Towards Transparent Real Estate Markets – Assessing Investment Opportunities in Times of Artificial Intelligence

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

eingereicht an der Fakultät für Wirtschaftswissenschaften der Universität Regensburg

vorgelegt von:

JONAS WILLWERSCH

Berichterstatter:	Prof. Dr. Wolfgang Schäfers
	Prof. Dr. Stephan Bone-Winkel
Tag der Disputation:	09. Juni 2021

Towards Transparent Real Estate Markets – Assessing Investment Opportunities in Times of Artificial Intelligence

Jonas Willwersch

Acknowledgments

Without the support of various people, this dissertation would not exist in its current form. I would like to take this opportunity to thank you for all your help.

First, I would like to thank my supervisor Prof. Dr. Wolfgang Schäfers, who made the doctoral project possible and provided constant motivation as well as helpful criticism at all times. In addition, I would also like to thank my second supervisor, Prof. Dr. Stephan Bone-Winkel, for his willingness and support for supervision.

However, without my co-authors, the individual papers could not have been realized either. Special thanks go to Felix Lorenz and Dr. Marcelo Cajias, as well as to Dr. Cay Oertel, Prof. Dr. Wolfgang Schäfers and Prof. Dr. Franz Fürst for countless hours of constructive collaboration and scientific exchange.

My colleagues and friends at the IRE|BS Institute, above all Felix Lorenz, Carina Kaiser, Marina Kölbl and Leopold Winkler, have enriched the last years not only professionally, but especially on a personal level. One could not have asked for a better team. Thank you!

Despite my academic work, I was fortunate enough to maintain a connection to the industry through my employment at Competo Capital Partners GmbH. For the many years, I would like to thank the two managing directors Ralf Simon and Thomas Pscherer. Their ongoing support along all important professional decisions since my bachelor's degree was never taken for granted and is highly appreciated.

Throughout the entire time of my dissertation, I have received unconditional support and emotional backing from my family and friends. For this, I would like to thank my parents Dr. Mathias and Ulrike Willwersch, who have bolstered me up in many conversations. Thank you to my sisters Maja and Philippa Willwersch, and grandparents Lilo Willwersch and Ulrich Bader for their unwavering confidence. Knowing that there always is a safe and peaceful haven to call home is a great privilege. Special thanks also to my uncle Stefan Willwersch, without whom my path most probably would not have led to the real estate industry in the first place and who has supported me with professional advice throughout the entire time. Moreover, I thank Florian Pichl, who has motivated me on this journey from the very beginning. Finally, a big thank you to Leonie Hartmann for her boundless optimism and her support throughout all phases of the dissertation.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation and Background

Real estate plays an important role in today's world in various ways. In addition to its housing function, it is in greater demand than ever as an asset class. Due to continuing low interest rates and excess liquidity, investors are scrambling for investment opportunities, even in times of major societal crises like the worldwide COVID-19 pandemic. In addition to private individuals, the institutional world[1] shows an increased interest in real estate, offering them predominantly stable returns and benefitting not only pension plans of many people, but also helping to build up and preserve capital. In fact, real estate makes up 5.1% of any institutional portfolio, on average (Andonov et al., 2013).

However, investing in real estate is not an easy task. More precisely, market participation is challenging due to several reasons. First, real estate markets are typically characterized by large lot sizes as real estate is usually sold at high prices due to its locational value, physical structure, and size. Second, markets are rather illiquid and slow-moving, which arises from long transaction processes including search, due diligence, negotiation, and closing phases (Crosby & McAllister, 2004; Bond et al., 2007). Third, real estate markets are economy-dependent and cyclical. The state of the national economy as well as of individual industries determines the demand for space, which directly impacts prices and rents (Bone-Winkel, Focke, & Schulte, 2016). Fourth, traded assets are highly heterogeneous as they are individual goods with almost no standardized appearance. Hence, all information about a respective property is often only available to a very limited circle of people (Wong et al., 2012). All these characteristics contribute to real estate markets and their assets being considered non-transparent. Although 'transparency' is frequently used in the real estate context, it lacks a unified definition. K.-W. Schulte et al. (2005) therefore propose:

"*Real estate markets can be described as transparent when it becomes clear how the market mechanisms and the variables behind these mechanisms work, i.e. when there is as much information as possible available at any point in time.*" (K.-W. Schulte et al., 2005, p.91)

Hence, if information asymmetries are present or information is entirely missing when assessing real estate investments, lacking transparency can quickly become the biggest

[1] According to Bone-Winkel, Schulte et al. (2016), this includes insurance companies, pension funds, open and closed-end real estate funds, real estate operating companies and real estate investment trusts, among others.

challenge. At market level, it immensely impedes research and selection processes and thus, investment activities, which is confirmed by different findings in the literature (i.e., Fuerst et al., 2015). Authors such as Adair et al. (2006), Falkenbach (2009), Lieser and Groh (2014) and Sadayuki et al. (2019) argue that opaque markets attract less capital inflows from foreign investors. Absent and incomplete market information often represents major entry barriers, either triggering high search costs or making an investment impossible from the start. Moreover, Wong et al. (2012) show that there is a link between transaction volume of real estate and the degree of information asymmetry between a buyer and a seller. The greater the imbalance, the more difficult it is for the parties to agree on a transaction price. Thus, the transaction process is negatively influenced by lacking or unevenly distributed information. In the given context, Chau and Wong (2016) further highlight that information asymmetry between landlord and tenant has a direct impact on the development of vacancy rates and rents.

At asset level, the picture is similarly obstructive. Occupants or sellers often know considerably more about both their properties and the surrounding environment than prospective buyers, resulting in information asymmetries between both sides (Firoozi et al., 2006; Wong et al., 2012). Rutherford et al. (2005) and Levitt and Syverson (2008) find that real estate agents sell their own houses for higher prices than similar ones they sell for their customers. The authors suggest that this is due to the informational advantage of the agents. Qiu et al. (2020) and Siebert and Seiler (2021) add that non-local homebuyers pay more than informed locals do. Thus, it seems intuitive that unevenly distributed information has a direct impact on individual transactions leading to adverse effects for one or both sides.

How is it possible that even in today's increasingly digital society, there continues to be a lack of transparency in many markets? The answer is admittedly rather simple: The present data and thus information situation is often insufficient. In the literature, many researchers report such data problems. For example, housing markets are known for repeatedly suffering from limited data availability (Rondinelli & Veronese, 2011). At the same time, also commercial and other sectors regularly perceive information as not available or incomplete (Devaney et al., 2017). Yet, prospects of improvement exist. While the degree of transparency still depends on the country or region considered, a general trend toward greater transparency has been observed in recent years. Eichholtz et al. (2011) attribute this to the internationalization of service providers, the rise of common investment vehicles, and more sophisticated and mature performance benchmarks, among others. By analyzing the Jones Lang Transparency Index, Newell (2016) confirm the above while additionally highlighting the importance of transparent market fundamentals, transaction

processes as well as of regulatory and legal issues. Many of these developments can be attributed to the continuously improving availability of data. In fact, one can note the emergence of several new data sources in the recent past. Whilst government data collections provide more and more macro- and microeconomic content, companies such as Real Capital Analytics or CoStar offer detailed information on capital flows, transactions and deal related details gathered by market participants. Asking data from mostly the residential sector can be retrieved from so-called multiple listing systems that source from real estate portals on which landlords advertise their properties. Although this list of emerging data sources is only an excerpt, the underlying potential becomes apparent. The commonly associated phenomenon is often simply referred to as the advent of Big Data (see e.g. Kok et al., 2017; DeLisle et al., 2020). Such data offers the prerequisites to more transparent markets. To access these markets however, it further requires appropriate techniques. To give one example, Arbia et al. (2019) show that data samples exceeding 70,000 observations overexert conventional statistical methods when using standard computer capabilities, which then lead to unreliable results. This finding represents most scientists' and practitioners' limits. Thus, in order to realize the full potential of data available today, there is a need to develop new and powerful statistical tools. As outlined below, the past decades have created a good foundation for this.

In the early nineteenth century, Legendre and Gauß introduced the first form of linear regression. Later, in the 1970s, adaptions and extensions such as generalized linear models with logistic regressions followed. When computational power started to improve in the 1980s, non-linear methods became feasible. While generalized additive models were able to handle non-linear relationships for the first time in history, classification and regression trees were developed simultaneously, marking the basis for what nowadays is called statistical machine learning. Since then, many so-called supervised and unsupervised machine learning (ML) methods have emerged, that are capable of processing large amounts of data.[2] The latter are often summarized under the term Artificial Intelligence (AI). However, what drives statistical or artificial learning to be advantageous? Conventional methods rely on predefined relationships in the data and thus may produce inaccurate results if these relationships do not fully correspond to reality. AI methods, on the contrary, have the ability to learn these relationships and rules from the data itself and thus, often achieve better results (Mullainathan & Spiess, 2017).

In one's everyday life, society already is confronted with AI, often without realizing it. Personalized web advertisements, spam filter for email accounts, webpages for translation

[2] For the complete history of statistical learning, see James et al. (2013).

and many more applications use AI techniques on a regular base. In the real estate industry, these methods have also found their areas of application. The Royal Institution of Charted Surveyors (RICS)[3] reported increased use of Big Data in conjunction with AI (RICS, 2017). Most of its use, however, can be found in the area of valuation, in which large data volumes enable ML techniques to precisely quantify real estate prices. Similarly, in the field of real estate research, authors such as Zurada et al. (2011), Mayer et al. (2019), and Bogin and Shui (2020) provide different insights on how ML methods are capable of identifying property prices. Aside from valuation topics, many problems and applications around the assessment of real estate investments are practicable yet immature.

The objective of this dissertation is to take a further step towards closing this gap. More precisely, this thesis aims to demonstrate how real estate investments can be assessed in times in which the use of Big Data together with AI methods evolves rapidly. The three individual papers show how increasing transparency allows a more thorough analysis of investment opportunities at market and asset level by using appropriate data and methods. With that, this dissertation supports and guides real estate stakeholders towards enhanced market knowledge, which allows for more in-depth and accelerated decision-making.

The first paper focuses on hidden investment mechanisms on market level by investigating European countries and cities. Cross-border capital flows are gradually part of many international real estate markets, determining prices and transactions. However, the factors influencing these flows have not been fully uncovered to date. In other words, it is still not entirely clear which market characteristics attract foreign investors. While existing literature has focused on the 'absolute attractiveness' of target markets such as sound economic growth or political stability, this paper breaks new ground. A unique dataset compiled of private transaction and real estate data as well as of public economic data enables the exploration whether 'relative attractiveness' of markets affect foreign capital flows by means of linear and nonlinear methods. The results indicate that relative risk premiums attract investors and that target markets are benchmarked against each other before an investment decision is made. This paper highlights that, by using appropriate transaction data, linear and nonlinear approaches are able to uncover market mechanisms.

The second paper links market to asset level and investigates how investment determinants such as rents can be better examined and predicted by adopting an increased focus on statistical learning methods. Hedonic regression models delivered price estimates depending on the underlying asset for many years in the past, whereas new technologies

[3] RICS is an important professional association for the real estate industry regularly providing industry standards, research and opinions.

such as AI estimate actual achievable prices more accurately. As presented before, these methods have already proven their worth and are demonstrably capable of supporting valuation procedures. While rents can be estimated with hedonic techniques, this paper is the first to investigate their derivation with statistical learning methods. Hence, it addresses the question whether ML methods are capable of precisely predicting residential rents. A large dataset on the city of Munich, covering residential rents as well as hedonic structural and locational characteristics of respective apartments, serves as basis for the analysis at hand. A comparison of a classic linear hedonic regression with several ML techniques reveals that the latter are able to outperform traditional methods. A second step involves the analysis of a residential portfolio and shows that ML methods also are applicable to detect underrent scenarios. This paper's findings encourage the application of statistical learning methodologies in order to increase understanding at market and asset level.

The third and last paper directs the view towards transparency regarding assets. While ML methods can outperform rather traditional linear methods at certain tasks[4], they are far from being a panacea. In fact, these methods are often criticized for their 'black box' character. As mentioned earlier, the power of ML arises from learning patterns and relationships from the data. This ability stems from complex algorithms which are able to produce consistent results but prohibit to see how the machine derives at such results (McCluskey et al., 2013). In the recent past, interpretable machine learning (IML) tools have been designed to analyze this exact inner working of such algorithms ex-post. In a first step, this paper uses residential rents from apartments in Frankfurt am Main together with structural and locational characteristics, allowing an ML algorithm to precisely predict rents. The second step involves the application of two IML methods, namely feature importance and feature effects. The former provides a ranking of the most important hedonic characteristics affecting rental estimation. The latter is able to highlight relationships between the rent and the individual characteristics. Subsequently, the IML methods are able to show on which economically comprehensible basis the ML algorithm came to its conclusion. Hence, this paper sheds light on how opening up a statistical learning methodology leads to enhanced economic understanding and increased transparency of the fundamental characteristics of real estate.

[4] ML methods are known to perform exceptionally well in predictive tasks, however, their inferential capabilities, which can be roughly translated to ‚why the prediction took place' are rather limited.

1.2 Research Questions

This section presents all important research questions, which are being addressed throughout the course of the different papers. While non-transparent markets are the basis and thus the unifying element of all papers, each one covers individual issues at a different level. A top-down approach serves as the framework for this dissertation. While the first paper focuses on questions at market level, the second paper investigates both, market and assets, and the third paper concentrates mainly on the asset level.

Paper 1: Do Cross-Border Investors Benchmark Commercial Real Estate Markets? Evidence from Relative Yields and Risk Premia for a European Investment Horizon

- Are international real estate markets transparent enough to investigate comparative investment behavior?
- Do investors benchmark international real estate investments, i.e., is the relative attractiveness of a target market decisive before market entry?
- Can relative attractiveness be expressed as relative yields and relative risk premia of real estate markets?
- Does benchmarking behavior induce a non-linear relationship between cross-border capital flows and relative yields and relative risk premia?
- Do transparent market mechanisms offer advantages for investors?

Paper 2: Rental Pricing of Residential Market and Portfolio Data – A Hedonic Machine Learning Approach

- Can artificial intelligence in combination with appropriate data contribute to a better understanding of real estate markets?
- Are machine learning methods suitable for estimating residential rents?
- Do hedonic machine learning methods provide a better predictive performance than conventional linear methods?
- Can conclusions be drawn on how investors currently determine their rents?
- If machine learning improves transparency, will investors benefit in re-letting scenarios?

Paper 3: Peeking inside the Black Box: Interpretable Machine Learning and Hedonic Rental Estimation

- Given the data available, can interpretable machine learning disclose rental mechanisms on asset level?
- Which hedonic characteristics are most important when determining rent?
- What are the economic and statistical relationships between these characteristics and the corresponding rent?
- Do machine learning algorithms estimate rents based on economically sound principles?
- Can interpretable machine learning methods represent algorithmic decision making in a comprehensible way?

1.3 Co-Authors, Submissions and Conference Presentations

The following overview provides information about co-authors, journal submissions, publication status and conferences presentations.

Paper 1: Do Cross-Border Investors Benchmark Commercial Real Estate Markets? Evidence from Relative Yields and Risk Premia for a European Investment Horizon

---

Authors:

Dr. Cay Oertel, Jonas Willwersch, Dr. Marcelo Cajias

Submission Details:

Journal: Journal of European Real Estate Research
Current Status: accepted (01/04/2020) and published in Volume 13, Issue 1 (02/03/2020)

Conference Presentations:

This paper was presented at:

- the 35th Annual Conference of the American Real Estate Society (ARES) in Paradise Valley, USA (2019)
- the 26th Annual Conference of the European Real Estate Society (ERES) in Cergy-Pontoise Cedex, France (2019)

Paper 2: Rental Pricing of Residential Market and Portfolio Data – A Hedonic Machine Learning Approach

Authors:

Dr. Marcelo Cajias, Jonas Willwersch, Felix Lorenz, Prof. Dr. Wolfgang Schäfers

Submission Details:

Journal: Real Estate Finance
Current Status: accepted for publication (04/23/2020)

Conference Presentations:

This paper was presented at:

- the 26th Annual Conference of the European Real Estate Society (ERES) in Cergy-Pontoise Cedex, France (2019)
- the 2nd Workshop "Artificial Intelligence and Finance" of the Center of Finance of the University of Regensburg, online (2020)
- the 37th Annual ARES Conference of the American Real Estate Society (ARES), online (2021)

This paper will be presented at:

- the 27th Annual Conference of the European Real Estate Society (ERES) in Kaiserslautern, Germany (2021) (submission accepted)
- the Verein für Socialpolitik (VfS) 2021 Annual Conference in Regensburg, Germany (2021) (submission accepted)

Paper 3: Peeking inside the Black Box: Interpretable Machine Learning and Hedonic Rental Estimation

Authors:

Felix Lorenz, Jonas Willwersch, Dr. Marcelo Cajias, Prof. Dr. Franz Fürst

Submission Details:

Journal: Real Estate Economics
Current Status: Under review (04/21/2021)

Conference Presentations:

This paper will be presented at:

- the 3rd Workshop "Artificial Intelligence and Finance" of the Center of Finance of the University of Regensburg, online (2021)
- the 27th Annual Conference of the European Real Estate Society (ERES) in Kaiserslautern, Germany (2021) (submission accepted)

1.4 References

Adair, A., Allen, S., Berry, J., & McGreal, S. (2006). Central and Eastern European property investment markets: issues of data and transparency. *Journal of Property Investment & Finance*, *24*(3), 211–220.

