The Impact of Online Real Estate Listing Data on the Transparency of the Real Estate Market

Using the Example of Vacancy Rates

Vom Fachbereich Bau- und Umweltingenieurwissenschaften der
Technischen Universität Darmstadt
zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)
genehmigte

Dissertation

vorgelegt von

Philip Detlev Gärtner (M.Sc.)

Erstreferent:	Prof. Dr.-Ing. Hans-Joachim Linke
Korreferentin:	Prof. Dr.-Ing. Alexandra Weitkamp

Tag der Einreichung:	20.06.2023
Tag der mündlichen Prüfung:	15.09.2023

Darmstadt 2023

Abstract

Despite the increasing digitization of the real estate market and the accompanying greater availability of data, as evidenced, for example, by the proliferation of online real estate listing platforms, there are still deficiencies in market transparency associated with a variety of negative aspects. This study aimed to investigate the impact of online real estate listing data on market transparency by examining the suitability of these data for scientific use in general and for the example of estimating vacancy rates in particular. Therefore, a comprehensive data set consisting of more than seven million listings was collected over one and a half years and analyzed with regard to all available features in terms of their quality and quantity. Furthermore, their explanatory power for estimating vacancy rates was tested by their application in different regression models.

The features specified in online real estate listings showed an average completeness of 85.97 % and, most widely, plausible feature specifications. Exceptions were information regarding energy demand, which were only available in 20.79 % of listings, and the specification of the building quality and condition, which showed indications of being positively biased. The estimation of vacancy rates on the district level, based on online real estate listing data, showed promising results, being able to explain vacancy rates with a goodness of fit of a pseudo R^2 of 0.81 and a mean absolute error of 0.84 percentage points. These results suggest that information contained in online real estate listing data are a good basis for scientific evaluation and are specifically well suited for estimating vacancy rates. The findings imply the utilization of online real estate listing data for a diverse range of purposes, extending beyond the current focus of price-related research.

Keywords: online real estate listing data, real estate market transparency, vacancy, linear regression, spatial regression

Kurzzusammenfassung

Trotz der zunehmenden Digitalisierung des Immobilienmarktes und der damit einhergehenden besseren Datenverfügbarkeit, die sich beispielsweise durch die Verbreitung von Onlineimmobilienplattformen zeigt, gibt es noch immer Defizite der Markttransparenz, die mit einer Vielzahl von negativen Aspekten verbunden sind. Ziel dieser Studie war es, den Einfluss von Onlineimmobilienangebotsdaten auf die Markttransparenz zu untersuchen, indem die Eignung dieser Daten für die wissenschaftliche Nutzung im Allgemeinen und für das Beispiel der Schätzung von Leerstandsquoten im Besonderen geprüft wurde. Dazu wurde ein umfangreicher Datensatz von mehr als sieben Millionen Inseraten über einen Zeitraum von eineinhalb Jahren erhoben und hinsichtlich aller verfügbaren Merkmale qualitativ und quantitativ analysiert. Darüber hinaus wurde ihre Erklärungskraft für die Schätzung von Leerstandsquoten durch ihre Anwendung in verschiedenen Regressionsmodellen getestet.

Die in Onlineimmobilienangeboten spezifizierten Merkmale wiesen eine durchschnittliche Vollständigkeit von 85,97 % und weitestgehend plausible Merkmalsangaben auf. Ausnahmen waren die Angaben zum Energiebedarf, die nur in 20,79 % der Inserate vorhanden waren, und die Angaben zu Gebäudequalität und -zustand, die Hinweise auf eine positive Verzerrung zeigten. Die Schätzung der Leerstandsquoten auf Kreisebene, auf Grundlage von Onlineimmobilienangebotsdaten, zeigte vielversprechende Ergebnisse, gemessen an einer Anpassungsgüte der Leerstandsquotenschätzung von einem Pseudo-R^2 von 0,81 und einem mittleren absoluten Fehler von 0,84 Prozentpunkten. Diese Ergebnisse deuten darauf hin, dass die in den Onlineimmobilienangebotsdaten enthaltenen Informationen eine gute Grundlage für wissenschaftliche Auswertungen darstellen und sich insbesondere gut für die Schätzung von Leerstandsquoten eignen. Die Ergebnisse implizieren die Nutzung von Onlineimmobilienangebotsdaten für eine Vielzahl von Anwendungsfällen, die über den derzeitigen Fokus der preisbezogenen Forschung hinausgehen.

Stichworte: Onlineimmobilienangebotsdaten, Immobilienmarkttransparenz, Leerstand, lineare Regression, räumliche Regression

Acknowledgments

The title page of this dissertation lists a single author. However, this work would not have been possible without the support of many others. For this I would like to express my deepest gratitude.

Special thanks go to Prof. Dr. Hans-Joachim Linke. He supported me unconditionally during the whole process of writing this dissertation, gave me all the freedom to develop the work and was the best supervisor one could wish for, especially on a personal level. I would also like to thank Prof. Dr. Alexandra Weitkamp for her extensive support. The regular meetings to discuss the topic and her comments were particularly encouraging, provided valuable input and contributed to the improvement of the work.

My time at TU Darmstadt would not have been such an unforgettable experience without all my colleagues at the Department of Landmanagement. You all contributed to the success of this dissertation not only with your valuable insights during the doctoral seminars or coffee breaks, but you also made the time fly by, for which I would especially like to thank you.

Markus and Bingxiang, you welcomed me warmly in your office and made it easy for me to settle into the department. Kim, Martina, Max and Damian, I would love to share the floor and especially the coffee with you again. Raphael, Jana, Luisa, and Christian, even though I was not able to enjoy our view on your floor, I was always happy to come to you because I knew that not every long night in the library meant a lot of reading. I would also like to thank all the other members of our department and all the student assistants, especially Cedric, I had the pleasure to work with. Without your extraordinary support, many more working days would have turned into nights.

Finally, I would like to thank the most important people in my life. Thank you, Mom, Dad and Luise, for always being there for me, no matter what. Diana, I would have to fill the pages of this work again if I wanted to write down what I am grateful to you for.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACS	American Community Survey
AGS	Amtlicher Gemeindeschlüssel (Official Municipality Key)
AHS	American Housing Survey
AIC	Akaike information criterion
AK VGR	Arbeitskreis Volkswirtschaftliche Gesamtrechnungen der Länder (Work Group National Account of the Federal States)
ANOVA	Analysis of variance
API	Application programming interface
BauGB	Baugesetzbuch (German Building Code)
BBSR	Bundesinstitut für Bau-, Stadt- und Raumforschung (German Federal Office for Building and Regional Planning)
BDEW	Bundesverband der Energie- und Wasserwirtschaft e.V. (German Association of the Energy and Water Industry)
BDSG	Bundesdatenschutzgesetz (German Federal Data Protection Act)
BetrKV	Betriebskostenverordnung (German Utility Costs Regulation)
BGB	Bürgerliches Gesetzbuch (German Civil Code)
BKG	Bundesamt für Kartographie und Geodäsie (German Federal Agency for Cartography and Geodesy)
BP	Breusch-Pagan
c.p.	ceteris paribus (all other things being equal)
CLT	Central limit theorem
CPS	Current Population Survey
DCF	Discounted cash flow
e.g.	exempli gratia (for example)
FDZ	Forschungsdatenzentrum (German Research Data Centre)
GDP	Gross domestic product
GEG	Gebäudeenergiegesetz (German Buildings Energy Act)
HVS	Housing Vacancy Survey
i.e.	id est (that is)
JLL	Jones Lang LaSalle
kWh	Kilowatt-hour
LM	Lagrange multiplier
MAE	Mean absolute error
ML	Maximum likelihood
MLR	Multiple linear regression

NLTK	Natural Language Tool Kit
OLS	Ordinary least square
ORL	Online real estate listing
PCA	Principal component analysis
RMSE	Root mean square error
RQ	Research question
RWI	Leibniz-Institut für Wirtschaftsforschung
SEM	Spatial error model
SLM	Spatial lag model
SLR	Spatial lag regression
sqm	Square meter
U.K.	United Kingdom
U.S.	United States of America
UrhG	Urheberrechtsgesetz (German Act on Copyright and Related Rights)
USPS	United States Postal Service
VIF	Variance inflation factor
WoFlV	Wohnflächenverordnung (German Living Space Ordinance)
ZHVI	Zillow Home Value Index
ZRI	Zillow Rent Index
ZTRAX	Zillow Transaction and Assessment Data Set

Variable Abbreviations

city	Large City Not Attached to an Administrative District
cr	Cold Rent per SQM
gdp	GDP per Capita
ls	Living Space
nnl	Normalized Number of Listings
pop	Population Change
rural	Rural District With Beginning Concentration Processes
sparse	Sparsely Populated Rural District
st	Settlement Type
urban	Urban District
vac	Vacancy Rate

1 Introduction

Real estate, consciously or unconsciously, plays a central role in the lives of almost all of us. At work and for leisure, we spend a large part of our time in real estate, real estate is an important component of private as well as macroeconomic wealth, and many of us identify with our home. This multifaceted importance contributes to the real estate market being a regularly discussed topic. Despite differing knowledge in the overall population, conversations about the level of rents or purchase prices, the location of properties, or the real estate market in general are familiar to many of us.

However, the importance of real estate markets becomes particularly apparent when parts of the market malfunction and express themselves in phenomena such as housing shortages, ghost towns, or housing price bubbles. These malfunctions remind us that our understanding and management of real estate markets is still far from perfect and that improvements to increase market transparency are worth striving for. Particularly criticized for years by scientists and practitioners is the availability of relevant and recent real estate market data of good quality (Cellmer & Szczepankowska, 2014, p. 2; L. Li & Wan, 2021; Lizieri, 2003, p. 1151). This lack of data is especially astonishing in current times, as more and more data is being collected and stored through digitalization.

1.1 Motivation and Research Gap

The digitalization process has also found its way into the real estate markets, in an eminent manner, through online real estate listing platforms. These platforms collect large amounts of real estate data and could provide an opportunity to solve the data availability problem. Some of these platforms, such as Zillow for the market in the United States of America (U.S.) or Immobilien Scout for the German market, provide or have previously provided data for specific scientific use cases.

Thus, the research field utilizing these data has been growing considerably in recent years and research profited not only from the better availability of the data compared to the previous data provision, e.g., through expert committees or private data collections, but also from the specific characteristics of online real estate listing (ORL) data. These characteristics include the broad spatial market coverage, the additional information compared to transaction data, and the existence of real-time information, if collected from the website itself. Furthermore, the market

for online real estate platforms is not very fragmented, but often develops clear market leaders due to economies of scale, which reduces self-selection processes and resulting bias problems of the listings. Although current market shares for Germany are not published, the 2016 opinion of the German Federal Cartel Office (Bundeskartellamt, 2016, p. 80) can be used as a reference point, assuming a market share of more than 70 % for Immobilien Scout, which has probably increased since then. In combination with the evidence that specific information contained in ORL data reflect the true characteristic with sufficient accuracy, e.g., the asking prices approximate the transaction prices (Lyons, 2019; Thomschke, 2015, p. 92), the suitability of the data for scientific purposes can be assumed in principle.

However, what is missing is a comprehensive examination of ORL data, considering the variety of different included characteristics and their quality and quantity as a comprehensive basis for future research. Thus, this research aims to compile a representative data set from which a general ORL data assessment can be derived and upon which future research can build. This approach can also be seen as an extension of the existing literature regarding real estate market transparency since, up to now, transparency research typically examines topics such as comparisons between countries (Bienert, 2017, pp. 47–50), the general relevance of real estate market transparency (Tormanski, 2012, p. 150), or the consequences of increasing transparency (Gholipour et al., 2020, pp. 1–2) and regularly refers to the Global Real Estate Transparency Index published by the company Jones Lang LaSalle (JLL). However, what is not examined is the influence of ORL data on real estate market transparency, which is done by this study. To not only assess but also test the practical applicability of the data, a concrete example is chosen. For this example, it is tested whether the application of ORL data could lead to an increase in transparency.

Recent research making use of ORL data, to a large extent, focuses on the examination of prices and price-related topics and includes, among many others, the work of Gupta et al. (2022) examining pandemic-induced price effects, the study of Taruttis & Weber (2022) analyzing the impact of energy efficiency on single-family home prices, and the work of Eilers et al. (2021) focusing on differences in rent prices for migrants and natives. Due to the centrality of the asking rent or asking price in the ORL data, these are obvious use cases. However, due to the large amount of information contained in ORL data, a variety of other applications are conceivable for which research is missing. Especially information for which the existing supply is weak and the acquisition of which is usually associated with high effort seem to be beneficial choices.

These aspects particularly apply to the case of large-scale vacancy data. High-quality surveys that are freely available take place throughout Germany only every ten years as part of the decennial census. The only similarly comprehensive alternative is the CBRE-empirica-vacancy-index, calculated annually but not freely available to the general public and not stating the exact underlying methodology. For a variety of reasons, this is especially unfortunate for vacancy data, e.g., as it complicates the political assessment of vacancy and countermeasures in general and vacancy is related to various severe consequences, including the loss in property values, the increase in the risks for fire and crime rates, and the reduction of the visual attractivity of vacancy influenced areas (BBSR, 2017, p. 3; Franz, 2001, p. 263; Konomi et al., 2017, p. 130; Manville & Kuhlmann, 2018, p. 471; K. Wang & Immergluck, 2019, p. 513).

Despite these incentives for the provision of more and better vacancy data, the availability was already criticized decades ago (Franz, 2001, p. 263) and is still insufficient, for some authors even to the extent that they argue that vacancy may be the least understood characteristic of housing markets (BBSR, 2017, p. 21). Reasons for the lack of these data could, for example, include the great effort or missing raw data, which are both required by most methods to estimate vacancy. Thus, this thesis attempts to address two issues simultaneously. First, to investigate the suitability of ORL data for scientific analysis, and second, to improve the availability of large-scale vacancy data, precisely nationwide vacancy rates. Both aspects contribute to assessing the impact of ORL data on real estate market transparency. Finally, the methodological approach applied to improve vacancy data availability could also be transferred to other similar problems.

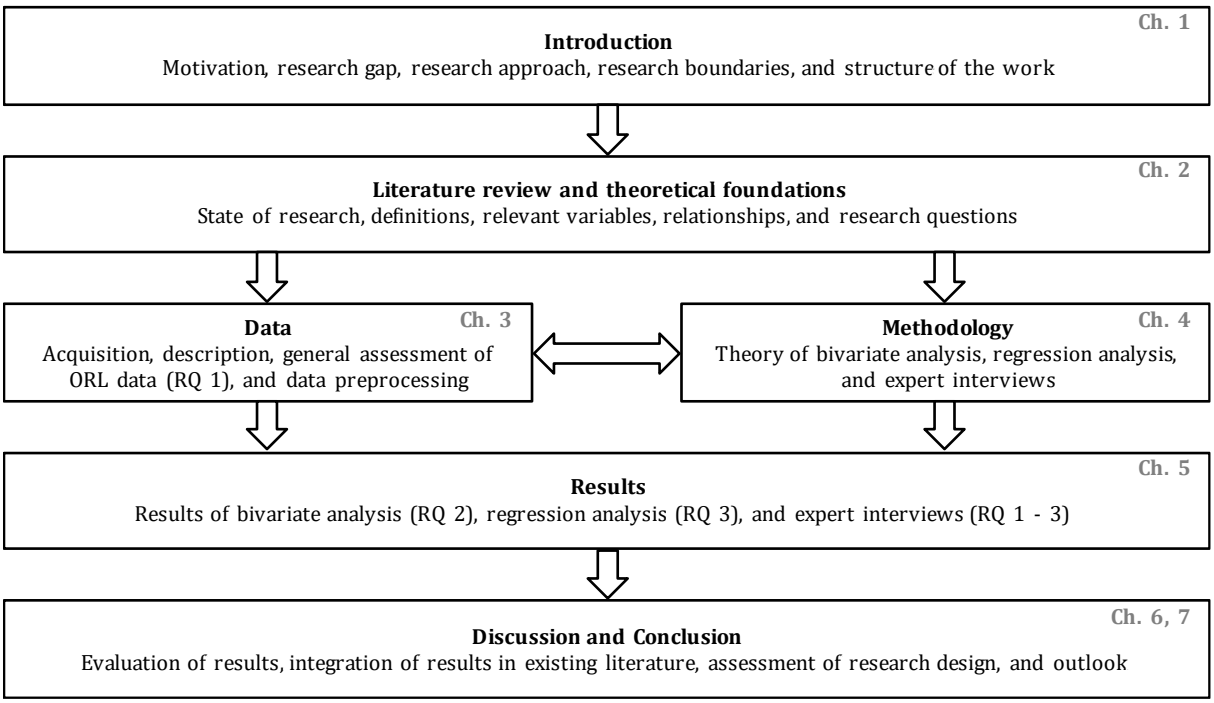
1.2 Research Approach and Structure of the Work

The choice of the research approach is crucial for the analyses conducted and the results that can be derived. It should be guided by the research questions and the information that can be obtained to answer them. The choice of the research approach for this study integrates theories of the widely applied work of Creswell & Creswell (2018), who differentiate between qualitative, quantitative, and mixed method approaches. Of these approaches, they recommend the qualitative approaches for research in a completely new area for which there is no or little prior work and where the crucial influences are unknown. In contrast, quantitative approaches are particularly well suited when presumed relationships of known variables are to be tested (Creswell & Creswell, 2018, pp. 40–57).

To contribute to closing the previously worked-out research gap, the massive amounts of data, online real estate platforms provide, can be used for analysis. Additionally, the closely related fields of vacancy research and research using ORL data provide an extensive knowledge base on which this research can be built and from which fundamental relationships between vacancy and real estate markets can be derived that are already well understood. Thus, a mainly quantitative research approach is suitable and chosen, which furthermore allows the derivation of objective results based on statistical tests. However, the more concrete research design is not often purely quantitative or qualitative and regularly exhibits elements of both (Creswell & Creswell, 2018, p. 41). Such elements can also be found in this study, e.g., in the application of a literature review to identify relevant relationships and in expert interviews for validation purposes. The following Figure 1 serves as an overview of the structure of the work and simultaneously gives an overview of the methodological framework.

Figure 1

Structure of the Work



Note. Own research.

The introduction builds the foundation of the work by motivating the topic and deriving the research gap. The choice of the applied research approach is explained based on the given prerequisites, especially the existing literature and data. Finally, to show the dependencies and relationships of the different chapters, the structure of the work is explained. Consisting of four subchapters, the second chapter focuses on examining the existing literature, which allows for assessing the current state of research regarding the topics of transparency, ORL, and vacancy. Particularly the research areas of ORL data and vacancy are fundamental for the subsequent evaluations and therefore more extensive than the transparency chapter, as they allow integrating the existing knowledge, e.g., for the precise definition of the research object, the identification of potentially relevant explanatory variables, and their relationships with vacancy. Considering all these different aspects, the research questions (RQs) 1 to 3 can be derived at the end of the chapter.

The literature review equally influences the data and methodology chapters. To answer the specific research questions, the data to be acquired and the methodology to be chosen must be adequate, individually and together. Besides the description of the data acquisition process, the acquired data, consisting of all variables available to an ordinary user of the platform, is described in detail. Based on this description, an assessment of the data is performed. Thus, a comprehensive ORL data set is evaluated to be able to answer the first research question thoroughly. Only a part of this data set is needed to answer the second and third research questions. Further data preprocessing steps are explained, which are additionally necessary to apply the chosen methodology.

The choice of methodology considers existing research and particularly the data contained in ORLs. The methodology consists of three parts, building the foundation for answering the second and third research question and evaluating the derived results. The structure of the description of the methodology is also applied to Chapter 5, which presents the derived results that are subsequently discussed and integrated into the existing literature. A research outlook identifies potential areas for further research and serves as the concluding chapter of this study.

1.3 Research Boundaries

A critical aspect of any research study is to identify and acknowledge the boundaries that may impact the scope and generalizability of the findings. The boundaries of this work are presented subsequently, particularly concerning the area examined, the market segment considered, and

the integrated time frame. These boundaries directly impact the limitations, i.e., the transferability to other potential cases of application.

The area examined is restricted to Germany in this study since Germany constitutes a large market with a most widely homogeneous regulatory framework at the federal level. This homogeneity can be crucial for examining market relationships, as distinctly varying legal regulations or preferences of the people, e.g., between different countries, can lead to changing relationships that are hardly simultaneously integrable into one model. According to the JLL Global Real Estate Transparency Index, Germany belongs to the ten most transparent countries (JLL, 2022a, p. 4). However, especially compared to the most transparent markets of the U.S. and the United Kingdom (U.K.), there is still criticism regarding the state of research, particularly concerning the residential real estate market (Cajias et al., 2020, p. 1023). This combination constitutes a good starting point, as on the one hand, there is a need for research, but on the other hand, there are generally data available that can be analyzed.

In addition, the appropriate level of analysis must be defined since vacancy rates refer to areas that can be of different levels, e.g., municipality, district, or state. Mainly due to the data availability, the district level is used, also offering the possibility to check the transferability to the level up, state level, or down, municipality level.

Also guided by the data availability, the rental market for apartments is particularly well suited, as it is more homogeneous than other market segments and exhibits the highest density of advertisements, which provide more information than other market segments, such as apartments for sale, single-family houses for rent, or commercial properties in general (Colonnello et al., 2022, p. 6). Finally, the reference date or reference period needs to be defined, which in the case of this work is the year 2019, likewise influenced by the data availability, which is further described in Chapter 3.1.

2 Literature Review and Theoretical Foundations

The objective of this work, outlined in the introduction, is to contribute to filling the research gap regarding the influence of ORLs on real estate market transparency by examining ORL data quality in general and testing if these data can be used to increase the accessibility of hardly available information in particular. Therefore, a comprehensive literature review is conducted to motivate the topic of real estate market transparency in Chapter 2.1, to analyze the research related to ORLs in Chapter 2.2, and vacancy in Chapter 2.3, in order to narrow the research gap, lay the theoretical foundations, choose an appropriate research example, and formulate specific research questions in Chapter 2.4.

Although both the data used and the example examined, refer to Germany, international literature is also analyzed due to the availability and quality of the information. This literature regularly uses U.S. examples, which is why this market also receives special attention in this study. However, caution needs to be paid to the possible transferability of results from these studies, as real estate markets deviate from study area to study area, e.g., due to differing legislation, preferences, and physical circumstances.

2.1 Transparency of Real Estate Markets

Transparency is a topic regularly addressed in the context of real estate markets and is considered so important that the German legislator describes the increase of transparency in the German Building Code (BauGB[1]) as desirable and explicitly defines the existence of expert committees that shall contribute to the increase of real estate market transparency (§§ 192-193, 198 BauGB). The fact that transparency is of particular interest in real estate markets is related to their characteristics, which make them less transparent compared to other markets. These characteristics include the heterogeneity of the products, which is inter alia caused by their unique location and immobility, high transaction costs due to taxes and advisory fees, and long durations of development processes (Lucius, 2001, p. 76). Taken together, these characteristics result in low liquidity and poorly comparable products, which in consequence, cause low transparency.

[1] German Building Code in the version promulgated on November 3, 2017 (Federal Law Gazette | p. 3634), last amended by Article 2 of the Act of January 4, 2023 (Federal Law Gazette 2023 | No. 6).

2.1.1 Transparency Definitions and Transparency Measurement

Despite the commonly perceived importance of real estate market transparency and the general interest in increasing real estate market transparency, there exists no all-encompassing and widely accepted definition of transparency (Bloomfield & O'Hara, 1999, pp. 8–9; Schulte et al., 2005, p. 90) and what can be described as transparent in one context may not apply to another. This heterogeneity of the transparency concept is also caused by the fact that transparency is a term that is subject to interpretation and not a simply measurable figure, which is again caused by the missing definition, as the definition of the object to be determined is a fundamental prerequisite for measurement. So, defining and measuring transparency are closely related topics.

Well-known approaches to describing transparency in a general context include the work of Florini (2007, p. 5), who broadly defines transparency by focusing on the availability and usefulness of information. Approaches related to real estate markets include the work of Schulte et al. (2005, p. 91), who focus their attempt to define transparency on the understanding of the underlying market mechanisms and variables and conclude that a market is transparent when the maximum amount of information is available. Therefrom, it can be deduced that the availability of additional meaningful information increases transparency in real estate markets. Similar to this approach, Triantafyllopoulos (2006, p. 61) also bases his definition of real estate market transparency on the understanding and traceability of market mechanisms, e.g., through an established regulatory framework and low information costs, both of which imply the availability of a large amount of high-quality information. Additionally, Lindqvist (2012, pp. 110–111) also highlights the importance of comprehensive information in the conclusion of her exhaustive literature review regarding transparency in European housing markets, which can be consulted for further general reading regarding housing market transparency.

To summarize, all these definitional approaches have commonalities, but a widely accepted universal definition would be desirable. However, for this study, the existing commonality of multiple definitional approaches that an increasing amount of relevant information increases real estate market transparency is sufficient since this work aims to examine the quantity and quality of ORL data and their use for estimating hardly available information. Thus, the informative value of ORL data is examined by assessing their reliability and the additional information provided, which allows concluding if and what kind of additional meaningful information are provided. Thereby, it can be assessed if ORL data contribute to increasing real estate market transparency. However, this approach does not allow transparency to be

measured on an absolute or relative scale and does not allow different levels of transparency to be compared, which is the approach taken by the JLL Global Real Estate Transparency Index widely used in research, e.g., in the studies of Farzanegan & Gholipour (2014), Ionaşcu et al. (2021) and Newell (2016).

2.1.2 Relevance of Real Estate Market Transparency

The interest in defining, measuring, and increasing transparency is also driven by the perceived relevance of transparency for various related issues and the associated consequences of changing transparency. Regularly mentioned, and therefore exemplarily described hereafter, are the influence of transparency on investment decisions, market liquidity, and market stability.

Gholipour et al. (2020, p. 8), for example, show that higher transparency can reduce mortgage defaults and generally improves the stability of financial markets. This contributes to the fact that institutional investors consider transparency an essential factor in strategic decision-making (Newell, 2016, p. 418). A logical consequence is that more transparent markets create higher confidence of the market participants and especially professional investors are attracted by more transparent markets (Schulte et al., 2005, p. 90). Thus, higher transparency leads to higher investments, according to Sadayuki et al. (2019, p. 512) especially by foreign investors, and higher market liquidity. The effect of higher transparency on foreign investors seems plausible, as their necessary effort to obtain information is typically higher than that of local investors and information asymmetries are reduced by the resulting decrease in information deficits (Eichholtz et al., 2011, p. 153). This reduction of information asymmetries is also confirmed by Bleck & Liu (2007, p. 229), who additionally mention the resulting increased market efficiency, which is also confirmed by Jetzek et al. (2019, p. 703). The increase in investments in more transparent markets is also reflected through the countercheck, which shows that a lack of transparency leads to a reduction in investment allocations (Newell, 2016, p. 408).

Besides the investment and liquidity perspective, low transparency also leads to longer marketing periods, as due to the missing information, sellers fear to sell at a price that is too low and therefore start their marketing with a highly overpriced offering, which, in effect, leads to a lower selling price and a more extended marketing period (Nikiforou et al., 2022, pp. 383–384). This finding is also confirmed by Beyerle (2006, p. 126), who states that increasing transparency leads to shorter reaction times.

Furthermore, increasing transparency reduces uncertainty for market participants and thereby eliminates the basis for speculation, which can cause real estate bubbles that in turn can cause financial crises (Farzanegan & Gholipour, 2014, p. 327) and it generally improves the fairness on the markets (Jetzek et al., 2019, p. 703).

However, increasing transparency is not only related to positive consequences, but there are examples of negative effects of transparency. Pavlov et al. (2016), for example, show that certain types of opaqueness lead to a decrease in real estate price volatility due to an insurance-like effect that is based on the pooling of real estate, which, however, would not take place if complete information for all properties were available. The resulting price volatility from increased transparency is usually perceived as unfavorable, and especially homeowners, who typically own only one property, cannot diversify and are therefore negatively impacted. Furthermore, Farzanegan & Gholipour (2014, p. 327) mention that against the previously discussed view, transparency's influence on foreign real estate investments is unclear and that increasing transparency could even lead to decreasing investments.

To summarize, the disadvantages of increasing transparency are in the minority and most of the consequences of increasing transparency are positive, which can also be seen from multiple research articles that describe the influence of online real estate platforms on increasing real estate market transparency positively. These articles include the work of Granados et al. (2010, p. 220), who mention that transparency is increased by providing historical price developments, options to filter offerings, and price estimations. Thus, these platforms not only provide the information given by the offering parties but also contribute to increasing market transparency by evaluating these information, presenting the results, such as price estimates, and thereby creating insights for market participants (Jetzek et al., 2019, p. 703). In general, increasing market transparency through the publication of information by online real estate platforms is mentioned in various articles (Asriyan et al., 2017, p. 2025; Decker, 2021, p. 71; Di Maggio & Pagano, 2018, p. 133; Stamsø, 2015, p. 174). Since this research aims to reassess the hypothesis that online real estate platforms increase real estate market transparency through well-accessible publications of ORLs, the existing ORL-related literature is reviewed in detail in the following chapter.

2.2 Online Real Estate Listings

An online real estate listing is a non-standardized and non-uniform but valuable real estate market information source. The possible applications of ORLs are diverse and range from small-

scale use as comparables for valuation purposes to the large-scale derivation of house price indices. The non-standardization and heterogeneity result from the actual use intended by the originators of ORLs, leading to limitations in their application beyond this use. The originators of ORLs are website operators with well-known examples, such as the Immobilien Scout GmbH, the Zillow Group, Inc., and the Redfin Corp.[2], together with individuals or legal entities who use these websites to market real estate for rent or sale. The influence on the information provision of both groups, website operators and individual or legal entities, is motivated by their objective, which primarily is not the evaluability of ORLs for other purposes than for the information regarding a potential sale or rent, a fortiori not the evaluability by third parties for third-party uses. Therefore, the information given come with restrictions for applications other than the actual intended use of marketing real estate, primarily limitations regarding data availability and quality.

Not only in the context of ORLs, but more generally, the data availability and quality are significantly influenced by the data acquisition process. In the case of ORLs, the primary data acquisition is done by the individual or legal entity that offers the property by filling in a predefined online form. In terms of data availability, this has different implications. An apparent influence is exerted by the party submitting the listing, as it decides about the entries made in particular fields. However, the influence of the website operators is also substantial and can have a positive or negative impact on the information provided, as they decide for which information they include which fields in their online input form and how these fields are designed, thus setting the general framework. Concrete implementations influencing the data availability are, for example, the use of mandatory fields or the provision of convenient input possibilities such as drop-down menus or fields with autocompletion. Similar to the data availability, the data quality is also influenced by both parties. Again, the offering party exerts the more obvious and direct influence, as it consciously or unconsciously decides which inputs are made and in what quality. Still, the influence of the website operators is significant. The implementation of restrictions regarding the input type, e.g., the exclusive admission of numbers in the field for the year of construction, or the inclusion of plausibility checks, e.g., the restriction of the building age to a positive range, can positively influence the data quality and thus simplify the evaluability of the data.

In principle, both parties, the individual or the legal entity and the platform operator, are interested in good data availability and quality, as both aspects positively influence

[2] Reliable estimates for the U.S.-American market share of the Zillow Group, Inc. and the Redfin Corp. are not available; however, their use in related real estate research is widespread.

marketability and, thus, their primary objective. However, these data availability and quality requirements differ for the sale or rental of a property from a scientific evaluation. The data availability is essentially limited to the data that are also important in the marketing process. Depending on the research question, however, further information, e.g., socio-economic data, could be of interest but are typically missing and can only be supplemented by approximate values from additional sources that are regularly not property-specific. Regarding the data quality, in particular, biases within the provided data pose a problem for scientific evaluation. Such biases include the non-randomness of the offered properties, the tendency for an overly positive description of one's property, and the difference between the asking price and realized price. Since these aspects can be of considerable relevance, the limitations regarding ORLs must be considered and weighed against the limitations of alternative sources of information.

Although no study could be found that has attempted to predict vacancy with a similar approach using ORL data, the research using ORLs is diverse and has been increasing in recent years. Thus, the application of information from ORLs goes beyond the use in practical applications, and the existing research provides foundations for this study. Conclusions can be drawn about the different types of ORL data used and the data acquisition possibilities typically applied in Chapter 2.2.1 for the applied data acquisition in this study. The advantages and disadvantages of using these data described in other articles are mostly not specific to the respective research question but are rather generalizable and are therefore also relevant in this study, so they are shown in Chapter 2.2.2 and help to evaluate the use of ORLs compared to other alternatives. A general overview of the types of studies using ORLs in Chapter 2.2.3 provides the opportunity to assess and classify the approach applied in this study into the general ORL research frame.

2.2.1 Sources of Online Real Estate Listings

The introduction described the primary data acquisition by the operators of the ORL websites. In order to use the information from ORLs, typically, a secondary data acquisition has to be done by the person who wants to evaluate the information contained in ORLs, since the researcher usually does not have direct access to the database underlying the website. For most scientific publications that use ORLs and in which this secondary data acquisition is described, this secondary data acquisition can be divided into three categories. First, the website operator can provide the data directly, e.g., through downloadable files or Application Programming Interface (API) access. Second, the data can be acquired through intermediaries, e.g., intermediaries that the website operator contracts to provide the data for specific types of use

by predefined contractual conditions or intermediaries that collect ORL data themselves without having a contractual relationship with the website operator. Third, the user may collect the data of the information contained in ORLs, e.g., through manual website inspection or an automated solution.

The first category includes research from an de Meulen et al. (2011, p. 7), who use house price data derived from ORLs provided by Immobilienscout[3]. Besides them, Deschermeier et al. (2014, p. 2) also mention the use of ORL data directly provided by Immobilienscout during a program called Transparenzoffensive, which intended to increase real estate market transparency, for their construction of commercial rent price indices. Deschermeier & Seipelt (2016, p. 64) also state that they used an Immobilienscout database consisting of all listings in specific cities over a specified period without reference to the Transparenzoffensive but in cooperation with Immobilienscout.

Even more widespread than the use of data from Immobilienscout is the use of data from its U.S.-American counterpart Zillow[4,5], which includes research from Goodman (2018, pp. 640–643), who examine the influence of the Great Recession on property taxes by using so-called Zestimates, which are value approximations derived by Zillow based on ORLs. Besides the Zestimates, in the past, Zillow provided broad access to their data via an API, which was used extensively, for example, by Heidari & Rafatirad (2020, pp. 323–324), who downloaded information of more than five million properties in their encounter to develop a model based on natural language processing to assess the safety of a real estate investment. Especially for research with methods from the field of artificial intelligence, such as natural language processing, large amounts of data are needed and can be conveniently provided via an API, as Zillow has done in the past. This need for data can also be found in the example of the development of a new model to automatize ontology creation in the real estate sector by Heidari et al. (2021), who also base their data acquisition on the previously mentioned Zillow API. The advantages of using data provided directly by the website operator are convenient access to the data, the possibility to request metadata not displayed on the website, and the provision of data in a preprocessed form. The main disadvantage is the dependence on the website operator. Depending on the contract design, the website operator can terminate or restrict the data provision at any time, which has already happened in the past. The data provision by

[3] Immobilienscout hereinafter refers to the website immobilienscout24.de and the company Immobilien Scout GmbH.

[4] Zillow hereinafter refers to the website Zillow.com and the company Zillow Group, Inc.

[5] A Google Scholar search on October 7, 2022 returned 458 hits for the term *Immobilienscout* and 9,710 hits for *Zillow*.

Immobilienscout via the Transparenzoffensive was not permanent and was not replaced for a considerable time until the introduction of the data provision in cooperation with the Leibniz-Institut für Wirtschaftsforschung (RWI), described in the following passage. Likewise, Zillow plans to stop providing data through its Zillow Transaction and Assessment Data set (ZTRAX), stating that they can no longer adequately service the growing group of researchers (Zillow, n.d.). Therefore, the reliance on the website operator to provide data is particularly problematic when the need for continuous data provision is high and the incentives for the website operator to provide data are low. In addition, the website operator can restrict the provision of data by self-selected arbitrary criteria, such as in the study by Berger & Schmidt (2019, p. 15), in which Immobilienscout chose to provide data for no more than two cities.

As an alternative to the direct provision of the data by the website operator, the data can be acquired from an intermediary that is either contracted by the website operator to provide the data or that collects and provides the data on its own initiative. Immobilienscout uses this possibility in cooperation with the Forschungsdatenzentrum (FDZ) Ruhr, a part of the RWI, which provides access to data sets obtained from Immobilienscout at different levels of detail (Schaffner, 2020). By now, multiple studies rely on these data, e.g., Baldenius et al. (2020, p. 205), who use the data to examine the development of rent and sale prices in rural and urban markets. Furthermore, recent research from Eilers et al. (2021) on the social justice of rent prices for people with migration backgrounds or research from Neumann & Taruttis (2022) on the influence of local demographic change on rent or housing prices apply the data provided by Immobilienscout through the RWI. These examples show the widespread use of the Immobilienscout data via the RWI data provision and, thus, on the one hand, the high acceptance of this service and, on the other hand, the resulting considerable degree of dependency on it.

A restriction imposed by Immobilienscout on the use of their data is the authorization for non-commercial use only (RWI, n.d.). For this reason, additional alternative offers have become established, such as the paid provision of data by third-party providers, which do not impose such restrictions. The use of such data provisions can be seen, for example, in the study by Winke (2017) about the effects of aircraft noise on owner-occupied asking prices, who uses ORL data partially provided by the IDN ImmoDaten GmbH, which does not disclose its way of data acquisition. Furthermore, there is the possibility of data provision by third parties who do not cooperate with the website owner but do also not sell the data. An example of such a data provision is the website kaggle.com, where various real estate data sets are provided free of charge. Gupta et al. (2019, p. 3) present an example of using such data sources attempting to

quantify the impact of the Great Recession, as they obtained their Zillow data from kaggle.com. The advantages and disadvantages of the data provision by an intermediary vary significantly depending on the exact form of data provision and can be free of charge or charged. A common advantage of all forms is that they are more convenient than independent manual or automated data collection. Furthermore, the RWI points out that they provide geo-referenced data and perform plausibility checks that are lacking if the data are collected independently, manually, or automatically (RWI, n.d.). A common disadvantage is the dependence on the intermediary. This dependency manifests itself either in the cost of the data or in the form of restrictions on use or limitations on data provision.

Many advantages and disadvantages regarding effort on the one side and dependency on the other are reversed in independent data collection. When ORLs are collected independently, they can either be collected manually, especially if the required data set is not too large, or through an automated solution that is initially more complex but allows for the collection of a more extensive data set. In several studies using ORLs, the data collection method is not disclosed, which can be because the authors do not consider this information necessary or do not want to publish it. An argument for refusing publication can be the legal gray area around web scraping. However, the characteristics of the data analyzed and the formulations used, sometimes indicate the data acquisition method. Püschel & Evangelinos (2012), for example, estimate the costs of airport noise using data they obtained from Immobilienscout. Since they do not name either the Transparenzoffensive or the RWI as the source of their data and the data set is relatively small, with 1,370 apartments evaluated (Püschel & Evangelinos, 2012, p. 599), the data could have been collected manually. Explicitly mentioned is the manual collection of 500 listings by Thießen (2013, p. 11), referring to an untraceable source of Haase (2013).

Larger data sets are evaluated, for example, by Kholodilin & Mense (2011) in their study about the relationship between homeownership rates and ORLs. In their work, they mention that they downloaded the data from three different German websites (Kholodilin & Mense, 2011, p. 3), which could be a chosen formulation that indicates scraping. De Nadai & Lepri (2018) mention that in their attempt to estimate the impact of neighborhood characteristics on housing prices, they collected more than 70,000 listings from an Italian website for ORLs at one point in time. Given the amount of data collected and the lack of mention of a source (De Nadai & Lepri, 2018, pp. 323–324), the most plausible option is that they also scraped the data. Similarly, in their approach to estimate real estate prices based on the visual impression of photos derived from ORL data, Poursaeed et al. (2018) state that they collected an extensive data set from Zillow, consisting of more than 9,000 listings, which also indicates that they scraped the data.

Besides the articles where it can only be assumed that the data acquisition is based on web scraping, others explicitly disclose their approach. Bernstein et al. (2019, pp. 7–8) even go into detail and explain that they used a python-based web scraper to collect data from trulia.com. From the same website, Bhuiyan & Al Hasan (2016, p. 471) scraped 7,216 ORLs of houses, including full-text information such as textual descriptions and number-based information such as the price or the year of construction. Web scraping was also explicitly applied by Han & Lee (2018, p. 521), who, similarly to the previously mentioned work of Poursaeed et al. (2018), tried to incorporate visual data into real estate valuation by using a self-generated data set consisting of 13,049 ORLs.

To summarize, the different data acquisition possibilities have advantages and disadvantages regarding the data itself and the data acquisition process. These advantages and disadvantages, for example, concern the simplicity of the data acquisition process, the integrity of the data set, or the independence from the website operator. Therefore, the choice of the data acquisition process should be made under consideration of the specific requirements of the research question. To be able to principally assess if ORL data are the appropriate choice, the general advantages and disadvantages of ORL data, going beyond the data acquisition process, frequently cited in the scientific literature, are outlined.

2.2.2 Advantages and Disadvantages of Online Real Estate Listings

The German real estate market occupies a unique position in an international comparison with regard to the collection and provision of information, as there exist official institutions in Germany, the Gutachterausschüsse called expert committees, which collect, process, and provide specific real estate market information nationwide for respectively delimited areas based on the German Building Code (§§ 192-193 BauGB). These information include all notarized purchase prices and data derived from those purchase prices (§ 193(5) BauGB), which is possible as the notaries are obliged to send a copy of the real estate purchase contract to the expert committee by the German Civil Code (BGB[6]) and except from rare cases[7], all real estate purchase contracts must be notarized (§ 311b BGB in conjunction with § 128 BGB). This situation is different from the U.S.-American real estate market, where the collection and publication of real estate purchase data are regulated not at the federal but at the state level,

[6] German Civil Code in the version promulgated on January 2, 2002 (Federal Law Gazette | page 42, 2909; 2003 | page 738), last amended by Article 1 of the Act of August 10, 2021 (Federal Law Gazette | p. 3515).

[7] E.g., the partial transition of company assets by the purchase of shares. For further information, note § 1 GrEStG.

which leads to significant deviations regarding the information provision from state to state (Bekkerman et al., 2021, p. 9). Since the research of this study is based on German ORLs, the advantages and disadvantages of ORLs are outlined compared to the purchase price collections of the German expert committees, a commonly used data source for real estate market research in Germany.

An important difference between the price information of ORLs and the purchase price information contained in the purchase price collection of the expert committees is that the price information of ORLs are information about the asking price and not the purchase price. Thus, the purchase price can deviate from the asking price. The deviations can be positive or negative depending on general influences such as market conditions and specific influences such as the asking price of comparable offerings. From those deviations, several potential disadvantages arise as most research questions typically address problems associated with the realized purchase price, not the asking price, and the asking price only serves as an estimate for the purchase price, as can be seen in the examples of Chapter 2.2.3.

With regard to the difference between asking and purchase price, Bauer et al. (2013, p. 7) emphasize that even though statements regarding the absolute price level of real estate markets are unreliable, statements about price trends are not as strongly affected thereby. The disadvantage regarding the absolute price difference is also described by an de Meulen et al. (2011, p. 4), who not only describe the price difference as problematic but also highlight that this price difference may vary over time. This variation is particularly problematic because it has the consequence that it is not possible to determine the difference once and apply this correction to future samples. Bauer et al. (2017, p. 98) complement the argument that the drawback is even more severe if the difference between the listing price and the purchase price varies not only over time, as described before, but systematically with specific characteristics of the listings. Such omitted systematic variations depending on specific characteristics lead to skewed differences between the purchase and listing price, even at the same point in time, and this skewness, in turn, leads to biased estimators. A similar criticism regarding the use of ORLs is expressed by Bauer et al. (2013, pp. 7–10), as they point out the potential lack of representativeness of ORL data samples, which could lead to biased results. Finally, Bauer et al. (2013, p. 7) suggest that the measurement error of ORL data could be more severe than the measurement error in official data.

Contrary to those disadvantages, using ORL data also comes with multiple advantages. An obvious advantage is the easy availability of an alternative to transaction data when transaction data is unavailable (Kay et al., 2014, p. 139). Furthermore, an de Meulen et al. (2011, pp. 4–

5), Dinkel & Kurzrock (2012, p. 6), and Bauer et al. (2013, p. 7) emphasize the actuality of ORL data. This advantage is particularly evident in comparison to official purchase price data. Depending on the applied data acquisition method, ORL data can become available from the moment a listing goes online. In comparison, official purchase price data is less up-to-date due to several factors, including the time required for marketing, contracting, notarization, reporting to the expert committee, and recording and evaluation by the expert committee. Taken together, these factors can result in a delay of several months from the time the offering party considers the initial pricing.

Besides this time-related advantage, there are also content-related advantages. Since it is in the interest of the offering party to provide as much information as possible about the property in order to make a transaction, the information contained in ORLs are often more detailed compared to the information available to the expert committees through other sources (an de Meulen et al., 2011, p. 5). These data availability benefits affect not only the amount of data provided for each property but also the number of properties listed. Market-leading platforms such as Immobilienscout provide information on a large share of all offered properties so that even for submarkets, a relevant number of listings are available, thus making it possible to examine market segments, e.g., in terms of regional distribution or property characteristics (Bauer et al., 2013, p. 7). The possibility to cover large areas with ORL data seems particularly valuable, as purchase price data is not centralized and research covering large areas becomes effortful due to the need for coordination with multiple expert committees and probable deviations in their data collection and processing, thus, in their data provision.

Contrary to the criticism of Bauer et al. (2013, p. 7) and an de Meulen et al. (2011, p. 4) regarding the difference between listing price and purchase price, Dinkel & Kurzrock (2012, p. 18) found that this difference does not vary systematically with other variables they studied. Besides that aspect, even more contradictory literature can be found with regard to the previously described negative effect of varying differences between asking price and purchase price, as Deschermeier et al. (2014, p. 4) state that it can be assumed that the deviation of these prices is stable. Thus, it would be possible to use ORL data to approximate realized purchase prices and actual rents. Nevertheless, the arguments provided by related research are not conclusive. In the context of this study, two other differences between ORL data and other data sources become relevant. First, ORL data is most widely homogeneous in its structure[8]. Second, depending on the applied data acquisition approach, the acquisition of ORL data can be both

[8] Regarding the characteristics collected and the input type for each characteristic, e.g., numerical input or text input.

convenient and inexpensive. The decision to use ORL data should therefore be made under consideration of the listed advantages and disadvantages of ORL data and go beyond the different methods of data acquisition described in the previous chapter. For research questions concerning the German real estate market, especially the consideration of purchase price data from the local expert committees seems relevant.

2.2.3 Common Fields of Application of Online Real Estate Listings

The scientific evaluation of ORLs is a relatively new field of research that has its natural origin in the introduction and spread of online real estate platforms. The first mentions of the websites Immobilienscout and Zillow found in the scientific literature initially deal with the description of these portals and date back to 2001 for Immobilienscout (Böhm, 2001) and 2006 for Zillow (McDonald, 2006). Since then, research using data from these and similar platforms has intensified, however, focusing on topics different from vacancy research. Despite this different research focus, these studies provide additional valuable information beyond the results presented in the previous two chapters, e.g., on the spatial aspect of ORL data or the data cleansing techniques applied. The following chapter provides an overview of such recent and recognized studies and classifies them into more general research areas, depicted in Figure 2 and described in the following.

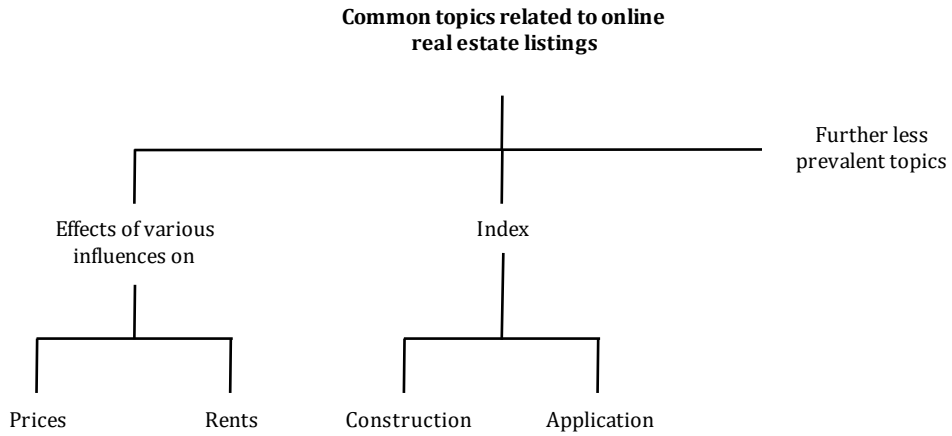
The most prominent range of topics includes research examining the effects of different characteristics and influences on asking prices and asking rents. Prices and rents are convenient research subjects in the context of online real estate platforms due to their data availability, as both information are crucial to the offered property and, therefore, usually not missing. This data availability is essential since missing values in relevant parts of the examination data set can lead to a massive reduction of the sample size or the need for complex data imputation with the associated uncertainty about the data imputation quality.

The distinction between studies on asking rents and asking prices is more of a substantive distinction than a distinction regarding the methods used and the results obtained since they are often similar in both groups, as displayed by the following examples. Substantially more literature is available covering asking price related topics, which can be further subdivided. A wide field of research tries to estimate property values based on ORL data. Jafari & Akhavian (2019), for example, use data derived from Redfin to evaluate a standard hedonic pricing model, which they have built based on American Housing Survey (AHS) data. They especially

criticize the missing information on property characteristics in the AHS data set and emphasize this advantage of ORL data compared to AHS data (Jafari & Akhavian, 2019).

Figure 2

Common Topics Related to Online Real Estate Listings



Note. Own research.

A relatively new trend in property price estimation research is visual data integration. Evident to the human surveyors and considered highly relevant, it is complex to integrate visual information on the property condition in a fully automatized valuation model. This gap can be closed when visual data is available, meaningful, and integrated into automatized valuation models. Poursaeed et al. (2018), for example, were able to build a convolutional neural network to classify property images and integrate the information derived into a valuation model. The property images and further information integrated were derived from Zillow. By integrating these information, they created a model that performed even better than the estimates calculated by Zillow (Poursaeed et al., 2018). This example demonstrates an additional advantage of ORL data over traditional data sources because, in contrast to other data sources, such as purchase price collections, real estate images are regularly included in ORL data and contain meaningful information. Heidari & Rafatirad (2020) complement a study that focuses on asking rents. Similar to the previously mentioned study using visual data, they also integrate alternative information sources using textual data like housing descriptions for their rent price prediction model. These examples show that it is the differentness of ORL data compared to

traditional data sources that offers an opportunity for new analysis approaches providing information beyond that obtained from traditional analyses.

Another group of articles examines the impact of different types of infrastructure on the asking price. Bauer et al. (2015) provide an example analyzing both rent and sales prices by investigating the influence of the Emscher Conversion, an ecological reconstruction of the river Emscher in the Ruhr area, on housing for sale and rent. They use georeferenced data provided by Immobilienscout and apply a hedonic price function to estimate the effect of the renewal and find that there is no significant effect for the development of rent prices but a stabilizing influence of the measure on housing prices.

Similarly, a positive relationship between prices and infrastructure can be assumed if the infrastructure leads to positively connoted characteristics like vitality and walkability. This presumption is confirmed by the study of De Nadai & Lepri (2018) for an example in which they use data from the largest Italian online real estate website Immobiliare to now-cast housing prices and find that vitality and walkability play a crucial role. Kay et al. (2014) supplement these findings by examining the effect of transit-oriented development on housing values. Their study confirms common hypotheses regarding infrastructure, e.g., the positive influence of nearby railway stations, high-quality schools, and park access on property values (Kay et al., 2014, p. 139).

On the other side, there are also infrastructure measures that are more likely to be associated with negative impacts on property values that can be analyzed using ORLs. Due to the associated perceived risk and experienced emissions, the effect of infrastructure measures such as wind turbines and nuclear power plants is expected to be negative. Frondel et al. (2019) confirm this hypothesis using data from Immobilienscout to build a hedonic price model that reveals the negative impacts of wind turbines on property values. Nevertheless, the effects are complex and depend on further circumstances of the specific case. In rural areas, for example, they are stronger than in urban areas and, as expected, the influence of wind turbines fades out with increasing distances (Frondel et al., 2019). Besides this distinct relationship, Bauer et al. (2017) show by the example of nuclear power plants that in some cases, there are also more complex effects, e.g., negative impacts due to perceived risks can be outweighed by other positive influences such as the positive economic impact of a nuclear power plant for the adjacent region.

Climate change, a topic that is receiving increasing attention in various research fields, also plays an important role in some studies dealing with the impact on real estate values. Bernstein

et al. (2019) address the effect of rising sea levels on property values by using data obtained from Zillow, including information on the exact location for the flood risk assessment, but also on property characteristics like the availability of sea view or the square footage of living space and find that risk exposed homes sell for a lower price than comparable properties without a similarly high risk of flooding. Miller & Pinter (2022) inherently confirm this price discount for properties with a high risk of flooding. They also show that an individual flooding event does not change the perceived value. Instead, especially in areas with a well-educated population, the value discount is more permanent and not event-driven (Miller & Pinter, 2022). These studies show that depending on the research objective, due to a large number of influences on property values and their interdependencies, it is essential to have access to a comprehensive database. Otherwise, it is not possible to isolate specific impacts, such as the risk of flooding, from other characteristics, such as ocean view. ORLs typically provide a large amount of information and their scope can be further enhanced by intersection with other data sets.

In addition to such thematic focus areas, several solitary studies deal with other specific influences on asking or purchase prices. B. Wang (2021), for example, examined the effect of COVID-19 on house prices using self-collected data from Redfin, which included information on housing characteristics integrated as control variables. Busch (2016) was able to show a correlation between the change in the net migration of German citizens and the asking prices in a study examining the influence of migration on asking prices for apartments from Immobilienscout. Micheli et al. (2019) also used data from Immobilienscout and additionally listing data from the Dutch service FUNDA to estimate the impact of the Dutch-German border on property prices and find that changes in tax rules lead to arbitrage effects, implying that property markets are even interconnected across borders. The differences between these studies emphasize the variety of existing influences on property values. In order to incorporate many of these influences into existing and future research, a broad database is needed, which can be provided in part by ORL data.

A second major branch of research focuses on real estate price indices. Similar to research on asking prices, ORL data related research on real estate price indices is attractive from multiple points of view. Index construction for real estate prices, especially based on regression analysis, is data intensive since numerous characteristics affect property values and thus must be integrated (Bailey et al., 1963, p. 934). Therefore, ORL data is well suited for this task since it provides information on many property characteristics that can be used in this context. Additionally, real estate price indices are convenient to use in their application for other related

research topics as they consolidate a large number of different and complex price information of individual subsegments into a single figure that is easy to use and interpret.

The construction of real estate price indices is described in detail for asking prices and asking rents. An de Meulen et al. (2011) provide one of the first studies describing the use of ORL data for index construction. In their research contribution, they use data from Immobilienscout to measure recent price changes and go even further than constructing a simple price index as they attempt to develop a price prediction model by integrating other macroeconomic information (an de Meulen et al., 2011). Bauer et al. (2013) complement a comprehensive overview of the motivation to introduce an index based on asking price data from Immobilienscout. Besides explaining the construction of this index, known as the IMX and now published by Immobilienscout, they also emphasize the advantages, especially the data availability, of ORL data for property price index construction (Bauer et al., 2013, p. 7).

Reflected in the larger number of offerings, the data availability is even better for rental prices than for sales prices and allows constructing indices for subsegments such as commercial properties[9]. This data availability enables studies like the work of Deschermeier et al. (2014), who describe the construction of rental price indices for commercial real estate and even further subdivide their analysis into segments for office and retail, and in regional markets. The potential areas of application are diverse and go beyond commercial real estate. It is possible, for example, to create topic-specific indices depending on current research trends or problem areas. Student rental apartments are only one example taken up by Deschermeier & Seipelt (2016), who explicitly motivate their development of a price index for student rental apartments with the tightening market for student living.

The diverse applications of ORL data based real estate price indices can be seen in the exemplary studies of Holt & Borsuk (2020), who use the Zillow Home Value Index (ZHVI) to estimate the influence of green space on housing values and of Gupta et al. (2019) who also use the ZHVI, in a completely different context, to examine the impacts of the Great Recession on housing values. Besides the application of indices representing home values, another widely applied group includes indices describing the development of rents. Examples are the study by Bao & Shah (2020), who use the Zillow Rent Index (ZRI) in their examination of the influence of Airbnb offerings on rents, and by Glynn & Fox (2019), who use the ZRI in their examination of homelessness in urban America.

[9] A search on Immobilienscout on November 2, 2022 returned the following rent/sale proportion of search hits for the German market: Office 54,202/2,321; Retail 8,030/1,099; Warehouse 16,018/1,335.

From the studies presented on the use of ORL data in conjunction with indices, several conclusions can be drawn for the use of ORL data in general. Due to the increasing amount of data, several indices, even for small segments of the real estate market, are available and allow a convenient integration if no detailed information about the individual properties need to be considered. The indices cover both the development of rents and sales and, in some cases, are published directly by online real estate platforms, being especially helpful when ORL raw data is not supplied. The possible applications of these indices are various, as can be seen from the presented existing research, also indicating which limitations or other aspects need to be considered when using such ORL index data.

In addition to the studies that fall into one of the two main research areas of prices and rents or indices, several other studies show the numerous possible applications of ORL data. The services offered by companies like Immobilienscout or Zillow, in addition to pure brokerage services, are diverse and include price estimations since their early years. Hollas et al. (2010) tested the accuracy of the price estimates by Zillow by comparing them with actual sales prices, which is specifically important, as those estimates are the input data for many other studies. In contrast to the previously presented studies, Bhuiyan & Al Hasan (2016) do not have the typical focus on housing value or rent prices, instead, they broaden the perspective by examining the time required to market a property depending on its characteristics using Trulia data. Similar to the previously mentioned study of Bao & Shah (2020), Coles et al. (2017) also investigate the consequences of the existence of Airbnb in specific areas and intersect Airbnb data with Zillow ORL data, but contrary to the other study, they do not only concentrate on rents but on more general usage patterns such as the geographic dispersion of Airbnb listings. Just as different is the work of Bauer et al. (2011), who use Immobilienscout data in their study that examines the influence of neighborhood unemployment on individual unemployment, or the study by Goodman (2018), who use Zillow data to examine the influence of the Great Recession on property taxes.

The large variety of the previously described studies applying ORL data shows the great potential of ORL data for various fields of research. Despite this variety of topics, no study explicitly formulated the hypothesis of a relationship between vacancy rates and the information contained in ORLs. However, this correlation appears to be inherent, as one can posit that a significant portion of vacant properties is marketed to generate future income streams, which would typically be the rational reaction for individual market participants endeavoring to optimize their profits.

2.3 Vacancy

Vacancy data, in general, and housing vacancy data, in particular, are a valuable source of information from various points of view depicted in detail in Chapter 2.3.1, e.g., for general market assessments or for describing and solving vacancy-related problems. Nevertheless, research on housing vacancy is underrepresented in real estate market research compared to other topics, such as research on housing prices, measured by the number of publications[10]. Therefore, this work not only examines the influence of online real estate listing data on the transparency of real estate markets in general but also tries to extend knowledge in the field of vacancy research by using the vacancy rate example. The scientific and application-oriented relevance of vacancy data is shown in detail in the following, not only to prove its relevance but also to show the current state of scientific knowledge and to be able to derive requirements on vacancy data from typical scientific and non-scientific use cases.

Despite the considerable relevance shown below, there is no universal and generally accepted definition of the term vacancy. Instead, vacancy is not even explicitly defined in many articles, but a vacancy definition is instead assumed implicitly, which can lead to ambiguities. To avoid such ambiguities in this study, to emphasize the multiple different characteristics of different vacancy data provisions, and to provide an approach for a future vacancy definition classification, common categories of definitions are outlined and the definition applied in this study is classified.

The different categories of definitions are, to some degree, caused by different methods of vacancy data collection. To contextualize and assess the method of vacancy rate estimation developed in this study, these other common approaches are presented subsequently. The description of the basic principles these methods are based on and the characterization of their advantages and disadvantages allow to assess the results of the method developed in this study in terms of its practical utility. Finally, the comprehensive analysis of vacancy-related literature enables the identification of potentially relevant data sources for explaining vacancy rates in this study.

[10] Google Scholar Results: ”Housing vacancy“ – 4,980 search hits; ”Housing price” – 163,000 search hits.
Web of Science Results: ”Housing vacancy“ – 96 search hits; ”Housing price” – 2,165 search hits.
Retrieved August 19, 2022.

2.3.1 Relevance of Vacancy Data

The process of vacancy data collection is both time- and labor-intensive, which also becomes evident from the description of the commonly used methods to measure vacancy in Chapter 2.3.3. However, the relevance of the availability of vacancy data can be justified by various arguments. First, the general scientific interest in vacancy data is shown. Second, the problems associated with vacancy are outlined, implicitly including the intention to use vacancy data to understand these issues better and potentially take countermeasures. Third, specific conceivable or existing use cases of vacancy data are demonstrated.

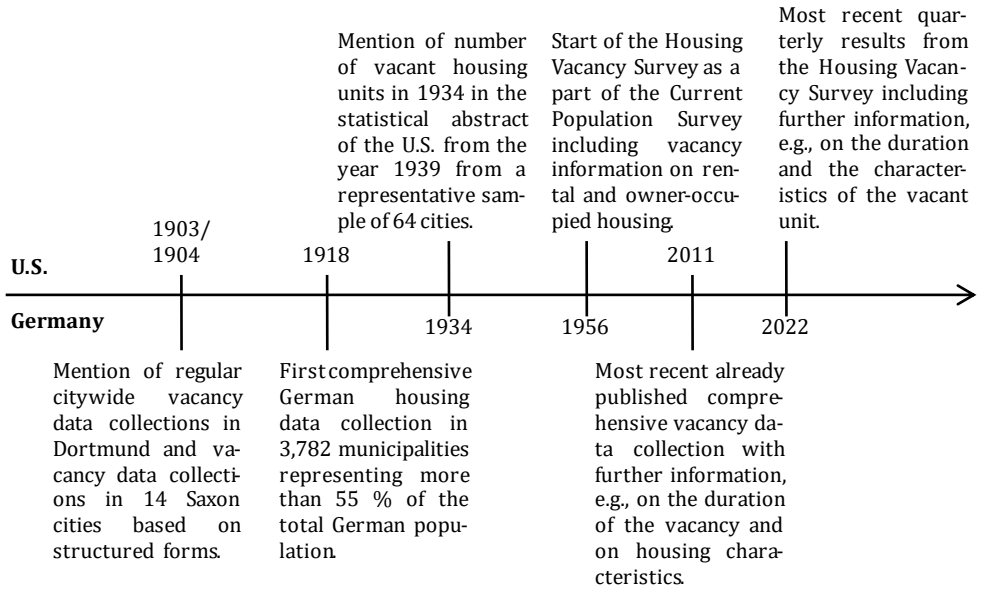
The general scientific interest in vacancy data becomes apparent through both the data supply and the data demand side. An enduring provision of vacancy data, associated with the aforementioned high effort for the data collection, on the one side, and the emphasis on the importance of these data in the scientific literature, on the other side, indicate this assumption.

On the supply side, the provision of vacancy data is regularly ensured by governmental agencies, as can be seen in Figure 3 below, providing an overview of selected examples of the development of vacancy data collection in Germany and the U.S. The figure demonstrates the long history and strong development thereof, as efforts to collect data in both countries have not weakened over time and surveys are conducted to the day.

On the demand side, the need for vacancy data expresses itself in the general reference to the importance of vacancy data (J. Li et al., 2019, p. 3) and, in particular, in demand for further additional data, dividable into different information needs. Deilmann et al. (2009, p. 667), for example, point out the importance of more detailed data on vacancies regarding the regional scale. This argument is also supported by Braun et al. (2014), who argue that contrary to the still recent headlines on housing shortages, there is fundamentally sufficient housing space in Germany, which does not match demand regarding regional distribution. The monitoring and steering of region-specific countermeasures is therefore of particular importance and dependent on the continuous nationwide collection of vacancy data that has been lacking to date in their opinion (Braun et al., 2014, p. 5). Others emphasize the need for additional information in the form of more detailed content. S. A. Gabriel & Nothaft (2001, pp. 122–123) suggest collecting a separate vacancy incidence and duration, allowing for gathering further information on fluctuation and thereby highlighting the importance of high-frequency surveys.

Figure 3

Selected Examples of the Development of Vacancy Data Collection



Note. Own research. The information depicted are retrieved from Statistisches Amt Dortmund (1904, p. 1), Statistisches Bundesamt (1956, pp. 5–17), Statistisches Bundesamt (1975, p. 12), U.S. Census Bureau (2022), U.S. Department of Commerce Bureau of the Census (1940, p. 877), and Weinberg (2006, p. 122).

In addition to these concrete application-oriented motivations, there is additionally a more, but not exclusively, scientifically based interest in understanding vacancy. A typical question is, for example, about the key driving factors of the vacancy rate (J. Li et al., 2019, p. 2; Newman et al., 2019, p. 809). This interest is supported by Gentili & Hoekstra (2019, p. 444), who mention the importance of the different vacancy causes and emphasize that different causes require different reactions. Thus, they bridge the gap from the purely scientific-oriented interest to the application orientation. Besides these explicit demands for additional data, the description of the problems associated with vacancies makes the importance of providing vacancy data even more apparent, as it highlights the opportunities associated with vacancy data availability, e.g., in the context of investigating causes of or evaluating countermeasures against vacancy.

The problems associated with vacancy are manifold and affect not only the individual but also impact society as a whole. To demonstrate the positive impact that could arise from understanding vacancy better, the most prominent problems are collected in the following. At the level of the individual homeowner, the loss of revenues is the economically most direct

consequence of vacancy (BBSR, 2017, p. 3; Braun et al., 2014, p. 6; Dransfeld & Lehmann, 2007, p. 22). Further direct negative effects are experienced from increasing costs, e.g., for the protection of the vacant property, the search for new tenants (Couch & Cocks, 2013, p. 507), or minimum heating that needs to be employed for maintenance reasons and which cannot be allocated to a tenant. As a result of the missing income and the sometimes increasing expenses, the cash flow turns negative during the vacancy period, leading to a reduced value compared to a situation where the property is rented. The increasing costs and losses of revenue thus directly lead to a decreasing monetary value, e.g., estimated by a discounted cash flow (DCF) calculation, of the affected property. Interestingly a loss in value can also be observed for rented properties in the same area (Kumagai et al., 2016, p. 709; Manville & Kuhlmann, 2018, p. 471; Nam et al., 2016, p. 1; Whitaker & Fitzpatrick IV, 2013, p. 85). This effect does not arise as an immediate reaction in the DCF calculation. Rents and discount rates are lagging values and are stable for some time. Thus, the effect is indirect and could be caused by specific characteristics associated with high vacancy areas. Well studied, for example, is the change in the crime rate, which also increases when vacancy increases (Couch & Cocks, 2013, p. 507; Du et al., 2018, p. 2; Konomi et al., 2017, p. 130; Kumagai et al., 2016, p. 709; Nam et al., 2016, p. 1; K. Wang & Immergluck, 2019, p. 513). Besides the increase in the crime rate, an increase in the susceptibility to fires and damages through natural disasters such as strong winds can be observed (Konomi et al., 2017, p. 130; Kumagai et al., 2016, p. 709). This phenomenon could be caused by an inadequate maintenance status of some vacant buildings. Together with the increasing probability of illegal waste disposal, leading to effects such as bad smell (Konomi et al., 2017, p. 130), significant vacancy leads to a degradation of the urban landscape (BBSR, 2017, p. 3), decreasing the attractiveness of the total area and thereby also affecting other properties that are still rented.

Besides these apparent, owner-affecting effects, the losses in value also affect the municipalities and the banking industry as they reduce the quality of life for the inhabitants, thereby, their satisfaction with the public sector, the income of the city due to shrinking property taxes and they additionally increase the risk of credit defaults (Accordino & Johnson, 2000, p. 303; Braun et al., 2014, p. 5; Manville & Kuhlmann, 2018, p. 471; Spehl, 2011, p. 41). Furthermore, the valuation of properties, e.g., for taxation purposes, gets more complicated due to the lack of easy-to-gather comparables, as areas of high vacancy are often at the same time areas of low liquidity (J. B. Hollander & Hartt, 2019, p. 256; Spehl, 2011, pp. 38–41).

On a broader level, the impact on society as a whole reveals itself in a decrease in the functional effectiveness of urban structures (Jin et al., 2017, p. 98) caused by the lower degree of

utilization of the overall capacity of those systems, increasing the cost per person. Using the example of waste disposal or water supply, it can be seen that a higher degree of utilization does not only lower the cost per person due to the fixed costs effect, e.g., the costs for once occurring acquisitions split up on a larger population, but that even the variable costs can decrease if larger and potentially more efficient facilities get profitable. In addition to such directly measurable cost effects, vacancy can be seen as a classical misallocation of resources (Couch & Cocks, 2013, p. 507; Zhang et al., 2016, p. 4). Vacant housing units represent a misallocation of physical resources to construct the buildings and financial resources to finance them. Both of these resources could have been used in areas with a higher demand for housing space, thereby increasing the economic welfare of the affected community.

The aforementioned effects of individual vacant properties and the concomitant reduction of the vitality of communal life (Konomi et al., 2017, p. 130) are not limited to the status quo. Vacant buildings can infect neighboring buildings, inducing a downward spiral and causing wider urban decay (Braun et al., 2014, p. 6). This effect is especially important to notice as vacancy can be seen as a precursor to the even more severe abandonment, being a significant variable in the prediction thereof (Hillier et al., 2003, pp. 92, 101) and thereby strengthening the justification for an early vacancy detection, which provides the opportunity to take countermeasures. These effects are not theoretical issues but a current topic given the demographic change in many industrialized countries and the trend toward urbanization. Thus, vacancy monitoring is one of many potential applications of vacancy data, further explained subsequently.

According to K. Wang & Immergluck (2019, p. 512), vacant and abandoned buildings have been a concern in urban planning as well as in urban management for decades and Franz (2001, p. 263) puts it straight by stating that scarcity of vacancy data exacerbates the political assessment thereof, including the development of countermeasures. Urban planning structures the development of land use in the built environment and tries to ensure that the right amount of the correct type of designated building land is available in the right place from a strategic point of view. The goals to be considered in this process are diverse and sometimes contradictory. In Germany, the law provides a guideline on which interests specifically need to be considered. Explicit mention of vacancy is made, for example, in § 1a(2) BauGB, urging to consider the possibilities that arise from using vacant properties before designating, e.g., formerly agriculturally used areas, as new building land. This legal requirement presupposes the availability of knowledge about vacant properties.

Besides such specific justifications for the need for vacancy data, there are also more generalizing rationales. Y. Pan et al. (2021, p. 2) argue that knowledge about vacant properties is generally important for effective urban planning. This argument is supported by L. Wang et al. (2019, p. 8568), who mention that accurate information about vacancy rates help city planners to develop better land-use strategies. Better information could be used, for example, to decide which locations are preferable for which uses, or which building types are preferable compared to others based on their demand. Thus, vacancy information could improve land use designations by providing additional information in the weighing process.

Contrary to urban planning, urban management does not represent the strategic perspective but tries to achieve the objectives predefined by the results of the urban planning process. Therefore, its measures that could profit from better vacancy data are more of an operative nature and include the possibility of the introduction of cadasters where continually monitored vacancies could be merged and which can be used to bring the vacant buildings actively back into use by supporting the owners in their marketing attempts (Spehl, 2011, p. 26). As an example of another type of use case, Spehl (2011, p. 26) describes a German case of financially supported demolitions of buildings based on the entries of a vacancy cadaster. Analogical to the specific vacancy cadasters in Germany, a similar approach is already well established in the U.S., where several major cities provided universal information about their community, including vacancy data or estimates of vacancy risk, almost 20 years ago (Hillier et al., 2003, pp. 91–92).

Besides these concrete examples of application, there are also more generalizing rationales. J. B. Hollander & Hartt (2019, p. 251), as well as Y. Pan et al. (2021, p. 2), emphasize the importance of understanding vacancy for policymakers to reduce the negative effects of vacancy and foster sustainable urban development. Although many of the previously mentioned arguments specifically refer to urban areas, they can presumably be adapted to rural land management, which is, for instance, supported by J. Li et al. (2019, p. 2), who particularly mention the possibility of optimization under the precondition of accurate and timely information concerning vacancy.

The share of vacant housing can moreover be seen as a generally important source of information, providing insights regarding the healthiness of the housing market and thereby supplying information for policy-making (Z. Chen et al., 2015, p. 2188). Such information could help politicians to evaluate the success of counter-measures (Dransfeld & Lehmann, 2007, pp. 21–22). Due to the many ways, the vacancy rate can be used as an indicator, Rink & Wolff

(2015, p. 311) criticize that it has not yet been systematically recorded and evaluated for Germany.

