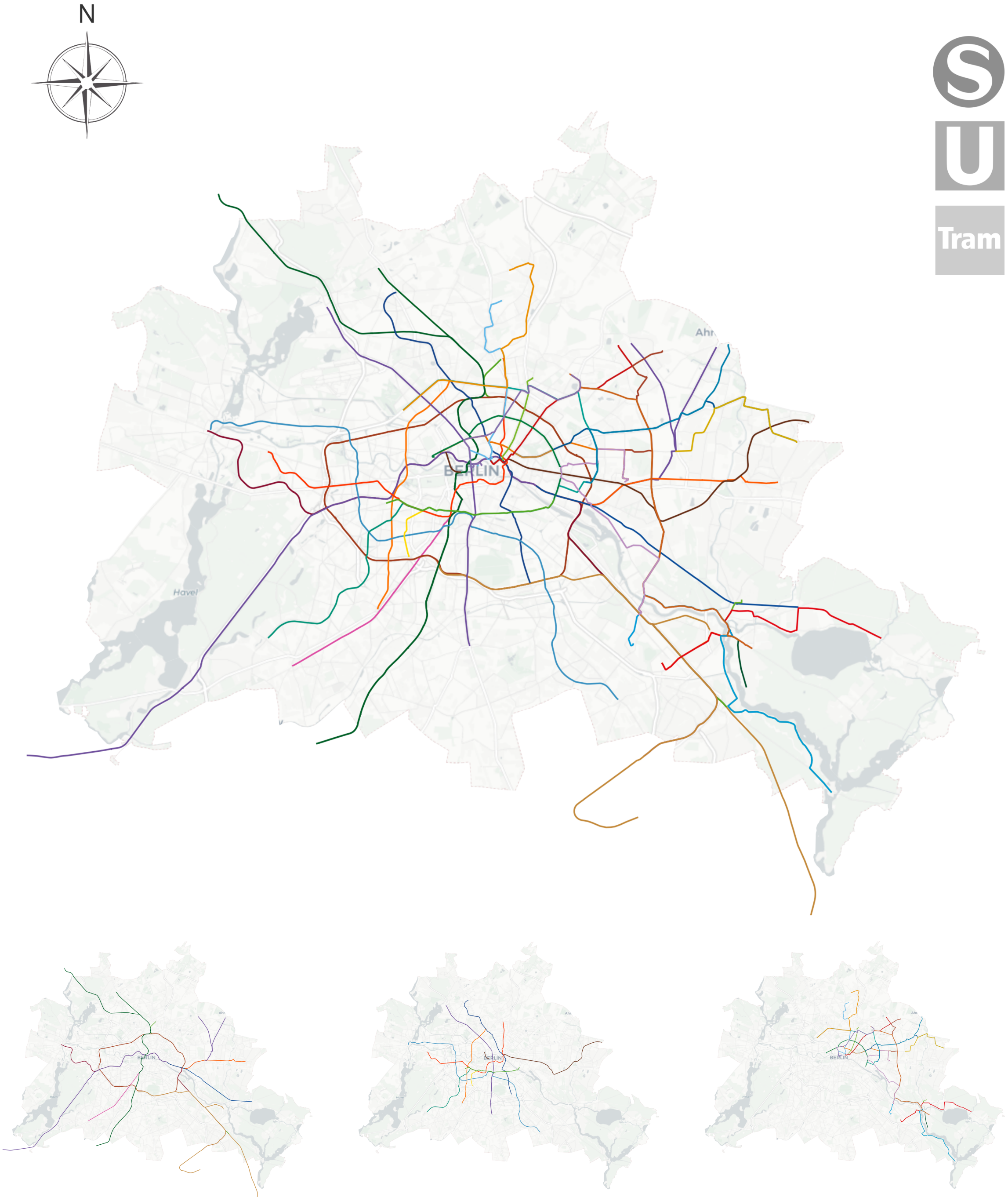

Figure 3.7: Berlin Transport Network (VBB, 2020) Transport

## 3.5 Final Dataset

Prior to modelling, the data are examined for collinearities and normalised such that they are represented in the same range of values. This is also in service of the fact that GWR generally performs better upon normalised data. The final dataset with all sources combined makes up a featurespace of 19 independent variables, dropping the feature *Modern Ratio* to reduce collinearity in the building age features as well as *Nearest Shopping* & *Running Water Access* after finding that they do not prove significant in modelling. This final working set is shown below in Table 3.8:

| **Variable** | **Description** | **Feature Set** | **Exp. Sign** |
|---:|:---:|:---:|:---:|
| Floor | Floor in Building | Structural Data | + |
| Area | Area in $m^2$ | Structural Data | + |
| Rooms | Count of Rooms | Structural Data | + |
| Accommodates | Object Capacity | Structural Data | + |
| Elevator | Presence of Elevator | Structural Data | + |
| Nearest Grocery | Distance (km) | Social Data | - |
| Nearest Grundschule | Distance (km) | Social Data | - |
| Nearest Medical | Distance (km) | Social Data | - |
| Nearest Kita | Distance (km) | Social Data | - |
| Nearest Oberschule | Distance (km) | Social Data | - |
| Recreation Access | Count Establishments | Social Data | + |
| Recreation Popularity | Sum Review Count | Social Data | + |
| Green Space Popularity | Sum Review Count | Senate Data | + |
| Standing Water Access | Sum Polygon Areas (km) | Senate Data | + |
| Green Space Access | Sum Polygon Areas (km) | Senate Data | + |
| Altbau Ratio | Ratio of Total | Senate Data | + |
| Pre-War Ratio | Ratio of Total | Senate Data | - |
| Post-War Ratio | Ratio of Total | Senate Data | - |
| Nearest Transport | Direct Distance (km) | VBB Data | + |

Table 3.8: Final Feature Set

# 4 Methodology

The methodology of this thesis progresses from exploratory methods at the outset to quantitative analysis via various regression techniques. Each of these approaches is laid out in detail in the following subsections, in which the theory behind the method is explored along with the general outcomes that the technique produces.

## 4.1 Performing & Visualising Geospatial Analyses

In the Data section (Section 3), a number of maps are used to visualise the data being described. This bridge between abstract numbers and human understanding is central not only to this piece of research but all research that intends to find a broader audience and communicate its findings in simpler terms. In the case of this thesis, the data is spatial in nature, that is it is expressed primarily in coordinates on a map. To handle such data in Python, special object types are required, namely the GeoDataFrame. GeoDataFrames are objects from the Geopandas Python package (Jordahl, 2014), which is built atop the original Pandas library (McKinney et al., 2010) in combination with the geospatial library Shapely (Gillies et al., 2020). What differentiates a normal Pandas DataFrame object from a GeoDataFrame is the *geometry* column in the frame. By definition, the column is reserved for data in the form of a Shapely geometry object, which follow the conventions set by the GeoJSON encoding format (Butler et al., 2016). This includes data types like *polygons*, *multi-polygons*, *points*, and *linestrings*, all of which are utilised in this thesis. With a standardised geometry series in the GeoDataFrame, methods can then access those geometries to perform a myriad of spatial operations. Many packages that work with GeoDataFrames require geometry coordinates to be expressed as a projection in a particular coordinate reference system (CRS).

While the well-known latitude and longitude coordinate system provides spatial precision, it also inherently exists in a 3-dimensional space, given that the Earth is an ellipsoid and this system is defined in the unit of degrees. This problem has, however, long since been solved by what in cartography are known as map projections, described by Skupin

(2000) as well as Lawhead (2013), who provides a thorough introduction to geospatial analysis with Python in general. A cartographic projection is a method by which to flatten the 3-dimensional globe onto a 2-dimensional space, requiring a transformation of latitude and longitude points into coordinates on a plane. This process, by nature, results in some distortion of the real-world geography. In some cases, this distortion is minimal and therefore negligible, while in others, it is too far of a departure from the subject matter being mapped. When projections consider a limited space on the globe, however, the projection distortion is likewise limited, as the Earth's curvature is, in smaller areas, less and less apparent. Considering geospatial data in 2-dimensional terms of x and y coordinates allows for performing meaningful statistical analyses and thus underlies the importance of (suitable) map projections. This thesis utilises Python statistical and geospatial packages which are built atop this methodology.

## 4.2 Hotspot & Coldspot Analysis

One basis of this study's exploration is the hypothesis that higher-property-value areas, and therefore areas with higher rent prices, are in urban localities characterised by higher-than-average recreational activity. To consider how well the data at hand helps in exploring this, exploratory examination of the clustering of certain POIs is carried out. These clusters can be referred to as *hot spots* and can be derived in two different ways, outlined by Xie and Yan (2008). In one sense, hot spots can be defined by the geographic and/or spatial clustering of locations. The method by which to examine this is called Kernel Density Analysis (4.2.1). While this accounts for numerous POIs that form spatial clusters, popularity of locations is not taken into account. The method to this, the more salient feature for price correlation, is called the Getis-Ord statistic (4.2.2). Both of these considerations of spatial hot spots are included in the final regression dataset in the two features *Recreation Access* and *Recreation Popularity / Green Space Popularity*.

### 4.2.1 Kernel Density Estimate Analysis

Kernel Density Estimation (KDE) is a method commonly used to explore the distribution of a given feature in a dataset (Kalinic and Krisp, 2018). Take, for example, the most common method to observe such distributions – the histogram. A histogram describes discreet subsets or bins of a feature. A kernel density estimator differs in that it smooths the observations, often with a Gaussian kernel, resulting in a continuous representation of the underlying data. In this thesis, the geospatial distributions of points in 2-dimensional space are observed, namely the positional data for restaurants, cafés, and bars. The general form the KDE is as follows:

$$f(s) = \sum_{i=1}^{n} \frac{1}{\pi h^2} k\left(\frac{d_{is}}{h}\right) \tag{4.1}$$

Here, *f(s)* represents the estimator as a function at location $s$ with $h$ as the bandwidth. Furthermore, $d_{is}$ is the distance from point $i$ to the location of interest $s$, and the function $k$ represents the function relating to $d_{is}$ and the bandwidth $h$. The KDE function is employed via the KDE method from the Python *geoplot* package (Bilogur, 2020). This method is based on the *seaborn kde_plot* method (Waskom et al., 2017), which is furthermore built upon the *scipy* statistical method *gaussian_kde* (Virtanen et al., 2020). This produces a visualisation layer atop a map generated by other mapping packages, which plots the result of a Gaussian kernel density estimator. KDE is a non-parametric method. The bandwidth $h$, however, is important in determining how the smoothing function transforms the original distribution. A larger $h$ results in a more generalised representation of the data, while a smaller $h$ does the opposite – it is a trade-off between high bias (simply re-representing the original data) and low variance (over-smoothing the data). Scipy's *gaussian_kde* notes the importance of bandwidth selection and states that it uses a common rule-of-thumb method for this task known as Scott's Rule (Kamalov, 2020). Manual scalar adjustments can be made to the auto-generated $h$ value; however, after testing higher and lower multipliers, this study finds that the default values produce satisfactory results and represent the expected spatial distributions well.

### 4.2.2 Getis-Ord Hot-Spot Analysis

KDE analysis addresses the spatial clustering of points. To understand value-based clustering, however, that is areas where a particular feature has a grouping of significantly higher or lower values than average, a statistic called the *Getis-Ord* $G_i^*$ is used. The $G_i^*$ statistic deals with local auto-correlation in contrast to the rest of the data (Ord and Getis, 2010). In the case of spatial data where the x and y variables are coordinates, this conceptually means how self-similar localities are and how they compare to map-level averages. The statistic is calculated on the basis of a individual location $i$ in the context of data at other locations $x_j$ and with a spatial weighting of $w_{ij}(d)$. Typically, the result of the distance weight function is binary, that is $w_{ij}$ is equal to 1 when the distance between points $i$ and $j$ is less than bandwidth $d$. When the opposite is true $w_{ij}$ returns a 0. Formally the $G_i^*$ statistic can be represented as follows:

$$G_i(d) = \frac{\sum_j w_{ij}(d) x_j}{\sum_j x_j} \tag{4.2}$$

Again, the bandwidth is central to this statistic in helping to determine what qualifies as being considered for in-cluster and what does not. In Python, $G_i^*$ is generated via the *PySAL* package (Python Spatial Analytics Library) (Rey and Anselin, 2007). Here, the bandwidth is selected heuristically by the user and is in the units of the geography of the GeoDataFrame, which, in turn, is based upon the projection of the spatial data points. In the case of this thesis, that projection is EPSG:3068 DHDN / Soldner Berlin (Pridal et al., 2004), whose unit of measurement is metres. In keeping with consistency, this threshold is set to 650 metres as it is in the feature construction in Section 3.3. Hotspot and coldspot consideration is therefore done on the basis of gathering neighbouring points (within 650 metres) for a given point $i$ and determining if the values for this "neighbourhood" are significantly higher or lower than the global average of the particular feature being examined. This significance is determined via Z-scores & corresponding p-values derived from *PySAL* in computation of the $G_i^*$ statistic itself.