Andonov, A., Kok, N., & Eichholtz, P. (2013). A Global Perspective on Pension Fund Investmentsin Real Estate. *The Journal of Portfolio Management*, *39*(5), 32–42.

Arbia, G., Ghiringhelli, C., & Mira, A. (2019). Estimation of spatial econometric linear models with large datasets: How big can spatial Big Data be? *Regional Science and Urban Economics*, *76*, 67–73.

Bogin, A. N., & Shui, J. (2020). Appraisal Accuracy and Automated Valuation Models in Rural Areas. *The Journal of Real Estate Finance and Economics*, *60*(1-2), 40–52.

Bond, S. A., Hwang, S., Lin, Z., & Vandell, K. D. (2007). Marketing Period Risk in a Portfolio Context: Theory and Empirical Estimates from the UK Commercial Real Estate Market. *The Journal of Real Estate Finance and Economics*, *34*(4), 447–461.

Bone-Winkel, S., Focke, C., & Schulte, K.-W. (2016). Begriff und Besonderheiten der Immobilie als Wirtschaftsgut. In K.-W. Schulte, S. Bone-Winkel, & W. Schäfers (Eds.), *Immobilienökonomie* (5th ed., pp. 4–24). De Gruyter Oldenbourg.

Bone-Winkel, S., Schulte, K.-W., Schulte, K.-M., & Pfrang, D. (2016). Bedeutung der Immobilienwirtschaft. In K.-W. Schulte, S. Bone-Winkel, & W. Schäfers (Eds.), *Immobilienökonomie* (5th ed., pp. 26–43). De Gruyter Oldenbourg.

Chau, K. W., & Wong, S. K [S. K.] (2016). Information Asymmetry and the Rent and Vacancy Rate Dynamics in the Office Market. *The Journal of Real Estate Finance and Economics*, *53*(2), 162–183.

Crosby, N., & McAllister, P. (2004). *Liquidity in commercial property markets: Deconstructing the transaction process* [Working Paper]. The University of Reading Business School, Reading.

DeLisle, J. R., Never, B., & Grissom, T. V. (2020). The big data regime shift in real estate. *Journal of Property Investment & Finance*, *38*(4), 363–395.

Devaney, S., McAllister, P., & Nanda, A. (2017). Determinants of transaction activity in commercial real estate markets: evidence from European and Asia-Pacific countries. *Journal of Property Research*, *34*(4), 251–268.

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*(1-2), 152–173.

Falkenbach, H. (2009). Market selection for international real estate investments. *International Journal of Strategic Property Management*, *13*(4), 299–308.

Firoozi, F., Hollas, D., Rutherford, R [Ronald], & Thomson, T. (2006). Property Assessments and Information Asymmetry in Residential Real Estate. *Journal of Real Estate Research*, *28*(3), 275–292.

Fuerst, F., Milcheva, S., & Baum, A. (2015). Cross-border capital flows into real estate. *Real Estate Finance*, *31*(3), 103–122.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 103). Springer New York.

Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *The Journal of Portfolio Management*, *43*(6), 202–211.

Levitt, S. D., & Syverson, C. (2008). Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions. *Review of Economics and Statistics*, *90*(4), 599–611.

Lieser, K., & Groh, A. P. (2014). The determinants of international commercial real estate investment. *The Journal of Real Estate Finance and Economics*, *48*(4), 611–659.

Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research*, *12*(1), 134–150.

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, *30*(4), 239–265.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Newell, G. (2016). The changing real estate market transparency in the European real estate markets. *Journal of Property Investment & Finance*, *34*(4), 407–420.

Qiu, L., Tu, Y., & Zhao, D. (2020). Information asymmetry and anchoring in the housing market: a stochastic frontier approach. *Journal of Housing and the Built Environment*, *35*(2), 573–591.

RICS. (2017). *The Future of Valuations*. Royal Institute of Chartered Surveyor.

Rondinelli, C., & Veronese, G. (2011). Housing rent dynamics in Italy. *Economic Modelling*, *28*(1-2), 540–548.

Rutherford, R. C., Springer, T. M., & Yavas, A. (2005). Conflicts between principals and agents: evidence from residential brokerage. *Journal of Financial Economics*, *76*(3), 627–665.

Sadayuki, T., Harano, K., & Yamazaki, F. (2019). Market transparency and international real estate investment. *Journal of Property Investment & Finance*, *37*(5), 503–518.

Schulte, K.-W., Rottke, N., & Pitschke, C. (2005). Transparency in the German real estate market. *Journal of Property Investment & Finance*, *23*(1), 90–108.

Siebert, R. B., & Seiler, M. J. (2021). Why Do Buyers Pay Different Prices for Comparable Products? A Structural Approach on the Housing Market. *The Journal of Real Estate Finance and Economics.* Advance online publication.

Wong, S. K [Siu Kei], Yiu, C. Y., & Chau, K. W. (2012). Liquidity and Information Asymmetry in the Real Estate Market. *The Journal of Real Estate Finance and Economics*, *45*(1), 49–62.

Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research*, *33*(3), 349–387.

2 Do Cross-Border Investors Benchmark Commercial Real Estate Markets? Evidence from Relative Yields and Risk Premia for a European Investment Horizon

2.1 Abstract

The purpose of the study is to introduce a new perspective on determinants of cross-border investments in commercial real estate, namely the relative attractiveness of a target market. So far, the literature has analyzed only absolute measures of investment attractiveness as determinants of cross-border investment flows. The empirical study uses a classic OLS estimation for a European panel data set containing 28 cities in 18 countries, with quarterly observations from Q1/2008 – Q3/2018. After controlling for empirically proven explanatory covariates, the model is extended by the new relative measurement based on relative yields/cap rates and relative risk premia. Additionally, the study applies a generalized additive mixed model (GAMM), to investigate a potentially nonlinear relationship. The study finds on average a c.p., statistically significant lagged influence of the proxy for relative attractiveness. Nonetheless, a differentiation is needed; relative risk premia are statistically significant, whereas relative yields are not. Moreover, the GAMM confirms a nonlinear relationship for relative risk premia and cross-border transaction volumes. The results are of interest for both academia and market participants as a means of explaining cross-border capital flows. The existing knowledge on determinants is expanded by relative market attractiveness, as well as an awareness of nonlinear relationships. Both insights help to comprehend the underlying transaction dynamics in commercial real estate markets. Whereas the existing body of literature focuses on absolute attractiveness to explain cross-border transaction activity, this study introduces relative attractiveness as an explanatory variable.

Keywords – Cross-Border Real Estate Investment, Direct Property Investment, European Real Estate, Panel Data, Nonlinear Regression

Acknowledgments: The authors especially thank PATRIZIA AG for contributing to this study. All statements of opinions are those of the authors and do not necessarily reflect the opinion of PATRIZIA AG or its associated companies.

2.2 Introduction

Direct cross-border investments in commercial properties have increased steadily over the past two decades. Accordingly, the related research and market participants have demonstrated an increased interest in understanding the determinants of capital flows across national borders.

Institutional economics theory defines the attractiveness of a target investment market as a function of its socio-economic environment and institutional framework (Fuerst et al., 2015). In line with this theory, Lieser and Groh (2014) provided empirical evidence of the importance of economic growth, demographics, urbanization or political stability of a particular country. However, other authors highlight the importance of additional factors on cross-border capital flows. Yet, the literature has described the attractiveness of an investment location solely with absolute measures of potential determinants.

The present article introduces a new approach to explaining inflowing cross-border capital into real estate market, namely relative attractiveness for a target investment horizon. As opposed to previous studies, it sheds light on whether cross-border investors benchmark investment opportunities against each other. More precisely, the study investigates whether relative attractiveness in the form of relative yields or relative risk premia determines the capital allocation of investors. In this context, the analysis concentrates on European real estate markets, as classic prime European investment markets represent rather homogenous, substantially economically integrated by the EU and geographically densely located competing investment markets. Thus, relative attractiveness appears to be a potential driver, but solely for geographical reasons. At the same time, as outlined by Devaney et al. (2017a), data availability issues in Europe especially at the city level have hampered research on cross-border transaction activity. Consequently, work at this level requires new empirical evidence.

The paper is structured as follows: A comprehensive literature review builds the foundation for the empirical study. The essence is the existing body of literature on the one hand, while legitimating the approach of introducing relative attractiveness as a further driver of cross-border investment activity on the other. The section concludes with a statement of the hypotheses for the empirical work. Subsequently, the paper outlines the data set and research design, including the new target variables for measuring the relative attractiveness of a city. It also reports the descriptive statistics. Since macroeconomic models on cross-border investment activities in real estate markets are subject to severe methodological challenges and data availability issues, the variable selection process and

the econometric approaches are discussed extensively. Afterwards, the empirical results are presented, and some conclusions drawn.

2.3 Related Literature and Hypotheses Derivation

Several studies have tried to identify common determinants of cross-border real estate investment in the view of various investor types and investment styles. From a portfolio point of view, national and regional diversification benefits are often perceived as one of the driving forces behind international capital allocation in real estate. Amongst others, Sirmans and Worzala (2003) and Holsapple et al. (2006) argued that the diversification of country-specific economic drivers is decisive. The abovementioned literature on diversification, however, often suffers from data unavailability on investors who cause the transaction flows. Thus, the relationship between investor, relevant portfolio and investment flow cannot be established. Accordingly, a growing body of literature has focused on investor-unrelated and general institutional and macroeconomic determinants of cross-border investment flows. Hence, several studies have investigated investment drivers and barriers on global and regional levels.

A comprehensive empirical study on the economic and institutional environment was conducted by Lieser and Groh (2011). First, they defined six relevant areas for cross-border investments, namely economic activity, real estate investment opportunities, the depth and sophistication of capital markets, investor protection and the legal framework, administrative burdens and regulatory limitations, as well as the socio-cultural and political environment. In a second step, they quantified the attractiveness of countries via a composite index approach. In a second paper, Lieser and Groh (2014) analyzed which of these country characteristics impact on foreign real estate investment volumes. After investigating 47 countries, they illustrated a significant relationship between foreign real estate investment activity as the dependent variable and real estate investment opportunities, the depth and sophistication of capital markets, investor protection and the legal framework, administrative burdens and regulatory limitations as independent variables. In line with this study, Devaney et al. (2017a) found that in European and Asian Pacific countries, the size and wealth of a country, the specific country risk, and property rights, as well as the performance of the real estate markets, mainly determine transaction activity.[5]

[5] Transaction activity was measured from turnover rates of the total transaction volume taking foreign and domestic investments together.

A second stream of papers narrowed the geographic focus and carried out empirical studies on national or city-level determinants. Chin et al. (2006) and Pi-Ying Lai and Fischer (2007) identified patterns in Asian regions and cities. They highlighted that political stability and legal regulations, as well as sound financial and economic structures, and the strength and stability of the current economy, are of major importance for investments in these areas. He and Zhu (2010) added that aside from a favorable institutional environment, Chinese cities and their real estate markets attract capital through population and market size. For Eastern Europe, McGreal et al. (2001) argued that foreign real estate investment activity can be affected negatively, especially by non-transparency, overall economic conditions, corruption, and bureaucracy. Salem and Baum (2016) found that foreign money flows into real estate markets in the Middle East and northern African countries are mainly influenced by political stability. Devaney et al. (2017b) investigated transaction activity in U.S. metropolitan office markets. Economic growth and market size were positively related to turnover rates, whereas vacancy rates and risk showed a negative relationship.

The studies presented thus far indicate that the institutional framework and the macroeconomic conditions shape cross-border investment. However, real-estate-related factors also influence cross-border capital flows, since investment success is not only linked to country characteristics, but also to the underlying real estate market and the property itself. A number of authors have therefore included various proxies of real estate markets into their investigations. Ford et al. (1998) found that market activity and rent levels of US real estate markets determine foreign investment behavior. Moreover, according to Laposa and Lizieri (2005) office construction attracts foreign investment in Eastern Europe. For China, He and Zhu (2010) showed that aside from satisfactory demographic conditions, already invested foreign capital attracts both foreign developers as well as more cross-border investors. In addition, Rodríguez and Bustillo (2010), Gholipour Fereidouni and Ariffin Masron (2013) and Farzanegan and Fereidouni (2014) observed market-specific property prices to be influential. Interestingly, Gholipour Fereidouni and Ariffin Masron (2013) found real estate market transparency to be an important determinant for foreign investors, but Farzanegan and Fereidouni (2014) did not confirm this finding. Fuerst et al. (2015) established a positive relationship between market liquidity and cross-border capital inflows, since the ability to sell properties increases. Devaney et al. (2017a) noted a negative relationship between office vacancy rates and turnover rates. With particular respect to property characteristics, Devaney et al. (2018) demonstrated that cross-border investors in U.S. gateway cities favor large and new buildings close to CBD locations.

To gauge investment potential and to explain capital flows, risk characteristics such as the previously documented institutional, macroeconomic and real-estate-related variables constitute crucial considerations. Nonetheless, income opportunities, which may additionally influence investors, can be assessed by an ex ante analysis of yields and pricing. A common method of early real estate investment evaluation is the capitalization (cap) rate. It is usually computed as the ratio of a property's net operating income to its price and therefore serves as an opportunity to compare assets and markets. When assessing the main determinants of cap rates, the literature refers to the Gordon-growth model (see e.g. McAllister & Nanda, 2016):

$$Cap.\ Rate = Nominal\ risk\text{-}free\ rate + \ Risk\ premium\ - Income\ growth \quad (1)$$

The nominal risk-free rate is often approximated by a long-term government bond, whereas the risk premium marks the difference between the government bond and an individual asset yield. The income growth measures the growth of rents or net operating income. Research companies, brokers and other market participants regularly provide cap rates and therefore enable investors to measure and compare *ex ante* investment potential. To the best knowledge of the authors, only a little research has investigated the cap rate, or related initial yields, and investment flow relationship, even though a direct relationship between both seems reasonable.

With respect to foreign investment, McAllister and Nanda (2016) and Oikarinen and Falkenbach (2017) detected that foreign capital decreases cap rates. For the present study, the subsequent question of whether the reverse relationship holds true and that cap rates impact investment activity has barely been analyzed. Considering American real estate, Ford et al. (1998) argued that foreign investors react to changes in cap rates. Considering turnover rates in international office markets, Devaney et al. (2019) could not prove that cap rates influence general investment activity. To shed more light on this topic, we suggest a new approach to analyzing cap rates and cross-border flow dynamics. So far, potential and actual determinants in the literature were taken into account in order to display the absolute attractiveness of real estate markets. However, we are interested in whether cross-border investors not only look at specific market characteristics representing the absolute attractiveness, but also benchmark certain key determinants such as initial yields against neighboring and competing markets.

More precisely, when cross-border investors choose among target locations, we expect them to look for outperformance opportunities within a predefined investment horizon. Thus, investors search for relative attractiveness among a given set of markets at the time of deploying capital. A straightforward way to evaluate outperformance is to benchmark

key metrics such as yields and risk compensation. Therefore, the present study analyzes empirically whether cap-rate-based relative yields and relative risk premia contribute to the relative attractiveness affecting cross-border investments. This leads to the first hypothesis:

Hypothesis 1: *The relative attractiveness of real estate markets affects cross-border capital inflows.*

The vast majority of the abovementioned articles use classic linear models, based mainly on panel models and OLS estimations. The present study aims at contributing to the existing body of literature by relaxing the assumption of a constant effect of the explanatory variables on cross-border investment volumes, as proposed by Devaney et al. (2017a). Possible reasons are potential investor heterogeneity with regard to risk appetite, differences in funding and investor herding behavior. Inspired by the real estate literature on hedonic pricing models (Cajias & Ertl, 2018), the present paper uses a generalized additive mixed model (GAMM). Accordingly, a potential nonlinear relationship between the variable of interest and the dependent variable will be assessed, addressing the following hypothesis:

Hypothesis 2: *The relative attractiveness of real estate markets has a nonlinear relationship with cross-border capital inflows*.