2.3.2 Categorization of Vacancy Definitions

In general, it can be assumed that there is a consensus on the basic concept of vacancy. Braun et al. (2014, p. 6) and also Hoekstra & Vakili-Zad (2011, p. 56) summarize this basic concept as the state of a building not being occupied. Beyond that, unfortunately, there is no widely acknowledged and more specific standard definition of vacancy (Hillier et al., 2003, p. 102; Y. Pan et al., 2020, p. 3, 2021, p. 3; Zhou et al., 2021, p. 3). Instead, multiple studies explicitly describe or implicitly assume a vacancy definition that best serves their study objective, resulting in numerous gradually different definitional approaches and leading to difficulties in processing, e.g., comparing results or merging data from different vacancy-related studies. In short, each individual vacancy definition leads to an individual vacancy count, which complicates the comparison of different research results. To clarify and emphasize these differences, as well as to enable an informed decision regarding the use of a particular concept of vacancy, a categorization of vacancy descriptions and definitions is conceptualized in this chapter. The terms definition and description are both used in this context since descriptions of the vacancy data collection process or even implicitly made assumptions can have a definitional character if no definition is explicitly stated. The detailed examination of vacancy definitions and descriptions particularly aims to raise awareness related to comparing vacancy numbers from different sources. Furthermore, the derived framework enables to classify the vacancy numbers derived from the method developed in this study into this general framework and contributes to increasing real estate market transparency with regard to vacancy.

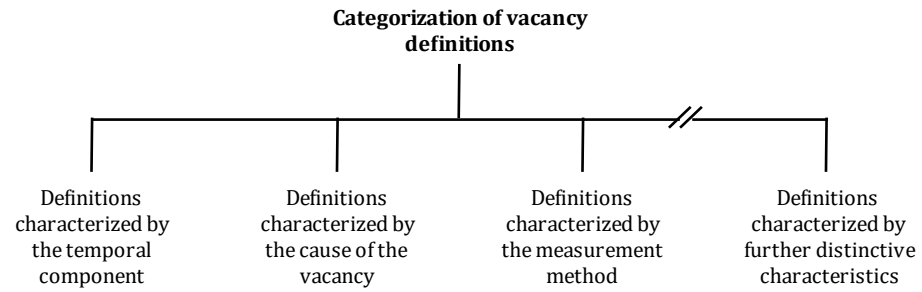
This approach of developing a framework for the classification of vacancy definitions is not the first approach to categorizing vacancy definitions or descriptions, as Rink & Wolff (2015) have developed an approach that tries to distinguish vacancy definitions specifically for housing vacancy by three criteria. The first of their three derived criteria concerns the question of what is defined as a housing unit. The second is what constitutes residential use and the third is how long a vacancy must prevail to be counted (Rink & Wolff, 2015, p. 315). Unfortunately, many definitional approaches of established vacancy-related studies cannot be classified by such a kind of systematization, as they neither provide explicit definitions nor information on all relevant criteria but rather contain implicit vacancy descriptions. Furthermore, they sometimes contain other types of information instead, which are not considered by the previously described

approach. To account for this diversity of descriptions and ultimately be able to classify the definition of this study, an attempt is made to systematically group these vacancy descriptions and definitions according to the categories central to their definitional approach. Due to the extensive amount of literature on vacancy, this attempt to categorize and summarize explicit and implicit vacancy definitions and descriptions is limited to recent and widely accepted research in the field of vacancy detection and estimation, the research area of this study. Therefore, the categorization provided in this chapter should be understood as a starting point for future, more extensive research specifically focused on developing a comprehensive framework for classifying vacancy definitions and descriptions by a more extensive literature review.

The categories of topics that are central to the definitional approach of the analyzed articles and that allow vacancy descriptions to be distinguished are the temporal aspect of vacancy, the cause for vacancy, and the method used to measure vacancy. Vacancy numbers are typically influenced by all of these aspects and thus, those definitional approaches should not be regarded as isolated vacancy definitions but rather as different perspectives that can be applied to vacancy. A comprehensive vacancy definition should provide information on all of these different aspects. Figure 4 depicts this non-exhaustive categorization.

Figure 4

Categorization of Vacancy Definitions



Note. Own research.

Hereafter, the shown categories are further described and subcategorized, as those categories again are collections of different subtypes of vacancy definitions within those categories. Using

the temporal component to distinguish vacancy definitions and descriptions is obvious since it is a concretely measurable variable simultaneously well suited for differentiating vacancy in various aspects. One of those aspects is that different vacancy durations are typically associated with different consequences and causes. The consequences influenced by time are, for example, the differing value losses and secondary effects such as rising crime rates or difficulties in re-letting. Difficulties in re-letting can particularly be assumed when the duration of a market-active vacancy exceeds a certain duration.

Conversely, short-term vacancy can sometimes be assessed as beneficial as it is inevitable to provide efficient housing markets since it enables relocation of tenants and prevents the formation of unbalanced positions of power between property owners and tenants in the form of oligopolies or monopolies. This idea is formalized in German literature as the fluctuation reserve (Bangert et al., 2006, pp. 179–180; Dransfeld & Lehmann, 2007, p. 21) and similarly in international literature as the natural vacancy rate (Rosen & Smith, 1983, pp. 781–782) which is defined as the rate at which there is neither excess demand nor excess supply. In conclusion, it can be stated that depending on the duration of the vacancy, the overall evaluation of the consequences can change from a positive to a negative assessment, as short-term vacancies are sometimes considered beneficial, while long-term vacancies are seen as rather detrimental. Thus, depending on the intended use, the duration measurement can be crucial for the quality of the conclusions regarding the vacancy.

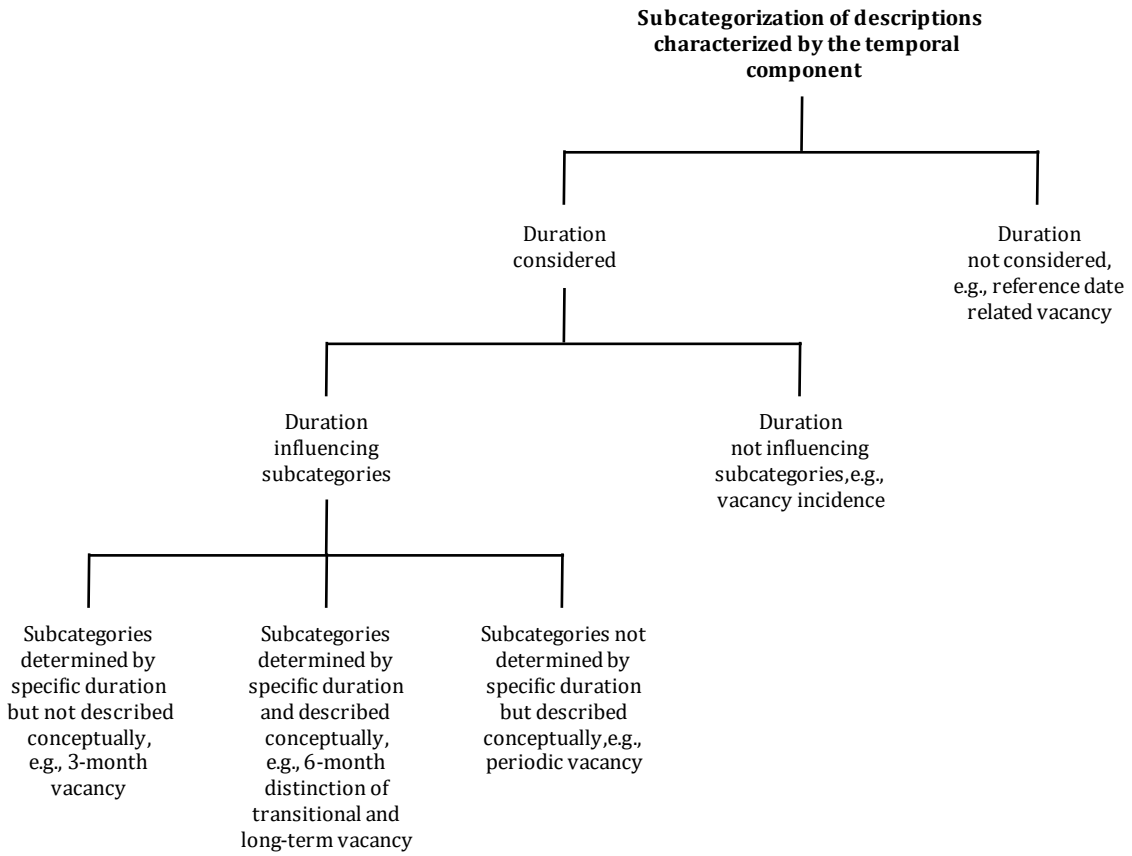
The most basic distinction related to vacancy durations one can make and which is made explicitly or implicitly in all vacancy definitions is the distinction of vacancy recorded with or without consideration of its duration. Figure 5 below provides an overview of the multi-level categorization, initially differentiating between definitions considering the duration or not. Examples for all subcategories are explained below using examples from recent research.

In their study examining vacancy in Berlin, Bangert et al. (2006, p. 179) explicitly mention both approaches, differentiating reference date related vacancy that does not take the duration of a vacancy into account from duration dependent vacancy, which considers vacancy durations by requiring a certain number of months of being vacant to be recorded. Zhang et al. (2016, p. 6) provide an example where reference date related vacancy is used, as they calculate the housing vacancy rate as the ratio of the sum of the area of all vacant housing units to the sum of the area of all housing units at a specific date. Thus, a differentiation regarding the consequences associated with different vacancy durations cannot be made due to the missing information. Recording the duration of a vacancy complements the determination of whether or not a building is vacant by providing additional information, which allow further conclusions to be

drawn. That temporal information can be used to subcategorize types of vacancies further or can be taken into account without any duration-based distinction.

Figure 5

Subcategorization of Vacancy Descriptions by the Temporal Component



Note. Own research.

Three different categories can be derived if the duration is considered for a further subdivision. First, a specific period can be given to serve as a criterion for distinguishing different types of vacancies without associating these types of vacancies with contextual interpretations. Second, the vacancies can, complementary to the description of the period of time, also be associated with an interpretative term that can imply consequences or causes of the specific type of vacancy, such as short-term or periodic vacancy. Finally, there is the possibility to describe

vacancies solely by such an interpretative term without giving more detail about how the authors define this term in measures of time.

Bangert et al. (2006, p. 179) and Spehl (2011, p. 29) provide examples of studies of the first group of definitions, which solely lists typical time ranges that serve as minimum periods, e.g., 3, 6, or 24 months to be counted as vacant. Despite that, they do not connect those individual groups of vacancies with descriptive terms and the vacancies are simply described as 3-month, 6-month, or 24-month vacancy.

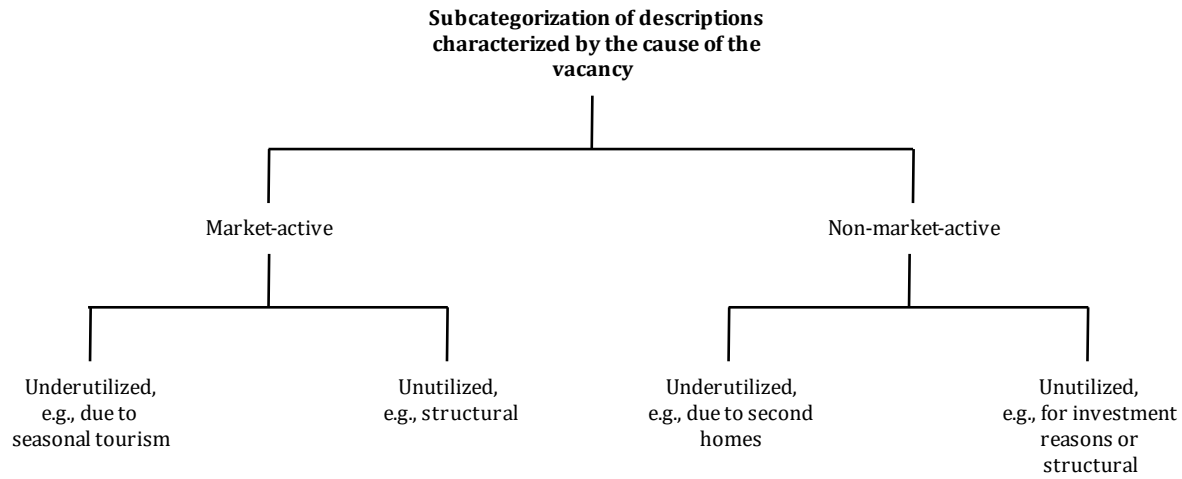
Explaining changes in long-term vacancy after the U.S. foreclosure crisis, K. Wang & Immergluck (2019) specifically distinguish between long-term vacancy and vacancy of a shorter duration, which they designate as transitional vacancy if it is prevalent for less than six months and thereby provide an example for the second group. Based on a United States Postal Service (USPS) data set with a different cutoff of 90 days instead of half a year, Whitaker & Fitzpatrick IV (2013, p. 85) distinguish between seasonal and long-term vacancies. J. Li et al. (2019) try to estimate housing vacancy from secondary data, i.e., power consumption data. Contrary to the aforementioned approaches, they do not distinguish between vacancies of shorter duration and vacancies of longer duration but argue that housing can only be labeled vacant when it has been unoccupied for at least one year. As a distinction to this, they introduce the concept of underutilization for buildings with only low utilization (J. Li et al., 2019, p. 5). Lee et al. (2022, p. 5) follow this example with reference to Korean legislation, which in Article 2 (1) 1. of the Korean Act on Special Cases Concerning Unoccupied House or Small-scale Housing Improvement also defines properties as vacant when they are unoccupied for at least one year. Despite the differences in content, all the definitions have in common that they specify a concrete duration, which separates different types of vacancies, further described by an interpretative term.

Numerous studies fall into the third category in which different types of vacancies are named without any further explanation in terms of duration. Rink & Wolff (2015, p. 314), for instance, distinguish between temporary, periodic, or structural vacancy, Hillier et al. (2003, p. 93) add permanent vacancy to the already mentioned temporary and long-term vacancy, and Y. Pan et al. (2021, pp. 7–8) distinguish long-term, short-term, seasonally and occasionally vacant buildings based on a clustering approach. However, none of them add specific temporal boundaries that separate these subcategories. In summary, it can be concluded that information regarding the temporal characteristics of a vacancy can provide valuable insights, e.g., on the consequences or causes of a vacancy.

Since the cause of a vacancy is a crucial aspect when evaluating measures against vacancy, many definitions and descriptions focus thereon and a possibility to distinguish these definitions and descriptions is given subsequently. Due to the large variety of possible causes, the possibilities to structure them are also various. However, for many potential areas of application of vacancy data, it is especially important to distinguish if the vacant properties are intended to be brought back into use or intentionally left vacant. This approach is similar to that of Rink & Wolff (2015, p. 314), who distinguish market-active and non-market-active vacancy, where market-active vacancy denotes vacancy available at the market for use and non-market-active vacancy, which is not. Furthermore, several studies such as the ones by Braun et al. (2014, p. 6), J. Li et al. (2019, p. 5), Spehl (2011, p. 30), and Zhou et al. (2021, p. 3) mention the phenomenon of underutilization and thereby distinguish underutilized units from unutilized units, additionally differentiating the description of vacancy causes and being particularly important for assessing measures to prevent or oppose vacancies. By this differentiation, depicted in Figure 6, a basic classification of vacancies regarding their causes can be made, allowing a primary assessment. However, within these categories, there is still a large number of different possible causes that can be distinguished. Examples of the first category include seasonal vacancy, described by Y. Pan et al. (2021, p. 10), due to tourism or seasonal industries, which encompasses apartments for workers and vacation.

Figure 6

Subcategorization of Vacancy Descriptions by the Cause of the Vacancy



Note. Own research.

These apartments are typically also offered outside the working or holiday season but are not in demand and can therefore be classified as market-active underutilization in terms of the temporal dimension. The second category of unutilized market-active vacancies includes housing units that are vacant for structural reasons, which are taken up by Rink & Wolff (2015, p. 314), but also include the typical case of change of tenant mentioned, for example, by Gentili & Hoekstra (2019, pp. 426–427).

The third and fourth categories of non-market-active vacancies are of particular interest, as they appear unnatural from the perspective of a rational, economically acting person. The usual assumption is that any owner tries to rent out their property if it is vacant in order to maximize profits. Nevertheless, various causes can lead to vacancies without being market-active, e.g., the case of second homes mentioned by Spehl (2011, p. 30) that are not tried to be sublet in vacancy periods. Zhou et al. (2021, p. 3) describe a typical example of a tenant changing homes between workdays and weekends. Since such second homes are sometimes seen as a waste of resources from a societal perspective, China introduced a regulation that attempts to stop this practice (J. Li et al., 2019, p. 3). The fourth category encompasses the vacancy of, in principle, utilizable units that are not in use but still not offered. Knowledge about such vacancies is not extensive, but Gerald (2005, p. 15) mentions investment considerations as possible reasons for such vacancies and Rink & Wolff (2015, p. 314) additionally bring up vacancies due to dilapidation. In the context of vacancies induced by the structural conditions of a building, there are also intentionally induced vacancies to be able to refurbish or demolish properties (Braun et al., 2014, p. 6; Dransfeld & Lehmann, 2007, p. 21).

The categories related to the temporal component or the cause of the vacancy provide information about the characteristics of the building vacancies, whereas a classification of definitions and descriptions characterized by the measurement method adds information regarding the data collection process, thereby simultaneously getting a definitional character. These data collection processes can be divided into general measurement methods, leading to different information that can be retrieved depending on these methods. Since these methods are relevant not only to categorizing vacancy definitions but also to classifying and assessing the quantitative method developed in this study, they are described in more detail in a separate Chapter 2.3.3.

In conclusion, the developed framework for classifying vacancy definitions and descriptions highlights the variety of aspects that need to be considered when working with vacancy figures due to their large potential for deviations that can directly influence derived results.

2.3.3 Methods to Measure Vacancy

The various methods to measure vacancy or vacancy rates have different advantages and disadvantages and thereby influence not only the vacancy definition but also multiple other aspects, e.g., the possible conclusions that can be drawn, the necessary effort, or the typical regional scale of application.

Therefore, for each method, the basic principles on which it is based are explained and further typical characteristics of the methods are mentioned. Those descriptions of characteristics include widely perceived advantages and disadvantages and the typical geographical scale of application since these criteria allow the comparison of different approaches. Examples of scientific literature are given, where those methods are studied or results from those methods are used to illustrate the current state of scientific knowledge regarding each method. Concluding, other noteworthy information about each approach are given, e.g., an outlook on trends in development for new approaches. Finally, a summary is provided as a basis for the justification of developing an additional method in this work. It is important to note that the described methods are not entirely separable in all cases. For instance, the site visit method can contain survey elements, i.e., when the experts conducting the site visits perform additional interviews to validate or enhance their results.

Site visits are a widely recognized method for collecting vacancy data, as can be seen from the studies referring to this methodology analyzed in the following. The fundamental idea of classical site visits is that an expert person performs an inspection of the study area and makes a vacancy assessment of each building based on the visual impression and, if applicable, on further information. The visual impression is created by assessing the building and its surroundings in general, but also by more detailed inspections, which can be looking into windows and assessing indicators such as the presence of flowers, window curtains, or furnishings in general (Dransfeld & Lehmann, 2007, p. 22). Contrary to many of the other examined approaches, the site visit method requires a data collection process specifically for measuring vacancy instead of relying on existing data, which causes advantages and disadvantages.

As the only approach that attempts to observe vacancy directly and not estimate it from secondary data, according to Spehl (2011, pp. 29, 31), it generates the most precise information. This finding is in line with the findings of Hillier et al. (2003, p. 102), who found that the results of site visits contributed the most to their model in predicting housing abandonment. The most obvious disadvantage is the high effort required for the data collection.

Additionally, the data collection accuracy depends on the local and subject-specific expertise of the person conducting the site visits and the attempts of the owner to hide the occupancy status, as simple actions, such as leaving window shutters down, can easily affect the possibility of an assessment. In addition, site visits cannot provide more detailed information about the characteristics of vacancies, such as duration or cause (Dransfeld & Lehmann, 2007, pp. 22–23), which other methods are capable of providing, such as the use of accurate utility data being collected repeatedly and containing time stamps.

Due to the high effort of this method, the method of estimating vacancy using site visits is often applied at a smaller scale in terms of regional distribution or market coverage, i.e., at the level of small cities or city blocks or on a sample basis when applied at larger levels, as illustrated by the examples below. Typical examples of using site visit data to estimate vacancy include the American Community Survey (ACS). The U.S. Census Bureau, compiling the ACS, describes the ACS as the cornerstone of its efforts to provide continuously updated and relevant data on population and housing characteristics, highlighting the overall relevance of the ACS and justifying the significant effort invested in generating reliable vacancy numbers (Torrieri et al., 2014, pp. 1, 85). Another example is a study by Hillier et al. (2003), who used data on building vacancies from site visits and other information from the Philadelphia Neighborhood Information System to build a model that predicts the likelihood of building abandonment. In addition to the beforementioned applications, Kumagai et al. (2016, p. 711) also conducted site visits to collect information on vacancies and validate their approach of estimating vacancies from utility data, thereby implicitly acknowledging the quality of site visits.

In addition to the previously described classical approach, new approaches have emerged with improvements in certain parts of the method. Konomi et al. (2017, p. 133), for example, developed a method that detects WiFi signals, which can be collected by smartphones that have been specifically prepared, instead of the more effortful individual evaluation by visual inspection. Following a similar line of thought is the proposal to use smartphones to capture images and geocode them to make the data collection process more efficient mentioned by Spehl (2011, p. 31). However, such adjustments to the method require careful consideration, as they can have the disadvantage of decreasing accuracy.

Already mentioned in the context of the study by Kumagai et al. (2016, p. 711), vacancy estimations can also be based on utility company data. For billing purposes, water and electricity suppliers use meters, which collect consumption data from their customers to determine the amount payable based on the amount of water or energy used by the customer. Therefore, these data are already available and it can be assumed that the occupancy of a

building should correlate with power and water usage. According to Dransfeld & Lehmann (2007, p. 23), two approaches can be considered for evaluating vacancies. On the one hand, the number of registered meters can be compared with the number of deregistered meters, and on the other hand, consumption data can be analyzed (Dransfeld & Lehmann, 2007, p. 23). Analyzing consumption data is based on the argument that there should be little or no water and electricity consumption in vacant apartments and, conversely, apartments with significant electricity and water consumption are in use (Zhou et al., 2021, pp. 4–5).

Therefore, the advantages include the low cost and low time consumption compared to site visits or surveys, as well as the high accuracy of the collected data compared to other methods with good scalability, e.g., remote sensing (J. Li et al., 2019, pp. 2, 10). Additionally, compared to methods based on observations at a single or a few points in time, such as the census method, it offers the advantage that vacancies can be evaluated in a temporal context, as consumption data are collected regularly (J. Li et al., 2019, p. 10). These advantages make methods based on data from utility companies a frequently used approach (Spehl, 2011, p. 32). In contrast to those advantages, there are two major disadvantages.

First, there are problems concerning a complete and consistent coverage of data from utility companies for the defined research area, as integrating such secondary data from often multiple sources can be a major challenge (Hillier et al., 2003, p. 92). In particular, this problem arises due to the heterogeneous landscape of utility companies (Spehl, 2011, p. 32) and their willingness and ability to provide reliable data (Hillier et al., 2003, p. 95). Furthermore, contradicting business interests and differences in data collection procedures contribute to the difficulties of using these data. The differences in the data collection process can range from deviating electricity and water meter reading dates to deviating data management solutions and, thus, different data export and data format options (Hillier et al., 2003, p. 95). However, it is particularly the complete integration of these data that is necessary, as Bangert et al. (2006, p. 178) formulate the full coverage of all electricity or water meters in the study area as an essential requirement for a correct methodological approach. Additionally, utility company data may sometimes be unreliable depending on the circumstances of their collection. Kumagai et al. (2016, pp. 710–711), for instance, describe a situation where water meters are sometimes not excluded from the evaluation even if the house no longer exists because of missing incentives to deregister the meters officially.

Second, for reasons closely related to the different types of vacancy definitions, the estimated amount of vacancy may be biased in either direction. On the one hand, underutilized units, e.g., for demographic reasons or due to the availability of a second home, could lead to an

overestimation of vacancy if underutilized units are not included in the deployed vacancy definition (Spehl, 2011, p. 32). On the other hand, permanently vacant units that consume water and energy, e.g., for maintenance purposes, could lead to an underestimation of vacancy. Finally, one must find a threshold or decision rule as a decision criterion to determine whether the unit is in use. Haramati & Hananel (2016, p. 111), for example, use a decision rule that assumes vacancy if there is a use of less than three cubic meters of water in each two-month period during six consecutive months. If such a decision rule is not to be assumed but derived from empirical data, e.g., through additional field surveys, questionnaires, in-depth interviews, or by applying Big Data analysis methods such as unsupervised learning to the existing data, this can be challenging and time-consuming, reducing the benefits of the low effort of using this method (J. Li et al., 2019, p. 4; Zhou et al., 2021, p. 5).

Due to the before mentioned advantages and disadvantages of the approach of using data from utility companies, it is suitable for application at a medium scale in terms of regional distribution or market coverage, i.e., at the level of a city rather than at the level of an individual city block or country. Examples of using supplier data to approximate vacancy include the study by Bangert et al. (2006) for Berlin or the study by J. Li et al. (2019) for the Songshan District of Chifeng City using electricity consumption. Besides electricity consumption, the work of Haramati & Hananel (2016) for Tel-Aviv and Jerusalem, as well as the work of Hillier et al. (2003) for Philadelphia and the investigation of Y. Pan et al. (2021) for Changshu use water consumption data. In principle, it is also conceivable to use other types of utility company data than water and electricity consumption for estimating vacancy rates, as long as the data allow the conclusion that little or no usage indicates vacancy. Such data could include data from telecommunications providers (Kumagai et al., 2016, p. 709) or waste disposal companies. However, because of their lower accuracy compared to water or energy consumption data, they should rather be used as supplementary data than as stand-alone data because they are inherently more unreliable, e.g., since waste containers in multifamily buildings are less likely to be household-specific than water or electricity meters (Spehl, 2011, p. 32).

Due to their regular application in censuses, survey-based methods are well-known for vacancy data collection. The different types of survey-based methods for determining vacancy have in common that a sample of building owners are questioned about the occupancy status of their buildings. Besides this commonality, surveys for determining vacancy can be conducted in different variations, e.g., regarding the scope of the survey and the methodology used. The scope can vary in terms of the study area by regional level from small to large areas and in terms of the sample size by the proportion of the population surveyed. The methodology can

vary from paper-based questionnaires to online or interview-based surveys, and participation can be voluntary or mandatory.

Using the example of vacancy estimation in the U.S., it can be seen that survey-based approaches are widely used as they include several official statistics, such as the decennial census, the ACS, the Current Population Survey (CPS)/Housing Vacancy Survey (HVS), and the AHS. The following table provides an overview of these different statistics in terms of the sample size, the obligation to cooperate, and the type of survey and thereby provides an impression of the effort required and information available.

Table 1

U.S. Surveys Collecting Vacancy Data

Variable	Decennial Census (short form)	ACS	CPS / HVS	AHS
Sample Size	Full census (140,498,736 housing units in 2020)	~ 250,000 addresses per month	~ 72,000 addresses per month	~ 120,000 addresses in 2019 (varying/budget dependent)
Mandatory/ Voluntary	Mandatory	Mandatory	Voluntary	Voluntary
Type of survey	Telephone/personal-visit interviews Internet/ paper-based questionnaires	Telephone/ personal-visit interviews	Telephone/ personal-visit interviews	Telephone/ personal-visit interviews

Note. Own research. The information depicted are retrieved from U.S. Census Bureau (2011, 2020).

The considerable effort of multiple surveys illustrates, in addition to the arguments elaborated in Chapter 2.3.1, the relevance attributed to vacancy data collection by governmental institutions. Advantages include that the survey method is already long-established and accepted in real estate research for collecting vacancy data (Charles et al., 1991, p. 159). In Germany, for example, vacancy data collection dates back to the 1950s (Rink & Wolff, 2015, p. 318). The principally straightforward idea of collecting the occupancy status of a building by questioning the owner makes this approach an obvious possibility. In the case of governmental

surveys, the possibility of mandatory implementations can lead to a market coverage that can be better compared to other methods, such as the use of real estate company data, which is inherently restricted by the market coverage of the real estate company.

The beforementioned advantages are countered by the relatively high effort involved (Z. Chen et al., 2015, p. 2188), which becomes evident when compared to methods where already existing data can be used, such as data from utility companies or real estate companies. In particular, the high personnel and infrastructural effort required (Charles et al., 1991, p. 159), also caused by the fact that household interviews are often necessary, in addition to paper surveys (Kumagai et al., 2016, p. 709), are accompanied by high costs and thereby contribute to the disadvantages of the method. This required effort contributes to increasing time intervals between surveys, resulting in data with low timeliness, criticized in the scientific literature (Bangert et al., 2006, p. 178), or to reducing sample sizes. Since the number of vacant buildings is generally relatively small compared to the total building stock, sampling errors are a relevant problem, especially if the sample is based on voluntary participation, which is prone to the problem of self-selecting participants. This problem is mentioned by J. Pan & Dong (2021, p. 360), as well as by Molloy (2016, p. 122), who even decided to use USPS data in addition to AHS data for this reason. Besides sampling problems, Gentili & Hoekstra (2019, p. 443) highlight the importance of the type of questions asked.

Because of its widespread use as a federal census, the scale of a typical vacancy survey is often the national level. Nevertheless, micro-census surveys cover smaller subsections of areas, e.g., several or individual cities. Examples in which vacancy data, determined by governmental surveys, are used are numerous and include recent studies from various countries. These examples include the study by Monkkonen (2019), who describes the role of government mortgage lending on vacancy rates in Mexico, and the study by Demers & Eisfeldt (2022), who use AHS vacancy data in their analysis of total returns for single-family rentals in the U.S. For the Asian market, similar data is collected and Zhang et al. (2016) use data from the China Urban Household Survey, which can be considered the Chinese counterpart to the U.S. CPS, to examine the impact of income inequality on the housing vacancy rate. Less common, but not precluded, as shown by the example of Charles et al. (1991, p. 159), is the use of vacancy data collected through nongovernmental surveys. In this example, vacancy data from Coldwell Banker's are used for examining the relationship between inflation and commercial real estate performance. The data consists of yearly national vacancy rates estimated from smaller samples collected in different cities (Charles et al., 1991, p. 159), showing a typical problem of nongovernmental surveys. Since there is no public funding for those surveys, there is an even

higher incentive to reduce costs by collecting small samples, leading to sampling-associated problems.

In conclusion, it can be stated that there is widespread use and, thus, high relevance of survey-based methods in estimating vacancy. In this context, it should be noted that the definition of survey-based vacancy estimation is not conclusive and different approaches fall into the category of survey-based methods, e.g., there is overlap between survey-based methods, on the one hand, and other methods, such as site visits, on the other.

The method of collecting vacancy data by involving mail carriers takes advantage of the fact that they are one of the few groups of people who see the vast majority of all residential buildings in a country with high frequency. Implemented in the U.S., buildings are recorded as vacant in the USPS address database if the mail has not been collected for 90 or more days (Office of Policy Development and Research, n.d.). Additionally, so-called no-stat addresses are recorded. These addresses include addresses where the USPS anticipates that mail will not be delivered in the foreseeable future for various reasons, such as ongoing construction works or an uninhabitable building condition. This distinction underscores the primary purpose of this recording, which is to avoid accumulating mail or redundant return transports, which is not always in line with a scientific understanding of vacancy, but may be helpful for specific use cases (Office of Policy Development and Research, n.d.).

Besides the obvious advantage of using data that are already available, there are further advantages and disadvantages mentioned in other studies. K. Wang & Immergluck (2019, p. 515) emphasize the broad coverage of the USPS data set, combined with the more frequent collection compared to the decennial census or the ACS. Molloy (2016, p. 122) points out the fine geographic granularity that allows conclusions to be drawn at the small-scale level, and besides such quantitative aspects, there is also a quality assurance of the data, as the address database is routinely audited and maintained (Whitaker & Fitzpatrick IV, 2013, p. 85).

Criticism of the method using mail delivery ranges from fundamental rejection based on assumed unsuitability (Spehl, 2011, p. 30) to detailed criticism of the case-specific implementation of the method. Molloy (2016, p. 122) recognizes the problem that vacancies collected by this method do not conform to traditional vacancy definitions because households using PO boxes would be recorded as vacant on the one side, and vacancies would be detected delayed due to the 90-day time limit on the other. This criticism highlights the importance of the method of vacancy measurement as a separating dimension in categorizing vacancy definitions and descriptions in Chapter 2.3.2.

The typical scale at which this method is applied can be derived from the geographic area for which a postal service is typically responsible, in many cases nationwide, such as in the U.S. or Germany. Examples of using estimations from mail carriers to approximate vacancy include the study by Newman et al. (2016, p. 143) that examines whether urban expansion contributes to higher vacancy rates or the study by Whitaker & Fitzpatrick IV (2013, p. 79) that describes spillover effects of vacant properties on values of neighboring properties. Conclusively, a study by Molloy (2016, p. 119) specifically summarizes and points out the potential of USPS data in examining the geographic distribution of vacancies at the national level, which requires both detailed geographic information and a large-scale database. Similar requirements can be seen in all of the above examples, which make them predestined for using vacancy estimates from mail carriers due to the previously mentioned remarks.

Another method which is making use of the existence of data that are already available is the use of real estate company data. Different types of companies operating in the real estate sector have individual knowledge of real estate submarkets in general and of specific parts of those submarkets in particular, based on their business activities and experience. In many cases, directly or indirectly, vacancies play a role in their operations, and they are in a position where they could provide information thereon.

The advantage of using already existing data are supplemented by the type of data basis, often consisting of lease contracts or rental offers, that may be very detailed and may contain information that are usually not published. Additionally beneficial is that companies providing such data can be assumed to have real estate sector-specific or area-specific expertise and are therefore well-suited to validate the results of such estimates. Besides these positive effects associated with the provision of vacancy estimates by real estate companies, there are also limitations. The published data are limited regarding the real estate sector and area, which is usually restricted to their sectoral and regional area of operation. Furthermore, these assessments are based on sampling, as companies regularly do not have access to information representing the entire building stock. Finally, only a share of these information is freely available, e.g., market reports by brokerage houses like JLL (2022b), but others are not, e.g., the CBRE-empirica-vacancy-index (empirica ag, 2021, p. 8).

The methodology is not intrinsically specific to any particular regional level since that level usually does not affect the effort required. However, its implementation requires that the companies have the necessary data, which are usually limited to certain areas. Examples utilizing vacancy estimations based on data from real estate companies include studies based on the CBRE-empirica-vacancy-index, which is used regularly for examinations regarding the

German market, e.g., in a general market study by Just et al. (2017, p. 30). This index is provided by a cooperation of the companies CBRE, a commercial real estate company focused on real estate service and real estate investment, providing the data, and empirica, a company focused on economic and social science research and consulting, extending and analyzing these data. To summarize, using data from real estate companies is a valuable addition to existing methods due to its possibility of using regularly undisclosed information. However, the restricted data availability of each company simultaneously is a downside of this approach.

Especially well-suited for large-scale vacancy estimations is the use of remote sensing data, as they are available on much larger scales than all other previously discussed data sources. The fundamental idea of remote sensing data in vacancy research is that active human inhabitants and their social and occupational activities lead to nighttime light, which is less or not present in uninhabited areas. A high correlation between active inhabitants and their social and occupational activities, which reflects itself in different levels of nightlight brightness, is already proven (L. Wang et al., 2019, p. 8568). The intensity of this nightlight can be measured from satellites by collecting information on the electromagnetic waves in the visible near-infrared range (L. Wang et al., 2019, p. 8568). These information can then be evaluated and used as a vacancy estimator by combining them with additional data such as known urbanized areas (Z. Chen et al., 2015, p. 2188).

Advantages of this method include that using remote sensing data does not generate costs that are similarly high as those of other methods (Z. Chen et al., 2015, p. 2188; L. Wang et al., 2019, p. 8568), especially compared to site-visits or surveys. Additionally, remote sensing data offers the possibility to provide a detailed spatial distribution of the estimated vacancy compared to methods that use spatially aggregated data, e.g., on the city level (Z. Chen et al., 2015, p. 2188). On the other side, it can be criticized that the use of remote sensing data is less accurate than other methods since inaccuracies result from different problems, e.g., non-residential light sources such as cars or streetlamps or the sometimes subjective choice of reference areas for light emission for algorithm training purposes (L. Wang et al., 2019, p. 8569). One approach to address accuracy issues could be to use finer-resolution data (Zheng et al., 2017, p. 120).

The typical scale of current applications varies from individual cities (Du et al., 2018, p. 4) to multiple metropolitan areas (Z. Chen et al., 2015, p. 2189). Thus, the hope for a method for collecting vacancy data with a low effort at a very large scale is not yet fulfilled, which is also reflected in the focus of these studies that still try to improve the accuracy. These studies include, for example, the work by Z. Chen et al. (2015, p. 2188), who combine nighttime light data with land cover information for fifteen U.S. metropolitan areas, and the study by Zheng et

al. (2017, p. 112) who combine nighttime light data with information about the population and land use for the Yangtze River Delta.

The previously presented methods include widely accepted and widely used approaches. Nonetheless, this list is not exhaustive and there are other less common approaches to measuring vacancy, e.g., using rental bond data collected due to government regulations in Australia (Wood et al., 2006, pp. 445–446). For the example of Germany, Spehl (2011, pp. 31–32) and Dransfeld & Lehmann (2007, p. 22) mention the possibility of calculating the vacancy rate from the difference between the estimated number of households and the estimated number of existing housing units.

The previously given descriptions of the approaches to measuring or estimating vacancies reveal their advantages and disadvantages. Except for the site visit method, most methods can be criticized for accuracy (Spehl, 2011, p. 32). This downside can be improved by combining multiple methods (Spehl, 2011, pp. 32–33, 60–62). One example of this approach is the study by Zhou et al. (2021, p. 4), who take advantage thereof by combining utility, night light, and governmental data. Furthermore, they apply land use data to increase accuracy, as the information that urban residential areas are typically less light emitting per person than commercial areas can be used by including land use data (Zhou et al., 2021, p. 17). Therefore, it can be concluded that any additional method that has advantages in a specific characteristic, e.g., requiring less effort compared to other methods, particularly has the potential to contribute to the increase of the overall accuracy without itself being more accurate than other methods.

Besides the general criticism concerning the accuracy, another aspect that pertains to all methods is the problem of the high effort required for detailed vacancy estimations at a large scale. According to the analyzed literature, no existing method can detect housing vacancy at a large scale with high accuracy and low effort simultaneously. As a consequence of those problems, the overall data situation remains inadequate with respect to comprehensive and detailed data provision, also for well-studied real estate markets, like the German and the U.S. market (BBSR, 2017, p. 4; K. Wang & Immergluck, 2019, p. 512). To conclude, any additional method of collecting or estimating vacancy data has the potential to improve the vacancy data basis. In particular, methods capable of capturing vacancy on a large geographic scale requiring low effort are needed.

2.3.4 Identification of Potentially Relevant Data

The previous chapters have demonstrated the importance of vacancy data for several reasons. The differences in vacancy definitions stem from the different uses of vacancy data, which require different relevant information, and from the different methods of vacancy measurement, which tend to dictate the vacancy definition based on their method of vacancy measurement, usually making it impossible to derive information on different types of vacancies from one measurement. Correspondingly, the variables commonly used to estimate vacancy are also diverse.

To provide an overview of commonly used types of variables for vacancy estimation that can serve as a basis for the variable selection of this study, commonly used variables are grouped into classes that contain variables with similar thematic priorities and are depicted in Table 2. As these variables are used in different research areas, contexts, and with different models, no conclusions should be drawn for this research regarding their particular relevance and values. However, the set of variables can serve as an indication of which variables could be included in vacancy models in general. This set of variables includes locational variables that are widely accepted to play an essential role in influencing real estate markets in general, which also applies to their influence on the vacancy rate.

Examples of location-describing or generally location-related variables in vacancy research are diverse and include the elevation (J. Li et al., 2019, p. 9), the rate of cultivated land divided by total land (J. Li et al., 2019, p. 9), and the percentage of rental housing in the respective area (Bentley et al., 2016, p. 10). Furthermore, Du et al. (2018, p. 14) specifically mention the relationship between locational attractiveness and vacancy and therefore specify variables intended to measure the attractiveness of the location, such as the number of parklands, the amount of industry, and the percentage of residential space.

Table 2

Variables Used in Vacancy Research Models

Variable group	Exemplary variables	Sources
Locational	Distance to urban center	J. Li et al. (2019, p. 10)
	Elevation	J. Li et al. (2019, p. 9)
	Share of residential area	Du et al. (2018, p. 14)

Variable group	Exemplary variables	Sources
Economic	GDP	L. Wang et al. (2019, p. 8582)
	Normalized housing prices	L. Wang et al. (2019, p. 8582)
	Rent	Baba & Shimizu (2023, p. 35)
Socioeconomic	Unemployment rate	Bentley et al. (2016, p. 10)
	Household income	Bentley et al. (2016, p. 10)
Sociodemographic	Population	L. Wang et al. (2019, p. 8582)
	Development of population	Deilmann et al. (2009, p. 667)
	Age	Bentley et al. (2016, p. 10)
Property characteristics	Plot area	Yue et al. (2022, p. 12)
	Floor area ratio	Yue et al. (2022, p. 12)
	Number of bathrooms	Nadalin & Igliori (2017, p. 3096)

Note. Own research.

The attractiveness is typically reflected in house prices, which belong to the group of economic variables that also include the Gross Domestic Product (GDP), for example, used to estimate vacancy in the work by L. Wang et al. (2019, p. 8582). In addition to house prices, rents reflect the attractiveness or demand for residential buildings, and recent research examines the relationship between vacancies and rents in detail (Baba & Shimizu, 2023). However, this relationship has been known for a long time, and its importance has already been demonstrated in the past, for example, by the work of Rosen & Smith (1983), suggesting that the information on rent levels contained in ORL data could have explanatory power for estimating the vacancy rate.

Furthermore, Bentley et al. (2016, p. 10) integrate closely related socioeconomic variables, such as household income and unemployment rate. Since these variables are closely related, integrating both must be carefully considered due to the risk of multicollinearity. A fundamentally similar approach is incorporated by including sociodemographic variables, which are considered generally important for the estimation of vacancy by Kumagai et al. (2016, p. 710), and which are included by L. Wang et al. (2019, p. 8582) in the form of the population as an explanatory variable. In addition, Deilmann et al. (2009, p. 667) describe that not the population but, in particular, the development of the population, influences the vacancy rate. Bentley et al. (2016, p. 10) show that it is not only the population as a whole that matters in this context but also the characteristics of the population, such as age and ethnic groups.

In contrast to the other vacancy-explaining variables, which focus on more general descriptive characteristics of the location and its residents, Yue et al. (2022, p. 12) found that variables directly describing the properties like the plot area and the floor area ratio significantly influence the housing vacancy rate. The influence of such property-specific variables is also confirmed by Nadalin & Igliori (2017, p. 3096), who found a relationship between the number of bathrooms and the vacancy rate. Finally, it can be summarized that the variety of variables used in models to estimate vacancy is large and can thereby serve as a good indication of which variables to include in vacancy estimations.

2.4 Derivation of Research Questions

The literature review highlighted the importance of real estate market transparency, the potential of ORL data, the need for further research on vacancy rates, and the interrelationships between these topics. Despite the importance of real estate market transparency, particularly in the context of vacancy research and the extensive literature on each of the topics, there is a lack of an approach that fully explores and exploits the potential of ORL data in the transparency and vacancy context. Thus, this study aims to connect these topics and explore the contribution that ORL data can make toward increasing transparency in the real estate market by explaining vacancy rates. In order to answer this overarching research question, it is subdivided into further operationalizable and concrete questions. These questions also allow for structuring the subsequent work and selecting the research methods to be used.

While ORL data are already widely used in scientific studies, there is a lack of a comprehensive, fundamental, and generalizable assessment of this data basis, which leads to the first concretizing research question, which is also the basis for assessing the transferability of the results to similar problems:

1. How can online real estate listing data be assessed in terms of quality and quantity?

The second and the third research question focus on the example of vacancy rates and their answers need to integrate insights gained from addressing the first research question. Since no research linking ORL data to vacancy could be found, the first step is to examine correlations that could be caused by relationships and answer the question:

2. Which correlations and possible relationships exist between online real estate listings and vacancy rates?

Based on the identification of these correlations and possible relationships, they can be further evaluated and their potential utility regarding the explanation of vacancy can be assessed by answering the third research question:

3. To what extent can online real estate listings contribute to the estimation of vacancy rates?

3 Data

Due to the centrality of data in this study, an explanation of the data collection, a description of the raw data, an explanation of the data preprocessing, and a description of the preliminary analysis data is provided. Particular emphasis is placed on the aspects related to ORL data, as these specifically form the basis for investigating the research questions and answering the first research question.

3.1 Data Acquisition and Raw Data Description

The documentation of the data collection process ensures the traceability and assessability of the transferability of the creation of the data basis, based on which the data are described in detail. The thorough examination of the data enables the first research question to be answered and builds the necessary foundation for model building (Chatterjee & Hadi, 2015, p. 101), which is needed to be able to answer the second and third research question. Depending on each variable, typical graphs and univariate statistical methods such as box plots and histograms (Fox, 2016, pp. 28–40), medians and means are used to describe the data and thereby provide a general overview of its overall characteristics and distributions. From these descriptions, it can be deduced whether methods of data preparation, such as data transformations or outlier cleansing, are to be used (Chatterjee & Hadi, 2015, p. 101) and which variables should be included in the construction of the model (Bursac et al., 2008).

3.1.1 Web Scraping

Considering Chapter 2.2.1, there are different possibilities for deriving ORL data for scientific purposes. For the German market, neither comprehensive ORL data sets provided by online real estate platforms nor APIs for convenient access of ORL data were available at the beginning of the research underlying this study. Hence, an approach relying on an individually developed automated data collection was selected. This approach offers additional benefits, including independence from data providers and the ability to transfer it to other comparable applications, as it is universally adaptable.

The goal of the data collection was to collect an extensive data set, including as much detailed information as possible from each listing, as presented on the website. First, to assess the quantity and quality of the data, which is relevant to answering the first research question, and

second, to use this information to assess its potential explanatory power for the vacancy rate, which is relevant to the second and third research question. Due to the previously formulated research objective and the required amount of listings, manual data collection is impossible and an automatized solution is needed, which can be implemented by web scraping.

Web scraping describes the process of extracting information from web pages in a targeted manner (Uzun, 2020, p. 61726) and is not to be confused with web crawling or web mining, which are terms sometimes used in similar contexts (vanden Broucke & Baesens, 2018, p. 155). While web scraping describes programs that automatize the extraction of information from HTML code by parsing the web pages data and further processing, e.g., converting and saving them in a structured way (Malik & Rizvi, 2011, p. 467), web crawling instead describes programs that also perform automatable tasks on the web, such as indexing (Kobayashi & Takeda, 2000, p. 153). However, in contrast to web scrapers, web crawlers particularly have the ability of autonomous web page navigation, sometimes including the ability of undirected searches without an explicit purpose (vanden Broucke & Baesens, 2018, p. 155). Web mining is a more general description of the process of the steps of discovering data, automatically extracting specific information, uncovering general patterns at individual websites and across multiple sites, and analyzing the mined patterns (Etzioni, 1996, pp. 65–66; Kosala & Blockeel, 2000, p. 2).

A topic regularly related to these approaches is legality, as particularly web scraping is often considered a legally gray area (Mitchell, 2018, p. 215; Rat für Sozial- und Wirtschaftsdaten, 2019, p. 5). Generally, it is assumed to be legal (Mitchell, 2018, p. ix) and for the case of scientific use in Germany, the prevailing opinion based on the German Act on Copyright and Related Rights (UrhG[11]) is clear towards web scraping being a legitimate practice (§60d UrhG). Nevertheless, during the design of a web scraping process, the consequences of the specific implementation should always be considered. For websites from market-leading online real estate platforms, it can be assumed that they are designed for and used to high traffic, so it is unlikely that the increase in traffic is noticed at all (Mitchell, 2018, pp. 267–268). Nevertheless, scraping was conducted when low utilization was assumed to minimize the effects on the website by avoiding additional traffic during times of high utilization.

Described slightly differently by several authors, e.g., Milev (2017, pp. 480–481), Krotov & Silva (2018, p. 2), and Krotov & Tennyson (2018, p. 170), the development and application of a web

[11] German Act on Copyright and Related Rights in the version promulgated on September 9, 1965 (Federal Law Gazette | page 1273), last amended by Article 25 of the Act of June 23, 2021 (Federal Law Gazette | p. 1858)

scraper can be subdivided into several steps, the most relevant for this research being the analysis of the website, the development of the web scraper, the data collection, and the saving of the data. The analysis of the data source is necessary for multiple reasons. First, the accessibility of the website by automated agents needs to be checked, as different websites require different ways of accessing, e.g., sending specific headers or mimicking browser behavior that can be implemented using different libraries, e.g., urllib or selenium (vanden Broucke & Baesens, 2018, pp. 119–120, 127). Second, the structure of the website needs to be examined to implement web page navigation if multiple pages shall be accessed. Third, the structure of the individual web pages needs to be analyzed to extract the information from the data retrieved through web scraping, as this data is typically only loosely structured since it is embedded in the source code of the web pages (Macias & Stelmasiak, 2019, p. 11).

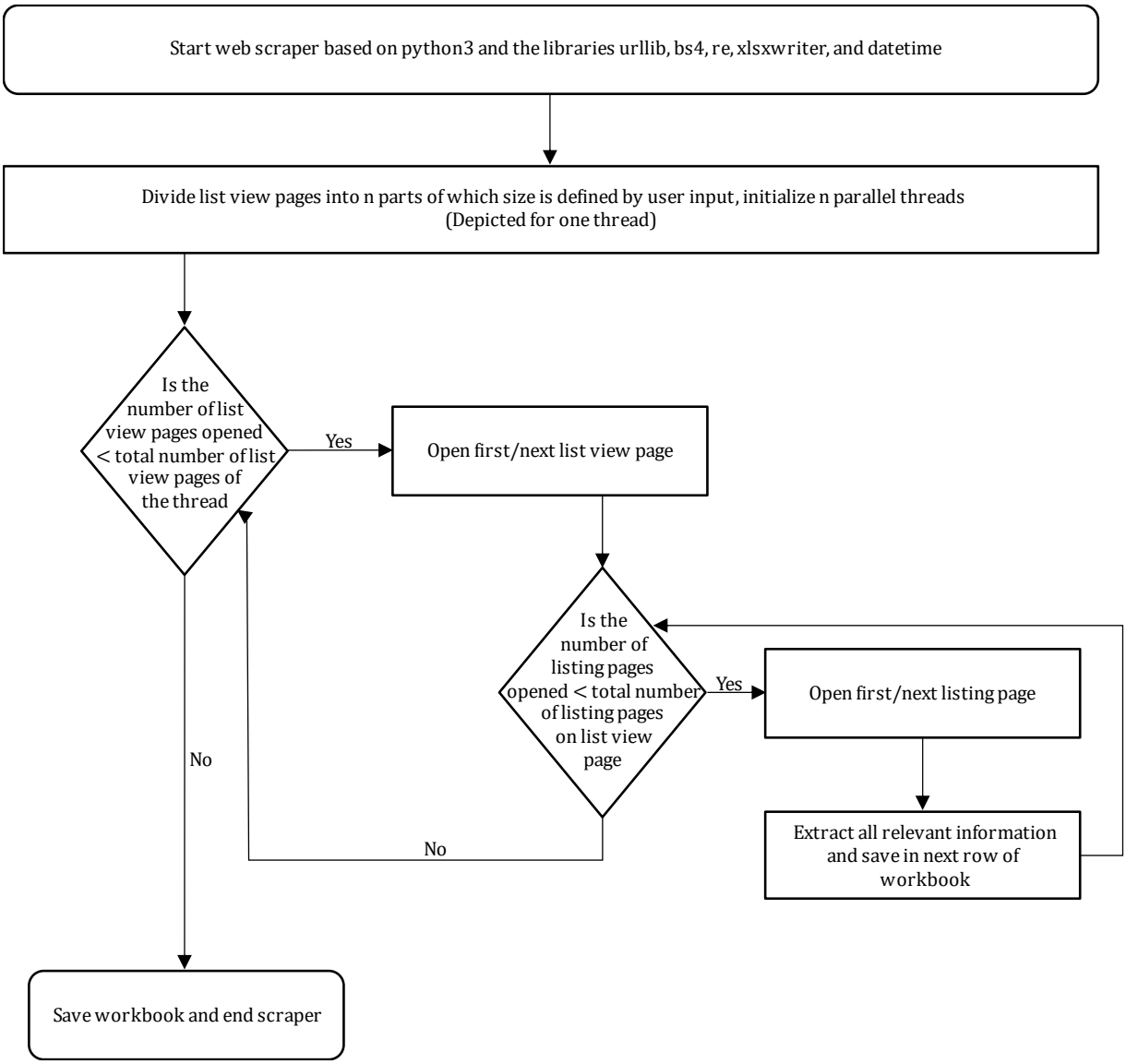
The data collection process of the web scraper developed for this research can be subdivided into downloading the web page information, parsing the downloaded information, and saving the parsed information into a database (Cavallo, 2018, p. 107; Macias & Stelmasiak, 2019, pp. 11–12). The web scraper is based on the programming language python since python is one of the standard languages used in data science and provides the necessary libraries to create a web scraping program (Krotov & Silva, 2018, p. 2; Massimino, 2016, p. 36). The most important libraries used are urllib for retrieving the data, bs4 and re for parsing the data, xlsxwriter for storing the data in a workbook, and datetime for monitoring the entire process. Fortunately, the online real estate platform from which the offers are retrieved provides a structured overview of the listings in a list format containing all individual listings. Unfortunately, this list cannot be opened at once but is divided into many list view pages with a certain number of listings per page. To get all individual listings, it is necessary to open all list view pages and, from each list view page, all individual listings, which are then downloaded and processed further. To speed up this process, the total number of list view pages is, depending on its total count, divided into n parts that are processed in parallel by starting the program multiple times with adjusted user input for the first list view page and the total number of list view pages to scrape.

Figure 7 shows the subsequent process for one of these structurally identical threads. The parsing of the information contained in the downloaded source code of the website is based on the libraries bs4 and re that allow identifying the relevant information by their class and tag names or regular expressions. The result of the process is stored in a standard spreadsheet software file to be checked. Similar to a study by Macias & Stelmasiak (2019, p. 14), error logs are inspected, the sizes of the output files are compared, and a high-level inspection of the

output file is conducted. After this validation, the data are inserted into the database containing all data collections.

Figure 7

Flowchart of Web Scraping Process



Note. Own research.

This process was conducted weekly and consisted of a full sample collection of the apartments category in Germany from a market-leading online real estate marketplace from 01/28/2019 to 08/03/2020. The resulting database includes 7,007,571 listings prior to data cleansing. For

the determination of the frequency of the data collection, different aspects, such as the increasing amount of information, but also duplicates, their processability, and the load on the online real estate platform, have to be weighed against each other. Based on these considerations, a weekly data collection approach is applied, which is also used in other studies, e.g., by Bhuiyan & Al Hasan (2016, p. 471), as this approach is robust to weekly fluctuations. The weekly web scraping was terminated when the ORL platform made considerable changes to the website that would have required the development of a new program that can handle sophisticated bot detection mechanisms and thereby is in the legally gray area.

The amount of data collected is extensive in comparison to other published data collections by web scraping, e.g., the study by Bhuiyan & Al Hasan (2016, p. 471), who scraped 7,216 houses during a period of five weeks, Bernstein et al. (2019, p. 32) who scraped 17,678 rental offerings, or Han & Lee (2018, p. 521) who scraped 13,049 data sets from an online real estate platform.

3.1.2 Description and Analysis of Online Real Estate Listing Data

Due to the approach of a weekly full sample data collection, the raw data set contains a large share of entries with similar information at different points in time, which is specifically addressed in the detailed description of the *Identifier* variable in this chapter. The following description of the raw data focuses on the overall appearance of the data set, such as the completeness of the various individual variables and the overall quality of these variables, to the extent that this can be assessed without access to official validation data. This general evaluation of the raw data forms the basis for assessing the overall quality of the ORL data and is thus the foundation for answering the first research question. To answer the second and third research questions, not all of these data are relevant and included, but only a selected subset, which is then explicitly processed in Chapter 3.2 and described in Chapter 3.3 with regard to the respective research questions. The data collection included 29 primary variables that are depicted in Table 3. The total number of observations containing a numeric or textual specification of each variable is given for each variable. Based thereon, the share of observations containing a specification of that variable is calculated. Conclusively the data type of each variable is specified.

ORLs are typically accessed by users of online real estate marketplaces using a web browser. In order to navigate through the listings that match the user criteria, some type of aggregation is usually displayed, e.g., in the form of a map or a list containing a high-level overview of the filtered listings. In this context, the variable *Title* is essential, as it is conventionally displayed

in these aggregated overviews and at the top of the description of each individual ORL web page. Thus, the incentive to choose a title that generates positive attraction is high. For the collected data set, the specification of a title is mandatory and allows free text entry.

Table 3

Collected Variables

	Observations		
Variable	n	%	Data Type
Title	6,997,464	99.86	Text
ZIP Code	7,007,571	100.00	Integer
Designation of Municipality	7,007,571	100.00	Text
Street	5,219,537	74.48	Text
Cold Rent	7,007,571	100.00	Float
Utilities	6,842,858	97.65	Float
Heating Costs	6,613,618	94.38	Text/Float
numeric specification	2,280,675	32.55	Float
non-numeric information	4,332,943	61.83	Text
Total Rent	7,007,571	100.00	Float
Deposit	6,463,841	92.24	Text
Number of Rooms	7,007,571	100.00	Float
Living Space	7,007,571	100.00	Float
Facilities	6,363,417	90.81	Categorical
Type of Apartment	6,017,616	85.87	Categorical
Floor	5,680,789	81.07	Integer
Year of Construction	5,515,966	78.71	Integer
Condition	5,226,534	74.58	Categorical
Quality	4,067,567	58.05	Categorical
Type of Heating	5,804,190	82.83	Categorical
Energy Source	5,527,478	78.88	Categorical
Energy Performance Certificate	4,433,375	63.27	Categorical
Energy Demand	1,457,099	20.79	Float
General Textual Description	6,477,638	92.44	Text

	Observations		
Variable	n	%	Data Type
Textual Description of Facilities	5,637,519	80.45	Text
Textual Description of Location	5,999,872	85.62	Text
Textual Description of Miscellaneous	4,941,260	70.51	Text
Apartment Provider	6,941,425	99.06	Text
Availability	6,409,828	91.47	Text
Identifier	7,007,571	100.00	Integer
Date of Data Acquisition	7,007,571	100.00	Date

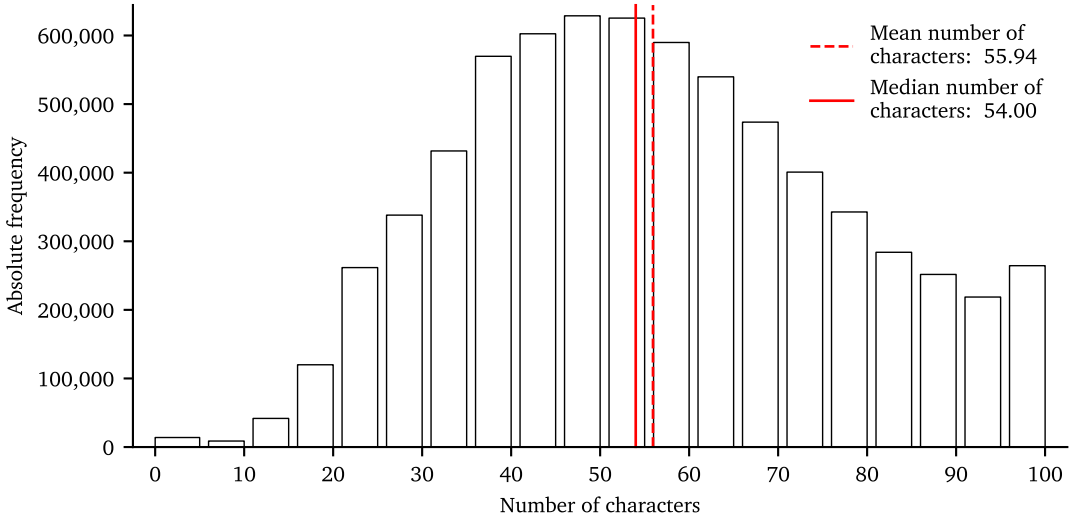
Note. Own research.

For 99.86 % of the observations, a non-empty title could be collected. The 0.14 % of observations with a missing title, i.e., a title length of 0 characters, are due to technical problems during the collection process and concern two collection dates[12]. Figure 8 shows the distribution of title lengths, and while the largest groups contain an average number of characters, the general distribution of the title lengths is broad, with users taking advantage of the full range of possible lengths.

[12] March 4, 2019: 469 observations; March 18, 2019: 9,630 observations

Figure 8

Distribution of Title Length



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 1.

Confirming the incentive to choose a title that generates positive attraction, words that are regularly included in titles are words that either describe positive features of the listing, e.g., 'balcony', 'built-in kitchen', and 'beautiful', or specify fundamental characteristics of the listing, e.g., '3-room-apartment'. The following Table 4 shows the words, in this context, words are defined as strings that are separated by spaces, occurring most often, and a cleaned version thereof, in which words are removed that do not carry a contextual meaning, a process called stopword removal that is regularly applied by the use of customized standard programming libraries such as the Natural Language Tool Kit (NLTK) library, to increase the information density in texts (Sarica & Luo, 2021, p. 1).

Table 4

Most Frequent Words in Titles

All space-separated strings	Observations	Only space-separated strings with contextual meaning	Observations
In (in)	2,782,768	Wohnung (apartment)	1,277,428
Mit (with)	2,690,500	Balkon (balcony)	1,033,777

All space-separated strings	Observations	Only space-separated strings with contextual meaning	Observations
Wohnung (apartment)	1,277,428	Zimmer (room)	640,348
-	1,190,106	2	602,251
Und (and)	1,188,608	3	477,109
Balkon (balcony)	1,033,777	Schöne (nice)	358,798
Im (in)	871,651	Lage (location)	355,493
Zimmer (room)	640,348	Wohnen (reside)	336,877
2	602,251	Einbauküche (built-in kitchen)	276,901
3	477,109	3-Zimmer-Wohnung (3-room apartment)	272,059

Note. Own research. Created from raw data set.

The variables *ZIP Code*, *Designation of Municipality,* and *Street* give information regarding the location of each listing. The specification of a numeric ZIP Code is mandatory, as can be seen by the share of 100.00 % of observations having the variable *ZIP Code* specified that complies with this requirement. According to expectations, the ZIP Codes that appear most often are ZIP Code areas in cities where a higher population density and a higher number of buildings can be assumed compared to rural areas. The most frequently appearing ZIP Codes are those of cities in the east of Germany, where a weak demand and a high vacancy rate can be assumed (Jacobs & Diez, 2018, Chapter 0.2). ZIP Codes that appear least often include ZIP Codes of municipalities with a small number of inhabitants and incorrect ZIP Codes. The ZIP Codes appearing most and least often can be found in Table 5 and in an extended version with the 100 most frequent ZIP Codes in Appendix A - 2.

Table 5

Most and Least Frequent ZIP Codes

ZIP Code	Observations	Municipality	Comment
09130	59,112	Chemnitz	8,4 % vacancy; 246,334 inhabitants
09126	56,402	Chemnitz	8,4 % vacancy; 246,334 inhabitants
09131	46,877	Chemnitz	8,4 % vacancy; 246,334 inhabitants
09112	46,370	Chemnitz	8,4 % vacancy; 246,334 inhabitants

ZIP Code	Observations	Municipality	Comment
09113	43,675	Chemnitz	8,4 % vacancy; 246,334 inhabitants
08056	32,620	Zwickau	12,2 % vacancy; 88,690 inhabitants
06217	31,026	Merseburg	5,8 % vacancy; 33,873 inhabitants
97522	1	Sand am Main	Correct ZIP Code, 3,142 inhabitants
97529	1	Sulzheim	Correct ZIP Code, 2,011 inhabitants
97647	1	Nordheim vor der Rhön	Correct ZIP Code, 1,110 inhabitants
98997	1	Fürth	Generally erroneous ZIP Code
99054	1	Erfurt	Typing error, correct: 99094
09917	1	Chemnitz	Generally erroneous ZIP Code
99687	1	Gotha	Typing error, correct: 99867

Note. Own research. Created from raw data set combined with population and county vacancy data based on the year 2019 (empirica ag, n.d.; Statistisches Bundesamt, 2020).

Despite the requirement for a numeric entry, incorrect entries are possible, as shown in Table 5. Such inputs can be either ZIP Codes that do not exist or ZIP Codes that do not match the specified designation of the city. Examples of non-existent ZIP Codes include those with less than five digits. Examples of listings with shorter ZIP Codes are the ZIP Codes 461 and 0852, which are both the beginning of the correct ZIP Code for their specific listing locations in Oberhausen and Plauen but are missing two additional digits. Besides the wrong number of digits, another common problem is the one of typing errors, e.g., 99687, which does not exist, instead of 99867, which does. These possibilities for error explain why the total number of ZIP Codes in the data set is higher, 10,908 observations, than the number of ZIP Codes actually assigned to municipalities, 8,181 (Deutsche Post DHL Group, 2018)[13].

The variable *Designation of Municipality* provides additional spatial information and can be used to further specify the location or correct errors in the ZIP Code input. Specifying a municipality name is mandatory and allows free text input, resulting in 100.00 % of the observations having a municipality name specified. The most and least frequent municipality names can be found in Table 6, and the 50 most frequent ones in an extended version in Appendix A - 3.

[13] The actual number of different ZIP Codes sums up to 28,278 wherefrom 16,137 are assigned to PO Boxes, 3,095 to major customers and 865 for special uses as lotteries besides the 8,181 ZIP Codes for municipalities.

Table 6

Most and Least Frequent Municipality Designations

Designation of Municipality	Observations
Chemnitz, Sonnenberg	46,450
Chemnitz, Kaßberg	40,623
Chemnitz	34,909
Chemnitz, Schloßchemnitz	33,429
Leipzig	31,898
Chemnitz, Hilbersdorf	31,382
Dresden	30,809
Mühlau b Chemnitz, Sachs, Mühlau	1
Nürnberg - Aussenstadt-Sued, Katzwang, Reichelsdorf Ost, Reichelsdorfer Keller	1
Bergisch Gladbach/Schildgen/Kalmünten, Bergisch Gladbach	1
-- Bitte wählen --, Reinickendorf (Reinickendorf)	1
1, Innenstadt	1
3-Zimmer-Mietwohnung in Peine - Nr. 2730, Peine	1
Berlim, Friedenau (Schöneberg)	1

Note. Own research. Created from raw data set.

Table 6 confirms the results from the examination of the ZIP Codes since the most frequent designations mainly refer to cities in the east of Germany where a weak demand and a high vacancy rate can be assumed. Since cities can be much larger than ZIP Code areas in terms of area, population, and number of residential units, the size of the city becomes more important for the absolute number of listings compared to the number of listings based on ZIP Code areas, where population density and vacancy rate are assumed to be more critical. This influence of the city size can also be seen in Appendix A - 3, which includes designations of the ten largest German cities, measured by inhabitants[14], within the 50 most frequently occurring municipality designations. Combined with the examples of the single occurrences, the disadvantages of free text entry become apparent. One of these disadvantages is that multiple designations are

[14] Berlin, Hamburg, München, Köln, Frankfurt am Main, Stuttgart, Düsseldorf, Leipzig, Dortmund, Essen (Statistisches Bundesamt, 2020)

possible for the same city. The term Berlin appears in 1,047 different designations, and even smaller cities show a significant variance, e.g., the city of Remagen with 27 different designations. Besides many different designations per municipality, incorrect spellings, additional information, and a non-standardized structure contribute to the deterioration of a precisely identifiable spatial localization. Nevertheless, the overall impression of the data set concerning the designations of the municipality is that, in most cases, the designations contain relevant information. However, these information must be extracted from the unstructured data to derive a distinct location of the listing.

The *Street* variable is not mandatory and allows free text entry, usually consisting of textual and numerical characters, with the textual characters specifying the street name and the numerical characters specifying the house number. As the specification of the S*treet* is not mandatory, the number of observations with the variable *Street* specified is significantly lower, with a share of 74.48 % of all observations, compared to the mandatory variables *Title*, *ZIP Code,* or *Designation of Municipality*. One of the reasons for not specifying the exact location by street and house number is that advertising real estate agents try to avoid direct contact between the owner and the potential tenant since this could make some of their services superfluous and reduce their income. The examples in Table 7 show that, apart from not specifying the variable *Street*, there are also apparent misspecifications, for example, when the variable *Street* is misused to provide other types of information. The number of characters ranges from 1 to 104. As can be seen from Table 7, the extreme values of the number of characters are due to erroneous specifications that list additional information in the *Street* variable that do not correspond to the intended use thereof. However, within the major groups of 11 – 30 characters, which include 98.45 % of all listings having a *Street* value defined, most of the observations seem to be correctly defined. Thus, the completeness of the *Street* variable is improvable, and, similar to the variable *Designation of Municipality*, the information must be extracted from the unstructured data.

Table 7

Exemplary Specifications of the Street Variable

Character Count	Observations	Example Correctness	Example
0	1,788,034	-	-
1 – 10	27,582	Correct Incorrect	Hörn 2, xyz 0,

Character Count	Observations	Example Correctness	Example
11 – 20	3,405,785	Correct	Auricher Straße 8,
		Incorrect	Ernst-Eger-Str. -,
21 – 30	1,732,977	Correct	Louise-Schroeder-Straße 11 A,
		Incorrect	Nördliches Paulusviertel XXX,
31 – 40	50,054	Correct	Annette-von-Droste-Hülshoff-Straße 54,
		Incorrect	Einsteinstraße / Steinhauserstraße XX,
41 – 50	2,631	Correct	Bürgermeister-Doktor-Schleicher-Straße 6,
		Incorrect	die genaue Adresse erfahren Sie beim Vermieter -,
51 – 60	407	Correct	-
		Incorrect	Zur Zeit keine weiteren Besichtigungstermine !! 0,
61 – 70	62	Correct	-
		Incorrect	untere mühlengasse ,östliche richtung nicht an der baustelle 2b,
71 – 80	8	Correct	-
		Incorrect	**RESERVIERT** Altstadt Kupferberg Nähe Schillerplatz und Gaustrasse +,
81 – 90	29	Correct	-
		Incorrect	rufen Sie uns an und vereinbaren Sie einen Besichtigungstermin 03 45 78 28 303 1,
91 – 100	1	Correct	-
		Incorrect	BESICHTIGUNG: Freitag um 18.00 Uhr ....Alte Döhrener Straße (links neben Blumen-STANGE) 89,
101 – 110	1	Correct	-
		Incorrect	Diese attraktive, sanierte Wohnung in der achten Etage besticht durch eine gehobene Innenausstattung 0,

Note. Own research. Created from raw data set.