The basis for this thesis' application of the $G_i^*$ statistic comes from Levi (2019b).

## 4.3 Regression Analysis

KDE and the Getis-Ord $G_i^*$ provide fundamental exploratory analysis of the primary input features as well as the target distribution (see results Section 5.1 for further detail). This provides the basis for map-based visual analysis of activity hotspots and price hotspots, but these relationships need to also be represented more concretely, that is quantitatively. Furthermore, the degree to which local activity plays a role in price development is important to clarify in the context of other variables. Regression analysis provides the means by which to derive these statistics. This thesis focuses on geographically-weighted regression (GWR), using a global hedonic regression as a benchmark against which to measure and qualify the GWR results. The hedonic pricing method is first introduced at a conceptual level (4.3.1), followed by a more granular explanation of GWR. The GWR modelling in Python produces both localised models and a global model. Given this thesis' feature set, these can be considered hedonic regressions. The mathematical methodology for both is therefore explained in the GWR section (4.3.2).

### 4.3.1 Hedonic Pricing Method

The simplest regression models which deal with housing prices focus solely on the structural makeup of the object (number of rooms, how big it is, etc.). The limitations of this, however, are easy to see, in that this simplistic approach entirely ignores the interplay between a property and its environment. The hedonic pricing method is a standard method for accounting for these environmental variables and is, in fact, no different than a standard regression in its mathematical makeup (Thériault et al., 2003). What defines a hedonic regression is simply that the independent variables used to predict the target are explicitly intended to represent the contextual features that make up the property's surrounding environment. Consider that the word *hedonic* comes from from Greek *hēdonikos*, from *hēdonē* meaning "pleasure".

In this sense, this thesis fulfills this description quite completely, in that the majority of the predictor features focus on the spatial surroundings of the properties studied.

Indeed the most dynamic and important data sources of the study deal directly with quality of life and recreational pleasure in specific localities. These recreational POIs, concentrations of which make up the most desireable neighbourhoods, represent, in turn, the highest value property locations. In the following section describing the methodology of geographically-weighted regression, the structure of an ordinary least squares (OLS) regression is explored, which also accounts for the foundation of a hedonic regression.

### 4.3.2 Geographically-Weighted Regression

Geographically-Weighted Regression (GWR) is based fundamentally on the idea that effects from independent variables on a dependent variable of study are not necessarily the same over the study space at large. This concept is known as non-stationarity, and addressing it is an important cornerstone of many popular and effective spatial analysis models. Given the case of real estate analysis, it is easy to imagine that effects between features and target in one neighbourhood are not the same as in another. Especially in a city like Berlin with such a unique history and modern-day makeup, it is important to consider that the context in which a given property lies is dynamic and must be treated as such. GWR is a method that provides a means by which to model variable effects in a non-stationary space like a large metropolis. This is achieved by subsetting the original dataset of properties into local datasets unique to specific localities upon which regressions are run. The basic form of a regression model looks like this:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{4.3}$$

GWR takes this structure and modifies it by subsampling the target ($y$) and independent variable input spaces ($X$) for each location $i$. The point of this is to focus on data near to an observation $i$ (i.e. within the same neighbourhood) rather than observations far away, like properties in a district with an entirely different social or environmental makeup. In order to identify which observations are eligible "neighbours", GWR employs a spatial weighting matrix $W$ that is $N\ x\ N$ where $N$ is the total number of observations. For a given property location $i$, its spatial weighting matrix $W_i$ is a diagonal matrix, having elements only on the diagonal and elsewhere values of 0. Each observation in the dataset apart from $i$ is represented by a value $j$ on this diagonal, whereby its value (between 0 and 1) represents its importance in reference to the original $i$. When $W^i_j$ is near a value of 1, this indicates geographical nearness, and the opposite is true for values closer to 0. $W^iy$ is therefore the subset of target-feature observations where responses are near to $i$, with $W^iX$ representing the same but for predictor (independent) variables. This is formulated as follows:

$$\hat{\beta}^i = (X'W^iX)^{-1}X'W^iy \tag{4.4}$$

This now represents a model, which is focused specifically on the property location $i$ and nearby observations. Here, $\beta^i$ is the localised effect of $X$ on $y$. Translating this now to the original OLS regression statement above, the final GWR form is prepared:

$$W^iy = W^iX\beta^i + \epsilon^i \tag{4.5}$$

With this, GWR provides localised effects $\beta^i$ at each individual site $i$. In practice, this results in $N$ models, which are, in turn, interpreted in their relation to a more general relationship between normalised versions of $X$ and $y$ where each has a mean of 0 and variance of 1, respectively. In this approach, GWR examines the local effects in their strength relative to the global average effect over the entire set of observations.

Just like with KDE and the Getis-Ord statistic, the degree to which GWR considers locality represents a trade-off between overly-specific and overly-general representations of the modelled behaviour in the data. In order to capture local relationships between predictors and target, the GWR must be sensitive enough to exclude non-relevant data points. If too sensitive, however, the model may perfectly capture the $\beta^i$ effect, producing an error of 0 and ultimately over-fitting. In the opposite direction, if the GWR is too agnostic, it will, at its extreme, simply become a global OLS model, negating the original use case. This trade-off again hinges on yet another bandwidth parameter to determine what qualifies as a neighbouring point. In the GWR model provided by the Python PySAL package (Rey and Anselin, 2007), automatic bandwidth selection is performed using the Golden Section Search method (Koupaei et al., 2016) in conjunction with the adjusted Akaike Information Criterion (AIC$_c$) selection criterion (Bozdogan, 1987). The GWR produces both a global results and the $N$ location-specific regressions where he global $R^2$ takes no spatial weighting into account. The summary output of the weighted regressions is a weighted average of local $R^2$ values, which carries the implication that certain areas of the study produce better- or worse-than-average local models. The ability to inspect a specific model $GWR$ at location $i$ and its respective $R^2$ is helpful in that it allows for decomposition of model performance for specific localities.

The basis for this thesis' application of GWR comes from Levi (2019a).

# 5 Results & Discussion

In this section, the methods introduced in Section 4 are applied to the data presented in Section 3 with specific focus on the recreation and green space activity features. Firstly, the analyses on POI point densities (KDE 4.2.1) as well as POI and target-value hotspots and coldspots (Getis-Ord $G_i^*$ 4.2.2) are explored. Given the additional insights produced here in conjunction with introductory information from Section 3, a foundational understanding of Berlin's structural and environmental makeup is established in regard to centres of activity and agglomerations of housing value. These hot- and coldspots in reference to property price for the target value, achieved via the $G_i^*$ statistic, allow a cursory correlation to be drawn between local activity and the housing value in the same area. To substantiate this, GWR results analysis provides quantitative support of effects observed in exploratory mapping & visualisation. This regression includes, as well, additional features that are not included in the hotspot analysis but are important in producing as precise model as possible. These are features derived from the Berlin Senate data on green space, waterways, and building ages, the VBB data on transport accessibility, and the remaining features from the ratings & reviews dataset (Section 3.3). These additional features help in making the case supporting the value of municipal data and other aggregators (i.e. Google and other social media platforms).

## 5.1 Hotspot & Coldspot Analysis

Hotspots and coldspots are identified either by high or low densities of observations themselves or by high and low densities of a particular value for each observation. In the following subsections (5.1.1 & 5.1.2), visualisations of the KDE and $G_i^*$ are examined and described in detail. For those with a good knowledge of the city structure of Berlin in physical, cultural, and economical terms, these results will confirm priors that one would expect to see, that is where activity is focused in different city sub-localities.

### 5.1.1 Kernel Density Estimate Results

Kernel density estimation (KDE) is a process that smooths the distribution of observations to give a continuous representation of the underlying data. In 2-dimensional space, this means mapping coordinate locations with high numbers of observations and then representing how and in which direction the density of data points tapers off.

In Figure 5.1, the full city map of Berlin is displayed. The orange ring represents the S-Bahn Ring (an important geographic delineator in the city), and the coloured swatches atop the map show the resulting densities from the KDE function. Localities of interest are numbered here in ascending order by size or by the context they add to the interpretability of the map. Particularly interesting to note is the extreme variability in POI density between the city within the S-Bahn Ring and the remaining localities outside of the Ring. Zooming into this space where the vast majority of spatial activity is, these sub-centres of Berlin can be more closely examined in Figure 5.2.