In order to provide insight into the abovementioned hypotheses, the following sections describe the data and the applied methodology. Subsequently the empirical results are presented, which lay the foundation for the assessment of hypotheses. The latter is stated in the conclusion section.

2.4 Data, Sample Description and Methodology

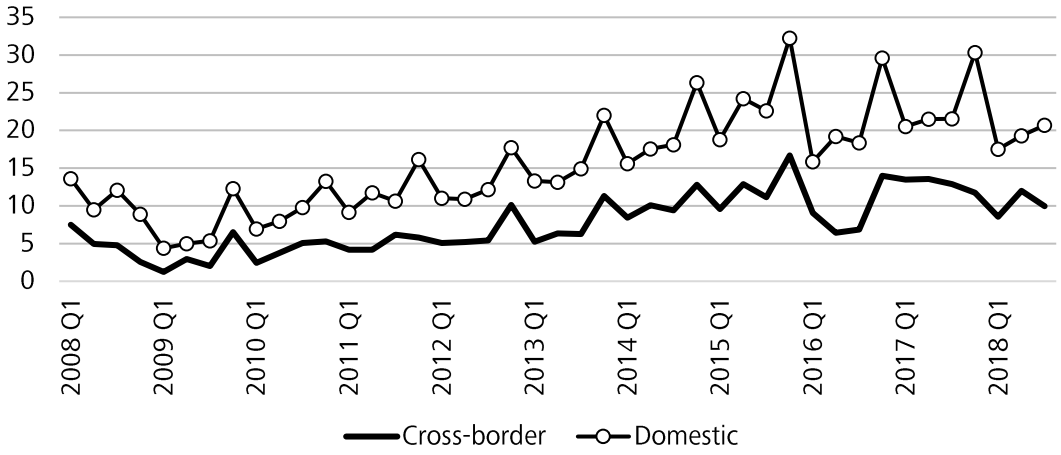
2.4.1 Data Source of the Dependent Variable

The analyzed data sample contains 28 European cities (n = 28) across 18 countries [2], with quarterly observations of transaction volumes for office properties from Q1 2008 to Q3 2018 (t = 43). The data is from various data providers. The dependent variable covers quarterly aggregated cross-border transaction volumes of office buildings provided by Real Capital Analytics, Inc. (RCA). RCA is a data-specialist that tracks commercial real estate transactions worldwide including single properties, portfolios and units which mainly consist of commercial real estate. The company applies a standard price floor of 5 Mio. EUR or greater in Europe to consider them in its statistics. Moreover, it sources information

about transactions from a variety of investors, brokerage firms, media companies and others. RCA labels a transaction as "foreign" or "cross-border" if the buyer's or the major capital partner's headquarter is not situated in the same country as the property. The buyers typically consist of institutional (equity and pension funds, insurances, banks, etc.), listed (REITs, REOCs and listed funds) and private investors (high net worth individuals, non-traded REITs, developers/owners/operators) as well as others (governments, corporates, non-profit, educational and religious users).

Figure 2.1 depicts aggregated quarterly investment volumes of cross-border and domestic investors in the European sample from Q1/2008 to Q3/2018. A visual inspection indicates a positive correlation of both capital types showing a trough in Q1/2009 and a peak in Q4/2015. However, the domestic volumes are continuously greater than the cross-border ones.

Figure 2.1: Agg. European Real Estate Investment in Bn. Euro (Q1/2008 – Q3/2018)



The use of RCA data itself is increasing in the real estate literature. Nonetheless, with respect to the measurement of investment activity, there is a debate on what constitutes the right measure to incorporate these flows into econometric models. Devaney et al. (2017b) argue that pure transaction volumes can be driven not only by activity, but also by price inflation. Instead, they suggest turnover rates measured as the appropriate value of the traded properties purchased by domestic and foreign investors, divided by the value of all properties in the market, so as to more accurately capture investment activity. We cannot follow this procedure, since the information on how much of the value of all properties in the market belongs to foreign investors is not accessible. Instead, we stick to the common transaction volumes, but control for inflation as proxy for the overall asset price development of the target market.

2.4.2 Research Design, Definition and Selection of the Explanatory Variables

Since we aim at replicating average investor behavior, the markets specified in the panel data set appear to be the key ones for global investors looking for investment opportunities in Europe (PricewaterhouseCoopers & Urban Land Institute, 2019). However, a decisive methodological point needs to be highlighted: By defining the panel data set as such, we model the included cities across Europe as a closed investment horizon in which only the markets specified compete for inflowing capital. Thus, the relative attractiveness relates to the benchmark of these investment locations only, assuming other markets beyond this horizon to be irrelevant for cross-border investors. However, the study expects these investors predominantly to target the specified main investment markets.

Since investors typically compare *ex ante* profitability metrics when looking for investment opportunities, prime initial yields form the base of relative attractiveness. The term *relative* indicates the comparison of one market with all others in the sample, which essentially creates a benchmark. To describe the relative attractiveness, we decide to measure the impact of two variables on the aforementioned foreign capital flows: relative yields as well as relative risk premia. The formula is thereby based on the relative return measure of MSCI (2019), which is frequently used, for example, in performance analysis. As the first variable, the relative net mean yield ($RNMY_{i,t}$) is defined as follows (see equation 2):

$$RNMY_{i,t} = \left(\frac{1 + PY_{i,t}(1 - CIT_{i,t})}{1 + \frac{1}{n}\sum_{i=1}^{n} PY_{i,t}(1 - CIT_{i,t})} - 1 \right) * 100 \quad (2)$$

$PY_{i,t}$ denotes the prime initial yield of best located assets in city i for period t.[6] The data stem from CoStar. Additionally, $CIT_{i,t}$ stands for the average corporate income tax of the respective country, obtained from the OECD. The taxation is additionally introduced, since the yield of an investment will eventually be capitalized as an *ex post* return and thus taxed. The domestic net yield in the numerator is calculated by multiplying $PY_{i,t}$ by one minus the specified tax. Although taxation issues are often neglected in related studies, we incorporate them in order to account for taxation-driven investment decisions and since we expect the average investor to be affected by a homogenous average tax rate, and so

[6] Although our sample presumably includes not only core investors, we do not consider average yields or cap rates. Instead, we use prime yields since they are from a cost and effort perspective relatively easy to obtain in early market research. Cross-border investors have higher search costs (see McAllister and Nanda, 2016b), suggesting that prime yields offer an inexpensive way to obtain an early market indication. In addition, several practitioners informed us that prime yields are often included in order to assess investment potential in foreign markets.

the focus is on these net yields.[7] The denominator provides the average net prime yield, which is defined by the mean of the net prime yields across all individuals. Thus, the denominator can also be interpreted as the abovementioned benchmark. Excess attractiveness of a city in comparison to the mean is c.p. expected to trigger inflowing capital.

Whereas a relative yield benchmarks the sole real-estate-related income potential, investors may be also affected by how much related risk premium a real estate market offers as an excess in relation to the country-specific risk-free alternative. In other words, is there an excess yield, justifying the capital allocation in a property market? Since investors expect risk premia when allocating capital to a risky asset, we specify the relative net mean risk premium ($RNMRP_{i,t}$) as such:

$$RNMRP_{i,t}^{GOV(5|10)} = \left(\frac{1+(PY_{i,t}(1-CIT_{i,t})-GOV_{i,t})}{1+\frac{1}{n}\sum_{i=1}^{n}(PY_{i,t}(1-CIT_{i,t})-GOV_{i,t})} - 1 \right) * 100 \quad (3)$$

The nominal risk free rate is approximated by long-term country-specific government bonds. To account for different investment horizons, we include 10 year government bonds to obtain long-term and 5 year government bonds for medium-term risk premia. The bond data is from Thomson Reuters Datastream. For both variables, the calculation works as follows: if Amsterdam's net domestic net yield or risk premium respectively was 4.5% in Q2/2013, and the European mean was 4.0% in that quarter, the city's relative attractiveness was 0.4% times 100 above the benchmark, which equals 4.0 in the data set.

In summary, the numerator of the two target variables represents a city's absolute attractiveness. The denominator denotes the constructed benchmark. For both relative attractiveness measures, a ratio above (below) 0 shows relative more (less) attractiveness in the respective city than can be found on European average. The expected signs for both measures of relative attractiveness are positive.

The remaining covariates are macroeconomic and real-estate-related controls, which are in line with the literature described above. In the estimation procedure, all controls are considered in absolute values (e.g. the GDP growth is measured by the value of the individual itself), meaning that the relative form only applies to the relative attractiveness measurements. Table 2.1 summarizes the macroeconomic and real estate controls.

[7] Yet, we are aware of the fact that especially in Europe certain fund and firm structures prevent taxation payments for real estate investments. Since we do not know which structures are implemented in the analyzed transactions, we include corporate income taxes in our models.

Table 2.1: Control Variables

Variable	Description	Proxy for	Level	Source
GDP growth	Amongst others, Lieser and Groh (2014) argue that a sound and healthy economy is a driving factor for direct real estate investments. Hence, we control for economic stability by including quarter-on-quarter GDP growth in the econometric analysis.	Economic stability	Country	OECD
CPI growth	Inflation is added in order to control for price movements with respect to cross-border transaction volumes. Consequently, quarter-on-quarter CPI growth serves as a control variable that adjusts for changes in the dependent variable due to market conditions.	Asset price inflation	Country	OECD
Unemployment rate	Employment is often perceived as another indicator of economic health and success. Fuerst et al. (2015) state that foreign investors are attracted by good employment conditions. Thus, we use the unemployment rate to capture the labor market and income situation.	Labor market and income	Country	OECD
Global Competitiveness Index (GCI)	In line with previous literature, a condensed country risk measure is central when choosing among international investment opportunities. We decide to control for country risk by using the GCI. The construction of the index is based upon twelve core areas, which cover e.g. institutions, infrastructure, the adoption of information and technologies and others (World Economic Forum, 2018).[8]	Country risk	Country	World Economic Forum
Vacancy	Office vacancy serves as an indication of the current state of demand in a real estate market. According to Devaney et al. (2017b), vacancy captures conditions in the space market.	Office demand	City	CoStar
Stock	Stock indicates the available office floor space and therefore shows the size of the market and / or the building activity. We incorporate it to control for the office supply.	Office supply	City	CoStar
Prime rent growth	Year-on-year prime rent growth shows the income growth potential of prime office buildings in the respective market.[9]	Income expectations	City	CoStar

[8] Some researchers such as Devaney et al. (2017b) use government and or corporate bonds spreads to control for country risk. To avoid multicollinearity, we cannot include this proxy, since the second target variable relative risk premium is constructed based on government bonds. Additionally, the JLL Global Real Estate Transparency Index series may be an alternative proxy to control for country specific risk factors. Nonetheless, the specified index was not used, because the study incorporates a macroeconomic index to account for effects on a broader and national economic level.

[9] We also controlled for non-prime rent growth. The results stayed robust but were not reported.

2.4.3 Descriptive Statistics

The following section reports the univariate analysis for the above-mentioned constituents of the data set. Table 2.2 displays the descriptive statistics of the dependent and target variables as well as the covariates.

Table 2.2: Descriptive Statistics

Variable	n	Unit	Mean	SD	Min.	Max.
Dependent Variable						
Cross-border transaction volume	1204	T €	277,195	705,530	0	7,384,621
Macroeconomic variables						
GDP growth	1204	%	0.298	0.957	-6.842	9.928
CPI growth Δ	1176	%	-0.007	0.868	-5.035	4.575
Unemployment rate Δ	1176	%	-0.045	1.192	-8.997	8.996
GCI	1204	Index	5.168	0.439	4.153	5.858
Real-estate-related variables						
Vacancy	1204	%	10.581	4.218	2.310	25.474
Stock Δ	1176	sqm	347,905	454,964	-1,199,068	3,797,188
Prime Rent Growth	1204	%	0.926	7.725	-54.930	48.072
1. Target Variable						
RNMY	1204	%	0.000	24.793	-44.784	134.013
2. Target Variable						
RNMRP 10y	1204	%	0.000	8.985	-29.913	19.943
RNMRP 5y	1204	%	0.000	8.956	-29.858	19.999

Notes: Δ indicates the first differences of the variable. Sqm stands for square meters.

From the descriptive statistics table, the need for a natural logarithm transformation of cross-border transaction volume and the stock is apparent, because variables vary substantially with regard to their absolute values. Since the origin of the investment volumes is not available, we are unable to control for exchange rate stability. Yet, we incorporate all monetary values in Euros (€) to form a uniform currency base. Additionally, a correlation matrix provides insights into the common movement of the covariates (see Table 2.3):[10]

[10] Correlations between timely lagged covariates are not reported. However, the indication of the contemporary realizations sufficiently reveals the potential of crucial correlations.

Table 2.3: Correlation Matrix

		1	2	3	4	5	6	7	8	9	10	11
1	Cross-border transaction volume	1.000										
2	GDP growth	0.057	1.000									
3	CPI growth Δ	0.005	0.037	1.000								
4	Unemployment rate Δ	-0.014	-0.063	0.023	1.000							
5	GCI	0.122	0.045	0.014	0.021	1.000						
6	Vacancy	-0.165	0.048	0.001	0.003	-0.451	1.000					
7	Stock Δ	0.095	-0.116	0.000	0.003	-0.269	0.104	1.000				
8	Prime rent growth	0.033	0.076	-0.018	-0.014	0.156	-0.197	-0.215	1.000			
9	RNMY	-0.233	-0.003	-0.013	-0.005	-0.621	0.552	0.279	-0.183	1.000		
10	RNMRP 10y	-0.103	0.080	-0.005	-0.011	0.242	0.063	-0.119	0.027	0.244	1.000	
11	RNMRP 5y	-0.102	0.080	-0.006	-0.011	0.244	0.060	-0.131	0.031	0.234	0.999	1.000

Note: Δ indicates the first differences of the variable.

In line with previous research, we define absolute values greater than 0.25 define as threshold for any econometric issues. The target variable RNMY shows critical correlations with the GCI, vacancy and stock. Among the controls, GCI yields correlations with stock and vacancy below -0.25. Stock and vacancy show a correlation above 0.25.

Even though we estimate a base model with all correlated variables, we try to control for multicollinearity by comparing the results of the specified model with the results of model variations. These variations individually exclude one of the correlated variables.

Lastly, since panel data models may be subject to potential non-stationarity, we carried out panel unit root test to check for econometric distractions (see Table 2.7 in the Appendix). For those covariates which suffer from non-stationarity, we used a differencing procedure, in order to generate a stationary time series. After a first differencing, we observe stationary covariates, denoted Δ(x).

2.4.4 Methodology

To assess the outlined hypotheses, two different methodologies are applied: Pooled OLS, as well as a GAMM. Firstly, an OLS model estimates the linear predictors to evaluate the first hypothesis. The model specification yields the following equation (4):

$$\ln(v_{i,t}) = \beta_{i,t-k} m_{i,t-k} + \beta_{i,t-k} r_{i,t-k} + \beta_{i,t-k} relative\ attractiveness_{i,t-k} + \beta_t time_t + \beta_i city_i + \varepsilon_{i,t} \quad (4)$$

Here, the natural logarithm of the cross-border transaction volume $\ln(v_{i,t})$ observed in a market i in quarter t is a function of the abovementioned domestic macroeconomic variables captured in vector $m_{i,t-k}$, real-estate-related variables in the vector $r_{i,t-k}$ and

one of the measurements for relative attractiveness, captured in vector $relative\ attractiveness_{i,t-k}$.

$$v_{i,t} = Cross - border\ transaction\ volume \tag{5}$$

$$m_{i,t-k} = \begin{cases} GDP\ growth \\ \Delta(CPI\ growth) \\ \Delta(Unemployment\ rate) \\ Global\ Competitive\ Index \end{cases} \tag{6}$$

$$r_{i,t-k} = \begin{cases} \text{Vacancy} \\ \ln(\Delta(Stock)) \\ Prime\ rent\ growth \end{cases} \tag{7}$$

$$relative\ attractiveness_{i,t-k} = \begin{cases} RNMY \\ RNMRP\ 10\ y \\ RNMRP\ 5\ y \end{cases} \tag{8}$$

To control for temporal heterogeneity, we use dummy variables labeled as $time$ for each period of the sample. The base year is 2008. City heterogeneity is captured in all models by including $city$ dummies, with Frankfurt representing the reference, considering its approximate geographic European centrality within the sample. $\varepsilon_{i,t}$ represents the error which is not captured in the model.