The variables *Cold Rent, Utilities, Heating Costs,* and *Total Rent* give information about the total amount payable to the apartment owner, which consists of different parts for different purposes. The first component from these, the cold rent, which defines the monthly amount demanded for using the offered property without utilities, is given in Euro. The specification of the variable *Cold Rent* is mandatory, as can be seen by the share of 100.00 % of listings having a cold rent defined. Nevertheless, 2,429 observations, representing a share of 0.03 % of all listings, specify a cold rent of €0. Although these listings do not specify a cold rent, most have a total rent of more than €0 defined. Reasons for the specification of a cold rent of €0 in this regard are, for

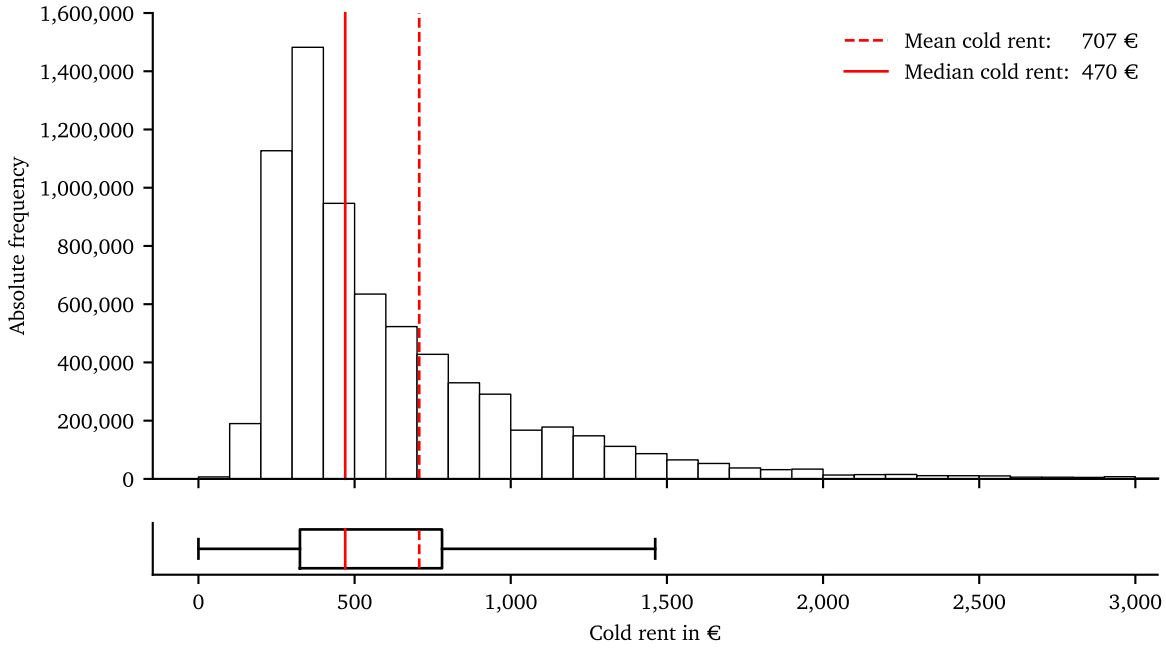
example, the definition of a total rent that is not subdivided into a cold rent component and utilities. Beyond these listings, some listings remain that also define a total rent of €0, which are attributable to reasons such as erroneous user inputs or the deliberate attempt of the landlord or real estate agent to not publicly state the requested rent and thereby disguise the price setting process. On the other side, there are a few listings with very high rents, as the maximum cold rent that can be observed is €111,111,111.11, and 775 listings, representing a share of 0.01 %, have a cold rent defined higher than €10,000. Obviously, rents similar to the maximum rent are erroneous values, which can be confirmed when inspecting their titles and descriptions, including terms like *test real estate* or *test listing*. A few listings with such high rents also appear due to confusion about thousand and decimal separators. Nevertheless, not all of these listings seem to be erroneous. Some listings with rents between €10,000 and €15,000 have detailed descriptions that allow concluding that they describe very luxurious penthouse or urban villa offerings in the best locations of large German cities in high demand like Munich or Berlin.

Due to the clustering of most rents in a range of expected standard rents and a very small share of extreme outliers, it is not possible to depict all listings in an expressive figure, especially with a constant linear scale. Therefore, a combined histogram and boxplot allow both an estimation of the range where most values can be found with the boxplot and a detailed assessment of the distribution in this range with the histogram. Figure 9 shows that the distribution is positively skewed. The observation of the positively skewed distribution fits the falling apart of the mean absolute cold rent and the median absolute cold rent. This skewness is a common phenomenon for asset returns in general (Adcock et al., 2015, p. 1253) and, therefore, also for rents caused by reasons such as the natural lower bound of €0 on the one side and fat tails, especially in combination with a small fraction of severe outliers on the other.

Due to the apparent deviation of the absolute cold rent distribution from a normal distribution, caution should be exercised when interpreting the boxplot, as the share of the data within the range from lower to the upper whisker is smaller, containing 94.51 % of observations, than it would be in the standard assumption of a normal distribution which would include 99.30 % of the observations in this range. Nevertheless, 99.42 % of the collected listings are in a range from €100 to €3,000. This range is wider than the standard range of the whiskers, usually defined by the interquartile range multiplied by a factor of 1.5 and added to each side of the box, thereby defining outliers. In this context, this rule of thumb does not seem to hold, as cold rents higher than the upper end of the whiskers seem to be plausible values and should probably be considered for further evaluation.

Figure 9

Distribution of Cold Rent



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 4.

In summary, it can be stated that a large majority of the cold rent values seem to be in an inherently realistic range and even values at the very lower and very upper end of the distribution can sometimes be considered meaningful. Depending on the intended use of the data and due to the distributional characteristics, a transformation of the variable by the natural logarithm, as is common practice (McMillen, 2008, p. 574), should be considered to better approximate a normal distribution.

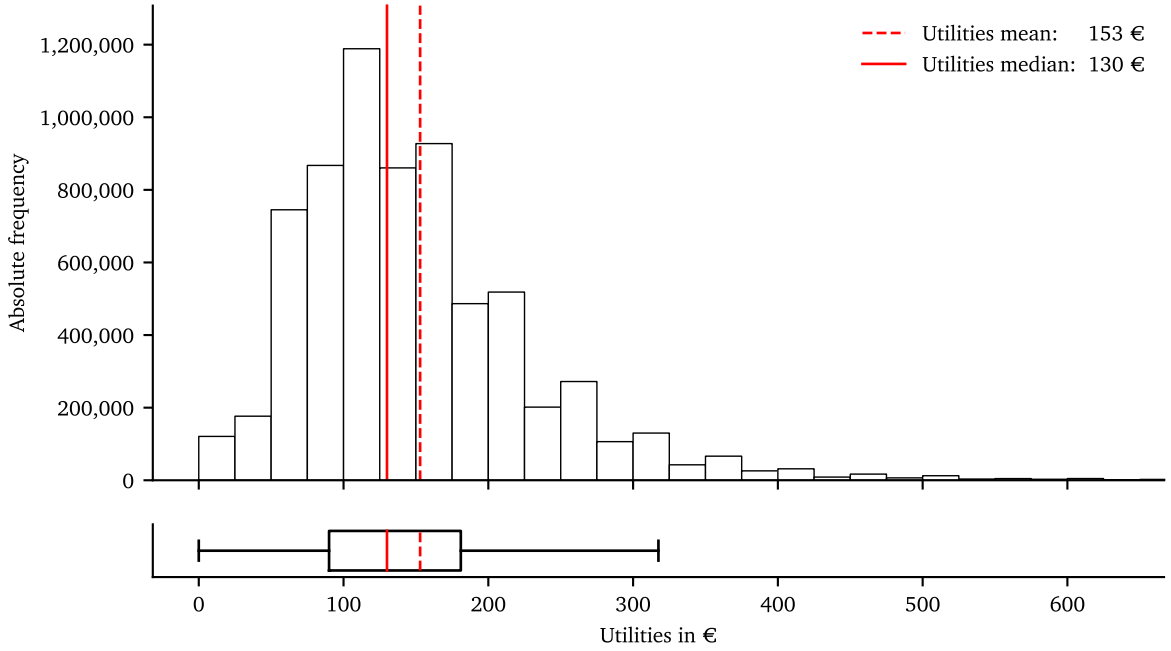
The variable *Utilities* intends to capture the cost positions according to the German Utility Costs Regulation (BetrKV[15]), which has its legal basis in § 556 BGB. According to § 2 BetrKV, several different components can be part of the utilities, including costs for water supply and heating. Typically, the advance payment for utilities, based on estimated utility costs, is reported separately in ORLs. Since it is sometimes common practice to further subdivide the reported utilities into heating costs and other costs, the online real estate marketplace allows reporting

[15] German Utility Costs Regulation in the version promulgated on November 25, 2003 (Federal Law Gazette | page 2346, 2347), last amended by Article 15 of the Act of June 23, 2021 (Federal Law Gazette | p. 1858).

separate heating costs by assigning a separate value to the variable *Heating Costs*. Different users use this option heterogeneously, making it difficult to compare the different values of the utilities. However, the share of listings with a numeric input is the large majority, with 97.65 % of all observations. The remaining 2.35 % of listings explicitly state that they do not specify the cost of utilities. Figure 10 shows that the distribution of the *Utilities* variable has a generally unsuspicious shape.

Figure 10

Distribution of Utilities



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 5.

Expectedly, and for similar reasons as the distribution of the cold rent, the distribution is positively skewed, which can also be observed by the divergence of the mean and the median, although it is closer to a bell-shaped curve. This effect is expected since utility costs do not vary as much with supply, demand, and quality as cold rents but are more closely related to size. Therefore, the share of utilities is typically decreasing for higher total rents and as a consequence, the right tail of the distribution should be smaller in comparison with the cold rent distribution like it can be observed. The number of listings specifying utilities of the

minimum value of €0 at the left end of the distribution is 113,917 and the maximum stated is €10,000,000. The maximum stated is apparently attributable to an erroneous input. However, the share of listings that specify a value for the utilities that seems clearly suspicious is small, as the share of listings that specify utilities that are within the range of the whiskers adds up to 96.20 % and the share of listings that specify utilities, which are in the range from €0 to €500 is 99.37 %. Very high values of utilities above €3,000 are rare and can only be observed for 3,044 listings. Therefore, except for a small share of listings, it can be summarized that most of the utility data appear principally plausible.

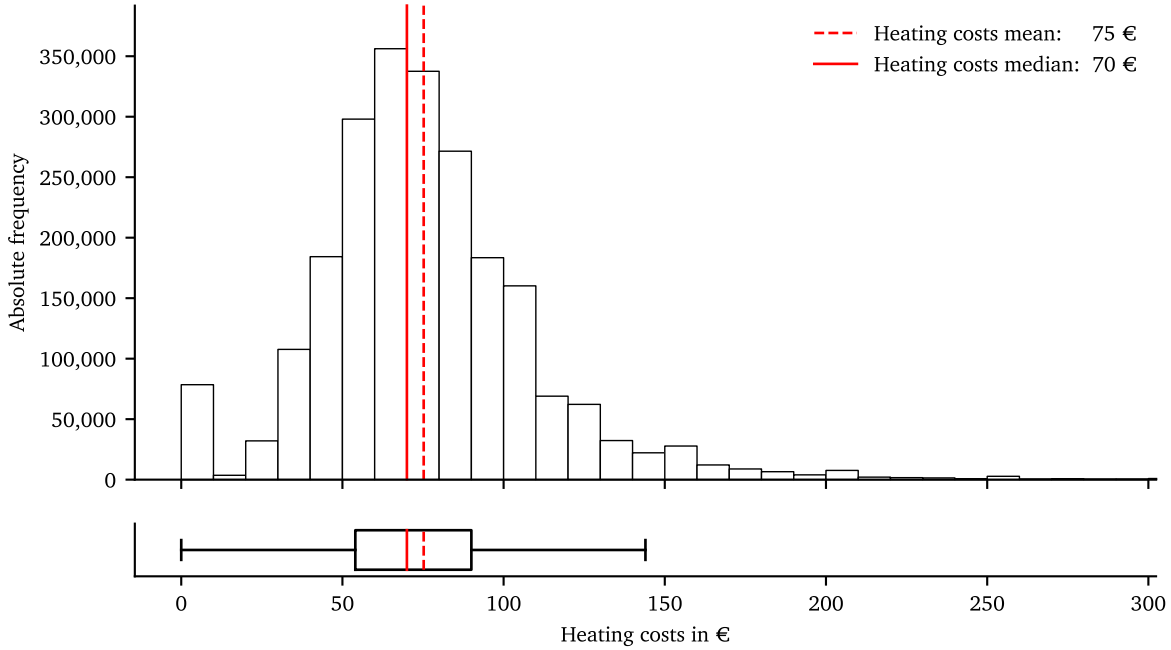
The *Heating Costs* variable is intended to provide specific information on heating costs, often a relevant part of the utility costs. In addition to being a large part of the utilities, heating costs are particularly interesting to a prospective tenant because they are usage-based, vary significantly from tenant to tenant, and can be influenced by the user. The major share of listings, 94.38 %, provide numeric or non-numeric information on heating costs. The remaining 5.62 % explicitly state that they do not specify any information on heating costs.

The listings giving information on heating costs are divided into 61.83 % of all observations giving non-numeric information and 32.55 % of all observations giving numeric information. The listings giving non-numeric information can be subdivided further into 52.16 % of all observations stating that heating costs are included in the given costs for utilities and 9.67 % of all observations stating that heating costs are not included in utilities but are also not further specified.

The listings giving numeric information on heating costs can also be subdivided into a group of listings that solely state an amount of money given in Euro typically payable for the heating costs, which represent 23.05 % of all listings, and a group containing 9.49 % of all listings, which has the additional term *inkl.* that stands for including. However, the possibility of further specifying the *Heating Costs* variable with this term seems to lead to ambiguities. Verification calculations of the given total rent show that in 5.54 % of all observations the augmentation is interpreted as included in utilities and the total rent, and in 3.77 % of all observations as not included in utilities but in the total rent, leaving a share of 0.19 % of all observations where the total rent cannot be calculated in any form taking into account the variables *Cold Rent*, *Utilities,* and *Heating Costs*. However, besides all these limitations regarding the data quality of the heating costs, the general distribution of the *Heating Costs* seems unsuspicious according to Figure 11, which can be confirmed by additional descriptive statistics.

Figure 11

Distribution of Heating Costs



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 6.

The vast majority of specified heating costs lie within a range from €0 to €250 with 99.60 %, wherefrom 1.09 % report heating costs of €0. At the upper end of the heating cost range, 505 observations, or 0.02 % of all observations, have heating costs greater than €1,000, with some obviously erroneous entries, such as observations with heating costs of €70,000 and €12,075 that have titles like TEST IS24, or listings that indicate a mix-up of thousand separators and decimal separators. In summary, the vast majority of listings that have a numeric specification of the *Heating Costs* variable appear to be unsuspicious. However, due to the possibility of ambiguous interpretations of some of the information provided and the large share of listings that do not provide specific information or no information at all, the explanatory power of the *Heating Costs* variable is limited and needs to be carefully considered depending on the intended use.

The specification of the *Total Rent* is mandatory, resulting in a share of 100.00 % of listings having a total rent defined. The total rent usually results from the sum of the cold rent, the utilities, and, if given, the separately specified heating costs. Nevertheless, the user can add

descriptive additives that textually complement that the total rent does not include utilities, heating costs, or both. As a result, the listings can be grouped into four different types of total rent statements. 83.89 % of all observations are part of the by far largest group that includes listings without any descriptive additive, 13.76 % of all observations state that heating costs are not included, 0.82 % of all observations state that utilities are not included, and 1.53 % of all observations state that utilities as well as separately specified heating costs are not included. Despite opening up the possibility of giving more precise information about the total rent, these possibilities seem either not to be detailed enough or rather confuse the users of the online real estate marketplace, as verifying calculations of the total rent from the cold rent and its other different components lead to ambiguous results. For the largest group of observations with no additional textual specification of what is included in the total rent, 85.24 % match the sum of the cold rent and, if reported, utilities and heating costs. For the group of observations that state that heating costs are not included, 92.79 % match the sum of cold rent and, if given, utilities, and for the group of observations that state that utilities are not included, 62.52 % match the sum of cold rent and heating costs. For the observations that state utilities and heating costs must be added separately to the total rent, the total rent equals the cold rent without any other costs in 66.55 % of the cases.

These calculations show that the total rent is sometimes not reasonably comparable between different listings, as the interpretation of what is and what is not part of the total rent seems to differ and this differing view of the total rent gets not revealed through other information, which leads to the problem that it is not possible to compare the total rent even when back-calculation is applied. Thus, descriptive statistics on the total rent have limited informative value but are shown for the sake of completeness in Figure 12.

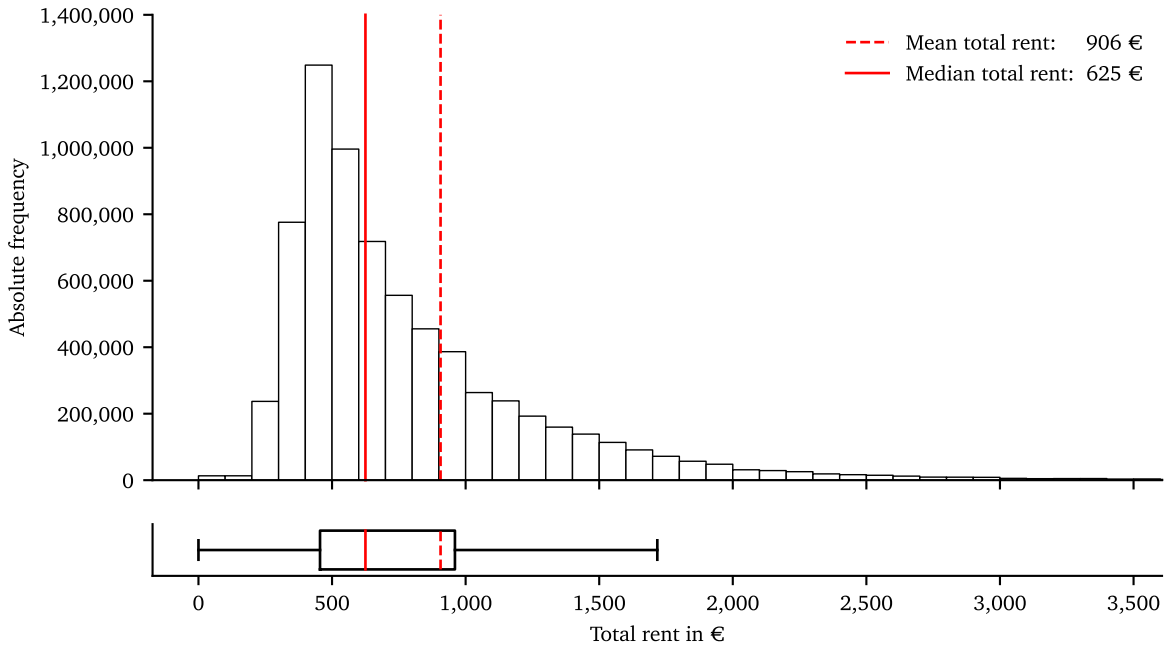
9,105 observations specify a total rent of €0, almost four times more than cold rent observations with a value of €0, but still representing only 0.13 % of all listings. Although these listings do not specify a total rent, most reveal a cold rent of more than €0. The number of observations with both a cold rent and a total rent of €0 decreases to 849, which are erroneous observations or observations that are not used in the intended way, e.g., listings using one listing for multiple properties and specifying the rent in a free text field. On the other hand, there are very few listings with extremely high rents, such as the maximum total rent, which is equal to the maximum cold rent of €111,111,111.11. In addition to these extreme outliers, 1,178 listings, or 0.13 %, have a total rent defined above €10,000.

These are similar results to the cold rent analysis, only slightly shifted towards higher values, which can also be observed for the mean total rent of €906 being higher than the mean cold

rent of €707 and for the median total rent of €625 being higher than the median cold rent of €470. Similar to the description of the cold rent, it can be concluded that most of the total rent values in the listings seem to be in an inherently realistic range, and even values at the very low and very high ends of the distribution can sometimes be considered meaningful. Nevertheless, there appears to be a slightly higher level of uncertainty associated with the total rent values than with the cold rent values, particularly in relation to the various possible components that may or may not be included in the total rent. For possible further evaluations, the shape of the distribution implies the use of a transformation by the natural logarithm.

Figure 12

Distribution of Total Rent



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 7.

The letting of real estate for use by third parties entails, on the one hand, the collection of rental income and, on the other, considerable risks for the real estate owner, such as the risk of carelessness by the tenant or the risk of uncollectible rent due to bankruptcy of the tenant. To reduce these risks, a security deposit is usually requested. For Germany, § 551 BGB regulates the security deposit and, among other aspects, the maximum security deposit (§ 551(1) BGB)

of three cold rents that can be demanded as well as the type of storage and investment of the deposit. The *Deposit* variable allows inserting free text to specify the required deposit. This option is used in 92.24 % of all listings. Due to the possibility of inputting any type of text input, the entries are very heterogeneous and complicate the comparison. Table 8 shows typical deposit descriptions of different lengths and their specific frequencies.

The deposit descriptions range in length from a single character to 50 characters, which appears to be the maximum number of characters possible. Examples of single-character entries include single numbers such as the number three, which occurs 12,035 times and could indicate the number of cold rents, but also characters such as X, which could indicate that no deposit is required, or characters such as €, which is more likely to indicate missing additional input. The specifications with very high numbers of characters are typically used to give additional or more specific explanations about the deposit, such as when it is due or that cooperative shares must be purchased instead of a deposit. Some listings, 2.82 % of all observations, use specifications such as 'MM', which stands for monthly rent, or 15.15 % of all observations use the word fragment 'rent', which often specifies a specific number of cold rents to be paid as a deposit. Finally, 89.35 % of all observations contain a numerical character in their deposit specification, which is at least an indication of a reasonable input for the *Deposit* variable, since deposit demands, regardless of whether they are given in a number of rents or in an exact amount of money, typically need to contain a numerical specification thereof.

Table 8

Typical Specifications of the Deposit Variable

Specification of variable *Deposit*	Variable Length	Variable frequency
2 Kaltmieten (2 cold rents)	12	184,022
2 KM (2 cr, abbreviation for cold rent)	4	48,693
nach Vereinbarung (by arrangement)	17	27,895
Genossenschaftsanteile (cooperative shares)	22	13,383
700,00 EUR	10	10,419
300,00 € Genossenschaftsanteil zzgl. 30,00 €Gebühr (300,00 € cooperative share plus 30,00 €fee)	50	249
-	1	168

Note. Own research. Created from raw data set.

In summary, it can be stated that the majority of the offers contain information about the deposit and that this information seems to be reasonable to a large extent. However, this information is very heterogeneous regarding both format and content. Thus, using the *Deposit* variable seems possible in principle, but the effort required to prepare the data and the resulting inaccuracies must be weighed against the benefits of its use.

Besides the previously described locational and price-related characteristics, the size and partitioning are usually especially relevant for the choice of an apartment. The data set contains two variables, the *Number of Rooms* and the *Living Space*, providing information thereon. Typically, the German Living Space Ordinance (WoFlV[16]) serves as the definitional basis when addressing living space in Germany. However, it should be noted that the WoFlV only directly applies to subsidized housing. In all other cases, it is not binding and only applied equivalently for simplicity and consistency regarding the calculation of the living space.

The determination of the number of rooms, though, has not been defined in the WoFlV or any other legal basis since the DIN 283 was repealed. Consequently, it is up to the user to determine the number of rooms and thus to interpret what defines a room, what defines half a room, and what is no room. A common approach is to count small rooms from about 6 square meter (sqm) to 10 sqm as half rooms. The numeric specification of the number of rooms is mandatory on the analyzed online real estate marketplace, resulting in all of the 7,007,571 observations having the number of rooms defined. Similar to other variables, most of the observations are in an expected range, with 99.29 % of the listings ranging from 1 to 5.5 rooms. A more detailed insight can be gained from Figure 13, which shows the distribution of the frequency of the number of rooms specifications for all listings.

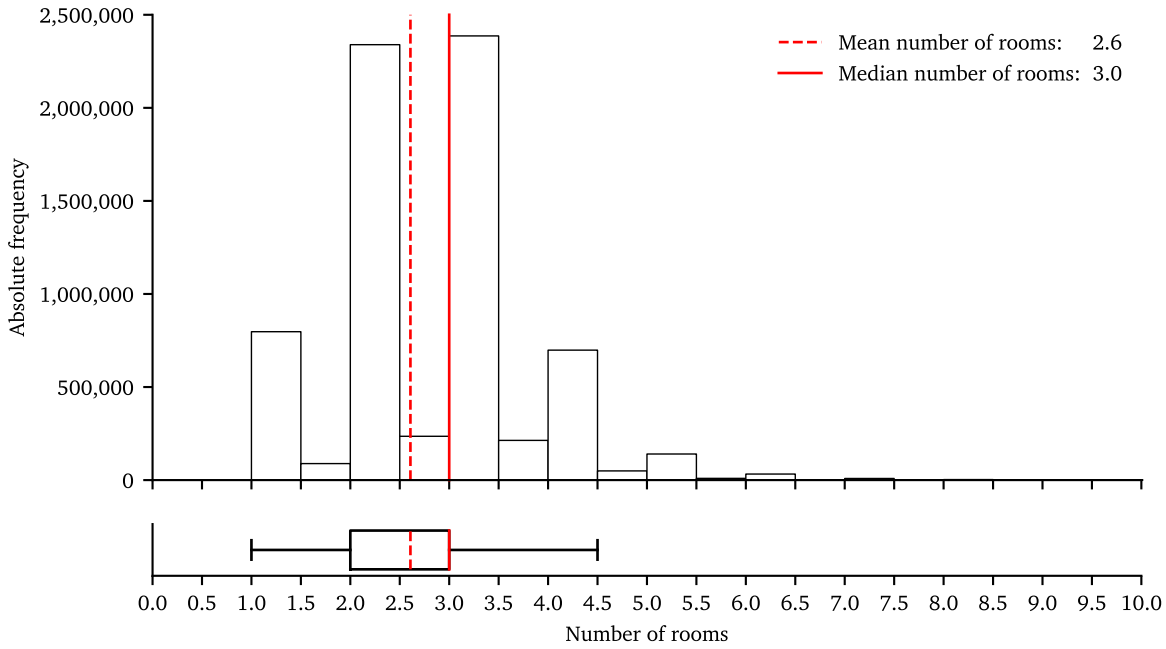
Despite the fact that smaller fractions than half rooms are not reasonable, the input field for the *Number of Rooms* variable allows floating point numbers leading to specifications like 1.51 rooms. To include these listings in the depiction, the bars include all listing specifications from the lower end, which is included, to the upper end, which is excluded. The distribution shows an expected behavior with one-, two-, three-, and four-room apartments as the most common categories. A small number of 1,943 listings, representing 0.03 % of all observations, specify more than 10 rooms. These listings include apparently incorrect *Number of Rooms* specifications, e.g., due to a mix-up of the number of rooms and size, but also a few listings that specify a correspondingly large size of the living space. Overall, however, the information on the number of rooms seems consistent and can be effectively validated by the living space. It

[16] German Living Space Ordinance promulgated on November 25, 2003 (Federal Law Gazette | page 2346)

can therefore be assumed that the data are generally utilizable for evaluations. In detail, however, the problem remains that the recording of fraction rooms, especially half rooms, is not regulated by law, leading to uncertainties and, therefore, heterogeneous data. Additionally, a small share of extreme outliers could bias the distribution.

Figure 13

Distribution of Number of Rooms



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 8.

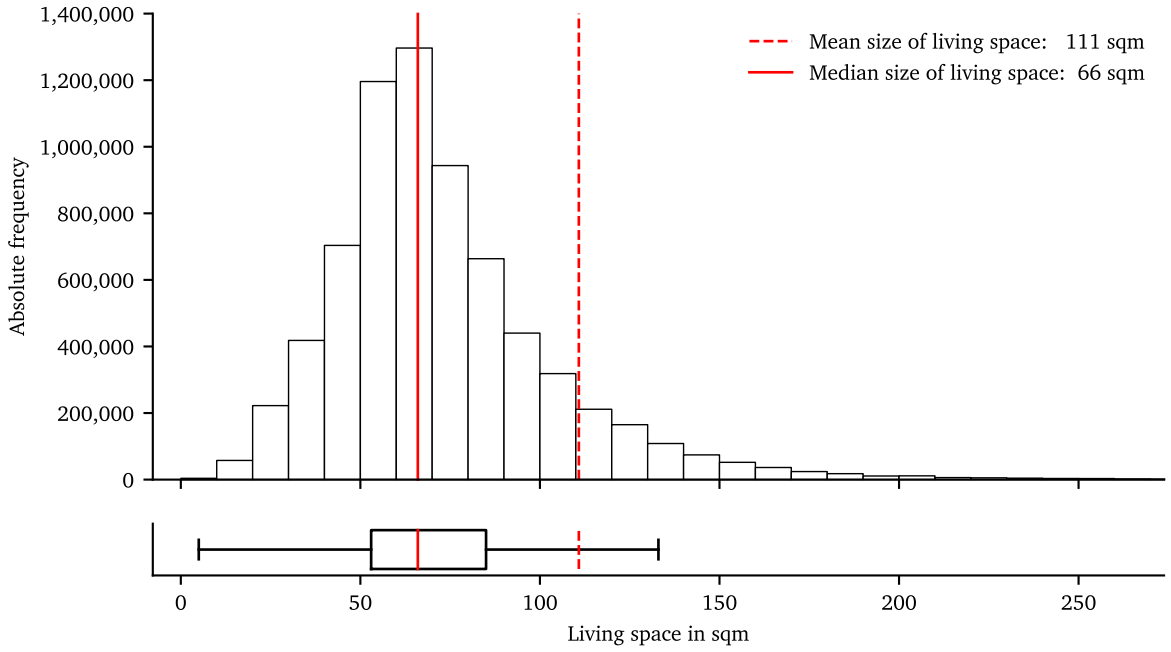
Besides the layout of an apartment, for which the *Number of Rooms* provides helpful information, the absolute size, usually, in particular, the living space, is an essential factor in the choice of an apartment and is provided by the *Living Space* variable. The specification of the variable *Living Space* is mandatory, as can be seen by the share of 100.00 % of listings having the size of the living space defined.

Nevertheless, 1,769 observations, representing a share of 0.03 % of all listings, specify a size of zero square meters. Reasons for specifying a living space of zero square meters are, among others, unintentional erroneous entries, which can be identified by their textual description and

specification of the number of rooms. On the other side, there are a few listings with very high living space specifications, as the maximum living space that can be observed is 90,000,000 sqm, and 821 listings, which represents 0.01 % of all observations, have a living space defined higher than 500 sqm. Living spaces similar to the maximum living space are erroneous values and should be classified as outliers, which can be confirmed when inspecting their titles and descriptions that describe one- to four-room apartments or test listings. The histogram and boxplot in Figure 14 show that the distribution is slightly positively skewed.

Figure 14

Distribution of Living Space



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 9.

Caused by the extreme outliers close to the maximum living space of 90,000,000 sqm, the mean and median are falling apart considerably and the interpretation of what represents a standard apartment should be made considering the median instead of the mean. Despite that bias, 98.49 % of all listings fall into a living space range from 20 sqm to 200 sqm. Summarizing, the living space information seem principally reliable despite a small share of apparently erroneous listings. Due to the slight positive skewness, a transformation using the natural logarithm may

be considered, depending on the intended use of the data. Since both variables are numeric, *Living Space* and *Number of Rooms* provide a good opportunity to exploit information and their reliability regarding the size and partitioning of an apartment without the need for extensive data preprocessing.

In addition to the characteristics already described, apartments regularly differ in their non-standard facilities, which are typically advertised and highlighted separately, as they can positively impact the realizable price. These information are recorded in the variable *Facilities*, which is not mandatory and categorical, allowing the selection of predefined categories of facilities. A share of 9.19 % of all listings do not specify any additional facilities. Since no validation category defines the non-existence of additional facilities, it is impossible to determine whether only the information is missing or there are no additional facilities. For the listings for which additional facilities are specified, Table 9 displays all possible categories and their frequency in all observations.

Table 9

Frequency of Different Facilities

Facility	Observations	
	n	%
Keller (basement)	4,448,261	63.48
Balkon/ Terrasse (balcony/ terrace)	4,197,427	59.90
Einbauküche (built-in kitchen)	2,335,509	33.33
Personenaufzug (elevator)	1,645,677	23.48
Garten/ -mitbenutzung (garden/ for shared use)	1,354,804	19.33
Gäste-WC (guest bathroom)	1,046,237	14.93
Stufenloser Zugang (stepless entrance)	639,995	9.13
WG-geeignet (suitable for shared living)	533,774	7.62
Wohnberechtigungsschein erforderlich (certificate for the entitlement for financial help required)	128,171	1.83

Note. Own research. Created from raw data set.

The frequencies of the different types of facilities do not reveal any unexpected patterns, only the feature *Wohnberechtigungsschein erforderlich*, which states that a certificate for the entitlement for financial help of the tenant is required, being rather a requirement regarding the prospective tenant than a feature of the apartment. In summary, the information contained in the *Facilities* variable are conveniently evaluable due to its categorical character and seem to be overall reliable. Nevertheless, the problem remains that there is no catch-all category for listings without facilities, resulting in opacity with respect to intentional and unintentional missing facility information and thus regarding the completeness of the data.

Besides the apartment-specific characteristics and facilities, an offered apartment can also be classified according to more general categories, regularly considered when describing listings, like the variables *Type of Apartment* and *Floor*. The *Type of Apartment* variable describes the offered apartment by a predefined category. These categories are not officially formalized by any regulation but are common practice when describing apartments and include categories such as the penthouse and attic categories. For 85.87 % of all observations, a type of apartment is defined. Since a catch-all category, designated 'other' exists, the remaining unspecified observations are either intentionally or unintentionally not described, and the missing information should not be attributable to a missing category. The distribution of the frequencies of the existing categories is generally unsuspicious and is depicted in Table 10. The sum of the most common categories, standard apartment, attic apartment, and ground floor apartment, together with the share of listings without specification of the apartment type, add up to 87.64 % of all listings. Rarely offered apartment types are apartment types that are also rare in the total housing stock, e.g., due to scarce features such as penthouses and apartments with terraces.

Table 10

Frequency of the Type of Apartment Variable

Type of apartment	Observations	
	n	%
Etagenwohnung (standard apartment)	3,462,771	49.41
Dachgeschoss (attic apartment)	887,217	12.66
Erdgeschosswohnung (ground floor apartment)	801,875	11.44
Sonstige (other)	250,463	3.57

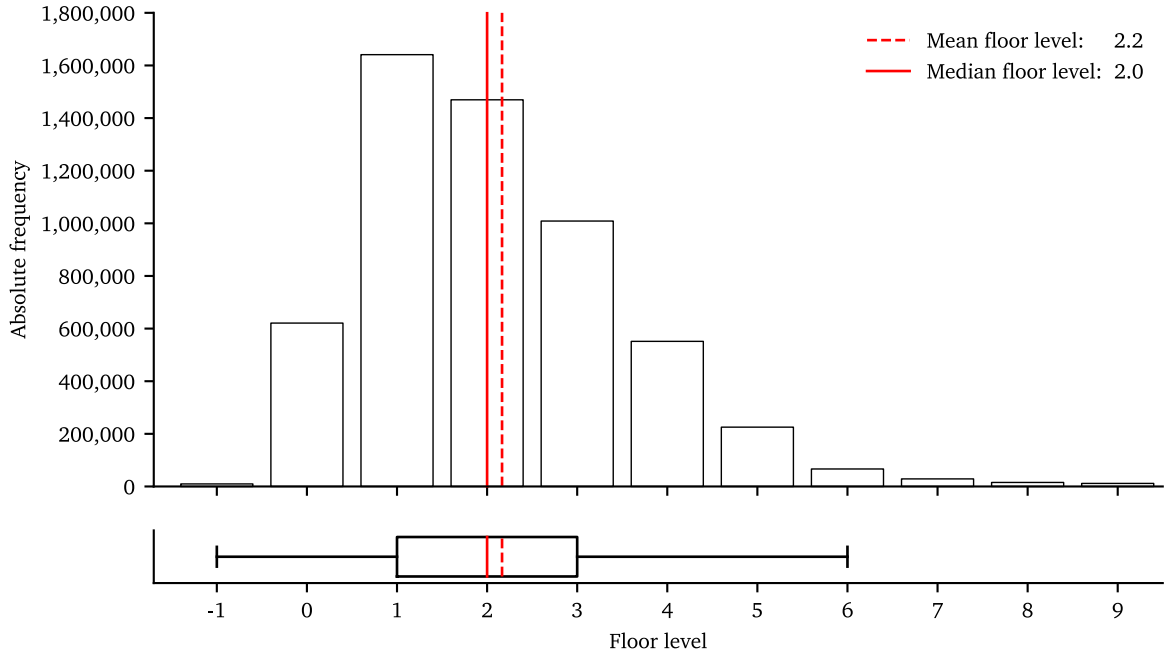
Type of apartment	Observations	
	n	%
Maisonette (maisonette)	223,218	3.19
Hochparterre (raised ground floor apartment)	142,938	2.04
Penthouse (penthouse)	89,228	1.27
Terrassenwohnung (apartment with terrace)	84,066	1.20
Souterrain (souterrain)	51,425	0.73
Loft (loft)	24,415	0.35

Note. Own research. Created from raw data set.

The categorical character of the *Type of Apartment* variable facilitates a quantitative evaluation without the need for a complex data cleansing in contrast to variables that allow free text entry. Compared to the other variables, the completeness regarding the full data set is at a medium level and the unsuspicious distribution of the *Type of Apartment* specifications, in combination with a low incentive to give false specifications, as they become apparent during the first site visit, benefit high correctness of the data. Similar to the previously described *Type of Apartment*, the *Floor* variable can be easily specified without ambiguity, in this case, by a numerical description. The online real estate marketplace from which the data was derived enables two possibilities to specify the floor, either by solely stating the floor or by additionally stating the total number of floors of the building in which the apartment is located. Of all the observations, 59.71 % specify the floor of the apartment offered and the total number of floors of the building, and 21.36 % specify exclusively the floor of the apartment offered. Thus, 18.93 % do not provide any information on this variable. The distribution of the floor, from the observations that provide information thereon, can be seen in Figure 15.

Figure 15

Distribution of Floor



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 10.

Apart from the fact that most observations are concentrated between the first floor and the 5th floor, there is still a small number of entries indicating uncommon specifications of the *Floor* variable. A share of 0.14 % of all listings state a negative floor -1, and 0.03 % specify a floor above 20. From these listings, the descriptions show that some listings are correctly specified, e.g., by describing a basement apartment or an apartment in a high-rise building, but others show that the floor specification is incorrect and that the listings are test listings or listings used to advertise multiple apartments together. Similar to the *Type of Apartment* variable, wrong information become immediately apparent during a site visit, which, together with the unsuspicious distribution, argues for generally reliable information concerning the floor. The completeness of the floor data is at a medium level compared to the other variables and the fact that the *Floor* variable can be analyzed efficiently facilitates its use in a variety of research articles, e.g., in the work of Danton & Himbert (2018) or Liu & Chang (2004) that both describe the relation of the floor of an apartment and the rent.

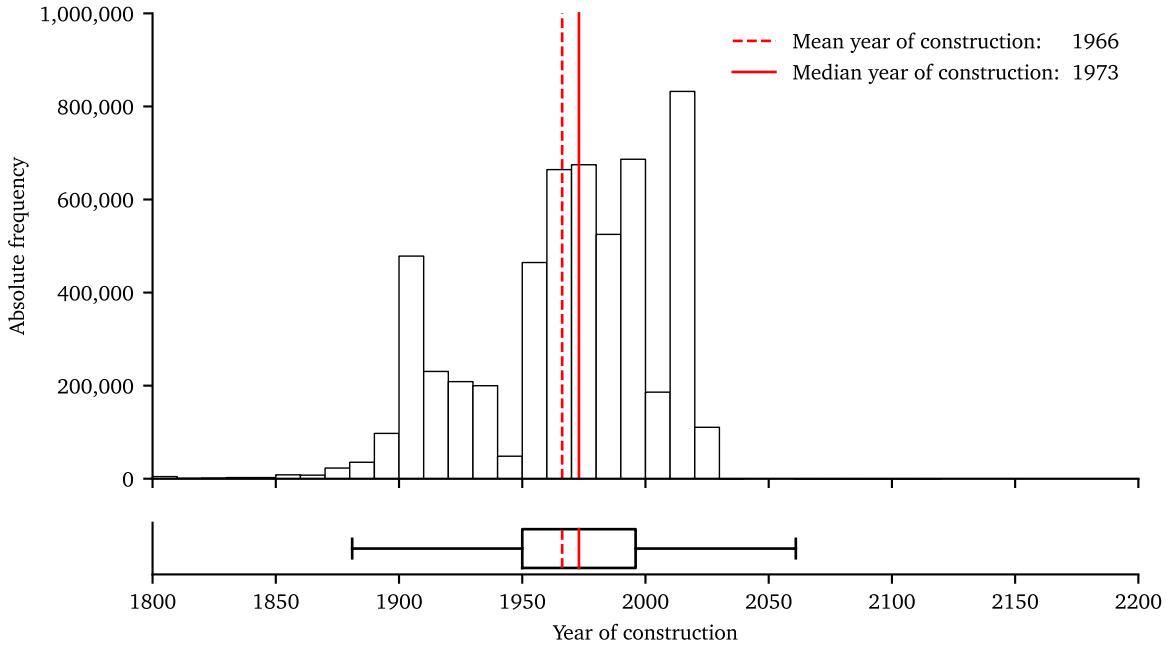
In contrast to the previously described variables, which mainly provide information about the existence or measurable size of an apartment characteristic, the variables *Year of Construction*, *Condition,* and *Quality* provide information regarding the apartment quality. Especially when taken together, these variables can reveal a picture of the quality of the offered apartments. The *Year of Construction* variable is specified for 78.71 % of all observations. A large share of the reported construction years, 89.25 %, is in the range from 1900 to 2020, and the preeminent part of the listings, 97.51 %, is in an extended range from 1800 to 2020.

Nevertheless, 3,594 listings specify a year of construction before 1500, and 1,465 listings specify a year of construction after 2020, thus, in the future from the date of data acquisition. Taken together, these listings represent 0.09 % of all observations with a year of construction stated. Some of these are obviously erroneous test listings or listings where the *Year of Construction* variable has not been used as intended by the online real estate platform. For example, a range for the year of construction is given by improperly splitting the four mandatory digits into two digits for the assumed lower end and two digits for the assumed upper end of the decade in which the building was constructed by assuming the century, e.g., 6572 for the years 1965/1972.

However, there are also listings with construction years before 1500 or after 2020 that seem to be correctly reported, e.g., listings that specify the exact year of construction before 1500 and describe facts of the building, like the characteristic of being a monument. On the other side of the distribution, some listings already offer properties before their completion for future rent, which seem to be correctly specified. Compared to most of the other variable distributions, the distribution of the *Year of Construction* is less steady, with significant drops in some categories of the histogram, e.g., a drop in the category from 1940 to 1950, in which the Second World War was a reason that led to a lower number of building completions, and a drop from 2000 to 2010, which could be caused by the effect that users who have the possibility to move into a newly constructed apartment change their apartment less frequently. The overall appearance of the distribution in Figure 16 looks generally plausible, with an increasing number of listings with decreasing age of the building, although there is one bar of the histogram from 1900 to 1910 that is a clear exception.

Figure 16

Distribution of Year of Construction



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 11.

This spike in the 1900-1910 bar can be explained by the fact that a large proportion of observations offering an apartment in a building built at the end of the 19th or beginning of the 20th century cannot determine the exact year of construction, leading to an unrealistically high number of 276,546 listings that specify precisely the year 1900 as the year of construction[17]. As a result, it can be concluded that a large share of the observed *Year of Construction* specifications is within a realistic range, but for a significant share of these observations, the exact year is subject to uncertainty and should instead be interpreted as an indication of the year of construction. Compared to the other variables, the completeness of the *Year of Construction* is at a medium level, similar to the *Floor* variable. Therefore, the use of the year of construction must be weighed under consideration of these limitations regarding completeness and precision.

[17] 4,180,995 observations specify a year of construction between the years 1900 (included) and 2000 (excluded). Assuming a uniform distribution of the years of construction, thus, 41,810 observations would be expected for one specific year instead of the observed 276,546 listings.

The *Condition* variable, as well as the *Quality* variable, are particularly susceptible to advertiser subjectivity, as the assessment of the condition and quality are even more in the eye of the beholder than it is the case with variables such as the *Number of Rooms* or *Facilities*, for which there is no official regulation but it is possible to rely upon tangible attributes. Table 11 gives an overview of the frequencies of the different possible specifications of the *Condition* and *Quality* variable.

Table 11

Frequencies of Condition and Quality Specification

Condition	Observations		Quality	Observations	
	n	%		n	%
Not specified	1,781,037	25.42	Not specified	2,940,004	41.95
Gepflegt (well-maintained)	1,711,732	24.43	Normal (normal)	2,199,816	31.39
Saniert (refurbished)	733,946	10.47	Gehoben (superior)	1,624,315	23.18
Renoviert (renovated)	714,382	10.19	Luxus (luxury)	190,466	2.72
Neuwertig (as new)	556,280	7.94	Einfach (ordinary)	52,970	0.76
Erstbezug (first occupancy)	552,416	7.88			
Modernisiert (modernized)	449,565	6.42			
Erstbezug nach Sanierung (First occupancy after refurbishment)	412,112	5.88			
Nach Vereinbarung (by arrangement)	60,157	0.86			
Renovierungsbedürftig (in need of renovation)	35,866	0.51			
Abbruchreif (condemned)	78	0.001			

Note. Own research. Created from raw data set.

Both variables describe the quality of the listings but with a different focus. While the *Condition* variable has a temporal component since it is mainly influenced by the age of the building

construction and materials used, the *Quality* variable focuses on the quality of these materials. For both variables, the groups with the most observations are those whose listings do not specify the condition or quality of the offered apartment. While the *Condition* variable is specified for 74.58 % of all observations, this number reduces to 58.05 % for the *Quality* variable. For both variables, there is a clear bias in the distribution toward positive classifications. Descriptions associated with a standard below average are scarce, e.g., in need of renovation for the *Condition* variable in 0.51 % of all observations or ordinary for the *Quality* variable in 0.76 % of all observations. This scarcity suggests that the information provided by the person offering the apartment could be positively biased, or the listings not specifying these variables could be negatively biased, or both. As a result, caution should be exercised when examining these variables due to their completeness, which is at the middle to lower end compared with other variables, as well as their correctness, which can hardly be evaluated.

The variables *Type of Heating*, *Energy Source*, *Energy Performance Certificate,* and *Energy Demand* all give information regarding energy-related topics. In times of climate change, rising energy prices, and increasing awareness of these issues, their influence on other variables, such as rent, can be expected to increase. This trend is also reflected in the scientific literature, as recent studies consistently show that energy efficiency has a positive impact on prices and rents (Chegut et al., 2020, p. 182; Im et al., 2017, p. 1031; Y. Li et al., 2019, p. 2), while the evidence on the impact of energy efficiency has been more mixed in the past (Gabe & Rehm, 2014, pp. 343–344; Hyland et al., 2013, pp. 947–948). Thus, a broader data base on energy-related aspects would be desirable.

The *Type of Heating* variable provides the possibility to specify the primary heating type of the offered apartment, which is done in 82.83 % of all listings. As the variable is categorical and only one category can be chosen, it is not possible to specify a combination of heating types if present. This missing possibility is in contrast to the *Energy Source* variable, which is also categorical but allows the specification of multiple types of energy sources. At least one category is specified by 78.88 % of all observations, and only a minor share of 1.29 % of all listings specify two or more energy sources. Table 12 and Table 13 show the frequencies of the different categories of heating types and energy sources for both variables.

Table 12

Frequency of the Type of Heating Variable

Type of heating	Observations	
	n	%
Zentralheizung (central heating)	3,394,208	48.44
Not specified	1,203,381	17.17
Fernwärme (district heating)	667,657	9.53
Gas-Heizung (gas heating)	498,659	7.12
Etagenheizung (single-story heating)	466,721	6.66
Fußbodenheizung (underfloor heating)	441,576	6.30
Öl-Heizung (oil heating)	122,686	1.75
Wärmepumpe (heat pump)	71,236	1.02
Blockheizkraftwerk (cogeneration plant)	48,716	0.70
Nachtspeicheröfen (night storage heater)	34,219	0.49
Holz-Pelletheizung (wood pellet heating)	24,228	0.35
Elektro-Heizung (electric heating)	22,336	0.32
Ofenheizung (stove heating)	8,833	0.13
Solar-Heizung (solar heating)	3,115	0.04

Note. Own research. Created from raw data set.

Both tables show that there is a predominant category. The most common heating type in the collected data set is central heating, with a share of 48.44 % of all observations, and the most prevalent energy source is gas, with a share of 41.60 %. The categorical data type allows the specification of a wide range of different heating types and energy sources. However, a few categories dominate the respective distributions, which seems plausible when comparing data from the German Association of Energy and Water Management (BDEW). The BDEW data shows similar categories as the predominant categories for the type of heating, e.g., central heating, single-story heating, and district heating, as well as for the energy sources, e.g., gas, oil, and district heating. However, the share of the different categories deviates significantly, e.g., according to the BDEW, 70.2 % of central heating (BDEW, 2019) compared to 58.48 %[18] central heating according to the collected listing data.

[18] 48.44 % / (1-0.1717) = 58.48

Table 13

Frequency of the Energy Source Variable

Energy source	Observations	
	n	%
Gas (gas)	2,915,317	41.60
Not specified	1,480,093	21.12
Fernwärme (district heating)	1,366,531	19.50
Öl (oil)	460,561	6.57
Erdgas leicht (low-caloric gas)	267,461	3.82
Strom (electricity)	178,762	2.55
Erdgas schwer (high-caloric gas)	126,255	1.80
Holzpellets (wood pellets)	65,858	0.94
Erdwärme (geothermal heat)	64,248	0.92
Nahwärme (local heat)	31,973	0.46
KWK fossil (CHP fossil)	24,937	0.36
Solar (solar)	24,593	0.35
Umweltwärme (ambient heat)	24,125	0.34
Fernwärme-Dampf (district heating steam)	20,029	0.29
Flüssiggas (liquid gas)	12,660	0.18
Wärmelieferung (heat supply)	10,487	0.15
KWK erneuerbar (CHP renewable)	5,633	0.08
Holz (wood)	5,369	0.08
Holz-Hackschnitzel (wood chips)	5,331	0.08
Bioenergie (bioenergy)	3,624	0.05
KWK regenerativ (CHP renewable)	3,191	0.05
Kohle (coal)	1,925	0.03
Wasserenergie (hydro power)	1,504	0.02
Windenergie (wind power)	466	0.01
KWK bio (CHP bio)	353	0.01
Kohle/Koks (coal/coke)	259	0.004

Note. Own research. Created from raw data set. Listings can contain multiple energy sources. Thus, the observations do not add up to 100.00 %.

Biases in the listing data could cause these deviations. Effects like deviating homeownership rates in urban and rural areas and related differences in the share of offerings with associated differences in the heating types in urban and rural areas could be one of many causes that lead to a deviation of both data sets. Thus, despite the general plausibility, uncertainty remains concerning the correctness of the heating type and energy source specification of the individual listings. Compared to the other variables, the completeness is at a medium level, and due to the categorical nature of the variables, their application for quantitative analyses seems to be well realizable.

From an economic point of view, the declared energy demand of the offered apartment and the way this energy demand is calculated are important, as the energy demand has a direct impact on the utility costs that the tenant has to bear, and this declared energy demand directly depends on the type of calculation. According to § 79 in conjunction with § 82 of the German Building Energy Act (GEG[19]), energy performance certificates that inform about the energy demand of a building can be calculated either based on the actual consumption of the last three years or based on a calculative demand. These types of energy performance certificates vary significantly in their general approach and, thus, in their results, the energy performance certificates based on the calculative demand usually stating a higher demand than the certificates based on actual consumption data (Ackermann, 2020, p. 9). The information contained in the *Energy Performance Certificate* variable provides this information and is depicted in Table 14.

Table 14

Distribution of the Energy Performance Certificate Variable

Energy Performance Certificate	Observations	
	n	%
Verbrauchsausweis (based on consumption data)	2,909,361	41.52
Not specified	2,574,196	36.73
Bedarfsausweis (based on calculative demand)	1,524,014	21.75

Note. Own research. Created from raw data set.

[19] German Building Energy Act in the version promulgated on August 8, 2020 (Federal Law Gazette | page 1728), last amended by Article 18a of the Act of July 20, 2022 (Federal Law Gazette | p. 1237).

Despite the legal requirement of an energy performance certificate for almost all types of apartments, a large share of 36.73 % of all observations do not specify the type of the energy performance certificate, which must then be shown to the potential tenant at the latest when viewing the apartment (§ 80(4) GEG). As suggested by the name, energy performance certificates based on a calculative demand have to be created by calculations with several different inputs, whereas energy performance certificates based on consumption data can be prepared comparatively uncomplicated, which could explain their further diffusion within the listing data. However, in combination with the comparatively low completeness of the variable, with a specification as low as 63.27 % of all observations, the energy performance certificate data must be used carefully. Nevertheless, especially due to the information regarding the typically occurring differences in the designation of the energy demand, the combined evaluation of the *Energy Performance Certificate* variable and the *Energy Demand* variable could be valuable.

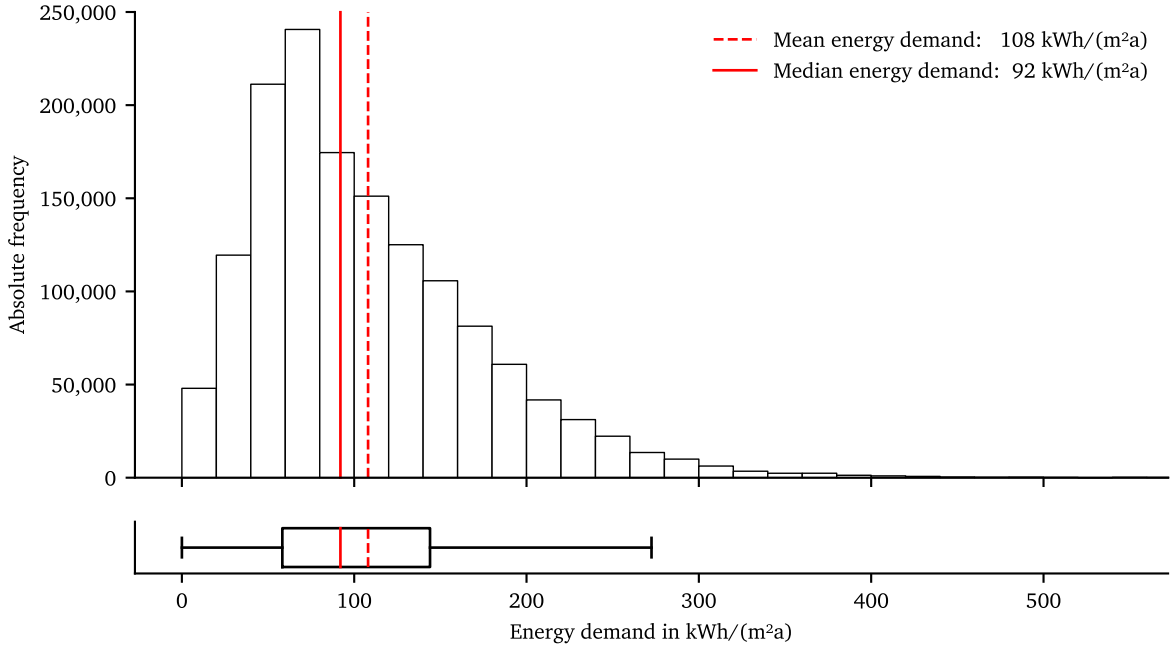
The *Energy Demand* variable specifies the energy demand according to the methodology defined by the type of energy performance certificate in kWh/(m²a). This unit states the energy demand in kilowatt hours per square meter and year and, combined with the *Energy Source* variable information, offers a good basis for the potential tenant to estimate the energy costs. However, the *Energy Demand* variable is only specified in 20.79 % of all observations. The individual energy demand varies enormously depending on factors such as the energy efficiency of the building and, if calculated by actual consumption data, by the individual usage patterns. McKenna et al. (2013, p. 83) show a variation of the calculated average energy demand of the German building stock up to the construction year of 2010, depending on location in the new or old federal states, the type of building, and the year of construction between 74 and 393 kWh/(m²a). However, these are calculative averages and individual energy demand can deviate significantly from these values, e.g., depending on altitude and exposure or, especially for newer buildings, depending on newer and higher energy efficiency standards of the building materials. First introduced in the early 1990s, the passive house concept leads to net heating energy demands lower than 15 kWh/(m²a). The distribution of the energy demand depicted in Figure 17 thus shows a generally unsuspicious positively skewed distribution, with the major share of listings, 98.39 %, in the range from 15 to 400 kWh/(m²a).

At the level of individual listings, the correctness cannot be assessed by the given data. However, the legal obligation to provide an official energy performance certificate to the potential tenant during the apartment inspection should result in most of the given values being correct. The very low completeness of the data, compared with the other variables, needs to be considered

carefully when using the data, especially as it could be caused by a self-selection process of more energy-demanding apartments, which would lead to a biased distribution.

Figure 17

Distribution of Energy Demand



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 12.

The variables *General Textual Description*, *Textual Description of Facilities*, *Textual Description of Location,* and *Textual Description of Miscellaneous* provide the possibility to add complementary textual information. The frequency with which these variables are used varies from 92.44 % of all observations for the variable *General Textual Description*, which is regularly used to further describe general characteristics of the apartment offered, to 70.51 % for the variable *Textual Description of Miscellaneous*, which is used for a wide variety of purposes, ranging from disclaimers to additional information regarding the deposit or information to be provided by the prospective tenant. Detailed depictions of the distributions of the characters for all variables are provided from Appendix A - 13 to Appendix A - 16. The character distributions show unsuspicious positively skewed distributions, and the maximum values of these distributions imply an upper limit in the range of 4,000 characters. The most common words include

conjunctions, articles, and prepositions, usually considered non-informative in textual analysis and are therefore cleaned as stop words (Silva & Ribeiro, 2003, p. 1662). Table 15 shows a stop word cleaned list of the 10 most frequent words for each variable.

Table 15

Most Frequent Words in Textual Descriptions

General Textual Description	Textual Description of Facilities	Textual Description of Location	Textual Description of Miscellaneous
Wohnung (apartment)	Wohnung (apartment)	befindet (is located)	Angaben (details)
befindet (located)	Bad (bathroom)	sowie (as well as)	Wohnung (apartment)
sowie (as well as)	Küche (kitchen)	Minuten (minutes)	Exposé (exposé)
verfügt (has)	Balkon (balcony)	ca. (approx)	Uhr (clock)
Küche (kitchen)	Fenster (window)	liegt (is located)	Informationen (information)
liegt (is located)	Dusche (shower)	Nähe (near)	vorbehalten (reserved)
Bad (bathroom)	Einbauküche (fitted kitchen)	Einkaufsmöglichkeiten (shopping opportunities)	Bitte (Request)
ca. (ca.)	sowie (as well as)	unmittelbarer (directly)	Besichtigungstermin (viewing appointment)
Balkon (balcony)	Badezimmer (bathroom)	befinden (are located)	Haftung (liability)
Haus (house)	ausgestattet (equipped)	erreichen (reach)	Richtigkeit (correctness)

Note. Own research. For further analysis, a more detailed list of the 100 most frequent stopword-cleaned words can be found in Appendix A - 17.

The most frequent words in each descriptive category indicate the respective focus areas. These focus areas appear consistent with the intended use of the text fields, as indicated by their designations, and thus provide an indicative confirmation of their contentual quality. The general textual descriptions, for example, seem to focus on a general description of the apartment offered, which can range from the location of the apartment to the rooms and facilities available, while the location descriptions expectedly contain mostly words related to locational information. Besides this indicated coherence of the content, there is also an appropriate amount of text. The average character count ranges from 269 characters for the facility description to 472 characters for the general description, including observations with no description, and from 334 to 510 characters excluding observations with no description. Thus, there is an overall basis, both in quantity and substance, upon which methods of preprocessing, such as stemming and lemmatization, and analysis, such as natural language processing techniques, can be applied.

In addition to information on the characteristics of the apartment, other complementary information, such as the provider of the apartment and the date of availability, are regularly given. The information on the provider of the apartment contain the names of real estate agents but also of private apartment providers, which is personal data and falls under the scope of the German Federal Data Protection Act (BDSG[20]) due to the automated nature of the data collection process (§ 1(1) BDSG). Therefore, only whether a housing provider is reported, being the case in 99.06 % of observations, is recorded, and no other information about the provider are collected. In principle, it would also be possible to investigate whether the provider is a professional or private apartment provider. For further evaluations, however, conformity with the BDSG must always be considered regarding the variable *Apartment Provider*.

The additional complementary text field for the *Availability* variable allows free text entry and is specified in 91.47 % of all observations. Thus, the *Availability* variable allows the provider of the apartment to define the date from which the apartment is available, to match the temporal component in addition to the attribute-related component. The typical specifications of this variable include an eight-digit date in the format dd.mm.yyyy, 41.22 %, the term sofort (immediately), 37.58 %, and the terms Vereinbarung/Absprache (agreement), 9.11 %, making up 87.91 % of all observations with a specification of the *Availability* variable. Due to the possibility of free text entry, there are also various individual declarations of the date of availability as six- or seven-digit declarations or only the month. However, most observations

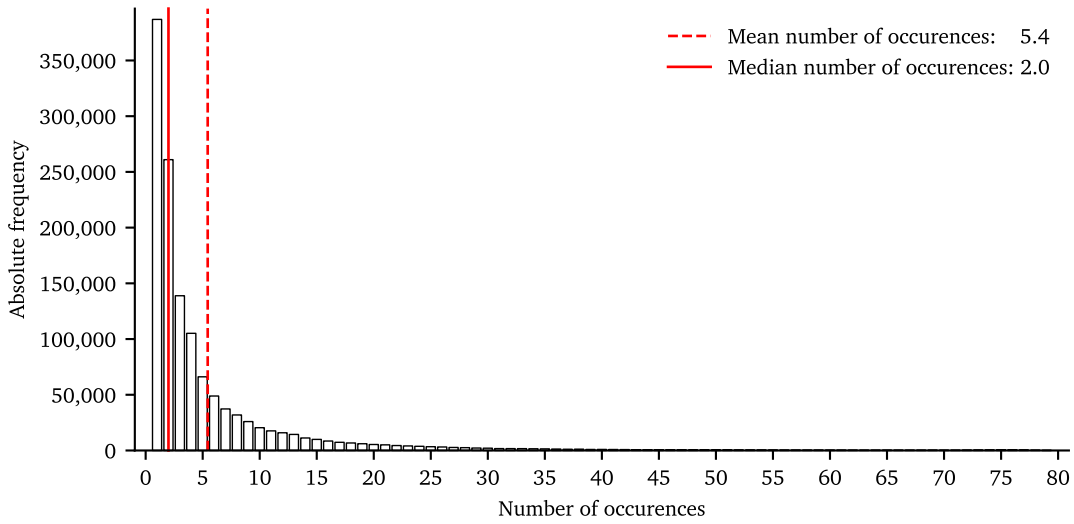
[20] German Federal Data Protection Act in the version promulgated on June 30, 2017 (Federal Law Gazette | page 2097), last amended by Article 10 of the Act of June 23, 2021 (Federal Law Gazette | p. 1858; 2022 | p. 1045).

seem to contain a reasonable date of availability. Therefore, the *Availability* variable, especially together with the date of data acquisition, can provide a useful basis for further evaluation but may require extensive data preparation and cleansing.

In addition to the variables directly related to the ORL, supplementary data such as a unique identifier and the date of data acquisition were recorded. The *Identifier* variable is automatically defined by the online real estate marketplace and is therefore available in 100.00 % of observations. One purpose of this variable is to make each listing uniquely identifiable, even if parts of the listing are corrected or changed. This unique identifiability is particularly helpful when the data is collected directly from the website and not provided by the website operator or an intermediary, as it provides a way of estimating the time on the market of a particular offer when the ORLs are collected repeatedly at different points in time. From early 2019 to mid-2020, ORLs were collected repeatedly on 79 dates during the data collection period. Figure 18 shows that most listings are offered for short periods, with 30.03 % for a maximum of two weeks or less, 50.29 % for three weeks or less, and 87.13 % for 11 weeks or less.

Figure 18

Distribution of Identifier Frequency



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 18.

However, the mean number of occurrences also indicates that within the large data basis of 7,007,571 listings, a large fraction of observations results from duplicates in many, but not necessarily all, aspects of these listings. Typically, most variables do not change from one observation to the following observation of the same ORL, identified by its unique identifier, but even the observation of the same ORL at another point in time contains information, e.g., the time on market. Besides this aspect, it is also not possible to detect changes in variables, such as the *Cold Rent* variable, unless comparable values are available. In the full data set, 1,288,255 unique observations of the *Identifier* variable were collected, which is a similar number of collected listings to the study of Barron et al. (2020, pp. 14–15), who collected 1,097,697 listings from Airbnb and assumed this to be the most comprehensive data set of the U.S. home-sharing market. Excluding commercially available data sets and data sets provided by intermediaries of the website from which the data were collected, it is also reasonable to assume that the data set acquired for and used in this study is one of the largest or the largest data set for individual publicly available research regarding the particular online real estate website.

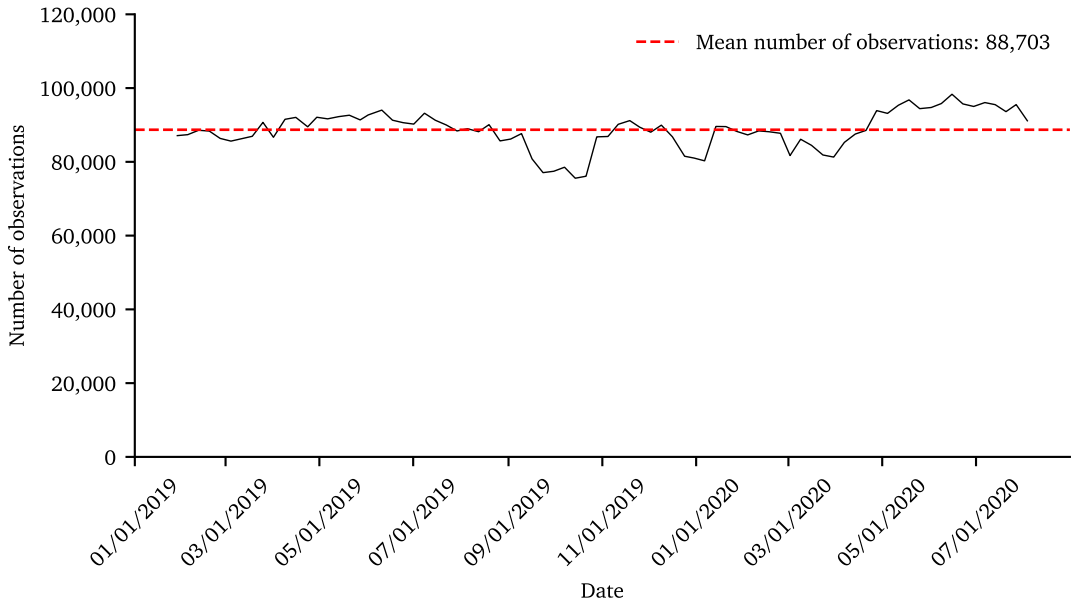
It is noticeable that important content such as the address or size of the apartment can change considerably, especially in the case of offers with not changing *Identifier* variables that are available for a very long time. Furthermore, very similar listings may appear at different points in time with different specifications of the *Identifier* variable. In this regard, various aspects, such as the pricing of the platform, can influence the *Identifier* variable. These ambiguities demonstrate that the *Identifier* variable can only be an indicator for identifying unique offers. Depending on the importance of the *Identifier* variable for the specific type of evaluation, additional duplicate cleansing, for example, based on an algorithm for comparing different characteristics, may be necessary. However, if the clear identification of duplicates is not the focus of the evaluation, the *Identifier* variable is a good indicator for the identification of unique listings, also due to its completeness of 100 %.

In addition to the *Identifier* variable, the *Date of Data Collection* variable also allows conclusions to be drawn about the temporal component of the observations. The individual timestamp is recorded for each observation that contains the information of a listing at a certain point in time. Since the web scraping data collection process is regularly implemented as an iterative process iterating over all web pages of interest, the exact timestamp varies from observation to observation, even within a single collection cycle. The information in this timestamp can be used for various applications. First, it is possible to identify the total number of collection cycles and assign each observation to one of these cycles. Second, the time required for each collection

cycle can be calculated since this time can vary significantly depending on many factors, such as the response time of the servers and the number of listings available. Third, additional information can be derived, such as how many observations were available during each collection cycle. Figure 19 shows these variations in the number of observations.

Figure 19

Distribution of Observations per Collection Cycle



Note. Own research. Created from raw data set. The table with the underlying values can be found in Appendix A - 19.

The number of observations ranges from 75,558 listings on 10/14/2019 to 98,330 listings on 06/15/2020. In addition to the number of observations, the time required for each collection cycle also varies substantially, from 2 hours 20 minutes and 33 seconds to 11 hours 55 minutes and 5 seconds. On average, the time between two data collection dates is 7 days and 2 hours. However, due to significant changes to the website and the associated need to adapt the data collection program, one of the time differences is greater, with 13 days and 21 hours, almost exactly two weeks[21]. In principle, however, the data collection was conducted on a very regular basis to minimize deviations due to the day of the week, even though irregular surveys are not

[21] More details can be found in Appendix A - 19

uncommon when using web scraping (Barron et al., 2020, pp. 14–15). Thus, the *Date of Data Collection* variable is a valuable source of information, especially as a foundation for derivative information, such as the completeness of each data collection cycle. In this respect, the behavior of the distribution of the number of observations does not show any suspicious anomalies. Finally, it can be summarized that ORLs provide a large variety of information that can be evaluated on a large scale to assess their quality and quantity, summarized in Chapter 6.1.

3.1.3 Data Selection for Vacancy Analysis

The previous chapters have discussed in detail the acquisition and quality of online real estate listing data, which serve as the foundation of this research. However, to evaluate the potential of online real estate listing data to increase the transparency of the real estate market using the example of estimating vacancy rates, the integration of additional data sources is necessary. Additional data sources are needed since not all groups of variables commonly used in models for estimating vacancy can be found in or derived from the ORL data. Omitting these sources of information would increase the imprecision of the estimates and also complicate the comparison of the relationship between ORL data and vacancies with alternative data sources and vacancies. Therefore, the aim is to integrate at least one variable from each of the groups of variables, which are typically used in the context of vacancy estimation, according to Chapter 2.3.4, and at most two variables, due to the risk of multicollinearity. The selection of specific variables is based on their availability and presumed explanatory power for vacancy rates. Variables that can be derived from ORL data are given priority over alternative data sources because it is the objective of this work to verify their quality and, in contrast to alternative data sources, they are permanently available in up-to-date form.

The selection of the vacancy data is particularly important since the vacancy rate is the variable to be explained and it serves as the dependent variable in the model. Therefore, the data must be appropriate in several respects, including the temporal aspect, the spatial scale, and the accuracy. Additionally, as derived in Chapter 2.3, the used vacancy data predetermine the type of vacancy that can be estimated. The results for this type of vacancy are not necessarily transferable to other types.

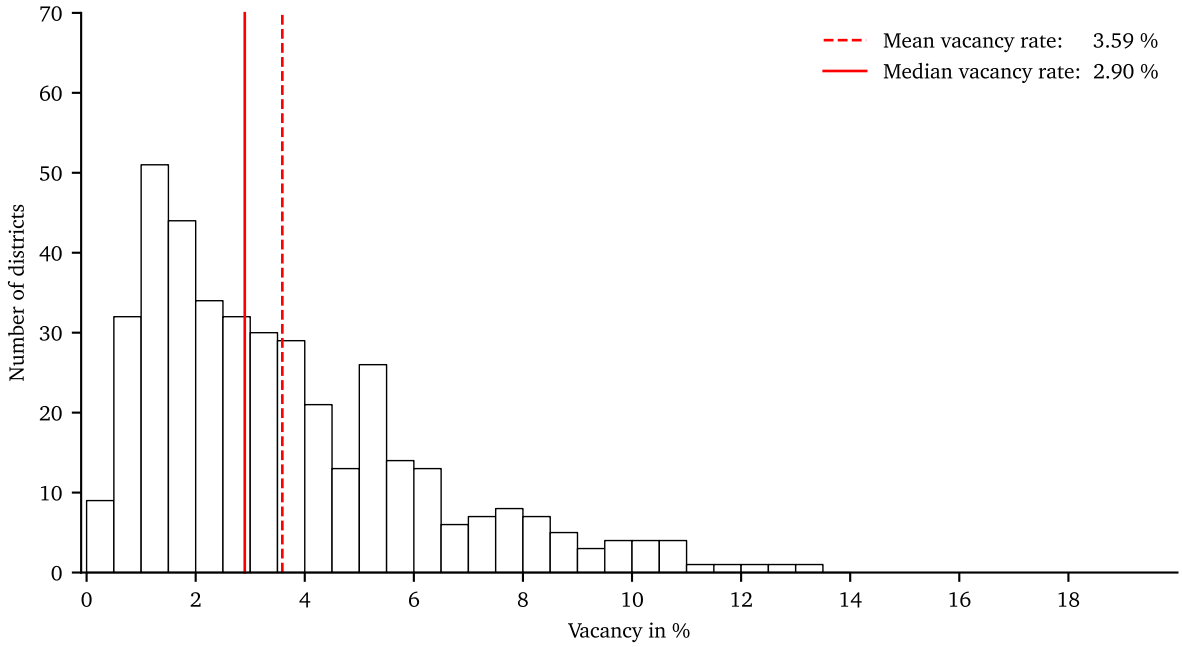
Considering the results of Chapter 2.3.3, it would be desirable to estimate vacancy at large spatial scales without similarly great effort since combining these results with other low-effort methods could increase the overall accuracy of vacancy estimations. The large-scale acquisition of online real estate listing data, covering the entire German market for apartments from the

beginning of 2019 to mid-2020, allows testing the applicability of ORL data for that purpose. However, German vacancy estimates at this level are rare, as can be seen from the example of studies such as that of Lerbs & Teske (2017, p. 59), still relying on vacancy data from the German decennial census of 2011, since vacancies are officially recorded only about every ten years in Germany (Jacobs & Diez, 2018, Chapter 0.2, p. 2). In addition to the 2011 census data, which is the most recent official vacancy data at the time of this study, there is also the CBRE-empirica-vacancy-index data from the empirica ag, which is used in vacancy-related studies, e.g., in the study by Berger & Schmidt (2019) and which provides more recent vacancy estimates. The integration of vacancy data that is more up-to-date than the census data is essential since vacancy rates tend to adjust slowly, but the period of about 10 years has already resulted in significant changes thereof, as can be seen at the example of Leipzig, where the market-active vacancy has reduced from 9 % in 2011 to 1.5 % in 2017 (Rink, 2021, p. 11). Therefore, the provision of the CBRE-empirica-vacancy-index data for this study, including vacancy rate estimates for all 401 German districts for 2019, the year with the most ORL data observations, is gratefully appreciated.