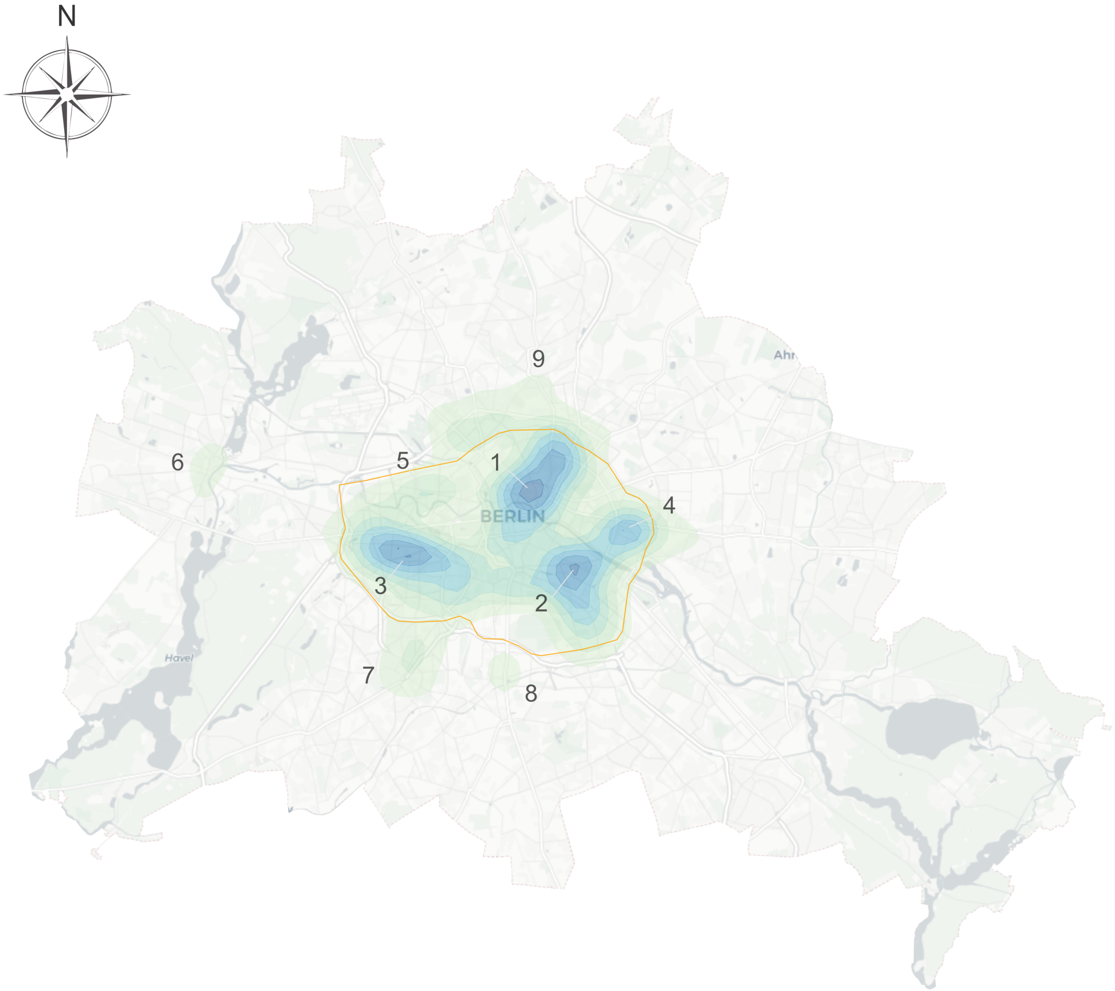

Figure 5.1: KDE Visualisation

The city-wide KDE results are encouraging, as they effectively capture or represent the parts of Berlin which are anecdotally the most popular and active. **Area 1** is the largest, high-density area and represents the fashionable, central district of Berlin Mitte. In this northern part of the city's centre is a high concentration of bars and restaurants stretching from Alexanderplatz, through Hackescher Markt & Weinmeisterstraße and up to area of Rosenthaler Platz & Torstraße. This area, which following the fall of the Berlin wall, was rather run down, is now one of the most prosperous locations in the city. The observation density extends somewhat north-eastward, encapsulating the central neighbourhoods in Prenzlauer Berg. **Area 2** is Berlin-Kreuzberg and the northern tip of Berlin-Neukölln, another popular area due to its alternative attitude and traditionally lower cost of living. This popularity has also translated to increasing demand and property speculation, resulting in higher property values & rent prices. **Area 3** shows the slightly-less-centralised, western centre of Berlin – the Kurfürstendamm area in Charlottenburg. **Area 4**, finally, represents Berlin-Friedrichshain, which appears not to sport quite the same density of recreational (food & drink) establishments as the other centres. Points **5, 6, 7, 8, 9** note other recognisable localities and help to lend credibility to the KDE results, and, by extension, the underlying data. These locations are Moabit, Spandau, Friedenau, Tempelhof, and where the northern band of Prenzlauer Berg meets the district of Berlin-Wedding, respectively.

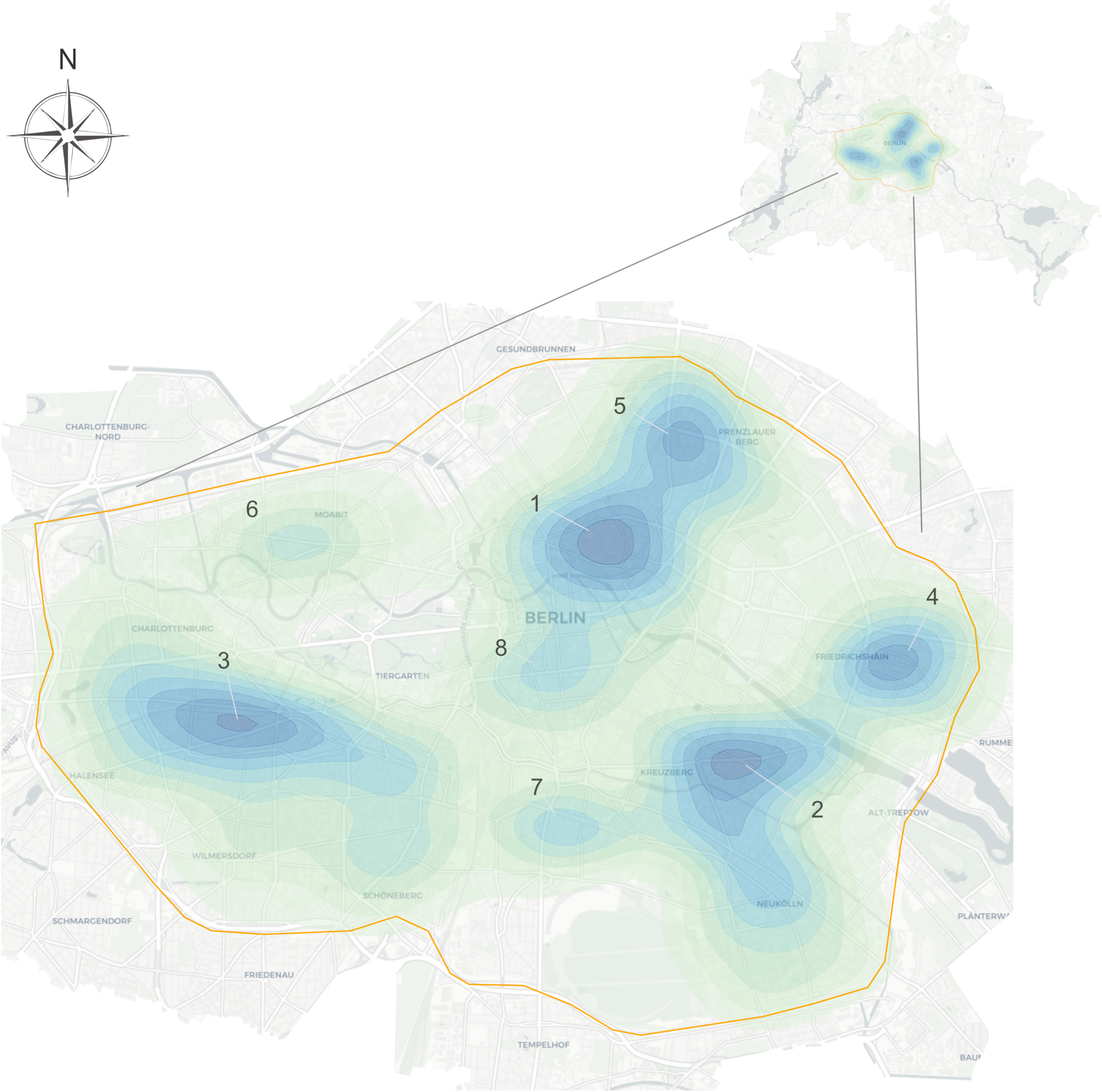

Figure 5.2: KDE Visualisation (within Ringbahn)

Looking more closely at the inner-city districts, pockets with lesser but still notable density become more clear. Areas **1, 2, 3 & 4** remain the same with added detail in how their high, central distributions drop off and in which directions. For example, **Area 7** in this map now shows the popular Kreuzberg district of Bergmannkiez, also known for its culinary offerings, but with less volume than the more easterly corner of Kreuzberg as a whole. Also more apparent is the distinction between areas **1, 5 & 8**. **Area 5** now shows the area around Eberswalder Str. U-Bahn station and Schönhauser Allee S-Bahn station as its own nucleus, independent from the central **Area 1**. **Area 8** captures the business district surrounding Friedrichstraße. The eastward direction in which **Area 3** stretches indicates the centre of the Berlin-Schöneberg district. The expectation from this exploration of the point density of recreational POIs is twofold: one, that the popularity distribution of these POIs should also follow a similar distribution, and two, that the value of interest, property price, will also mirror this pattern.

### 5.1.2 Getis-Ord Hot-Spot Results

The Getis-Ord $G_i^*$ statistic helps to identify points within a distribution which differ significantly from the overall distribution of the feature values. In cases where groupings of data points are notably smaller than the expected value, these groupings can be considered "coldspots". In the other direction, where values are higher than the the distribution average, a "hotspot" can be identified. As explained in Section 4.2.2, what determines membership for these groups is based on the *bandwidth* parameter for the method, which, in this study is 650 metres. Three feature distributions are examined in the context of the $G_i^*$ statistic: the regression target, price, and two primary independent variables of interest: recreation & green space activity data, respectively. The spatial distribution of properties within the Berlin S-Bahn ring is more stable in comparison to Berlin's edge districts. It is also where the most significant variance occurs within the studied features' distributions. For this reason, the $G_i^*$ analysis takes place primarily within the Ringbahn.

Figure 5.3 shows the $G_i^*$ statistic results for the price distribution of listings mapped over the central districts of Berlin. For each listing, its price, averaged along with that of its

$n$ neighbours, is compared to the global price average. Specific points are determined as members of hot- or coldspots. These clusters are abstracted by postcode in consideration of data sensitivity. The trends remain clear to see and show a distribution of high and low price consistent with expected results.

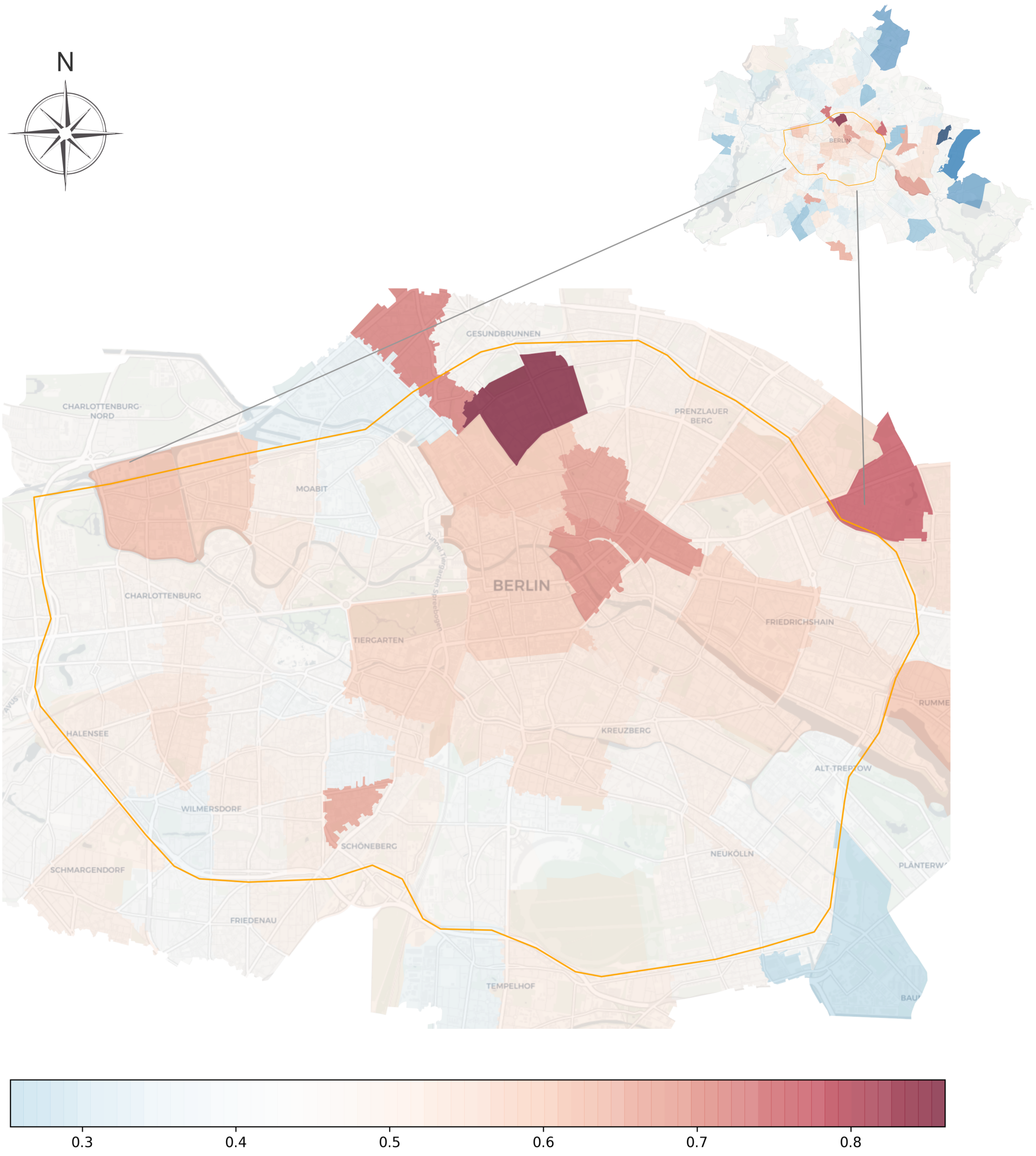

Figure 5.3: $G_i^*$ Price (within Ringbahn)