Since real estate markets are prone to timely delayed effects, we estimate lagged terms up to four quarters for each included covariate (k = 4). Some authors have addressed the influence of transaction activity on cap rates (see e.g. McAllister & Nanda, 2016 and Oikarinen & Falkenbach, 2017) who ran their econometric analysis as differently to our procedure). Accordingly, we check our data sample by first carrying out a Granger causality test to evaluate a potentially inverse relationship between the dependent and the target variables.

Even though the abovementioned pooled OLS estimation procedure is capable of testing the first economic hypothesis by isolating a linear c.p. effect on average across the data set of the relative attractiveness, the second hypothesis requires a different approach. To further explore potential nonlinearity we use a second and semiparametric model. The GAMM allows nonlinear as well as linear relationships of the covariates (see equation 9):

$$\ln(v_{i,t}) = \beta_{i,t-k} m_{i,t-k} + \beta_{i,t-k} r_{i,t-k} + f_{i,t-k}(relative\ attractiveness_{i,t-k}) + \beta_t time_t + \beta_i city_i + \varepsilon_{i,t} \tag{9}$$

Here, the function $f_{i,t-k}$ denotes the smoothing function for the relative attractiveness proxy, which is used to check for potential nonlinearity. Thus, we do not estimate a linear predictor for the variables of interest, in contrast to the OLS model. Since the potential nonlinear behavior of the macroeconomic and real-estate-related controls is of minor interest, we introduce a smoothing function only for the target variables. The number of knots is set equal to 20, in order to allow for enough flexibility of the smoothing terms.

2.5 Empirical Results

Due to potential inverse relationships between the relative attractiveness proxies and capital flows, we firstly conduct a Granger causality test to detect potential simultaneity bias in our sample (see Table 2.4):

Table 2.4: Granger Causality Test / Inverse Relationship

Dependent	Independent	F statistic	p-value	
RNMRP 10y	ln(Cross-border transaction volume)	0.6708	0.6123	
RNMRP 5y	ln(Cross-border transaction volume)	0.6542	0.6240	
RNMY	ln(Cross-border transaction volume)	2.1227	0.0753	*

As displayed above, we find strong empirical proof against a potential inverse relationship between cross-border volumes and both RNMRPs. Only for the RNMY is the relationship inversely statistically significant at the 10 percent level and may therefore cause simultaneity. The standard methodical procedure for accounting for simultaneity is to use an instrument variable approach such as two stage least squares. However, since target variables are of particular interest – unlike controls – we do not search for instruments, but emphasize the potential presence of simultaneity bias with regard to the RNMY.

To test the first hypothesis, we run pooled OLS estimations. The results can be found in Table 2.5. For each of the three target variables, we estimate the same four specifications. The base model includes all control variables, whereas the second, third and fourth models individually exclude the variables GCI, vacancy and stock, to check for robustness. The selected variables were systematically exchanged, due to the findings within the correlation matrix and to account for potential multicollinearity.

Table 2.5: Pooled OLS Estimation Results

Dependent variable: ln (cross-border transaction volume)												
Model	Model 1	Model 1.1	Model 1.2	Model 1.3	Model 2	Model 2.1	Model 2.2	Model 2.3	Model 3	Model 3.1	Model 3.2	Model 3.3
Controls:												
Macroeconomic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Real estate	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Excluded control	None	GCI	Vacancy	Stock	None	GCI	Vacancy	Stock	None	GCI	Vacancy	Stock
Target variables:												
RNMY	-0.094	-0.091	-0.086	-0.102								
	(0.082)	(0.082)	(0.087)	(0.083)								
RNMY (-1)	0.086	0.083	0.048	0.099								
	(0.110)	(0.110)	(0.117)	(0.111)								
RNMY (-2)	0.088	0.086	0.111	0.083								
	(0.102)	(0.102)	(0.107)	(0.103)								
RNMY (-3)	-0.163	-0.158	-0.175	-0.180								
	(0.113)	(0.113)	(0.114)	(0.113)								
RNMY (-4)	0.087	0.083	0.093	0.108								
	(0.075)	(0.075)	(0.072)	(0.076)								
RNMRP 10y					0.164	0.158	0.203 *	0.183				
					(0.113)	(0.112)	(0.111)	(0.111)				
RNMRP 10y (-1)					-0.222	-0.230	-0.229	-0.217				
					(0.157)	(0.155)	(0.159)	(0.159)				
RNMRP 10y (-2)					0.335 **	0.339 **	0.306 **	0.333 **				
					(0.155)	(0.152)	(0.155)	(0.155)				
RNMRP 10y (-3)					-0.265	-0.257	-0.247	-0.260				
					(0.173)	(0.169)	(0.173)	(0.174)				
RNMRP 10y (-4)					0.108	0.086	0.116	0.093				
					(0.140)	(0.136)	(0.139)	(0.142)				

Table 2.5 continued from previous page

RNMRP 5y									0.111		0.106		0.141		0.129	
									(0.117)		(0.117)		(0.117)		(0.116)	
RNMRP 5y (-1)									-0.149		-0.161		-0.149		-0.141	
									(0.159)		(0.158)		(0.162)		(0.161)	
RNMRP 5y (-2)									0.311	**	0.320	**	0.282	*	0.307	**
									(0.154)		(0.152)		(0.155)		(0.154)	
RNMRP 5y (-3)									-0.259		-0.252		-0.241		-0.254	
									(0.173)		(0.169)		(0.174)		(0.174)	
RNMRP 5y (-4)									0.102		0.081		0.110		0.088	
									(0.140)		(0.137)		(0.139)		(0.143)	
Time dummies	YES	YES		YES	YES	YES	YES		YES	YES	YES		YES		YES	
City dummies	YES	YES		YES	YES	YES	YES		YES	YES	YES		YES		YES	
Constant	10.141	18.708	***	-0.386	5.780	3.670	17.658	***	-8.171	-0.999	4.206		17.750	***	-7.742		-0.636	
	(14.429)	(1.779)		(14.277)	(14.527)	(14.854)	(1.964)		(14.522)	(14.856)	(15.012)		(1.959)		(14.685)		(15.008)	
Observations	1064	1064		1064	1064	1064	1064		1064	1064	1064		1064		1064		1064	
R^2	0.450	0.450		0.438	0.445	0.452	0.451		0.440	0.447	0.451		0.451		0.439		0.446	
Adjusted R^2	0.408	0.410		0.397	0.405	0.410	0.412		0.399	0.407	0.409		0.411		0.399		0.407	

Notes: The estimations are based on pooled OLS panel regressions with year and city dummies. "(-t)" denotes the t-th lag of the covariate. The estimation results of the control variables are available upon request. Dummies are included but not reported. Heteroscedasticity and autocorrelation-robust standard errors were used. ***, ** and * represent statistical significance at 0.01, 0.05 and 0.10 levels, respectively. Standard errors are displayed in parentheses.

Firstly, the explanatory power of the models is in line with related research, ranging around an adjusted R^2 of 0.35 – 0.40. However, when we exclude binaries for the city individuals, we observe estimations (not reported) with declining adjusted R^2 values around 0.10, showing city-specific heterogeneity. Both specified findings are in line with related literature e.g. Devaney et al. (2019). The temporal binaries are predominantly statistically insignificant, indicating temporal homogeneity.

Focusing on the linear predictors of interest, we find on average a positive and statistically significant, c.p. relationship for the RNMRP 10y and 5y within the base models 2 and 3 for the second lag. The model variations 2.1 – 2.3 and 3.1 – 3.3 provide similar results, emphasizing the robustness of the results. One can derive two insights from these findings. First, cross-border investors favor higher risk premia when looking for investment opportunities in Europe. Interestingly, this also applies to investors who anticipate long- and medium-term holding-periods. The models report a c.p. effect on average around 0.3% per base point relative risk premium (since betas range around 0.3).

Second, if a city offers a relative risk premium above the European mean, it generally takes six months until cross-border capital flows into the respective market. The specified finding is in line with expectations due to search and transaction phases in direct markets. Crosby and McAllister (2004) and Bond et al. (2007) state an average transaction period in UK commercial real estate markets of approximately six to nine months. Model 2.2 also shows a statistically significant positive sign for RNMRP 10y for the contemporary covariate (lag = 0), which however is not investigated any further.

Considering the target variable RNMY, no statistically significant relationship between relative yield and inflowing transaction volume could be revealed. This finding adds to the study of Devaney et al. (2019), in which cap rates were not found to impact on general transaction activity in commercial real estate markets. Concluding the OLS result section, the first hypothesis can be confirmed after differentiating between yields and risk premia. Thus, relative attractiveness contributes to the existing absolute measures of determinants of cross-border transactions. However, relative attractiveness of cross-border investors is only perceived in terms of relative risk premia and not relative yields.

In addition to the fully parametric model, we assess hypothesis two by specifying semi-parametric GAMMs. We use smoothing functions for the RNMRP only, since the RNMY has not shown significance in the fully parametric approach (models 1 – 1.3). The GAMM specifications are identical to the linear ones and denoted with a "G", to ensure easy comparability with the OLS peer. All other covariates are still included with a linear predictor. However, we do not report the coefficients of still parametrized lags of the

RNMRP, since they are already reported above (see Table 2.5). Instead, Table 2.6 presents the estimated degrees of freedom and the statistical significance for the smoothing functions of the covariates as an expression of nonlinear behavior.

Table 2.6: GAMM Estimation for Spline Functions of Non-parametric Covariates

Dependent variable: ln (cross-border transaction volume)

Model	Model G.2	Model G.2.1	Model G.2.2	Model G.2.3	Model G.3	Model G.3.1	Model G.3.2	Model G.3.3
Controls:								
Macroeconomic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Real estate	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Excluded control	None	GCI	Vacancy	Stock	None	GCI	Vacancy	Stock
Target variables:								
RNMRP 10y	-	-	7.272***	-				
			(2.887)					
RNMRP 10y (-1)	-	-	-	-				
RNMRP 10y (-2)	6.92***	6.999***	2.089*	6.853**				
	(2.721)	(2.753)	(2.397)	(2.605)				
RNMRP 10y (-3)	-	-	-	-				
RNMRP 10y (-4)	-	-	-	-				
RNMRP 5y					-	-	-	-
RNMRP 5y (-1)					-	-	-	-
RNMRP 5y (-2)					6.905***	6.983***	7.427***	6.843**
					(2.636)	(2.664)	(3.238)	(2.520)
RNMRP 5y (-3)					-	-	-	-
RNMRP 5y (-4)					-	-	-	-
Time dummies	YES	YES	YES	YES	YES	YES	YES	YES
City dummies	YES	YES	YES	YES	YES	YES	YES	YES
Observations	1064	1064	1064	1064	1064	1064	1064	1064
Adjusted R^2	0.421	0.423	0.416	0.418	0.420	0.423	0.414	0.407

Notes: The estimations are based on GAMM regression, using penalized splines and the Gaussian link family. "(-t)" behind the name of the covariate denotes the t-th lag. The estimated degrees of freedom of the smooth terms are reported. The joint significance of the smoothing terms expressed by the F-test values is displayed in parentheses. The remaining parametrized covariates are not reported, but are identical to the specifications displayed in Table 2.5. Heteroscedasticity and autocorrelation-robust standard errors were used. ***, ** and * represent statistical significance at 0.01, 0.05 and 0.10 levels, respectively.

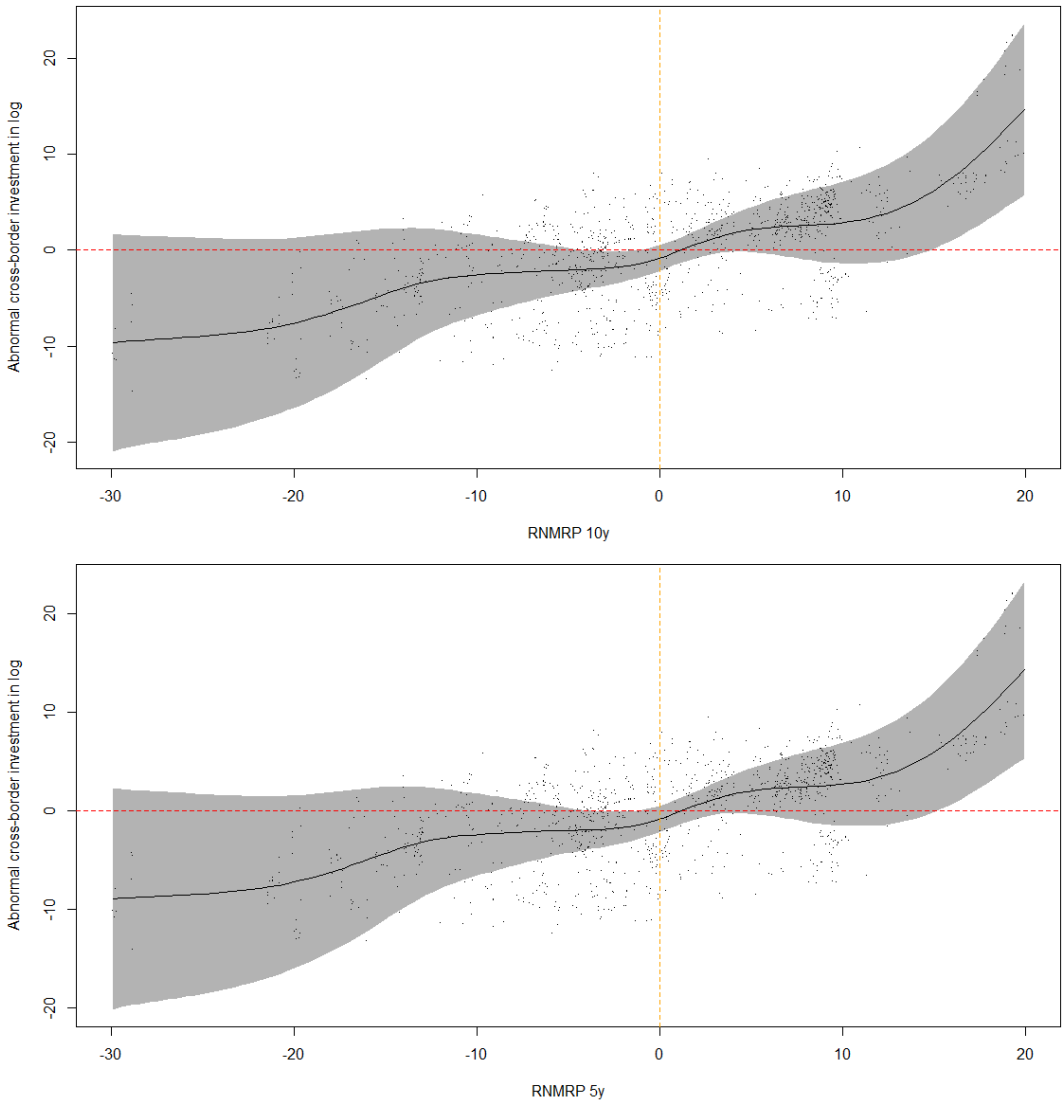
Firstly, the results reveal slight differences for the adjusted R^2. Since we only use smoothing functions for a single or two covariates per specification, a few models (e.g. models G.2, G.2.2, G.3, and G.3.2) show increased explanatory power by about one percentage point.

More importantly, the specified smooth terms unanimously show statistical joint significance, as expressed by the F-test values for each smooth term. The individual nonlinear behavior for each of the target variables can be assessed by the estimated

degrees of freedom of the respective smooth term. The interpretation works as follows: Estimated degrees of freedom equal to one represent an entirely linear relationship between the dependent and independent variables across the entire distribution. Hence, the larger the difference of the estimated degrees of freedom from one, the stronger the nonlinearity within the relationship becomes.

Here, most smoothing functions show estimated degrees of freedom around 7. Only the second lag of the RNMRP 10y in model G.2.2 shows a much smaller value. Nonetheless, the function is still statistically significant. Thus, we can confirm the nonlinear behavior of the target variables across all specifications. Since smoothing functions of GAMMs do not report a single estimate, a numerical interpretation is not possible. Instead, we report the graphical illustration of selected and representative functions and the respective partial residuals, as displayed below (see Figure 2.2). The functions were chosen from the base models G.2 and G.3. Other models show similar results.

Figure 2.2: Smooth Functions of RNMPR 10y and RNMRP 5y – Models G.2 & G.3



The graphic display shows the smoothing function across the distribution of the RNMRPs on the x-axis. The y-axis represents the divergence of the smoothing function from the mean of the dependent variable. Accordingly, we can derive various findings from the graphic inspection of the functions. Firstly, we can confirm the linear models and their predictions of a constantly increasing trend across the entire bivariate distribution. Secondly, the almost perfect match of the smooth term and the vertical and horizontal line indicate market efficiency, because a relative premium of zero matches a smooth term value of zero. The horizontal line also shows a negative effect of the risk premia on the mean cross-border transaction volume below zero.