The publisher of the CBRE-empirica-vacancy-index provides a data description that includes information on the data collection methodology and the vacancy definition used (empirica ag, 2021). The calculation basis of the estimated vacancy rates is based on information on the management of more than 733,000 residential units provided by CBRE, supplemented by regression estimates and expert knowledge. The objective of these estimations is to provide information on the market-active vacancy of apartments for rent. Consistent with the ORL data collected, this excludes apartments for sale. Market-active vacancies are, in this context, defined as vacancies that are immediately available, as well as vacancies that are temporarily unavailable due to problems that can be resolved within six months. This definition excludes apartments that are part of the ORL data set, which are still in use but are already being marketed to new tenants (empirica ag, 2021, pp. 6–7). The distribution of the vacancy data can be seen in Figure 20 and shows a positively skewed distribution.

Figure 20

Distribution of the Vacancy Rate Variable



Note. Own research. Created based on data from empirica ag (n.d.). The table with the underlying values can be found in Appendix A - 20.

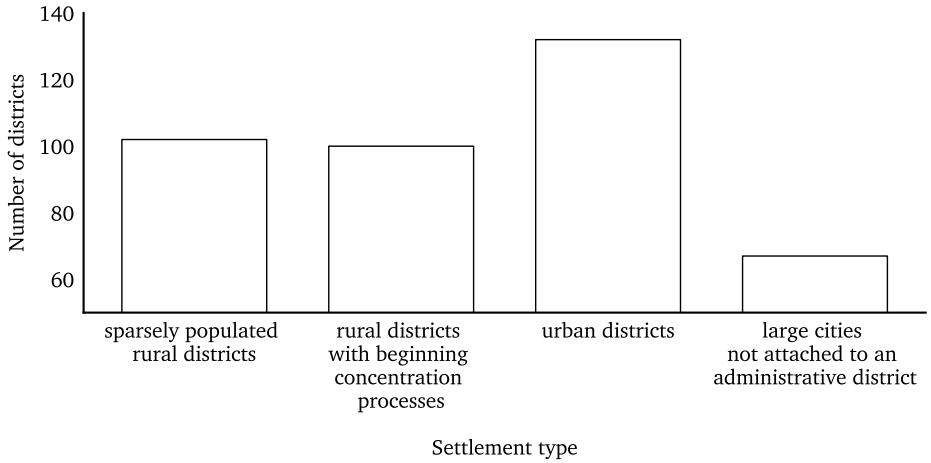
The variable groups typically included in vacancy research and identified in Chapter 2.3.4 include variables that describe locational, economic, sociodemographic, socioeconomic, and property characteristics. The ORL data include locational information in variables such as the *Designation of Municipality* and *Textual Description of Location*. However, these variables are not suitable to describe the situation at the aggregated district level since they contain either a plain spatial localization of the individual objects or a textual description unsuitable for this type of evaluation.

Trends that influence the structure of the population and that are regularly discussed include urbanization and urban exodus and imply that the settlement types, in particular, could be relevant for vacancy research. One of the most established approaches to distinguish settlement types, not focusing solely on particular subcategories and therefore being particularly suitable for this study that covers the entirety of Germany, is an approach of the German Federal Office for Building and Regional Planning (BBSR) that distinguishes settlement types at the district level (Küpper & Milbert, 2020, pp. 88, 99–100). This approach uses the proportion of the

population living in large and medium-sized cities and the population density as criteria for the delimitation of the district-level settlement types (Küpper & Milbert, 2020, p. 89). It distinguishes cities not attached to an administrative district, urban districts, rural districts with beginning concentration processes, and sparsely populated rural districts (Küpper & Milbert, 2020, p. 90). Figure 21 shows the absolute frequency of the different categories.

Figure 21

Distribution of the Settlement Type Variable



Note. Own research. Created based on data from BBSR (2020). The table with the underlying values can be found in Appendix A - 21.

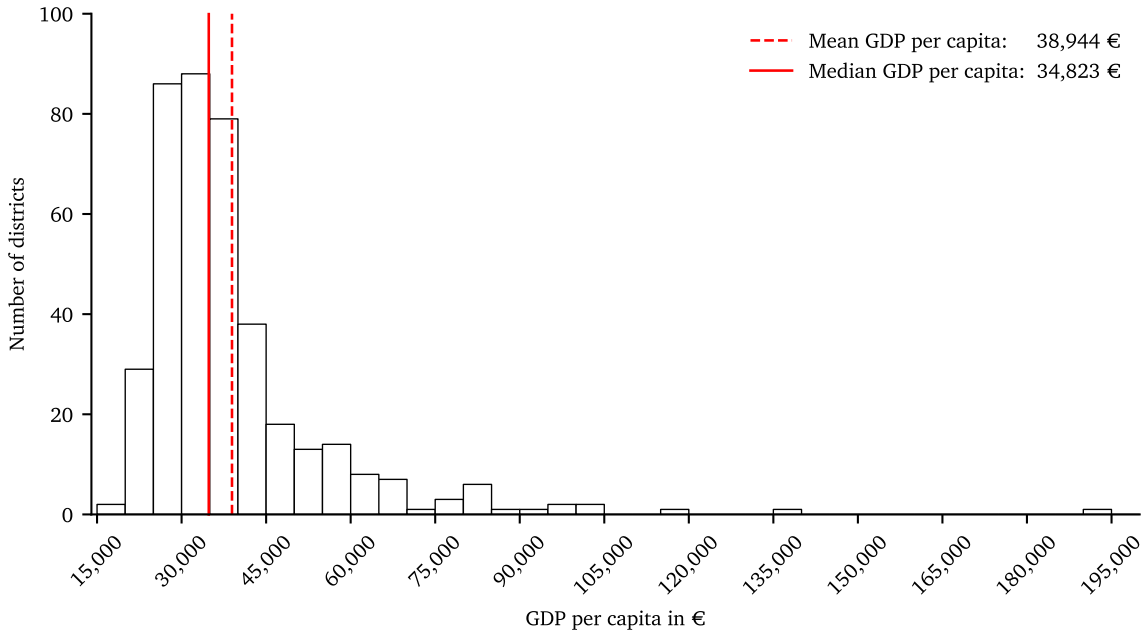
The relationship between economic variables and the vacancy rate has been confirmed several times, as shown in Chapter 2.3.4. However, the distinction between economic and socioeconomic variables is not clear-cut, so they cannot be assigned to just one of the groups in all cases and are therefore discussed together in the following. The ORL data set primarily includes the rent as an economic variable that can be used to indicate the scarcity of the good housing in the respective area. A detailed description of the variable *Cold Rent* has already been given in Chapter 3.1.2. Since the cold rent is usually almost proportionally related to the living space and it is assumed that the price per area unit rather than the total rent carries the economically more relevant information, the total rent is divided by the living space of the listing.

In addition to the rather economic information on rent, the GDP per capita is included and can

be considered a more socio-economic variable, as it contains information on the economic strength of the population within the respective area. The GDP per capita data are regularly issued by the German Work Group National Account of the Federal States (AK VGR). As can be seen from Figure 22, the distribution of the GDP per capita in the different districts is positively skewed, implying a transformation of the data by the natural logarithm.

Figure 22

Distribution of the GDP per Capita Variable



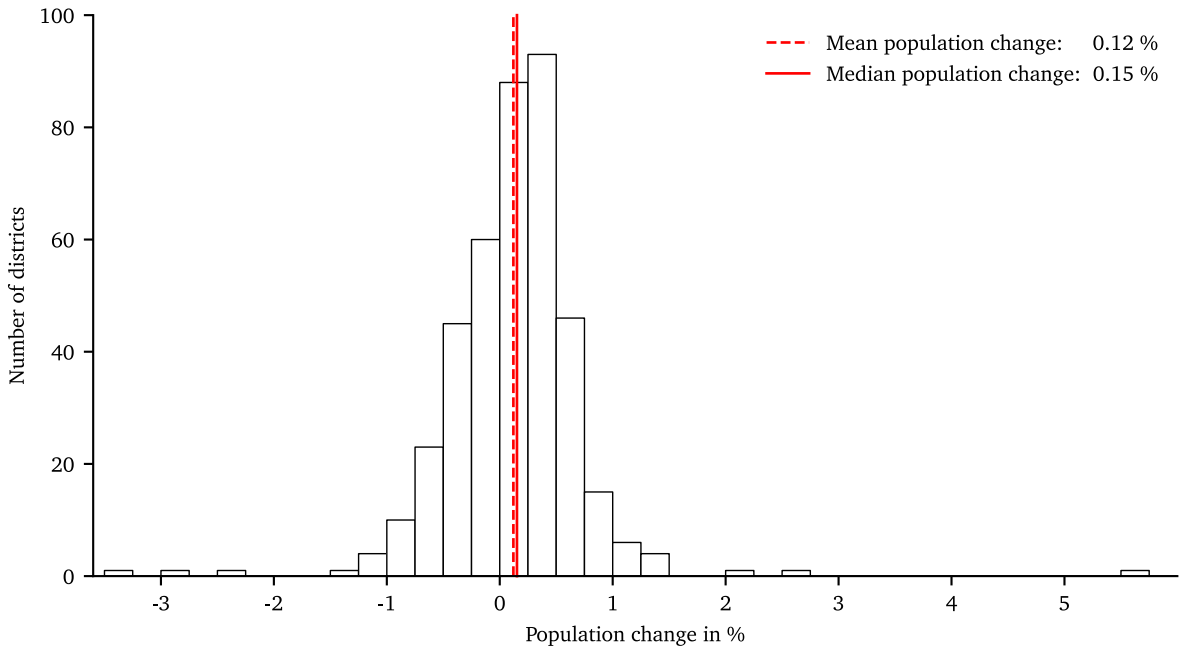
Note. Own research. Created based on data from AK VGR (Ed.) (2022). The table with the underlying values can be found in Appendix A - 22.

Sociodemographic variables describe the characteristics and developments of populations or groups. In the context of vacancy, a wide variety of sociodemographic variables could be relevant, as characteristics such as age or social background may influence the usual demand for housing. The migration balance, birth rate, and death rate may be even more relevant to market-active vacancy since the adjustment of the housing stock is slow, and population inflows or outflows directly impact vacancy rates. The annual change in district population reflects all of these effects in a single number and is therefore included using data from the federal and state statistical offices.

In total, the population increased by almost 150,000 inhabitants in Germany and Figure 23 shows the distribution of the rates of change in percent for all districts.

Figure 23

Distribution of the Population Change Variable



Note. Own research. Created based on data from Statistische Ämter des Bundes und der Länder (n.d.-b). The table with the underlying values can be found in Appendix A - 23.

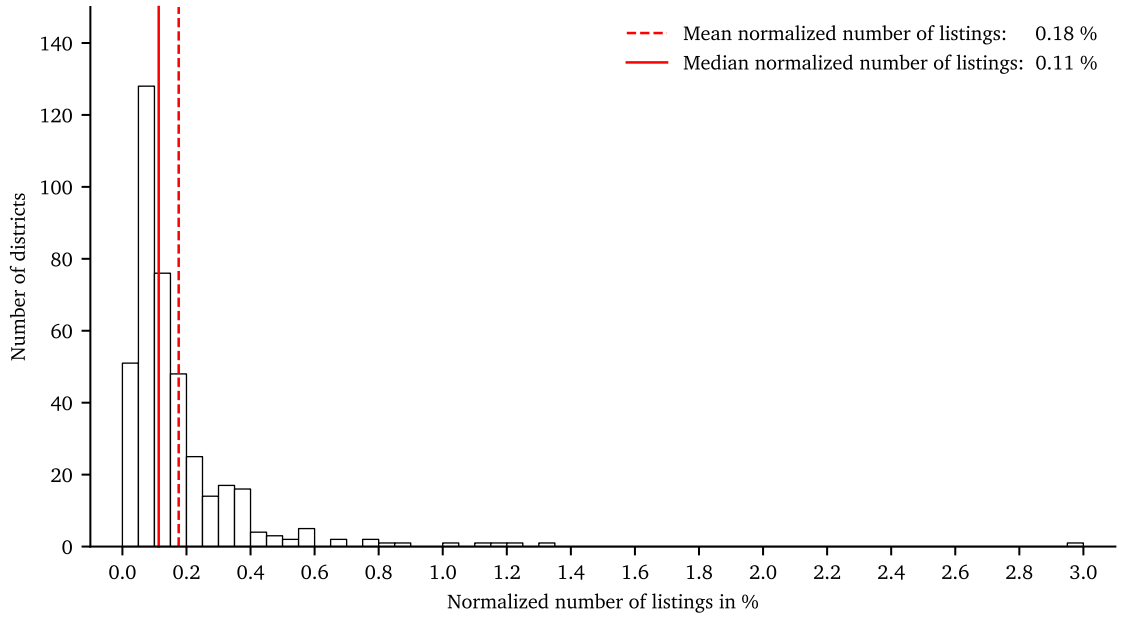
The identification of potentially relevant data for explaining vacancy revealed that studies typically include property characteristics at the individual building level and attempt to find relationships between these individual characteristics and particular vacancies. Since this research examines vacancy at the aggregate district level, the then aggregate property characteristics may be less useful in explaining vacancy, and instead, more general characteristics, descriptive of the entire building stock, may become more critical. Thus, an aggregate property-specific variable but also a variable describing the entire building stock were included to account for both types of variables. The ORL data set includes different variables that describe property characteristics at the level of individual listings. Referring to Table 3, which presents an overview of all variables in Chapter 3.1.2, the *Living Space* variable is well suited for this purpose, as it is available for all listings, is continuous, quantitative, and therefore

well evaluable, and describes an essential characteristic of real estate offers. A detailed description of the *Living Space* variable can also be found in Chapter 3.1.2.

Not included in the ORL data set as a variable that is defined by the user of the online real estate platform but derivable from the ORL data set is the number of listings per district, which is displayed in Figure 24[22]. Set in relation to the housing stock, the then normalized number of listings provides information about the tightness of the market.

Figure 24

Distribution of the Normalized Number of Listings Variable



Note. Own research. Created from raw data set and based on data from Statistische Ämter des Bundes und der Länder (n.d.-a). The table with the underlying values can be found in Appendix A - 24.

Considering the assumption of a proportional relationship between supply and vacancy from the chapter where the vacancy variable is described, this normalized number of listings could

[22] The derivation of the values depicted in Figure 24 requires a prior data preprocessing, in particular a preparation of the spatial information and a data aggregation, which are described in Chapter 3.2.2. However, in order to give a complete and coherent overview of the selected variables, based on which the further data analysis is described, the values are presented here.

have a particularly strong relationship with the vacancy rate. For a more rigorous interpretation, this normalized number of listings is scaled by the 48 data acquisitions to show the average number of listings per district per data acquisition in relation to the housing stock of that district. Since this can be interpreted as a fraction of the total housing stock, it is shown as a percentage. Summarizing, Table 16 shows that various data sources are integrated and almost half of the explanatory variables are based on the ORL data set.

Table 16

Variable Overview

Variable	Variable type	Source(s)
Vacancy Rate	Endogenous	empirica ag (n.d.)
Settlement Type	Exogenous	BBSR (2020)
Cold Rent per SQM	Exogenous	ORL data set
GDP per Capita	Exogenous	AK VGR (Ed.) (2022)
Population Change	Exogenous	Statistische Ämter des Bundes und der Länder (n.d.-b)
Living Space	Exogenous	ORL data set
Normalized Number of Listings	Exogenous	ORL data set; Statistische Ämter des Bundes und der Länder (n.d.-a)

Note. Own research.

3.2 Data Preprocessing

The intended bivariate analysis, as well as the regression analysis, require preprocessing for some of the different exogenous and endogenous variables. The preprocessing mainly concerns the ORL data, as, due to their generation process, the ORL data exhibit characteristics that require data exclusion, preparation of their spatial information, and data aggregation. Especially for ORL data, these are typical steps that are necessary since the beginning of their use for research (R. Gabriel et al., 2008, p. 291), as ORL data in parts show the, by Hernes et al. (2021, p. 15) defined, typical characteristics of big data that include a spontaneous and fast user input that is not quality assured.

3.2.1 Data Exclusion

The exclusion of listings is based on existing literature on online real estate listings, the general plausibility of the listing data, and, in particular, the plausibility of the listing data used to estimate the vacancy rate and the time component and exclude listings that carry no reasonable information. As described in Chapter 3.1, data collection was conducted from early 2019 to mid-2020. For answering the first research question, it was beneficial to have as large a data base as possible. However, to examine the contribution that ORL data can provide to the estimation of vacancy rates, it is necessary to consider the role of the time component specifically. As motivated in the introduction, it would be particularly desirable to provide a basis for further research aimed at developing an instrument for providing timely vacancy data or even the nowcasting of vacancy data. This objective presupposes the existence of the data at the reference date of the vacancy estimation. The reference date of the vacancy data set, described in the previous chapter, best suited for this research is 12/31/2019. Thus, ORLs collected prior to 12/31/2019 are included and all listings acquired after that date are excluded. This exclusion reduces the data set from 7,007,571 to 4,206,464 listings.

Existing research using ORL data mainly relies on absolute or relative values of the variables contained in the data sets to clean the data for outliers, as can be seen in Table 17. Under consideration of these approaches, a similar approach is taken in this study to exclude listings that would bias the distribution due to the fact that they are not serious offerings but test or fake offerings. Since the objective of this data cleansing process is explicitly the exclusion of not existing test offerings, a search for the term ‘test’ followed by an arbitrary character string is conducted on the variables *Title*, *Street*, *Designation of Municipality*, *Deposit*, *General Textual Description*, *Textual Description of Facilities*, *Textual Description of Location*, *Textual Description of Miscellaneous* and *Availability*. This procedure reduces the data set from 4,206,464 to 4,206,172 listings.

Outlier cleansing of ORL data based on absolute or relative variable values is described in various studies, such as the study from Boeing & Waddell (2017, p. 460), the study from Breidenbach et al. (2022, pp. 6–7), or the study from Colonnello et al. (2022, pp. 6–7). Most of these studies define cut-off values for data cleansing but do not provide an explanation for the derivation of these values. To derive a more transparent procedure for data exclusion, Table 17 provides a comprehensive overview of previously used values for outlier cleansing of ORL and sales data, based on which a decision can be made for appropriate values relevant to this study.

Table 17

ORL and Sales Data Variable Limits Used in Previous Research

Variable	Limit		Source
	Lower	Upper	
General variable unspecific limits	1st percentile	99th percentile	Breidenbach et al. (2022, p. 7)
	2.5th percentile	97.5th percentile	Jafari & Akhavian (2019, p. 518)
	0.2th percentile	99.8th percentile	Boeing & Waddell (2017, p. 460)
	0.5th percentile	99.5th percentile	R. Gabriel et al. (2008, p. 291)
	Exclusion of listings for which (research-relevant) variables are not specified		Colonnello et al. (2022, p. 7)
			Baldenius et al. (2020, p. 205)
			Bernstein et al. (2019, p. 257)
			Jafari & Akhavian (2019, p. 518)
			Bauer et al. (2017, p. 98)
			Deschermeier et al. (2016, p. 303)
Absolute rent	50 €	99.5th percentile	Colonnello et al. (2022, p. 7)
	10 €	10,000 €	Baldenius et al. (2020, p. 205)
	2 €	115,670 €	R. Gabriel et al. (2008, p. 291)
Rent per sqm	3 €	30 €	Deschermeier et al. (2016, p. 303)
Total price	50,000 $	5,000,000 $	Miller & Pinter (2022, p. 6)
	10,000 €	99.5th percentile	Colonnello et al. (2022, p. 7)
	5,000 €	5,000,000 €	Baldenius et al. (2020, p. 205)
	50,000 $	10,000,000 $	Bernstein et al. (2019, p. 257)
	20,000 €	2,000,000 €	Frondel et al. (2019, p. 5)
	1,000 €	10,000 €	Bauer et al. (2017, p. 98)
	0 €	10,000,000 €	Bauer et al. (2013, p. 11)
	55 €	11,250,000 €	R. Gabriel et al. (2008, p. 291)
Living space (apartments)	10 sqm	99.5th percentile	Colonnello et al. (2022, p. 7)
	5 sqm	400 sqm	Baldenius et al. (2020, p. 205)
Living space (houses)	40 sqm	800 sqm	Frondel et al. (2019, p. 5)
	25 sqm	500 sqm	Bauer et al. (2017, p. 98)
	5 sqm	500 sqm	Bauer et al. (2013, p. 11)
Number of rooms	1	20	Frondel et al. (2019, p. 5)

Variable	Limit		Source
	Lower	Upper	
	-	11	Bauer et al. (2017, p. 98)
	1	10	Bauer et al. (2013, p. 11)
	1	500	R. Gabriel et al. (2008, p. 291)
Building age	-	200 years	Bauer et al. (2017, p. 98)
	- 3 years	199 years	Bauer et al. (2013, p. 11)
Floor	-	88th	Bauer et al. (2013, p. 11)
Plot size	-	20,234 sqm	Miller & Pinter (2022, p. 6)
	20 sqm	5,000 sqm	Frondel et al. (2019, p. 5)

Note. Own research.

Considering the research objective of this work, i.e., the examination of vacancy, the goal is to exclude clearly implausible values since such values probably do not represent actual apartment offerings or substantially bias the distributions of the potential explanatory variables. The detailed examination of the ORL data in Chapter 3.1.2 has shown for the variables *Cold Rent* and *Living Space* that implausible values tend to be found in the lower range below the 0.05th percentile and the upper range above the 99.95th percentile. However, applying these percentiles to both variables yields absolute limits of 15 € and 5,900 € as well as 10 sqm and 329 sqm, which are in the range of the absolute limits applied by other studies. Thus, the 0.05th percentile and the 99.95th percentile are applied for the data cleansing regarding the variables *Cold Rent* and *Living Space*. Since some listings violate the limits for both variables simultaneously, the exclusion is not exactly equal to 0.2 % of the listings, but 4,198,317 listings remain, which equals a reduction of 0.19 % of the listings.

The exclusion of the outliers is based on the full sample. However, for the analysis of the data on the district level, it is beneficial if the sample size per district is not based on a very small number of listings since remaining unrecognized outliers would have a strong influence that could be reduced in larger samples. In order to obtain these information, the unstructured data contained in the geographically descriptive variables *ZIP Code*, *Designation of Municipality*, and *Street* need to be processed to derive precise and automatically analyzable information, described in the following chapter. Examining the result of this data cleansing process shows that all of the 401 districts provide a solid base of listings. The district of Kronach has the lowest number of listings, with 196, and the district of Chemnitz has the highest number of listings,

with 222,741. Compared to the study of Baldenius et al. (2020, p. 205), this exceeds their suggested lower limit of 25 listings per aggregation unit by far.

3.2.2 Preparation of Spatial Information

The examination of the ORLs per district requires the listings to be distinctly locatable at this level of examination, e.g., for assessing the district sample reliability and the derivation of the *Normalized Number of Listings*. Furthermore, structured and automatized evaluable information are needed for many applications. Since the information contained in ORLs do not often exhibit this characteristic when being collected from the platforms due to unstructured free text input, instead of being provided by the platforms, a method for such a mapping is developed. The *ZIP Code* variable and the *Designation of Municipality* variable are provided for all listings, which enables mapping at the municipality level. In principle, mapping at the municipality level is not required for this work due to the availability of vacancy data at the district level. However, it does not cause additional effort and offers the possibility of transferring the developed approach to the municipality level when the data availability improves.

At the municipality level, the German official municipality key (AGS) is an appropriate classification for this purpose, as it distinctly delimits the different municipalities from each other, even if they have identical names. Furthermore, it allows an easy aggregation to higher levels as it simultaneously contains information regarding the district level and the state level, which is needed for this study. To achieve this goal of mapping at the AGS level, a method must be developed to transform the information in the given variables into information corresponding to a correct AGS. Considering the available variables with particular relevance for mapping at the municipality level, *ZIP Code* and *Designation of Municipality*, a list containing all possible correct combinations of ZIP Codes, municipality designations, and AGS enables such a mapping.

Considering only one of the variables given in the ORL data set is not sufficient since various ZIP Codes contain multiple municipalities with unique AGS, e.g., the ZIP Code 68723 contains the municipalities Schwetzingen (AGS: 08226084), Oftersheim (AGS: 08226062), and Plankstadt (AGS: 08226063). Furthermore, the municipality designation alone is not sufficient since there are municipalities with very similar names, e.g., Frankfurt am Main (AGS: 06412000) and Frankfurt (Oder) (AGS: 12053000), or even the same name, Asbach

(AGS: 07134004) and Asbach (AGS: 07138003). However, using both pieces of information makes a unique mapping possible in almost all cases[23].

Since such a list did not exist, a shapefile for Germany containing all ZIP Code areas and a shapefile with the data set VG250-EW from the German Federal Agency for Cartography and Geodesy (BKG) containing all officially existing municipalities, their designation, and their AGS are intersected (BKG, 2019).

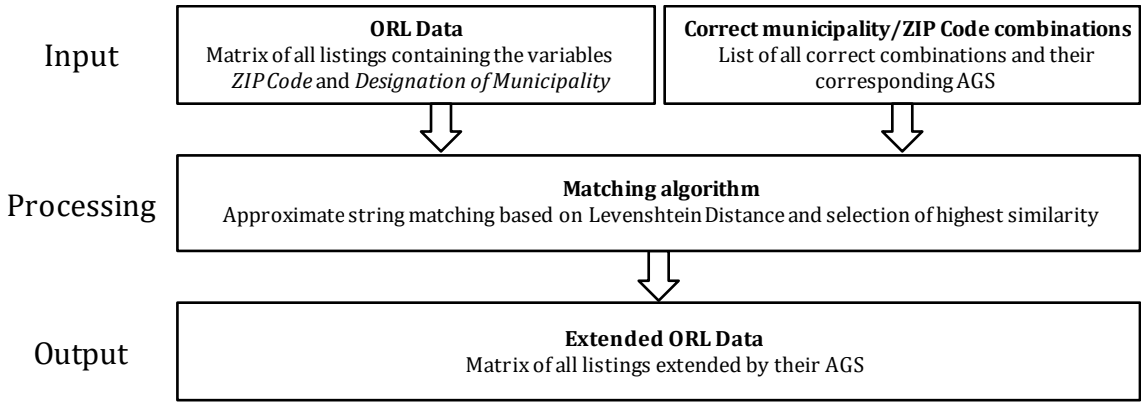
Based on the list from that intersection, a comparison of the variables given, *ZIP Code* and *Designation of Municipality*, of each listing with this list enables the derivation of an AGS for each listing. This comparison, however, is not trivial to be conducted, particularly in an automated manner, since the ZIP Codes can contain errors and the *Designation of Municipality* variable is a free text field that is used very differently by the users. Thus, a comparison method needs to be applied that is able to match, e.g., the following free text input, 'Lützelsachsen, Weinheim' or 'Weinheim - Ortsteil Rippenweier, Weinheim' with the ZIP Code 69469 to the correct specification that consists of the ZIP Code 69469, the designation 'Weinheim', and the AGS 08226096. This can be done by jointly comparing the ZIP Code and the municipality designation based on approximate string matching. Therefore, the commonly used and recognized python library fuzzywuzzy (Bosker, 2021; Gonzalez, n.d.), based on the Levenshtein Ratio (Levenshtein, 1966), is implemented and returns a similarity metric between the user input and ZIP Code prefiltered list of the correct combinations. Thereof, the highest similarity is chosen for each listing and the AGS of this match is assigned to the listing. Thus, the mapping process can be described as shown in Figure 25.

To test the quality of this algorithm, accuracy, defined as a ratio of the number of correct values and the total number of values (Cappiello et al., 2003, p. 84), is evaluated as a performance measure. This evaluation consists of two steps. First, a randomized subset of the database is drawn and individually evaluated. Second, the significance of that evaluation is estimated.

[23] Exceptions are the small municipalities Hamfelde, with the identicial ZIP Code 22929, but with the AGS 01053049 and the AGS 01062026, as well as the municipalities Köthel, located in the same ZIP Code area, but with the AGS 01053070 and the AGS 01062040. However, due to their small size, this concerns only one listing of the entire data set, which was mapped to the larger of both municipalities (AGS 01062040).

Figure 25

AGS City Mapping Procedure



Note. Own research.

For the first step, a sample of 1,000 entries is randomly drawn from the entire database and manually checked for the correctness of the assigned AGS for every individual listing. The correct AGS is determined by an individual manual examination of all characteristics contained in the advertisement that indicate the actual location, e.g., the title, the stated name of the location, the stated ZIP Code, and, if given, the street name and house number, as well as textual descriptions. Those characteristics are combined and evaluated by feeding them into databases, e.g., the ZIP Code database of the Deutsche Post AG, the community directory of the statistics portal by the federal and state statistical offices, and web mapping sites such as Google Maps and OpenStreetMap. By combining the results of those queries, it is assumed that human interpretation enables the derivation of highly reliable results. For the drawn sample of 1,000 listings, an unequivocal mapping was possible in all cases. In 988 cases, the matching process was correct regarding the distinct municipality, corresponding to 98.8 % of the total listings examined. As a relevant share of the incorrect mappings still matched a municipality in the same district, the district correctness is even higher with 995 correct matches, corresponding to 99.5 % of the total listings examined.

To be able to assess the robustness of these ratios regarding the total data base, a binomial test can be used. The variable of interest *correctness of match* is Bernoulli distributed with the possible outcomes $correct\ or\ 1$ with unknown success probability π and $false\ or\ 0$ with unknown probability $1 - \pi$ for each reviewed listing which allows estimating exact P-values with the binomial test as explained in M. Hollander et al. (2014, pp. 11–38), with the null-hypothesis

that the ratio of the sum of the correctness of match variables to the sample size is smaller or equal to a given success rate. For the given validation numbers, the binomial test indicates that at a significance level of 5 %, more than 98 % of listings are mapped to the correct municipality and more than 98.5 % of listings are mapped to the correct district.

3.2.3 Data Aggregation

The data acquisition process has made it possible to collect extensive data on online real estate listings due to the design of the weekly data collection. As a result of this design, the data have characteristics of pooled cross-sectional data and would, in principle, also allow analyses that take the time component into account. However, vacancy rates are assumed to adjust slowly (Rosen & Smith, 1983, p. 780) and would therefore require data over a substantially longer period and additional data on vacancy rates over the same period. Since neither the listing data nor the vacancy data meet this condition, and since there is only one observation of the vacancy rate for each district over the entire data acquisition period, the information contained in the listing data of multiple data acquisitions must be reduced or aggregated to fit the vacancy data.

There are different possible methods to implement this data aggregation, particularly when deciding how to handle recurring listings and the duration that they are listed. The longer a listing is offered, the more relevant it is for the vacancy rate. On the one hand, because it is available for a longer period of time and therefore directly has a greater impact on the calculation of the vacancy rate compared to a listing that is listed for a shorter period of time and on the other hand, because these listings seem to have characteristics leading to the long vacancy duration and should therefore be considered particularly. Since it can be assumed that the characteristics generally influencing the vacancy rate also influence the duration of vacancies, they are likely to be very similar to the characteristics described in Chapter 2.3.4 and selected in Chapter 3.1.3 for the estimation of the vacancy rate. Thus, the vacancy duration is a derived variable whose causal variables are already integrated, which would lead to a high degree of multicollinearity if it were integrated separately in estimating the vacancy rate. Therefore, instead of directly integrating the duration, it can be considered by including the listings according to their offered duration. This inclusion can be achieved by using all observations within the relevant period and by averaging their relevant characteristics without excluding listings that can be found in more than one data collection. An approach similar in the result is also applied in a study for creating a hedonic real estate price index by Bauer et al. (2013, p. 10), who assume that listings collected at different points in time are different listings

even if they concern the same property. Considering the previously described consideration of the importance of longer available listings and the different research objective, which does not include the creation of a price index, a similarly strict assumption is not necessary for this study. Due to the regularity of the data collection, this corresponds to a consideration of the characteristics of the listings weighted by their vacancy duration and can be expressed for every district j as the average of the particular characteristic $\bar{c_j}$ of all listings from i=1 to n that are located in district j:

$$\bar{c_j} = \frac{\sum_{i=1}^{n} c_{i,j}}{|\{i|c_{i,j}\ is\ defined\}|}\ for\ all\ districts\ j\ from\ 1\ to\ 401 \quad (1)$$

which results in a vector $\begin{bmatrix} \bar{c}_1 \\ \dots \\ \bar{c}_{401} \end{bmatrix}$.

Based on this procedure for data aggregation, the frequency distributions of the *Cold Rent per SQM* and the *Living Space* variable can be derived at the district level. Both figures can be found in Appendix A - 25 and Appendix A - 26. However, the type of each of the distributions, in particular, their deviation from the normal distribution, is similar to the results depicted in Chapter 3.1.2. While the rent-related distribution is severely positively skewed, such a substantial deviation can not be found for the living space-related distribution.

3.2.4 Data Transformation

The decision to use transformed or untransformed data must be carefully considered, but generally, the transformation of variables is common practice, e.g., to achieve variables that are closer to normality and to increase the linearity of regression models (Chatterjee & Hadi, 2015, p. 163; Cohen et al., 2003, pp. 249–251). Negative aspects of a transformation can include that the originally existing relationship between two variables is transformed away (Cohen et al., 2003, p. 250) and that a back transformation may be necessary for the interpretation of the results. However, in disciplines such as economics, based on gained experience, some standard transformations have evolved that are usually applied to similar types of variables and can be considered reasonable (Cohen et al., 2003, p. 223). While, depending on the specific application, there are various possible transformations, including, among others, the logarithmic transformation, the reciprocal transformation, and the Box-Cox transformation (Montgomery et al., 2012, pp. 176–182), especially the logarithmic transformation is useful

and thus widely applied since it often has positive effects, such as the removal of heteroscedasticity and the reduction of asymmetry (Chatterjee & Hadi, 2015, p. 180).

Probably the most common reason for using a logarithmic transformation is to reduce positive skewness, resulting in a distribution that is closer to normality and thus typically also resulting in more normally distributed error terms in regression analysis (Hannon & Knapp, 2003, p. 1428). Such positively skewed data is regularly found in the econometric analysis since variables, such as the *Vacancy Rate*, the *Cold Rent per SQM*, or the *Living Space*, integrated into this study have a natural lower bound of zero that cannot be undercut and no upper limit, which causes a longer tail on the right side of the distribution. Thus, it is common practice to apply the logarithmic transformation, in particular the natural logarithm, to similar variables that are positively skewed in the econometric analysis, as can be seen from examples like the study of Bernstein et al. (2019, p. 259) for price data, or the study of K. Wang & Immergluck (2019, p. 521) for vacancy rates. However, this approach is not undisputed, as Wooldridge (2020, p. 210) states that the logarithmic transformation is an unfrequent approach for data given in percent or as proportions. This critique is also reflected in the articles of Du et al. (2018, p. 9) and L. Wang et al. (2019, p. 8577), who, contrary to K. Wang & Immergluck (2019, p. 521), use the vacancy rate in an untransformed form.

Taking all these aspects into account and considering the distribution plots of the different variables provided in Chapter 3.1.3, in Appendix A - 25 and Appendix A - 26, the distinctly positive skewed variables *Vacancy Rate*, *Cold Rent per SQM*, *GDP per Capita* and *Normalized Number of Listings* are transformed by the natural logarithm. However, since the transformation is only indicated but negative consequences can not be excluded and to reflect the critique of the transformation of variables that are given as proportions, the descriptive statistics and the bivariate analysis are provided for the original and the transformed variables. The decision of whether to include the transformed or the untransformed variables in the regression analysis can then be based on an assessment of which form better models a linear relationship.

3.3 Description of Analysis Data

Standard descriptive statistics are provided in the following for the variables included in the analysis data set and the variables initially derived but transformed, which serve as the basis for the subsequent bivariate analysis on which the regression analysis is built. Since it can be assumed that spatial aspects may influence the results of the regression analysis beyond what is indirectly represented by the values of the individual characteristics contained in the data set,

additional information on the spatial distribution of the information contained in the data set are provided.

3.3.1 Basic Descriptive Statistics

Descriptive statistics used in similar contexts to evaluate the data typically include the mean, minimum, maximum, and standard deviation (Baba & Shimizu, 2023, p. 34; Frondel et al., 2019, p. 5; Glumac et al., 2018, pp. 15–16; Hollas et al., 2010, p. 29). However, to evaluate the impact of the data transformation, especially concerning the derivation of more normally distributed variables, additional statistics evaluating skewness and kurtosis are included and displayed in Table 18. Since these descriptive statistics are not reasonably applicable to the categorical data of the *Settlement Type* variable, it is described separately.

Table 18

Basic Descriptive Statistics

Variable	Minimum	Maximum	Mean	Standard deviation	Skewness	Kurtosis
Vacancy Rate in %	0.20	13.30	3.59	2.59	1.16	1.04
ln(Vacancy Rate)	-1.61	2.59	1.00	0.80	-0.44	-0.09
GDP per Capita in €	16,670	194,876	38,943	17,232	3.50	20.80
ln(GDP per Capita)	9.72	12.18	10.50	0.34	1.19	2.42
Number of Listings	196	222,741	10,470	21,428	6.17	49.49
Normalized Number of Listings	0.01	2.99	0.18	0.23	6.36	64.58
ln(Normalized Number of Listings)	-4.55	1.10	-2.12	0.82	0.42	0.48
Cold Rent per SQM in €	4.18	23.71	8.40	2.65	1.43	3.64
ln(Cold Rent per SQM)	1.43	3.17	2.08	0.29	0.45	-0.02
Population Change in %	-3.29	5.61	0.12	0.61	1.09	20.71
Living Space in sqm	57.73	107.81	78.33	9.83	0.01	-0.50

Note. Own research. Information regarding the *Settlement Type* variable can be taken from Figure 21 in Chapter 3.1.3.

The descriptive statistics provide information on various aspects and can be used to check the plausibility of the aggregated data set and the results of the data transformation. In particular, for the untransformed variables, the minimum and maximum values indicate whether the data are plausible even at the tails of the variable distributions. The *Vacancy Rate* variable takes the minimum value for the districts of two of the largest German cities in prosperous metropolitan areas, Frankfurt and Munich, and the maximum value for the district of Greiz, which is, measured by its *GDP per Capita*, economically weak and shows a declining population. The *GDP per Capita* is minimal for the district Südwestpfalz, a sparsely populated district with a declining population as well as a high *Vacancy Rate*, and maximal for the district of Wolfsburg, where the headquarter of the large German car manufacturer Volkswagen is located (Volkswagen AG, 2022, p. 5).

The *Normalized Number of Listings* is minimal for the district Neustadt a.d. Waldnaab, which is with 239 listings in total one of the districts with the least listings and sparsely populated, according to the BBSR classification. In contrast, the district of Chemnitz shows the highest *Normalized Number of Listings,* which is in line with its high *Vacancy Rate* and declining population, however, suspicious in its magnitude. Compared to the districts with the next highest values of Leipzig, 1.31, and Gera, 1.21, as well as compared to the mean of 0.18 and the standard deviation of 0.23, the maximum value of 2.99 is exceptional. A random evaluation of listings shows no abnormality, however, when conducting the bivariate analysis and regression analysis, this value needs to be considered carefully, as it could bias the results for this variable severely. The *Cold Rent per SQM* takes its minimum for the district Lüchow-Dannenberg, which is sparsely populated, has an above-average *Vacancy Rate*, and a below-average *GDP per Capita*. The highest value for the *Cold Rent per SQM* is obtained in the district of Munich, where the *Vacancy Rate* is minimal and the tight housing market is regularly reported. The minimum value for the population change variable is set by the district of Wartburgkreis, which is sparsely populated and has been declining in population for years. The maximum value is set by the district of Suhl, which had a total population of 36,789 in 2019 and can therefore be strongly influenced by relatively small absolute changes in population (Statistische Ämter des Bundes und der Länder, n.d.-b). Similar observations can be made for the *Living Space* variable, for which the extreme values are defined by the districts of Prignitz, minimum, and the district of Neustadt a.d. Aisch-Bad Windsheim, maximum, which both contain a relatively small number of listings. Therefore, a smaller number of extreme listings can influence these districts severely. In summary, the descriptive statistics contribute to the

fact that not only the overall data set but also the adjusted and aggregated analysis data set appears to be generally adequate for further evaluation.

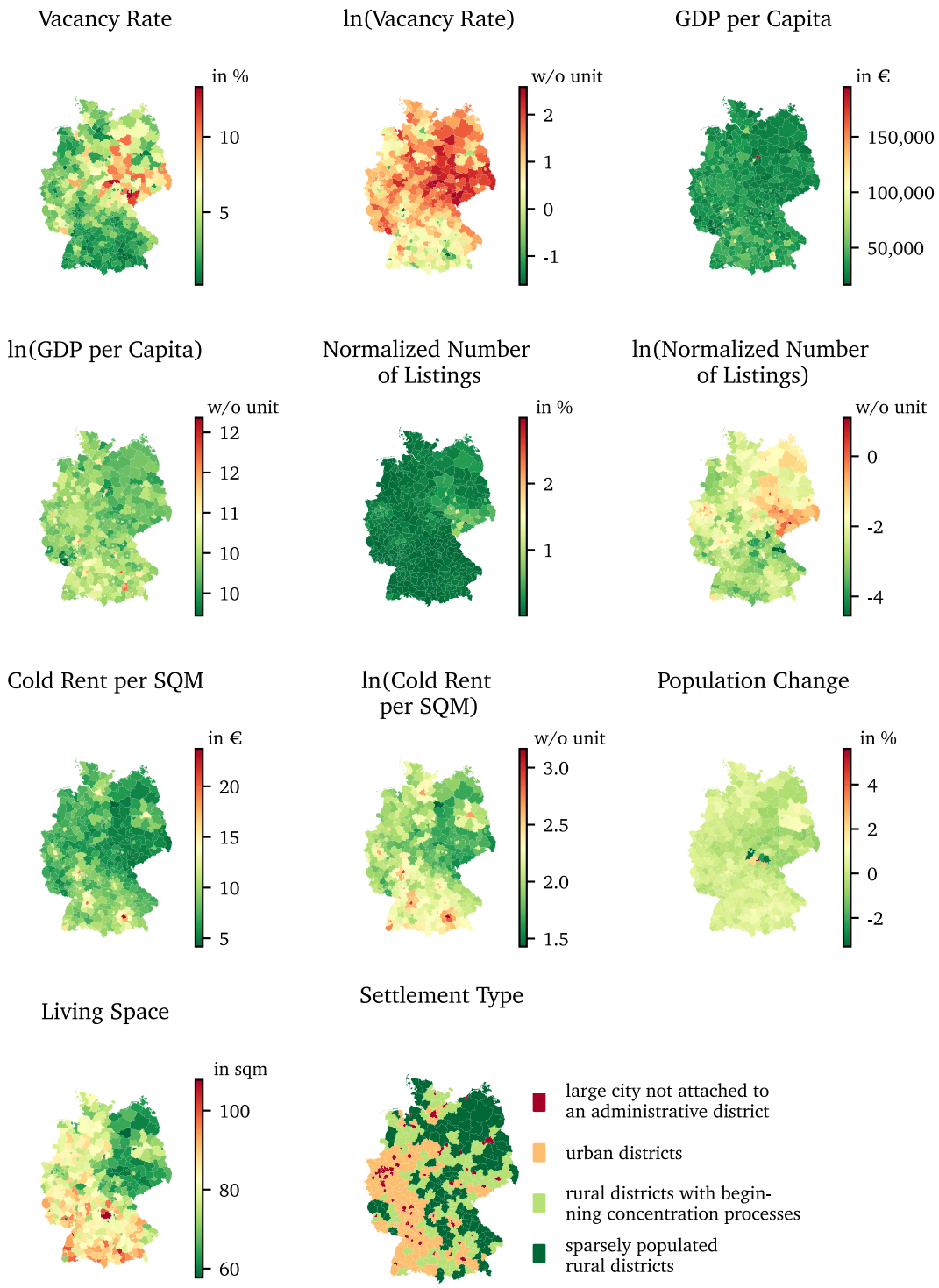
Contrary to the untransformed variables, the transformed variables do not provide intuitively interpretable information. However, the assumption derived from the literature that the logarithmic transformation can reduce skewness and kurtosis (Bishara & Hittner, 2015, p. 788) is confirmed, as all variables transformed by the natural logarithm show a smaller skewness after the transformation. In all cases where the logarithm was applied, both skewness and kurtosis can be classified as moderate according to the assessments by Bishara & Hittner (2015, p. 793) and Bishara et al. (2018, p. 172).

3.3.2 Spatial Description

The real estate market is a market with geographically fixed assets and the significance of the location is widely recognized across a variety of perspectives. The commonly referenced adage that ‘location, location, location’ are the three most important attributes of real estate is widely accepted and known internationally by both, buyers and sellers. Moreover, the significance of the location in the valuation of real estate is even reflected in governmental regulations, such as the German Building Code (§194 BauGB), and is supported by the establishment of a dedicated scientific field, spatial econometrics. According to Tobler’s first law of geography, ”everything is related to everything else, but near things are more related than distant things” (Tobler, 1970, p. 236). In the context of this study, it can be assumed that this quote also applies to the variables in this analysis since geographically close districts regularly show similar behavior, e.g., for their economic or sociodemographic development. Thus, the description of the variables should not be limited to the usual descriptive statistics alone, but the included variables are additionally visualized in Figure 26 in order to evaluate different consequential effects. To enable a comparison of the correlations of the different variables and an assessment of the effect of the transformation of variables, no further classification or transformation of the color scale is applied.

Figure 26

Spatial Distribution of Analysis Data



Note. Own research. Created based on analysis data set.

It is equally apparent that all variables are not randomly distributed, but similar values occur in spatial clusters. This similarity of values indicates that spatial autocorrelation could be prevalent and it is necessary to control the regression analysis results for this phenomenon. Furthermore, utilizing the natural logarithm appears to yield a beneficial outcome in terms of enhancing the visibility of differences, particularly for the variables *GDP per Capita* and *Normalized Number of Listings*. This improvement can be attributed to the impact of highly positive values being mitigated, resulting in better visible relative differences among the ranges where most of the values lie. Nevertheless, it is evident that even in the depictions of the transformed variables, a few extreme values significantly influence the overall spectrum of values, e.g., Wolfsburg with a logarithmized *GDP per Capita* of 12.18 compared to a mean of 10.50 with a standard deviation of 0.34. As these values are not considered outliers due to being erroneous, they will remain in the data set. Nevertheless, conducting further analyses that omit these variables could reveal whether they are unique cases that deviate from the typical correlations and should be removed from the model.

Furthermore, it is not only the distribution within the individual maps that is relevant but also the distributions of the different maps in comparison. Similar structures of the explanatory variables in comparison to the explained variable indicate that a relationship between these variables could exist. These similarities in structure can be seen on a large scale, e.g., the states of Saxony, Thuringia, Saxony-Anhalt, Brandenburg, and Mecklenburg-Vorpommern stand out clearly from the neighboring states in several variables. However, they can also be seen at a smaller scale, for example, in the comparison of *ln(Vacancy Rate)* and *ln(GDP per Capita)* in Munich and the surrounding area. In summary, both the descriptive statistics and the spatial description do not reveal any unusual or unexpected effects, and the data seem suitable for further analysis.

4 Methodology

The overarching research question is subdivided into three concretizing questions. The first of these research questions relates to the characteristics and individual features of the ORL data, each of which could be considered separately and is therefore addressed in Chapter 3.1.2, using univariate analysis methods, which, taken together, provide an overall picture of the ORL data quality. The second and third research questions, concerning the example of vacancy rates, however, examine relationships of two or more variables and require more complex bivariate and multivariate methods of analysis. The choice of these methods exhibits various interdependencies, there is a large variety of possible methods, and the methods themselves require more explanation than the univariate analysis. Thus, the description of the general, quantitative research approach and methodological framework given in Chapter 1.2 is complemented in this chapter by a description of the particular methods to answer the research questions related to the example of the vacancy rate, their requirements, and their interdependencies.

The second research question aims to find relationships between vacancy and other individual variables, while the third research question goes beyond this goal and aims to determine how good vacancy can be estimated using ORL data and, therefrom, derived variables in combination with other data. The quantitative analysis of such a problem, which involves several variables that can be considered as independent variables, in this case, the ORL data and additional data, and a single dependent variable to be explained or estimated, in this case, the vacancy rate, suggests the use of linear regression analysis since it is considered in fundamental econometric literature as the most useful tool for this type of problems (Greene, 2018, p. 13). Furthermore, the possible application of linear regression analysis is already shown for vacancy research (Du et al., 2018; Hagen & Hansen, 2010) and for research utilizing ORL data (Holt & Borsuk, 2020; Kholodilin et al., 2017). However, in this work, the standardized multiple linear regression is used as a first step, and a spatial regression model is additionally utilized, as it can be assumed that location significantly impacts vacancy rates beyond the application of dummy variables.

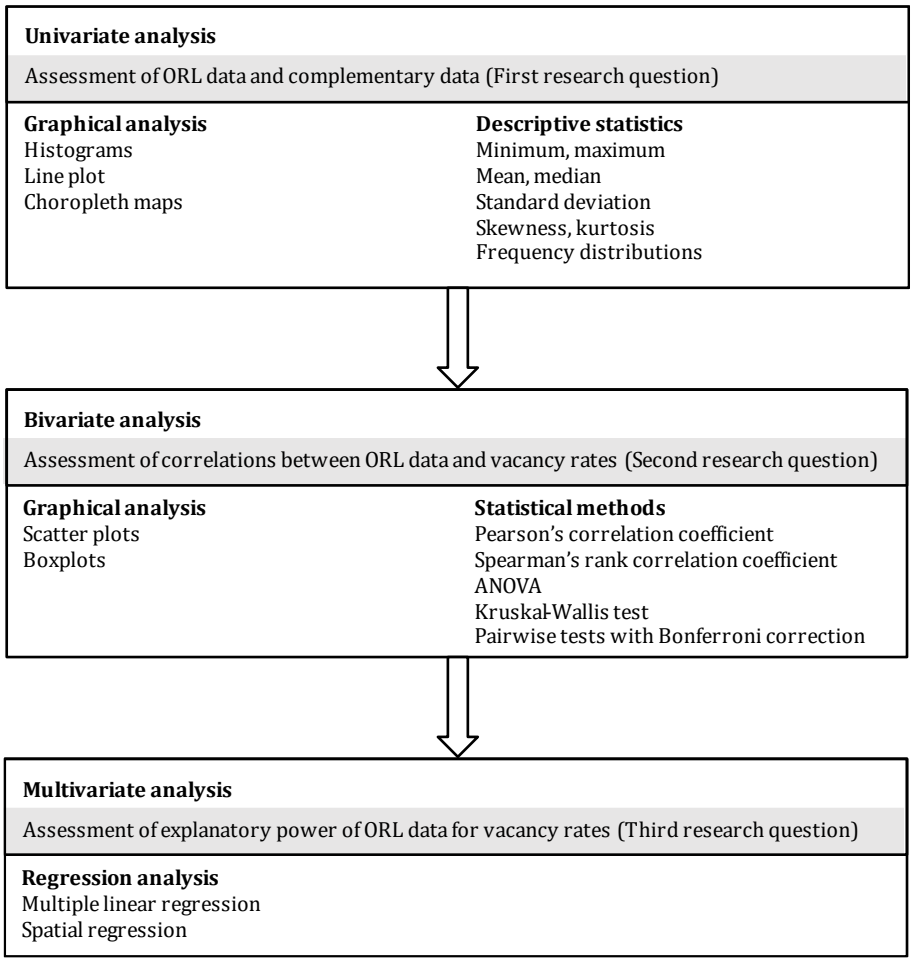
In general, a careful inspection of the data before conducting a regression analysis is fundamental to good data analysis practice and will greatly increase confidence in the regression analysis results (Cohen et al., 2003, p. 101). Such careful inspection is usually conducted by univariate analysis at the level of individual variables, e.g., using histograms or descriptive statistics such as the determination of the skewness and standard deviation to check

the variable distribution, and by bivariate analysis, e.g., using scatter plots and correlation coefficients to check the direction, strength, and linearity of relationships. Similar approaches can, for example, be found in the work of Jafari & Akhavian (2019, pp. 518–521) using ORL data for the analysis of influences on U.S. housing prices and in the study of L. Wang et al. (2019, p. 8573) trying to estimate the housing vacancy rate based on satellite data.

Figure 27 brings together all the quantitative methodological approaches in a comprehensive overview and shows how the descriptive studies from the previous chapter and the analysis conducted for the vacancy estimation example are related and which methods are part of them.

Figure 27

Statistical Methodological Framework



Note. Own research.

4.1 Bivariate Analysis

The main objective of the bivariate analysis is to identify relevant variables and examine their relationship with the vacancy rate, which is central in the early stages of modeling (Fox, 2016, p. 47). The examination of the relationship allows conclusions to be drawn about the existence, type, e.g., linear or non-linear, direction, and strength of the relationship. A bivariate analysis can be conducted by calculating statistical measures that provide a standardized and objective statement, but usually, an additional graphical analysis is strongly recommended (Field, 2017, p. 464; Fox, 2016, p. 47) since especially the visual inspection allows detecting nonlinear relationships (Schober et al., 2018, p. 1767), which can lead to similar correlation coefficients compared to imperfect linear relationships but require different types of models.

The literature review, particularly in Chapter 2.3.4, and the univariate analysis in Chapter 3.1.2 have laid the groundwork for the variable selection and provided a preselection that is depicted in Table 18 and that includes continuous[24] as well as categorical variables. The bivariate analysis examines these variables for a more detailed analysis of which variables to include in which form in the model. Since the methods applicable to continuous and categorical variables differ, they are described separately.

For the case of continuous variables, a visual inspection can be conducted using scatter plots since scatter plots are referred to as the most useful statistical figure for visualizing bivariate data and are regularly recommended (Fox, 2016, p. 40; Schober et al., 2018, p. 1767). They especially allow assessing the relationship between the two depicted variables (Cohen et al., 2003, p. 21). A basic scatter plot consists of two axes, on each of which the values of one variable are defined. Scale and measurement units can vary between the axes. A marker depicts each observation. The entirety of the markers results in a shape that contains information about the behavior of the represented variables in relation to each other. It can show, for example, that an in- or decrease in one variable is typically associated with an in- or decrease in the other variable. Furthermore, it allows for detecting the shape of such a relationship and thereby contributes to the model specification. However, since a basic scatter plot has only two dimensions, it is important to note that the influences of other variables are neglected, which can be particularly problematic when both depicted variables and another omitted variable are correlated. This limitation of scatter plots is considered in this study by choice of the variable

[24] Mathematically correct, these are bounded continuous variables, but for the sake of simplicity they will be referred to as continuous variables in the following, since the difference does neither influence the bivariate analysis nor the multivariate analysis. However, if a model with the purpose of prediction shall be built, this difference needs to be considered for the dependent variable.

combinations depicted, especially by the additional depiction of independent variables among themselves. However, the main objective of the bivariate analysis is to determine which variables to include in which way in the regression analysis. Thus, the potentially independent continuous variables, *GDP per Capita*, *ln(GDP per Capita)*, *Normalized Number of Listings*, *ln(Normalized Number of Listings)*, *Cold Rent per SQM*, *ln(Cold Rent per SQM)*, *Population Change*, and *Living Space*, are plotted against the potentially dependent continuous variables, *Vacancy Rate* and *ln(Vacancy Rate)*. In addition, scatter plot matrices are provided for a model consisting only of the transformed variables and for a model consisting only of the untransformed variables in order to assess the relationships among the independent variables and to address the aforementioned correlation problem. To further increase the objectivity and reliability of the findings derived from the scatter plots, they are supplemented with correlation coefficients and associated significance levels.

The choice of the specific correlation coefficients depends on the given data. The most common correlation coefficients applied in the context of regression analysis include Pearson's r, Spearman's ρ, and Kendall's τ (Field, 2017, p. 464). Of these, Pearson's r is probably the best-known, and Spearman's ρ the best-known alternative when Pearson's r should not be applied due to violations of application requirements (Bishara & Hittner, 2012, p. 400; de Winter et al., 2016, p. 284). Kendall's τ is less common and instead recommended for small sample sizes (Bishara & Hittner, 2012, p. 400), which is not relevant in this study, why only Pearson's r and Spearman's ρ are taken into consideration since the sample size can be regarded as large with 401 observations. In principle, in the given case of continuous data, there are no prerequisites or constraints for applying a correlation coefficient (P. Chen & Popovich, 2002, p. 13; Schober et al., 2018, p. 1764). However, for expressing the significance of Pearson's r through a significance level, the data needs to be representative and both variables need to jointly follow a normal distribution (Schober et al., 2018, p. 1764). The representativity can be neglected as the 401 districts represent a full sample and, according to a variety of recognized publications, the requirement of jointly normal distributed variables can often be relaxed, and Pearson's r is quite robust to non-normally distributed variables (Bishara & Hittner, 2012, p. 402), especially when the sample size is not small (Field, 2017, p. 463). As described in Chapter 3.3.1, the transformed variables are close to normality since their skew can be assessed as weak, which implies the use of Pearson's r. However, since this view is not entirely undisputed and, depending on the specific application, both coefficients have their advantages (Bishara & Hittner, 2012, pp. 402, 411; de Winter et al., 2016, p. 286; van den Heuvel & Zhan, 2022, p. 50), Pearson's r as well as Spearman's ρ are calculated.

Pearson's r is calculated as follows (Rodgers & Nicewander, 1988, p. 61):

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}} \tag{2}$$

with m_x as the mean of x and m_y as the mean of y.

Spearman's ρ is calculated identically except for the difference that for x and y, the fractional ranks x_r and y_r are used (de Winter et al., 2016, p. 276):

$$\rho = \frac{\sum(x_r - m_{x_r})(y_r - m_{y_r})}{\sqrt{\sum(x_r - m_{x_r})^2 \sum(y_r - m_{y_r})^2}} \tag{3}$$

Finally, both correlation coefficients provide similar but not identical information. While Pearson's r can provide information about the degree of linearity between two variables, Spearman's ρ provides information about their joint monotonicity (de Winter et al., 2016, p. 284; Schober et al., 2018, p. 1767). Thus, the advantages of both correlation coefficients are considered and both measures are integrated into the scatter plots with their related significance measures.

In contrast to the continuous variables, the *Settlement Type* variable defines the type of settlement for each of the 401 districts according to the classification of the BBSR that differentiates between the types 'sparsely populated rural district', 'rural district with beginning concentration processes', 'urban district' and 'large city not attached to an administrative district'. This differentiation is a differentiation into categories that has no natural order. Thus, the variable has a nominal scale and the tests applied for the continuous variables are not applicable. However, it is relevant to examine if the dependent variable differs significantly from category to category since an even distribution of the dependent variable between the categories implies that the categories can not contribute to the explanation of the dependent variable.

Such an analysis is typically conducted by the Analysis of Variance (ANOVA) (Kim, 2017, p. 22), which is applied, for example, in the vacancy research context by Morckel & Durst (2023, pp. 316, 319). The application of ANOVA requires random sampling, independent measures, normality of the different categories, and equality in variance (Doncaster & Davey, 2007, pp. 14–15). Due to full coverage, the first two assumptions are not problematic. For the *ln(Vacancy Rate)* variable, all measures indicate normality. However, for the untransformed *Vacancy Rate*

variable, a slight skew is indicated[25]. The sample size of each category is at least 67, which implies that due to the central limit theorem (CLT), ANOVA can also be applied to the untransformed variables with respect to the normality assumption.

The assumption of equal variance between the groups can be tested by the regularly recommended Bartlett's test (McGuinness, 2002, p. 682). Despite the test implying heterogeneity of variance between the groups[26], the extent of heterogeneity does not seem to be problematic, as McGuinness (2002, p. 681) summarizes standard literature implying relative robustness of ANOVA against heterogeneity of variance, leading to a rule of thumb of a difference of the smallest and largest standard deviation by the factor 2, which is fulfilled for the *Vacancy Rate* case as well as the *ln(Vacancy Rate)* case, as can be seen in Appendix B - 11. The F-test statistic for the ANOVA is calculated as follows (Kim, 2017, p. 24):

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2/(N-k)} \tag{4}$$

with N as the total sample size, k denoting the number of groups, n_i as the sample size of the particular group, $\bar{Y}_i$ as the sample mean of group i, $\bar{Y}$ as the overall mean and Y_{ij} as the j^{th} observation in the i^{th} group.

However, consistent with the continuous variable approach, an additional approach is performed to confirm the results, where the assumptions of the ANOVA do not have to be met. For multiple groups, the non-parametric Kruskal-Wallis test is a common choice (Field, 2017, pp. 388–389, 417–418). The Kruskal-Wallis test statistic is calculated as follows (Ostertagová et al., 2014, p. 116):

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{{R_i}^2}{n_i} - 3(N+1) \tag{5}$$

with N as the total sample size, k denoting the number of groups, n_i as the sample size of the particular group and R_i as the sum of the ranks.

Both tests, the F-test for the ANOVA and the Kruskal-Wallis test, only provide information about the joint difference of all group means. However, one is usually interested in which groups, in particular, differ from each other, which can be done by pairwise comparison and consideration

[25] In the appendix, histograms, Appendix B - 1 to Appendix B - 8, q-q plots, Appendix B - 9 and Appendix B - 10, and a table containing information regarding skewness and kurtosis, Appendix B - 11, are provided for this assessment.

[26] Bartlett's test statistic: 11.6939, P-value: 0.0085

of the Bonferroni correction (Field, 2017, p. 715; Kim, 2017, p. 26). The Bonferroni correction adjusts the critical α-value by dividing it through the number of pairwise tests.

Similarly to the previous case of exclusively continuous variables, it is also reasonable to conduct an additional visual inspection of the data, which can, in this case, be done by a figure that contains multiple boxplots, one for each category (Backhaus et al., 2018, pp. 166–167). The boxplots give information about the distribution of the dependent variable for each subgroup by displaying the median, quartiles, and extreme values.

4.2 Regression Analysis

The regression analysis aims to provide results to answer the third research question by providing information about the informational value of ORL data. The variety of potentially applicable regression models is enormous, but existing research in similar fields of application can provide a guideline for choosing appropriate models. In particular, the possible influence of spatial information derived from the spatial description in Chapter 3.3.2 leads to an increasing amount of modeling approaches that include multiple different types of spatial regression models. However, according to a rule known as Occam's Razor, models should always be kept as simple as possible, as these models are typically better generalizable and easier to interpret (Blumer et al., 1987). The interpretation aspect becomes particularly relevant for the spatial regression models, as their estimates are known to be more difficult to interpret (LeSage & Pace, 2009, p. 34). Thus, for models including spatial information, it is not an uncommon approach to test standard multiple linear regression analysis or to include the spatial information in an easy-to-interpret way, e.g., by dummy variables, as can be seen from the following examples, prior to more complex methods of spatial modeling.

In their analysis of the influence of vacancy on rent prices, for example, Baba & Shimizu (2023, pp. 31–32) conduct an ordinary least square (OLS) approach before conducting a regression model that includes a matrix with spatial weights. Similarly, Yue et al. (2022, p. 7) test a standard OLS model before applying a spatial error model (SEM) and a spatial lag model (SLM) in their research examining the influence of floor area ratio and plot area on vacancy rates. A similar approach is taken in this work by testing an OLS model prior to a more sophisticated spatial regression model. The OLS model created includes spatial information, the settlement types defined by the BBSR as dummy variables, a common approach in spatial modeling, as can be seen from the following examples.

Controlling for effects on a small scale, Jafari & Akhavian (2019, pp. 518–520) use dummy variables to test if the availability of specific local amenities within a certain distance affects real estate prices. On a larger scale, an de Meulen et al. (2011, p. 7) include regional dummies that correspond to the division of municipalities in Germany in their approach of forecasting house prices, which is similar to the approach of Bauer et al. (2015, pp. 273–274) who control for the location of municipalities within a specific region or outside of that region by dummy variables in their study examining the effects of a large infrastructure project on real estate prices. The influence of being located in a particular municipality is also considered by dummy variables by the approach of Frondel et al. (2019, pp. 9–12), examining the effects of wind turbines on housing prices. Hollas et al. (2010, pp. 28–32) use ZIP Code dummies to check if the location in a specific ZIP Code area influences the difference between price prediction and sales prices. The study by Micheli et al. (2019, pp. 770–771) integrated dummy variables on an even larger scale to control if houses are located in one country or another. Based on these approaches, the methodology of multiple linear regression and the implementation of a spatial regression model are described. In particular, the requirements are addressed, and the choice of models is explained.

4.2.1 Multiple Linear Regression

The linear regression model, considered in fundamental statistical textbooks as the tool that is most useful for general econometric analysis, regularly represents the starting point for further analysis (Greene, 2018, p. 13). However, several assumptions must be fulfilled that linear regression can be applied without constraints. Deviations from these assumptions can have different consequences, including biased estimates or unreliable test statistics (Backhaus et al., 2018, p. 103). These assumptions and the rationale for the methods used to verify that they are met are given below, and the order of the examination is based on Backhaus et al. (2018, pp. 89–103). However, diagnosing violations of these assumptions regularly requires the regression to be already conducted, so in the results chapter, the regression results are reported prior to subsequent testing of the satisfaction of these assumptions.

The first assumption, the model specification assumption, consists of three parts regarding linearity, included variables, and the number of parameters that need to be estimated. The linearity assumption requires the relationship between the independent variables and the dependent variable to be, at least approximately, linear (Backhaus et al., 2018, p. 90; Cohen et al., 2003, pp. 117–119; Greene, 2018, p. 17; Montgomery et al., 2012, p. 129). This means that

the relationship needs to be linear in its parameters, not necessarily in its variables, which enables the previously discussed transformations of the variables by the natural logarithm (Backhaus et al., 2018, p. 91). To detect model misspecifications, graphical analysis is particularly recommended, as it allows for visualizing potential misspecifications, which is the basis for handling them (Cohen et al., 2003, p. 126). Regularly used for this purpose are scatter plots that plot the residuals against the predicted variable, including a smoothed residual line (James et al., 2021, pp. 93–94) and scatter plots that plot the residuals against the explanatory variables with an additional lowess line (Cohen et al., 2003, p. 125). Both types of plots are given in the results. Finally, tests for autocorrelation and heteroscedasticity, both presenting violations of model assumptions given below, can hint at a non-linear relationship (Backhaus et al., 2018, p. 93).

Since vacancy research is an established field of research, as demonstrated by the large number of articles in Chapter 2.3, the selection of variable groups in Chapter 2.3.4 is based on a broad knowledge base and the probability of omitting an essential group of variables can be considered low. The last part of the model specification assumption requires the number of parameters to be estimated to be smaller than the number of observations (Backhaus et al., 2018, p. 90), which is fulfilled as the number of observations is equal to the number of 401 German districts.

The second assumption states that the disturbances have an expected value of zero, which is implicitly integrated into the regression model by the forced zero mean of the residuals (Backhaus et al., 2018, p. 90). Deviations from this assumption usually occur when there are systematic errors in the measurement of the dependent variable, particularly when no constant term is integrated into the analysis (Backhaus et al., 2018, p. 93; Greene, 2018, pp. 22–23). In this study, there is no evidence of systematically biased values of the dependent variable nor a constant term left out, so no further verification of this assumption is necessary.

Furthermore, there should be no correlation between independent variables and disturbances (Backhaus et al., 2018, p. 90). However, since disturbances cannot be measured in contrast to the error terms, this can only be detected indirectly, e.g., by inspecting the residual scatter plots (Greene, 2018, pp. 22–23). Additionally, this only poses a problem if the first assumption is violated. In this case, the consequence would be biased parameter estimates (Backhaus et al., 2018, pp. 93–96). Both aspects, variable selection and inspection of residual visualizations, are already addressed for the previous assumptions.

A common problem in regression analysis is the occurrence of heteroscedasticity, which is associated with inefficient parameter estimates (Backhaus et al., 2018, pp. 94–96, 103). In the presence of heteroscedasticity, the estimators are still unbiased but no longer efficient, i.e., they no longer have the smallest possible variance and the test statistics may be incorrect (Backhaus et al., 2018, p. 103). Equivalently to the inspection of the linearity of the relationship between the dependent variable and the independent variables, a visual inspection of the scatterplot of the residuals and the estimated values is recommended (Backhaus et al., 2018, p. 95; Chatterjee & Hadi, 2015, p. 171; Field, 2017, p. 357). When conducting additional quantitative tests, large sample sizes can lead to erroneously significant test statistics (Field, 2017, p. 359) so visual inspection should be prioritized. However, there are several tests to test for heteroscedasticity, including the commonly mentioned Breusch-Pagan (BP) test and White's test (Cohen et al., 2003, p. 133; Greene, 2018, pp. 313–314; Wooldridge, 2020, pp. 270–273). For the sake of completeness, both tests are applied additionally to the inspection of the scatter plots. The BP test is calculated by regressing the obtained residuals from the regression analysis and regressing them against the independent variables from the base regression analysis resulting in an R^2 that is used to derive either a Lagrange multiplier (LM) test statistic or an F-test statistic (Greene, 2018, p. 314; Wooldridge, 2020, pp. 270–271). The test's null hypothesis assumes homoscedasticity (Wooldridge, 2020, p. 270). The White's test is similar to the BP test, however more general, since it additionally includes squares of the independent variables and cross-products, but except this difference is calculated similarly (Greene, 2018, p. 314; Wooldridge, 2020, pp. 271–272).

Typically, a problem in time series data, but also in spatial data, is the presence of correlation between the error terms of different observations, known as autocorrelation (Backhaus et al., 2018, p. 96; Chatterjee & Hadi, 2015, pp. 209–210; Cohen et al., 2003, pp. 134–136). Autocorrelation does not bias the parameter estimates, but the regression becomes inefficient and standard errors, and consequently test statistics, become invalid (Backhaus et al., 2018, p. 103; Field, 2017, p. 514; Montgomery et al., 2012, p. 475). Usually, the Durbin-Watson test is recommended to test for autocorrelation, which, however, is only reasonable for residuals with a natural order, i.e., time series data (Backhaus et al., 2018, p. 97; Chatterjee & Hadi, 2015, pp. 212–214; Cohen et al., 2003, pp. 136–137; Montgomery et al., 2012, p. 475). The standard procedure for diagnosing the lesser-known alternative of spatial autocorrelation is the calculation of Moran's I (Dubin, 1998, p. 319; Fischer & Getis, 2010, p. 257; Greene, 2018, p. 423) as follows:

$$I = \frac{N(e'We)}{S(e'e)} \tag{6}$$

with N as the total sample size, e as the residual vector, S as the standardization factor, and W as the row-standardized weights matrix.

Nevertheless, other prominent alternatives and enhancements exist, including the differentiation between global and local autocorrelation and visualizations such as the Moran scatter plot (Fischer & Getis, 2010, p. 82). Central to these approaches is the definition of a spatial weights matrix that defines the type of spatial relationship between the different observations, for which the possibilities are numerous, including the prominent distance and neighborhood matrices (Fischer & Getis, 2010, p. 260). The main criterion for selecting a spatial weights matrix should be the objective to model the spatial dependencies as accurately as possible, meaning explicitly that the weights matrix should reflect the understanding of the examined spatial dependencies or the assumed theoretical relationships (Fischer & Getis, 2010, p. 260). A purely quantitatively driven approach, for example, an optimization of the R^2, is usually opposed (Kelejian & Piras, 2017, p. 87). A distance decay function seems more natural for the problem of district vacancy rates than a contiguity function. Reasons for that assumption are the varying size of the districts, the varying number of neighboring districts, and especially that the distance from one district to another seems more important for choosing a housing location than the shared border with another district. Thus, a distance-dependent weight matrix is used.