Postcodes in the Mitte district of Berlin show the highest clustering of extreme hotspot values with the magnitude decreasing in an outward-radiating fashion. Other noteworthy areas are the warmer areas in northern Berlin-Charlottenburg, and southward in the centre of Charlottenburg and where it meets Berlin-Schöneberg. Throughout Friedrichshain-Kreuzberg, price values are also higher. Interestingly, these price hotspots are well in line with the KDE distribution seen in Figure 5.2, indicating that property value shares a correlation with density of recreational (food & drink) establishments in the area. Perhaps more interesting, however, is how this price distribution relates to that of the popularity of the recreational establishments shown in Figure 5.4:

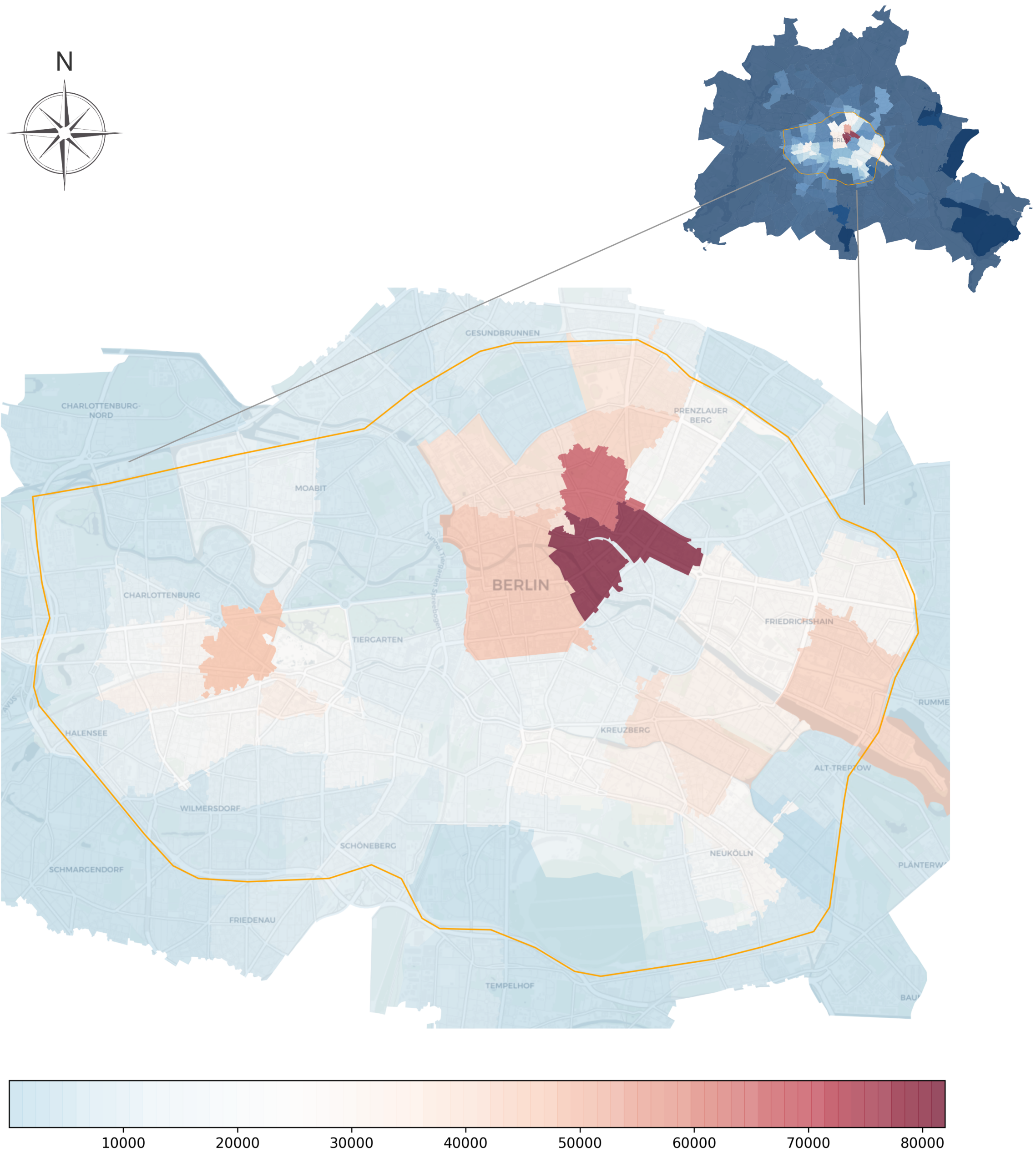

Figure 5.4: $G_i^*$ Recreation (within Ringbahn)

Firstly, note the sparse presence of non-blue postcodes as well as the abundance of very dark blue postcodes in the city-wide map. This hotspot-coldspot distribution illustrates highly concentrated centres of activity in the gastronomy space confined primarily to Berlin-Mitte, Prenzlauer-Berg, the Ku'Damm area in Charlottenburg, and the eastern section of Kreuzberg. This kind of concentration suggests the existence of a power-law or similarly long-tailed distribution amongst restaurant, bar, and café popularity (Wilhelm and Hänggi, 2003), that is that a very small number of establishments receive the overwhelming majority of attention and traffic from customers. This kind of behaviour self-enforces and continues to concentrate, in this case business, amongst a small number of locations. Similar to the case of the KDE visualisation, price-value hotspots appear to share the same areas of the city as the high-popularity areas for food & drink. Quite interestingly, it appears that the high-priced postcodes, according to Figure 5.3, are not necessarily those with the highest concentrations of recreational establishments but rather the neighbouring postcodes. This suggests, perhaps, that a sort of medium proximity to these recreation hotspots in contrast to immediate proximity is appealing.

The $G_i^*$ statistic result for access to green space in metres is calculated as well and is shown below in Figure 5.5. Here, the urbanisation of Berlin can be seen in the less-coloured postcodes. Access to green space is again calculated as a sum of the area in m$^2$ of all parks or similar spaces in the vicinity (650 metres) of a given apartment. At the very high end of the scale, for example, listings nearby the former Tempelhof Airfield record access to the park's 2.682.234 m$^2$ of space. Otherwise, greenery is more abundant along the edges of the inner city, with the exception of Tiergarten and Volkspark Friedrichshain. While this analysis of green space hot and coldspots provides a further insight into the distribution of access to nature throughout Berlin, a clear visual relationship to the value of housing is not apparent. More concrete, quantitative relationships between inputs (independent variables) and the variable of study, price, are examined in the final section of this thesis.

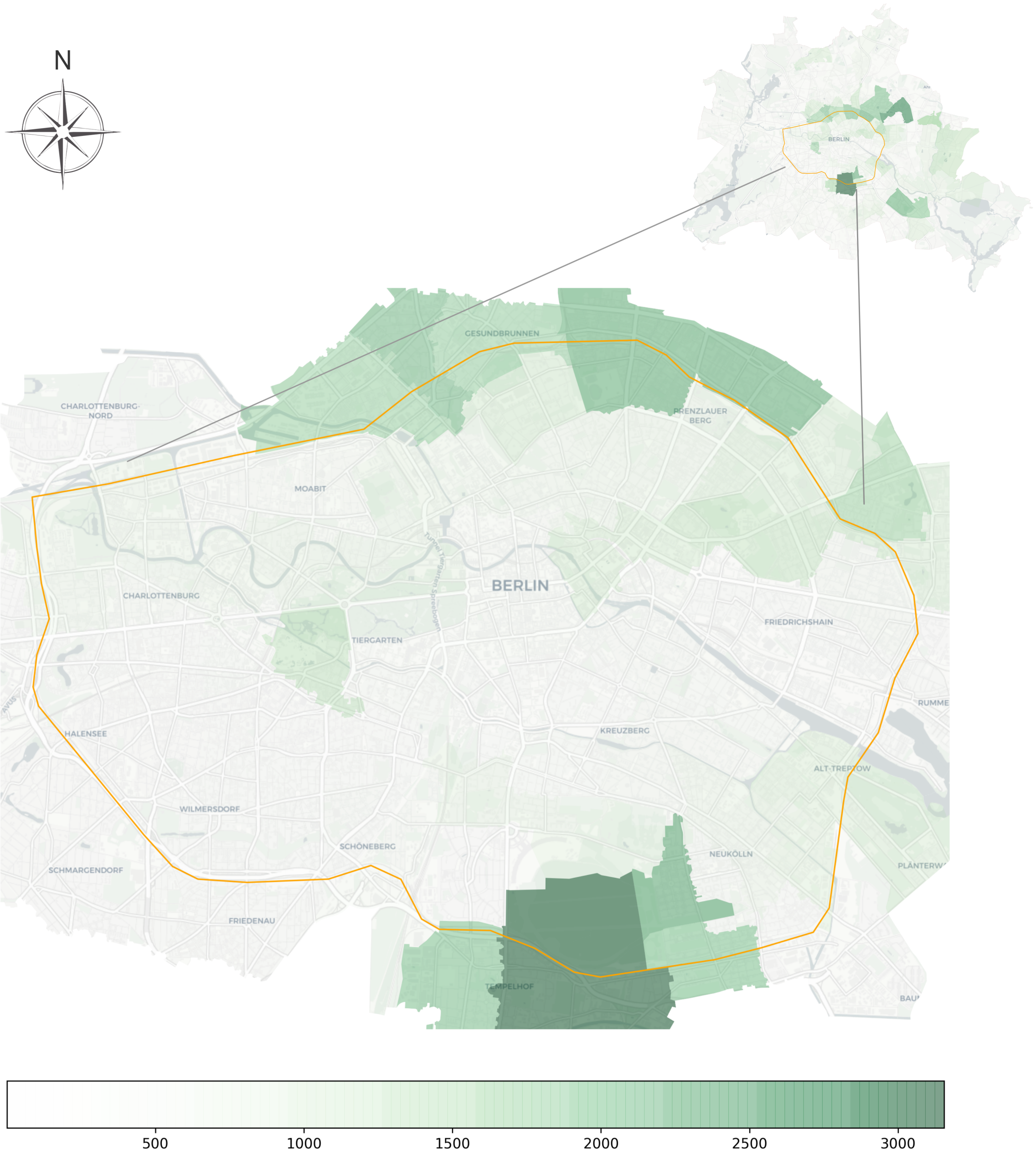

Figure 5.5: $G_i^*$ Green Space (within Ringbahn)

## 5.2 Analysis of Price Modelling

Results of the global regression and the aggregated local regressions are examined in the two subsections below. Section 5.2.1 observes the high-level metrics for evaluating the overall model fit and how the spatial sensitivity of the GWR affects the quality of the insights produced. The GWR model is benchmarked against the global regression (evaluating all points equally) in order to substantiate the value of localised regressions. It is also compared to a similar GWR model which does not contain the novel data points gathered for this study. This is to quantify the informational value of these additional features. Coefficients for each of the input features are also explored here. Section 5.2.2 looks further into the behaviour of specific independent variables and how the strength of a feature's signal varies over space in its relationship to the prediction target.

Data for this regression consists of observations covering all of Berlin. The clarity of individual feature effects are, however, more reliable and clear to interpret when limiting the space of study to properties within the Berlin S-Bahn Ring. Results in sections 5.2.1 and 5.2.2 will therefore represent the regression estimates within this geographic scope, trimming the number of observations from 4.021 to 3.410.