However, most interestingly we observe a convex shape of the smoothing terms for values in the right tail of the bivariate distribution. This indicates that markets with extreme risk premia also attract extremely high capital inflows, since especially the upper tail of the distribution has a convex shape. The same applies inversely to the lower tail, causing transaction volumes to decline heavily below the mean. The results are similar for the RNMRP 5y.

Lastly, the combination of the two approaches can be summarized in terms of the following empirical results: We find both, statistical significance for the linear predictor as well as the corresponding smoothing terms of risk premia. Thus, the combination of the empirical findings reveals a linear c.p. effect and also nonlinear behavior in the tails of the bivariate distributions of the risk premia and independent covariate.

2.6 Conclusion and further Aspects

This study presents a new approach to explaining commercial cross-border transaction activity, namely the development of a city's relative attractiveness in comparison to its peers. We find an on average, c.p. and statistically significant relationship between relative risk premia and inflowing cross-border capital into office properties in Europe. We thus confirm the existence of a risk-premium-chasing behavior of cross-border investors with regard to relative city attractiveness. Moreover, we find empirical evidence for a timely lagged effect, since statistical significance can be observed predominantly for the two-quarter-delayed covariates.

However, a decisive differentiation for the economic finding is needed. The measurement of relative yields is unanimously statistically insignificant, underlining the importance of risk premia instead of pure yields as explanatory variable for cross-border inflows. Nonetheless, we conclude and also extend the existing body of literature by showing the relationship between the investor calculus of relative attractiveness and capital flows as a new determinant. Thus, we can partly justify our first hypothesis, while highlighting the importance of differentiating between relative yields and risk premia.

Moreover, we find evidence for a potential nonlinear behavior of the relative attractiveness measures, expressed by the statistical significance of the smoothing terms in the GAMM. Consequently, we can justify the second hypothesis. Interestingly, we find a curvy or convex shape of the smoothing functions, especially in the tails. This finding indicates extreme capital inflow behavior for locations which also offer extreme relative premia. Thus, we conclude, that especially risk-affine cross-border investors trigger abnormally high investment flows into real estate markets.

Some limitations apply to the used data. Firstly, we analyze European data only. The same analysis on a larger scale appears promising, as proposed on a global level by Devaney et al. (2019). Secondly, the depth of the data can be discussed. As outlined by Lieser and Groh (2014), a large variety of covariates show a statistically significant relationship with foreign investment volumes. However, since we control for the most important types of impact variables, the explanatory power of the models are in line with previous studies. Thus, we perceive the selected controls as a sufficient set of variables. Econometric robustness tests also confirm the stability of the results across various specifications.

Practical implications can be derived from an investment management and risk controlling perspective. The understanding of determinants in market transaction volumes is an important factor for anticipating inflowing capital and potential capital value changes.

Therefore, for example positive divergence from the benchmark is expected to cause on average higher inflows of cross-border capital. Equity investors can use the insights especially for their disinvestment strategies. In this context, they can specifically address foreign buyers, when risk premia in markets of their existing property investments move above the European mean, since cross-border investors are expected to invest in these locations. Financing debt investors on the other hand can expect sales of standing investments to cross-border investors in advance of their expected maturity, if risk premia of the market move above the European mean. Secondly, financing institutions can benefit from cross-border investors by offering funds to them and consequently expect new business opportunities. Early anticipation of potential financing requests will help to plan refinancing and money allocation activities.

Further useful research may be undertaken by differentiating between the geographic origin and type of investor. Considering the return chase behavior, Devaney et al. (2018) noted that different nationalities may matter. Moreover, focusing on other property types would reveal whether the relative attractiveness is generally applicable to other markets and not a phenomenon unique to the commercial- and office sector. Lastly, extending the present approach of relative attractiveness not only to the yield and risk premium side of a market, but to other covariates, may provide further insights.

2.7 Appendix

Table 2.7: Results Levin-Lin-Chu Test for Stationarity

Variable	Test statistic	P-Value	Variable with Δ	Test statistic	P-Value
Dependent Variable					
Cross-border transaction volume	-4.036	0.000			
Macroeconomic variables					
GDP growth	-11.571	0.000			
CPI growth	0.904	0.900	CPI growth Δ	-21.245	0.000
Unemployment rate	14.910	1.000	Unemployment rate Δ	-7.503	0.000
GCI	-3.552	0.000			
Real-estate-related variables					
Vacancy	-8.816	0.000			
Stock	1.276	0.899	Stock Δ	-5.293	0.000
Prime Rent Growth	-1.958	0.030			
1. Target Variable					
RNMY	-1.226	0.110			
2. Target Variable					
RNMRP 10y	-1.355	0.090			
RNMRP 5y	-1.694	0.050			

2.8 References

Bond, S. A., Hwang, S., Lin, Z., & Vandell, K. D. (2007). Marketing Period Risk in a Portfolio Context: Theory and Empirical Estimates from the UK Commercial Real Estate Market. *The Journal of Real Estate Finance and Economics*, *34*(4), 447–461.

Cajias, M., & Ertl, S. (2018). Spatial effects and non-linearity in hedonic modeling. *Journal of Property Investment & Finance*, *36*(1), 32–49.

Chin, W., Dent, P., & Roberts, C. (2006). An exploratory analysis of barriers to investment and market maturity in southeast Asian cities. *Journal of Real Estate Portfolio Management*, *12*(1), 49–57.

Crosby, N., & McAllister, P. (2004). *Liquidity in commercial property markets: Deconstructing the transaction process* [Working Paper]. The University of Reading Business School, Reading.

Devaney, S., Livingstone, N., McAllister, P., & Nanda, A. (2019). Capitalization rates and transaction activity in international office markets: A global perspective. *Global Finance Journal*, (in press).

Devaney, S., McAllister, P., & Nanda, A. (2017a). Determinants of transaction activity in commercial real estate markets: evidence from European and Asia-Pacific countries. *Journal of Property Research*, *34*(4), 251–268.

Devaney, S., McAllister, P., & Nanda, A. (2017b). Which factors determine transaction activity across U.S. metropolitan office markets? *The Journal of Portfolio Management*, *43*(6), 90–104.

Devaney, S., Scofield, D., & Zhang, F. (2018). Only the best? Exploring cross-border investor preferences in US gateway cities. *The Journal of Real Estate Finance and Economics*, *32*(1), 1–24.

Farzanegan, M. R., & Fereidouni, H. G. (2014). Does real estate transparency matter for foreign real estate investments? *International Journal of Strategic Property Management*, *18*(4), 317–331.

Ford, D. A., Fung, H.-G., & Gerlowski, D. A. (1998). Factors affecting foreign investor choice in types of U.S. real estate. *Journal of Real Estate Research*, *16*(1), 99–111.

Fuerst, F., Milcheva, S., & Baum, A. (2015). Cross-border capital flows into real estate. *Real Estate Finance*, *31*(3), 103–122.

Gholipour Fereidouni, H., & Ariffin Masron, T. (2013). Real estate market factors and foreign real estate investment. *Journal of Economic Studies*, *40*(4), 448–468.

He, C., & Zhu, Y. (2010). Real estate FDI in Chinese cities: Local market conditions and regional institutions. *Eurasian Geography and Economics*, *51*(3), 360–384.

Holsapple, E. J., Ozawa, T., & Olienyk, J. (2006). Foreign "direct" and "portfolio" investment in real estate: An electic paradigm. *Journal of Real Estate Portfolio Management*, *12*(1), 37–47.

Laposa, S., & Lizieri, C. (2005). *Real estate capital flows and transitional economies* [Conference Paper]. American Real Estate Society Meeting, Santa Fe, NM.

Lieser, K., & Groh, A. P. (2011). The attractiveness of 66 countries for institutional real estate investments. *Journal of Real Estate Portfolio Management*, *17*(3), 191–211.

Lieser, K., & Groh, A. P. (2014). The determinants of international commercial real estate investment. *The Journal of Real Estate Finance and Economics*, *48*(4), 611–659.

McAllister, P., & Nanda, A. (2016). Do foreign buyers compress office real estate cap rates? *Journal of Real Estate Research*, *38*(4), 569–594.

McGreal, S., Parsa, A., & Keivani, R. (2001). Perceptions of real estate markets in Central Europe: A survey of European Investors. *Journal of Real Estate Literature*, *9*(2), 147–160.

MSCI. (2019). *MSCI Property Indexes Methodology: Index construction objectives, guiding principles and methodology for the MSCI Property Indexes.*

Oikarinen, E., & Falkenbach, H. (2017). Foreign investors' influence on the real estate market capitalization rate – evidence from a small open economy. *Applied Economics*, *49*(32), 3141–3155.

Pi-Ying Lai, P., & Fischer, D. (2007). The determinants of foreign real estate investment in Taiwan. *Pacific Rim Property Research Journal*, *13*(3), 263–279.

PricewaterhouseCoopers, & Urban Land Institute. (2019). *Emerging trends in real estate: Creating an impact. Europe 2019.*

Rodríguez, C., & Bustillo, R. (2010). Modelling foreign real estate investment: The spanish case. *The Journal of Real Estate Finance and Economics*, *41*(3), 354–367.

Salem, M., & Baum, A. (2016). Determinants of foreign direct real estate investment in selected MENA countries. *Journal of Property Investment & Finance*, *34*(2), 116–142.

Sirmans, C. F., & Worzala, E. (2003). International direct real estate investment: A review of the literature. *Urban Studies*, *40*(5-6), 1081–1114.

World Economic Forum. (2018). *The global competitiveness report: 2018.*

3 Rental Pricing of Residential Market and Portfolio Data – A Hedonic Machine Learning Approach

3.1 Abstract

Artificial intelligence (AI) and especially machine learning (ML) methods increasingly offer valuable alternatives to answer questions in real estate research and practice. This study comprises two components: First, we investigate whether ML methods are suitable of estimating residential rents by comparing a conventional hedonic model with four ML algorithms, namely Support Vector Regression (SVR), Random Forest Regression (RFR), Gradient Tree Boosting (GTB) and eXtreme Gradient Boosting (XGB). We find ML methods to model rental values more precisely than traditional linear regression. While RFR shows the highest predictive performance, GTB appears to be most robust to overfitting. Second, we use these findings to estimate rental values for an institutionally managed portfolio and match them with their corresponding contract rents. On average, we find the apartments to be underrented, with ML models indicating higher deviation of estimated and contract rents than linear Ordinary Least Squares (OLS) models. Thus, our findings indicate that investors rather rely on traditional methods to derive contract rent levels within their portfolio. With that, this study reveals potential benefits when applying ML hedonic models in the area of residential markets and portfolios.

Keywords: Machine learning, hedonic models, residential real estate, rent prediction, multiple listing systems

Acknowledgments: The authors especially thank PATRIZIA AG for contributing to this study. All statements of opinions are those of the authors and do not necessarily reflect the opinion of PATRIZIA AG or its associated companies.

3.2 Introduction

The role of residential rents is of central importance in the real estate industry for both tenants and landlords. Considering the former, rents often account for the largest portion of their monthly spending. For the latter, they mark the fundamental determinant of the value of housing (Gallin, 2008; Genesove, 2003). Consequently, literature has long concentrated on the question of how rental prices develop within a market. Today, more than ever, this is of great importance given urbanization and demographic changes leading to thriving residential markets especially in metropolitan areas (IMF, 2018; ULI, 2020).

In the case of a common house or apartment itself, its rent is "a single-dimensional summary of the market's valuation of all the physical, service and locational attributes […]" (J. Goodman, 2004; Verbrugge et al., 2017). In other words, every single characteristic of a residential property should be priced in and thus, ultimately contributes to the rent that the market will accept. However, prices for individual attributes are not fixed. Researchers have long tried to fathom the connections between the characteristics of a property and its associated rent. While rather conventional statistical methods such as Ordinary Least Squares (OLS) still represent the preferred statistical tool, new possibilities arise from the field of artificial intelligence (AI).

While those methods are increasingly used in several areas of real estate research and practice, they have only been applied in the derivation and analysis of residential rents in a limited way yet. Given the above, this paper investigates whether hedonic machine learning (ML) methods are capable of providing new insights and applications in residential rental markets. Recent research in the field of real estate applying ML methods focuses predominantly on valuation aspects. Authors such as Lindenthal (2020), Hamilton and Johnson (2018) and Lindenthal and Johnson (2020) apply such techniques to investigate whether aesthetics and architectural styles affect real estate prices. Using ML, Chin et al. (2020) estimate the benefit of infrastructural investments on property values while Pérez-Rave et al. (2019) apply ML to big data for predictive and inferential purposes. The subject of rents and how market participants can use AI to assess and verify investment decisions has, however, not yet been investigated in depth.

Consequently, literature on this topic is scarce even though new tools seem to have capabilities that may outperform conventional hedonic methods. From a practical point of view, next to its relevance for the institutional sector, our findings may be useful to governments, for whom such methods can serve as additional instruments to engage in housing markets. Consequently, we attempt to shed light on which ML methods are best

suitable for capturing and processing price formations in rental markets. ML methods differentiate from traditional regression methods in their underlying predefined assumptions. The former presuppose a linear or non-linear relationship between rental values and the hedonic characteristics whereas artificial intelligent learning methods 'think' differently. More precisely, there is no such predetermined prerequisite, but an algorithm. Hence, an econometrician takes advantage of letting the machine decide the steps necessary to model the relationship between the response and the explanatory variables in several training steps.

The aim of this study is to shed light on the application of algorithm-driven methods in rental markets. In addition, we aim to assess the value that market participants might obtain when managing a residential real estate portfolio based on ML methods as opposed to fundamental OLS analysis. Consequently, we

(1) assess how accurate linear and algorithm-driven hedonic models predict rents based on a large data set from multiple listing systems (MLS). For this purpose, a variety of performance metrics (error measures) is used.

(2) transfer the findings from (1) to a dataset of an institutionally managed residential portfolio. Using the previous model specifications, we estimate rental values an investor could expect for the portfolio apartments in re-lettings scenarios. Further, we compare them to their corresponding contract rents to find out whether the different models would estimate a potential (or need) for rental adjustments.

The paper is structured as follows. Section 3.3 contains an overview of the literature in the field of real estate and ML. Section 3.4 explains the composition of the two data sets. In section 3.5, the ML methods used for rent analysis throughout the paper are introduced. The results are presented in section 3.6. The sixth and final section summarizes the conclusions.

3.3 Hedonic Modelling in the Real Estate Literature

The aim of hedonic modelling is to better understand the fundamental factors affecting property rents and prices. By expressing the rent or price of an apartment as the sum of its estimated individual characteristics, hedonic modelling can be used for inferential and predictive purposes. Traditionally, a hedonic model employs multiple linear regression to establish the relationship between the response and the corresponding hedonic characteristics (Rosen, 1974; A. C. Goodman, 1978). Depending on the spatial characteristics of the market under investigation and the data structure, a hedonic model

needs to fulfil a minimum number of assumptions (see e.g. Sirmans et al., 2005 and Bourassa et al., 2007).[11] However, several authors such as Lai et al. (2008), Bourassa et al. (2010) and Cajias (2018) have demonstrated the limited explanatory power of traditional hedonic models and shown that statistical developments such as the inclusion of spatial and non-linear effects lead to significant enhancements in model accuracy (see more: Fik et al., 2003; Lin et al., 2009; Banzhaf & Farooque, 2013).

Over the last decade, advances in computational power and ML algorithms have enabled the development of modern regression techniques. By abandoning the previously mandatory functional form of the relationship between the response and the covariates, a variety of ML algorithms emerged – such as Gradient Boosting Trees (GTB) (Friedman, 2001), Random Forest Regression (RFR) (Breiman, 2001) and Support Vector Regression (SVR) (Smola & Schölkopf, 2004). Given the goal of ML methods is to maximize explanatory power and prediction accuracy, real estate literature has identified these to be well suited for predictive questions.[12]

3.3.1 Hedonic Analysis of Property Prices – Mass Appraisal and Automated Valuation

Aside from traditional valuation models, automated valuation methods (AVM) based on ML algorithms are becoming even more popular (Kontrimas & Verikas, 2011). Commonly, authors focus on a specific ML method and look at its predictive power on real estate prices to draw conclusions on the stand-alone improvements. In this context, Yoo et al. (2012) use transaction data on 4,469 houses in Onondaga County, NY (USA) to demonstrate the superior model accuracy of RFR, compared to traditional regression techniques due to the ability of modelling non-linear relationships. The findings are in line with Antipov and Pokryshevskaya (2012), who investigate a dataset of 2,848 transactions relating to apartments in St. Petersburg (Russia). Both call for a more frequent application of ML methods in predicting property prices. Moreover, Yao et al. (2018) use RFR to map fine-scale housing prices in Shenzhen (China). By analyzing residential property transactions in Hong Kong and Nanjing (China), Lam et al. (2009) apply SVR to predict property prices. Moreover, the investigation of 100 house transactions in Lithuania by Kontrimas and Verikas (2011) shows that SVR is well suited due to its ability to capture

[11] The assumptions are intended to correspond to the variables to be included in the model, controlling for spatial characteristics and nearby amenities. The data structure is generally either cross-sectional, time serial, panel or pooled cross-sectional and determines the normality assumptions of the residuals.