Furthermore, there should not be a perfect linear relationship between the explanatory variables, which would make the estimation of the parameters technically impossible (Backhaus et al., 2018, p. 98; Cohen et al., 2003, pp. 419–420; Greene, 2018, pp. 20–22). A certain degree of correlation between the independent variables is much more common and unavoidable than perfect multicollinearity, which can lead to problems such as lower precision of the parameter estimates, depending on the level of correlation (Backhaus et al., 2018, pp. 98–99; Field, 2017, pp. 533–534; Montgomery et al., 2012, p. 286). This is because it is no longer possible to determine which explanatory variables contribute to which extent to the explanation of the dependent variable (James et al., 2021, p. 100). Common approaches for detecting multicollinearity are inspecting correlation matrices (Montgomery et al., 2012, pp. 292–294) and calculating the variance inflation factor (VIF) and the condition number (Greene, 2018, p. 95). The VIF is calculated by regressing each independent variable on all other independent variables and using the R^2 from this regression in the following formula (Greene, 2018, p. 95):

$$VIF = \frac{1}{1 - R^2} \tag{7}$$

According to Greene (2018), the condition number is calculated from the matrix of all independent variables X and the derived matrix X'X. The columns of the matrix X'X are scaled to unit length. From this matrix, the largest eigenvalue is divided by the smallest eigenvalue and from this value, the square root is taken (Greene, 2018, p. 95). Typical limits that indicate a problematic degree of multicollinearity are 5 to 10 for the VIF and 1000 for the condition number (Montgomery et al., 2012, pp. 296–298). The examination of the correlation matrix is provided within the results of the bivariate analysis and the values for the VIFs and condition number are provided together with the results of the regression analysis.

Techniques in dealing with multicollinearity are various and the choice of the specific technique depends on the desired results. Since it is highly effective and the underlying data suggest that no major relevant information is lost, variable elimination is considered in this work (Montgomery et al., 2012, p. 304). When applying variable elimination, it is essential to reconsider whether the model is still adequately specified or if the first assumption that it is correctly specified could be violated. Therefore, an explanation of the choice of the eliminated variables is given in the discussion of the results. The advantages of variable elimination, e.g., in comparison to principal component analysis (PCA), include that, if carefully chosen, the variables in the regression analysis are straightforward to interpret and it is unlikely that the technically constructed latent variables of the PCA correspond to the assumed real-world relationship (Greene, 2018, p. 97). However, if the results suggest that the model should include some ambiguously defined variables that are mixtures of included variables, a PCA should be considered (Greene, 2018, p. 97). This could particularly be the case if multiple correlated explanatory variables show a similar degree of explanatory power.

Finally, the disturbances should be normally distributed (Backhaus et al., 2018, p. 90). This assumption is usually checked by visual inspection of the residuals via scatter plots, histograms, and q-q plots (Chatterjee & Hadi, 2015, p. 105; Cohen et al., 2003, pp. 137–139). However, the normality assumption is not strict in this work since the sample size is large and therefore, even significant deviations from normality do not lead to problems of the regression analysis (Field, 2017, p. 346). Applying statistical tests to check for normality is specifically unrecommended for large samples as they often lead to significant results due to the sample size but do not have any practical implications for the regression analysis (Field, 2017, p. 346).

Assuming all previous assumptions are met, the multiple linear regression model can be specified based on the results of the bivariate analysis, which is especially relevant for the choice

of how the variables are included, e.g., transformed or untransformed. Worth mentioning is that the *ln(Normalized Number of Listings)* is not included singularly corresponding to the other variables but, based on the results of the bivariate analysis that suggest a changing relationship for the different settlement types, in the form of an interaction term in combination with the *Settlement Type* variable. This approach is common in vacancy research, as can be seen from the examples from Baba & Hino (2019, p. 370) or Morckel (2014, pp. 13–15) and explicitly allows modeling changing relationships over different areas. Thus, taking into account the results of the bivariate analysis and using abbreviations for the written-out variable names for the sake of conciseness[27], the base regression model can be formulated as follows:

$$\begin{aligned}\ln(vac) = \beta_0 + \beta_1 \ln(gdp) + \beta_2 \ln(cr) \; + \beta_3 pc + \beta_4 ls + \beta_5 sparse + \beta_6 rural \\ + \beta_7 urban + \beta_8 (\ln(nnl) \text{ x } sparse) + \beta_9 (\ln(nnl) \text{ x } rural) \\ + \beta_{10} (\ln(nnl) \text{ x } urban) + \beta_{11} (\ln(nnl) \text{ x } city) + \varepsilon\end{aligned} \tag{8}$$

In this base model, the settlement type *city* is used as the reference category to avoid creating a perfect linear relationship by including all settlement type categories and violating the assumption regarding multicollinearity. Additionally, it can only be seen as a starting point because after checking the regression assumptions, there are usually changes in the model, which are described in Chapter 5.2.1. Furthermore, a spatial model is tested as an extension, for which this base model also represents the initial model.

For both the standard linear regression model as well as the spatial regression model, the mean absolute error (MAE) of the retransformed *Vacancy Rate* variable is evaluated as this measure can easily be interpreted as the average absolute deviation from the actual vacancy rate value and is thus better suited for the validation by experts. However, it needs to be mentioned that this measure is less widespread and always smaller or equal to the more complex to interpret but more common root mean square error (RMSE). Thus, the MAE of the *Vacancy Rate* variable is calculated as follows:

$$MAE = \frac{\sum_{i=1}^{N} |e^{\ln(\widehat{vac}_i)} - vac_i|}{N} \tag{9}$$

[27] vac: Vacancy Rate; gdp: GDP per Capita; cr: Cold Rent per SQM; pc: Population Change; ls: Living Space; sparse: Sparsely Populated Rural District; rural: Rural District With Beginning Concentration Processes; urban: Urban Districts; city: Large City Not Attached to an Administrative District; nnl: Normalized Number of Listings.

4.2.2 Spatial Regression

The common assumption of the cause of spatial autocorrelation is that spatial elements close to each other often influence each other (J. B. Hollander et al., 2018, p. 600) and that this influence decreases with increasing distance (Tobler, 1970, p. 236). In some cases, these effects are explainable by other variables that are similar for elements close to each other. However, in some cases, the variables included in an analysis cannot fully cover the observable effect that adjacent elements exhibit similar patterns. Tests for spatial autocorrelation of the residuals are used to detect such spatial effects that the model does not explain. Moran's I, explained in the previous chapter and implemented in Chapter 5.2.1, as the standard procedure for detecting spatial autocorrelation, is also used in related research, e.g., in the study of J. B. Hollander et al. (2018, p. 600) examining vacancy in shrinking downtowns, the work of K. Wang & Immergluck (2019, p. 521) trying to explain changes in long-term vacancy, or the examination of housing abandonment of Morckel (2014, p. 10).

When spatial dependencies are assumed or diagnosed by a test for spatial autocorrelation like Moran's I, they should be integrated, which was regularly done through the inclusion of spatial dummies in the past (Glumac et al., 2018, p. 8). However, more elaborate spatial models have been developed that can model spatial effects in more detail and that include spatial dummies only to complement these models. From these models, the SLM and SEM belong to the most recognized spatial models, are widely used, and thus present a reasonable basis for spatial regression analysis (Dell'Anna & Bottero, 2021, p. 3; Grekousis, 2020, p. 452). The SLM is appropriate when it is assumed that the dependent variable influences surrounding values of its own, and the SEM, if it is assumed that critical explanatory variables are missed (Locke & Baine, 2015, p. 398). Due to the extensive amount of existing literature, based on which the variables were chosen, and the high value for the adjusted R^2, it is not assumed that relevant variables are omitted. However, it seems plausible that high values for the vacancy rate in one district lead to, c.p., higher values in surrounding districts by themselves, as higher vacancy rates have a price-decreasing effect that could exert a pull effect on close districts, arguing for a SLM. In addition to this substantive rationale, a decision can also be made between the SEM and the SLM using a quantitative selection procedure from Anselin & Rey (2014, pp. 109–111). They recommend conducting LM error and LM lag tests and, additionally, robust LM error and robust LM lag tests, if both are significant, to choose the correct model (Anselin & Rey, 2014, p. 110). The tests are implemented as described in Anselin et al. (1996).

Taking into account the results from the standard OLS regression, implemented in Chapter

5.2.1, and from the specification tests, implemented at the beginning of Chapter 5.2.2, the thereof derived SLM can be formulated as follows:

$$\ln(vac) = \beta_0 + \beta_1 \ln(cr) + \beta_2 pc + \beta_3(\ln(nnl) \text{ x } sparse) + \beta_4(\ln(nnl) \text{ x } rural) + \beta_5(\ln(nnl) \text{ x } city) + \rho W ln(vac) + \varepsilon_i \quad (10)$$

with a reduced form of the base regression model and $\rho W ln(vac)$ as the spatial lag term.

Due to the inclusion of the dependent variable as a spatially lagged explanatory variable, a closed-form solution is not possible and it needs to be specified which method for parameter estimation is used. Usually, the maximum likelihood (ML) method and the spatial two-stage least squares method are considered, of which it can be assumed that the ML method may lead to better results (Grekousis, 2020, p. 467). Thus, the ML method is chosen for parameter estimation.

The interpretation of SLMs is more complicated than the interpretation of standard OLS models as these models include spatial spillover effects with a cascading nature due to the inclusion of the spatial lag term, resulting in self-influence of the dependent variable (Golgher & Voss, 2016, p. 180; LeSage & Pace, 2009, p. 34). Thus, changes in individual observations do not only affect the region where they are located directly but potentially also all other regions indirectly by these spillovers (LeSage & Pace, 2009, p. 33). Therefore, the interpretation of the coefficients should not be based directly on the parameter estimates but on direct, indirect, and total effects derived from them (Golgher & Voss, 2016, p. 180). According to LeSage & Pace (2009, pp. 34–39) and Holt & Borsuk (2020, pp. 5–6), they can be calculated as follows:

$$S_k(W) = (I_n - \rho W)^{-1} I_n \beta_k \quad (11)$$

$$Total\ effect = \frac{1}{n} i_n' \big(S_k(W)\big) i_n \quad (12)$$

$$Direct\ effect = \frac{1}{n} tr(S_k(W)) \quad (13)$$

$$Indirect\ effect = Total\ effect - Direct\ effect \quad (14)$$

with $S_k(W)$ as the partial derivative matrix for variable k, I_n the identity matrix with dimension (n x n), ρ the parameter from the regression analysis for the variable $W * \ln(vac)$, β_k the parameter of the variable k, and i_n a vector of ones with the dimension (n x 1) (Holt & Borsuk, 2020, pp. 5–6; LeSage & Pace, 2009, pp. 34–39).

A comparison of the SLM with the standard linear regression model seems reasonable since the SLM is considerably more complicated and, remembering Occam's razor, should only be used when necessary, i.e., when the standard OLS estimation leads to clearly wrong results or when

the more complicated model leads to distinctly better results. Such a comparison is common and, for example, also conducted by Locke & Baine (2015, pp. 398–403), who compare an OLS model against an SLM and an SEM. Similarly to the interpretation of the parameter estimates, the interpretation of the goodness of fit becomes more complicated, as the standard measure of fit, the R^2 becomes invalid and too optimistic due to the inclusion of the dependent variable as an explanatory variable (Grekousis, 2020, p. 464). Commonly used alternatives include the pseudo R^2 and the Akaike information criterion (AIC), of which the AIC is the more useful alternative according to Grekousis (2020, p. 467). However, since the R^2 and the pseudo R^2 are widespread and scaled to a range from 0 to 1, which allows an interpretation of the absolute value and, thereby, a comparison of the goodness of fit with other, only loosely related models, they are included. An increasing value of the pseudo R^2 and a decreasing value of the AIC indicate a better fit (Anselin, 2005, p. 175). Although some of the measures are not widely used, they have all already been applied in studies of either vacancy research or real estate market research in general, e.g., the MAE in the study by De Nadai & Lepri (2018, p. 327) examining neighborhood influence on real estate prices, or the pseudo R^2, also in price related research by Glumac et al. (2018, p. 16) or Kay et al. (2014, pp. 135–136), and the AIC in the work by Morckel (2014, p. 10) for selecting the best model in their attempt to explain housing abandonment causing factors.

4.3 Expert Interviews

The results obtained by the quantitative analysis are based on an extensive literature review and can be compared with previously found relationships in the literature. Furthermore, it is possible to evaluate whether the results are coherent by comparing the different quantitative approaches and their interpretations. These approaches make it possible to validate the results, however, they do not allow to evaluate them in terms of expectability and comprehensibility of the content.

Thus, expert interviews are applied to provide these additional information, which is a common follow-up procedure after using quantitative methods (DeJonckheere & Vaughn, 2019, p. 2). Beyond that, the expert interviews offer more valuable insights. For example, they allow the relevance of the topic to be evaluated, the variables included and their effect to be checked, and the quality and transferability to be assessed. To summarize, expert interviews are an additional approach to triangulate the results by adding another method and multiple expert perspectives,

which is regarded as a powerful strategic approach for improving internal validity and credibility (Merriam & Tisdell, 2015, pp. 244–245).

The semi-structured interview is the most widespread among the different types of interviews (DiCicco-Bloom & Crabtree, 2006, p. 315). Fostered by the various advantages of semi-structured interviews, which include the possibility to further explain questions and thereby prevent misunderstanding and to ask follow-up questions, this technique is also applied in this study. It especially allows tackling problems in understanding the questions caused by the complexity of the topic and to follow up on answers that provide new perspectives. Thus, the semi-structured interview approach was applied. The interview guide development was based on the five phases approach described by Kallio et al. (2016), with the phases:

1. identifying prerequisites
2. including previous knowledge
3. designing interview guide
4. pilot testing of the interview guide
5. presenting the final guide.

The objective of the first phase is to identify the suitability of semi-structured interviews for the planned purpose (Kallio et al., 2016, p. 2959). Since the semi-structured interview is used in this study as a supplementary approach to the other applied quantitative approaches, the previously described advantages justify the application as an additional method for triangulation. However, this does not allow the conclusion to be drawn that it would be the best approach in isolation. The second phase aims to incorporate the previous knowledge inherent in this work as the results from different methodological approaches, including a comprehensive literature review, are validated by the interviews.

According to the review by Kallio et al. (2016, pp. 2959–2960), the design of the interview guide should be guided by the goal of generating questions that serve as a tool for collecting interview data and, at the same time, these questions should exhibit various characteristics, including being neutral and clearly worded (DeJonckheere & Vaughn, 2019, p. 5), which was considered for the formulation of the questions. The last step to derive the final interview guide is the pilot testing of the draft, mainly aiming to identify potential needs for revision (Kallio et al., 2016, p. 2960). For this purpose, the interview guide was discussed with the supervisor of this thesis and pilot tested in an interview with an expert for real estate markets holding a doctoral degree, which led to the final interview guide included in Appendix B - 12 (German

original) and Appendix B - 13 (English translation), structured into multiple categories. The structure of the interview guide exhibits a similar order as the work itself, starting with the relevance of the topic, moving on with the evaluation of the derived results, and finally checking their quality and transferability. Thereby, the questions build up on each other, making it easier for the experts to follow the topic without providing detailed introductions.

The interview is introduced by requesting the experts to describe themselves to clarify the perspective of each expert and allow them to find their way into the interview, followed by their perception of the relevance of the topics of vacancy data and ORL data, to assess and validate the importance of the research. The second category of questions asks about possible factors explaining vacancies in order to validate and potentially enhance the categories derived through the literature review in Chapter 2.4. Subsequently, the variables used in the quantitative analysis and their relevance are validated by requiring the experts to rank their relevance and importance in terms of explaining vacancy, which presents valuable insights regarding the added value and predictability of the results from the quantitative analysis in comparison to the experts' opinions. Finally, the quality and transferability are evaluated. Therefore, questions are asked to evaluate the achieved accuracy of the estimation by asking for the accuracy expected and needed by the experts. Furthermore, the presumed temporal and spatial transferability is queried.

Besides the questions that are asked, the choice of the interviewed experts is essential to achieve reliable results. Appropriate experts should have knowledge about the topic and be able and willing to express it (Whiting, 2008, p. 36). Thus, purposeful sampling of these experts is necessary, even though it is typically not the goal to derive a statistically representative sample, which would require extensive groups of experts (DeJonckheere & Vaughn, 2019, pp. 3–4). Concerning this study and the objective to validate the results, purposeful sampling means in particular that the experts should be professionally qualified in the topic and the group of experts should be diverse in terms of their perspectives on the topic. This diversity ensures that no information relevant to explaining the derived results are missed. As it is common practice to protect the participants from unwanted consequences from the interviews and to encourage them to speak freely, the interviews are anonymized (Creswell & Poth, 2018, p. 250).

As a measure of the experience of the experts, the length of time they have been working in their profession is used. The professional experience of the seven experts interviewed ranges from 8 to 30 years, with an average experience of 19.3 years and a median experience of 20 years. More than half of the experts are members of at least one expert committee for real estate valuation. Combined with their profession, focusing on real estate markets, this experience

implies their familiarity with the research topic, particularly their understanding of real estate market interrelationships, and their aptitude for the interviews.

The experts can be divided into three groups with different real estate market perspectives. Two experts (S1, S2) work as academics focusing on real estate markets, one with a doctoral degree and the other with a doctoral degree and a professorship. Another two experts work as private sector real estate appraisers (P1, P2), one with an international focus and the other focusing on Germany, both in leading positions. The other experts work for publicly funded institutions (I1, I2, I3), one for a research institution, one in a leading position in an expert committee at the state level, and one in a state ministry focused on real estate valuation.

The evaluation of the interviews follows the standard procedure of data preparation and organization, data reduction into themes, and presentation of the results (Creswell & Poth, 2018, p. 251). Since the expert interviews are conducted as an additional method for triangulation purposes, their focus is validation and supplementation rather than exploring a new field, which influences the applied methodology of evaluation considerably. The four question categories predetermine the structure of the evaluation and organization of the data. The answers to the questions and their additional remarks within these categories are confronted with each other and subsequently condensed by summarizing similar statements. The results of this process are presented as a textual summary, which is the basis for the final discussion in Chapter 6 that compares, combines, and confronts the results from the different methodological approaches.

5 Results

The previous chapter, explaining the methodology to answer the vacancy rate-related research questions, outlined how the results are derived. These results are presented in this chapter consistent with the structure of the methodology and are likewise divided into the conducted bivariate analysis and regression analyses that subsequently allow to answer the second and third research questions.

5.1 Bivariate Analysis

As induced by the results from the univariate analysis, in Chapter 3.2.4, the variables *Vacancy Rate*, *GDP per Capita*, *Normalized Number of Listings,* and *Cold Rent per SQM* were transformed by the natural logarithm and descriptive statistics are provided for the transformed as well as for the untransformed forms of the variables. Based on these results, it must be decided in which form, transformed or untransformed, the variables are to be included and which are to be integrated generally. The bivariate analysis allows a more in-depth examination of the suitability of these different variables for regression analysis, as it allows conclusions to be drawn about the nature and extent of the relationship between two variables, as opposed to the previous univariate analysis, which only provided information about the potential explanatory variables in isolation. For examining the explanatory power of ORL data with regard to vacancy rates, in particular, the relationships between potential explanatory variables and vacancy rates or a transformed form of vacancy rates are relevant. Thus, the selected bivariate analysis methods are applied to all possible variable combinations of the previously chosen potential explanatory variables, the *Vacancy Rate* variable, and the transformed *ln(Vacancy Rate)* variable[28]. The visualizations and the correlation coefficients are presented in a joint graph to allow a combined, comprehensive evaluation. The different variable combinations are examined in the order listed in Table 18.

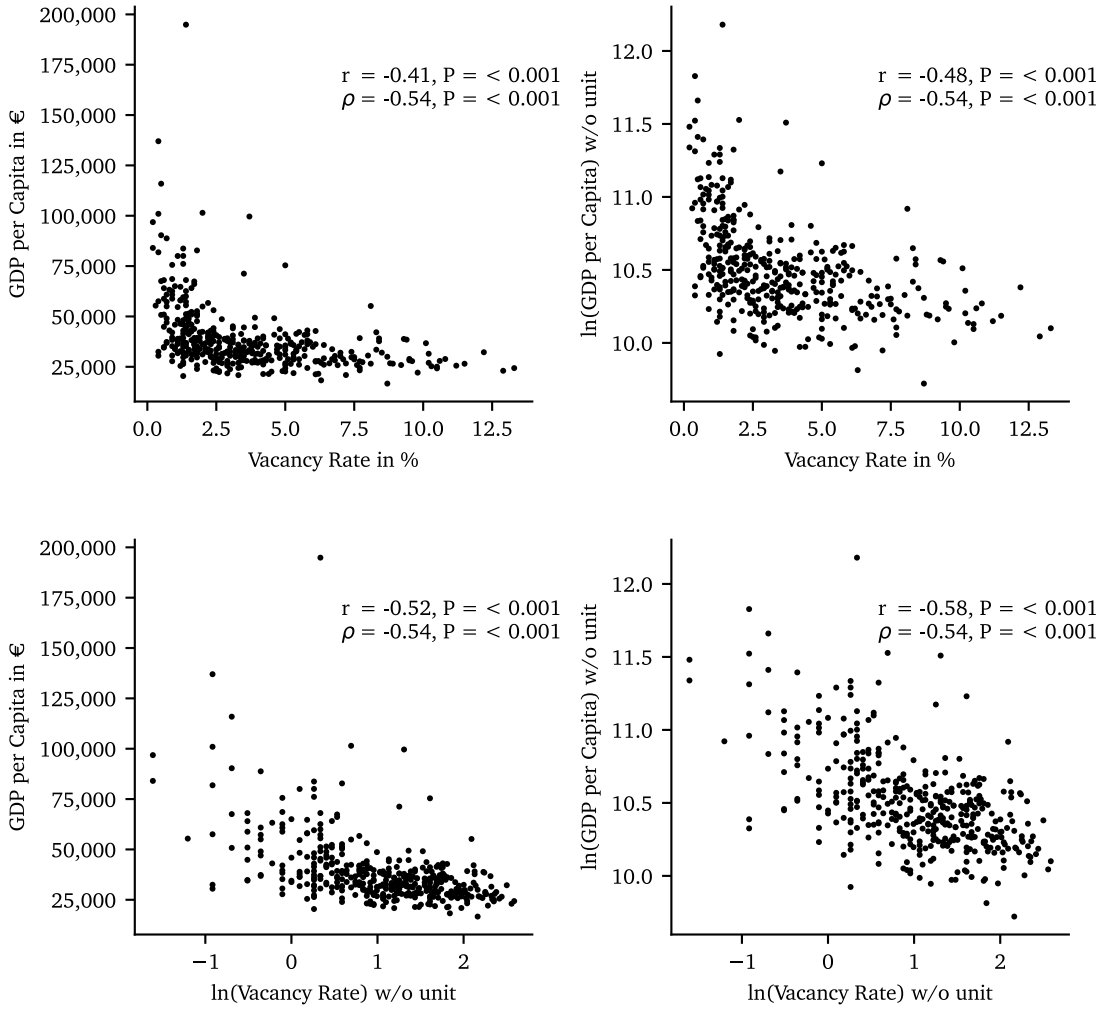
The *GDP per Capita* variable is included as a variable from the (socio-)economic variables group and showed a considerable deviation from a normal distribution, implying a logarithmic transformation. Therefore, four possible variable combinations result from the transformed and untransformed combinations of both variables, which can be seen in Figure 28. All combinations containing one variable or both variables in untransformed form show a noticeable deviation

[28] For an overview of all variables, see Table 18.

from a linear relationship to a curved monotonically decreasing curve. The combination of both untransformed variables, shown in the upper left quadrant, shows the most obvious deviation, while the combination of both variables transformed by the natural logarithm appears to be mostly linear. The decreasing shape of the relationship is also expressed in the correlation coefficients, both expressing that an increase in one variable is connected with a decrease in the other variable. Since the transformation by the natural logarithm does not change the rank order of the observations, Spearman's ρ is invariant to it and has the same value and P-value for all possible combinations.

Figure 28

Relationship of GDP per Capita and Vacancy Rate



Note. Own research. Created based on analysis data set.

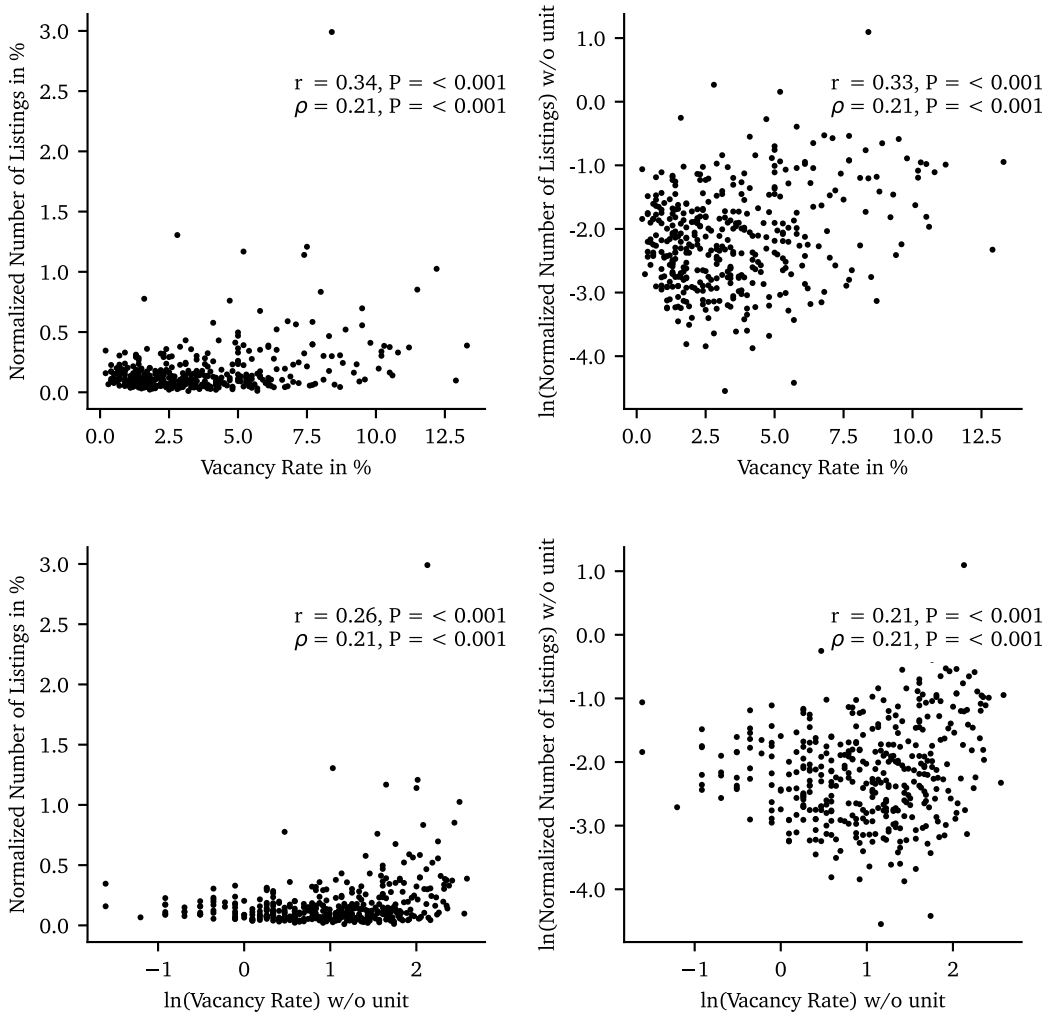
Pearson's r can be calculated for all possible combinations, however, the combinations containing an untransformed variable seem to be rather curved than linear, which implies a violation of the requirement of the joint normal distribution (Schober et al., 2018, p. 1764). Thus, the P-value is reported for consistency only for all combinations and should not be interpreted as a reliable measure of significance, at least not in magnitude, for the non-linear combinations. Nevertheless, the linear relationship expressed by Pearson's r is strongest for the combination of both transformed variables, which is in line with the visual inspection. For this combination, the P-values for both correlation coefficients are orders of magnitude smaller than the significance level of $\alpha = 5$ % consistently used in this paper[29]. As a logical conclusion, considering the combination of the variables *Vacancy Rate* and *GDP per Capita* in isolation, a combination of both transformed variables should be included in the regression analysis.

The *Normalized Number of Listings* cannot be assigned to one of the categories of variables commonly used in vacancy research but is included because of the presumed contextual relationship between the number of listings and the vacancy rate. It showed an even greater deviation from the normal distribution regarding skewness and kurtosis than the previously described *GDP per Capita*. However, the transformation resulted in a variable that is most widely normally distributed. Thus, the relationship between the *Vacancy Rate* and the *Normalized Number of Listings* is examined for both variants, leading to four possible combinations depicted in Figure 29. Contrary to the presumed contextual relationship, none of the combinations presented shows a strong linear relationship. Both correlation coefficients imply an increasing relationship for all combinations. However, especially for the variants containing the untransformed *Normalized Number of Listings*, this rather seems to be due to an increasing variance than to the explanatory power of the relationship. The combinations containing the transformed *Normalized Number of Listings* have a weak positive trend that seems slightly more pronounced for the correlation with the untransformed *Vacancy Rate*. However, no strong explanatory power is apparent overall, also reflected in the correlation coefficients. Due to the shape of the scatter plots, only Spearman's ρ should be interpreted, which is significant at the 5 % confidence level, but shows only a weak correlation of 0.21 compared to other variables. Concluding, the *Normalized Number of Listings* should be included in the transformed form and ideally with an untransformed *Vacancy Rate*.

[29] Due to the large number of observations, the P-Values can become very small implying highly significant relationships. However, for consistency reasons it is common practice to stick to the previously chosen significance level, why the comparison is always made with the values for the 5 % significance level.

Figure 29

Relationship of Normalized Number of Listings and Vacancy Rate



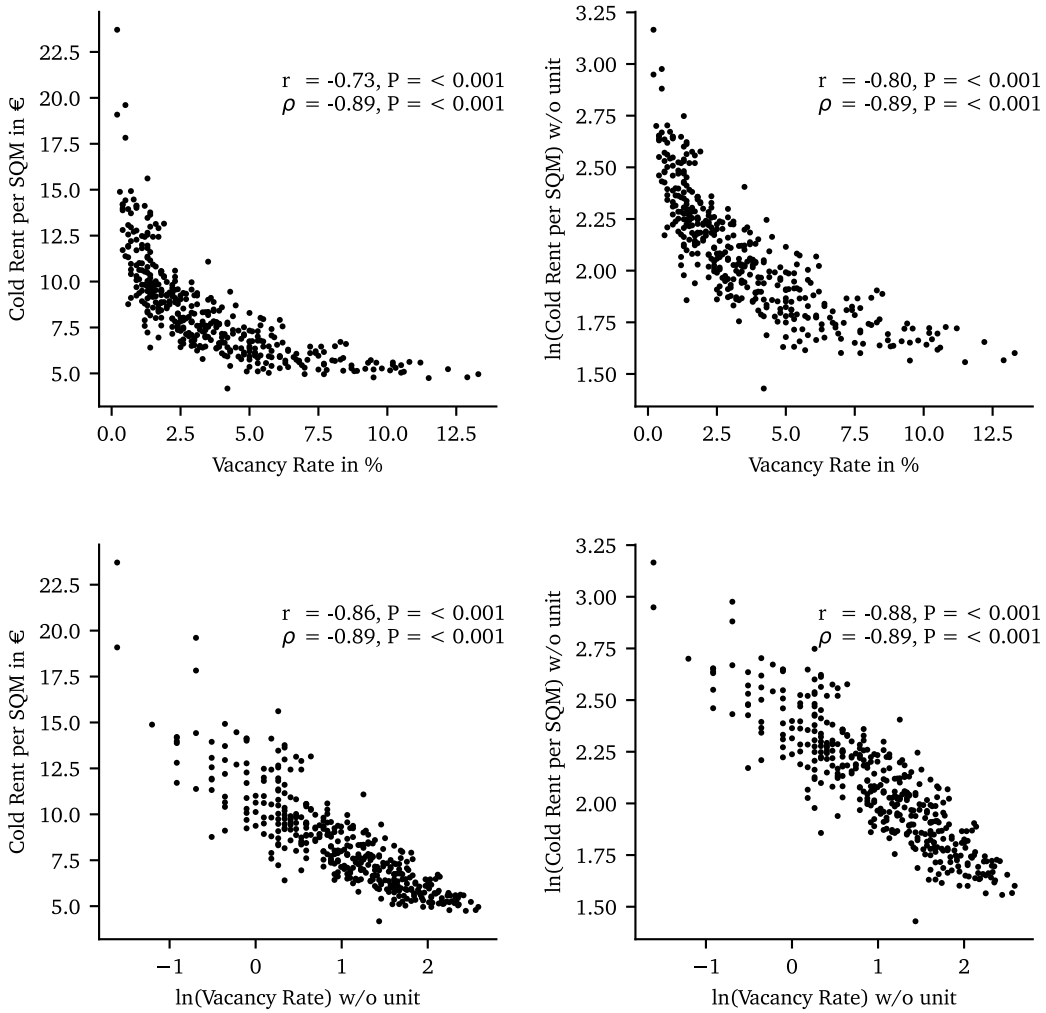
Note. Own research. Created based on analysis data set.

However, due to the weakness of the relationship, other variables are considered more important and should therefore be treated with priority when it comes to decisions regarding the design of the regression model. Finally, due to the unexpectedly weak relationship of the *ln(Normalized Number of Listings)* variable with the *ln(Vacancy Rate)* variable, further examinations of this relationship were conducted and indicated that the relationship is not stable over the different settlement types. Thus, this relationship was further analyzed, subdivided into the different settlement types. This analysis can be found in Appendix C - 1 and shows that taking the settlement type into account in this relationship considerably increases the correlation.

The *Cold Rent per SQM* variable complements the *GDP per Capita* as the second variable from the (socio-)economic variables group and showed slight deviations from the normal distribution. Therefore, a transformation by the natural logarithm was considered and resulted in a distribution that is very close to normal. Thus, four possible combinations are evaluated and shown in Figure 30.

Figure 30

Relationship of Cold Rent per SQM and Vacancy Rate



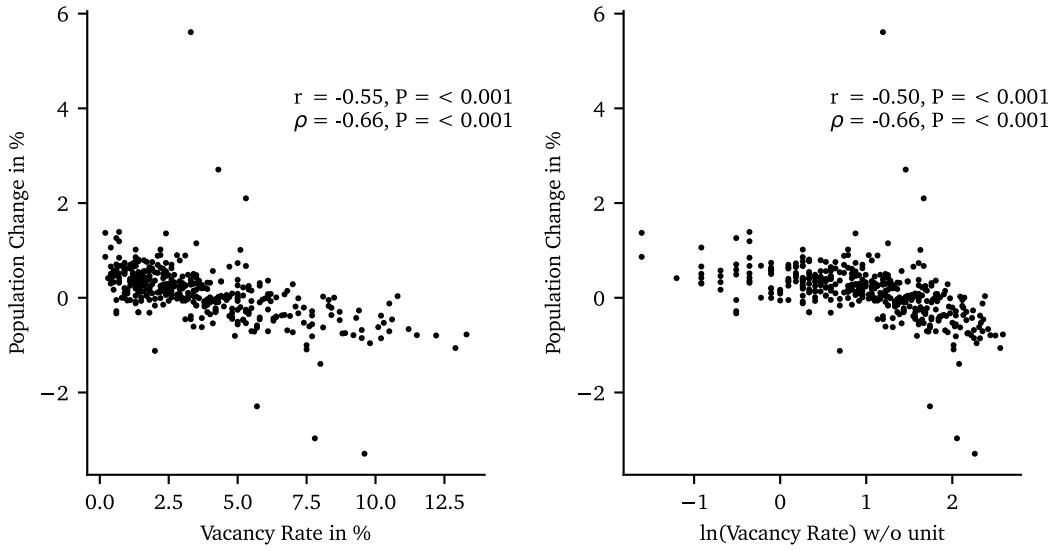
Note. Own research. Created based on analysis data set.

The graphs containing the *Vacancy Rate* in an untransformed variant show a rather curved shape, which is clearly more linear for the combinations integrating the transformed *ln(Vacancy Rate)*. Of these combinations, the combination of both transformed variables shows an even closer fit to a linear shape, as it is less curved, especially for the upper end of the values of the *ln(Vacancy Rate)*. These findings are also reflected in the correlation coefficients that, without exception, show absolute values larger than 0.7. Since the correlation is negative, expressed by the negative values of the correlation coefficients, an increase in one variable leads to a decrease in the other. This relationship is as expected, since higher vacancy rates are associated with lower attractivity and with lower demand which leads to lower rents. This distinctive relationship is also confirmed by the P-values for Spearman's ρ that are significant at the level of 5 %. Since the relationship is very close to linear, particularly for the combinations containing the *ln(Vacancy Rate)* variable, Pearson's r can also be interpreted as highly significant. Based on this significance, small differences in the values for Pearson's *r* are relevant, why the choice of the preferred variable combination for the regression analysis is based on the highest *r*, which is reached by the combination of both transformed variables.

The *Population Change* variable represents a variable from the sociodemographic variables group regularly included in vacancy research and is depicted in Figure 31.

Figure 31

Relationship of Population Change and Vacancy Rate



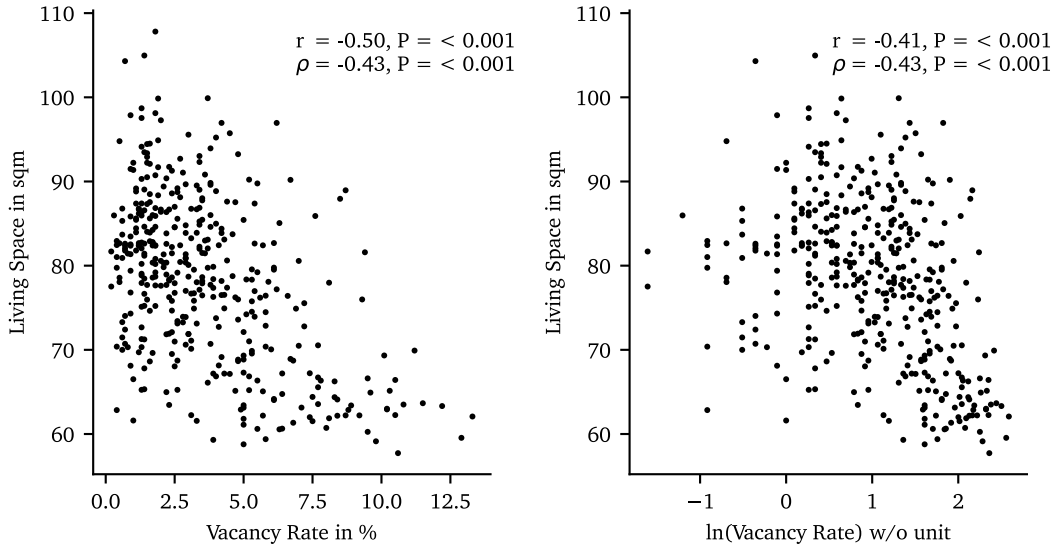
Note. Own research. Created based on analysis data set.

It is chosen explicitly since a strong negative correlation is assumed due to the characteristic of the building stock to adjust slower than the population. The univariate distribution did not imply a relevant deviation from a normal distribution. Hence, no transformation was applied, leading to the comparison of two combinations. The relationship of both combinations appears to be approximately linear and decreasing, which is consistent with the previously assumed correlation. These results are confirmed by the correlation coefficients, which are both significant at the 5 % level but only imply a moderate correlation between -0.5 and -0.66. Due to the linear relationship of both variables and the significance of the correlation coefficients, Pearson's *r* can be used as a basis for decision-making regarding the recommendation for the regression analysis. Therefore, from the perspective of solely integrating the Population Change variable, a regression using the untransformed vacancy rate would be optimal.

The last selected continuous variable, the *Living Space* variable, falls into the group of variables that describe property characteristics, and it is shown in Figure 32.

Figure 32

Relationship of Living Space and Vacancy Rate



Note. Own research. Created based on analysis data set.

The relationship between vacancy and property characteristics seems more evident at the individual building level than at the aggregate level used in this study, as individual variations

in characteristics between properties could lead directly to vacancy, such as a dilapidated building condition compared to a well-maintained building. However, it is also possible that average characteristics for a given area approximate other relevant factors of that area, such as attractiveness or affordability, which could lead to explanatory power of these average characteristics and has been used, for example, in the research of Yue et al. (2022).

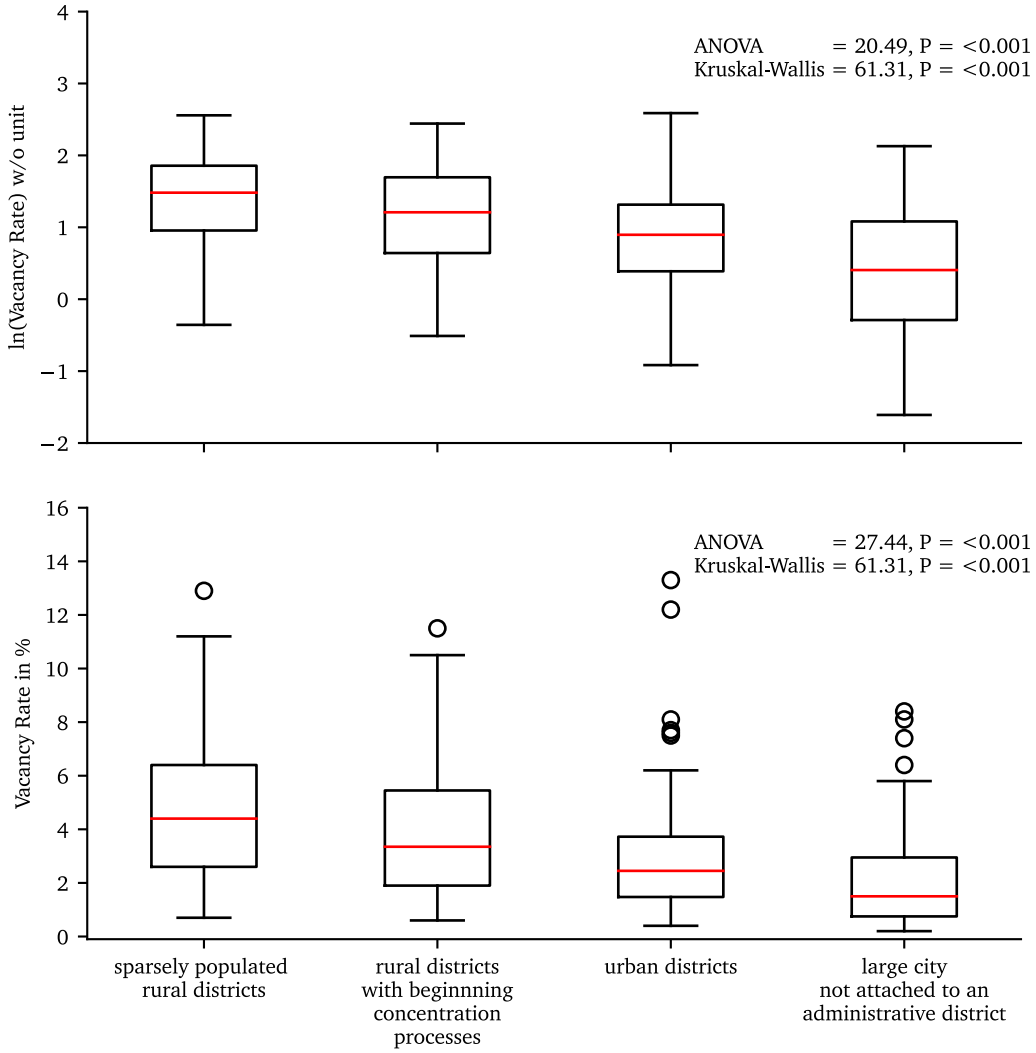
The univariate distribution of the *Living Space* variable did not imply a relevant deviation from a normal distribution. Hence, no transformation was applied, leading to the comparison of two combinations. Both combinations do not show a clear linear trend. Thus, the scatter plots imply not relying on Pearsons's r. However, Spearman's ρ of -0.43, which is significant at the 5 % level, fits the observation from the scatter plots of a negative relationship between the transformed or untransformed *Vacancy Rate* and the *Living Space*. This relationship expresses that the average offered living space gets smaller with an increasing vacancy rate. The reasons for this relationship could be that rents are lower in districts with higher vacancy rates and residents, therefore, tend to demand the segment of larger apartments or that the entire building stock in districts with typically high vacancy rates is skewed toward smaller apartments. Due to the ambiguous results of the analysis of the scatter plots and the resulting unreliable Pearson's correlation coefficients, no clear recommendation regarding the regression analysis can be derived from the Living Space variable, and the decision should instead be based on the other variables.

As with much real estate-related research, the location could also be important in explaining the vacancy rate. Therefore, a variable accounting for location is included, supplemented by other factors accounting for location in the spatial regression analysis. Since the analysis is conducted at the district level, locational characteristics that represent the entire district should be considered. At the district level, settlement type classifications are provided by the BBSR that could be particularly suitable since urbanization and rural exodus processes are strongly associated with vacancy. The *Settlement Type* variable is categorical and thus examined by other methods of analysis that are also explained at length in Chapter 4.1. For visual inspection, the box plots shown in Figure 33 help differentiate the vacancy distributions of the four settlement types. Both combinations clearly show a decreasing trend for the transformed and untransformed *Vacancy Rate* with higher settlement density. The logarithmic transformation has the typical damping effect for the values at the upper end, leading to the fact that the outliers of the untransformed distribution still fall into the whiskers in the transformed distribution. The box plots suggest that the different settlement types are not subject to the same distribution, as the representations differ significantly for both combinations. This

impression gets confirmed by the value of the ANOVA and the Kruskal-Wallis test, which are both significant at the 5 % level and thereby indicate different underlying distributions for the different settlement types. Therefore, pairwise tests should be conducted to determine if all settlement types are different from each other in each case or if it is only a subset of types that do not have an identical underlying distribution.

Figure 33

Relationship of Settlement Type and Vacancy Rate



Note. Own research. Created based on analysis data set.

For the application of the pairwise tests, the Bonferroni correction must be applied, which reduces the critical value of the 5 % significance level to 0.05 divided by the number of 6 tests that must be performed to test all possible combinations. Thus, the critical value to test against is approximately 0.0083. Table 19 below shows the test statistics and P-values for all possible combinations, confirming the impression of the visual inspection that the more rural and the more urban districts are more similar to each other than the other combinations for the untransformed *Vacancy Rate*.

Table 19

Pairwise ANOVA and Kruskal-Wallis Tests

Type of vacancy data	Settlement type combination	ANOVA F-test statistic	ANOVA P-value	Kruskal-Wallis test statistic	Kruskal-Wallis P-value
untransformed	sparse/rural	3.1699	0.077	3.8649	0.049
	sparse/urban	35.7241	<0.001	33.2885	<0.001
	sparse/city	43.5371	<0.001	41.9944	<0.001
	rural/urban	14.4348	<0.001	11.3955	<0.001
	rural/city	23.4564	<0.001	26.1101	<0.001
	urban/city	4.8990	0.028	10.0231	0.002
transformed	sparse/rural	4.1160	0.044	3.8649	0.049
	sparse/urban	35.8130	<0.001	33.2885	<0.001
	sparse/city	62.1239	<0.001	41.9944	<0.001
	rural/urban	13.2897	<0.001	11.3955	<0.001
	rural/city	36.5998	<0.001	26.1101	<0.001
	urban/city	13.8570	<0.001	10.0231	0.002

Note. Own research. Critical value for $\alpha = 5$ % according to Bonferroni transformation: 0.008. Settlement type abbreviations: sparse - Sparsely Populated Rural District; rural - Rural District With Beginning Concentration Processes; urban - Urban District; city - Large City Not Attached to an Administrative District.

For the transformed *Vacancy Rate,* the combination of the sparsely populated rural districts and the rural districts with beginning concentration processes stands out. However, no

recommendation can be derived from these statistics regarding what type of *Vacancy Rate* variable should be included. Since the settlement types S*parsely Populated Rural District* and *Rural District With Beginning Concentration Processes* differ considerably from each other in reality, it is not apparent that the correlations of the various settlement types and vacancy differ from each other but are the same for these two types, and since presumed correlations should always be given preference over purely quantitatively constructed correlations, all settlement types are initially considered as independent types in the regression analysis. However, if there is also no difference in these types in the regression analysis, they should be combined into one settlement type for vacancy estimation.

The results of the bivariate analysis of all continuous variables are heterogeneous regarding the use of the transformed or the non-transformed vacancy rate, and the different variables should be included in part transformed and in part non-transformed. Since the vacancy rate as the explained variable can only be included in one variant, transformed or untransformed, and the choice does not come to the same result for all variables, but both variants are preferred by two variables each, the variable with the strongest correlation, *Cold Rent per SQM*, is used as the basis for the decision to use the *ln(Vacancy Rate)* variable. This choice is supported by the results of the univariate analysis and the approach commonly used in comparable research of including similar variables transformed by the natural logarithm. Table 20 summarizes the correlations of the variables included in the regression analysis and the chosen *ln(Vacancy Rate)*.

Table 20

Correlations of Explanatory Variables and ln(Vacancy Rate)

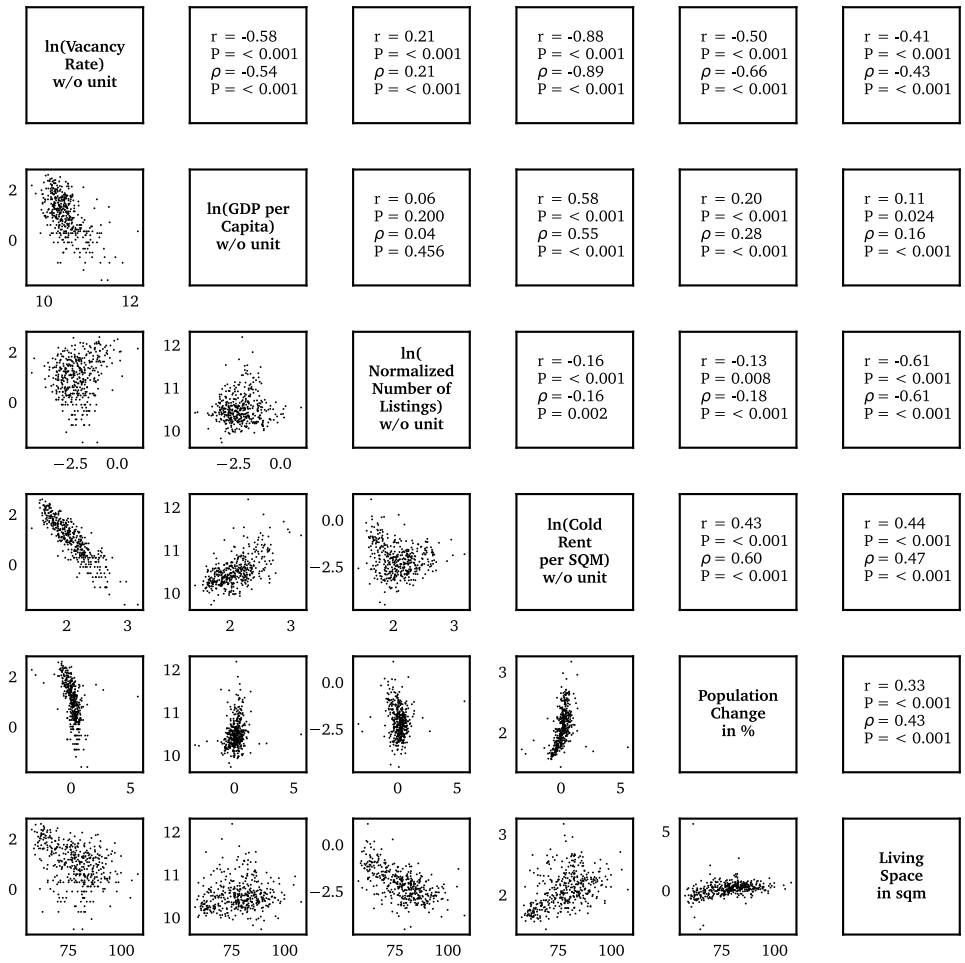
Variable	Pearson's r	P-value	Spearman's ρ	P-value
ln(GDP per Capita)	-0.58	<0.001	-0.54	<0.001
ln(Normalized Number of Listings)	0.21	-	0.21	<0.001
ln(Cold Rent per SQM)	0.88	<0.001	-0.89	<0.001
Population Change	-0.50	<0.001	-0.66	<0.001
Living Space	-0.41	-	-0.43	<0.001

Note. Own research. P-values are only reported when the requirements for examining significance are met.

However, in terms of regression analysis, it is crucial to evaluate not only the relationships between the explanatory variables and the dependent variable but also the relationships among the explanatory variables themselves. Strong correlations between the explanatory variables themselves imply possible problems with multicollinearity, which is undesirable in regression analysis as the consequences are unreliable regression coefficients with a large variance (Draper & Smith, 1998, p. 369). For the joint examination of all variables included in the regression analysis, the scatter plot matrix, sometimes also called draftsman's plot, is well suited and displayed in Figure 34, as it allows gaining an impression of the overall data set by the combined analysis of scatter plots and correlation coefficients (Chatterjee & Hadi, 2015, pp. 101–102).

Figure 34

Scatter Plot Matrix for all Continuous Variables



Note. Own research. Created based on analysis data set.

Besides the intended correlations of the dependent and the explanatory variables, particularly the variable combinations *ln(GDP per Capita)/ln(Cold Rent per SQM)*, *ln(Normalized Number of Listings)/Living Space*, and *ln(Cold Rent per SQM)/Population Change* seem suspicious with a Spearman's ρ larger than 0.5. Thus, in the regression analysis, particular focus is placed on the analysis of multicollinearity.

Summarizing, from the results of the bivariate analysis, it can be concluded that all variables included in the regression analysis show at least a weak correlation with the *ln(Vacancy Rate)* variable, most of them being even moderate or strong. However, caution must be paid to the relationships between the explanatory variables themselves, as they imply possible multicollinearity.

5.2 Regression Analysis

The bivariate analysis results suggest which variables to include and in what form and already indicate where potential problems might arise. Based thereon, a base regression model is formulated in Chapter 4.2.1. The results of the estimation of this model are shown in this chapter and a final multiple linear regression model is derived that considers the common assumptions required for a reliable model. This model builds the foundation for a spatial model that specifically attempts to include spatial effects not integrated by the variables integrated into the standard multiple linear regression model.

5.2.1 Multiple Linear Regression

Table 21 shows the results of the base regression model that is further evaluated by the different tests and techniques of visualization chosen in Chapter 4.2.1 to ensure conformity with the standard linear regression assumptions. A detailed interpretation, particularly of the regression coefficients, is given for the final model. The interpretation of the base model is restricted to what is required to finalize the multiple linear regression model. More detailed explanations, e.g., of the coefficients, are given for the final model. Notably, the *Living Space* variable is insignificant and none of the spatial dummy variables are significant, but three out of four interaction terms containing the spatial dummy variables are. The F-statistic indicates a highly significant model in total, which is in line with the high value of 0.814 for the adjusted R^2, stating that the model can explain 81.4 % of the variance in the dependent variable.

Table 21

Results Multiple Linear Regression Base Model

Variable	Coefficient	Standard error	t-statistic	P-value	[0.025	0.975]	VIF
Constant	7.404	0.675	10.963	<0.001	6.076	8.731	-
ln(gdp)	-0.210	0.069	-3.057	0.002	-0.344	-0.075	190.139
ln(cr)	-1.992	0.095	20.973	<0.001	-2.179	-1.806	123.085
pc	-0.199	0.032	-6.143	<0.001	-0.263	-0.135	1.367
ls	0.003	0.003	0.966	0.334	-0.003	0.008	149.938
Spatial dummies							
sparse	0.112	0.146	0.768	0.443	-0.174	0.398	17.491
rural	0.020	0.135	0.151	0.880	-0.245	0.286	14.880
urban	0.032	0.149	0.212	0.832	-0.261	0.325	24.438
Interaction terms							
ln(nnl) x sparse	0.134	0.047	2.863	0.004	0.042	0.226	11.555
ln(nnl) x rural	0.125	0.044	2.825	0.005	0.038	0.211	8.951
ln(nnl) x urban	0.101	0.051	1.962	0.051	-0.000	0.201	16.152
ln(nnl) x city	0.216	0.064	3.367	0.001	0.090	0.343	5.950

Model Summary

Dependent variable:	ln(vac)	F-statistic (P-value):	159.776 (<0.001)
Number of observations:	401	Condition number:	3,162
Degrees of freedom:	389	R^2:	0.819
BP LM test (P-value):	9.147 (0.608)	Adjusted R^2:	0.814
BP F-test (P-value):	0.825 (0.615)	MAE (Vacancy Rate):	0.858
White's LM test (P-value):	33.981 (0.980)	AIC:	293
White's F-test (P-value):	0.606 (0.986)	Moran's I (P-value):	0.125 (0.001)

Note. Own research.

The MAE indicates an average deviation from the actual values smaller than one percentage point by the model estimates. However, the values for the VIFs and the condition number

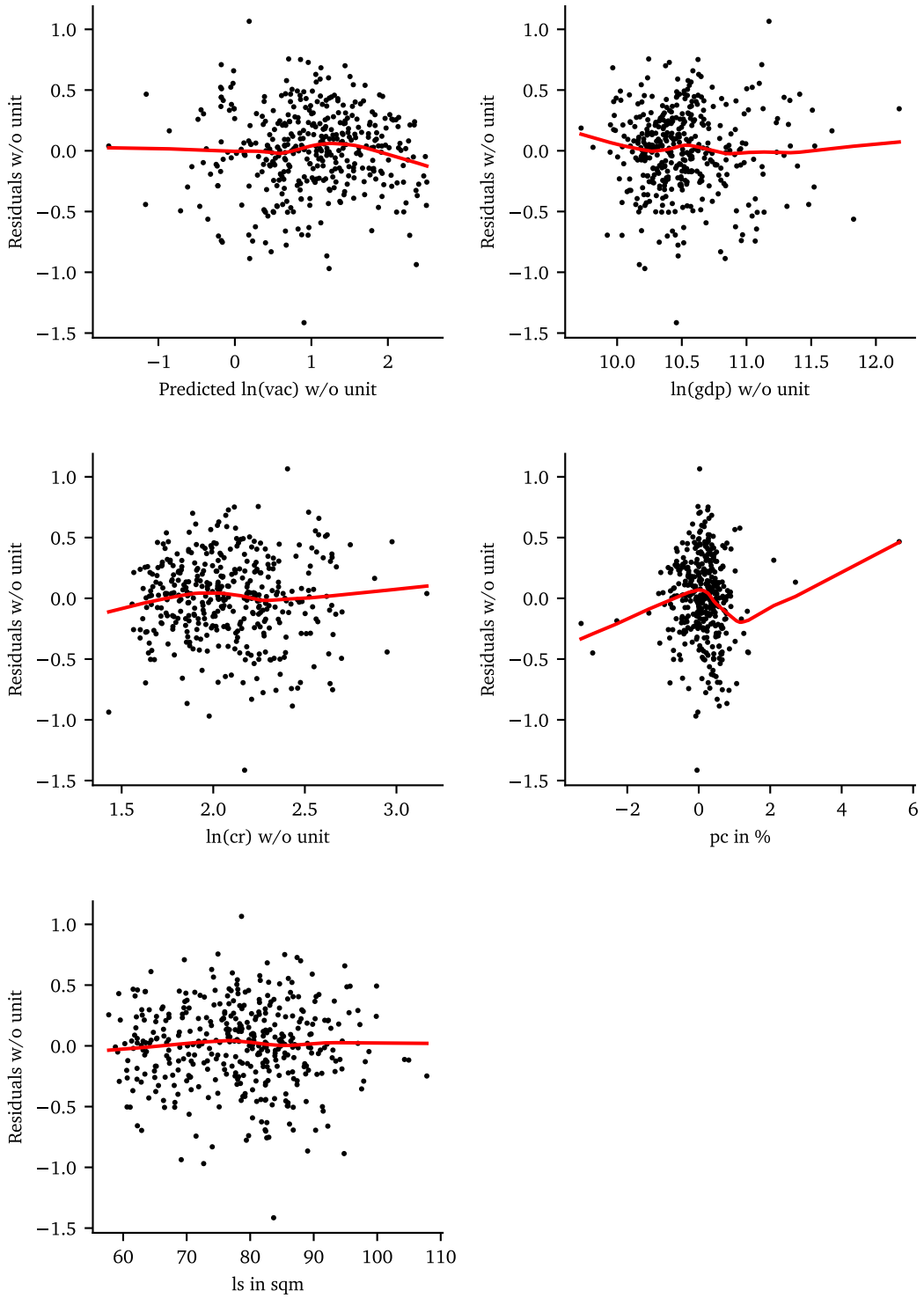
indicate multicollinearity problems, wherefore a detailed analysis of all assumptions underlying linear regression models is conducted. Residual plots of all continuous variables are given in Figure 35 to examine the requirement of a linear relationship between the dependent and the explanatory variables. The plot of the estimated dependent variable on the residuals, considered particularly important, is inconspicuous, as are most of the independent variables. However, the population change variable subplot shows conspicuous outliers likely to substantially bias the estimate concerning this variable. The elimination of these outliers should therefore be considered.

The variables with particular leverage due to their values for the population change and residual are the districts of Wartburgkreis (-3.29 % / -0.21), Landkreis Saalfeld-Rudolstadt (-2.97 % / -0.45), Ilm-Kreis (-2.29 % / -0.19), the city of Dessau-Roßlau (-1.40 % / -0.12), Landkreis Schmalkalden-Meiningen (2.10 % / 0.31), Landkreis Sonneberg (2.71 % / 0.13), and the city of Suhl (5.61 % / 0.47). Similar subplots are provided for the interaction term variables in Appendix C - 2. From these subplots, especially the subplot with the interaction terms ln(*nnl*) *x city,* shows suspicious values for the dimensionless interaction term and residual for the districts of Freiburg im Breisgau (-2.71 / -0.49), the city of Münster (-2.44 / -0.74), and the city of Chemnitz (1.10 / -0.27). For the final model, all of these potentially biasing observations are excluded. The consequences of this exclusion are taken up for the final interpretation and discussion of the results.

Besides their explanatory power in terms of model misspecification, the residual scatter plots can also be used to assess potential problems with heteroscedasticity. None of these plots shows a considerably changing variance of the residuals, which is confirmed by numerical tests. The BP- and White's test statistics on heteroscedasticity indicate that no heteroscedasticity is present, as the null hypothesis of homoscedasticity can not be rejected at the 5 % level since the P-values of all test statistics are consistently much higher.

Figure 35

Residual Plots of Continuous Variables Base Model

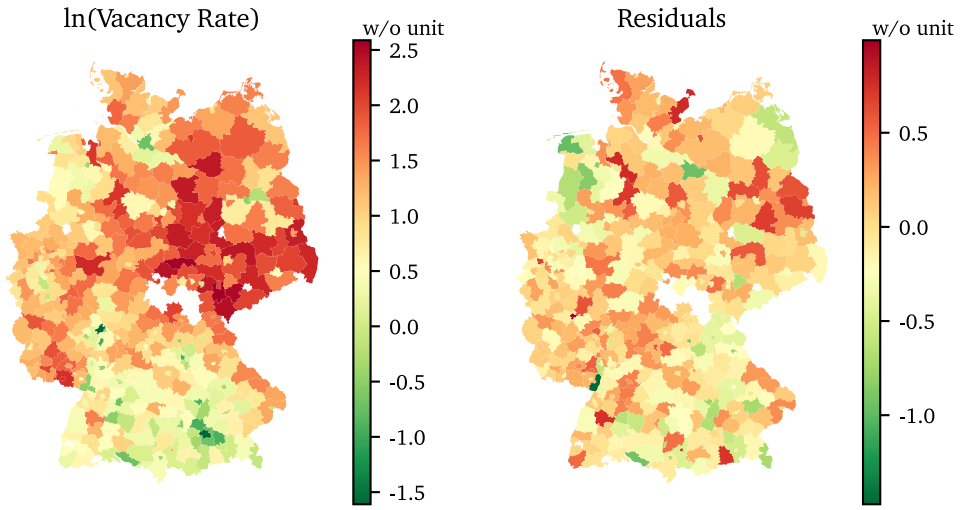


Note. Own research.

As a measure of spatial autocorrelation, the calculation of Moran's I of the residuals indicates whether explanatory power remains in the residuals hinting at the problem of spatial autocorrelation. As derived in Chapter 4.2.1, Moran's I is calculated using a row-standardized inverse distance weights matrix, which results in a value of 0.12 and a P-value of 0.001, indicating significance of weak positive spatial autocorrelation at the 5 % level. For comparison, the value for Moran's I for the explained variable *ln(Vacancy Rate)* is 0.51, also with a P-value of 0.001. Figure 36 visualizes both spatial distributions and clearly shows that the remaining spatial autocorrelation within the residuals is distinctly smaller than the spatial autocorrelation within the explained variable. However, some spatial dependencies remain, which can be addressed by a spatial regression approach.

Figure 36

Spatial Distribution of ln(Vacancy Rate) and Residuals



Note. Own research.

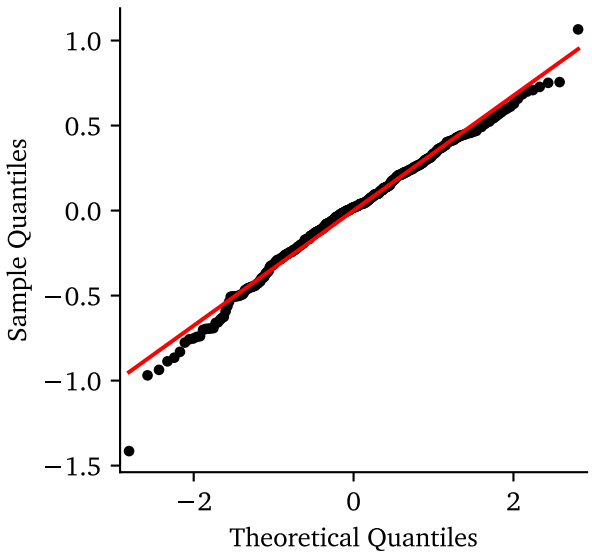
The VIFs, severely exceeding the value of 10 in multiple cases and the condition number exceeding the value of 1,000 clearly, hint at severe problems with multicollinearity. These problems are expectable, as the scatter plot matrix already showed distinct correlations between the variable *ln(Cold Rent per SQM)* and the variable *ln(GDP per Capita)*, as well as between the variable *ln(Normalized Number of Listings)* and the variable *Living Space*. The relationship between the economic variables seems natural, as the explanatory power of both variables is

rooted in the same group of variables. The relationship between the *Living Space* Variable and the *ln(Normalized Number of Listings)* variable is not that clear. However, the assumed weak relationship between vacancy and an average living space implies the elimination of the *Living Space* variable. Furthermore, the inclusion of the spatial dummies in the interaction terms and the minor differences shown in the ANOVA imply the exclusion of the individual spatial dummies. Additionally, the slope of the interaction terms is relatively flat and the difference in the slopes in the interaction terms is small, so similar to the spatial dummies, these have a risk of high multicollinearity and the interaction term with the highest VIF value, *ln(nnl x urban)*, is excluded.

Less important is the assumption of the normality of the residuals since the sample size is large and regression analysis is regarded as robust to non-normal residuals due to the CLT. Nevertheless, the scatter plots of the residuals against the predictor variables do not show any suspicious behavior and the other standard graphs, the Q-Q plot and a histogram are provided in Figure 37 and Appendix C - 3 subsequently. The Q-Q plot and the histogram confirm an approximate but not perfect linear relationship, as can be seen from the slight deviations from the straight line and the slightly left-skewed data in the histogram.

Figure 37

Q-Q Plot of Residuals of Base Model



Note. Own research.

Taking the results of the inspection of the linear regression assumption into account, the base model can be adapted to the final model. It should be noted that the problem of spatial autocorrelation is only addressed by the spatial regression analysis, so the effects of a weak spatial autocorrelation are still present, as can be seen from the value of Moran's I of 0.14 at a significance level of 5 %. The results for the final model are shown in Table 22 and the base model equivalent plots can be found in Appendix C - 4 to Appendix C - 7. None of these plots shows a remaining relevant violation of the discussed assumptions. The same applies to the assumption checking tests, which indicate the significance of all variables together by the F-test and in isolation by the t-tests.

Table 22

Results Multiple Linear Regression Final Model

Variable	Coefficient	Standard error	t-statistic	P-value	[0.025	0.975]	VIF
Constant	5.404	0.177	30.468	<0.001	5.055	5.753	-
ln(cr)	-2.053	0.085	-24.249	<0.001	-2.219	-1.886	2.437
pc	-0.311	0.050	-6.168	<0.001	-0.409	-0.211	1.154
Interaction terms							
ln(nnl) x sparse	0.036	0.018	2.018	0.044	0.001	0.071	1.450
ln(nnl) x rural	0.051	0.019	2.708	0.007	0.014	0.088	1.468
ln(nnl) x city	0.151	0.033	4.520	<0.001	0.085	0.217	1.407

Model Summary

Dependent variable:	ln(vac)	F-statistic (P-value):	321.711 (<0.001)
Number of observations:	390	Condition number:	28
Degrees of freedom:	384	R^2:	0.807
BP LM test (P-value):	8.048 (0.154)	Adjusted R^2:	0.805
BP F-test (P-value):	1.618 (0.154)	MAE (Vacancy Rate):	0.859
White's LM test (P-value):	21.177 (0.219)	AIC:	284
White's F-test (P-value):	1.256 (0.218)	Moran's I (P-value):	0.139 (<0.001)

Note. Own research.

Furthermore, the VIFs are far below the critical value from 5-10 and the condition number is magnitudes smaller than the critical value of 1,000. Both BP test statistics and both White's test statistics indicate that there is no heteroscedasticity at the 5 % level. Although the number of variables was halved from 12 to 6, the adjusted R^2 remained nearly stable and, with a value of 0.805, can explain about four-fifths of the variance of the independent variable. Similarly, the MAE is almost unchanged and, on average, vacancy rates can be estimated with the given model with a deviation smaller than one percentage point.

Due to the model specification, which is in parts log-log and log-linear for the *Population Change* variable, the interpretations of the parameter estimates differ. Changes in the logarithmized variables by one percent lead to percentage changes in the dependent variable by the parameter estimate. An increase of the *Cold Rent per SQM* variable by one percent leads, c.p., to a decrease of the *Vacancy Rate* variable by 2.05 %. Caution needs to be paid to the interpretation of the change of the *Vacancy Rate* variable since it is already given in percent. Thus, in the case of the cold rent, an increase or decrease in the *Vacancy Rate* variable is not equal to a change of the vacancy rate by 2.05 percentage points, but the rate is changed by this percentage, which is a severely smaller change of the vacancy rate. The effects of the other logarithmized variables, the interaction terms, are much smaller. Changes of the *Normalized Number of Listings* variable by one percent result in changes much smaller than one percent, i.e., 0.04 % for the sparsely populated rural districts, 0.05 % for the rural districts with beginning concentration processes, and 0.15 % for the large cities not attached to administrative districts. The interpretation of the *Population Change* variable differs from that. Changes in the *Population Change* variable by one unit, i.e., by one percentage point, lead, c.p., to changes in the *Vacancy Rate* variable by $100 * (e^{-0.311} - 1)\,\% = -26.73\,\%$. Due to the differences in the units, a direct comparison of the different estimates is impossible. However, it is obvious that the vacancy rate is much more sensitive to changes in the *Cold Rent per SQM* or the *Population Change* than to changes in the N*ormalized Number of Listings*.

For the interpretation of the importance of the individual parameter estimates, the intuitive and widespread approach of dominance analysis, introduced by Budescu (1993) and refined by Azen & Budescu (2003), is used to describe the relative importance of the different parameter estimates for the regression analysis. Based on this approach, the python library dominance-analysis allows the calculation of a so-called percentage relative importance that provides a general indication of the importance of the included independent variables in a more general way than a simple comparison of partial R^2 or similar measures (Shekhar et al., 2023). The results of the dominance analysis correspond to the previously made observation that the

interaction terms are less important than the other variables. By far the most crucial variable is the *ln(Cold Rent per SQM)* variable, 64.37 %, followed by the *Population Change* variable, 25.23 %, and the *ln(Normalized Number of Listings)* variable included in the interaction terms, 7.93 % for observations located in large cities not attached to administrative districts, 1.79 % for observations in sparsely populated rural districts, and 0.38 % for observations in rural districts with beginning concentration processes. Remarkable is the result for the adjusted R^2 for a regression of the dependent variable on the *ln(Cold Rent per SQM)* in isolation, which leads to a value of 0.774 and underlines the importance of the cold rent in estimating the vacancy rate.

5.2.2 Spatial Regression

The spatial regression analysis allows the integration of spatial effects that lead to weak spatial autocorrelation expressed by Moran's I of 0.14 in the final standard regression model. To choose between the SEM and SLM, introduced in Chapter 4.2.2, in addition to the substantive justification in that chapter, quantitative tests were conducted in this chapter to guide the model choice. According to Anselin & Rey (2014, p. 110), an LM lag and an LM error test are performed to identify the correct type of spatial model. Since both tests are significant, as can be seen from Table 23, additional robust LM lag and LM error tests are recommended.

Table 23

Specification Tests Spatial Model

Test	Test statistic	P-value
LM error test	30.997	<0.001
LM lag test	11.203	<0.001
Robust LM error test	18.634	<0.001
Robust LM lag test	0.971	0.324

Note. Own research.