### 5.2.1 Global Model vs. GWR

Section 4.3 introduced the concepts of a regression model following the hedonic pricing method (HPM) along with that of the spatially-sensitive geographically weighted regression (GWR). The GWR method in the Python *PySAL* library (Rey and Anselin, 2007) produces results for both a global regression and the average output of $N$ local regressions, where $N$ is the number of observations. Both the global and location-specific regressions can be considered as following the hedonic pricing method in that they model not only structural details but also contextual features relating to utility provided by a property's surrounding environment. Table 5.1 shows the adjusted and unadjusted values for the evaluation metrics AIC and R$^2$, respectively. Comparing these, the additional accuracy offered by the GWR model is clear. Most noteworthy is the 73% increase in the R$^2$ value between the global and weighted regressions. This indicates, and is shown in Figure 5.6, that effects from covariates are non-stationary throughout Berlin. Local models are therefore able to account for relationships within the selected bandwidth of the GWR that may not exist in the same way elsewhere in the sample. Bandwidth auto-selection returns an optimal bandwidth of 96 for the GWR, meaning that each of $N$ local regressions is done on a property $i$ and its 96 geographically nearest data points.

| **Regression Inputs & Parameters** | |
|---|---|
| Observations | 3410 |
| Covariates | 19 |
| GWR Bandwidth | 96 |

| **Global Regression Results** | |
|---|---|
| AIC | 7558.83 |
| AIC$_c$ | 7561,10 |
| R | 0,432 |
| R$^2$ | 0,429 |

| **GWR Averaged Local Models** | |
|---|---|
| AIC | 5728,94 |
| AIC$_c$ | 6894,11 |
| R | 0,828 |
| R$^2$ | 0,742 |

Table 5.1: Regression Summary Metrics

To make further sense of the local R$^2$ values, they are visualised in Figure 5.6 by postcode. This is a more intuitive way to see where the model has a better or worse fit. Take for example the postcode of 12049 in Berlin-Neukölln or Berlin-Mitte where the model fit is particularly good, held in comparison to the lighter postcodes in Berlin-Charlottenburg to the west where model fit drops off. The poorer fit in the western heart of the city is particularly interesting as it seems limited to the area around Ku'Damm and Zoologischer Garten. The most likely explanation for this is simply that this area of the city has a more complex dynamic to its housing sub-market, which is not fully captured by the features selected for this study. The areas in which the model performs well, on the other hand, could perhaps be described as more mono-dimensional, that is less complicated and therefore easier to model in regards to their attractive characteristics.

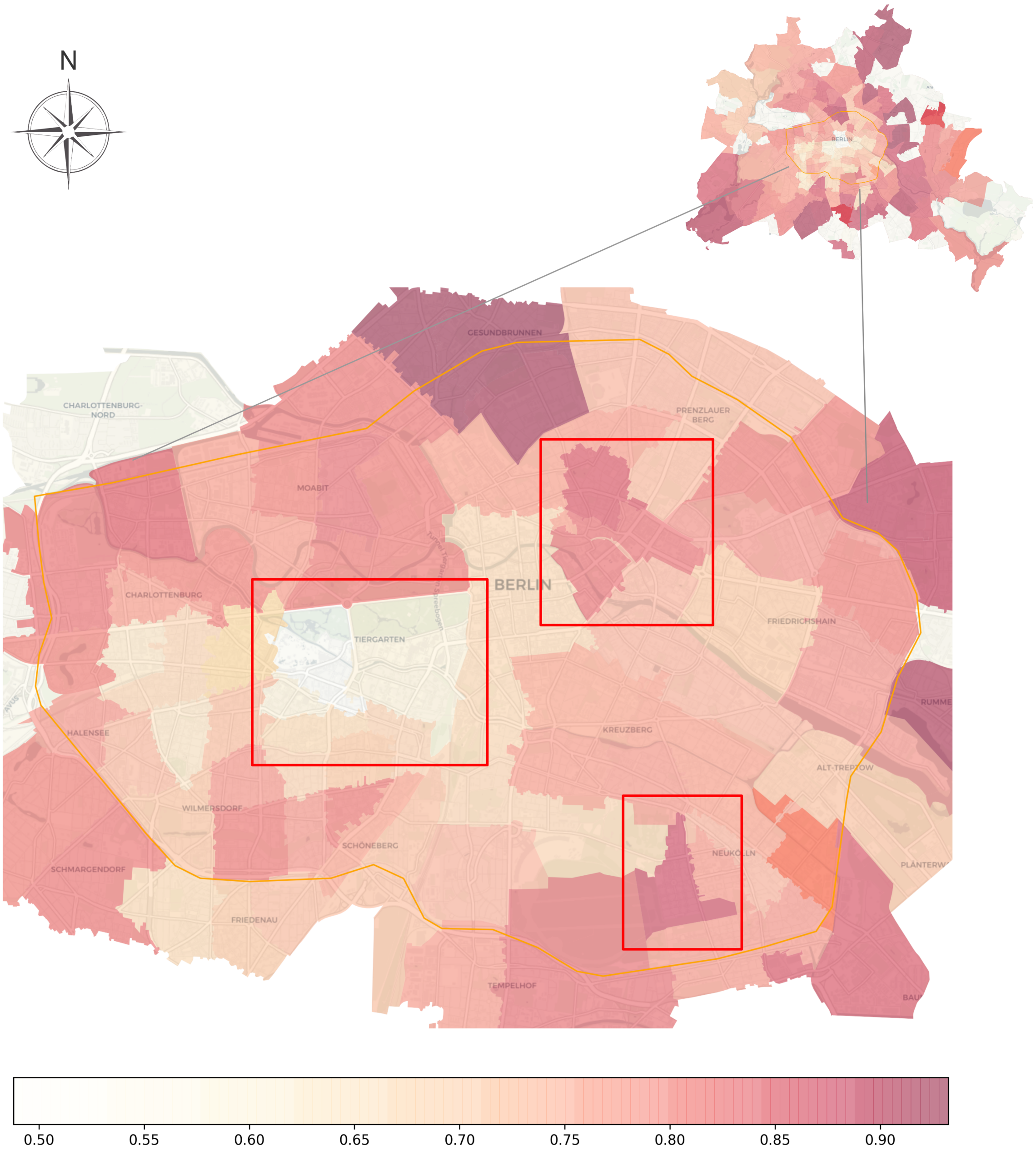

Figure 5.6: GWR Local R$^2$ Values

In addition to the comparison between the global regression and the GWR local average results, the GWR model is also run on a subset of covariates which does not include data from the social reviews or municipal datasets. Here, the selected bandwidth is 145 in contrast to the earlier 96. At this bandwidth, the model without the additional features produces an R$^2$ value of 0,62 vs. the complete model's 0,68. This illustrates a slight improvement from the addition of these novel features but also hints at the importance of the bandwidth in GWR modelling. At a wider bandwidth, the final model with all covariates included takes more neighbouring points into account and in doing so produces a lesser result. This is not always the case for every GWR model, though. For each unique problem and set of features, the bias / variance trade-off must be taken into account in which a bandwidth too large introduces too much variance and a bandwidth too small produces a model that over-fits. In the case of this thesis, the more covariates which are included in the regression, the more likely it is that a smaller bandwidth is necessary in order to capture the spatial relationships at play.

### 5.2.2 Visualisation of Effects

Beyond the higher-level, goodness-of-fit analysis of the GWR model, it is also important to observe the dynamics of specific covariates and their effects & significance values. In Table 5.2, these coefficients are laid out, representing the global regression. To consider local covariate effects, refer to the maps in Figures 5.7, 5.8, 5.9, 5.10, and 5.11. Note that the independent variables remain in their normalised form in addition to the fact that the target prices are scaled to obfuscate true values. Feature effects and their respective strengths in either direction are therefore best considered as relative to those of other localities rather than in terms of absolute value.

| Variable | Est. $\beta$ | Std. Error | P-Value |
|---:|:---:|:---:|:---:|
| Floor | -0,033 | 0,013 | 0,015 |
| Area | -0,715 | 0,022 | 0,000 |
| Rooms | 0,102 | 0,035 | 0,004 |
| Accommodates | 0,203 | 0,017 | 0,000 |
| Elevator | 0,236 | 0,014 | 0,000 |
| Nearest Grocery | 0,039 | 0,018 | 0,030 |
| Nearest Grundschule | 0,078 | 0,017 | 0,000 |
| Nearest Hospital | -0,052 | 0,018 | 0,005 |
| Nearest Kita | 0,031 | 0,015 | 0,047 |
| Nearest Oberschule | -0,071 | 0,016 | 0,000 |
| Recreation Access | -0,117 | 0,035 | 0,001 |
| Recreation Popularity | 0,226 | 0,030 | 0,000 |
| Green Space Popularity | 0,104 | 0,014 | 0,000 |
| Standing Water Access | -0,115 | 0,027 | 0,000 |
| Green Space Access | 0,034 | 0,013 | 0,011 |
| Altbau Ratio | -0,022 | 0,025 | 0,362 |
| Pre-War Ratio | -0,086 | 0,019 | 0,000 |
| Post-War Ratio | -0,059 | 0,022 | 0,006 |
| Nearest Transport | -0,038 | 0,024 | 0,116 |

Table 5.2: Feature Effects & Significance

Notice that all covariates are considered quite significant with the exception of the features *Altbau Ratio* & *Nearest Transport*. Considering that these are global coefficient estimates, these features might yet play a role in certain localities when plotted on a map. Indeed, when examined visually and spatially, the variable effect of each feature on the target becomes more easily interpretable. This visual analysis provides the necessary means by which to derive more nuanced insights from the GWR model, especially when paired with a confident understanding of the underlying geography and urban topography. In the following Figures (5.7, 5.8, 5.9, 5.10, and 5.11), a sample of the covariates are plotted by postcode, displaying the average effect of the feature in that postcode for instances in which the feature is found to be significant. These variables are *Recreation Popularity*, *Green Space Popularity*, *Altbau Ratio*, *Nearest Kita*, and *Nearest Transport*. This particular subset is selected to represent the various datasets introduced in Section 3.

Figure 5.7 below shows the local feature effects of the input *Recreation Popularity*. Interestingly, the strength of the effect is rather moderate in comparison to other features, with many postcodes showing lighter shades. In Table 5.2, however, this feature is very significant with a very large $\beta$ coefficient relative to other inputs. This map suggests an interesting interplay in which a given neighbourhood shows either a positive or negative relationship with having many popular restaurants, cafés, and bars in the area. For those familiar with Berlin, many of the more deeply shaded postcodes reflect popular sub-districts. What is curious here is the breakdown of those which trend blue vs. those which are more orange. One explanation for this behaviour can be that certain areas are considered more family-friendly or attractive for their more modest nightlife and recreational offerings. Other areas, on the other hand, are desired exactly for the fact that they are central hubs for these kinds of activities.

Of particular interest are the three contrasting sub-localities in Berlin-Charlottenburg, Berlin-Mitte & Prenzlauer-Berg, and in the district of Kreuzberg. Note the area around Ku'Damm, for example, which is a cluster of orange-shaded postcodes, while further northwest, postcodes near the Charlottenburg palace are more blue, indicating that being farther from the action is more desireable there. A similar dynamic can be seen between the Bergmannkiez neighbourhood in Kreuzberg and the neighbourhoods surrounding Oranienstraße and Maybachufer, which are well-known areas for long nights out. The same plays out as well in the area around Kollwitzplatz and Kastanienallee in contrast to surrounding postcodes. The cases on the edge of ring are curious as well; these may perhaps represent areas where the geographic border of the S-Bahn Ring is a delineator of more lively neighbourhoods within the Ring as opposed to quieter neighbourhoods outside of it. This would be a fitting explanation for the case in the far east of Berlin where the Ostkreuz transit hub is a noteworthy divider in the city's topography.