[12] The black box character of ML methods is often perceived as a disadvantage that prevents the econometrician from understanding and interpreting the influence of certain variables. However, if the goal of ML methods is prediction, this disadvantage is not very harmful, since the focus is not recognizing relationships between variables, but rather optimizing the predictive performance.

non-linear relationships. Regarding boosting methods, van Wezel et al. (2005) for example, deploy gradient boosting to predict automobiles as well as real estate sales prices in Boston (USA), Windsor and Essex (Canada), and the Netherlands. Moreover, Kok et al. (2017) demonstrate the performance of boosting as well as RF and OLS on property prices with a dataset containing 54,000 US multi-family houses.

Furthermore, recent literature in the field of modelling property prices compares several ML techniques. Zurada et al. (2011) apply OLS, further linear regression techniques, regression trees (RT) and SVR, using a sample of 16,366 transaction prices in Louisville, Kentucky (USA). Baldominos et al. (2018) show the performance of RFR and SVR on house prices in Spain, using online listings. Moreover, Mayer et al. (2019) analyze the accuracy of different hedonic valuations models – including RFR, GTB and OLS as well as further linear models – and propose the application of different data updating techniques for property price valuations. Pace and Hayunga (2020) compare the performance of spatial models to ML techniques using tree-based algorithms. Ho et al. (2020) apply different ML methods for a dataset of housing transactions in Hong Kong. Bogin and Shui (2020) find RFR to perform best in accurately estimating rural property prices. The authors conclude that ML is more appropriate in modelling property prices due to its ability to allow for non-linear effects, whereas traditional models might suffer from misspecification.

3.3.2 Hedonic Analysis of Residential Rents

Especially for ML methods, most studies within the hedonic modelling literature focus on real estate prices. Far less is known about explaining and modelling rental values by applying ML approaches. Early research estimated the determinates of rental values (Sirmans et al., 1989; Kee & Walt, 1996). Recent studies on the rental housing market, including Thomschke (2015), Zhang and Yi (2017) and Cajias and Ertl (2018), show that traditional methods are still able to estimate property rents properly. While, for example, V. James et al. (2005) use spatial models to predict apartment rents, Cajias (2018) shows that semi-parametric models are capable of improving model accuracy by accounting for non-linear relationships in rental markets. Although traditional models are limited in their ability to reveal and model non-normal complex relationships, a lack of research exists regarding the application of ML methods for modelling property rents, as Hu et al. (2019) state.

Given the relevance of rental estimation for tenants, investors and governmental bodies together, with the "potential of AI-based methods" (Zurada et al., 2011), it is important to also accurately model the underlying rental market. Even though there is a growing body of literature on the topic, further investigation is needed due to various reasons: First,

literature is rather silent when it comes to a holistic comparison of various ML approaches for evaluating the varying performance measurements of different algorithms. Second, to the best of our knowledge, property rents have not been analyzed in depth so far in an ML context. Third, the emerging velocity and volume of real estate data through MLS enables new insights to real estate markets and provides a promising field of research, since "one of the main approaches to face [such data sources] is machine learning" (Pérez-Rave et al., 2019). And finally, the potential of ML applications for market participants to derive well-founded decisions in real estate markets has not yet been fully explored nor used.

3.4 Data

This study encompasses the residential real estate market in Munich, Germany. The country is home to one of the largest and most active real estate markets in Europe. As it is well-known as a safe haven for national and international investors, it attracts both domestic and cross-border capital allocation. As of 2019, Germany consisted of 42 million occupied apartments while having one of the lowest owner-occupancy rates in Europe with 47%. With that, Germany is considered one of the most important hubs for capital allocation in residential real estate on the continent, and thus, offers an interesting market for an in-depth investigation. With approximately 1.5 million residents and an annual growth rate of about 0.75%, Munich is the third largest city in Germany. The city and its metropolitan areas have one of the most prospering economies in Germany, accommodating several globally active companies in sectors such as automotive, environmental techniques, information and communication, insurance, life sciences and medicine. Stable economic growth and good employment conditions have yielded a positive development of the residential market throughout the last decade.

To analyze the rental market in Munich, we use two different data sets: First, asking data from MLS enables us to estimate and compare the predictive performance of the applied hedonic models. Based on the derived values, we then estimate rental values for a residential portfolio of institutionally managed apartments and compare the estimates to the observed contract rents.

3.4.1 MLS Data

In contrast to comparable international real estate markets, Germany does not require either private or institutional landlords to publish rental information. Therefore, no general

database of contract rents exists. Consequently, asking rents from MLS serve as the main source of pricing information and are used to estimate the current rental level in the German residential market (see well-established applications, such as F+B Residential Index, Empirica Real Estate Index, etc.). The use of asking data can be advantageous as it offers the possibility to capture and rapidly reveal market movements. Y. Chen et al. (2016) and Baldominos et al. (2018) argue that it is more appropriate for modelling timely dynamics of housing markets on a fine-scale level.

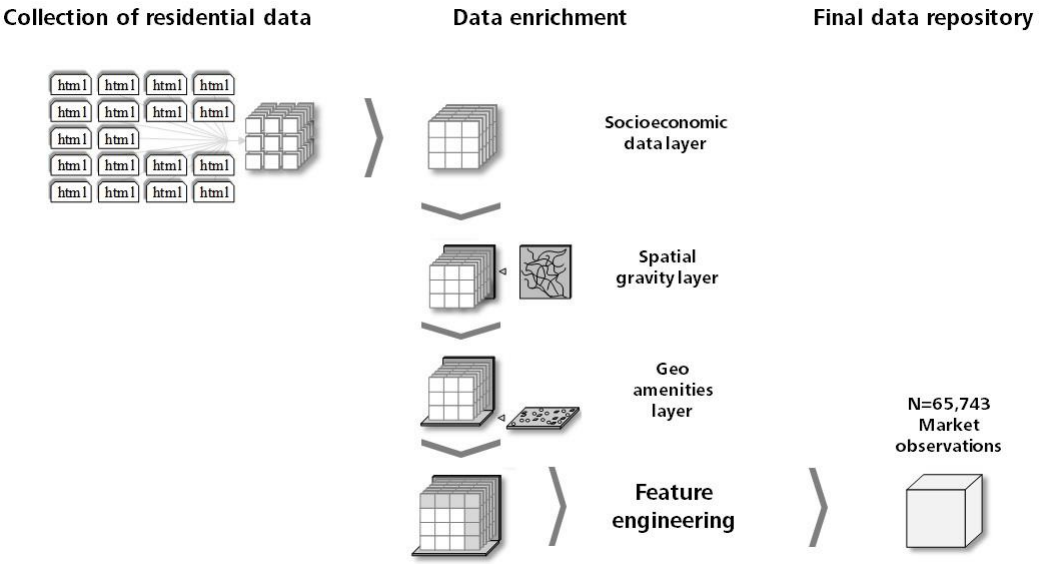
When it comes to real estate sales, early research has documented that differences between listing and transaction prices exist and that these are highly associated with market liquidity, measured by time on market (TOM). Jud and Winkler (1994) and Jud et al. (1996) found that both the degree of above market pricing and changes in the listing price affect TOM. Yavas and Yang (1995) state that overpricing increases the marketing time. Analyzing the German residential market, Cajias and Freudenreich (2018) show that Munich is subject to high market liquidity, with the degree of overpricing being comparably low. Since the German residential market is a renter's rather than a buyer's market, their findings indicate diminishing deviations between asking and market rents. Cajias (2018) suggests that "the deviation [of asking and market rents] is not expected to lead to error bias, especially after controlling for […] hedonic characteristics". As Gröbel (2019) suggests, asking data in Germany "reflects the currently prevailing overall market situation" since the price formation in the housing market is perceived to be determined by the offering party.

Moreover, MLS asking data can overcome the challenges raised by the general lack of European housing contract data which is mentioned, for instance, by Rondinelli and Veronese (2011). It is actively used for empirical research by several authors such as Hanson and Hawley (2011), Rae (2014), Gröbel and Thomschke (2018), Pérez-Rave et al. (2019) and Gröbel (2019) for studies in Germany, the US and the UK. As Pérez-Rave et al. (2019) state, MLS data shows important characteristics of big data in terms of volume, variety and value. This enables researchers and market participants to overcome temporal delays and limited analyses on market developments that are associated with, for example, official statistics. In this context, MLS are perceived as "one of the most significant feature of today's real estate industry" (Li & Yavas, 2015). Due to the characteristics of the Munich residential real estate market, we expect the asking rents to be a good approximation for market-conform rental values. Although asking data plays a significant role in housing markets (see e.g. Shimizu et al., 2016, Han & Strange, 2016), differences to transaction data can occur that need to be kept in mind.

To assess the performance of hedonic models, our study comprises a dataset of 65,743 residential apartments in Munich, including hedonic characteristics, socio-economic information and distance variables, from January 2013 to June 2019. To avoid sample bias for the investigation of Munich's residential market that is mainly dominated by apartments, we exclude single houses as well as semi-detached and terraced houses. Furthermore, highly specialized market segments like student apartments, senior living accommodations, furnished co-living spaces, and short-stay apartments are not considered.

We access Empirica Systeme, one of the largest providers of real estate data in the German residential market. It uses web-scraping techniques for collecting, preparing and integrating real estate listings from more than 120 different MLS with full hedonic characteristics.[13] Furthermore, we include socio-economic data from Growth from Knowledge (GfK), Germany's largest market research institute. We also add a gravity layer using data from Eurostat and the German statistical office to implicitly enable the models to account for spatial information. Finally, we complement each georeferenced residential data point by an amenities layer measuring the Euclidean proximity to important amenities. This information is gathered from Open Street Map (OSM) and Google via an API in R (R Core Team, 2020). Data preparation and processing is displayed in Figure 3.1.

Figure 3.1: Extraction-Load-Transform Process for Estimating Hedonic Market Models



[13] The Empirica Systeme GmbH is an established partner in data analytics solutions for the residential market in Germany and a data provider for brokers such as CBRE, Colliers, Engel&Voelkers, JLL, Savills as well as for banks, institutional real estate managers, cities and others.

This results in a dataset comprising eight structural characteristics (living area, age and whether the apartment has a bathtub, built-in kitchen, parking lot, terrace, balcony and an elevator), two socio-economic (number of households and households purchasing power in ZIP code area), and seven distance variables (proximity to bus station, park, school, subway, supermarket, neighborhood center and city center). Rent, living area, distances as well as both socio-economic characteristics are incorporated using their log-transformation to account for the distribution. Quarter and year dummies are used to control for time effects. Earlier studies found additional contract and market information to affect price formation in housing markets (see e.g. competition and listing density in Turnbull & Dombrow, 2006). However, since we do not have access to further information, our analysis is limited to structural, neighboring and locational characteristics.

Table 3.1 shows the descriptive statistics. We find a mean asking rent of 1,238 EUR/p.m. (euros per month), with rental values ranging from 123.97 EUR/p.m. up to 10,764 EUR/p.m. An average apartment is 76.49 sqm (square meters), comprises approximately three rooms, and was built in 1975. Each apartment is on average 1.44 km distant from the subway, 0.76 km from a supermarket and 0.56 km from the next school. Moreover, the city center is on average 4.62 km away, the center of the corresponding ZIP code is in 0.60 km distance. The mean number of households in a ZIP area accounts for 11,423 with a mean purchasing power of 59,855 EUR each.

Table 3.1: Descriptive Statistics of the MLS Data

Variable name	Unit	Spatial reference	Source	Mean	Median	SD	Min	Max
Living Area	sqm	Apartment	Portfolio	76.49	71.00	36.49	10.00	435.00
Age relative to 2017	Integer	Apartment	Portfolio	42.36	41.00	33.84	-2.00	118.00
Centroid ZIP	km	Distances	Google/OSM	0.60	0.53	0.38	0.00	2.43
Centroid NUTS	km	Distances	Google/OSM	4.62	4.57	2.08	0.22	12.33
Rent	EUR/p.m.	Apartment	Portfolio	1,238.00	1,079.34	721.82	123.97	10,764.00
Number of households (HH)	HH/ZIP	ZIP	GfK	11,423.00	11,768.00	3,305.76	1,860.00	16,978.00
Household purchasing power	EUR/HH/ZIP	ZIP	GfK	59,855.00	58,849.80	5,501.76	46,170.00	71,765.00
Bus	km	Distances	Google/OSM	1.14	0.75	1.10	0.00	6.20
Park	km	Distances	Google/OSM	0.79	0.44	0.92	0.00	4.75
School	km	Distances	Google/OSM	0.56	0.24	0.85	0.00	4.89
Subway	km	Distances	Google/OSM	1.44	0.75	1.67	0.00	11.76
Supermarket	km	Distances	Google/OSM	0.76	0.35	1.03	0.00	5.16
Bathtub	Binary	Apartment	Portfolio	0.54	1	0.5	0	1
Built-in kitchen	Binary	Apartment	Portfolio	0.68	1	0.47	0	1
Parking lot	Binary	Apartment	Portfolio	0.62	1	0.49	0	1
Terrace	Binary	Apartment	Portfolio	0.18	0	0.38	0	1
Balcony	Binary	Apartment	Portfolio	0.63	1	0.48	0	1
Elevator	Binary	Apartment	Portfolio	0.56	1	0.5	0	1

Notes: This table reports the summary statistics comprising data from January 2013 to June 2019. Age is calculated as the difference from building age to the year 2017. All distance variables are calculated as the distance to the specific apartment in kilometers. Binary variables report whether the apartment includes a certain characteristic (1) or not (0). Rent is presented as euro per month. Information on households is reported on ZIP level. SD: standard deviation, Min: minimum value, Max: maximum value.

3.4.2 Portfolio Data

In addition to the obtained data through MLS, a German asset manager granted access to portfolio data from institutionally managed residential real estate that is publicly not available. The portfolio consists of 716 apartments located in Munich, comprising contract rents and the same explanatory variables as presented in the previous section. Table 3.2 summarizes the descriptive statistics of the residential portfolio. An average apartment contains 71.99 sqm and yields a rental income of 1,009.37 EUR/p.m. The distance to the city center of 6.77 km is about 2 km further than the distance of an average apartment, but the distance to the center of the related ZIP code is with 0.50 km 200 m shorter. Moreover, the distances to all important infrastructure facilities is on average closer compared to the apartments in the previous dataset. Purchasing power and number of households are about the same. We again consider additional hedonic characteristics and time controls as dummy variables.

Table 3.2: Descriptive Statistics of the Portfolio

Variable name	Unit	Spatial reference	Source	Mean	Median	SD	Min	Max
Living Area	sqm	Apartment	Portfolio	71.99	75.56	30.59	20.92	179.79
Age relative to 2017	Integer	Apartment	Portfolio	37.91	46.00	29.64	1.00	90.00
Centroid ZIP	km	Distances	Google/OSM	0.50	0.50	0.28	0.20	1.00
Centroid NUTS	km	Distances	Google/OSM	6.77	6.00	5.24	1.70	19.00
Rent	EUR/p.m.	Apartment	Portfolio	1,009.37	938.61	469.33	204.52	3,179.67
Number of households (HH)	HH/ZIP	ZIP	GfK	13,200.98	13,662.00	2,321.27	9,720.00	16,256.00
Household purchasing power	EUR/HH/ZIP	ZIP	GfK	55,441.53	54,496.47	3,309.16	52,045.09	63,720.57
Bus	km	Distances	Google/OSM	0.92	0.64	0.83	0.13	2.77
Park	km	Distances	Google/OSM	0.65	0.68	0.26	0.29	1.14
School	km	Distances	Google/OSM	0.57	0.43	0.23	0.26	0.92
Subway	km	Distances	Google/OSM	0.60	0.53	0.26	0.13	1.01
Supermarket	km	Distances	Google/OSM	0.58	0.66	0.23	0.01	0.87
Bathtub	Binary	Apartment	Portfolio	0.50	1	0.10	0	1
Built-in kitchen	Binary	Apartment	Portfolio	0.21	0	0.41	0	1
Parking lot	Binary	Apartment	Portfolio	0.50	1	0.10	0	1
Terrace	Binary	Apartment	Portfolio	0.06	0	0.25	0	1
Balcony	Binary	Apartment	Portfolio	0.94	1	0.23	0	1
Elevator	Binary	Apartment	Portfolio	0.63	1	0.48	0	1

Notes: This table reports the summary statistics comprising data from June 2019. Age is calculated as the difference of the building age to the year 2017. All distance variables are calculated as the distance to the specific apartment in kilometers. Binary variables report whether the apartment includes a certain characteristic (1) or not (0). Rent is presented as euro per month. Information on households is reported on ZIP level. SD: standard deviation, Min: minimum value, Max: maximum value.