From these robust tests, only the robust LM error test remains significant, which implies applying a spatial error model. Thus, the substantive rationale given in Chapter 4.2.2 and the

quantitative tests lead to contradictory conclusions regarding the model to be applied. As already discussed in Chapter 4.2.1 with respect to the choice of the weight matrix, in econometrics, the substantive argument is regularly given preference over the purely quantitative one, especially in the case of contradictory statements as described here. Thus, the SLM is applied and presented in detail. For the sake of completeness, however, Appendix C - 12 also reports the results of the SEM, performing expectably worse concerning the objective of reducing spatial autocorrelation, measured by Moran's I.

The results of the SLM are shown in Table 24 and the examination of the regression assumptions can be found in Appendix C - 8 to Appendix C - 11.

Table 24

Results Spatial Lag Regression Model

Variable	Coefficient	Standard error	t-statistic	P-value	[0.025	0.975]	VIF
Constant	4.716	0.257	18.351	<0.001	4.211	5.221	-
ln(cr)	-1.811	0.107	-16.969	<0.001	-2.021	-1.601	1.914
pc	-0.265	0.051	-5.239	<0.001	-0.365	-0.166	1.639
Interaction terms							
ln(nnl) x sparse	0.025	0.018	1.428	0.154	-0.009	0.060	1.262
ln(nnl) x rural	0.036	0.019	1.947	0.052	-0.000	0.073	1.228
ln(nnl) x city	0.178	0.033	5.323	<0.001	0.112	0.244	0.814
Spatial lag term							
Wln(vac)	0.175	0.048	3.618	<0.001	0.080	0.269	1.256

Model Summary

Dependent variable:	ln(vac)	Degrees of freedom:	383
Number of observations:	390	F-statistic (P-value):	279.058 (<0.001)
BP LM test (P-value):	7.155 (0.209)	Pseudo R^2:	0.814
BP F-test (P-value):	1.435 (0.211)	MAE (Vacancy Rate):	0.838
White's LM test (P-value):	19.848 (0.282)	AIC:	273
White's F-test (P-value):	1.173 (0.284)	Moran's I (P-value):	0.079 (0.006)

Note. Own research.

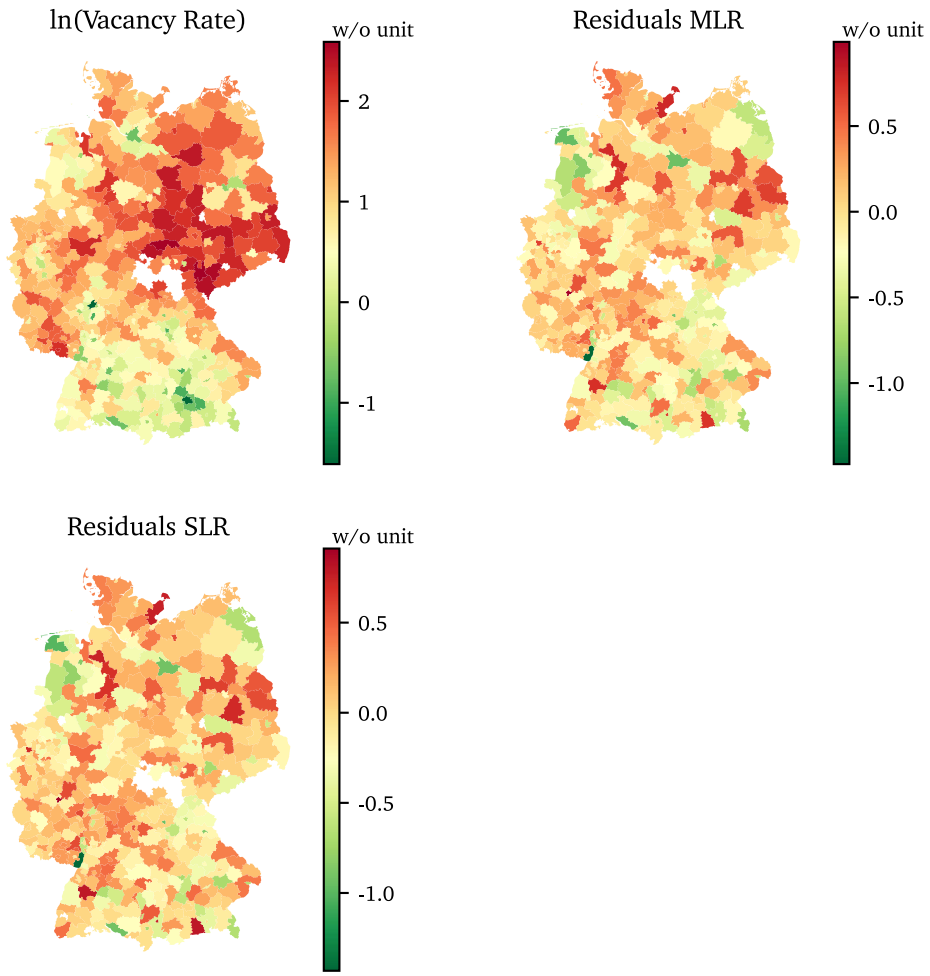
As for the final linear regression model, none of these plots of the spatial lag regression (SLR) shows a relevant violation of the regression assumptions. The same applies to most assumption-checking tests, which indicate the significance of all variables together by the F-test. However, the inspection of the variables in isolation by the t-tests reveals that the interaction term *ln(nnl) x sparse* and the interaction term *ln(nnl) x rural* are no more significant at the 5 % level. That is plausible, as both interaction terms cover spatial characteristics similar to the additionally included spatial lag term.

Furthermore, the VIFs are far below the critical value of 5-10. Both BP test statistics and both White's test statistics indicate that there is no heteroscedasticity at the 5 % level. The pseudo R^2 is not directly comparable with the standard R^2. However, it is also scaled from 0 to 1 and shows a high value of 0.814, which indicates a good fit of the model. The MAE improves slightly to 0.838, meaning that, on average, vacancy rates can be estimated with the given model with a deviation smaller than 0.85 percentage points. The AIC, as a standard criterion for the comparison of different models with the same objective, shows a slight decrease from 284, for the standard regression model, to 273 for the spatial regression model, indicating a positive development.

The spatial autocorrelation is reduced distinctly from a value for Moran's I of 0.139 to 0.079. However, it is still significant at the 5 % level. Thus, caution needs to be applied when interpreting test statistics. Figure 38 visualizes the spatial distribution of the dependent variable, the residuals of the multiple linear regression (MLR), and the residuals of the spatial regression analysis. It can be seen that the spatial autocorrelation cannot be eliminated entirely and the improvements due to the SLM are graphically weak, thus, better assessable by Moran's I.

Figure 38

Spatial Distribution of ln(Vacancy Rate), Residuals MLR and Residuals SLR



Note. Own research.

Due to the structure of the model with the lagged dependent variable as an explanatory variable, the parameter estimates cannot be interpreted directly but need to be decomposed into direct, indirect, and total effects. Table 25 shows these decomposed effects that can be interpreted as average effects regarding all observations similar to standard regression coefficients. Due to the varying effects caused by the weights matrix, the effects change from observation to observation. LeSage & Pace (2009, pp. 34–39) provide a more detailed explanation of the interpretation of these effects.

It is noticeable that the total effects are very similar to the estimated parameters of the MLR but

can be decomposed by the spatial regression analysis into direct and indirect effects caused by the spillover mechanisms. For all variables, it can be seen from Table 25 that the direct effects are more important than the spillover effects.

Table 25

Parameter Effects of Spatial Regression Model

Variable	Direct effect	Indirect effect	Total effect
ln(cr)	-1.818	-0.376	-2.194
pc	-0.266	-0.055	-0.321
ln(nnl) x sparse	0.025	0.005	0.030
ln(nnl) x rural	0.036	0.008	0.044
ln(nnl) x city	0.179	0.037	0.216

Note. Own research.

Furthermore, the strongest effects are caused by changes in the *Cold Rent per SQM* and in the *Population Change* variables. A change of the *Cold Rent per SQM* variable by 1 % leads, c.p., to a decrease in the *Vacancy Rate* variable by 2.19 %. Again, caution must be applied regarding the difference between percentage and percentage point changes. Changes in the *Population Change* variable by one unit, i.e., by one percentage point, lead, c.p., to changes in the *Vacancy Rate* variable by $100 * (e^{-0.321} - 1)\,\% = -27.46\,\%$. Changes in the *Normalized Number of Listings* variable by one percent result in changes much smaller than one percent, i.e., 0.03 % for the sparsely populated rural districts, 0.04 % for the rural districts with beginning concentration processes, and 0.22 % for large cities not attached to administrative districts. However, especially for the sparsely populated rural districts and the rural districts with beginning concentration processes, the robustness of these effects is subject to high uncertainty, as they are small and not significant at the 5 % level. Since the direction of the relationship is plausible and there exists a strong underlying intuition of a relationship between vacancy rates and the *Normalized Number of Listings* variable, more data is needed to finally derive a decision regarding the exclusion or inclusion of these variables in similar models.

5.3 Expert Assessment of the Results

The expert interviews provide additional information, in particular by incorporating knowledge about the non-quantifiable aspects of the research. These aspects include the relevance of the research, the approach chosen to answer the research question, the plausibility and quality of the results, and the potential transferability.

The relevance of the topics vacancy and ORL data is checked on the one hand by the extent to which these data have already been used by the experts and on the other hand by asking about potential areas of application. Questioning for potential areas of application takes into account that specific data could be unavailable to the experts but would be important from their perspective and could be used if they were available.

Vacancy data have already been used by several of the experts. The most common application is the use within valuation reports, but the data are also used in expert committees within the derivation process for standard land values (S2, P1, P2). In both processes, the data are especially valuable as additional information that allow to assess and justify the choices of other variables, e.g., the property rate or the risk for vacancy of the individual object. From the experts that have already used vacancy data, each expert mentioned negative aspects related to the availability of these data, i.e., that they sometimes try to gather these data by themselves or that they would use them more often if the availability would be better. Furthermore, of the experts that have not applied vacancy data in the past, two experts stated that they would apply vacancy data if the availability were better (I1, I2) and the remaining two experts know of vacancy data related projects (S1, I3). One of the experts also cited the use of vacancy data in federal policy decisions as particularly important (I2), which requires nationwide, uniform availability.

The availability, in combination with the available quality, was criticized by all experts. One of the experts explained that vacancy data are available in principle, but they come from a wide variety of sources, have to be compiled with great effort, and are not comparable (S1). Five experts stated that they are not aware of any comprehensive, nationwide source of vacancy data (S2, P1, P2, I1, I2), and one expert highlighted the Census, however, stating that it is only available with a frequency of 10 years and with delays, but is of high quality (I3). The same expert is also well familiar with the CBRE-empirica-vacancy-index and denoted it as an established source of information with no comparable alternative for the German market.

The application of ORL data is more common than the application of vacancy data within the group of experts, with six of the seven experts stating that they have already used ORL data in

the past (S1, S2, P1, P2, I1, I2). Of these experts, three said they had already used ORL data for research-related purposes (S1, I1, I2) and three mentioned the use for preparing valuation reports (S2, P1, P2). The advantages given are the large amounts of available data, the easy access, and the up-to-date availability (S2, P1, I1). However, none of the experts relies entirely on ORL data to prepare valuation reports, but they are instead applied for plausibility checks (S2, P1, P2, I1, I2). The sources from which they obtain the data vary from directly accessing the platforms for small-scale use, to using third-party data providers, to having the platform operator provide the data directly (S1, S2, P1, P2, I1, I2, I3). The seemingly most obvious alternative for large-scale use, direct provision by the platform operator, was criticized by the experts who have experience with this form of data access regarding the quantity and quality of the data provided and as not being long-lasting (S1, S2, I2). This criticism explains the use of third-party data providers by four of the experts (S1, P1, I2, I3). Furthermore, one expert explicitly criticized the high effort required to preprocess spatial information (I2). In summary, both types of data, vacancy as well as ORL data, are considered important by the experts, with particular potential for improvement in the availability of vacancy data and convenient access to ORL data.

To review the selection of variables used in the quantitative analysis, the following summarizes the data considered relevant by the experts in terms of vacancy estimation. To assess vacancy at the individual property level, typical methods such as evaluating electricity meter data, site visits, and using postal data were mentioned, which were also described in the literature review (S2, I1). Regarding the aggregate level of vacancy rates, which is particularly important in this study because of its use as a research example, one of the experts assumed that using other aggregate data is not possible because vacancy would be a local phenomenon (S2). However, six of the seven experts suggested that there is information in aggregate data that is suitable for estimating vacancy rates (S1, P1, P2, I1, I2, I3). Regularly mentioned were economic data, including general economic development and rent (S1, P1, I2, I3), demographic data (S1, P1, P2, I1, I2, I3), property characteristics (S1, P1, P2, I1, I3), the location (P2, I2, I3), real estate market data (S1, I3), and relationships in between these data (S1).

Variables that experts believe could be derived specifically from ORL data were also frequently mentioned. These included many of those already mentioned, such as information on the rent level (S1, P1, I1), the characteristics and condition of the properties (S1, I3), the location of the listings (S1, P2), but also others such as the time on market (S2, P1, I1), changes in the asking price (S2, P1) and the type of owner (I3). Thus, the experts mentioned variables from all the

variable groups derived in Chapter 2.3.4 and additional variables particularly related to ORL data, such as the time on market.

After asking for the variables the experts considered relevant, the variables used in the quantitative analysis were given to get the assumed relative importance and relevance of these variables from the experts. Since the interpretation of regression models is not routine for most experts and specific aspects of the model further complicate intuitive interpretation, such as the partial data transformation and the introduction of interaction terms, the questions referred to the simplified direct relationships between all explanatory variables and the dependent variable. These relationships are not directly comparable to the actual regression coefficients. Nevertheless, they imply the general assessment of the influence of the different variables by the experts. The experts were asked to rank the variables in terms of their importance in explaining vacancy rates from the highest importance rank 1 to the lowest importance rank 6. The results from this ranking can be found in Table 26 but need to be interpreted cautiously, as some experts ranked variables equally.

Table 26

Expert Assessment of Variable Importance

Expert	gdp	cr	pc	ls	st[30]	nnl
S1	4	3	1	6	5	1
S2	6	3	5	4	2	1
P1	5	3	1	4	6	2
P2	1	6	3	4	1	5
I1	3	3	1	6	5	2
I2	4	2	1	5	6	3
I3	6	4	2	5	3	1
Median rank	4	3	1	5	5	2
Mean rank	4.1	3.4	2.0	4.9	4.0	2.1

Note. Own research. Rank 1 = most important; rank 6 = least important.

[30] st: Settlement Type

In addition, one of the experts considered the classification of the settlement type in the ranking to be difficult and argued that this variable should instead be considered separately, which is why the ranking of it in fifth place is not strictly comparable with the other evaluations (I1). The calculation of the median and mean of the expert rankings consistently exhibit that the experts assessed the *Population Change* variable and the *Normalized Number of Listings* variable as most important in estimating vacancy. These are followed by the variable *Cold Rent per SQM*, which has medium importance, the variables *GDP per Capita* and *Settlement Type*, for which the ranking varies significantly between the experts, and the variable *Living Space*, which is ranked at the lower end in terms of importance by all experts.

Since the ranking of the variables does not allow any conclusions to be drawn about the distances regarding the importance between the variables, the relevance of the variables was also asked, which was divided into the categories very important, important, and unimportant. The results of this evaluation can be found in Appendix C - 13, but draw a similar picture compared to the evaluation of importance.

Besides the evaluation of the conducted analysis, the expert interviews provide the opportunity to draw conclusions with regard to the reached quality of the estimation and the potential transferability of the model. Therefore the experts were asked about the quality of an estimation that they would expect to be possible by the introduced variables. As a measure of quality the MAE was applied since an average absolute deviation is probably easier to guess by experts compared to more abstract concepts such as the RMSE. Three experts stated they could not assess the accuracy (S2, P2, I3). Three experts gave comparable estimates. One said that an MAE of 2 to 3 percentage points would probably be the maximum in terms of quality (S1). One said it is probably possible to reach an MAE of 2 percentage points or better in more than 50 % of the cases (P1), and one assessed an MAE of 2.5 percentage points as desirable (I1). All of the experts assumed these qualities as potential optima. Finally, one expert found it challenging to estimate an MAE and instead estimated that it is possible to achieve an accuracy of 50 % deviation from the percentage value in 95 % of cases (I2). Besides these estimates, the experts made additional remarks regarding the quality. One expert stated that the added value of the model is the relative comparability of the districts, and therefore the accuracy of the absolute estimates is not of central importance (S2), and another expert highlighted the importance of the data quality used to estimate the model (P2).

The possible transferability of the model can be divided into temporal and spatial transferability. However, the results are similar for both dimensions and it can be summarized that the closer in time or space the point to which transfer is to be made, the better the

transferability. All experts said that the fundamental relationships can probably be transferred in time, however, they made several restrictions. Changes in the overall contextual factors, such as evolving legislation, changing economic incentives, or shifting preferences, are of particular importance since they were mentioned by five experts (S1, S2, P1, P2, I2). Three experts assumed that it could be possible to transfer the estimates directly in the short term, but in the long term, the estimates need to be reviewed (S1, P1, I1).

For the assessment of the spatial transferability, it was distinguished between a transferability within the same area of examination, Germany, but on another level, e.g., on the municipality or state level, and a transferability at the same level but to another area, e.g., another country. A transferability within Germany on another level was assumed possible by all experts expecting fundamentally transferable relationships. However, several restrictions were given again, including changing parameter sizes (S1, I1, I2), availability of appropriate data at the intended level (S2), and possibly necessary weighting of input data (S1). The transferability to other countries was seen more critically. It was mentioned that other preferences and legislation could exacerbate the application of the same model in another country, and especially the size of the parameters should be verified (S1, S2, P1, I1, I3). In summary, the expert interviews provide additional information, both in terms of verifying the approach used and in terms of aspects not covered by the methodology, which will be discussed together with the other findings in Chapter 6.

6 Discussion

The discussion combines all findings to answer the research questions building upon the outcomes of the literature review, the description of the listing data, the results from the bivariate analysis, the results from the regression analysis, and the results from the expert interviews. Therefore, the findings are critically discussed, confronted, and interpretative explanations are provided. The structure of the discussion is based on the research questions.

6.1 Online Real Estate Listing Data

The literature review indicates that ORL data are well suited for application in research due to their widespread use in the scientific literature and their various advantages, including their availability, comprehensive market coverage with detailed information on the properties offered, and their homogeneous structure. However, these advantages are confronted by disadvantages like the unverified user input and the deviation of asking and transaction prices. What is missing so far is a comprehensive analysis of a large data set collected directly from an online real estate platform to be able to assess the completeness and quality of these data and evaluate their suitability for scientific use. Thus, the collection of an extensive ORL data set with more than seven million listings provides the opportunity to derive a fundamental quantity and quality assessment that can serve as a basis for future research, discussed in the following.

Concerning RQ 1, from the detailed analysis of the listing data that can be found in Chapter 3.1.2, it can be concluded that they generally provide a comprehensive, reliable, and thus valuable basis for scientific evaluations. However, this general assessment of the data can change upon closer inspection. Therefore, a more detailed assessment of the listing data is differentiated into an assessment of the quantity and an assessment of the quality of the listing data and the different variables included, as well as the resulting implications. The quality assessment extends not only to the correctness of the data but also to other characteristics, such as the possibility of automated processing. Since the sample covers a large part of the total data base, the size and content of the data set allow general conclusions to be drawn. At the same time, this means that conclusions may change for subsamples if they are not equally distributed, i.e., if subsamples are not selected randomly but based on criteria such as location or rent level.

In terms of the quantity of the data, Table 3 in Chapter 3.1.2 shows that the average completeness of the collected variables is 85.97 %, and the median completeness is 91.47 %, which can be assessed as high. Of the collected 29 variables, 27 provide information specified

by the person offering the apartment, one variable is defined by the online real estate platform, and one variable is added to capture the temporal perspective. 16 variables are available in more than 90 % of the observations, and 26 are given in more than 70 % of all observations. Only one variable, *Energy Demand*, is reported in less than 50 % of all observations. Therefore, the data availability can be considered generally good. Limitations in the availability of data for research purposes are probably less due to missing information on the available listings but to the fact that only a small part of the total building stock is offered. This effect of scarce offers is especially problematic in sparsely populated areas if the spatial distribution is simultaneously relevant, as the number of observations in these cases may be too small to draw conclusions.

Characteristics that can be used to assess the data quality include the data type, the overall plausibility of the majority of the observed variable values, and the individual plausibility of the observed variable values. The data type of the observed variables is an important indicator of data quality since its specification already affects essential features. These features include automated evaluability and customizability. Data fields that only allow numeric input are usually easier to evaluate automatically but offer limited possibilities for handling exceptional cases deviating from the standard. The opposite is true for free text fields. Exceptional cases can be described well, but an automated evaluation is only possible to a limited extent or requires considerably more complex evaluation techniques. The data types of the collected variables are heterogeneous and free text entry as well as numerical data can each be found 11 times, categorical data 7 times, and data in date format one time. *Cold Rent* and *Utilities*, for example, are easy to analyze due to their numerical format. In the case of the *Utilities* and *Heating Costs*, however, it is also apparent that misunderstandings can arise due to the lack of a possibility of more detailed explanations. The opposite is the case with the *Designation of Municipality* variable. In principle, it is possible to insert any municipality designation, supplemented by additional specifications and explanations, which allows a precise description of the location. However, the possibility of free text entries also leads to difficulties in their evaluation caused by intentional and unintentional misuse of the input field. Therefore, more standardization would be desirable for the location information.

For very large data sets, assessing the data quality by evaluating the correctness of the individual information is usually impossible if no validation sample exists or is collected with considerable effort. However, one possible approach is to examine the overall plausibility of the observed variables by evaluating the distribution characteristics for numeric variables, the frequency of the categories for categorical values, and the regularly occurring words for free text fields. The distributional characteristics and the word and category frequencies are in an expected range

for most variables. Exceptions are only the condition and quality descriptions, which seem to be positively biased. Besides these apparently positively skewed distributions, it cannot be ruled out that other variables are slightly skewed. For example, it cannot be precluded that the contracted cold rent systematically deviates positively or negatively from the asking rent, depending on the market situation. This bias is also reflected in the literature, as can be seen in the discussion about the differences between asking and transaction prices in the literature review in Chapter 2.2.2, but does not seem to be a severe problem according to the referenced previous research, wherefore the data quality can be assessed as generally good.

As an alternative to evaluating the individual plausibility of the observed variable values, it is also possible to try to specifically identify outliers and examine these outliers for plausibility in order to be able to filter out incorrect values. The detection of outliers is particularly possible for numerical variables. However, it does not seem plausible to rely solely on a standardized criterion for detecting outliers since the objective is not to filter out outliers that differ from other values by a certain margin but rather to filter out values that bias the observations due to their nature of not being a valid real estate listing. From the example of the *Cold Rent*, it can be seen that even the small minority of values at the very low and very high ends, which represent less than 0.05 % of all observations, contain a relevant part of observations that appear to be reliable. The further one moves away from these extremes, the more reliable the values appear to be. Thus, it seems generally possible to define a relative limit, depending on the particular accuracy requirement of the respective application, which serves as a boundary for choosing which values to include and which values to exclude. Generally, good indicators of erroneous observations are numeric values that show signs of being the result of a mix-up of English and German thousand separators or observations that contain the term test in any of the free text fields. In addition to checking the quality of the data using the approaches described above, it can also be argued that it is plausible that the majority of the data reflects reality, since most of the variables can be easily checked by the prospective tenant at the latest during the inspection. Thus, there is a high incentive to provide the correct information in advance.

Listing data thus offer great potential overall, as in most cases, they accurately describe reality. However, due to the limitations in quantity and quality discussed previously, appropriate data cleansing and data preparation methods have to be applied, depending on the intended use of the data and the requirements thereof. The results of the literature review and the analysis of the listing data are also supported by the statements of the experts. The use of ORL data in various fields of application can be confirmed by the fact that six of the seven interviewed experts have already used ORL data in the past. Furthermore, they endorsed the advantages

derived from the literature review and data analysis, highlighting the actuality, the easy access, and the good availability. However, they also criticized the inadequate provision of large-scale ORL data for scientific applications in the past. Summarizing all results, online real estate listing data can generally be assessed as a valuable source of information in terms of quantity and quality. However, depending on the intended application, the specific characteristics of the various variables included in ORL data need to be considered, and data preprocessing can become necessary, e.g., mentioned by one of the experts with regard to the spatial component of the included information. These aspects were considered in the choice of the variables that were examined regarding their suitability for estimating vacancy rates.

6.2 Relationships Between Online Real Estate Listing Data and Vacancy

Although there are various approaches to measuring or estimating vacancy, vacancy data is a highly relevant topic, and the use of ORL data has become common in research, the relationship between the information contained in ORL data and vacancy is underexplored. Based on the literature review, groups of essential variables in estimating vacancy could be found and verified by expert interviews.

Combining this knowledge regarding the relevant variables with the insights from the detailed analysis of the ORL data set led to the choice of the variables *Normalized Number of Listings*, *Cold Rent per SQM*, and *Living Space* from the ORL data set and the supplementation of the variables *Vacancy Rate*, *GDP per Capita*, *Population Change* and *Settlement Type* from additional publicly available data sources. Based on their descriptive statistics, transformations of the variables *Vacancy Rate*, *GDP per Capita*, *Normalized Number of Listings*, and *Cold Rent per SQM* by the natural logarithm were considered. A bivariate analysis was conducted to examine the isolated relationships between the potential explanatory variables and the *Vacancy Rate* variable to finally decide which of these transformed and untransformed variables to include in the regression analysis in which form. The results from the bivariate analysis are discussed for each of the variables in the order of the bivariate analysis in detail subsequently to add interpretations to the results and compare the different outcomes, which answers RQ2.

The *GDP per Capita* variable and the *Vacancy Rate* variable showed the closest fit for the combination of both variables transformed by the natural logarithm. The negative relationship expresses that an increase in one of the variables leads to a decrease in the other. This relationship is expectable, as it implies that districts with a well-functioning economy have lower vacancy rates, which can be justified by the fact that these districts typically attract

workers who, in turn, need and can afford housing. Looking at the economic perspective in isolation, this attraction should lead to the absorption of all available housing as long as households have an overall positive benefit from this economic strength that is not offset by increased costs, e.g., due to more expensive housing. Pearson's *r* takes a value of -0.58, implying a moderate correlation of the variables (Schober et al., 2018, p. 1765). Since L. Wang et al. (2019, pp. 8582–8583) included the GDP in their analysis trying to estimate vacancy rates of Chinese cities from night-time light data, a comparison of the fit of an R^2 of 0.39 derived in their study can be drawn. Due to the difference between the data sets, the values are not directly comparable. However, the R^2 of 0.39 is in a similar range compared to the R^2 of 0.33 resulting from the correlation coefficient of -0.58 in this study and thus contributes to the plausibility of the result. In conclusion, the GDP per capita shows a moderate fit to vacancy, which is also confirmed by other studies, which is why it can be assessed as a generally reasonable variable in modeling vacancy rates.

The *Normalized Number of Listings* variable did not show a strong linear trend for any of the variable combinations. However, a closer inspection of the *Normalized Number of Listings* variable, separated by the settlement type, revealed that linear trends can be found depending on the settlement type. Especially for the rural districts with beginning concentration processes and the large cities not attached to administrative districts, the scatter plots of the transformed variables displayed clear linear trends confirmed by values for Pearson's correlation coefficient of 0.48 and 0.61, which both imply moderate linear correlations. In principle, a relationship between the *Normalized Number of Listings* and the *Vacancy Rate* seems to be a logical consequence, as it seems reasonable to assume that vacant apartments are offered, which is also reflected in the positive correlation of these variables. However, various uncertainties are linked with the *Normalized Number of Listings*. First, it is possible that the supply structure is not homogeneous over the entire area of examination. Especially problematic in this context are changes in the percentages of vacant apartments offered in different districts. Such changes could be possible across different types of districts, especially in areas with very high and very low vacancy rates, as supply patterns could change for them. In areas with very low vacancy rates, for example, a relevant part of the offers could be facilitated via informal networks, and in areas with very high vacancy rates, some providers could give up and no longer advertise the apartments at all. Second, the data base of the housing stock is imperfect, as the underlying statistic does not differentiate the type of housing unit, and the structure of the housing market changes between different districts. Thus, separating the *Normalized Number of Listings* by the

Settlement Type appears to partially address these shortcomings by reducing the described biases.

The combinations of the *Cold Rent per SQM* variable and the *Vacancy Rate* variable show strong relationships and the application of the natural logarithm has a clearly visible positive effect in terms of linearization. Compared to the other variables, the effect of a correlation coefficient of -0.88 stands out and can be designated a strong to very strong relationship. However, this effect seems plausible as the *Cold Rent per SQM* variable is closely related to the estimated *Vacancy Rate* variable and additionally benefits from the applied data. First, because it is possible to derive rents specifically for apartments, which is the type of housing unit estimated in the CBRE-empirica-vacancy-index, in contrast to, e.g., the housing stock data, and second, because it is a variable that can be derived directly from the market and is not based on estimates, as most of the other variables included. The substantive relationship is primarily based on fundamental assumptions regarding economic processes already described, for example, in the fundamental work of Rosen & Smith (1983). Simplifying, it can be assumed that missing demand for housing leads to vacancy and housing demand can be increased by the housing owners through adjusting their asking rent downwards to an individually economically justified lower bound. Furthermore, in the opposite case of very low vacancy rates, it can be assumed that housing owners adjust their asking rents upwards to increase their earnings. These effects should be especially visible in asking rents as they are newly negotiated and, therefore, are a better measure than rents from existing contracts. Thus, vacancy rates should be directly reflected by asking rents. Compared to the R^2 of 0.39 for house prices, derived in the study of L. Wang et al. (2019, pp. 8582–8583), the *Cold Rent per SQM* variable leads to a distinctly better fit of an R^2 of 0.78 for the data set and rent variable used in this study.

The *Population Change* Variable reflects the development of the population, which is assumed to be an important estimator for housing vacancies, as the housing stock adjusts distinctly slower than the population, particularly when the population change is significantly influenced by positive or negative migration. Based on the univariate analysis, the bivariate analysis was only conducted for the untransformed *Population Change* variable, which displayed a linear shape that is also reflected in the correlation coefficients of -0.55 for the combination with the untransformed *Vacancy Rate* variable and -0.50 for the combination with the transformed *Vacancy Rate*. The correlation is at a moderate level and the relationship is negative. The direction of the correlation is as expected, implying that the lower or more negative the population change, the higher the vacancy rate, and vice versa. Not directly comparable, since their study is examining the relationship between the vacancy rate and some normalized total

population, L. Wang et al. (2019, pp. 8582–8583) found a linear relation with an R^2 of 0.72 that is remarkably higher than the R^2 of the *Population Change* variable in this study of 0.3 and 0.25. This deviation shows that housing markets in countries as different as Germany and China are not directly comparable in certain aspects and that caution is needed when transferring findings from one market to another.

The *Living Space* variable was considered because it is well available in the ORL data and represents one of the variables from the group of property characteristic variables that were regularly mentioned in vacancy research. However, property characteristics are assumed to rather influence vacancy at the individual building level and not at the aggregated district level, as this would imply that one chooses the district of his or her apartment based on which characteristics are typically available in this district, which seems unlikely on a large scale. The evaluation of the bivariate analysis only partially confirms this assumption by the scatter plots that do not reveal a clear linear relationship. However, the shape of the scatter plots shows a slightly decreasing trend, which is also reflected in the negative correlation coefficients at a moderate level. An underlying reason for this trend could be that rent levels are lower in areas with high vacancy rates, allowing tenants to rent larger apartments while the smaller apartments remain on the market. This, however, would imply a correlation between the *Cold Rent per SQM* variable and the *Living Space* variable, which indeed is indicated by the scatter plot matrix in Figure 34 that could cause multicollinearity problems in the regression analysis. Thus, the *Living Space* variable was initially considered for the regression analysis, but with a particular focus on a possible exclusion of it.

Finally, the *Settlement Type* variable was considered as a variable from the group of locational characteristics that could play a role in estimating vacancy rates. Since it is a categorical variable, the analysis examined the existence of significant differences between the *Vacancy Rate* variable values within the different settlement types. Significant differences between most settlement type combinations could be found, leading to its inclusion in the regression analysis. The underlying relationship could, for example, be related to phenomena like urbanization or rural exodus.

Comparing the results from the different data sources, the bivariate analysis indicates a contribution of the ORL data to increasing the transparency of the real estate market by explaining vacancy rates, particularly through the *Cold Rent per SQM* variable, but possibly also through the *Normalized Number of Listings* and further additional variables. Different regression

analyses were conducted to evaluate how well these data can explain vacancy rates in combination.

6.3 Vacancy Rate Estimation Using Online Real Estate Listing Data

The results from the univariate and bivariate analyses have laid the foundation for conducting the regression analyses with the objective of explaining vacancy rates of German districts based on ORL data and additional publicly available data sources. Therefore a base model including all previously derived variables was built and refined to a final model. Furthermore, regression models were tested that particularly consider the spatial aspect of the vacancy rate estimation. All models were validated by comparison of multiple statistical measures. The models were compared with each other by the AIC, assessed by the experts through the MAE, and compared to other models with a similar objective by the (pseudo) R^2. Furthermore, the models were derived based on a comprehensive literature review and the experts were questioned in detail about the model components.

The base regression model served as the starting point for deriving the final model. From this base model, it could be concluded that the variables *Living Space*, *ln(GDP per Capita)*, the spatial dummies, and one of the interaction terms should be excluded. The exclusion of these variables was based on their insignificance, multicollinearity problems, and, in particular, consistent explanations for these effects.

The low relevance of the *Living Space* variable was already assumed from the results of the bivariate analysis and the assumption that aggregate property characteristics do typically not affect, which district is chosen for living. Instead, possible correlations with the *ln(Vacancy Rate)* variable could be caused by its relation with economic effects that are already included in the *ln(GDP per Capita)* variable and the *ln(Cold Rent per SQM)* variable. Regarding an effect already included in another variable, a similar interpretation can be derived for the *ln(GDP per Capita)* variable, which is closely related to the *ln(Cold Rent per SQM)* variable. Since the *ln(Cold Rent per SQM)* is apparently more closely related to apartment vacancies and also shows a much better fit within the bivariate analysis, the *ln(GDP per Capita)* variable is excluded. The inclusion of both the spatial dummies and the interaction terms is not reasonable, as both seem to explain a similar effect. This could be caused by the *ln(Normalized Number of Listings)* variable exerting only a small but relevant effect, simultaneously implying the exclusion of one of the interaction terms with regard to the dummy variable trap. Using the significance of the different variables as decision support, the spatial dummies and the interaction term consisting of the

ln(Normalized Number of Listings) variable and the spatial dummy for the urban areas are excluded. Besides this general assessment of the variables, the residual plots also showed that a few districts exert relevant leverage on the included *Population Change* variable and some interaction terms. Considering their potential biasing effect, these data points were excluded, which is also reasonable from a substantive perspective as they are likely attributable to one-time effects. For instance, it is implausible that certain districts experience unusually large changes in population every year. Instead, it is more probable that such changes are isolated incidents.

Concerning RQ3, the derived final model is able to estimate vacancy rates with an adjusted R^2 of 0.81 and an MAE of 0.86, which is remarkably accurate compared to other results presented in the literature and the accuracy assumed by the experts. Other approaches to estimate vacancy rates led to comparable and lower accuracies, e.g., an adjusted R^2 of 0.66 in the study by Du et al. (2018, p. 1) using remote sensing and additional geospatial data, an R^2 of 0.78 in the study by L. Wang et al. (2019, p. 8583) using socio-economic data but restricting their sample to comparable cities, and an R^2 of 0.73 by the study of Z. Chen et al. (2015, pp. 2188–2194) using remote sensing and land cover data also restricted to metropolitan areas.

As expected, it was hard for the experts to estimate possible accuracies due to the complexity and missing experience regarding that specific case of vacancy estimation. Therefore, four experts could not estimate a possible accuracy as questioned. However, three experts gave estimations that were all close to each other and in the range from 2 to 3 percentage points for the MAE as a potential optimum of the model. The actual model performed distinctly better, with an MAE of 0.86 percentage points. This performance is influenced by some limitations that need to be considered when assessing the results, which are given at the end of this chapter.

An analogous observation can be drawn in relation to the evaluation of the significance of the various variables by the experts. In general, their assessments draw a similar picture to the regression analysis. The variables *Population Change*, *Normalized Number of Listings*, and *Cold Rent per SQM* are all included in the final model and the variables *Living Space* and *GDP per Capita* are excluded from the final model. This reflects the opinion of the experts, rating the three included variables as the three most important ones, while the excluded variables are two of the three variables rated as less important. In detail, however, the estimates vary widely and the relevance of the *Cold Rent per SQM* variable, in particular, is significantly underestimated and ranked only in third position regarding its importance while quantitatively being by far the most important variable. This underestimation of the importance of the *Cold Rent per SQM*

variable could be caused partially by the underlying data but also by the types of relationships between the variables. The data-related problem resulting from the data sources and their accuracy was already addressed in the previous chapter. Nevertheless, especially the relationship between the *Cold Rent per SQM* variable and the *Vacancy Rate* variable seems powerful and possibly underestimated.

While all other variables examined by the bivariate analysis describe relationships that in isolation either influence vacancy, e.g., the *Population Change* variable, or specify a characteristic that is assumed to be a good indicator of vacancy, e.g., the *Normalized Number of Listings* variable, the *Cold Rent per SQM* variable unites these information through the market mechanisms in one particular number that is directly observable in ORLs. From an economic perspective, the rent is the price that results from demand and supply, which in turn are influenced by various aspects, e.g., the housing stock, which is included in the *Normalized Number of Listings* variable, or the solvency of the potential tenants, which is partially explained by the *GDP per Capita* variable. Assuming that the rent unites these aspects in a concrete figure, approximating the abstract concept of attractivity, leads to the conclusion that districts that exhibit higher rents are more attractive and a higher attractivity leads to lower vacancy numbers as it supports owners in enforcing their rent claims on the one side and fuels demand on the other.

The final standard regression model still exhibits weak spatial autocorrelation measured by Moran's I of 0.14, which is significant at the 5 % level. Due to the potential negative influences of this spatial autocorrelation, specification tests were applied and spatial regression models were estimated. The SLM was able to reduce Moran's I as a measure of spatial autocorrelation further to a value of 0.08 but still being significant at the significance level of 5 %. Thus, caution needs to be paid when interpreting the significance of the parameter estimates of the model, particularly that of estimates close to the critical values. Nevertheless, this problem does not seem to be severe for most cases of application for two reasons elaborated previously. First, spatial autocorrelation is only weak, and second, spatial autocorrelation does not bias the estimates. Thus, the remaining spatial autocorrelation is unproblematic, for example, when building upon the model for deriving prediction models or for conducting a general assessment of the data or factors explaining vacancy rates, which includes the intended applications of this study. However, if the application of the results specifically requires a certain significance level, e.g., when attempting to prove causal effects at the given significance level of 5 %, further attempts should be made to reduce spatial autocorrelation, e.g., by testing other model types. Such an approach, however, comes with other downsides, e.g., data mining related problems.

The calculation of the total effects of the parameter estimates of the SLM regression supports these conclusions, showing values very similar to the parameter estimates of the final standard regression model. Returning to the concept of Occam's Razor, it can be concluded that the spatial regression model is not necessarily superior to the standard regression model. In particular, when ease of interpretation is additionally required, using the standard regression model can also be justified.

Finally, it can be concluded that ORL data can contribute considerably to increasing transparency through the estimation of vacancy rates. However, this contribution is influenced by limitations that need to be considered when assessing the results and possible applications.

6.4 Assessment of Research Design

To examine the influence ORLs can have on real estate market transparency in general and on explaining the vacancy rate in particular, a research design consisting of multiple methodological steps was developed and applied. The research design and the methodological steps are not inherent to the research question analyzed in this study and thus can be abstracted from the specific problem and can potentially be transferred. The evaluation of this research design, which is conducted by assessing the methodological steps individually and in their entirety subsequently, enables an assessment of its suitability for this study and in general, which facilitates drawing conclusions regarding transferability.

Based on the findings from the literature review and the therefrom resulting requirements, the specific data collection process was determined. Web scraping was used to collect a suitable, extensive ORL data set. Due to the unavailability and impossibility of collecting validation data for a data set of this scope, alternative data validation methods had to be found. Univariate analysis methods, in particular assessing the quantity and quality of the data based on investigations of aspects such as distributions, extreme values, and word frequencies, lead to the conclusion that a very large share of the ORL data appears to be plausible. Especially in combination with investigating random samples, this approach to data validation seems to be a useful alternative when validation data are not available or cannot be collected for large data sets, which is inherent for data derived through web scraping in the previously described setting.

Furthermore, the preprocessing of the data, including a detailed derivation of the data exclusion, preparation of spatial information, and data aggregation and transformation,

positively influenced the results and was found to be valuable in improving the quality of the data obtained through web scraping, assessed, e.g., by the strength of correlations between dependent and independent variables. These approaches of data preprocessing and their underlying ideas can be transferred, entirely or in parts, to other web scraped data sets, depending on the characteristics of these data, e.g., the inclusion of spatial information. By doing so, these approaches propose procedures for handling comparable data sets in order to obtain improved results.

Based on these preprocessed data sets, typical bivariate and multivariate quantitative methods can be applied for data exploration and analyzing potential relationships. The extensive bivariate analysis of relationships between the dependent and the independent variables proved to be particularly useful for exploring more detailed dependencies than the statements derived from the literature review and thereby laid the foundation for model building. Both the standard regression analysis and the spatial regression analysis showed that they could achieve remarkable results in terms of accuracy by combining data derived from web scraping with standard data sources, such as data from statistical offices. However, the data derived through web scraping not only added a valuable perspective but, according to dominance analysis, had the most important impact. Since no directly comparable studies exist, the expert interviews provided the possibility to validate the results and, furthermore, to assess the quality and transferability of the results as well as of the developed methodological framework.

Summarizing, in the absence of easily accessible data, web scraping and the methods building upon these data implemented in this study can provide a valuable alternative for closing this gap. For real estate market-related research questions, especially ORL data seem to be attractive due to their detailed information and good market coverage.

6.5 Limitations

The previous chapters provide answers regarding the research questions and related topics. However, it is important to acknowledge and address the inherent limitations that were partially indicated previously. In particular, multiple limitations arise from the research boundaries. The area examined was restricted to German districts for reasons such as the homogeneity of data and availability of the data. This implies that the results obtained are inherently valid for this specific area and level of examination and that their transferability needs to be tested for other areas or levels of examination. Nevertheless, the expert interviews

indicate that a transferability of the level could be possible, provided that the respective data are available. The transferability to other areas mainly depends on the similarity of the markets.

Furthermore, the results obtained also depend on the time component of the data used. The combination of the collected ORL data and CBRE-empirica-vacancy-index allowed the derivation of results for the year 2019. Three experts assumed that the results could be directly transferable in the short term, but most experts also assumed that a specific focus should lie on changing circumstances, e.g., evolving legislation or preferences. It can be concluded that the more distant in time or space a potential transfer is, the more attention needs to be paid to evaluating if this transfer is possible. Additionally, the results were derived for apartment vacancy rates, also attributable to the data that could be used. However, in this case, the transferability of the model or the results to other market segments should be strongly doubted, as the market mechanisms for different submarkets might differ substantially.

Besides these limitations arising from the research boundaries, the data used has characteristics that influence the results. The CBRE-empirica-vacancy-index was the best-suited vacancy data that could be acquired for the combination with the ORL data due to its actuality and quality, which could also be confirmed by the experts. However, this data comes with certain restrictions that become apparent when applying the classification of vacancy descriptions and definitions derived in the literature review. The vacancy data of this index refers to a specific point in time and thereby does not distinguish between vacancies of different durations. Furthermore, it refers to market-active vacancies, which are defined by the individual definitional approach given by this index, and the methodology of the vacancy estimation is based on the evaluation of data from a large real estate company with additional expert interviews. Thus, the derived results are directly valid only for this case. Nevertheless, it seems probable that vacancy rates of closely related definitions can be estimated by a similar approach and that the more a vacancy definition deviates, the more attention needs to be paid to evaluating the possibility of transferability.

Furthermore, although regression models are commonly used to describe causal relationships, this interpretation is not valid for all variables in the context of the models applied in this research and it is not permitted to interpret all independent variables as causal factors for vacancy. Although a substantive relationship between the variables and the connection does not seem coincidental, this relationship is not the classical causal one-directional relationship. Less attractive regions, for example, typically lead to a population outflow causing lower rents and higher vacancy rates. Thus, in this case, the rent can be considered an indicator of vacancies rather than the cause, which can also be seen from the example of a rent freeze that could lower

rents but would probably not increase vacancies, what a causal relationship would imply. In contrast, however, the change of the population could be interpreted as causal for vacancies.

Finally, the accuracy of the models needs to be discussed. Compared to the assessments of the experts and comparable results in the literature, the models performed very well. However, this performance is based on some restrictions that need to be considered and that could lead to a decreasing performance depending on other applications. On the one hand, the ORL data consists of information aggregated over space and time. Considering the weekly data acquisition and the examination at the district level, it becomes apparent that these effects considerably smooth the data and individual outliers become less relevant. Thus, using a similar approach on smaller levels or using data collected over shorter periods could decrease the model accuracy. Vice versa, it could also lead to improvements. On the other hand, a full sample estimation was conducted, a common approach for primarily examining relationships between variables, as it was done in this study. However, if the model is to be extended to a prediction model, a train/test split is necessary, which typically leads to slightly lower accuracies.

To summarize, the applied boundaries and the utilized data lead to limitations that restrict the direct transferability of the results to other applications, with the degree of restriction varying depending on the specific application. However, this simultaneously provides an excellent starting point for further research that can leverage new data sources, such as the upcoming German census.

7 Conclusion and Outlook

The research findings suggest that ORL data can contribute substantially to increasing transparency in the real estate market. The increase of transparency is furthermore not limited to the in recent times mainly used application area of price comparisons but can also be used for other important fields such as vacancy estimations, since the estimation of vacancy rates by ORL data has shown to be a valuable extension of existing research.

The research in this dissertation was motivated by the lack of real estate market information, especially with regard to vacancy data, despite the availability of comprehensive data through online real estate platforms facilitated by digitalization, as outlined in the introduction. The literature review shows that transparency is still a vague concept. However, existing definitions largely agree that increasing transparency can be assumed when additional information are provided and these information contribute to a better understanding of the market and market interrelationships. An increase in transparency by this understanding is mainly associated with positive consequences, such as higher market liquidity. Nevertheless, there are potential adverse effects, such as the possibility of increasing volatility. ORL data that can be used to increase real estate market transparency can be derived from multiple sources, including a provision by the platform itself or an entirely independent collection utilizing automated data collection methods, all with different inherent advantages and disadvantages that can influence the data quantity and quality. Reflecting the increasing availability of ORL data and their advantages, the amount of research based on ORL is also increasing. However, the combination of ORL data and vacancy research is underexplored, despite the importance of vacancy data, for example, for reducing vacancy-related problems. This research gap is remarkable as various methods for measuring or estimating vacancy or vacancy rates exist but still show a need for improvement or enhancement, particularly by methods that can estimate vacancy rates on a large scale with better accuracy than existing approaches. Nevertheless, the extensive amount of existing vacancy research can provide the most important groups of variables that should be included in vacancy estimation as a basis, consisting of variables describing locational, economic, socioeconomic, sociodemographic, and property-related aspects, also confirmed by experts in real estate markets.

Parts of these variables can be derived from ORL data and parts must be added by additional data sources. Caused by data availability reasons and to ensure a transferability of the overall research framework to similar problems, these necessary ORL data were collected by an automatized data collection approach that allowed a comprehensive data acquisition of more

than seven million listings over almost one and a half years. Based on this extensive data base, it can be shown that ORLs provide valuable information in terms of quantity and quality and enhance existing sources of information, e.g., data of expert committees with more detailed information. However, due to their nature of being unverified, they typically require preprocessing, like the exclusion of outliers and the preparation of certain information, such as spatial characteristics. Depending on the specific research application, further steps can become necessary and for the application in this research, this includes steps that need to be conducted for data aggregation and data transformation.

Complementing these preprocessed data with additional publicly available data sources builds a foundation for estimating vacancy rates consisting of variables from all variable groups commonly used in vacancy estimation according to the literature review and the expert interviews. A bivariate analysis of these variables and a variable describing the vacancy rate shows that they are a reasonable basis for estimating vacancy rates. Especially the relationship between asking rents per sqm and vacancy rates is apparent. This finding regarding the importance of the asking rent can be confirmed by more detailed regression analyses that allow additional insights to be derived. Despite the presence of weak spatial autocorrelation in the final standard regression model, this relatively simple model, consisting of six independent variables, is able to estimate vacancy rates with an R^2 of 0.81 and an MAE of 0.86. This performance is remarkably good compared with previous large-scale vacancy rate estimations and with the assumptions made by the experts regarding possible accuracies. However, it is also caused by aspects, such as data aggregation over time and space, that influence the potential transferability and can lead to decreased accuracy depending on the specific application.

The additionally tested spatial regression models can reduce spatial autocorrelation significantly. Nevertheless, the regression results stay very similar compared to the standard regression model. This finding implies that depending on the intended use of the results, careful consideration of applying different regression models should be made and more sophisticated models are not always better.

The expert interviews complement these findings with additional information. They confirm the relevance of ORL data and specifically highlight the shortage of suitable vacancy data. The general assessment of the importance of the different variables is similar to the quantitative results. However, in detail, the assumptions deviate considerably from the derived quantitative results, particularly the influence of the cold rent is underestimated. Thus, the quantitative analysis not only provides added value by deriving specific estimates and laying a foundation for further research, e.g., with the objective of a prediction model, but it also provides insights

with regard to the importance of different information for estimating vacancy rates that are so far misperceived.

The achieved results rely on a mainly quantitative research design that was chosen based on the initially identified research gap in the field of vacancy research applying the information contained in publicly available ORL data. To derive these information, a web scraper was programmed, the data were preprocessed, and various quantitative methods were applied to analyze these data combined with data from other publicly available data sources. Reflecting the previously outlined results, it was possible to derive comprehensive answers to the research questions. Furthermore, the research design is transferable to other research questions, either by applying ORL data to them or by exploiting new data sources for similar or other research questions.

Thus, these results and the applied research boundaries offer various linkages for further research regarding the data used, the approach taken, and the possible applications of the results. The most apparent possible improvement in terms of the data used is the application of newer and potentially better data for both types of variables, independent and dependent. Concerning the independent variables, the ORL data base of this study was extensive. However, it still offers room for improvement, for example, through a permanent instead of a weekly data acquisition that would allow calculating precise values for the time on market and would furthermore not miss listings offered only for very short periods. Of the additional variables supplemented, especially a higher level of detail of the information of the building stock and, generally, an improvement of the data quality of all variables supplemented would be helpful. For the vacancy rate, the coming census offers the opportunity to improve and extend results, as the census is more accurate due to its characteristic of being a full census and more detailed, as typically, the smallest area of aggregation that can be examined is at least the municipality level. Examining vacancy rates on a more detailed level could simultaneously require additional other data since the structure or components of the model could change.

Changes in the modeling approach generally offer a variety of possibilities that can be tested. For the objective of this study to explore the potential of ORL data for explaining vacancy rates, the common standard approaches were tested to fill the existing research gap and lay the foundation for further research. However, the variety of existing models, especially those that particularly consider spatial characteristics, is large and could be explored. Combined with the previously addressed missing data, more sophisticated models could also become feasible. Especially more frequent and accurate vacancy data could help to construct models that could also be able to detect trends in vacancy development. Furthermore, such data would help to

evaluate models like the model derived in this study and, with substantially increased amounts of vacancy data, neuronal network models could also be tested. Finally, data from other countries could be used to test the generalizability of the model derived in this work.

However, the results obtained in this dissertation could already form a basis for potential applications. Provided the required data can be derived regularly, a permanent vacancy monitoring could, for example, be introduced that is able to nowcast vacancy developments. Based on these findings, better decisions could be made in terms of designing funding programs or regulating zoning designations.

Finally, since the approach chosen proved to be a valuable source of additional information, the research framework of this study could be applied to other applications with a similar structure, meaning problems that are not completely solved or questions that are not conclusively answered but for which an extensive data base can be identified that could be collected by the method of web scraping and that could be analyzed with the methods applied in this study and the intention of deriving better solutions or answers.

References

Accordino, J., & Johnson, G. T. (2000). Addressing the Vacant and Abandoned Property Problem. *Journal of Urban Affairs*, *22*(3), 301–315. https://doi.org/10.1111/0735-2166.00058

Ackermann, T. (2020). Energiebedarf versus Energieverbrauch unter Einbeziehung von Langzeitmessungen zum Temperaturverlauf. *Bauphysik*, *42*(1), 1–10. https://doi.org/10.1002/bapi.201900030

Adcock, C., Eling, M., & Loperfido, N. (2015). Skewed distributions in finance and actuarial science: A review. *The European Journal of Finance*, *21*(13–14), 1253–1281. https://doi.org/10.1080/1351847X.2012.720269

AK VGR (Ed.). (2022). *Bruttoinlandsprodukt, Bruttowertschöpfung in den kreisfreien Städten und Landkreisen der Bundesrepublik Deutschland—1992 und 1994 bis 2020, Berechnungsstand November 2021* [Data set]. https://www.statistikportal.de/sites/default/files/2022-07/vgrdl_r2b1_bs2021_1.xlsx

an de Meulen, P., Micheli, M., & Schmidt, T. (2011). *Forecasting House Prices in Germany* (Ruhr Economic Papers No. 294). RWI - Leibniz-Institut für Wirtschaftsforschung. https://www.econstor.eu/handle/10419/61677

Anselin, L. (2005). *Exploring spatial data with GeoDaTM: A workbook*. Center for Spatially Integrated Social Science.

Anselin, L., Bera, A. K., Florax, R., & Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, *26*(1), 77–104. https://doi.org/10.1016/0166-0462(95)02111-6

Anselin, L., & Rey, S. J. (2014). *Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Asriyan, V., Fuchs, W., & Green, B. (2017). Information Spillovers in Asset Markets with Correlated Values. *American Economic Review*, *107*(7), 2007–2040. https://doi.org/10.1257/aer.20151714

Azen, R., & Budescu, D. V. (2003). The Dominance Analysis Approach for Comparing Predictors in Multiple Regression. *Psychological Methods*, *8*(2), 129–148. https://doi.org/10.1037/1082-989X.8.2.129

Baba, H., & Hino, K. (2019). Factors and tendencies of housing abandonment: An analysis of a survey of vacant houses in Kawaguchi City, Saitama. *Japan Architectural Review*, *2*(3), 367–375. https://doi.org/10.1002/2475-8876.12088

Baba, H., & Shimizu, C. (2023). The impact of apartment vacancies on nearby housing rents over multiple time periods: Application of smart meter data. *International Journal of Housing Markets and Analysis*, *16*(7), 27–41. https://doi.org/10.1108/IJHMA-03-2022-0035

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2018). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (15th ed.). Springer Gabler Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-56655-8

Bailey, M. J., Muth, R. F., & Nourse, H. O. (1963). A Regression Method for Real Estate Price Index Construction. *Journal of the American Statistical Association*, *58*(304), 933–942. https://doi.org/10.1080/01621459.1963.10480679

Baldenius, T., Kohl, S., & Schularick, M. (2020). Die neue Wohnungsfrage: Gewinner und Verlierer des deutschen Immobilienbooms. *Leviathan*, *48*(2), 195–236. https://doi.org/10.5771/0340-0425-2020-2-195

Bangert, D., Nelius, K., & Schmelcher, R. (2006). Wenn kein Strom mehr fließt: Erfassung des Wohnungsleerstands über Stromzähler in Berlin. *STANDORT – Zeitschrift für Angewandte Geographie*, *30*(4), 178–180. https://doi.org/10.1007/s00548-006-0352-7

Bao, H. X. H., & Shah, S. (2020). The Impact of Home Sharing on Residential Real Estate Markets. *Journal of Risk and Financial Management*, *13*(8). https://doi.org/10.3390/jrfm13080161

Barron, K., Kung, E., & Proserpio, D. (2020). The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.3006832

Bauer, T. K., Braun, S. T., & Kvasnicka, M. (2017). Nuclear power plant closures and local housing values: Evidence from Fukushima and the German housing market. *Journal of Urban Economics*, *99*, 94–106. https://doi.org/10.1016/j.jue.2017.02.002

Bauer, T. K., Budde, R., Micheli, M., & Neumann, U. (2015). Immobilienmarkteffekte des Emscherumbaus? *Raumforschung und Raumordnung | Spatial Research and Planning*, *73*(4), 269–283. https://doi.org/10.1007/s13147-015-0356-5

Bauer, T. K., Fertig, M., & Vorell, M. (2011). *Neighborhood Effects and Individual Unemployment* (SOEPpaper No. 409-2011). DIW Berlin. http://dx.doi.org/10.2139/ssrn.1965950

Bauer, T. K., Feuerschütte, S., Kiefer, M., an de Meulen, P., Micheli, M., Schmidt, T., & Wilke, L.-H. (2013). Ein hedonischer Immobilienpreisindex auf Basis von Internetdaten: 2007-2011. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *7*, 5–30. https://doi.org/10.1007/s11943-012-0125-7

BBSR. (2017). *Lücken in der Leerstandsforschung – Wie Leerstände besser erhoben werden können* (BBSR-Berichte KOMPAKT 02/2017). http://nbn-resolving.org/urn:nbn:de:101:1-201707183314

BBSR. (2020). *Raumabgrenzung: Referenzdaten—Gebietsstand 31.12.2019—Siedlungsstrukturelle Kreistypen* [Data set]. https://www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/Raumabgrenzungen/deutschland/kreise/siedlungsstrukturelle-kreistypen/siedlungsstrukt-kreistypen-2019.csv?__blob=publicationFile&v=3

BDEW. (2019). *Wie heizt Deutschland?* (Studie zum Heizungsmarkt September 2019). https://www.bdew.de/media/documents/BDEW_Heizungsmarkt_final_30.09.2019_3ihF1yL.pdf

Bekkerman, R., Cohen, M. C., Liu, X., Maiden, J., & Mitrofanov, D. (2021). The Impact of the Opportunity Zone Program on Residential Real Estate. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.3780241

Bentley, G. C., McCutcheon, P., Cromley, R. G., & Hanink, D. M. (2016). Race, class, unemployment, and housing vacancies in Detroit: An empirical analysis. *Urban Geography*, *37*(5), 785–800. https://doi.org/10.1080/02723638.2015.1112642

Berger, E. M., & Schmidt, F. (2019). *Inattention in the Rental Housing Market: Evidence from a Natural Experiment* (Discussion Paper Series Discussion Paper No. 1716). Gutenberg School of Management and Economics & Research Unit Interdisciplinary Public Policy. https://ipp-mainz.uni-mainz.de/files/2019/09/Discussion_Paper_1716.pdf

Bernstein, A., Gustafson, M. T., & Lewis, R. (2019). Disaster on the Horizon: The Price Effect of Sea Level Rise. *Journal of Financial Economics*, *134*(2), 253–272. https://doi.org/10.1016/j.jfineco.2019.03.013

Beyerle, T. (2006). Immobilienmarkt-Research in Deutschland: Ein Arbeitsmarkt für Geographen. *STANDORT – Zeitschrift für Angewandte Geographie*, *30*(3), 123–126. https://doi.org/10.1007/s00548-006-0326-9

Bhuiyan, M., & Al Hasan, M. (2016). Waiting to be Sold: Prediction of Time-Dependent House Selling Probability. *IEEE International Conference on Data Science and Advanced Analytics*, 468–477. https://doi.org/10.1109/DSAA.2016.58

Bienert, S. (2017). *Europäisierung der Immobilienkapitalmärkte* (Kompendium der DVFA Kommission Immobilien und der IREBS International Real Estate Business School Heft 18). DVFA Deutsche Vereinigung für Finanzanalyse und Asset Management e.V., IRE|BS International Real Estate Business School. https://epub.uni-regensburg.de/35730/1/Europ%C3%A4isierung%20der%20Immobilienkapitalm%C3%A4rkte.pdf

Bishara, A. J., & Hittner, J. B. (2012). Testing the Significance of a Correlation With Nonnormal Data: Comparison of Pearson, Spearman, Transformation, and Resampling Approaches. *Psychological Methods*, *17*(3), 399–417. https://doi.org/10.1037/a0028087

Bishara, A. J., & Hittner, J. B. (2015). Reducing Bias and Error in the Correlation Coefficient Due to Nonnormality. *Educational and Psychological Measurement*, *75*(5), 785–804. https://doi.org/10.1177/0013164414557639

Bishara, A. J., Li, J., & Nash, T. (2018). Asymptotic confidence intervals for the Pearson correlation via skewness and kurtosis. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 167–185. https://doi.org/10.1111/bmsp.12113

BKG. (2019). *Verwaltungsgebiete 1 zu 250000: VG250 und VG250-EW* [Dokumentation]. http://sg.geodatenzentrum.de/web_download/vg/vg250_0101/vg250_0101.pdf

Bleck, A., & Liu, X. (2007). Market Transparency and the Accounting Regime. *Journal of Accounting Research*, *45*(2), 229–256. https://doi.org/10.1111/j.1475-679X.2007.00231.x

Bloomfield, R., & O'Hara, M. (1999). Market Transparency: Who Wins and Who Loses? *The Review of Financial Studies*, *12*(1), 5–35. https://doi.org/10.1093/rfs/12.1.5

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's Razor. *Information Processing Letters*, *24*(6), 377–380. https://doi.org/10.1016/0020-0190(87)90114-1

Boeing, G., & Waddell, P. (2017). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. *Journal of Planning Education and Research*, *37*(4), 457–476. https://doi.org/10.1177/0739456X16664789

Böhm, J. (2001). ImmobilienScout 24—Case Study eines innovativen Immobilienmarktplatzes im Internet. In W. Rohmert & J. Böhm (Eds.), *E-Business in der Immobilienwirtschaft* (pp. 85–91). Gabler Verlag. https://doi.org/10.1007/978-3-322-88969-0_7

Bosker, H. R. (2021). Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies. *Behavior Research Methods*, *53*, 1945–1953. https://doi.org/10.3758/s13428-021-01542-4

Braun, R., Heising, P., & Schwede, P. (2014). *Aktuelle und zukünftige Entwicklung von Wohnungsleerständen*. BBSR. http://nbn-resolving.org/urn:nbn:de:101:1-2015021123162

Breidenbach, P., Jäger, P., & Taruttis, L. (2022). *Aging and Real Estate Prices in Germany* (Ruhr Economic Papers No. 953). RWI - Leibniz-Institut für Wirtschaftsforschung. https://doi.org/10.4419/96973117

Budescu, D. V. (1993). Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression. *Psychological Bulletin*, *114*(3), 542–551. https://doi.org/10.1037/0033-2909.114.3.542

Bundeskartellamt. (2016). *Marktmacht von Plattformen und Netzwerken* (Arbeitspapier Az. B6-113/15). https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Berichte/Think-Tank-Bericht.pdf;jsessionid=B1E6A734DD27BE0F63BFFF85C14E3B7F.1_cid387?__blob=publicationFile&v=2

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, *3*(1), 17. https://doi.org/10.1186/1751-0473-3-17

Busch, R. (2016). Inländische Wanderungen in Deutschland—Wer gewinnt und wer verliert? *Zeitschrift für Immobilienökonomie*, *2*, 81–102. https://doi.org/10.1365/s41056-016-0012-3

Cajias, M., Freudenreich, P., Freudenreich, A., & Schäfers, W. (2020). Liquidity and prices: A cluster analysis of the German residential real estate market. *Journal of Business Economics*, *90*, 1021–1056. https://doi.org/10.1007/s11573-020-00990-2

Cappiello, C., Francalanci, C., & Pernici, B. (2003). Time-Related Factors of Data Quality in Multichannel Information Systems. *Journal of Management Information Systems*, *20*(3), 71–92. https://doi.org/10.1080/07421222.2003.11045769

Cavallo, A. (2018). Scraped Data and Sticky Prices. *The Review of Economics and Statistics*, *100*(1), 105–119. https://doi.org/10.1162/REST_a_00652

Cellmer, R., & Szczepankowska, K. (2014). *Simulation modeling in a real estate market*. The 9th International Conference Environmental Engineering. http://dx.doi.org/10.3846/enviro.2014.113

Charles, W., Glenn, M., & Donna, M. (1991). The Impact of Inflation and Vacancy of Real Estate Returns. *Journal of Real Estate Research*, *6*(2), 153–168. https://doi.org/10.1080/10835547.1991.12090643

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example* (5th ed). Wiley.

Chegut, A., Eichholtz, P., Holtermans, R., & Palacios, J. (2020). Energy Efficiency Information and Valuation Practices in Rental Housing. *The Journal of Real Estate Finance and Economics*, *60*, 181–204. https://doi.org/10.1007/s11146-019-09720-0

Chen, P., & Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric Measures* (Quantitative Applications in the Social Science No. 07–139). Sage Publications.

Chen, Z., Yu, B., Hu, Y., Huang, C., Shi, K., & Wu, J. (2015). Estimating House Vacancy Rate in Metropolitan Areas Using NPP-VIIRS Nighttime Light Composite Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(5), 2188–2197. https://doi.org/10.1109/JSTARS.2015.2418201

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed). Lawrence Erlbaum Associates.

Coles, P., Egesdal, M., Ellen, I. G., Li, X., & Sundararajan, A. (2017). Airbnb Usage Across New York City Neighborhoods: Geographic Patterns and Regulatory Implications. *SSRN Electronic Journal*. https://ssrn.com/abstract=3048397

Colonnello, S., Marfè, R., & Xiong, Q. (2022). Housing Yields. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.3786959

Couch, C., & Cocks, M. (2013). Housing Vacancy and the Shrinking City: Trends and Policies in the UK and the City of Liverpool. *Housing Studies*, *28*(3), 499–519. https://doi.org/10.1080/02673037.2013.760029

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (Fifth edition). SAGE.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry & research design: Choosing among five approaches* (Fourth edition). SAGE.

Danton, J., & Himbert, A. (2018). Residential Vertical Rent Curves. *Journal of Urban Economics*, *107*, 89–100. https://doi.org/10.1016/j.jue.2018.08.002

De Nadai, M., & Lepri, B. (2018). The economic value of neighborhoods: Predicting real estate prices from the urban environment. *IEEE 5th International Conference on Data Science and Advanced Analytics*, 323–330. https://doi.org/10.1109/DSAA.2018.00043

de Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman Correlation Coefficients Across Distributions and Sample Sizes: A Tutorial Using Simulations and Empirical Data. *Psychological Methods*, *21*(3), 273–290. https://doi.org/10.1037/met0000079

Decker, N. (2021). Affordable Housing Without Public Subsidies: Rent-Setting Practices in Small Rental Properties. *Journal of the American Planning Association*, *87*(1), 62–72. https://doi.org/10.1080/01944363.2020.1798806

Deilmann, C., Effenberger, K.-H., & Banse, J. (2009). Housing stock shrinkage: Vacancy and demolition trends in Germany. *Building Research & Information*, *37*(5–6), 660–668. https://doi.org/10.1080/09613210903166739

DeJonckheere, M., & Vaughn, L. M. (2019). Semistructured interviewing in primary care research: A balance of relationship and rigour. *Family Medicine and Community Health*, *7*(2). http://dx.doi.org/10.1136/fmch-2018-000057

Dell'Anna, F., & Bottero, M. (2021). Green premium in buildings: Evidence from the real estate market of Singapore. *Journal of Cleaner Production*, *286*. https://doi.org/10.1016/j.jclepro.2020.125327

Demers, A., & Eisfeldt, A. L. (2022). Total returns to single-family rentals. *Real Estate Economics*, *50*(1), 7–32. https://doi.org/10.1111/1540-6229.12353

Deschermeier, P., Haas, H., Hude, M., & Voigtländer, M. (2016). A first analysis of the new German rent regulation. *International Journal of Housing Policy*, *16*(3), 293–315. http://dx.doi.org/10.1080/14616718.2015.1135858

Deschermeier, P., & Seipelt, B. (2016). Ein hedonischer Mietpreisindex für studentisches Wohnen. *IW-Trends - Vierteljahresschrift zur empirischen Wirtschaftsforschung*, *43*(3), 59–76. https://doi.org/10.2373/1864-810X.16-03-04

Deschermeier, P., Seipelt, B., & Voigtländer, M. (2014). Mietpreisentwicklung von Gewerbeimmobilien in deutschen Großstädten. *IW-Trends - Vierteljahresschrift zur empirischen Wirtschaftsforschung*, *41*(3). https://doi.org/10.2373/1864-810X.14-03-04

Deutsche Post DHL Group. (2018, June 29). *25 Jahre fünfstellige Postleitzahlen in Deutschland [Press release]*. https://www.dpdhl.com/de/presse/pressemitteilungen/2018/25-jahre-fuenfstellige-postleitzahlen-in-deutschland.html

Di Maggio, M., & Pagano, M. (2018). Financial Disclosure and Market Transparency with Costly Information Processing. *Review of Finance*, *22*(1), 117–153. https://doi.org/10.1093/rof/rfx009

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education*, *40*(4), 314–321. https://doi.org/10.1111/j.1365-2929.2006.02418.x

Dinkel, M., & Kurzrock, B. (2012). Angebots- und Transaktionspreise von selbstgenutztem Wohneigentum im Ländlichen Raum. *Zeitschrift Für Immobilienökonomie*, *13*(1), 5–25.

Doncaster, C. P., & Davey, A. J. H. (2007). *Analysis of variance and covariance: How to choose and construct models for the life sciences*. Cambridge University Press.

Dransfeld, E., & Lehmann, D. (2007). *Grundstückswertfragen im Stadtumbau* (1st ed.). Forum Baulandmanagement NRW. https://www.forum-bauland.nrw/wp-content/uploads/2018/07/grundstueckswertfragen.pdf

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118625590

Du, M., Wang, L., Zou, S., & Shi, C. (2018). Modeling the Census Tract Level Housing Vacancy Rate with the Jilin1-03 Satellite and Other Geospatial Data. *Remote Sensing*, *10*(12). https://doi.org/10.3390/rs10121920

Dubin, R. A. (1998). Spatial Autocorrelation: A Primer. *Journal of Housing Economics*, *7*(4), 304–327. https://doi.org/10.1006/jhec.1998.0236

Eichholtz, P. M. A., Gugler, N., & Kok, N. (2011). Transparency, Integration, and the Cost of International Real Estate Investments. *The Journal of Real Estate Finance and Economics*, *43*, 152–173. https://doi.org/10.1007/s11146-010-9244-5

Eilers, L., Paloyo, A. R., & Vance, C. (2021). Rental prices in Germany: A comparison between migrants and natives. *Scottish Journal of Political Economy*, *68*(4), 434–466. https://doi.org/10.1111/sjpe.12273

empirica ag. (n.d.). *CBRE-empirica-Leerstandsindex 2019* [Data set]. empirica regio.

empirica ag. (2021). *CBRE-empirica-Leerstandsindex 2021—Zeitreihe 2009-2020—Ergebnisse und Methodik*. empirica regio.

Etzioni, O. (1996). The World-Wide Web: Quagmire or Gold Mine? *Communications of the ACM*, *39*(11), 65–68. https://doi.org/10.1145/240455.240473

Farzanegan, M. R., & Gholipour, H. F. (2014). Does real estate transparency matter for foreign real estate investments? *International Journal of Strategic Property Management*, *18*(4), 317–331. https://doi.org/10.3846/1648715X.2014.969793

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th edition). SAGE Publications.

Fischer, M. M., & Getis, A. (Eds.). (2010). *Handbook of Applied Spatial Analysis*. Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03647-7

Florini, A. (Ed.). (2007). *The Right to Know: Transparency for an Open World*. Columbia University Press. https://doi.org/10.7312/flor14158

Fox, J. (2016). *Applied regression analysis and generalized linear models* (Third Edition). SAGE Publications.

Franz, P. (2001). Wohnungsleerstand in Ostdeutschland: Differenzierte Betrachtung notwendig. *Wirtschaft Im Wandel*, *7*(11), 263–267.

Frondel, M., Kussel, G., Sommer, S., & Vance, C. (2019). *Local cost for global benefit: The case of wind turbines* (Ruhr Economic Papers No. 791). RWI - Leibniz-Institut für Wirtschaftsforschung. https://doi.org/10.4419/86788919

Gabe, J., & Rehm, M. (2014). Do tenants pay energy efficiency rent premiums? *Journal of Property Investment & Finance*, *32*(4), 333–351. https://doi.org/10.1108/JPIF-09-2013-0058

Gabriel, R., Hoppe, T., & Pastwa, A. (2008). Intelligente Wohnungsmarktanalyse auf Basis einer Business Intelligence-Anwendung—Anforderungen, Modellierung und prototypische

Realisierung. In B. Dinter, R. Winter, P. Chamoni, N. Gronau, & K. Turowski (Eds.), *Synergien durch Integration und Informationslogistik* (pp. 287–298). Gesellschaft für Informatik e.V.