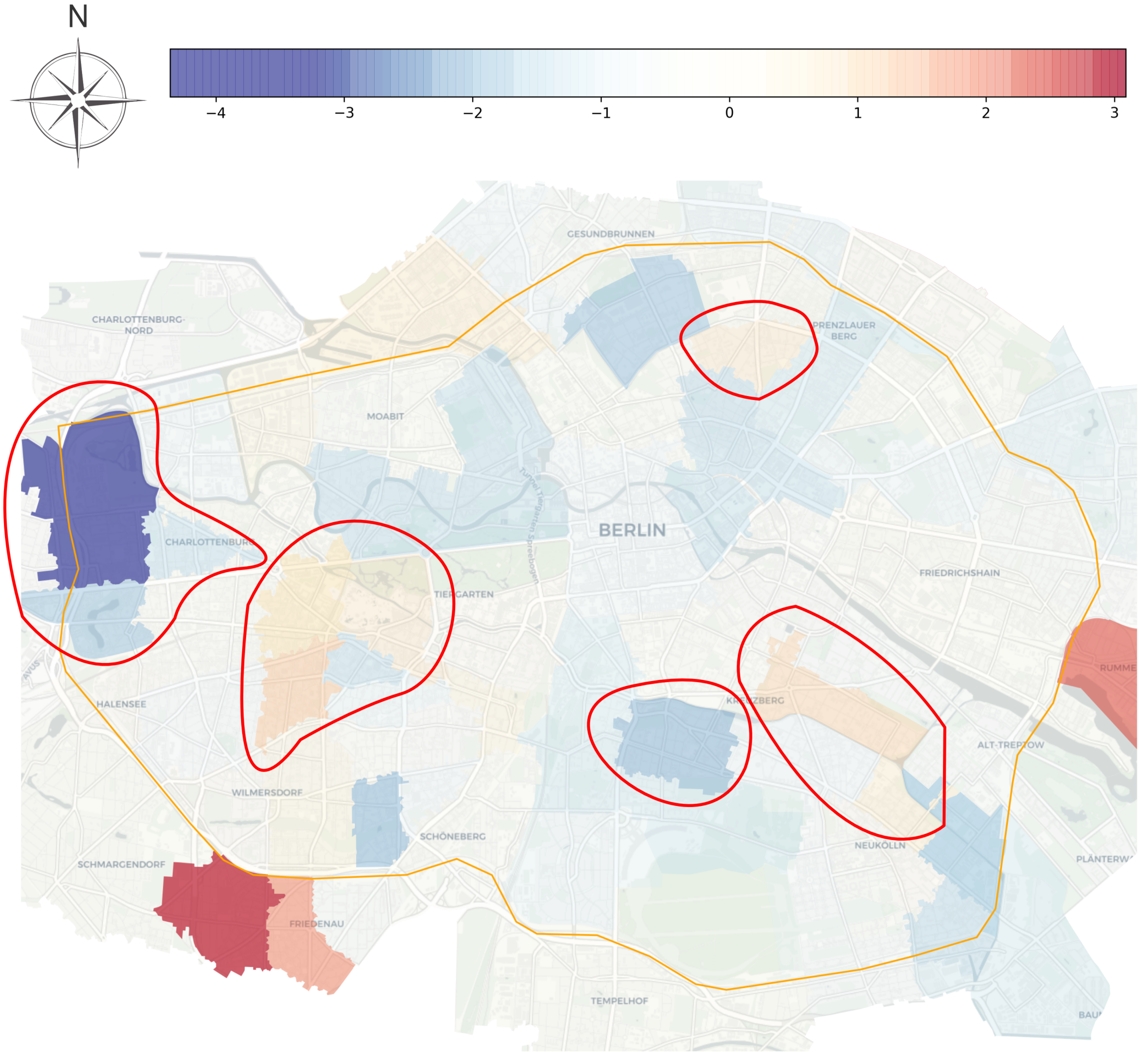

Figure 5.7: Recreation Popularity $\beta$

Figure 5.8 shows the effect of the feature *Green Space Popularity*. A reminder, this feature sums the counts of reviews for nearby parks and green spaces, similar to how the *Recreation Popularity* feature is constructed. A one-size-fits-all argument might assert that in densely populated corners of Berlin, green space is a more salient good in the context of housing value. This would apply, for example, in the Neukölln and Wedding districts in the north and south of the city, respectively. This also holds, albeit to a lesser degree, in the eastern half of the city in areas where the proportion of pre-war buildings is higher. The story in West Berlin is more complex but lends itself to some tailored explanations. The palace gardens and surrounding greenery in the northeast of

Charlottenburg, for example, appear to be a driver for higher prices in the immediate area. In the case of postcodes near Berlin's Tiergarten as well as the western bounds of the former Tempelhof Airfield (turned massive park), negative coefficients might be explained by the fact that actual access to neighbouring green spaces in the way of park entrances is limited. Communities (and housing value) in these areas may therefore develop farther away from neighbouring greenery.

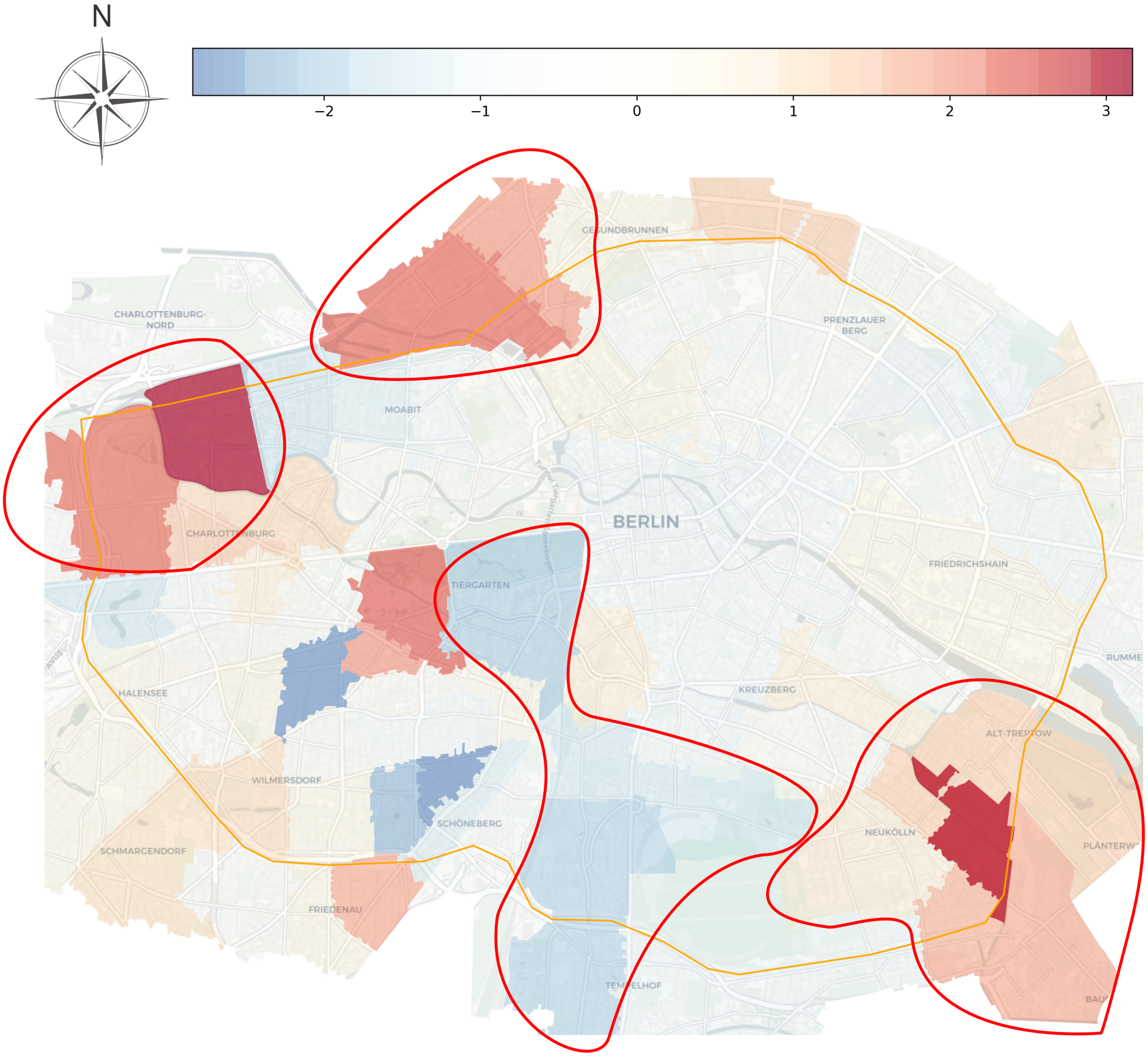

Figure 5.8: Greenspace Popularity $\beta$

Figure 5.9 displays the behaviour of the *Altbau Ratio* feature. While the GWR global model classified this feature as generally non-significant, the mapped results are rather encouraging and offer an interesting story on the role that old-build apartments play within the overall Berlin real estate environment. This visualisation shows that higher densities of Altbau apartments in a neighbourhood are correlated with higher property value, but only when these older buildings have been well renovated. Some of the most beautiful old houses in Berlin are in the western-most districts of Charlottenburg, Wilmersdorf, Schöneberg and western Kreuzberg. Refer again to Figure 3.6, in which these same areas are highlighted with high densities of Altbau houses. Note as well the warmer area in northern Berlin-Mitte in contrast to the central parts of the city where it might be the case that newer developments command higher prices rather than older apartments. In the case of Neukölln, this reasoning holds, as the edge-most postcodes of this district contain many old building which are yet to be properly renovated.

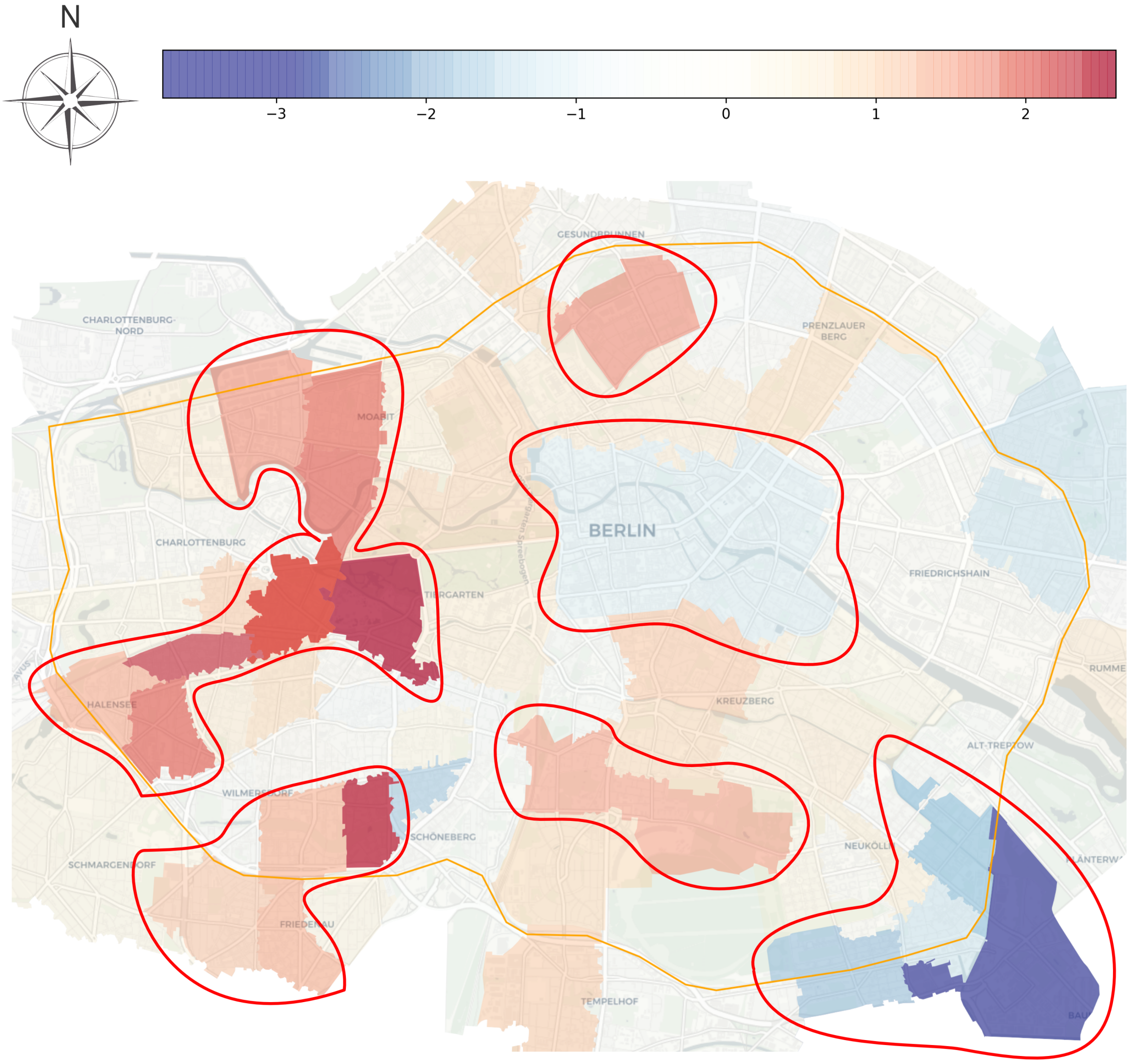

Figure 5.9: Altbau Ratio $\beta$

In Figure 5.10, the importance of proximity to kindergartens is examined. The results here indicate a fairly clear clustering of areas where such a proximity is either a good or a bad thing. The most likely hypothesis for this is that some parts of Berlin have older and more settled families with lower numbers of children, while others are host to younger families and families with many children. Take for example the clearly blue areas of central Charlottenburg and Schöneberg. These are two of the wealthier, more established communities in Berlin. It might make sense that in these areas, a kindergarten would be more a source of noise than a high-utility social good. Conversely, in northern Charlottenburg, northern Berlin-Mitte, and especially in Berlin-Neukölln, the effect is the opposite. In the current social topography of Berlin, these areas are experiencing growth in popularity amongst younger families and generally have higher-density populations where there are more children and therefore a greater need for kindergartens.