3.5 Methodology

Our analysis comprises two components. In the first part, we apply five hedonic models and estimate rental values based on the MLS data presented in section 3.4.1. Several error measures are used to compare the results to determine the model's predictive performance. The methods and error measures are presented throughout this section. In the second part, we transfer the findings and model specifications to the portfolio dataset discussed in section 3.4.2. Comparing the estimated rents to their contract rents enables us to identify to what extent a possible potential (or need) for rental adjustments exists as well as to highlight which new insights investors can get when applying ML methods in their rental estimation.

3.5.1 Hedonic Modelling with Traditional and Machine Learning Methods

The analysis encompasses one linear and four ML models. We follow Zurada et al. (2011) and Chin et al. (2020) by choosing OLS as the base case for the comparison of several algorithm-driven hedonic models. OLS is a widespread variant for hedonic modelling and

consequently a well-known and easy interpretable benchmark for performance analysis. SVR, RFR, GTB and eXtreme Gradient Boosting (XGB) represent the modern approaches that will be applied in our analysis. Except for XGB, all methods have been used for real estate related questions in areas such as valuation. XGB is a method developed in the last few years that shows computational advantages especially in large data sets. In the following, we discuss the basic structure of each hedonic method under investigation:

Ordinary Least Squares Regression – OLS

The most common approach is based on the traditional OLS regression. The rent y of property i is described as the sum of the predicted values of its j characteristics x_{ij}. By making use of OLS as a parametric optimization procedure, the estimated parameters β_j are achieved by minimizing the sum of the squared residuals as a loss function. The linear relationships between rents and the hedonic characteristics are valid for the entire population whenever the Gauss-Markov theorem is valid, that is, the estimators are the best linear unbiased estimators of the observed market values. Several statistical instruments can be further employed to increase the explanatory power, such as interaction terms, polynomial effects, and spatial effects.

Machine Learning Methods

ML techniques can identify complex structures and patterns. They provide high flexibility by avoiding the assumption of a specific functional form between the response and independent variables and are at the same time able to learn from the underlying data and optimize the predictive model. By dividing the dataset into a training and test set, overfitting within the training set (in-sample) is penalized by poor out-of-sample accuracy within the test set. Removing the test set during the learning process could mean that important patterns within the data remain unnoticed. Hence, z-fold cross validation is necessary. The resampling approach within this study makes use of a 5-fold cross-validation technique with a 75:25 ratio between the train and the test sets based on random sampling.[14]

Support Vector Regression – SVR

SVR is a modification of the Support Vector Machine, to categorize observations by finding a dividing hyperplane within an a-priori defined gap between the categories (Cortes & Vapnik, 1995). Instead of dividing the feature space by a certain gap, SVR attempts to fit observations within this specific threshold area to estimate a hyperplane – representing

[14] An in-depth description of cross validation is provided by Ho et al. (2021) with a discussion of possible advantages and disadvantages of selected ML methods. See Hastie et al. (2009) and G. James et al. (2013) for a more detailed description of the applied ML methods.

the regression line – that is able to capture the observed values. The threshold area is characterized by the soft margin ε. It defines the form of the hyperplane and is determined by choosing support vectors (SV) with respect to a specific loss function that allows an error margin tolerance. While error terms less than ε (and consequently within the threshold area) stay unconsidered, the part of the error exceeding the margin (ξ) is subject to a linear penalization. Consequently, SVs are chosen in a way that the threshold area includes as many observed values as possible while still accepting values exceeding the boundary through penalization. The model consequently tries to fit a hyperplane that on the one hand stays as flat as possible and on the other hand accounts for exceeding values within its functional form by estimating the amount up to which deviations larger than ε are tolerated.

Random Forest Regression – RFR

Characterized mainly by Breiman et al. (1984), RFR is a bagging method based on the concept of regression trees (RT). The idea of a RT is to divide the regression space into sub-intervals and provide a predicted value for each final interval, called leaf R_p. Starting with a specific input variable x_j, observations are binary partitioned at the node t_n into values being higher or lower than the chosen splitting value. The process of binary partitioning is iteratively applied at each resulting node, first choosing an independent variable x_j and the value s at which the splitting will take place. s is chosen in such a way that the sum of squared errors of the two inferior nodes is minimized. At each individual terminal node R_p, the predicted value $\hat{y}|t_n$ is a constant term that is equal to the average of the observed values with respect to the partition.

Partitioning can be applied any number of times to grow the tree and improve the approximation to the data. However, deep trees can be subject to noise as fewer observations in each terminal partition are available to estimate the predicted value. To avoid overfitting, penalty terms are used to identify, for example, the optimal number of nodes and to prune the tree. Since single pruned trees perform poorly in predicting observed values, a forest of trees is built by using several different trees simultaneously. The difference between the trees is ensured by using bootstrap aggregation. The overall predicted values are calculated by averaging the individual prediction rules.

Gradient Tree Boosting – GTB

Aside from bagging techniques, such as the abovementioned RFR, boosting methods are representatives of ensemble learners that combine the results of multiple models. The idea is to consolidate many so-called weak learners (standalone prediction rules that lead to imprecise results) into a meaningful and powerful so-called committee of predictions

(Hastie et al., 2009). The GTB, proposed by Friedman (2001), is a boosting concept with an ensemble of RTs as weak learners. In contrast to RFR, GTB does not consider the average prediction rule of the underlying trees but an ensemble of independent trees as the final predicted value. It uses the prediction rule of subsequent trees and an ensemble of trees that depend on the prediction of the preceding decision rule. Based on an initial decision rule, GTB proceeds with the prediction error of the initial (or preceding) rule as the target variable and iteratively builds a subsequent RT on the prediction error in order to incrementally enhance the final prediction rule.

Extreme Gradient Boosting – XGB

XGB is a scalable ML method for tree boosting. Moreover, it is an extension of the GTB algorithm. Developed by T. Chen and Guestrin (2016), it is a rather new approach to classification as well as regression, as it contains specific features that won several Kaggle[15] competitions in the recent past. The Gradient Boosting framework provides the foundation for the XGB algorithm, which offers several advancements.

The first involves a so-called regularized objective $\mathcal{L}(\phi)$ that penalizes complex models and therefore counteracts overfitting. Second XGB also contains a shrinkage parameter as a learning rate that rescales the predictions of individual trees to ensure further model improvements by following trees. A further addition of the method enables column subsampling that performs better in preventing overfitting than the traditional row subsampling. Split finding is one of the major challenges associated with tree learning. To find optimal split points, XGB offers the exact greedy algorithm and the approximate algorithm, both of which can be situationally applied. Since conventional approximate splitting algorithms may face difficulties in dividing data when the data points are not of equal weight, XGB adds the weighted quantile sketch algorithm. The latter ensures optimal splitting even when data is weighted. However, this method not only improves the computational procedures, it also increases the machine's system design via various features.

3.5.2 Error-based Comparison of Model Performance

Following Zurada et al. (2011), Schulz et al. (2014) and Mayer et al. (2019), we use mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and coefficient of determination R^2 to conclude on the accuracy of the applied methods. We furthermore investigate the precision regarding over- or underestimation by

[15] Kaggle is one of the leading online platforms for the data science community and regularly hosts data competitions, see https://www.kaggle.com/.

applying the mean percentage error (MPE). While similar research give little attention to the dispersion of the errors within the prediction, we discuss error buckets (PE10 and PE20), coefficient of dispersion (COD) and inter-quartile-range (IQR) to assess the magnitude of the estimation errors. By looking at the accuracy, precision and dispersion, we aim to derive further insights on the differences between the applied ML methods. Detailed descriptions of the error metrics can be found in Table 3.6 in the Appendix.

3.6 Econometric Results

In the first part of our analysis, we aim to investigate the predictive performance in terms of accuracy (how well models perform on average), precision (if models over- or underestimate observed values) and dispersion (the distribution and variance of estimation errors).

3.6.1 Predictive Performance of Hedonic Models

All results were obtained with the following model specifications. We used repeated cross-validation with five folds and five repetitions running on 72 central processing units (CPUs) simultaneously. GTB worked best with a tree depth of 6, a shrinkage rate of 0.07 and the number of trees being 438. SVR ran on the following specifications: $C = 0.9$, $\varepsilon = 0.0451$, $\sigma = 0.00679$. While the number of trees where 498 for RFR, XGB was trained with $\alpha = 0.112$, $\gamma = 0.601$ and $\eta = 0.216$.

Table 3.3: Error-based Comparison of Model Performance at Market Level

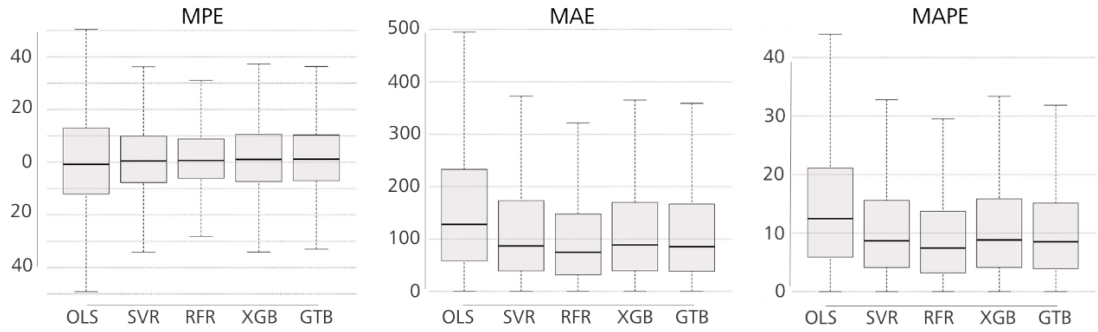
Error measure	Unit	OLS	SVR	GTB	XGB	RFR
MAE	EUR/p.m.	179.31	135.71	130.73	136.02	116.16
	EUR/sqm/p.m.	2.34	1.77	1.71	1.78	1.52
RMSE	EUR/p.m.	269.81	216.83	203.62	217.63	185.82
MAPE	%	15.60	11.63	11.36	11.72	10.16
R^2	%	81.65	87.79	89.32	87.87	91.35
ME	EUR/p.m.	18.81	13.07	21.16	22.7	22.91
MPE	%	1.65	1.40	2.01	2.05	1.56
PE10	%	40.65	56.02	56.92	55.50	62.62
PE20	%	71.67	84.82	86.11	84.97	88.49
IQR	EUR/p.m.	257.14	176.34	171.47	180.13	153.84
COD	%	-24.23	27.52	11.47	12.52	18.02

Notes: This table reports the error-based measurements on the predictive performance through MAE, RMSE, MAPE and R^2. ME and MPE indicate over- or underestimation. PE10, PE20, IQR and COD show the dispersion. All measures are out-of-sample (test set) and are based on the calculations presented in Table 3.6 in the Appendix. Absolute values are reported in euro per month. Relative values are reported in percent.

With respect to the results displayed in Table 3.3 we find all ML methods to be more accurate in modelling rents than traditional OLS regression.[16] While OLS provides on average highest absolute rental estimation errors (MAE), we find all ML methods to considerably increase the model accuracy, with RFR being most accurate. Figure 3.2 shows the boxplots of the error distribution. The graphical analysis regarding median and quantiles underpin the findings. To illustrate these results, we convert the MAE to EUR/sqm, dividing it by the size of an average apartment of 76.49 sqm. The estimation error decreases from 2.34 EUR/sqm (OLS) to 1.52 EUR/sqm (RFR). Regarding the RMSE, which differs from the MAE by penalizing extreme deviations, the results show a similar picture. Compared to OLS, all ML methods are more robust to extreme deviations. These findings complement the results of Bogin and Shui (2020) and Pace and Hayunga (2020) for property prices estimations, who likewise determine the highest prediction accuracy for RFR.

While OLS shows an R^2 of 81.65%, GTB and RFR are able to explain approximately 90% of the deviation. Ho et al. (2021) find similar results for housing transactions. Wu et al. (2008) and Y. Chen et al. (2016) show SVR to be robust and also accurate in modelling property prices and rents. It is therefore not surprising that SVR works well in our setting (R^2 of 87.79%) and is similar to ensemble learners such as XGB and GTB.

Figure 3.2: Graphical Error-based Comparison of Model Performance at Market Level



Notes: The box represents 50% of the data within the quantiles 25 and 75%. The line measures the median, that is, the quantile 50%. The antennas cover the 5% and 95% range of the data.

A look at the MAPE shows that traditional OLS misestimates the observed rents by 15.60% on average, while RFR improves model accuracy with an average misspecification of about 10%. These findings corroborate the results of Hu et al. (2019), who also show the tree-based bagging algorithm RFR to be most suitable for modelling property rents. Regarding transactions prices, Baldominos et al. (2018) likewise highlight ensembles of regression trees to perform best.

[16] The complete results of the OLS estimation are displayed in the Appendix in Table 3.7.

As Fik et al. (2003) state, Freddie Mac early suggested that at least 50% of the predicted sale prices of residential properties should be within 10% of the true value. In common real estate valuation practice, the estimated value of a property is allowed to vary 10% to 20% from its market value. Transferring this to rents, all our models yield satisfactory results. As Figure 3.2 shows, the median percentage deviation of all ML methods, as displayed in the boxplots, is below 10%. Therefore, we conclude ML algorithms to be capable of precisely modelling rents.

Aside from the previously analyzed accuracy, the quality of an estimation is additionally influenced by its precision which indicates whether hedonic models predict values that are on average above or below observed rents. In the field of property valuations, Bogin and Shui (2020) find real estate prices often to be overestimated, resulting in problems for mortgage lending. In the case of residential rents, we propose overestimated rental values to be less problematic for market participants, given that tenants are expected to react to landlords' high rental expectations with contract negotiations. In contrast, underestimations would lead to rental values that are below market level and mean landlords miss income. In Table 3.3, the positive MPEs indicate that all methods underestimate the observed rental value on average.

In addition, the dispersion of the estimation adds another possibility for investigation. The boxplots of MPE in Figure 3.2 show a symmetric distribution of all methods, indicating no general bias for traditional as well as ML variants. PE10 calculates the percentage of observations with a deviation of less than 10%. This metric can also be referred to as 'hit rate'. While OLS can estimate 40.65% of all observations within this range, algorithm-driven RFR models estimate 62.62% correctly. Within a deviation of +/-20%, we find all ML methods exceed 84%. The IQR draws a similar picture. While OLS estimates 50% of all observed values within a range of 1.68 EUR/sqm above or below the median, the ML models significantly decrease the range of deviations (+/-1.00 EUR/sqm).[17] The COD also confirms these results. Thus, ML methods are not only more accurate on average, but the error dispersion is also lower leading to a better predictive performance.

To verify the robustness of our results especially in terms of general applicability, we run all methods on an additional sample of rents from July 2019 to September 2019. The model specifications are the same as in the previous analysis (January 2013 to June 2019). The results are presented in Table 3.8 and Figure 3.4 in the Appendix. They consequently provide error-based measurements for a one-period-ahead out-of-sample forecast. Our

[17] The range as EUR/sqm is calculated by dividing the IQR by the average size of an apartment as reported in the descriptive statistics. Because the IQR displays the distance between the q25 and q75, we can therefore show the interval that comprises 50% of all estimations.

findings are equivalent to the findings in the original dataset. An upward shift in all error-based measurements can be traced back to thriving residential real estate markets in German metropolitan areas – especially in Munich. Bogin and Shui (2020) find RFR to be prone to overfitting. We can corroborate their results. While RFR performs best when it comes to the original dataset, we now find all other ML methods to be more accurate in forecasting future rents. Regarding RMSE as well as PE10 and PE20, the results indicate that RFR seems to show some misspecification for high deviations. We suggest RFR fits extreme values generally well (lowest RMSE in Table 3.3) but fails to explain them within new sample of future rents (as it shows the highest RMSE besides OLS, but good results for PE10 and PE20 in Table 3.8 in the Appendix).

To summarize, the key facts in the first part of our analysis are:

- In terms of accuracy, all ML methods are more accurate in modelling rents than OLS with RFR performing best.
- All methods underestimate observed values on average although the extent of underestimation is low.
- ML methods bear less risk than OLS due to a lower amount of misspecification.
- SVR shows similar results to the tree-based ML methods (RFR, GTB and XGB).
- RFR appears to be prone to overfitting whereas boosting methods (GTB and XGB) are more robust.