Gabriel, S. A., & Nothaft, F. E. (2001). Rental Housing Markets, the Incidence and Duration of Vacancy, and the Natural Vacancy Rate. *Journal of Urban Economics*, *49*(1), 121–149. https://doi.org/10.1006/juec.2000.2187

Gentili, M., & Hoekstra, J. (2019). Houses without people and people without houses: A cultural and institutional exploration of an Italian paradox. *Housing Studies*, *34*(3), 425–447. https://doi.org/10.1080/02673037.2018.1447093

Gerald, J. F. (2005). *The Irish Housing Stock: Growth in Number of Vacant Dwellings* (Quarterly Economic Commentary No. 24; pp. 1–22). ESRI.

Gholipour, H. F., Tajaddini, R., & Pham, T. N. T. (2020). Real estate market transparency and default on mortgages. *Research in International Business and Finance*, *53*. https://doi.org/10.1016/j.ribaf.2020.101202

Glumac, B., Herrera Gomez, M., & Licheron, J. (2018). *A Residential Land Price Index for Luxembourg: Dealing with the Spatial Dimension* (Working Paper Series 2018(07)). Luxembourg Institute of Socio-Economic Research. https://doi.org/10.2139/ssrn.3183160

Glynn, C., & Fox, E. B. (2019). Dynamics of homelessness in urban America. *The Annals of Applied Statistics*, *13*(1), 573–605. https://doi.org/10.1214/18-AOAS1200

Golgher, A. B., & Voss, P. R. (2016). How to Interpret the Coefficients of Spatial Models: Spillovers, Direct and Indirect Effects. *Spatial Demography*, *4*, 175–205. https://doi.org/10.1007/s40980-015-0016-y

Gonzalez, J. (n.d.). *fuzzywuzzy—Fuzzy String Matching in Python*. Retrieved May 19, 2023, from https://github.com/seatgeek/fuzzywuzzy

Goodman, C. B. (2018). House prices and property tax revenues during the boom and bust: Evidence from small-area estimates. *Growth and Change*, *49*(4), 636–656. https://doi.org/10.1111/grow.12261

Granados, N., Gupta, A., & Kauffman, R. J. (2010). Research Commentary—Information Transparency in Business-to-Consumer Markets: Concepts, Framework, and Research Agenda. *Information Systems Research*, *21*(2), 207–226. https://doi.org/10.1287/isre.1090.0249

Greene, W. H. (2018). *Econometric analysis* (8th ed.). Pearson.

Grekousis, G. (2020). *Spatial Analysis Methods and Practice: Describe – Explore – Explain through GIS* (1st ed.). Cambridge University Press. https://www.cambridge.org/core/product/identifier/9781108614528/type/book

Gupta, A., Mittal, V., Peeters, J., & Van Nieuwerburgh, S. (2022). Flattening the curve: Pandemic-Induced revaluation of urban real estate. *Journal of Financial Economics*, *146*(2), 594–636. https://doi.org/10.1016/j.jfineco.2021.10.008

Gupta, A., Nguyen, L., Dunning, C., & Chan, K. M. (2019). *Quantifying the Effects of the 2008 Recession using the Zillow Dataset* (arXiv: 1912.11341). https://doi.org/10.48550/arXiv.1912.11341

Haase, J. (2013). *Immobilienwerte und Fluglärm: Anwendung einer vom DIW erprobten Methode im Rhein-Main-Gebiet* (Wissenschaftliche Forschungsarbeit). Fakultät für Wirtschaftswissenschaften der TU Chemnitz.

Hagen, D., & Hansen, J. (2010). Rental Housing and the Natural Vacancy Rate. *Journal of Real Estate Research*, *32*(4), 413–434. https://doi.org/10.1080/10835547.2010.12091288

Han, J. M., & Lee, N. (2018). Holistic Visual Data Representation For Built Environment Assessment. *International Journal of Sustainable Development and Planning*, *13*(4), 516–527. https://doi.org/10.2495/SDP-V13-N4-516-527

Hannon, L., & Knapp, P. (2003). Reassessing Nonlinearity in the Urban Disadvantage/Violent Crime Relationship: An Example of Methodological Bias from Log Transformation. *Criminology*, *41*(4), 1427–1448. https://doi.org/10.1111/j.1745-9125.2003.tb01026.x

Haramati, T., & Hananel, R. (2016). Is anybody home? The influence of ghost apartments on urban diversity in Tel-Aviv and Jerusalem. *Cities*, *56*, 109–118. https://doi.org/10.1016/j.cities.2016.04.006

Heidari, M., & Rafatirad, S. (2020). Bidirectional Transformer based on online Text-based information to Implement Convolutional Neural Network Model For Secure Business Investment. *IEEE International Symposium on Technology and Society*, 322–329.

Heidari, M., Zad, S., Berlin, B., & Rafatirad, S. (2021). Ontology Creation Model based on Attention Mechanism for a Specific Business Domain. *IEEE International IOT, Electronics and Mechatronics Conference*.

Hernes, D., Lehrbass, F., & Maucy, K. (2021). *Big Data basierte Analyse des Einflusses traditioneller und neuartiger Faktoren auf Mietpreise in Düsseldorf* (ifes Schriftenreihe No. 25). ifes Institut für Empirie & Statistik, FOM Hochschule für Oekonomie & Management. https://dx.doi.org/10.2139/ssrn.3835668

Hillier, A. E., Culhane, D. P., Smith, T. E., & Tomlin, C. D. (2003). Predicting Housing Abandonment with the Philadelphia Neighborhood Information System. *Journal of Urban Affairs*, *25*(1), 91–106. https://doi.org/10.1111/1467-9906.00007

Hoekstra, J., & Vakili-Zad, C. (2011). High Vacancy Rates and Rising House Prices: The Spanish Paradox: High Vacancy Rates and Rising House Prices. *Tijdschrift Voor Economische En Sociale Geografie*, *102*(1), 55–71. https://doi.org/10.1111/j.1467-9663.2009.00582.x

Hollander, J. B., & Hartt, M. (2019). Vacancy and property values in shrinking downtowns: A comparative study of three New England cities. *Town Planning Review*, *90*(3), 247–273. https://doi.org/10.3828/tpr.2019.18

Hollander, J. B., Hartt, M. D., Wiley, A., & Vavra, S. (2018). Vacancy in shrinking downtowns: A comparative study of Québec, Ontario, and New England. *Journal of Housing and the Built Environment*, *33*(4), 591–613. https://doi.org/10.1007/s10901-017-9587-9

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric Statistical Methods* (3rd ed.). John Wiley & Sons.

Hollas, D. R., Rutherford, R. C., & Thomson, T. A. (2010). Zillow's Estimate of Single-Family Housing Values. *Appraisal Journal*, *78*(1), 26–32.

Holt, J. R., & Borsuk, M. E. (2020). Using Zillow data to value green space amenities at the neighborhood scale. *Urban Forestry & Urban Greening*, *56*. https://doi.org/10.1016/j.ufug.2020.126794

Hyland, M., Lyons, R. C., & Lyons, S. (2013). The value of domestic building energy efficiency—Evidence from Ireland. *Energy Economics*, *40*, 943–952. https://doi.org/10.1016/j.eneco.2013.07.020

Im, J., Seo, Y., Cetin, K. S., & Singh, J. (2017). Energy efficiency in U.S. residential rental housing: Adoption rates and impact on rent. *Applied Energy*, *205*, 1021–1033. https://doi.org/10.1016/j.apenergy.2017.08.047

Ionaşcu, E., Taltavull de La Paz, P., & Mironiuc, M. (2021). The Relationship between Housing Prices and Market Transparency. Evidence from the Metropolitan European Markets. *Housing, Theory and Society*, *38*(1), 42–71. https://doi.org/10.1080/14036096.2019.1672577

Jacobs, T., & Diez, B. (2018). *Wohnraumkonzept der Stadt Chemnitz—Fassung vom Oktober 2018*. Stadt Chemnitz. https://www.chemnitz.de/chemnitz/media/unsere-stadt/stadtentwicklung/konzepte/wohnraumkonzept_2018_10.pdf

Jafari, A., & Akhavian, R. (2019). Driving forces for the US residential housing price: A predictive analysis. *Built Environment Project and Asset Management*, *9*(4), 515–529. https://doi.org/10.1108/BEPAM-07-2018-0100

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Jetzek, T., Avital, M., & Bjorn-Andersen, N. (2019). The Sustainable Value of Open Government Data. *Journal of the Association for Information Systems*, *20*(6), 702–734. https://doi.org/10.17705/1jais.00549

Jin, X., Long, Y., Sun, W., Lu, Y., Yang, X., & Tang, J. (2017). Evaluating cities' vitality and identifying ghost cities in China with emerging geographical data. *Cities*, *63*, 98–109. https://doi.org/10.1016/j.cities.2017.01.002

JLL. (2022a). *JLL Global Real Estate Transparency Index 2022*. https://www.jll.de/content/dam/jll-com/documents/pdf/research/global/jll-global-real-estate-transparency-index-2022.pdf

JLL. (2022b, January). *Office Market Overview*. https://www.jll.de/en/trends-and-insights/research/office-market-overview

Just, T., Voigtländer, M., Eisfeld, R., Henger, R., Hesse, M., & Toschka, A. (2017). *Wirtschaftsfaktor Immobilien 2017* (Beiträge zur Immobilienwirtschaft Heft 19). IRE|BS International Real Estate Business School, Universität Regensburg.

Kallio, H., Pietilä, A.-M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, *72*(12), 2954–2965. https://doi.org/10.1111/jan.13031

Kay, A. I., Noland, R. B., & DiPetrillo, S. (2014). Residential property valuations near transit stations with transit-oriented development. *Journal of Transport Geography*, *39*, 131–140. https://doi.org/10.1016/j.jtrangeo.2014.06.017

Kelejian, H. H., & Piras, G. (2017). *Spatial econometrics*. Academic Press. https://doi.org/10.1016/C2016-0-04332-2

Kholodilin, K. A., & Mense, A. (2011). *Can Internet Ads Serve as an Indicator of Homeownership Rates?* (Discussion Papers No. 1168). DIW Berlin. https://dx.doi.org/10.2139/ssrn.1954628

Kholodilin, K. A., Mense, A., & Michelsen, C. (2017). The market value of energy efficiency in buildings and the mode of tenure. *Urban Studies*, *54*(14), 3218–3238. https://doi.org/10.1177/0042098016669464

Kim, T. K. (2017). Understanding one-way ANOVA using conceptual figures. *Korean Journal of Anesthesiology*, *70*(1), 22–26. https://doi.org/10.4097/kjae.2017.70.1.22

Kobayashi, M., & Takeda, K. (2000). Information Retrieval on the Web. *ACM Computing Surveys*, *32*(2), 144–173. https://doi.org/10.1145/358923.358934

Konomi, S., Sasao, T., Hosio, S., & Sezaki, K. (2017). Exploring the Use of Ambient WiFi Signals to Find Vacant Houses. *Proceedings of the 13th European Conference on Ambient Intelligence*, 130–135. https://doi.org/10.3233/AIS-180507

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM SIGKDD Explorations Newsletter*, *2*(1), 1–15. https://doi.org/10.1145/360402.360406

Krotov, V., & Silva, L. (2018). Legality and Ethics of Web Scraping. *Proceedings of the 24th Americas Conference on Information Systems*, 1–5.

Krotov, V., & Tennyson, M. (2018). Research Note: Scraping Financial Data from the Web Using the R Language. *Journal of Emerging Technologies in Accounting*, *15*(1), 169–181. https://doi.org/10.2308/jeta-52063

Kumagai, K., Matsuda, Y., & Ono, Y. (2016). Estimation of housing vacancy distributions: Basic bayesian approach using utility data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLI-B2*, 709–713. https://doi.org/10.5194/isprsarchives-XLI-B2-709-2016

Küpper, P., & Milbert, A. (2020). Typen ländlicher Räume in Deutschland. In C. Krajewski & C.-C. Wiegand (Eds.), *Land in Sicht: Ländliche Räume in Deutschland zwischen Prosperität und Marginalisierung* (pp. 82–97). Bundeszentrale für politische Bildung. https:// nbn-resolving.org/urn:nbn:de:0168-ssoar-71081-9

Lee, J., Newman, G., & Lee, C. (2022). Predicting Detached Housing Vacancy: A Multilevel Analysis. *Sustainability*, *14*(2), 922. https://doi.org/10.3390/su14020922

Lerbs, O., & Teske, M. (2017). Leerstände setzen Eigenheimpreise unter Druck. *Immobilien & Finanzierung*, *2017*(2), 58–60.

LeSage, J. P., & Pace, R. K. (2009). *Introduction to spatial econometrics*. CRC Press.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Li, J., Guo, M., & Lo, K. (2019). Estimating Housing Vacancy Rates in Rural China Using Power Consumption Data. *Sustainability*, *11*(20), 5722. https://doi.org/10.3390/su11205722

Li, L., & Wan, W. X. (2021). The Effect of Expected Losses on the Hong Kong Property Market. *The Journal of Real Estate Finance and Economics*. https://doi.org/10.1007/s11146-021-09851-3

Li, Y., Kubicki, S., Guerriero, A., & Rezgui, Y. (2019). Review of building energy performance certification schemes towards future improvement. *Renewable and Sustainable Energy Reviews*, *113*. https://doi.org/10.1016/j.rser.2019.109244

Lindqvist, S. (2012). The concept of transparency in the European Union's residential housing market: A theoretical framework. *International Journal of Law in the Built Environment*, *4*(2), 99–115. https://doi.org/10.1108/17561451211242486

Liu, H.-L., & Chang, H.-C. (2004). Variation of floor rent differentials of high-rise buildings. *Papers in Regional Science*, *83*, 495–503. https://doi.org/10.1007/s10110-004-0202-6

Lizieri, C. M. (2003). Occupier Requirements in Commercial Real Estate Markets. *Urban Studies*, *40*(5–6), 1151–1169. https://doi.org/10.1080/0042098032000074353

Locke, D. H., & Baine, G. (2015). The good, the bad, and the interested: How historical demographics explain present-day tree canopy, vacant lot and tree request spatial variability in New Haven, CT. *Urban Ecosystems*, *18*(2), 391–409. https://doi.org/10.1007/s11252-014-0409-5

Lucius, D. I. (2001). Real options in real estate development. *Journal of Property Investment & Finance*, *19*(1), 73–78. https://doi.org/10.1108/14635780110365370

Lyons, R. C. (2019). Can list prices accurately capture housing price trends? Insights from extreme markets conditions. *Finance Research Letters*, *30*, 228–232. https://doi.org/10.1016/j.frl.2018.10.004

Macias, P., & Stelmasiak, D. (2019). *Food inflation nowcasting with web scraped data* (NBP Working Paper No. 302). Narodowy Bank Polski. https://static.nbp.pl/publikacje/materialy-i-studia/302_en.pdf

Malik, S. K., & Rizvi, S. (2011). Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation. *Proceedings of the International Conference on*

Computational Intelligence and Communication Networks, 465–469. https://doi.org/10.1109/CICN.2011.97

Manville, M., & Kuhlmann, D. (2018). The Social and Fiscal Consequences of Urban Decline: Evidence from Large American Cities, 1980–2010. *Urban Affairs Review*, *54*(3), 451–489. https://doi.org/10.1177/1078087416675741

Massimino, B. (2016). Accessing Online Data: Web-Crawling and Information-Scraping Techniques to Automate the Assembly of Research Data. *Journal of Business Logistics*, *37*(1), 34–42. https://doi.org/10.1111/jbl.12120

McDonald, J. F. (2006). Market Value Websites: How Good Are They? *Journal of Real Estate Literature*, *14*(2), 223–230. https://doi.org/10.1080/10835547.2006.12090177

McGuinness, K. A. (2002). Of rowing boats, ocean liners and tests of the anova homogeneity of variance assumption. *Austral Ecology*, *27*(6), 681–688. https://doi.org/10.1046/j.1442-9993.2002.01233.x

McKenna, R., Merkel, E., Fehrenbach, D., Mehne, S., & Fichtner, W. (2013). Energy efficiency in the German residential sector: A bottom-up building-stock-model-based analysis in the context of energy-political targets. *Building and Environment*, *62*, 77–88. https://doi.org/10.1016/j.buildenv.2013.01.002

McMillen, D. P. (2008). Changes in the distribution of house prices over time: Structural characteristics, neighborhood, or coefficients? *Journal of Urban Economics*, *64*(3), 573–589. https://doi.org/10.1016/j.jue.2008.06.002

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation* (4th ed.). John Wiley & Sons.

Micheli, M., Rouwendal, J., & Dekkers, J. (2019). Border Effects in House Prices. *Real Estate Economics*, *47*(3), 757–783. https://doi.org/10.1111/1540-6229.12255

Milev, P. (2017). Conceptual Approach for Development of Web Scraping Application for Tracking Information. *Economic Alternatives*, *2017*(3), 475–485.

Miller, R. G., & Pinter, N. (2022). Flood risk and residential real-estate prices: Evidence from three US counties. *Journal of Flood Risk Management*, *15*(2). https://doi.org/10.1111/jfr3.12774

Mitchell, R. (2018). *Web Scraping with Python: Collecting More Data from the Modern Web* (2nd ed.). O'Reilly Media.

Molloy, R. (2016). Long-term vacant housing in the United States. *Regional Science and Urban Economics*, *59*, 118–129. https://doi.org/10.1016/j.regsciurbeco.2016.06.002

Monkkonen, P. (2019). Empty houses across North America: Housing finance and Mexico's vacancy crisis. *Urban Studies*, *56*(10), 2075–2091. https://doi.org/10.1177/0042098018788024

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). John Wiley & Sons.

Morckel, V. (2014). Spatial characteristics of housing abandonment. *Applied Geography*, *48*, 8–16. https://doi.org/10.1016/j.apgeog.2014.01.001

Morckel, V., & Durst, N. (2023). Using Emerging Hot Spot Analysis to Explore Spatiotemporal Patterns of Housing Vacancy in Ohio Metropolitan Statistical Areas. *Urban Affairs Review*, *59*(1), 309–328. https://doi.org/10.1177/10780874211065014

Nadalin, V., & Igliori, D. (2017). Empty spaces in the crowd. Residential vacancy in São Paulo's city centre. *Urban Studies*, *54*(13), 3085–3100. https://doi.org/10.1177/0042098016666498

Nam, J., Han, J., & Lee, C. (2016). Factors Contributing to Residential Vacancy and Some Approaches to Management in Gyeonggi Province, Korea. *Sustainability*, *8*(4). https://doi.org/10.3390/su8040367

Neumann, U., & Taruttis, L. (2022). Sorting in an urban housing market—Is there a response to demographic change? *Review of Regional Research*, *42*, 111–139. https://doi.org/10.1007/s10037-021-00158-7

Newell, G. (2016). The changing real estate market transparency in the European real estate markets. *Journal of Property Investment & Finance*, *34*(4), 407–420. https://doi.org/10.1108/JPIF-07-2015-0053

Newman, G., Gu, D., Kim, J.-H., Bowman, A. O. M., & Li, W. (2016). Elasticity and urban vacancy: A longitudinal comparison of U.S. cities. *Cities*, *58*, 143–151. https://doi.org/10.1016/j.cities.2016.05.018

Newman, G., Lee, R. J., Gu, D., Park, Y., Saginor, J., Van Zandt, S., & Li, W. (2019). Evaluating drivers of housing vacancy: A longitudinal analysis of large U.S. cities from 1960 to 2010. *Journal of Housing and the Built Environment*, *34*(3), 807–827. https://doi.org/10.1007/s10901-019-09684-w

Nikiforou, P., Dimopoulos, T., & Sivitanides, P. (2022). Identifying how the time on the market affects the selling price: A case study of residential properties in Paphos (Cyprus) urban area. *Journal of European Real Estate Research*, *15*(3), 368–386. https://doi.org/10.1108/JERER-11-2021-0051

Office of Policy Development and Research. (n.d.). *HUD Aggregated USPS Administrative Data On Address Vacancies*. Retrieved March 21, 2022, from https://www.huduser.gov/portal/datasets/usps.html

Ostertagová, E., Ostertag, O., & Kováč, J. (2014). Methodology and Application of the Kruskal-Wallis Test. *Applied Mechanics and Materials*, *611*, 115–120. https://doi.org/10.4028/www.scientific.net/AMM.611.115

Pan, J., & Dong, L. (2021). Spatial Identification of Housing Vacancy in China. *Chinese Geographical Science*, *31*(2), 359–375. https://doi.org/10.1007/s11769-020-1171-7

Pan, Y., Zeng, W., Guan, Q., Yao, Y., Liang, X., Yue, H., Zhai, Y., & Wang, J. (2020). Spatiotemporal dynamics and the contributing factors of residential vacancy at a fine scale: A perspective from municipal water consumption. *Cities*, *103*, 102745. https://doi.org/10.1016/j.cities.2020.102745

Pan, Y., Zeng, W., Guan, Q., Yao, Y., Liang, X., Zhai, Y., & Pu, S. (2021). Variability in and mixtures among residential vacancies at granular levels: Evidence from municipal water

consumption data. *Computers, Environment and Urban Systems*, *90*. https://doi.org/10.1016/j.compenvurbsys.2021.101702

Pavlov, A., Wachter, S., & Zevelev, A. A. (2016). Transparency in the Mortgage Market. *Journal of Financial Services Research*, *49*, 265–280. https://doi.org/10.1007/s10693-014-0211-9

Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. *Machine Vision and Applications*, *29*, 667–676. https://doi.org/10.1007/s00138-018-0922-2

Püschel, R., & Evangelinos, C. (2012). Evaluating noise annoyance cost recovery at Düsseldorf International Airport. *Transportation Research Part D: Transport and Environment*, *17*(8), 598–604. https://doi.org/10.1016/j.trd.2012.07.002

Rat für Sozial- und Wirtschaftsdaten. (2019). *Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement. Mit Gutachten "Web Scraping in der unabhängigen wissenschaftlichen Forschung"* (Output 4, 6. Berufungsperiode). https://doi.org/10.17620/02671.39

Rink, D. (2021). *Stadtentwicklung, Wohnungsmarkt und Wohnungspolitik in Leipzig* (UFZ Discussion Papers No. 06/2021). UFZ Helmholtz-Zentrum für Umweltforschung. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-75773-6

Rink, D., & Wolff, M. (2015). Wohnungsleerstand in Deutschland. Zur Konzeptualisierung der Leerstandsquote als Schlüsselindikator der Wohnungsmarktbeobachtung anhand der GWZ 2011. *Raumforschung und Raumordnung*, *73*(5), 311–325. https://doi.org/10.1007/s13147-015-0361-8

Rodgers, J. L., & Nicewander, A. W. (1988). Thirteen Ways to Look at the Correlation Coefficient. *American Statistician*, *42*(1), 59–66. https://doi.org/10.2307/2685263

Rosen, K. T., & Smith, L. B. (1983). The price-adjustment process for rental housing and the natural vacancy rate. *The American Economic Review*, *73*(4), 779–786.

RWI. (n.d.). *RWI-GEO-RED/X: Real Estate Data and Price Indices*. Retrieved October 10, 2022, from https://www.rwi-essen.de/forschung-

beratung/weitere/forschungsdatenzentrum-ruhr/datenangebot/rwi-geo-red-real-estate-data

Sadayuki, T., Harano, K., & Yamazaki, F. (2019). Market transparency and international real estate investment. *Journal of Property Investment & Finance*, *37*(5), 503–518. https://doi.org/10.1108/JPIF-04-2019-0043

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE*, *16*(8). https://doi.org/10.1371/journal.pone.0254937

Schaffner, S. (2020). *FDZ data description: Real-estate data for Germany Campus Files (RWI-GEO-RED city and RWI-GEO-RED cross)—Advertisments on the internet platform ImmobilienScout24 for teaching purposes* [RWI Projektberichte]. RWI - Leibniz-Institut für Wirtschaftsforschung. http://hdl.handle.net/10419/242995

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Schulte, K., Rottke, N., & Pitschke, C. (2005). Transparency in the German real estate market. *Journal of Property Investment & Finance*, *23*(1), 90–108. https://doi.org/10.1108/14635780510575111

Shekhar, S., Bhagat, S., Sivakumar, K., & Koteshwar, B. (2023, April 11). *Dominance-analysis 1.1.9*. Dominance-Analysis : A Python Library for Accurate and Intuitive Relative Importance of Predictors. https://pypi.org/project/dominance-analysis/

Silva, C., & Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. *Proceedings of the International Joint Conference on Neural Networks*, 1661–1666. https://doi.org/10.1109/IJCNN.2003.1223656

Spehl, H. (2011). *Leerstand von Wohngebäuden in ländlichen Räumen: Beispiele ausgewählter Gemeinden der Länder Hessen, Rheinland-Pfalz und Saarland* (E-Paper der ARL No. 12). ARL - Akademie für Raumforschung und Landesplanung. https://nbn-resolving.org/urn:nbn:de:0156-73022

Stamsø, M. A. (2015). Selling a house and the decision to use a real estate broker in Norway. *Property Management*, *33*(2), 173–186. https://doi.org/10.1108/PM-01-2014-0006

Statistische Ämter des Bundes und der Länder. (n.d.-a). *Bestand an Wohngebäuden und Wohnungen in Wohn- und Nichtwohngebäuden—Stichtag 31.12. - Regionale Tiefe: Kreise und krfr. Städte* [Data set]. Genesis-Online. Retrieved March 10, 2023, from https://www.regionalstatistik.de/genesis/online?operation=result&code=31231-02-01-4&zeiten=2019

Statistische Ämter des Bundes und der Länder. (n.d.-b). *Bevölkerung nach Geschlecht—Stichtag 31.12. - Regionale—Tiefe: Kreise und krfr. Städte* [Data set]. Genesis-Online. Retrieved March 10, 2023, from https://www.regionalstatistik.de/genesis/online?operation=result&code=12411-01-01-4&zeiten=2018,2019

Statistisches Amt Dortmund. (1904). *Die Zählung der leerstehenden und im Bau befindlichen Wohnungen vom 1. Dezember 1903* (Viertes Heft Nachtrag 1).

Statistisches Bundesamt. (1956). *Gebäude- und Wohungszählung in der Bundesrepublik Deutschland vom 13. September 1950 Einführung in die Methoden und die Organisation der Zählung* (Band 38 Heft 1).

Statistisches Bundesamt. (1975). *Gebäude- und Wohnungszählung vom 25. Oktober 1968* (Heft 1). W. Kohlhammer.

Statistisches Bundesamt. (2020). *Alle politisch selbständigen Gemeinden (mit Gemeindeverband) in Deutschland nach Fläche, Bevölkerung, Bevölkerungsdichte und der Postleitzahl des Verwaltungssitzes der Gemeinde. Ergänzt um die geografischen Mittelpunktkoordinaten, Reisegebiete und Grad der Verstädterung. Gebietsstand: 31.12.2019* [Data set]. Gemeindeverzeichnis-Informationssystem GV-ISys.

Taruttis, L., & Weber, C. (2022). Estimating the impact of energy efficiency on housing prices in Germany: Does regional disparity matter? *Energy Economics*, *105*. https://doi.org/10.1016/j.eneco.2021.105750

Thießen, F. (2013). *Fluglärmbelastung und Immobilien Ergebnisse neuer Studien für Deutschland* (WWDP : Diskussionspapiere der Fakultät für Wirtschaftswissenschaften der Technischen Universität Chemnitz No. 110). https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa-123122

Thomschke, L. (2015). Changes in the distribution of rental prices in Berlin. *Regional Science and Urban Economics*, *51*, 88–100. https://doi.org/10.1016/j.regsciurbeco.2015.01.001

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234–240. https://doi.org/10.2307/143141

Tormanski, A. (2012). Immobilienrichtwerte – Beitrag zur Markttransparenz, Wertermittlung und Besteuerung? *zfv – Zeitschrift für Geodäsie, Geoinformation und Landmanagement*, *137*(3), 149–155.

Torrieri, N., ACSO, DSSD, & SEHSD Program Staff. (2014). American Community Survey Design and Methodology. *U.S. Census Bureau*.

Triantafyllopoulos, N. (2006). Public sector transparency and tourism real estate investments in Greece. *Regional and Sectoral Economic Studies*, *6*(2), 57–72.

U.S. Census Bureau. (2020). *U.S. Decennial Census 2020—Occupancy Status* [Data set]. https://data.census.gov/cedsci/table?q=United%20States&y=2020&d=DEC%20Redistricting%20Data%20%28PL%2094-171%29&tid=DECENNIALPL2020.H1

U.S. Census Bureau. (2022). *Housing Vacancies and Homeownership (CPS/HVS): 2022* [Data set]. https://www.census.gov/housing/hvs/data/q122ind.html

U.S. Census Bureau. (2011, September 12). *Vacancy Rate Fact Sheet*. https://www.census.gov/topics/housing/guidance/vacancy-fact-sheet.html

U.S. Department of Commerce Bureau of the Census. (1940). *Statistical Abstract of the United States 1939* (Sixty-First Number).

Uzun, E. (2020). A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages. *IEEE Access*, *8*, 61726–61740. https://doi.org/10.1109/ACCESS.2020.2984503

van den Heuvel, E., & Zhan, Z. (2022). Myths About Linear and Monotonic Associations: Pearson's r, Spearman's ρ, and Kendall's τ. *The American Statistician*, *76*(1), 44–52. https://doi.org/10.1080/00031305.2021.2004922

vanden Broucke, S., & Baesens, B. (2018). From Web Scraping to Web Crawling. In *Practical Web Scraping for Data Science*. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-3582-9_6

Volkswagen AG. (2022). *Geschäftsbericht 2021*. https://www.volkswagenag.com/presence/investorrelation/publications/annual-reports/2022/volkswagen/Y_2021_d.pdf

Wang, B. (2021). How Does COVID-19 Affect House Prices? A Cross-City Analysis. *Journal of Risk and Financial Management*, *14*(47). https://doi.org/10.3390/ jrfm14020047

Wang, K., & Immergluck, D. (2019). Housing vacancy and urban growth: Explaining changes in long-term vacancy after the US foreclosure crisis. *Journal of Housing and the Built Environment*, *34*, 511–532. https://doi.org/10.1007/s10901-018-9636-z

Wang, L., Fan, H., & Wang, Y. (2019). An estimation of housing vacancy rate using NPP-VIIRS night-time light data and OpenStreetMap data. *International Journal of Remote Sensing*, *40*(22), 8566–8588. https://doi.org/10.1080/01431161.2019.1615655

Weinberg, D. H. (2006). How the United States Measures Well-being in Household Surveys. *Journal of Official Statistics*, *22*(1), 113–136.

Whitaker, S., & Fitzpatrick IV, T. J. (2013). Deconstructing distressed-property spillovers: The effects of vacant, tax-delinquent, and foreclosed properties in housing submarkets. *Journal of Housing Economics*, *22*(2), 79–91. https://doi.org/10.1016/j.jhe.2013.04.001

Whiting, L. S. (2008). Semi-structured interviews: Guidance for novice researchers. *Nursing Standard*, *22*(23), 35–40. https://doi.org/10.7748/ns2008.02.22.23.35.c6420

Winke, T. (2017). The impact of aircraft noise on apartment prices: A differences-in-differences hedonic approach for Frankfurt, Germany. *Journal of Economic Geography*, *17*(6), 1283–1300. https://doi.org/10.1093/jeg/lbw040

Wood, G., Yates, J., & Reynolds, M. (2006). Vacancy rates and low-rent housing: A panel data analysis. *Journal of Housing and the Built Environment*, *21*, 441–458. https://doi.org/10.1007/s10901-006-9059-0

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach* (7th ed.). Cengage Learning.

Yue, X., Wang, Y., & Zhang, H. (2022). Influences of the Plot Area and Floor Area Ratio of Residential Quarters on the Housing Vacancy Rate: A Case Study of the Guangzhou Metropolitan Area in China. *Buildings*, *12*(8), 1197. https://doi.org/10.3390/buildings12081197

Zhang, C., Jia, S., & Yang, R. (2016). Housing affordability and housing vacancy in China: The role of income inequality. *Journal of Housing Economics*, *33*, 4–14. https://doi.org/10.1016/j.jhe.2016.05.005

Zheng, Q., Zeng, Y., Deng, J., Wang, K., Jiang, R., & Ye, Z. (2017). "Ghost cities" identification using multi-source remote sensing datasets: A case study in Yangtze River Delta. *Applied Geography*, *80*, 112–121. https://doi.org/10.1016/j.apgeog.2017.02.004

Zhou, Y., Xue, L., Shi, Z., Wu, L., & Fan, J. (2021). Measuring Housing Activeness from Multi-Source Big Data and Machine Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3940180

Zillow. (n.d.). Zillow's Assessor and Real Estate Database. *Zillow Research*. Retrieved August 24, 2022, from https://www.zillow.com/research/ztrax/

Appendix

Appendix A: Data

Appendix A - 1

Distribution of Title Length Data

Lower bound	Upper bound	Observations
0	5	13,895
6	10	8,763
11	15	41,744
16	20	119,951
21	25	261,621
26	30	338,133
31	35	431,675
36	40	569,670
41	45	602,511
46	50	628,701
51	55	625,381
56	60	589,819
61	65	539,738
66	70	473,698

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 2

Most Frequent ZIP Codes Extended

ZIP Code	Municipality	Observations
09130	Chemnitz	59,112
09126	Chemnitz	56,402
09131	Chemnitz	46,877
09112	Chemnitz	46,370
09113	Chemnitz	43,675
08056	Zwickau	32,620
06217	Merseburg	31,026
39112	Magdeburg	27,587
04157	Leipzig	25,396
08523	Plauen	25,063
09599	Freiberg	24,354
08058	Zwickau	23,091
39108	Magdeburg	22,416
06124	Halle	22,164
01067	Dresden	21,344
38820	Halberstadt	21,266
02826	Görlitz	21,148
04229	Leipzig	20,992
09111	Chemnitz	20,956
08525	Plauen	20,187
04159	Leipzig	19,968
09116	Chemnitz	19,801
04103	Leipzig	19,725
08412	Werdau	19,362
07545	Gera	19,338
04315	Leipzig	19,037
06132	Halle	18,700
09119	Chemnitz	17,949
39104	Magdeburg	17,695
01662	Meißen	17,325

ZIP Code	Municipality	Observations
06110	Halle	17,291
01159	Dresden	16,887
06712	Zeitz	16,739
02625	Bautzen	16,715
04109	Leipzig	16,422
09456	Annaberg-Buchholz	16,421
04720	Döbeln	16,200
39576	Stendal	15,948
04177	Leipzig	15,609
04105	Leipzig	15,010
04205	Leipzig	14,338
39124	Magdeburg	14,253
01307	Dresden	14,231
06844	Dessau	13,794
07549	Gera	13,653
04209	Leipzig	13,585
01796	Pirna	13,461
01705	Freital	13,184
04277	Leipzig	13,087
04317	Leipzig	13,079
06108	Halle	12,959
39114	Magdeburg	12,853
99817	Eisenach	12,594
04179	Leipzig	12,552
09127	Chemnitz	12,545
04275	Leipzig	12,459
09648	Mittweida	12,457
04207	Leipzig	12,445
06667	Weißenfels	12,419
04600	Altenburg	12,336
04347	Leipzig	12,297
08527	Plauen	12,228
99867	Gotha	12,216

ZIP Code	Municipality	Observations
04155	Leipzig	12,155
06618	Naumburg	12,018
04318	Leipzig	11,850
06122	Halle	11,845
01309	Dresden	11,589
01169	Dresden	11,414
01099	Dresden	11,219
06128	Halle Saale	11,195
06295	Lutherstadt Eisleben	11,070
99974	Mühlhausen	10,913
45899	Gelsenkirchen	10,835
45881	Gelsenkirchen	10,795
01589	Riesa	10,673
08371	Glauchau	10,591
09120	Chemnitz	10,517
07546	Gera	10,290
04299	Leipzig	10,223
45896	Gelsenkirchen	10,041
04552	Borna	9,895
47798	Krefeld	9,823
08060	Zwickau	9,652
06842	Dessau	9,468
08529	Plauen	9,453
06886	Lutherstadt Wittenberg	9,422
02763	Zittau	9,306
01187	Dresden	9,222
39387	Oschersleben	9,218
58089	Hagen	9,071
10557	Berlin	9,019
07548	Gera	9,008
27568	Bremerhaven	8,957
17291	Prenzlau	8,924
06749	Bitterfeld	8,869

ZIP Code	Municipality	Observations
01097	Dresden	8,842
09212	Limbach-Oberfrohna	8,829
01277	Dresden	8,734
02977	Hoyerswerda	8,645

Note. Own research. Created from raw data set.

Appendix A - 3

Most Frequent Municipality Designations Extended

Designation of City raw	Observations	Designations of Cities cleaned and aggregated	Observations
Chemnitz, Sonnenberg	46,450	Chemnitz	363,162
Chemnitz, Kaßberg	40,623	Leipzig	350,200
Chemnitz	34,909	Berlin	215,568
Chemnitz, Schloßchemnitz	33,429	Dresden	194,116
Leipzig	31,898	Magdeburg	141,991
Chemnitz, Hilbersdorf	31,382	Frankfurt am Main	113,320
Dresden	30,809	München	113,275
Berlin	29,852	Essen	109,537
Merseburg, Saalekreis	28,178	Düsseldorf	88,271
Freiberg, Mittelsachsen (Kreis)	21,302	Hamburg	86,979
Chemnitz, Gablenz	20,600	Duisburg	86,355
Magdeburg, Stadtfeld Ost	20,070	Zwickau	79,759
München	19,957	Dortmund	74,426
Chemnitz, Lutherviertel	19,654	Plauen	67,318
Chemnitz, Bernsdorf	19,077	Köln	64,257
Halberstadt, Harz (Kreis)	18,857	Gera	60,044
Magdeburg, Sudenburg	18,521	Wuppertal	56,679
Dortmund, Innenstadt	18,257	Bremen	48,806
Chemnitz, Zentrum	17,715	Stuttgart	48,266
Gera, Stadtmitte	17,403	Nürnberg	44,308
Riesa, Meißen (Kreis)	16,721	Wiesbaden	40,605
Hamburg	16,320	Görlitz	32,814
Berlin, Tiergarten (Tiergarten)	15,698	Merseburg	30,976
Düsseldorf	15,557	Freiberg	24,499
Leipzig, Möckern	14,596	Halberstadt	21,244
Zwickau	14,490	Riesa	20,767
Zeitz, Burgenlandkreis	14,448	Werdau	19,247
Werdau, Zwickau (Kreis)	13,830	Meißen	17,359
Essen	13,778	Zeitz	16,772

Designation of City raw	Observations	Designations of Cities cleaned and aggregated	Observations
Berlin, Mitte (Mitte)	13,761	Bautzen	16,753
Chemnitz, Kappel	13,168	Kaiserslautern	16,246
Magdeburg, Neue Neustadt	13,064	Stendal	15,898
Meißen, Meißen (Kreis)	13,021	Döbeln	15,264
Frankfurt am Main	12,644	Saarbrücken	15,228
Döbeln, Mittelsachsen (Kreis)	12,579	Göttingen	14,766
Chemnitz, Altendorf	12,540	Recklinghausen	14,360
Göttingen, Göttingen (Kreis)	12,292	Gladbeck	13,217
Zwickau, Mitte-Nord	12,250	Freital	13,179
Kaiserslautern, Innenstadt	12,224	Pirna	13,026
Stendal, Stendal (Kreis)	12,015	Eisenach	12,611
Köln	11,983	Weißenfels	12,423
Bautzen, Bautzen (Kreis)	11,965	Altenburg	12,268
Recklinghausen, Recklinghausen (Kreis)	11,812	Gotha	12,207
Pirna, Sächsische Schweiz-Osterzgebirge (Kreis)	11,475	Mittweida	11,816
Stuttgart	11,404	Marl	11,729
Gladbeck, Recklinghausen (Kreis)	11,385	Gießen	11,542
Nürnberg	11,154	Lutherstadt Eisleben	11,062
Leipzig, Lausen-Grünau	11,058	Witten	10,949
Berlin, Charlottenburg (Charlottenburg)	10,991	Minden	10,671
Magdeburg	10,983	Glauchau	10,627

Note. Own research. Created from raw data set.

Appendix A - 4

Distribution of Cold Rent

Lower bound	Upper bound	Observations
0	100	6,825
100	200	189,938
200	300	1127,022
300	400	1482,221
400	500	946,327
500	600	634,725
600	700	523,064
700	800	427,813
800	900	330,059
900	1,000	290,943
1,000	1,100	167,302
1,100	1,200	178,162
1,200	1,300	147,983
1,300	1,400	111,706
1,400	1,500	86,715
1,500	1,600	65,305
1,600	1,700	53,039
1,700	1,800	37,617
1,800	1,900	31,775
1,900	2,000	33,648
2,000	2,100	13,288
2,100	2,200	14,753
2,200	2,300	15,257
2,300	2,400	11,113
2,400	2,500	10,890
2,500	2,600	10,277
2,600	2,700	5,997
2,700	2,800	5,822
2,800	2,900	5,152
2,900	3,000	7,484

Lower bound	Upper bound	Observations
3,000	3,100	2,551
3,100	1,000,000,000	32,798

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 5

Distribution of Utilities

Lower bound	Upper bound	Observations
0	25	120,812
25	50	176,516
50	75	745,098
75	100	867,293
100	125	1,188,707
125	150	860,499
150	175	927,426
175	200	486,499
200	225	518,416
225	250	201,477
250	275	271,982
275	300	106,393
300	325	130,011
325	350	42,566
350	375	66,386
375	400	25,893
400	425	31,532
425	450	8,758
450	475	16,946
475	500	6,569
500	525	12,496
525	550	3,314
550	575	4,600
575	600	2,744
600	625	4,728
625	10,000,000	15,197

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 6

Distribution of Heating Costs

Lower bound	Upper bound	Observations
0	10	78,481
10	20	3,593
20	30	32,044
30	40	107,673
40	50	184,300
50	60	298,014
60	70	356,197
70	80	337,551
80	90	271,502
90	100	183,458
100	110	160,189
110	120	68,997
120	130	62,218
130	140	32,307
140	150	22,182
150	160	27,781
160	170	12,147
170	180	8,828
180	190	6,562
190	200	3,956
200	210	7,630
210	220	2,074
220	230	1,644
230	240	1,420
240	250	726
250	260	2,764
260	270	504
270	280	723
280	290	475
290	300	488

Lower bound	Upper bound	Observations
300	310	924
310	10,000,000	3,323

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 7

Distribution of Total Rent

Lower bound	Upper bound	Observations
0	100	13,104
100	200	13,155
200	300	237,026
300	400	775,934
400	500	1,248,741
500	600	996,103
600	700	718,142
700	800	556,116
800	900	455,276
900	1,000	386,520
1,000	1,100	263,579
1,100	1,200	238,557
1,200	1,300	192,668
1,300	1,400	159,607
1,400	1,500	138,587
1,500	1,600	113,620
1,600	1,700	91,055
1,700	1,800	71,856
1,800	1,900	56,934
1,900	2,000	47,964
2,000	2,100	31,242
2,100	2,200	28,839
2,200	2,300	25,525
2,300	2,400	18,892
2,400	2,500	16,502
2,500	2,600	14,758
2,600	2,700	12,193
2,700	2,800	9,201
2,800	2,900	8,997
2,900	3,000	8,574

Lower bound	Upper bound	Observations
3,000	3,100	5,465
3,100	3,200	4,527
3,200	3,300	4,802
3,300	3,400	4,878
3,400	3,500	3,334
3,500	3,600	3,644
3,600	1,000,000,000	31,654

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 8

Distribution of Number of Rooms

Lower bound	Upper bound	Observations
0.0	0.5	0
0.5	1.0	0
1.0	1.5	797,279
1.5	2.0	88,713
2.0	2.5	2,339,401
2.5	3.0	235,397
3.0	3.5	2,386,486
3.5	4.0	213,344
4.0	4.5	698,250
4.5	5.0	49,396
5.0	5.5	140,216
5.5	6.0	9,332
6.0	6.5	32,509
6.5	7.0	1,434
7.0	7.5	8,727
7.5	8.0	466
8.0	8.5	2,834
8.5	9.0	98
9.0	9.5	910
9.5	10.0	11
10.0	10.5	825
10.5	1000.0	1,943

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 9

Distribution of Living Space

Lower bound	Upper bound	Observations
0	10	3,878
10	20	57,581
20	30	222,272
30	40	418,130
40	50	703,584
50	60	1,195,911
60	70	1,296,540
70	80	943,405
80	90	663,607
90	100	440,117
100	110	318,436
110	120	211,356
120	130	164,969
130	140	108,419
140	150	74,243
150	160	51,891
160	170	36,364
170	180	24,132
180	190	17,829
190	200	10,847
200	210	11,304
210	220	6,023
220	230	5,346
230	240	4,248
240	250	3,365
250	260	3,155
260	1000	10,619

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 10

Distribution of Floor

Floor	Observations
-1	9,632
0	620,924
1	1,641,063
2	1,469,450
3	1,008,709
4	551,465
5	225,441
6	66,480
7	28,752
8	15,272
9	11,602
10 - 999	31,999

Note. Own research. Created from raw data set.

Appendix A - 11

Distribution of Year of Construction

Lower bound	Upper bound	Observations
1000	1800	22,609
1800	1810	4,510
1810	1820	932
1820	1830	1,454
1830	1840	2,472
1840	1850	2,457
1850	1860	8,344
1860	1870	7,663
1870	1880	22,886
1880	1890	35,339
1890	1900	97,379
1900	1910	478,430
1910	1920	230,523
1920	1930	208,646
1930	1940	199,891
1940	1950	48,352
1950	1960	464,584
1960	1970	664,268
1970	1980	674,724
1980	1990	525,071
1990	2000	686,506
2000	2010	186,025
2010	2020	832,311
2020	2030	110,474
2030	6580	116

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 12

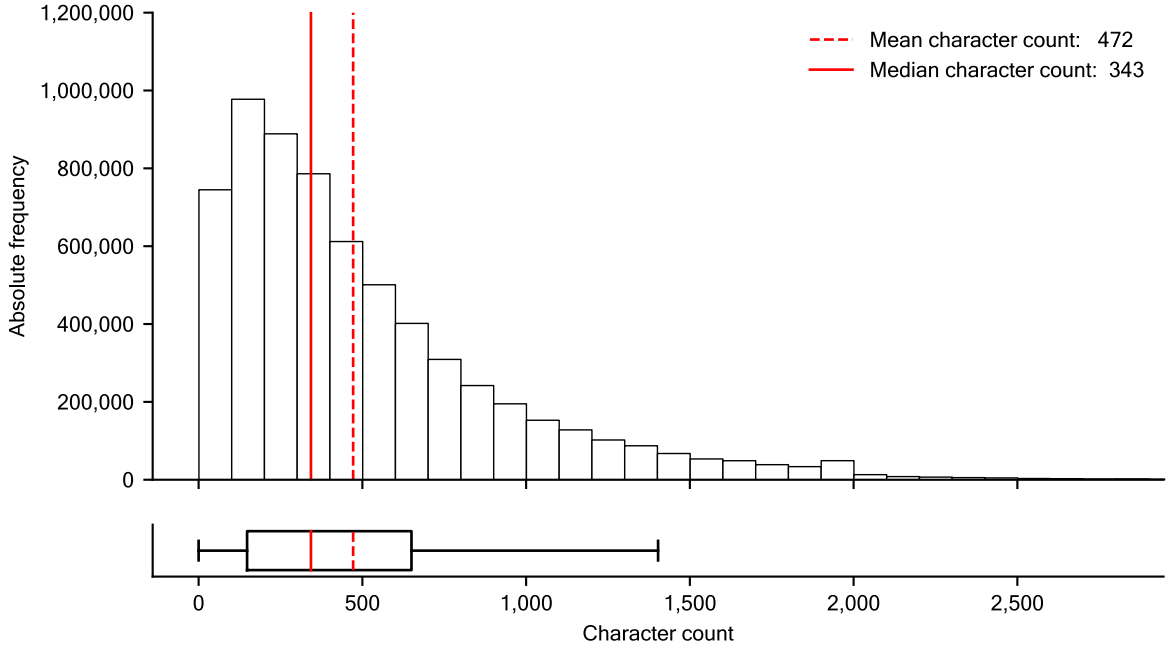
Distribution of Energy Demand

Lower bound	Upper bound	Observations
0	20	47,979
20	40	119,474
40	60	211,226
60	80	240,629
80	100	174,523
100	120	151,136
120	140	125,064
140	160	105,717
160	180	81,332
180	200	60,856
200	220	41,750
220	240	31,187
240	260	22,269
260	280	13,522
280	300	9,970
300	320	6,252
320	340	3,474
340	360	2,383
360	380	2,395
380	400	1,297
400	420	969
420	440	694
440	460	365
460	480	243
480	500	301
500	520	167
520	3,020	1,925

Note. Own research. Created from raw data set.

Appendix A - 13

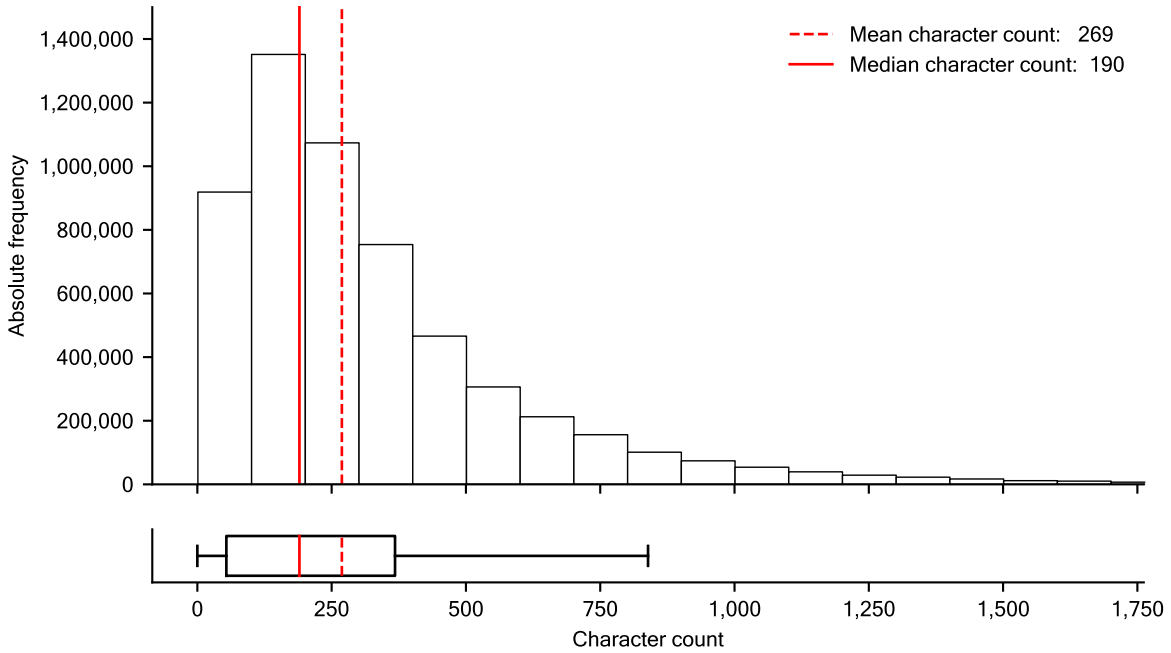
General Textual Description Data



Note. Own research. Created from raw data set.

Appendix A - 14

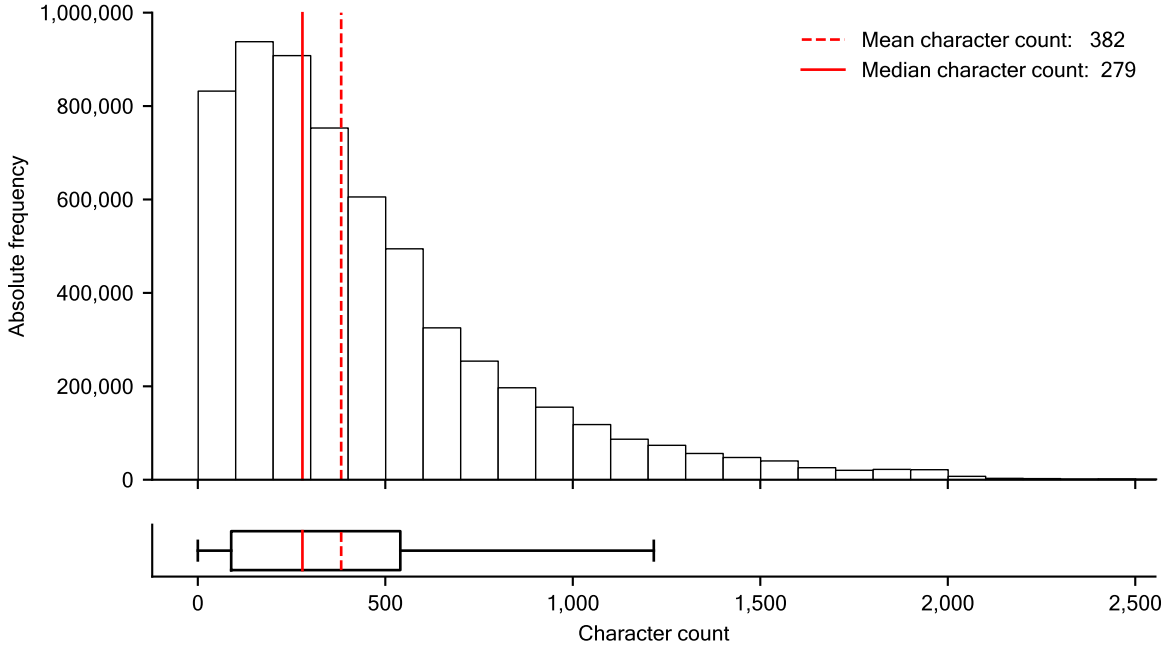
Textual Description of Facilities Data



Note. Own research. Created from raw data set.

Appendix A - 15

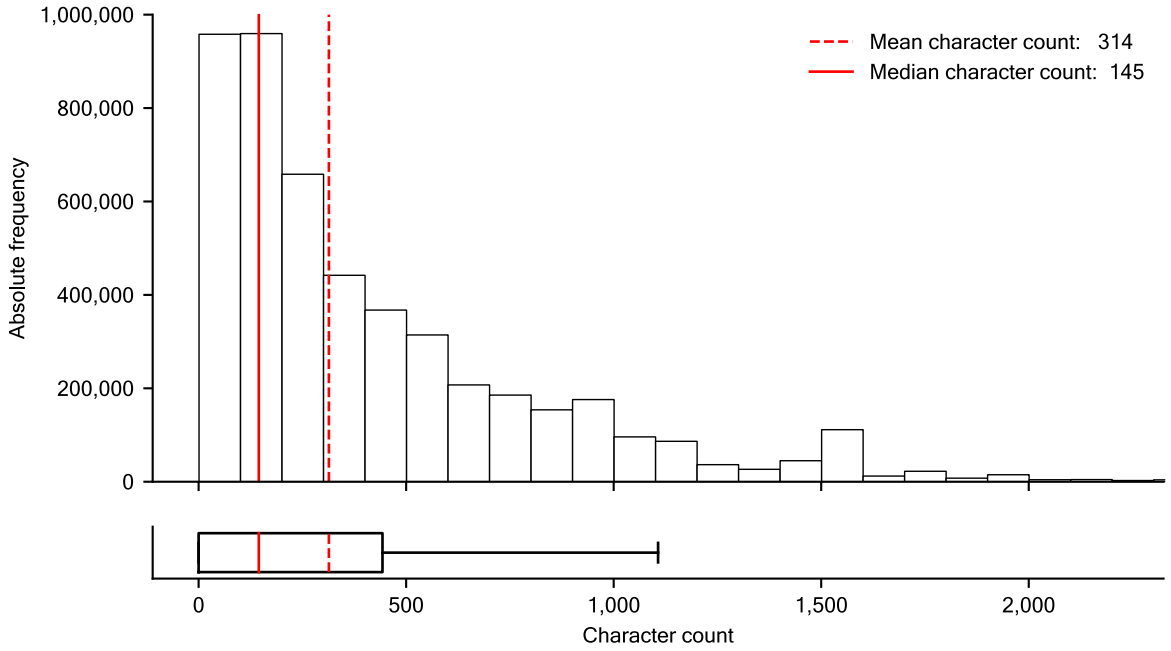
Textual Description of Location Data



Note. Own research. Created from raw data set.

Appendix A - 16

Textual Description of Micellaneous Data



Note. Own research. Created from raw data set.

Appendix A - 17

Textual Description of Miscellaneous Data

General Textual Description	Textual Description of Facilities	Textual Description of Location	Textual Description of Miscellaneous
Wohnung (apartment)	Wohnung (apartment)	befindet (is located)	Angaben (details)
befindet (located)	Bad (bathroom)	sowie (as well as)	Wohnung (apartment)
sowie (as well as)	Küche (kitchen)	Minuten (minutes)	Exposé (exposé)
verfügt (has)	Balkon (balcony)	ca. (approx)	Uhr (clock)
Küche (kitchen)	Fenster (window)	liegt (is located)	Informationen (information)
liegt (is located)	Dusche (shower)	Nähe (near)	vorbehalten (reserved)
Bad (bathroom)	Einbauküche (fitted kitchen)	Einkaufsmöglich-keiten (shopping opportunities)	Bitte (Request)
ca. (ca.)	sowie (as well as)	unmittelbarer (directly)	Besichtigungstermin (viewing appointment)
Balkon (balcony)	Badezimmer (bathroom)	befinden (are located)	Haftung (liability)
Haus (house)	ausgestattet (equipped)	erreichen (reach)	Richtigkeit (correctness)
Objekt (object)	verfügt (disposes)	täglichen (daily)	finden (find)
Zimmer (room)	Wohnzimmer (living room)	fußläufig (accessible by foot)	ausschließlich (exclusively)
bietet (provides)	Badewanne (bathtub)	erreichbar (available)	Daten (data)
Wohnungen (apartments)	Laminat (laminate)	Stadtteil (district)	freuen (rejoice)

General Textual Description	Textual Description of Facilities	Textual Description of Location	Textual Description of Miscellaneous
Wohnzimmer (livingroom)	Schlafzimmer (bedroom)	gute (good)	Gewähr (guarantee)
zwei (two)	Keller (basement)	wenigen (few)	daher (therefore)
Badezimmer (bathroom)	Wanne (bathtub)	Lage (location)	sowie (as well as)
Verfügung (disposal)	Flur (corridor)	Wohnung (apartment)	bitte (please)
Schlafzimmer (bedroom)	Fliesen (tiles)	Stadt (city)	Mietvertrag (rental agreement)
Einbauküche (fitted kitchen)	befindet (located)	Anbindung (connection)	bleiben (remain)
m^2 (square meters)	Fußbodenheizung (underfloor heating)	Innenstadt (city center)	möglich (possible)
ausgestattet (equipped)	Tageslichtbad (daylight bathroom)	entfernt (removed)	gerne (gladly)
Mehrfamilienhaus (multifamily house)	Zimmer (room)	bietet (provides)	ausgeschlossen (excluded)
stehen (are available)	Abstellraum (storeroom)	direkt (directly)	stehen (are available)
gepflegten (groomed)	weiß (white)	Fuß (foot)	Verfügung (available)
große (large)	WC (wc)	km (kilometer)	Weitere (more)
gehört (belongs)	ca. (approx)	Schulen (schools)	Interesse (interest)
steht (stands)	Wände (walls)	gut (good)	per (per)
Lage (location)	große (large)	öffentlichen (public)	erfolgen (occur)
handelt (acts)	Räume (rooms)	Gehminuten (walking minutes)	Objekt (object)

General Textual Description	Textual Description of Facilities	Textual Description of Location	Textual Description of Miscellaneous
direkt (direct)	Kellerraum (basement)	Geschäfte (stores)	€ (€)
ab (as of)	komplett (complete)	Bedarfs (demand)	vollständig (complete)
befinden (located)	bietet (provides)	finden (find)	rund (around)
großen (large)	moderne (modern)	Objekt (object)	geschlossenen (closed)
helle (bright)	helle (bright)	Zentrum (center)	Vereinbarungen (agreements)
Dusche (shower)	gefliest (tiled)	Umgebung (environment)	übernehmen (take over)
Obergeschoss (upper floor)	Laminatboden (laminate floor)	schnell (fast)	Mietpreisänderungen (rental price changes)
Fenster (window)	neue (new)	ebenfalls (also)	Angebot (offer)
schöne (nice)	Waschmaschinen-anschluss (washing machine connection)	wenige (few)	erhalten (receive)
Kellerraum (basement room)	m² (square meters)	Verkehrsanbindung (traffic connection)	Nebenkosten (service charges)
Flur (hallway)	Wohnräumen (living spaces)	Infrastruktur (infrastructure)	Besichtigung (visit)
vorhanden (available)	Zugang (access)	Bahnhof (train station)	Mieter (renter)
komplett (complete)	gehört (belongs)	Ärzte (doctors)	Energieausweis (energy certificate)
Platz (space)	Haus (house)	Kindergärten (kindergardens)	Immobilien (propertys)
Zugang (access)	Wohnungen (apartment)	gibt (there are)	personenbezogenen (person-related)

General Textual Description	Textual Description of Facilities	Textual Description of Location	Textual Description of Miscellaneous
Mieter (renter)	neu (new)	Einrichtungen (facilities)	weitere (more)
Stellplatz (parking space)	hochwertige (high quality)	Auto (car)	Angebote (offers)
neue (new)	vorhanden (available)	Autobahn (highway)	Bad (bathroom)
€ (€)	Aufzug (elevator)	Stadtzentrum (citycenter)	Ansprechpartner (contact)
moderne (modern)	Platz (space)	ruhigen (quiet)	führen (lead)
Blick (view)	Waschmaschine (washing machine)	zahlreiche (numerious)	bzw. (resp.)
finden (find)	großer (large)	öffentliche (public)	Maßgeblich (significantly)
außerdem (also)	Ausstattung (equipment)	Verkehrsmittel (transportation)	bereits (already)
Etage (floor)	Stellplatz (parking space)	Straße (street)	Hinweise (notes)
bezogen (related)	zwei (two)	vorhanden (available)	sorgfältig (carefully)
neu (new)	Terrasse (terrace)	Bus (bus)	enthalten (include)
Nähe (near)	Kellerabteil (basement compartment)	Bushaltestelle (bus stop)	steht (stands)
ebenfalls (also)	hell (bright)	Minuten (minutes)	Mitarbeiter (employees)
gut (good)	weiße (white)	Restaurants (restaurant)	Ort (location)
gibt (gives)	Fliesenspiegel (tile backsplash)	nahe (near)	Soweit (as far as)

General Textual Description	Textual Description of Facilities	Textual Description of Location	Textual Description of Miscellaneous
drei (three)	großen (large)	Haus (house)	Maßangaben (dimensions)
sofort (immediately)	Wohnräume (living space)	Hauptbahnhof (central station)	Wohnungen (apartments)
Keller (basement)	Gäste-WC (guests wc)	Richtung (direction)	benannten (named)
ruhigen (quiet)	€ (€)	ruhige (quiet)	jegliche (any)
großzügige (generous)	ebenfalls (also)	Wohngebiet (residential)	Anfrage (request)
angemietet (rented)	Bodenbelag (flooring)	bequem (convenient)	GmbH (GmbH)
OG (upper floor)	inkl. (inclusive)	Kindergärten (kindergardens)	bitten (request)
Badewanne (bath tub)	verlegt. (relocated)	erreicht (reached)	direkt (directly)
Jahr (year)	Blick (view)	Ort (city)	senden (send)
Erdgeschoss (ground floor)	Räumen (rooms)	verfügt (disposes)	Anfragen (requests)
Gebäude (building)	großzügige (generous)	direkter (direct)	vereinbaren (arrange)
Abstellraum (storage room)	Rollläden (shutters)	Bad (bathroom)	Vollständigkeit (completeness)
bieten (provide)	Parkett (parquet)	m (meter)	beträgt (amounts)
Garten (garden)	hochwertigen (high quality)	innerhalb (within)	Einrichtungen (facilities)
Räume (rooms)	modernen (modern)	Flughafen (airport)	Telefonnummer (phone number)
Immobilie (property)	modern (modern)	zentral (centrally)	Service-Center (service center)

General Textual Description	Textual Description of Facilities	Textual Description of Location	Textual Description of Miscellaneous
Wohnfläche (living space)	Wandfliesen (wall tiles)	Wohnlage (residential)	Rufnummer (phone number)
Tageslichtbad (daylight bathroom)	direkt (direct)	Leipziger (Leipziger)	Grundrissgrafiken (floor plan)
Terrasse (terrace)	Verfügung. (available)	etwa (approx)	gern (gradly)
modernen (modern)	Diele (hallway)	Entfernung (distance)	Gerne (gradly)
Wohnanlage (residential complex)	offene (open)	ÖPNV (local public transport)	handelt (acts)
erreichen (reach)	Wohnbereich (living area)	zahlreichen (numerous)	Fragen (questions)
bequem (convenient)	Warmwasser (hot water)	Kindergarten (kindergarden)	beruhen (are based)
schönen (nice)	Wannenbad (tub bath)	Autobahnen (highways)	Unterlagen (documents)
insgesamt (total)	Decken (ceiling)	min (min)	Kopien (copies)
Mehrfamilienhauses (multifamily house)	Bäder (bathrooms)	Apotheken (pharmacies)	Vermietung (renting)
Energieausweis (energy certificate)	neuen (new)	Bedarf (demand)	Irrtum (mistake)
neuen (new)	Mieter (renter)	gelegen (located)	basieren (are based)
gepflegte (neat)	Duschbad (shower bathroom)	S-Bahn (city train)	Vermieter (lessor)
ersten (first)	bzw. (resp.)	ebenso (also)	bieten (offer)
besteht (exist)	Tiefgarage (underground garage)	Altstadt (old town)	ab (as of)
Aufzug (elevator)	renoviert (renovated)	Ortsteil (part of the village)	Immobilie (property)

General Textual Description	Textual Description of Facilities	Textual Description of Location	Textual Description of Miscellaneous
gelangen (reach)	möglich (possible)	rund (around)	gemacht (made)
per (per)	großes (large)	diverse (various)	E-Mail (e-mail)
Bitte (please)	hochwertiger (high quality)	beliebten (popular)	erreichen (reach)
großer (large)	elektrische (electric)	ruhig (quiet)	jederzeit (anytime)
Dachgeschoss (top floor)	separate (separate)	guten (good)	darüber (about this)
großzügigen (generous)	bereits (already)	Leipzig (leipzig)	Eigentümer (owner)
ruhiger (quiet)	Raufasertapete (ingrain wallpaper)	ruhiger (quiet)	Vorhandensein (presence)
Wohnhaus (residence)	frisch (fresh)	zentrale (central)	beachten (note)

Note. Own research. Created from raw data set. Stopword cleaned with python standard library Natural Language Tool Kit (Sarica & Luo, 2021, p. 1).

Appendix A - 18

Distribution of Identifier Frequency

Individual listings	Occurences	Observations
386,857	1	386,857
260,999	2	521,998
138,865	3	416,595
105,145	4	420,580
66,115	5	330,575
48,883	6	293,298
37,281	7	260,967
31,881	8	255,048
25,944	9	233,496
20,443	10	204,430
17,597	11	193,567
15,896	12	190,752
14,409	13	187,317
11,227	14	157,178
9,952	15	149,280
8,485	16	135,760
7,331	17	124,627
6,741	18	121,338
6,052	19	114,988
5,375	20	107,500
5,051	21	106,071
4,438	22	97,636
4,086	23	93,978
3,751	24	90,024
3,381	25	84,525
3,139	26	81,614
2,712	27	73,224
2,529	28	70,812
2,201	29	63,829
2,060	30	61,800
1,700	31	52,700

Individual listings	Occurences	Observations
1,676	32	53,632
1,649	33	54,417
1,530	34	52,020
1,447	35	50,645
1,278	36	46,008
1,111	37	41,107
1,172	38	44,536
921	39	35,919
943	40	37,720
880	41	36,080
809	42	33,978
704	43	30,272
682	44	30,008
665	45	29,925
572	46	26,312
592	47	27,824
530	48	25,440
535	49	26,215
523	50	26,150
504	51	25,704
430	52	22,360
411	53	21,783
412	54	22,248
371	55	20,405
324	56	18,144
318	57	18,126
291	58	16,878
288	59	16,992
280	60	16,800
264	61	16,104
252	62	15,624
232	63	14,616
227	64	14,528

Individual listings	Occurences	Observations
256	65	16,640
228	66	15,048
225	67	15,075
237	68	16,116
265	69	18,285
226	70	15,820
237	71	16,827
312	72	22,464
393	73	28,689
466	74	34,484
629	75	47,175
635	76	48,260
458	77	35,266
243	78	18,954
96	79	7,584

Note. Own research. Created from raw data set.

Appendix A - 19

Distribution of Observations per Collection Cycle

Start collection	End collection	Observations
2019-01-28 13:27:34	2019-01-28 17:58:20	87,109
2019-02-04 09:14:17	2019-02-04 11:34:51	87,371
2019-02-11 09:57:49	2019-02-11 12:37:07	88,561
2019-02-18 09:10:01	2019-02-18 12:07:38	88,345
2019-02-25 09:48:01	2019-02-25 13:04:28	86,355
2019-03-04 10:29:18	2019-03-04 13:13:49	85,620
2019-03-18 07:36:14	2019-03-18 10:21:49	86,927
2019-03-25 06:11:11	2019-03-25 09:15:43	90,702
2019-04-01 01:21:17	2019-04-01 04:55:18	86,664
2019-04-08 14:19:58	2019-04-08 17:35:02	91,533
2019-04-15 15:04:17	2019-04-15 18:16:45	92,053
2019-04-23 08:13:06	2019-04-23 11:28:23	89,498
2019-04-29 06:33:45	2019-04-29 09:46:46	92,093
2019-05-06 07:49:33	2019-05-06 11:07:17	91,658
2019-05-13 11:24:39	2019-05-13 14:41:52	92,258
2019-05-20 10:12:23	2019-05-20 13:30:10	92,642
2019-05-27 10:09:42	2019-05-27 19:57:38	91,368
2019-06-01 15:47:40	2019-06-01 19:08:22	92,707
2019-06-10 12:24:47	2019-06-10 15:53:59	94,029
2019-06-17 10:53:55	2019-06-17 17:56:49	91,293
2019-06-24 08:11:07	2019-06-24 15:07:52	90,620
2019-07-01 07:45:25	2019-07-01 11:15:50	90,233
2019-07-08 06:44:53	2019-07-08 10:04:46	93,187
2019-07-15 08:00:32	2019-07-15 11:23:56	91,279
2019-07-22 09:50:41	2019-07-22 15:27:21	90,001
2019-07-29 10:27:11	2019-07-29 13:55:37	88,378
2019-08-05 07:02:35	2019-08-05 10:30:28	88,979
2019-08-12 09:25:51	2019-08-12 21:20:57	88,169
2019-08-19 06:08:19	2019-08-19 09:50:04	90,086
2019-08-26 09:52:50	2019-08-26 13:28:22	85,675
2019-09-02 07:40:42	2019-09-02 11:25:45	86,179

Start collection	End collection	Observations
2019-09-09 09:07:09	2019-09-09 12:38:56	87,657
2019-09-16 05:50:30	2019-09-16 09:14:46	80,795
2019-09-23 08:54:32	2019-09-23 12:16:21	77,072
2019-09-30 08:44:24	2019-09-30 12:10:31	77,446
2019-10-07 08:21:32	2019-10-07 11:49:44	78,543
2019-10-14 07:41:52	2019-10-14 11:03:17	75,558
2019-10-21 07:49:07	2019-10-21 12:54:09	76,102
2019-10-28 07:38:05	2019-10-28 11:04:21	86,777
2019-11-04 13:06:23	2019-11-04 19:51:11	86,874
2019-11-11 06:29:12	2019-11-11 09:51:23	90,178
2019-11-18 18:41:03	2019-11-18 22:14:17	91,182
2019-11-25 06:56:07	2019-11-25 10:25:26	89,386
2019-12-02 09:55:02	2019-12-02 13:26:57	88,040
2019-12-09 07:24:40	2019-12-09 11:25:02	89,948
2019-12-16 12:00:42	2019-12-16 15:35:06	86,794
2019-12-24 10:21:40	2019-12-24 15:08:23	81,515
2019-12-30 14:29:01	2019-12-30 17:46:08	81,025
2020-01-06 10:10:30	2020-01-06 13:36:31	80,271
2020-01-13 18:24:15	2020-01-13 21:59:43	89,594
2020-01-20 07:17:37	2020-01-20 10:45:20	89,557
2020-01-27 08:10:57	2020-01-27 11:36:04	88,252
2020-02-03 11:14:57	2020-02-03 14:36:55	87,303
2020-02-10 14:52:37	2020-02-10 18:24:37	88,394
2020-02-17 07:20:07	2020-02-17 10:53:04	88,161
2020-02-24 16:45:32	2020-02-25 00:20:28	87,746
2020-03-01 23:25:48	2020-03-02 03:22:56	81,705
2020-03-08 20:52:51	2020-03-09 00:54:11	86,128
2020-03-15 22:18:33	2020-03-16 02:18:50	84,477
2020-03-23 08:15:44	2020-03-23 12:36:41	81,863
2020-03-30 08:46:32	2020-03-30 12:59:36	81,290
2020-04-06 07:20:45	2020-04-06 11:30:42	85,276
2020-04-13 09:08:05	2020-04-13 13:19:41	87,533
2020-04-20 08:44:19	2020-04-20 12:52:08	88,527

Start collection	End collection	Observations
2020-04-27 07:10:36	2020-04-27 11:22:37	93,893
2020-05-04 08:48:27	2020-05-04 13:12:21	93,138
2020-05-11 10:06:40	2020-05-11 14:20:07	95,347
2020-05-18 07:55:40	2020-05-18 12:16:05	96,798
2020-05-25 07:20:12	2020-05-25 11:36:51	94,421
2020-06-01 18:11:20	2020-06-02 01:48:37	94,735
2020-06-08 08:30:52	2020-06-08 12:58:30	95,804
2020-06-15 06:58:24	2020-06-15 11:37:18	98,330
2020-06-22 08:51:37	2020-06-22 13:29:15	95,706
2020-06-29 09:13:17	2020-06-29 13:36:09	95,017
2020-07-06 15:20:30	2020-07-06 19:45:46	96,072
2020-07-13 09:45:27	2020-07-13 14:13:26	95,505
2020-07-20 09:36:36	2020-07-20 14:02:18	93,609
2020-07-27 00:01:08	2020-07-27 04:13:57	95,525
2020-08-03 07:40:36	2020-08-03 11:57:24	91,130

Note. Own research. Created from raw data set.