Figure 5.11 shows the final covariate effects map representing the spatial importance of nearness to public transport. In this case, all public transport lines in the city are considered equal, that is U-Bahn, S-Bahn, and Tram lines. As can be seen, most areas of the city show either slightly positive or slightly negative utility from proximity to public transport. This is likely due to correlations related to the orientation of the transport line to other, unrelated features in the area.

Take, for example, the areas which have much stronger signals. In Neukölln, the single postcode 12049 has a distinct disutility from being near to the nearest transport line, in this case, U-Bahn line U8. This is most likely due to the fact that, within the postcode, the further an apartment is from the U-Bahn line, the closer that apartment is to the largest public open space in the inner boundaries of the city. A similar effect can be seen in the right-most and north-most areas of the highlighted zones in Berlin-Charlottenburg. Here, proximity to transport means either being further from the district sub-centre or from other desirable natural amenities like the Landwehr Kanal. There is also the fact that certain parts of individual transport lines can be particularly undesirable in terms of hygiene or safety. This variability is especially non-stationary and can change from one station to the next. The red regions on display, on the other hand might be so because this area is a key transit, shopping, and commerce hub in the west of the city. Perhaps here, the utility of access to all of these things outweighs any related disutilities.

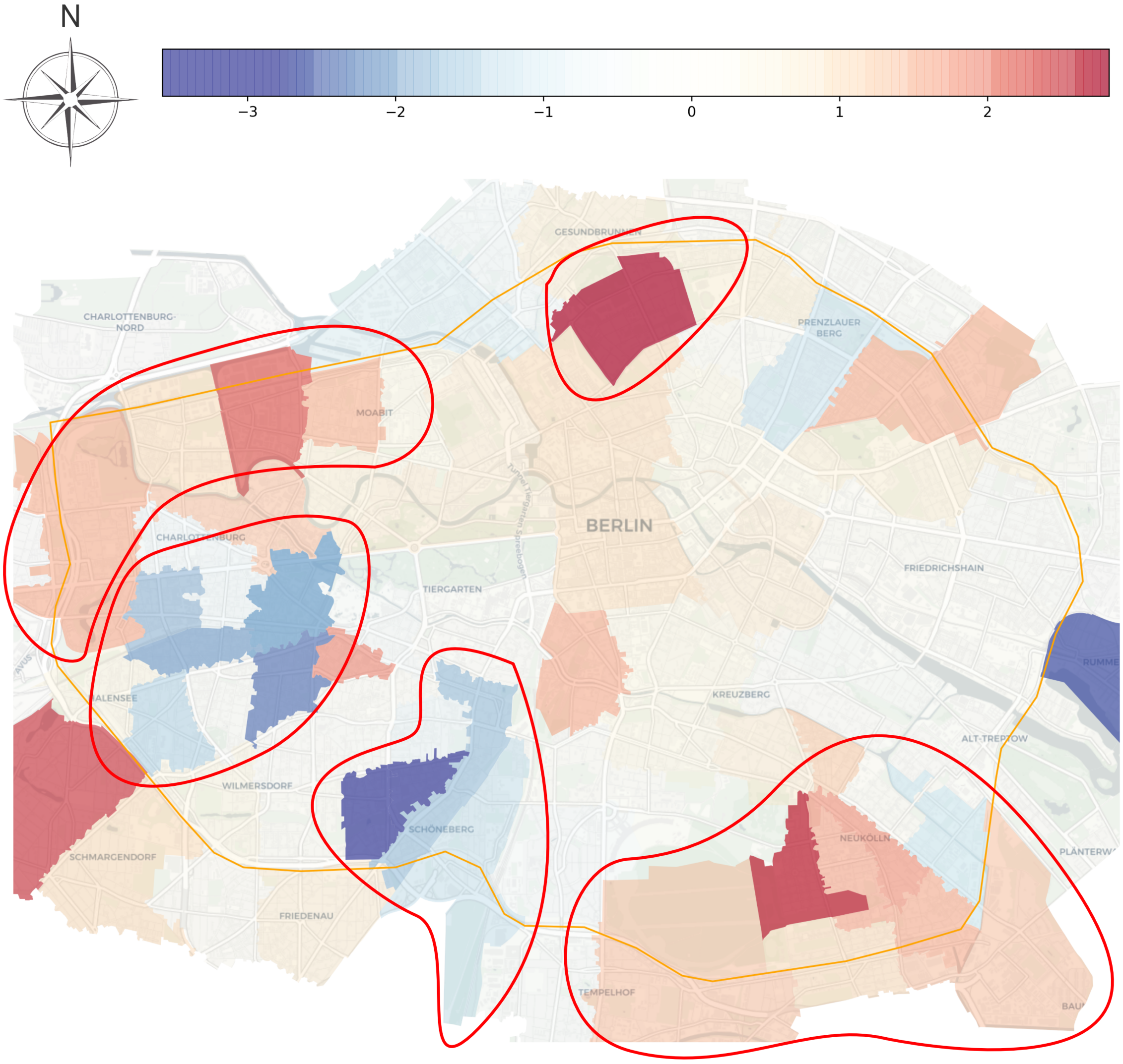

Figure 5.10: Nearest Kita $\beta$

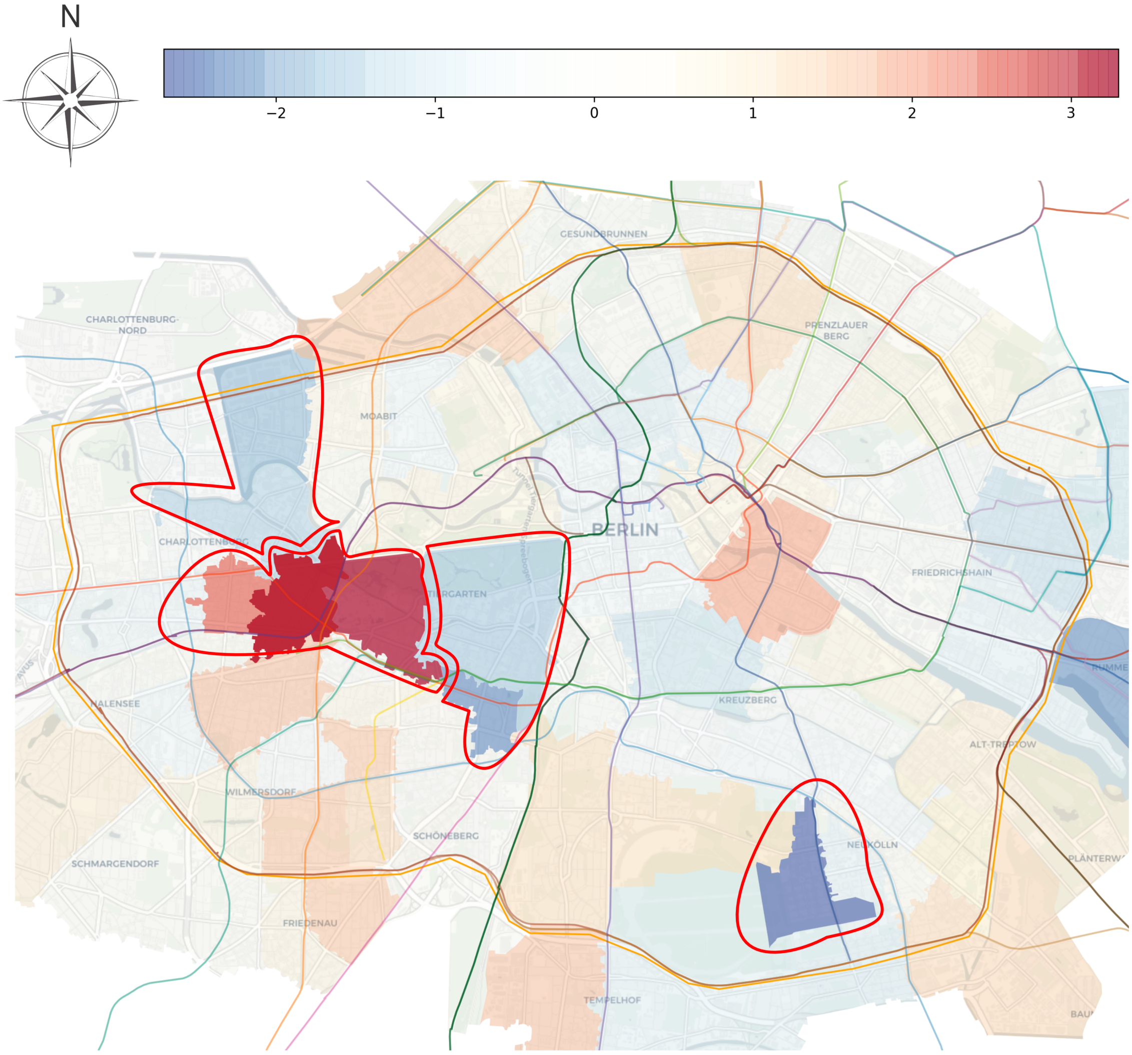

Figure 5.11: Nearest Transport $\beta$

# 6 Conclusion

Modelling real-estate value and, by extension, consumer preference is a challenging task. The variables that go into making up the value offering of an individual apartment are many, disparate, and run the gamut from the structural makeup of the property to the popularity of the neighbourhood it lies in. This thesis has built upon established methods by which to attempt such a modelling task with the additional value of novel data sources to help in modelling less-traditional features. The analysis takes place within the Berlin real estate market and is supported by big data / social media data from large aggregators (Google), along with municipal and state-affiliated statistical data (Berlin Senate & VBB). Through mapping of kernel density and Getis-Ord hotspot estimates, patterns in the distributions of certain attractions or points of interest (POIs) like nightlife or green space are visualised. Initial hypotheses and impressions formed from these exploratory methods along with individual priors are substantiated in the results of the geographically weighted regression, which accounts for spatially non-stationary effects. Here, given the significant increase in the $R^2$ statistic (0,43 to 0,74) from the global OLS regression to the geographically weighted regression, the reality of significant spatial variation in feature importance is made clear. Comparing a more bare-bones GWR using only structural features to the final GWR, which contains all features, the notable $R^2$ differential of 0,62 to 0,74 proves furthermore the value of the additional contextual features in the study. Further visual analyses by way of plotting local GWR covariate effects provide a glimpse into the depth of analysis made possible by the GWR methodology, given that the underlying data can support such fine-grain inspection.

This thesis outlines clear next steps in expanding the scope of research explored in this paper. A more expansive base dataset of properties and prices would provide more data points with which to capture latent trends in specific localities. Of further interest would be observing the change in spatial price dynamics over time in conjunction with the evolution of data involving restaurants, bars, and cafés. Considering that these businesses are rather mobile and appear and disappear more quickly than, say, public transport lines, there is an opportunity to track the growth and decline of neighbourhoods as their demographics change and fluctuate. This is particularly interesting, considering the movement of Berlin residents within the city as well as taking into account the inflow and outflow of Berliners from within Germany and abroad. For now, however, this thesis succeeds in providing a strong case for the value in including public and social reviews data to provide additional clarity in understanding spatially diverse real estate markets.

# Bibliography

Adair, A. S., J. N. Berry, and W. S. McGreal (1996). Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research 13*, 67–83.

Banerjee, S., S. Bhattacharyya, and I. Bose (2017). Whose online reviews to trust? understanding reviewer trustworthiness and its impact on business. *Decission Support Systems 96*, 17–26.