Appendix A - 20

Distribution of Vacancy Rate Variable

Vacancy Rate in %		Districts
Lower bound	Upper bound	
0.0	0.5	9
0.5	1.0	32
1.0	1.5	51
1.5	2.0	44
2.0	2.5	34
2.5	3.0	32
3.0	3.5	30
3.5	4.0	29
4.0	4.5	21
4.5	5.0	13
5.0	5.5	26
5.5	6.0	14
6.0	6.5	13
6.5	7.0	6
7.0	7.5	7
7.5	8.0	8
8.0	8.5	7
8.5	9.0	5
9.0	9.5	3
9.5	10.0	4
10.0	10.5	4
10.5	11.0	4
11.0	11.5	1
11.5	12.0	1
12.0	12.5	1
12.5	13.0	1
13.0	13.5	1
13.5	14.0	0

Note. Own research. Created from raw data set.

Appendix A - 21

Distribution of Settlement Type Variable

Settlement type	Districts
sparsely populated rural districts	102
rural districts with beginning concentration processes	100
urban districts	132
large cities not attached to an administrative district	67

Note. Own research. Created from raw data set.

Appendix A - 22

Distribution of GDP per Capita Variable

GDP per capita in €		Districts
Lower bound	Upper bound	
15,000	20,000	2
20,000	25,000	29
25,000	30,000	86
30,000	35,000	88
35,000	40,000	79
40,000	45,000	38
45,000	50,000	18
50,000	55,000	13
55,000	60,000	14
60,000	65,000	8
65,000	70,000	7
70,000	75,000	1
75,000	80,000	3
80,000	85,000	6
85,000	90,000	1
90,000	95,000	1
95,000	100,000	2
100,000	105,000	2
105,000	110,000	0
110,000	115,000	0
115,000	120,000	1
120,000	125,000	0
125,000	130,000	0
130,000	135,000	0
135,000	140,000	1
140,000	145,000	0
145,000	150,000	0
150,000	155,000	0
155,000	160,000	0
160,000	165,000	0

GDP per capita in €		Districts
Lower bound	Upper bound	
165,000	170,000	0
170,000	175,000	0
175,000	180,000	0
180,000	185,000	0
185,000	190,000	0
190,000	195,000	1
195,000	200,000	0

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 23

Distribution of Population Change Variable

Population change in %		Districts
Lower bound	Upper bound	
-3.50	-3.25	1
-3.25	-3.00	0
-3.00	-2.75	1
-2.75	-2.50	0
-2.50	-2.25	1
-2.25	-2.00	0
-2.00	-1.75	0
-1.75	-1.50	0
-1.50	-1.25	1
-1.25	-1.00	4
-1.00	-0.75	10
-0.75	-0.50	23
-0.50	-0.25	45
-0.25	0.00	60
0.00	0.25	88
0.25	0.50	93
0.50	0.75	46
0.75	1.00	15
1.00	1.25	6
1.25	1.50	4
1.50	1.75	0
1.75	2.00	0
2.00	2.25	1
2.25	2.50	0
2.50	2.75	1
2.75	3.00	0
3.00	3.25	0
3.25	3.50	0
3.50	3.75	0
3.75	4.00	0

Population change in %		Districts
Lower bound	Upper bound	
4.00	4.25	0
4.25	4.50	0
4.50	4.75	0
4.75	5.00	0
5.00	5.25	0
5.25	5.50	0
5.50	5.75	1
5.75	6.00	0

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 24

Distribution of Normalized Number of Listings Variable

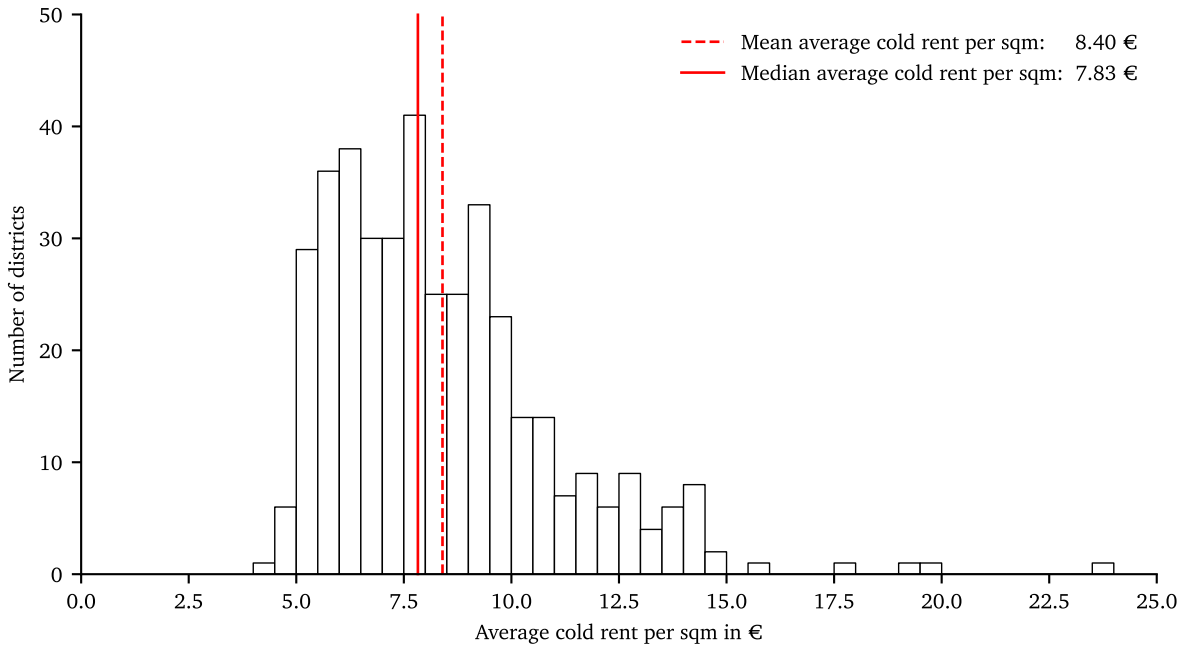
Normalized number of listings in %		Districts
Lower bound	Upper bound	
0.00	0.05	51
0.05	0.10	128
0.10	0.15	76
0.15	0.20	48
0.20	0.25	25
0.25	0.30	14
0.30	0.35	17
0.35	0.40	16
0.40	0.45	4
0.45	0.50	3
0.50	0.55	2
0.55	0.60	5
0.60	0.65	0
0.65	0.70	2
0.70	0.75	0
0.75	0.80	2
0.80	0.85	1
0.85	0.90	1
0.90	0.95	0
0.95	1.00	0
1.00	1.05	1
1.05	1.10	0
1.10	1.15	1
1.15	1.20	1
1.20	1.25	1
1.25	1.30	0
1.30	1.35	1
1.35	1.40	0
1.40	1.45	0
1.45	1.50	0

Normalized number of listings in %		Districts
Lower bound	Upper bound	
1.50	1.55	0
1.55	1.60	0
1.60	1.65	0
1.65	1.70	0
1.70	1.75	0
1.75	1.80	0
1.80	1.85	0
1.85	1.90	0
1.90	1.95	0
1.95	2.00	0
2.00	2.05	0
2.05	2.10	0
2.10	2.15	0
2.15	2.20	0
2.20	2.25	0
2.25	2.30	0
2.30	2.35	0
2.35	2.40	0
2.40	2.45	0
2.45	2.50	0
2.50	2.55	0
2.55	2.60	0
2.60	2.65	0
2.65	2.70	0
2.70	2.75	0
2.75	2.80	0
2.80	2.85	0
2.85	2.90	0
2.90	2.95	0
2.95	3.00	1

Note. Own research. Created from raw data set. Values included in lower bound excluded in upper bound.

Appendix A - 25

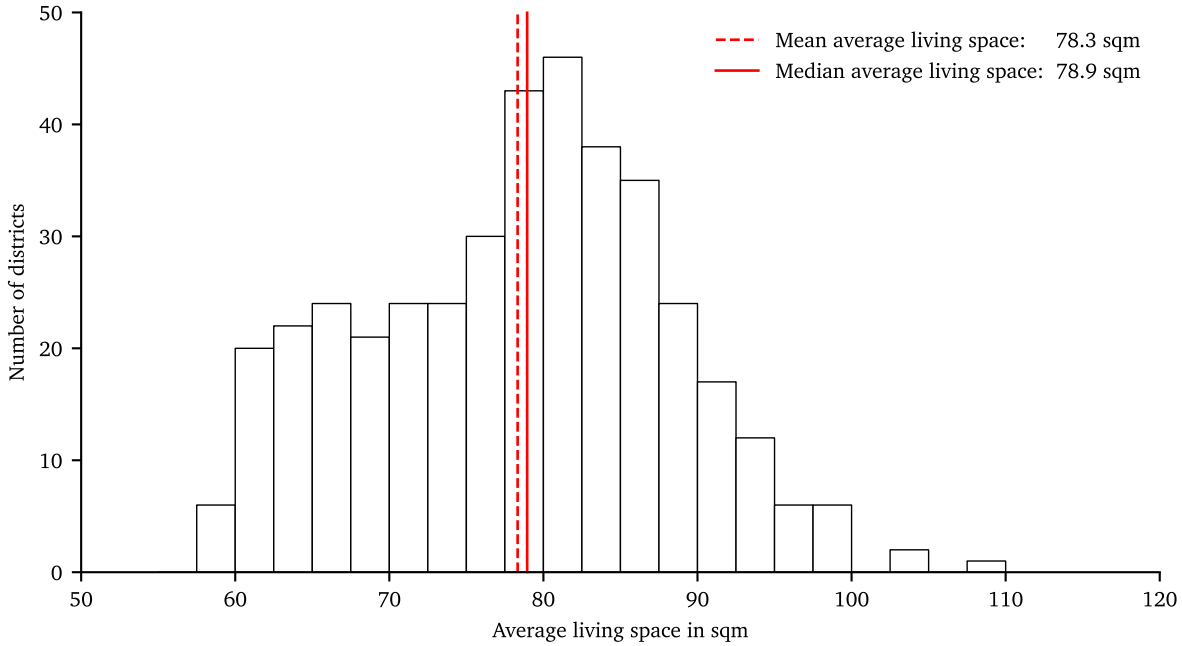
Distribution of Average Cold Rent per SQM



Note. Own research. Created from raw data set.

Appendix A - 26

Distribution of Average Living Space

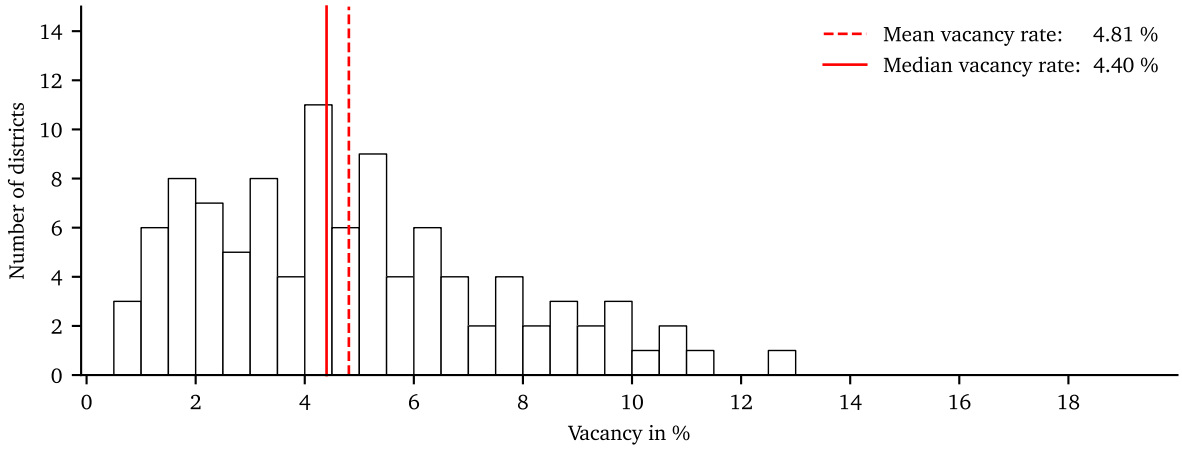


Note. Own research. Created from raw data set.

Appendix B: Methodology

Appendix B - 1

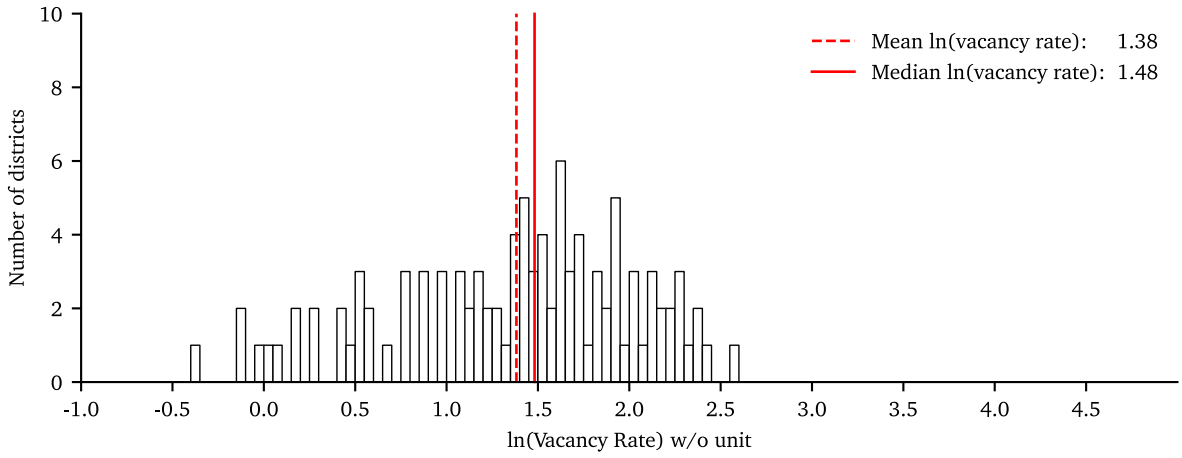
Histogram Vacancy Rate Sparsely Populated Rural District



Note. Own research. Created from raw data set.

Appendix B - 2

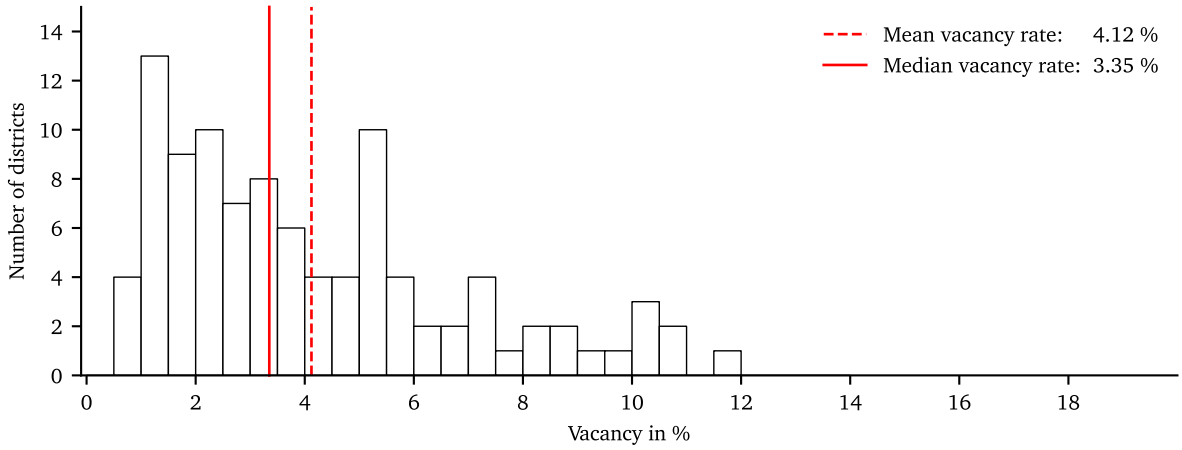
Histogram ln(Vacancy Rate) Sparsely Populated Rural District



Note. Own research. Created from raw data set.

Appendix B - 3

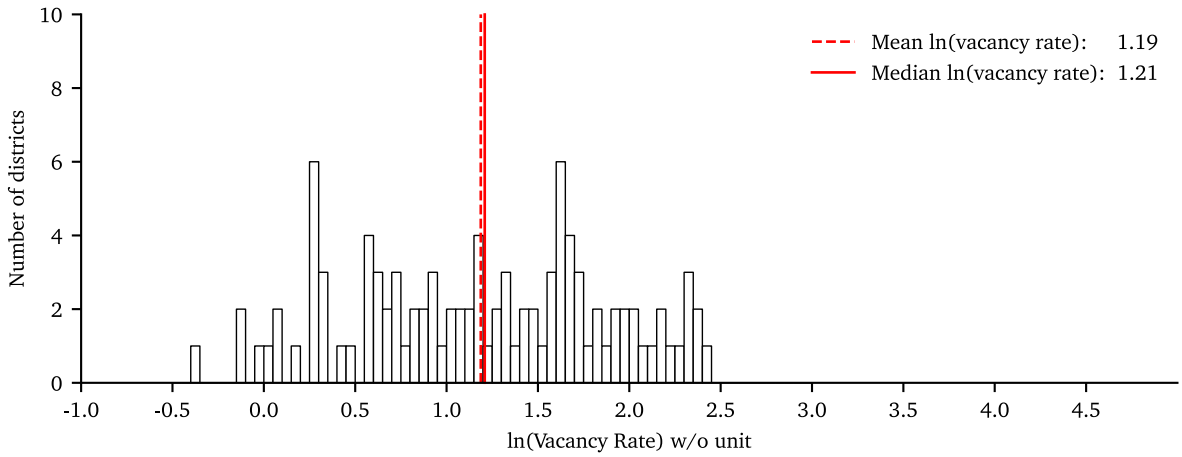
Histogram Vacancy Rate Rural District With Beginning Concentration Processes



Note. Own research. Created from raw data set.

Appendix B - 4

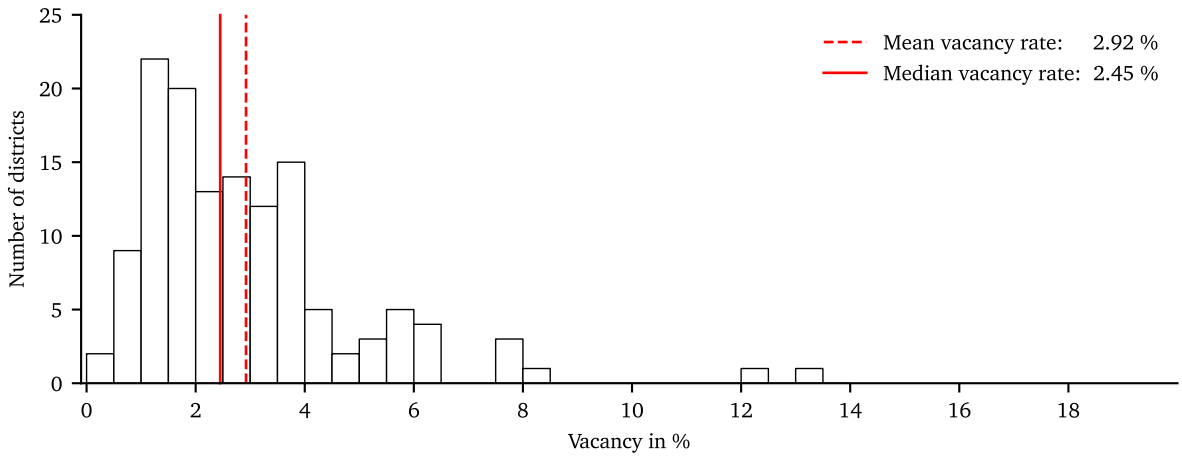
Histogram ln(Vacancy Rate) Rural District With Beginning Concentration Processes



Note. Own research. Created from raw data set.

Appendix B - 5

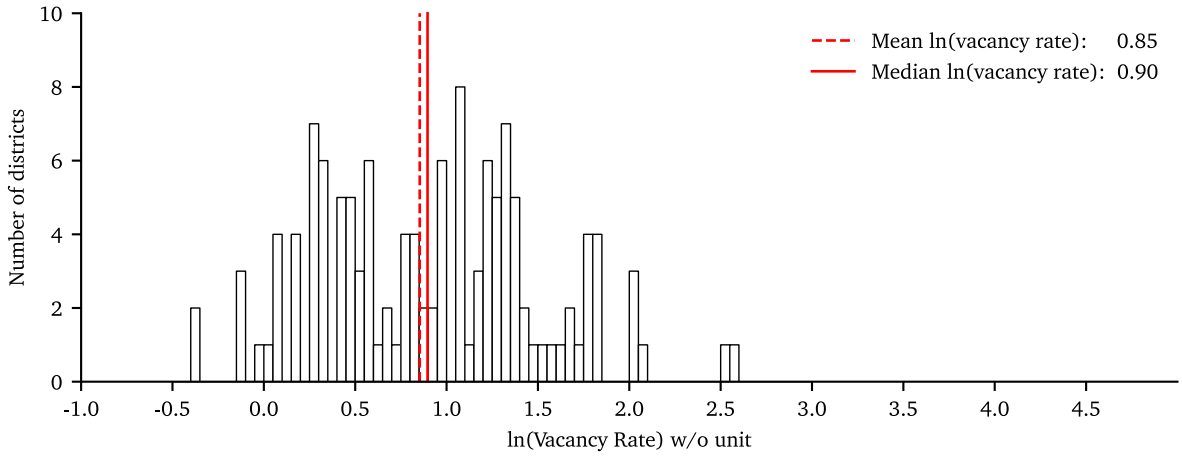
Histogram Vacancy Rate Urban Districts



Note. Own research. Created from raw data set.

Appendix B - 6

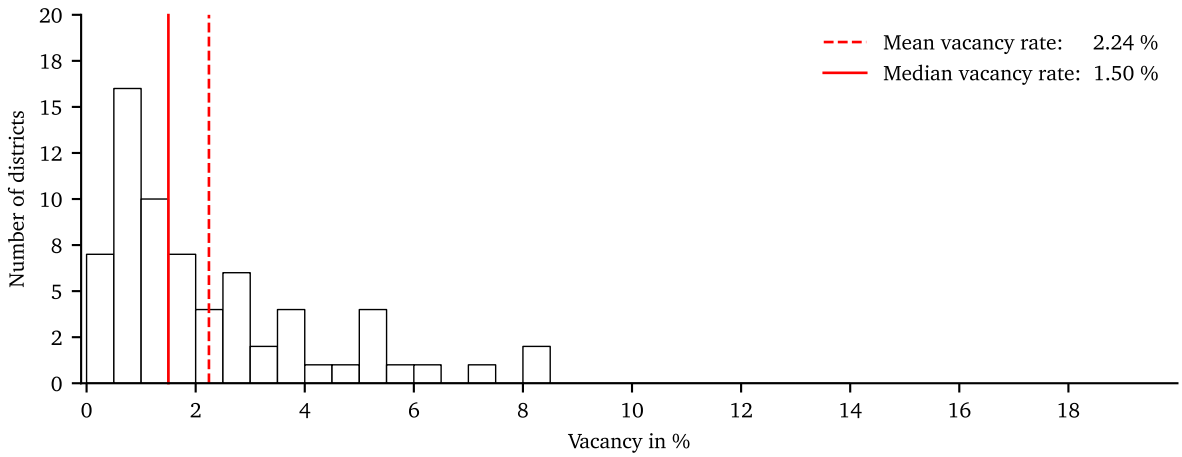
Histogram ln(Vacancy Rate) Urban Districts



Note. Own research. Created from raw data set.

Appendix B - 7

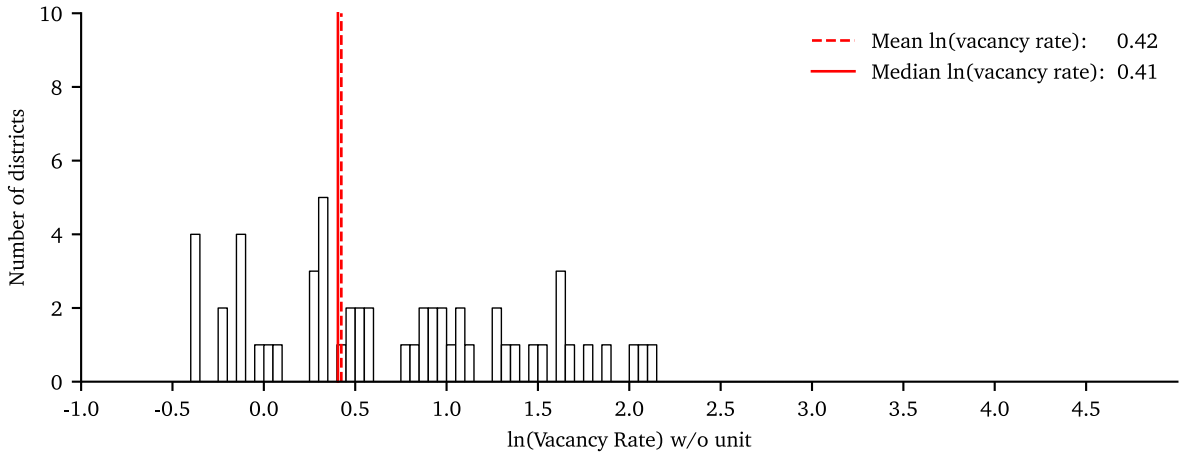
Histogram Vacancy Rate Large City Not Attached to an Administrative District



Note. Own research. Created from raw data set.

Appendix B - 8

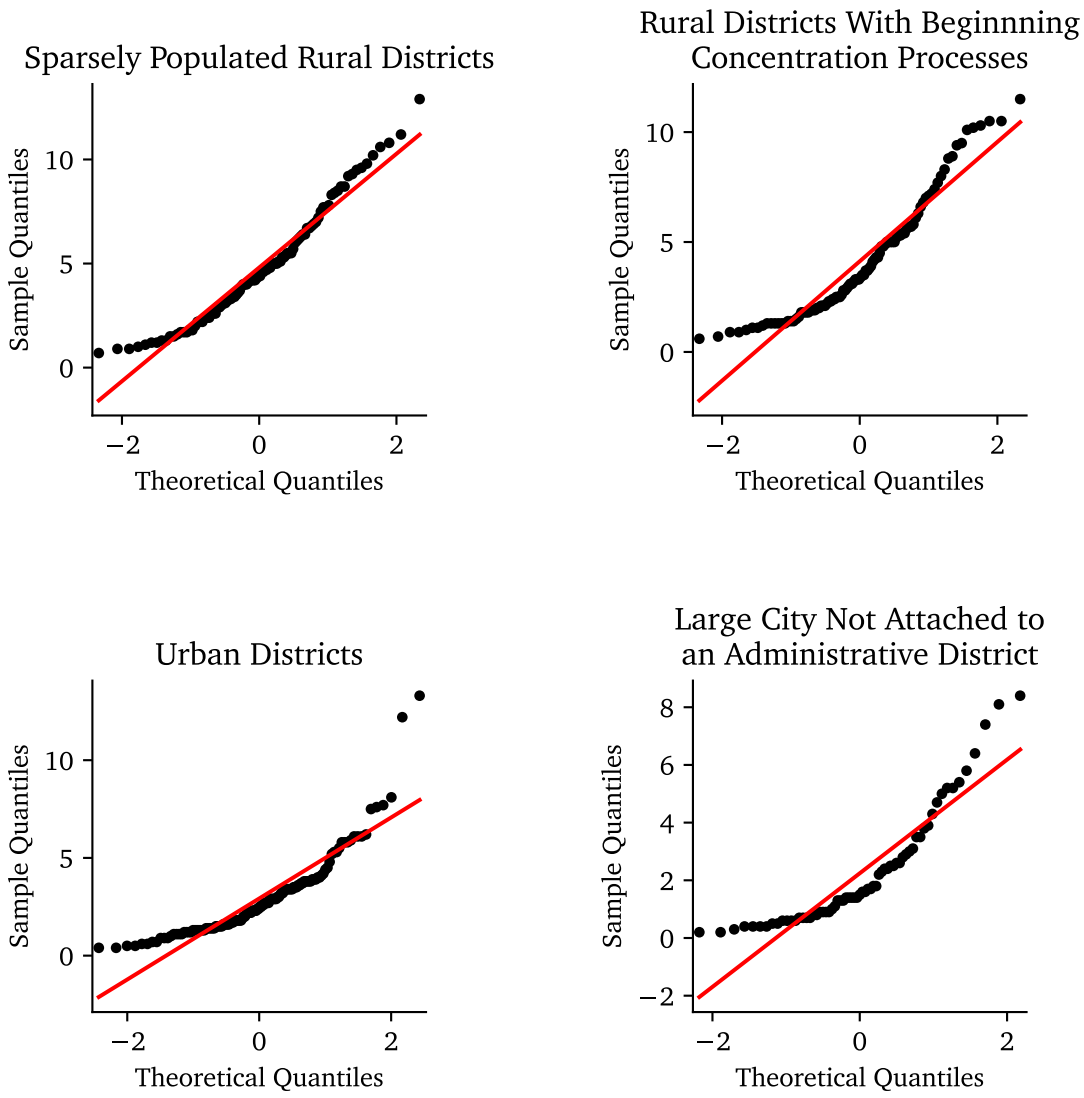
Histogram ln(Vacancy Rate) Large City Not Attached to an Administrative District



Note. Own research. Created from raw data set.

Appendix B - 9

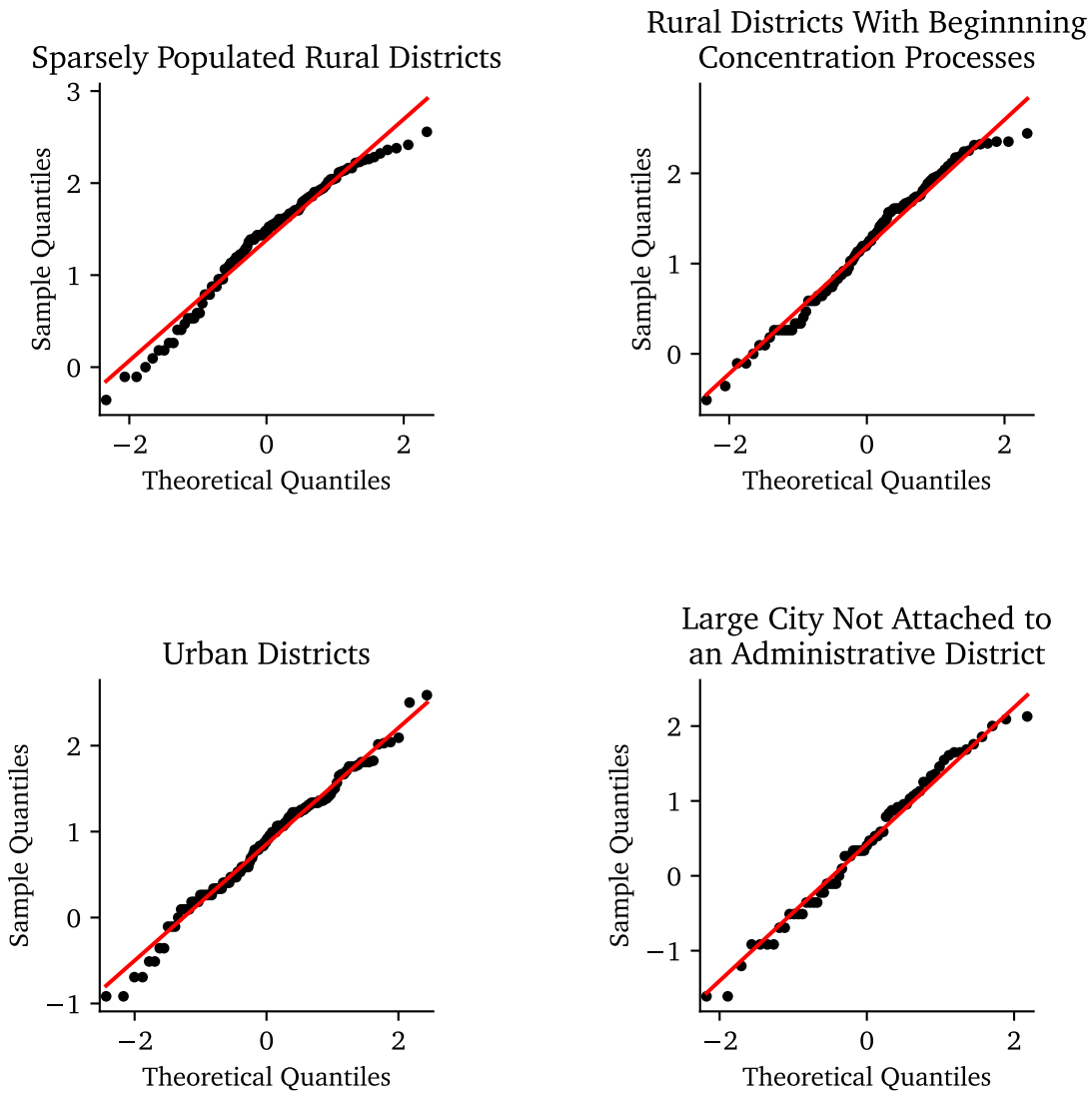
Q-Q Plot Vacancy Rate Settlement Types



Note. Own research. Created from raw data set.

Appendix B - 10

Q-Q Plot ln(Vacancy Rate) Settlement Types



Note. Own research. Created from raw data set.

Appendix B - 11

Normality of Variables Vacancy Rate and ln(Vacancy Rate) for Settlement Types

Settlement type	Variable	Skewness	Kurtosis	Standard deviation
Sparsely populated rural districts (n=102)	Vacancy rate	0.67	-0.12	2.75
	Ln(Vacancy rate)	-0.57	-0.30	0.66
Rural districts with beginning concentration processes (n=100)	Vacancy rate	0.91	0.01	2.73
	Ln(Vacancy rate)	-0.20	-0.77	0.71
Urban districts (n=132)	Vacancy rate	2.06	6.56	2.08
	Ln(Vacancy rate)	-0.19	0.06	0.68
Large cities not attached to an administrative district (n=67)	Vacancy rate	1.42	1.53	1.98
	Ln(Vacancy rate)	-0.13	-0.65	0.92

Note. Own research. Created from raw data set.

Appendix B - 12

Interview guide German

1. Fragen zum Experten und zur Relevanz der Themen Onlineimmobilienangebotsdaten und Leerstandsdaten

1a) Bitte beschreiben Sie überblicksartig Ihren Beruf und Ihre derzeitige Position, einschließlich Ihrer Hauptverantwortlichkeiten.

1b) Wie lange arbeiten Sie bereits in diesem oder einem ähnlichen Tätigkeitsfeld?

1c) Haben Sie in der Vergangenheit Leerstandsdaten verwendet? Wenn ja, könnten Sie bitte erklären, wozu und woher Sie diese Daten bezogen haben?

1d) Wie beurteilen Sie die Verfügbarkeit von Leerstandsdaten?

1e) Haben Sie in der Vergangenheit Onlineimmobilienangebotsdaten verwendet? Wenn ja, könnten Sie bitte erklären, wozu und woher Sie diese Daten bezogen haben?

2. Fragen zu möglichen leerstandserklärenden Faktoren

2a) Welche Faktoren könnten sich aus Ihrer Sicht allgemein zur Schätzung von Leerstandsquoten eignen?

2b) Welche Informationen könnten in Onlineimmobilienangebotsdaten enthalten sein, die die Schätzung von Leerstandsquoten unterstützen könnten?

3. Fragen zu den ausgewählten Variablen

Die folgenden Fragen beziehen sich auf marktaktiven Leerstand (Leerstandsquoten) nach der Definition des CBRE-Empirica-Leerstandsindexes. Nach dieser Definition umfasst marktaktiver Leerstand leerstehende Wohnungen, die unmittelbar disponibel sind, sowie leerstehende Wohnungen, die aufgrund von Mängeln derzeit nicht zur Vermietung anstehen, aber gegebenenfalls mittelfristig aktivierbar wären (<6 Monate).

Räumlich beziehen sich die Fragen auf die Ebene der Kreise und kreisfreien Städte.

Bitte schätzen Sie die Wichtigkeit der folgenden Variablen in Bezug auf die Schätzung von Leerstand ein, in dem Sie diese in eine Reihenfolge abnehmender Wichtigkeit bringen und eine

Wertung der Relevanz (++ (sehr wichtig), + (wichtig), 0 (unwichtig)) vornehmen.

- Bevölkerungsänderung des letzten Jahres
- BIP pro Kopf
- Durchschnittliche Angebotsmiete pro Quadratmeter
- Durchschnittliche Wohnfläche
- Normalisierte Angebotsanzahl (Anzahl Angebote im Verhältnis zum Gebäudebestand)
- Siedlungsstruktureller Raumtyp

4. Fragen zur Qualität und Übertragbarkeit der Ergebnisse

4a) Wie genau lässt sich Leerstand aus Ihrer Sicht mit Hilfe dieser Variablen schätzen?

4b) Lassen sich die geschätzten Zusammenhänge dieser Variablen aus Ihrer Sicht in der Zeit übertragen?

4c) Lassen sich die geschätzten Zusammenhänge dieser Variablen aus Ihrer Sicht räumlich übertragen?

Appendix B - 13

Interview guide English

1. Questions about the expert and the relevance of the topics online real estate supply data and vacancy data

1a) Please give an overview of your profession and your current position, including your main responsibilities.

1b) How long have you worked in this or a similar field of activity?

1c) Have you used vacancy data in the past? If so, could you please explain what for and where you obtained this data?

1d) How do you assess the availability of vacancy data?

1e) Have you used ORL data in the past? If so, could you please explain what for and where you obtained this data?

2. Questions on possible factors explaining vacancy rates

2a) In your view, what factors might be suitable for estimating vacancy rates in general?

2b) What information might be included in ORL data that could support the estimation of vacancy rates?

3. Questions about the selected variables

The following questions refer to market active vacancy (vacancy rates) as defined by the CBRE Empirica Vacancy Index. According to this definition, market active vacancy includes vacant apartments that are immediately available for disposition, as well as vacant apartments that are not currently available for rent due to defects, but could be activated in the medium term if necessary (<6 months). Spatially, the questions refer to the level of districts and large cities not attached to an administrative district.

Please estimate the importance of the following variables in relation to the estimation of vacancy by ranking them in order of decreasing importance and assigning a relevance score (++ (very important), + (important), 0 (unimportant)).

- Population change in the last year
- GDP per capita
- Average asking rent per square meter

- Average living space
- Normalized number of offers (number of offers in relation to the building stock)
- Settlement structure type of areaDurchschnittliche Wohnfläche

4. Questions on the quality and transferability of the results

4a) In your view, how accurately can vacancy be estimated using these variables?

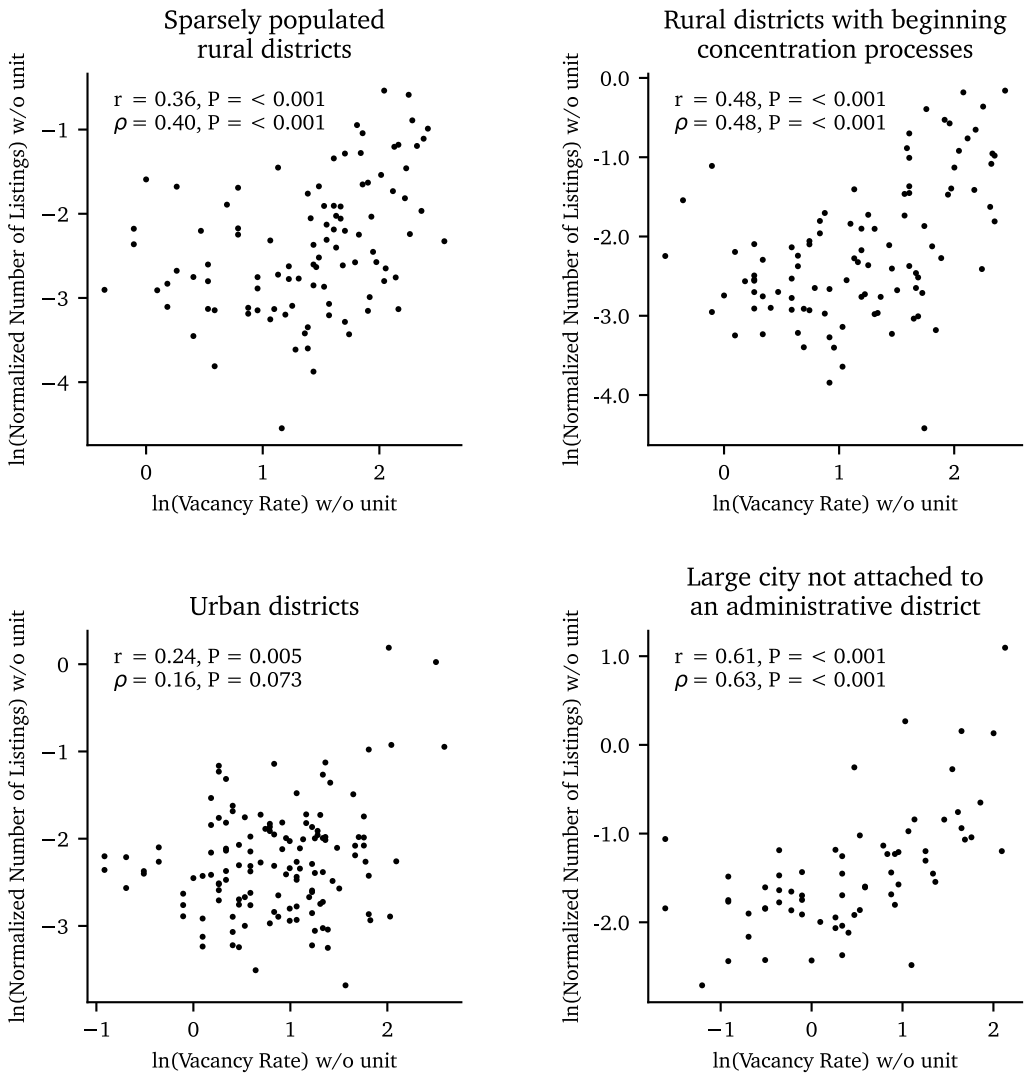
4b) In your view, can the estimated relationships of these variables be transferred in time?

4c) From your point of view, can the estimated correlations of these variables be spatially transferred?

Appendix C: Results

Appendix C - 1

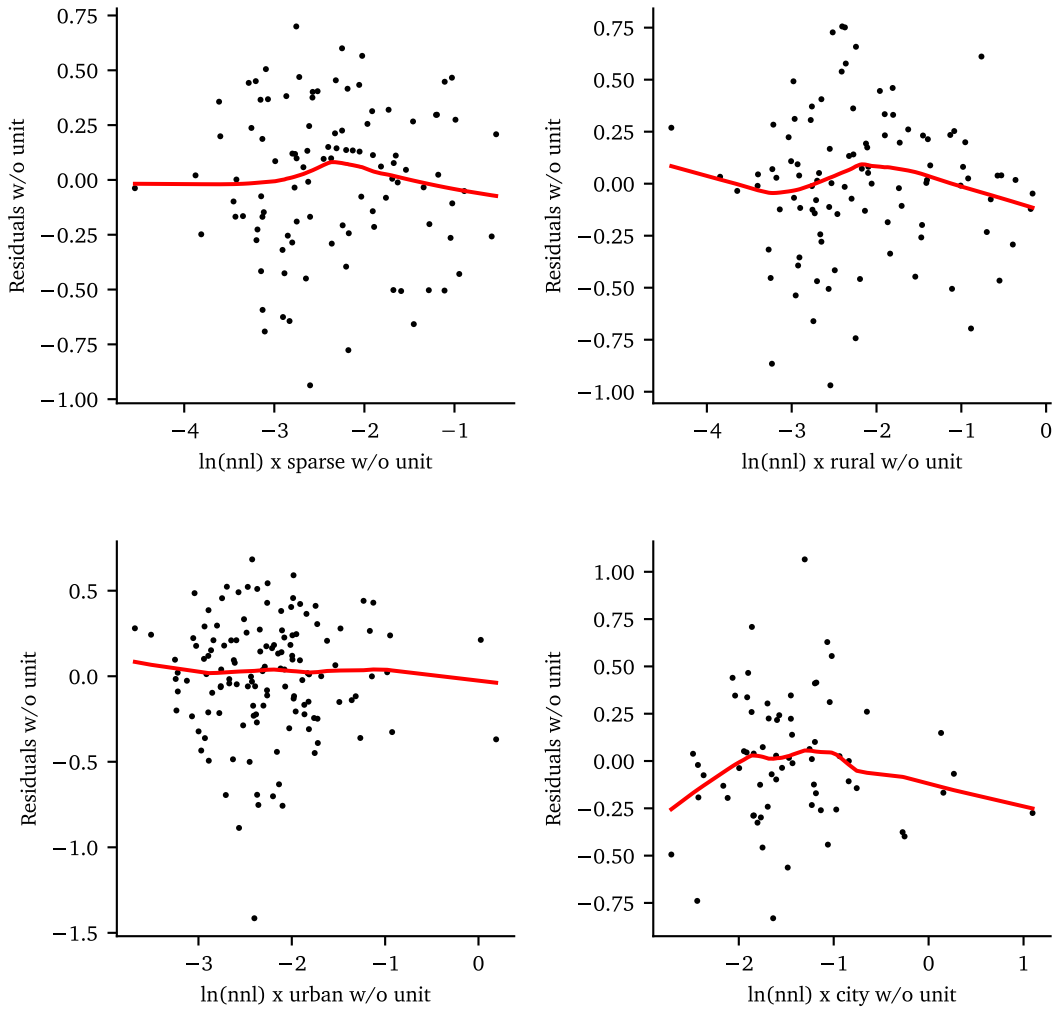
Relationship of Normalized Number of Listings and Vacancy Rate Subdivided



Note. Own research. Created from raw data set.

Appendix C - 2

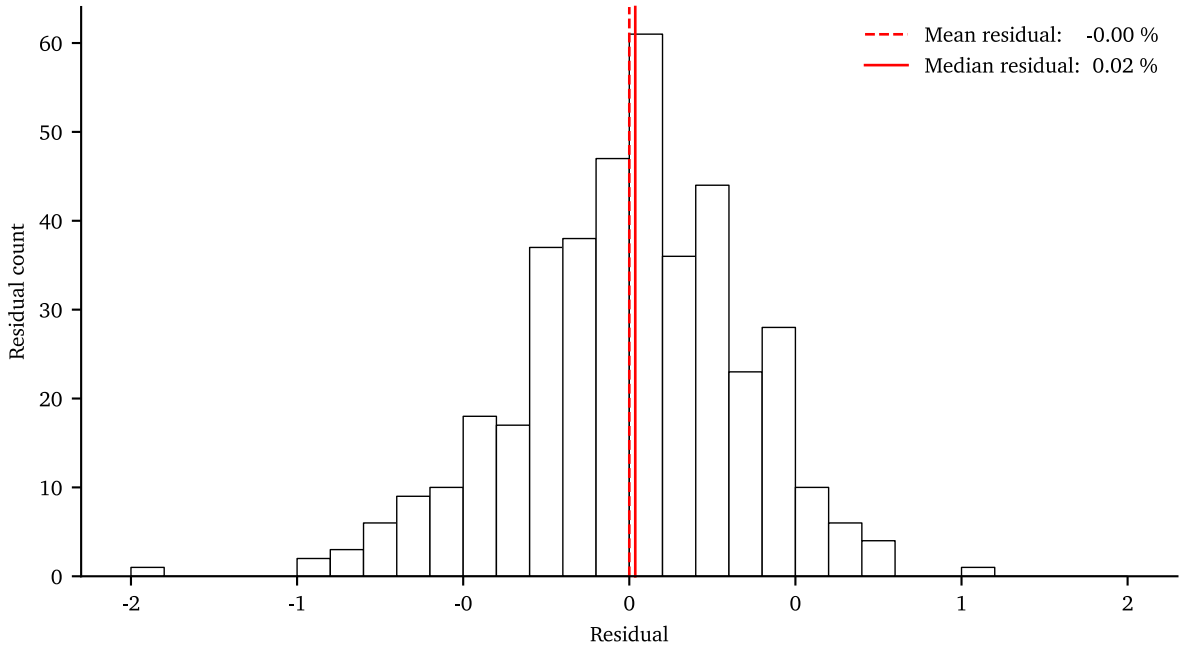
Residual Plots of Interaction Term Variables Base Model



Note. Own research. Created from raw data set.

Appendix C - 3

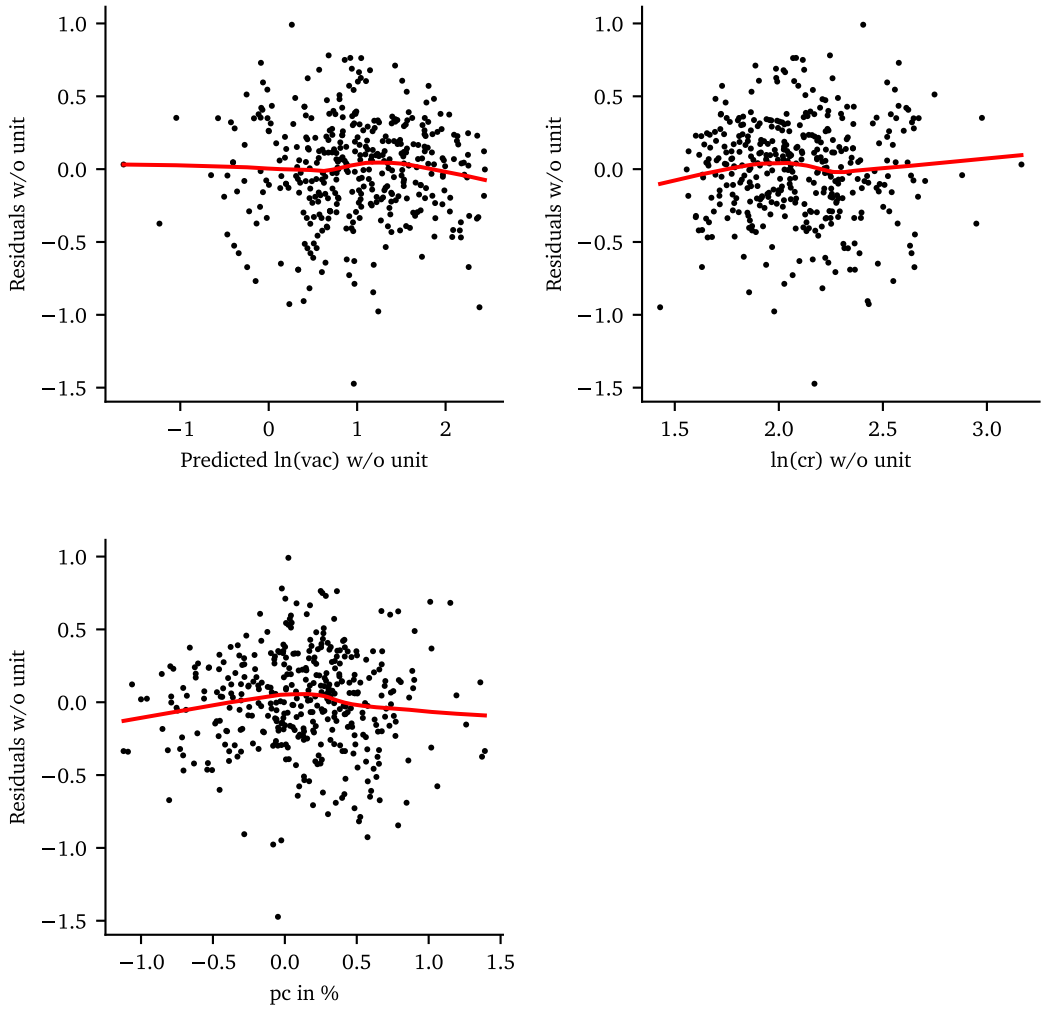
Histogram of Residuals of Base Model



Note. Own research. Created from raw data set.

Appendix C - 4

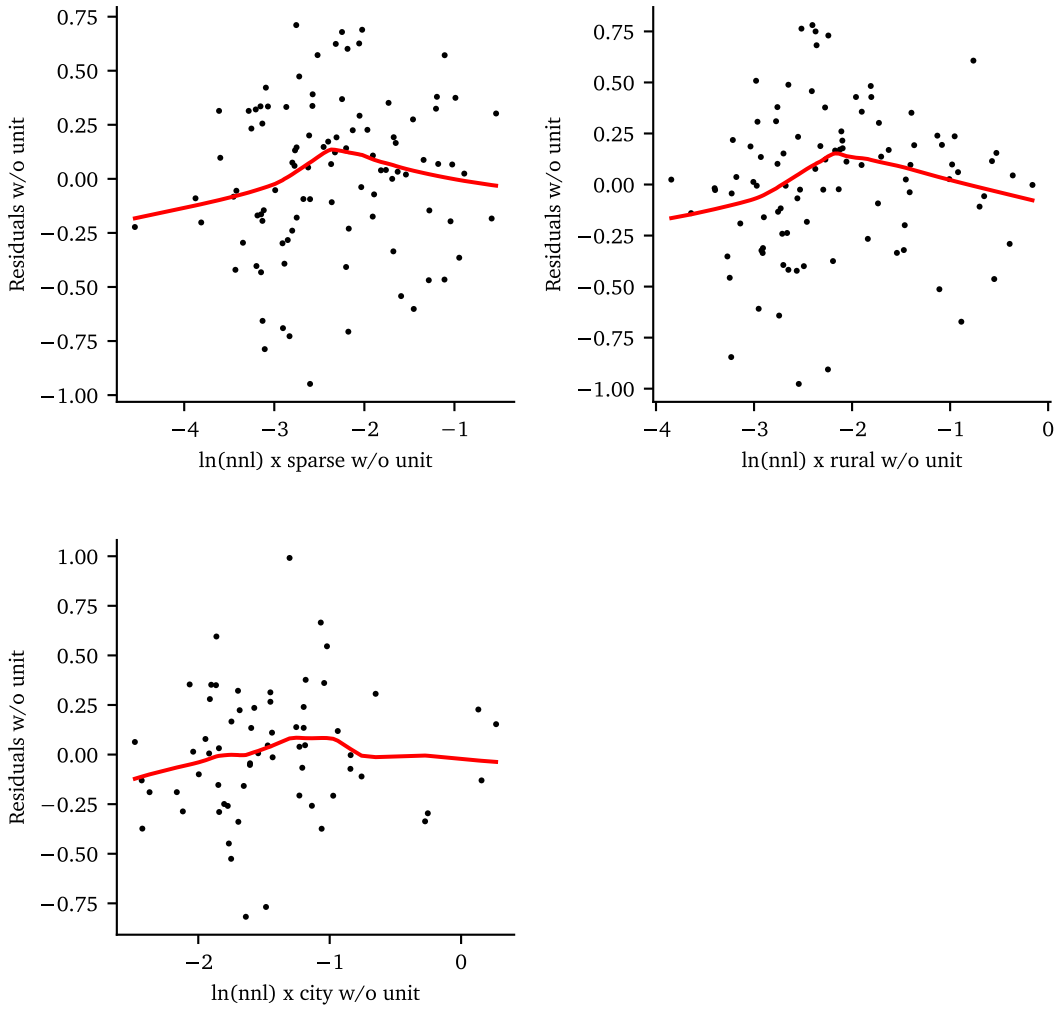
Residual Plots of Continuous Variables Final Model



Note. Own research. Created from raw data set.

Appendix C - 5

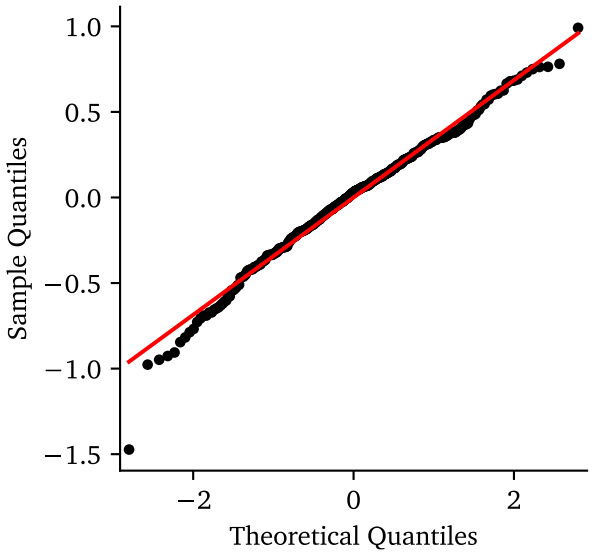
Residual Plots of Interaction Term Variables Final Model



Note. Own research. Created from raw data set.

Appendix C - 6

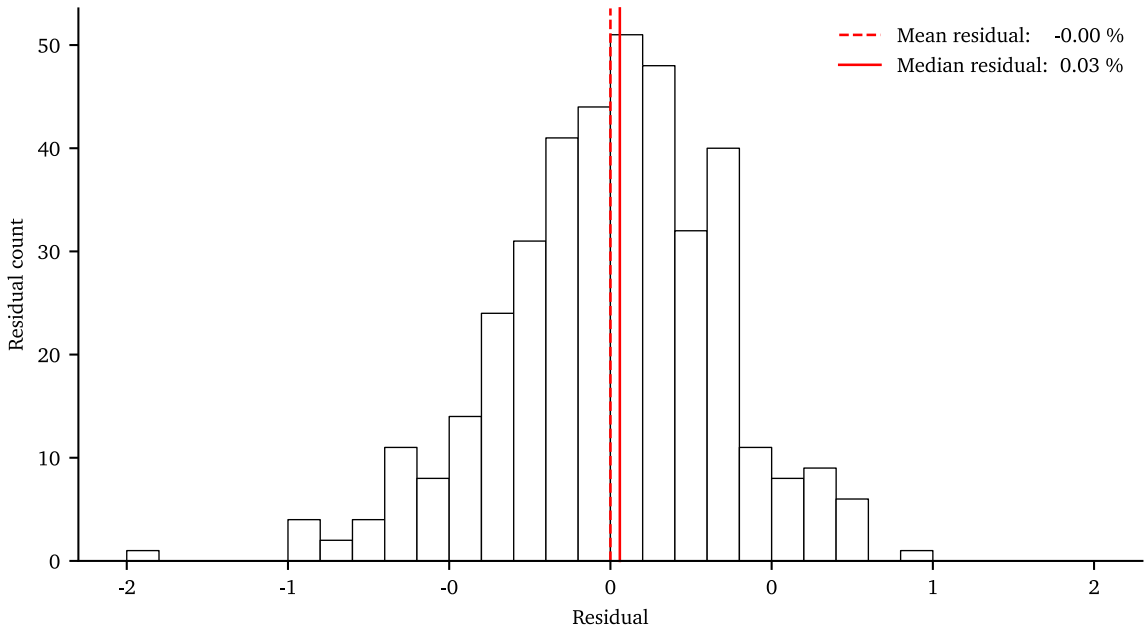
Q-Q Plot of Residuals of Final Model



Note. Own research. Created from raw data set.

Appendix C - 7

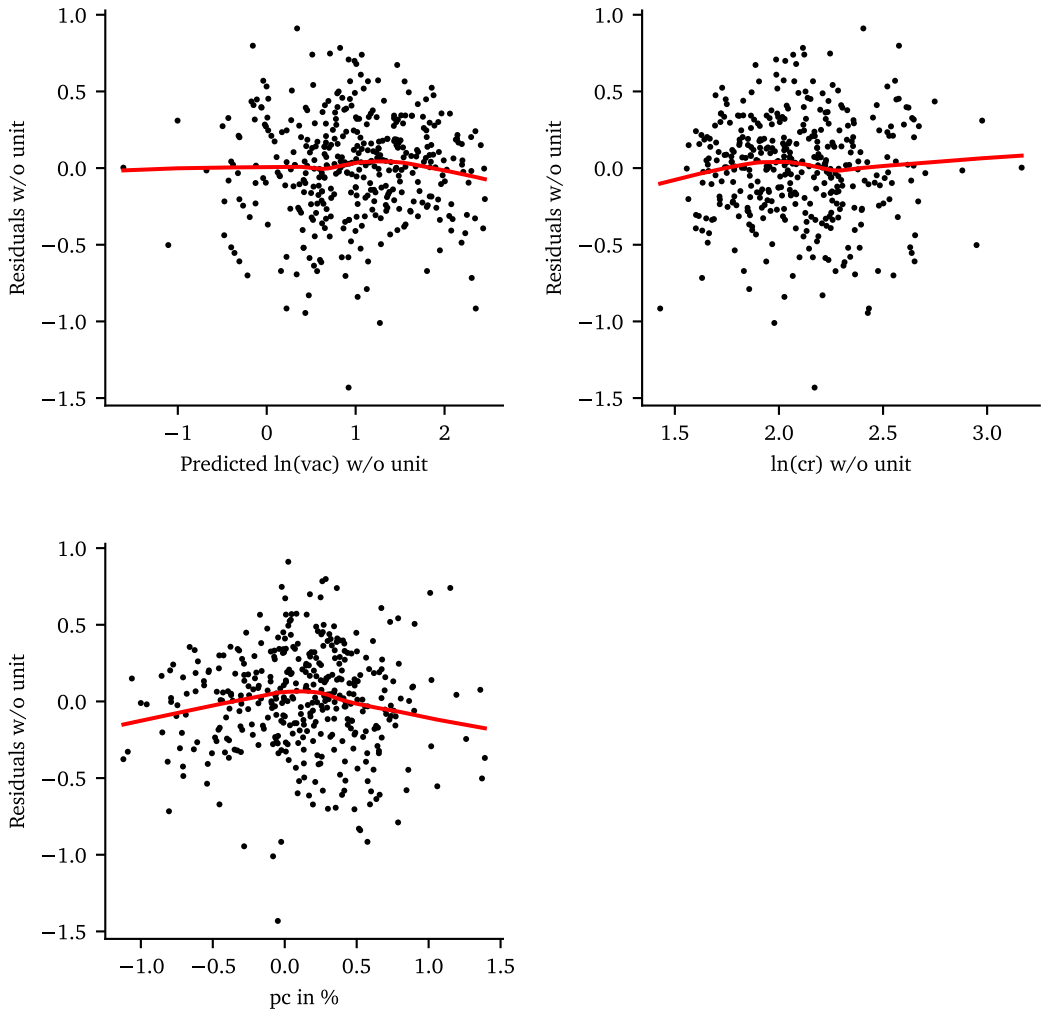
Histogram of Residuals of Final Model



Note. Own research. Created from raw data set.

Appendix C - 8

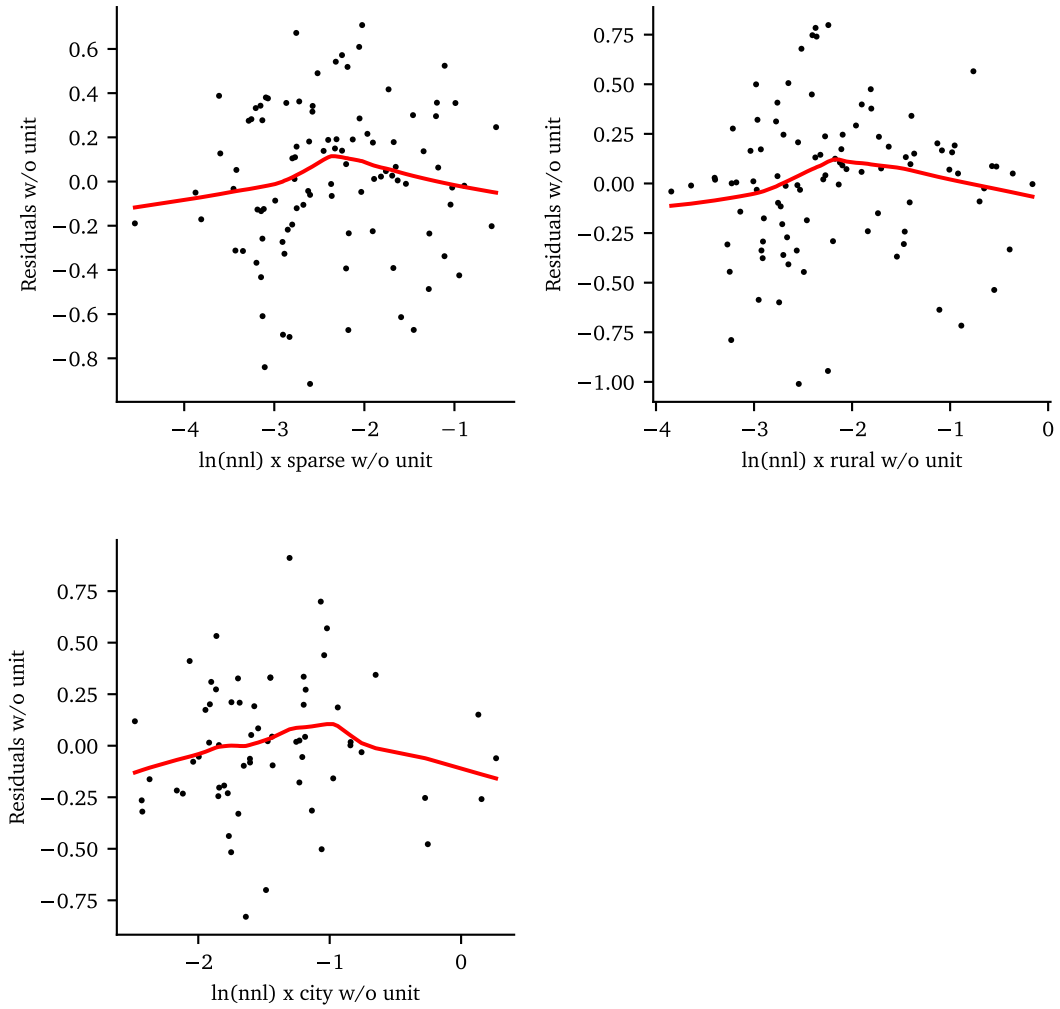
Residual Plots of Continuous Variables SLM



Note. Own research. Created from raw data set.

Appendix C - 9

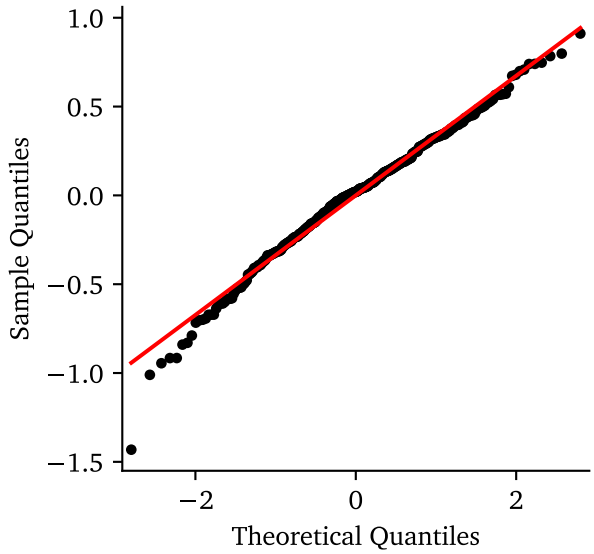
Residual Plots of Interaction Term Variables SLM



Note. Own research. Created from raw data set.

Appendix C - 10

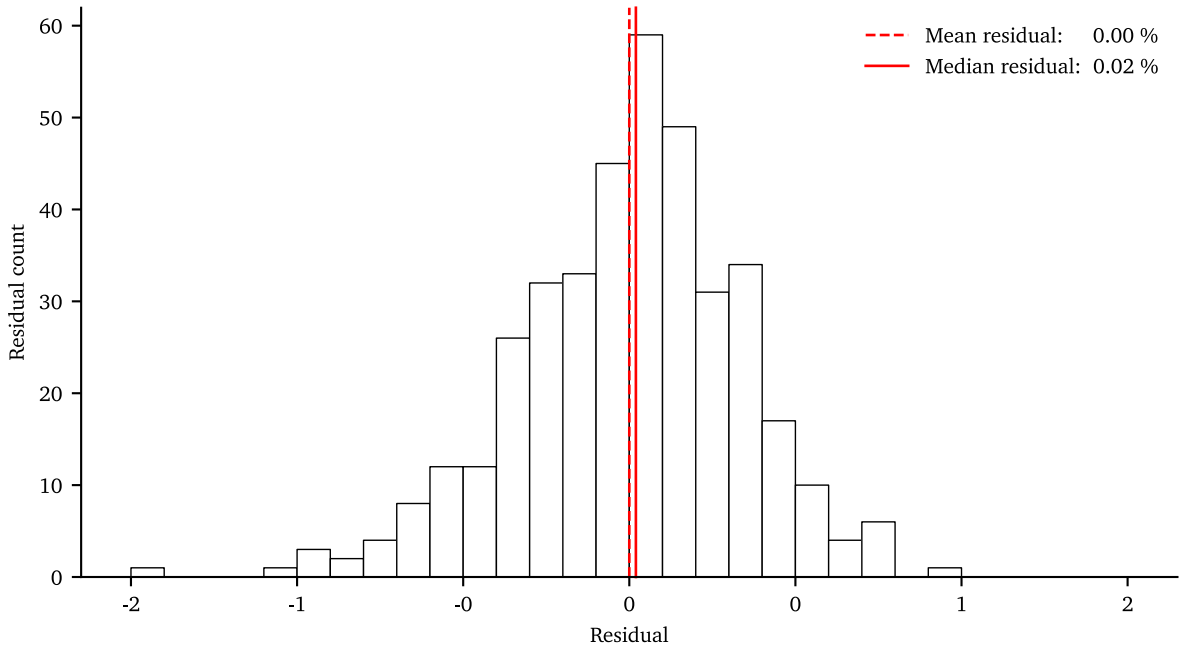
Q-Q Plot of Residuals of SLM



Note. Own research. Created from raw data set.

Appendix C - 11

Histogram of Residuals of SLM



Note. Own research. Created from raw data set.

Appendix C - 12

Results Spatial Error Regression Model

Variable	Coefficient	Standard error	t-statistic	P-value	[0.025	0.975]	VIF
Constant	5.246	0.207	25.351	<0.001	4.840	5.654	-
ln(cr)	-1.992	0.098	-20.426	<0.001	-2.184	-1.801	1.914
pc	-0.294	0.051	-5.722	<0.001	-0.395	-0.193	1.639
Interaction terms							
ln(nnl) x sparse	0.017	0.019	0.894	0.371	-0.021	0.056	1.262
ln(nnl) x rural	0.017	0.020	0.888	0.375	-0.021	0.055	1.228
ln(nnl) x city	0.157	0.032	4.869	<0.001	0.094	0.221	0.814
λ	0.433	0.076	5.718	<0.001	0.284	0.582	1.256

Model Summary

Dependent variable:	ln(vac)	Degrees of freedom:	384
Number of observations:	390	F-statistic (P-value):	304.930 (<0.001)
BP LM test (P-value):	8.844 (0.115)	Pseudo R^2:	0.805
BP F-test (P-value):	1.782 (0.115)	MAE (Vacancy Rate):	0.871
White‘s LM test (P-value):	20.899 (0.231)	AIC:	258.073
White’s F-test (P-value):	1.239 (0.231)	Moran’s I (P-value):	0.167 (<0.001)

Note. Own research. Created from raw data set.

Appendix C - 13

Expert Assessment of Variable Relevance

Expert	gdp	cr	pc	ls	st	nnl
S1	+	++	++	+	0	++
S2	+	+	+	+	+	++
P1	0/+	++	++	0/+	0/+	++
P2	+	0	+	0	+	0
I1	+	+	++	+	-	+
I2	+	++	++	0	0	+
I3	0	+	++	0/+	++	++

Note. Own research. ++: very important, +: important, 0: unimportant, -: can not be assessed