Barreca, A., R. Curto, and D. Rolando (2020). Urban vibrancy: An emerging factor that spatially influences the real estate market. *Sustainability 12*(1), 346.

Bilogur, A. (2016-2020). mwaskom/seaborn: v0.8.1 (september 2017).

Bozdogan, H. (1987, 02). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika 52*, 345–370.

Broadbent, G. (1990). *Emerging Concepts in Urban Space Design*. https://books.google.de/books/about/Emerging_Concepts_in_Urban_Space_Design.html [Online; accessed 2020-10-02].

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis 28*(4), 281–298.

Butler, H., M. Daly, A. Doyle, S. Gillies, T. Schaub, and T. Schaub (2016, August). The GeoJSON Format. RFC 7946.

Cellmer, R., A. Cichulska, and M. Belej (2020). Spatial analysis of housing prices and market activity with the geographically weighted regression. *International Journal of Geo-Information 9*.

Chen, T., E. C. Hui, J. Wue, W. Lang, and X. Li (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons 60*(3), 293–303.

Chen, T., E. C. Hui, J. Wue, W. Lang, and X. Li (2019). Identifying urban spatial structure and urban vibrancy in highly dense cities using georeferenced social media data. *Habitat International 89*.

Cholodilin, K. A. and C. Michelsen (2019). High risk of a housing bubble in germany and most oecd countries. *DIW Weekly Report 9*(32), 265–273.

Clark, S. and N. Lomax (2020). Rent/price ratio for english housing sub[U+2010]markets using matched sales and rental data. *Royal Geographical Society: Area 52*(1).

Colabianchi, N., M. Dowda, K. A. Pfeiffer, M. J. C. A. Dwayne E Porte and, and R. R. Pate (2007). Towards an understanding of salient neighborhood boundaries: Adolescent reports of an easy walking distance and convenient driving distance. *International Journal of Behavioral Nutrition and Physical Activity 4*(66), 5–20.

Daams, M. N., F. J.Sijtsma, and P. Veneri (2019). Mixed monetary and non-monetary valuation of attractive urban green space: A case study using amsterdam house prices. *Ecological Economics 166*.

Daniels, R. and C. Mulley (2013). Explaining walking distance to public transport: The dominance of public transport supply. *Journal of Transport and Land Use 6*(2), 5–20.

Development, S. D. U. and Housing (2020). Fis broker: Maps, data, services - online. https://www.stadtentwicklung.berlin.de/geoinformation/fis-broker/index_en.shtml. [Online; accessed 2020-10-02].

Dolls, M., C. Fuest, C. Krolage, F. Neumeier, and D. Stöhlker (2020). Ökonomische effekte des berliner mietendeckels. *IFO Schnelldienst 73*(3), 33–38.

Dou, X., J. A. Walden, S. Lee, and J. YoungLee (2012). Does source matter? examining source effects in online product reviews. *Computers in Human Behavior 28*(5), 1555–1563.

DW (2017). Berlin property price growth tops global list. https://www.dw.com/en/berlin-property-price-growth-tops-global-list-knight-frank-report/a-43351744. [Online; accessed 2020-10-04].

Emenlauer, D. R., V. Kantowski, and I. Solovyeva (2018). Hochhausentwicklungsplan – hochhausleitbild für berlin. https://www.parlament-berlin.de/adosservice/18/Haupt/vorgang/h18-1597.A-v.pdf. [Online; accessed 2020-10-18].

Fotheringham, A. S., W. Yang, and W. Kang (2017). Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers 107*(6), 1247–1265.

Ghania, N. A., S. Hamid, I. A. T. Hashem, and E. Ahmed (2019). Social media big data analytics: A survey. *Computers in Human Behavior 101*, 417–428.

Gillies, S. et al. (2007-2020). Shapely: manipulation and analysis of geometric objects.

Hannum, C., K. Y. Arslanli, and A. F. Kalay (2019). Spatial analysis of twitter sentiment and district-level housing prices. *Journal of European Real Estate Research 12*(2).

Hiebert, J. and K. Allen (2019). Valuing environmental amenities across space: A geographically weighted regression of housing preferences in greenville county, sc. *Land 8*, 1–16.

Hua, L., S. He, Z. Han, H. Xiao, S. Su, M. Weng, and Z. Cai (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy 82*, 657–673.

Icon, C. W. O., F. Ren, W. Hu, and Q. Du (2019). Multiscale geographically and temporally weighted regression: Exploring the spatiotemporal determinants of housing prices. *International Journal of Geographical Information Science 3*(3), 489–511.

Irwin, E. G. (2002). The effects of open space on residential property values. *Land Economics 78*(4), 465–480.

Jordahl, K. (2014). Geopandas: Python tools for geographic data. *https://github.com/geopandas/geopandas*.

Kain, J. F. and J. M. Quigley (2012). Measuring the value of housing quality. *Journal of Big Data 65*, 532–548.

Kalinic, M. and J. M. Krisp (2018). Kernel density estimation (kde) vs. hot-spot analysis - detecting criminal hot spots in the city of san francisco. *Conference on Geo-information Science 21*.

Kamalov, F. (2020). Kernel density estimation based sampling for imbalanced class distribution. *Information Sciences 512*, 1192–1201.

Kim, H.-S., G.-E. Lee, J.-S. Lee, and Y. Choi (2019). Understanding the local impact of urban park plans and park typology on housing price: A case study of the busan metropolitan region, korea. *Landscape and Urban Planning 184*, 1–11.

Koupaei, J. A., S. Hosseini, and F. M. Ghaini (2016). A new optimization algorithm based on chaotic maps and golden section search method. *Engineering Applications of Artificial Intelligence 50*, 201–214.

Lawhead, J. (2013). *Learning Geospatial Analysis with Python*. Packt Publishing.

Levi, J. (2019a). Learn about geographically weighted models in python using airbnb data in berlin residential districts (2018).

Levi, J. (2019b). Learn about local getis-ord gi in python using airbnb data in berlin residential districts (2018).

Long, R. (2020). Impacts of shopping malls on the housing price - evidence from stockholm. Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden.

Lütkepohl, H. and F. Xu (2012). The role of the log transformation in forecasting economic variables. *Empirical Economics 42*, 619–638.

Marti, P., L. Serrano-Estrada, and A. Nolasco-Cirugeda (2019). Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems 74*, 161–174.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Volume 445, pp. 51–56. Austin, TX.

Mossay, P., J. K. Shin, and G. Smrkolj (2020). Quality differentiation  spatial clustering among restaurants.

OECD, Eurostat, I. L. Organization, I. M. Fund, T. W. Bank, and U. N. E. C. for Europe (2013). *Handbook on Residential Property Price Indices*. https://www.oecd-ilibrary.org/content/publication/9789264197183-en.

Ord, K. and A. Getis (2010, 09). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis 27*, 286 – 306.

Owuor, I. and H. H. Hochmair (2020). An overview of social media apps and their potential role in geospatial research. *International Journal of Geo-Information 9*.

Pridal, P., T. Pohanka, and R. Kacer (2004). https://epsg.io/3068. [Online; accessed 2020-10-03].

Qi, L., J. Li, Y. Wang, and X. Gao (2019). Urban observation: Integration of remote sensing and social media data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12*(11).

Rey, S. J. and L. Anselin (2007). PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies 37*(1), 5–27.

Ridker, R. G. and J. A. Henning (1967). The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics 49*(2), 246–257.

Rivas, R., D. Patil, V. Hristidis, J. R. Barr, and N. Srinivasan (2019). The impact of colleges and hospitals to local real estate markets. *Journal of Big Data 6*(7).

Seo, S. (2006). A review and comparison of methods for detecting outliers in univariate data sets. Master's thesis, University of Pittsburgh, Pittsburgh, Pennsylvania.

Skupin, A. (2000). From metaphor to method: cartographic perspectives on information visualization. *IEEE Symposium on Information Visualization (INFOVIS)*.

So, H., R. Tse, and S. Ganesan (1997). Estimating the influence of transport on house prices: Evidence from hong kong. *Journal of Property Valuation and Investment 15*(1), 40–47.

Tamara Sliskovic, J. T. (2019). The importance of distance in hedonic housing price model – the case of zagreb. *Ekonomski Pregled 70*(5).

Tang, L. (2017). Mine your customers or mine your business: The moderating role of culture in online word-of-mouth reviews. *Journal of International Marketing 5*(2).

Thériault, M., F. Rosiers, P. Villeneuve, and Y. Kestens (2003, 03). Modelling interactions of location with specific value of housing attributes. *Property Management 21*, 25–62.

Timcke, M.-L., A. Pätzold, D. Wendler, and C. Möller (2018). Gebäudealter alt- oder neubau? so wohnt berlin. https://interaktiv.morgenpost.de/so-alt-wohnt-berlin. [Online; accessed 2020-10-06].

Tomal, M. (2020). Modelling housing rents using spatial autoregressive geographically weighted regression: A case study in cracow, poland. *International Journal of Geo-Information 9*.

VBB (2020). Datensätze (gtfs, haltestellen, linienfarben). https://www.vbb.de/unsere-themen/vbbdigital/api-entwicklerinfos/datensaetze. [Online; accessed 2020-10-03].

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, l. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors (2020). Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*.

Waskom, M., O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, Brian, C. Fonnesbeck, A. Lee, and A. Qalieh (2017). mwaskom/seaborn: v0.8.1 (september 2017).

Wen, H., Y. Zhang, and L. Zhang (2015). Assessing amenity effects of urban landscapes on housing price in hangzhou, china. *Urban Forestry  Urban Greening 14*(4), 1017–1026.

Wilhelm, T. and P. Hänggi (2003, 11). Power-law distributions resulting from finite resources. *Physica A: Statistical Mechanics and its Applications 329*, 499–508.

Wu, C., X. Ye, F. Ren, Y. Wan, P. Ning, and Q. Du (2016). Spatial and social media data analytics of housing prices in shenzhen, china. *PLoS ONE 11*(10).

Xie, Z. and J. Yan (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems 32*(5), 396–406.

Zhang, L., J. Zhou, and E. C. man Hui (2020). Which types of shopping malls affect housing prices? from the perspective of spatial accessibility. *Habitat International 96*.

Zhang, S., L. Wang, and F. Lu (2019). Exploring housing rent by mixed geographically weighted regression: A case study in nanjing. *International Journal of Geo-Information 8*.

Zhang, X., Y. Zheng, L. Sun, and Q. Dai (2019). Estimating the influence of transport on house prices: Evidence from hong kong. *Sustainability 11*.

Zhang, Z. (2019). The price premium for properties near shopping centers: Evidence from china. In *Proceedings of the 4th International Conference on Humanities Science, Management and Education Technology (HSMET 2019)*, pp. 393–397. Atlantis Press.

Zilisteanu, I. R., R. Muntean, S. C. Gherghina, G. Vintila, and T. C. Barbu (2019). Towards a hedonic pricing method for the bucharest private housing market. *Scientific Annals Of Economics And Business 66*(3).

Zurbarán, M., P. Wightman, M. Brovelli, D. O. M. Iliffe, M. Jimeno, and A. Salazar (2018). Nrand[U+2010]k: Minimizing the impact of location obfuscation in spatial analysis. *Transactions in GIS 22*(5), 1257–1274.

# Declaration of Authorship

I hereby confirm that I, Alex Truesdale, have authored this master thesis alone, independently and without outside assistance other than those indicated in the list of references. Where I have consulted the published work of others in any form (i.e. ideas, equations, figures, text, tables), explicit attribution to the source can be found.

November 11, 2020<br>
Berlin, Germany<br>
**Alex Truesdale**

---

Hiermit erkläre ich, Alex Truesdale, dass ich die vorliegende Masterarbeit allein und nur mithilfe der zitierten Quellen, die in der Referenzliste stehen, angefertigt habe. Die Prüfungsordnung ist mir bewusst. Ich habe dementsprechend weder eine bisherige Masterarbeit in meinem Studienfach eingereicht noch habe ich zuvor eine Prüfung einer Masterarbeit nicht bestanden.

November 11, 2020<br>
Berlin, Germany<br>
**Alex Truesdale**