Altogether, a reasonable explanation for the better performance of ML methods can be given by the fact that they are able to capture non-linear and non-normal relationships (Pace & Hayunga, 2020; Bogin & Shui, 2020). Because non-linearity is an important characteristic of real estate markets, the application of ML techniques provides more accurate estimates of residential rents.

3.6.2 Rental Prediction at Portfolio Level

The previous results demonstrate that both traditional and ML methods can mimic the price formation in residential rental markets. By means of the previous model specifications, the models can estimate a rental value an investor could expect in a re-letting scenario. We transfer this knowledge to the portfolio data described in section 3.4.2 to estimate a rent for every apartment based on their hedonic, socio-economic and spatial characteristics. A comparison of the estimated rent with the actual contract rent provides information on the feasibility of rental adjustments when re-letting apartments from the portfolio. In a first step, we use MAE, RMSE and MAPE to analyze the accuracy.

Table 3.4: Error-based Comparison of Model Performance at Portfolio Level

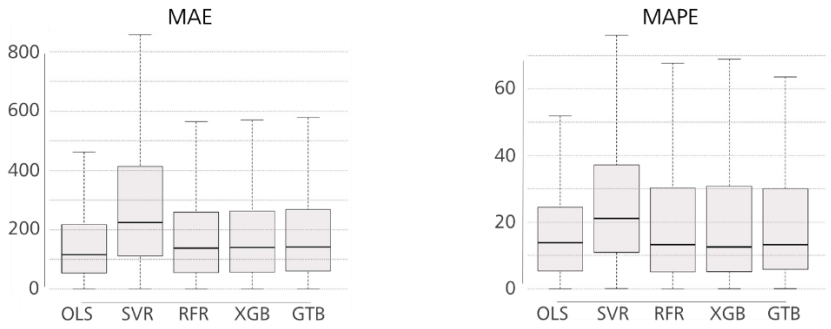
Error measure	Unit	OLS	SVR	GTB	XGB	RFR
MAE	EUR/p.m.	158.64	268.51	197.79	195.59	168.44
	EUR/sqm/p.m.	2.20	3.73	2.75	2.72	2.34
RMSE	EUR/p.m.	211.29	323.94	256.58	261.39	222.84
MAPE	%	15.70	25.83	17.74	17.64	16.24
PE20	%	68.44	45.39	62.43	62.43	63.39

Notes: This table reports the model accuracy through MAE, RMSE and MAPE. PE20 shows the dispersion. All measures are based on the calculations presented in Table 3.6 in the Appendix. Absolute values are reported in euro per month. Relative values are reported in percent.

In Table 3.4, OLS displays the lowest absolute error. All ML methods show a considerably higher deviation within their estimation. While OLS only allows for an average estimation error of 2.20 EUR/sqm, tree-based methods RFR, GTB and XGB result in an average deviation of 2.34 to 2.75 EUR/sqm. RMSE and MAPE underpin these findings. Interestingly, this is contrary to the previous findings using MLS Data. Hence, a look at the models' 'hit rate' reveals the following: While tree-based methods can estimate about 63% of all observed rents within a deviation of +/-20%, OLS is able to model 68.44% accurately. For the portfolio data, we can consequently conclude that linear OLS leads to more accurate estimates. The graphical illustration is shown in Figure 3.3.

Furthermore, it is noticeable that SVR shows the highest deviation of portfolio rents from estimated rents, with an MAE of 3.73 EUR/sqm, which requires a deeper discussion. SVR is very sensitive to the choice of support vectors and tends to neglect the informational content of observations within the threshold area that defines the hyperplane. Because investors usually follow predefined investment goals when acquiring their portfolio apartments, specifications in the portfolio dataset can result in biased estimations of rental values for the portfolio observations when applying SVR. We assume its poor performance to be attributed to the difficulties encountered in correctly modelling the portfolio data and therefore exclude SVR in the following comparison.

Figure 3.3: Graphical Error-based Comparison of Model Performance at Portfolio Level



Notes: The box represents 50% of the data within the quantiles 25% and 75%. The line measures the median, that is, the quantile 50%. The antennas cover the 5% and 95% range of the data.

Regarding the interpretation of the results in this section, however, one must keep the following in mind: A low error measurement (and therefore a low average deviation) indicates that estimated rents are to a large extent in line with observed contract rents. Because estimated rents represent a rental value a landlord could expect in re-lettings, OLS (with the lowest error measures) would indicate a low potential (or need) for rental adjustments. In contrast, ML models show considerably higher deviations. Because these models have confirmed a higher predictive performance in 3.6.1 on the MLS dataset, we would assume that estimates from ML models more accurately reflect the potential rental value in re-letting. An investor who bases the rental estimation on OLS would consequently underestimate possible rental changes in upcoming re-letting negotiations. Given the estimated rents from OLS are in line with contract rents to a higher degree, we assume investors to 'think linear'. The results indicate that investors use linear models within their rental estimation, although ML methods can identify higher rental potentials.

Table 3.5: Average Potential for Rental Increases

	As % of contract rents (MPE)			As rent in EUR/sqm (ME/sqm)		
Method	**All**	**q5 & q95**	**q10 & q90**	**All**	**q5 & q95**	**q10 & q90**
OLS	-4.95% *	-4.85%*	-4.75%*	-0.87*	-1.02*	-1.09*
GTB	-14.81% *,**	-14.13%*,**	-13.64%*,**	-2.29*,**	-2.32*,**	-2.34*,**
XGB	-14.56%*,**	-13.99%*,**	-13.59%*,**	-2.21*,**	-2.30*,**	-2.36*,**
RFR	-12.54%*,**	-12.10%*,**	-11.91%*,**	-1.67*,**	-1.82*,**	-1.92*,**

Notes: This table reports the average rental lift potential. Relative values are calculated as the difference between contract rent and estimated rent as % of contract rent. Absolute values are calculated as the same difference divided by the rental area. The column 'All' includes results for the whole sample, while q5 & q95 excludes observations of the highest and lowest 5% quantile and q10 & q90 of the highest and lowest 10% quantile, respectively.
*denotes whether the mean is significantly different from the observed mean on a significance level of 1%.
** denotes whether the mean is significantly different from the OLS mean on a significance level of 1%.

To assess to which extent this rental potential exists and consequently whether portfolio apartments are under- or overrented, we calculate the relative difference of estimated rents to contract rents. According to the results in Table 3.5, all models indicate that contract rents are below estimated rents. While OLS indicates portfolio apartments to be

underrented by 4.95% (0.87 EUR/sqm) on average, algorithm-driven hedonic models signal contract rents to be 12.54% (1.67 EUR/sqm) (RFR) to 14.81% (2.29 EUR/sqm) (GTB) below estimated rents. Our results are robust even if we exclude the highest and lowest 5%-quantile and 10%-quantile, respectively. The fact that all models show underrent situations is intuitive, especially in metropolitan areas in Germany, since rental growth in the residential real estate market exceeds inflation and hence, contract rents lag behind.

However, the difference between the methods is of special interest. An investor using OLS underestimates the rental-lift potential in his portfolio. By 'thinking linear' when researching the market, he assumes that contract rents are in line with estimated rents to a high extent. In contrast, our study reveals that ML methods show the potential for rental increases to be two to three times higher. In fact, we assume the potential to be at the level of the results of GTB and XGB, since boosting methods have shown to be more robust than RFR.

However, given current market practice, the following must be considered additionally: Contractual arrangements on lease term and rental adjustments, specific regulations in rental markets and further legal peculiarities between landlords and tenants impede the realization of the full rental potential. Nevertheless, the sole identification in this case provides investors with valuable possibilities to derive investment decisions. Aside from the linearity perception of an investor, another possible reason contributing to OLS' high performance, is the rather homogenous composition of the portfolio, whose data structure can be well captured by linear models. Moreover, considering the general economics of property management, another possible explanation becomes apparent: A residential manager is contractually not incentivized to achieve the highest rents but rather to focus on minimizing costs, again, favoring OLS which does not capture high rental deviations. These complementary explanations should be examined in more detail if the ML methods are to be used in real case scenarios.

3.7 Conclusion

In this study we investigate the predictive performance of traditional and algorithm-driven hedonic models and the added value an application of those methods can provide for market participants in the residential real estate market. In the first part of our analysis, both traditional linear and ML methods perform well in explaining residential rents. However, algorithm-driven models are more accurate: While OLS on average misestimates observed market rents by 15.60% (2.34 EUR/sqm), tree-based RFR shows the highest accuracy by reducing the absolute estimation error to 10.16% (1.52 EUR/sqm), followed by boosting methods. Hence, ML methods provide a valuable alternative for modelling market rents. However, it is important to bear in mind that these techniques tend to underestimate, resulting in below-market rental expectations in contract negotiations. Moreover, we find the bagging approach of RFR to be prone to overfitting. We suggest the use of boosting methods GTB and XGB to lead to more robust rental estimations.

Transferring these findings to an institutionally managed portfolio, we obtain the following insights: OLS indicates that contract rents are only 4.95% below estimated rents. In contrast, ML methods – which have shown to be more accurate in modelling rents – identify potential for rental increases that is two to three times higher. Given these contrasting results, we assume investors 'think linear' and make use of OLS findings when determining rental values; for example in contract negotiations. That being said, the application of ML methods can provide added value in residential portfolios by revealing considerable potential for rental adjustments that have not been identified by more traditional approaches. Nevertheless, complementary explanations for findings of this kind should be considered when applying ML to day-to-day operations.

Practical implications of our study are manifold. Whereas investment managers gain insights to rethink and structure their portfolios, governmental bodies and policy makers can evaluate housing policies in a timely manner by showing the impact on residential markets. Possible applications of artificial intelligence are consequently not limited to the private sector. Since almost every investigation is confronted with limitations, a thorough reflection is appropriate when comparing or applying findings in other scenarios. The first part of our analysis uses asking data, which is considered a valuable proxy for timely rents. However, deviations to transaction data can occur. Moreover, since our study covers a period with stable economic conditions, it would be interesting to see how the models react to stagnating or downturn markets. Also, our analysis solely focuses on the residential market, which further limits the general applicability since we assume that algorithms may behave differently when learning from office or retail data.

While traditional models remain an important and valid tool in hedonic modelling, ML models provide beneficial insights into rental markets and portfolios. Overall, we assume an increasing number of AI applications to lead to additional ideas and added value in research and practice. Future research in this area may further expand this knowledge since new algorithms and methods are constantly being developed. Expanding data sets by investigating other markets will strengthen the use of ML methods in the area of real estate in the future.

3.8 Appendix

Table 3.6: Error-based Measurements on the Predictive Performance

Accuracy		
Mean Absolute Error (MAE)	$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Average of all absolute errors. Lower MAE signals higher precision in units
Root Mean Squared Error (RMSE)	$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Average of squared residuals. In contrast to MAE, RMSE penalizes high deviations
Mean Absolute Percentage Error (MAPE)	$MAPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$	Average of all absolute percentage errors. Lower MAPE signals higher accuracy in percent
R^2	$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Goodness of fit of the model
Precision		
Mean Error (ME)	$ME(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$	Average of difference between observed and predicted value
Mean Percentage Error (MPE)	$MPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)$	Positive and negative errors cancel out due to the lacking absolute value operation. Positive (negative) MPE signals underestimation (overestimation)
Dispersion		
Error buckets (PE(x))	$PE(x) = 100\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert < x$	Percentage of predictions where the percentage error is less than x%, with x being set to 10 and 20
Coefficient of Dispersion (COD)	$COD = \frac{100}{n}\frac{\sum_{i=1}^{n}(\frac{\hat{y}_i}{y_i} - Median\left(\frac{\hat{y}_i}{y_i}\right))}{Median(\frac{\hat{y}_i}{y_i})}$	Ratio of the mean deviation from prediction errors to the median prediction error, divided by the median
Inter-Quartile Range (IQR)	$IQR = (y_i - \hat{y}_i)_{75} - (y_i - \hat{y}_i)_{25}$	Range in terms of the difference between the 75th and 25th percentile of the distribution of the prediction error

Table 3.7: Results of the OLS Estimation

Variable	Estimate	Std. Error	t value	sign. level
log Living Area	0.918	0.002	422.377	***
Age relative to 2017	-0.001	0.000	-16.385	***
log Centroid ZIP	-0.014	0.001	-9.572	***
log Centroid NUTS	-0.036	0.002	-18.420	***
log Number of households (HH)	-0.730	0.035	-23.916	***
log Household purchasing power	3.400	0.151	22.569	***
log Bus	-0.028	0.001	-19.727	***
log Park	-0.017	0.001	-12.122	***
log School	0.002	0.001	1.181	
log Subway	-0.020	0.001	-13.561	***
log Supermarket	0.001	0.001	0.806	
Bathtub	-0.012	0.002	-6.616	***
Built-in kitchen	0.052	0.002	25.868	***
Parking lot	0.019	0.002	8.126	***
Terrace	0.023	0.003	8.734	***
Balcony	-0.012	0.002	-5.811	***
Elevator	0.070	0.002	34.014	***
Intercept	-3.451	0.401	-8.601	***
Time dummies	Yes			

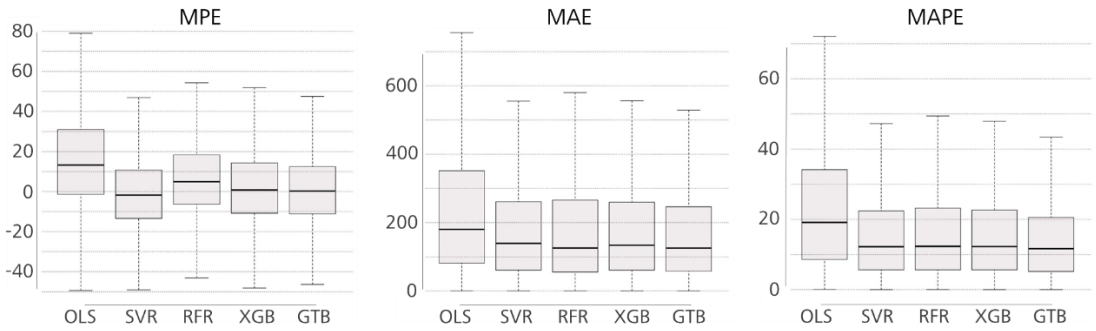
Notes: The dependent variable is log rent per month per apartment. The OLS model delivers an adjusted R^2 of 80.42% calculated in-sample (training set). ***, ** and * represent statistical significance at 0.01, 0.05 and 0.10 levels, respectively.

Table 3.8: Error-based Comparison of Model Forecasting at Market Level

Error measure	Unit	OLS	SVR	GTB	XGB	RFR
MAE	EUR/p.m.	271.14	201.84	189.52	203.61	212.35
	EUR/sqm/p.m.	3.54	2.64	2.48	2.66	2.78
RMSE	EUR	418.86	303.76	292.20	320.88	366.93
MAPE	%	24.00	15.69	15.02	16.08	16.77
R^2	%	80.12%	84.39%	86.39%	84.60%	83.93%
ME	EUR	177.59	10.54	26.13	42.26	108.20
MPE	%	15.47	1.16	1.13	1.94	6.61
PE10	%	29.54%	42.12%	44.38%	41.42%	43.21%
PE20	%	57.22%	70.46%	74.22%	70.89%	72.75%
IQR	EUR	322.61	275.89	258.27	274.44	273.39
COD	%	1.94	-9.73	79.99	25.80	4.09

Notes: This table reports the error-based measurements on the predictive performance through MAE, RMSE, MAPE and R^2. ME and MPE indicate over- or underestimation. PE10, PE20, IQR and COD show the dispersion. All measures are out-of-sample (test set) and are based on the calculations presented in Table 3.6. Absolute values are reported in euro per month. Relative values are reported in percent.

Figure 3.4: Graphical Error-based Comparison of Model Forecasting at Market Level



Notes: The box represents 50% of the data within the quantiles 25% and 75%. The line measures the median, that is, the quantile 50%. The antennas cover the 5% and 95% range of the data.

3.9 References

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778.

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. *Applied Sciences*, *8*(11), 2321.

Banzhaf, H. S., & Farooque, O. (2013). Interjurisdictional housing prices and spatial amenities: Which measures of housing prices reflect local public goods? *Regional Science and Urban Economics*, *43*(4), 635–648.

Bogin, A. N., & Shui, J. (2020). Appraisal Accuracy and Automated Valuation Models in Rural Areas. *The Journal of Real Estate Finance and Economics*, *60*(1-2), 40–52.

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, *35*(2), 143–160.

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence: a comparison of alternative methods. *Journal of Real Estate Research*, *32*(2), 139–159.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cajias, M. (2018). Is there room for another hedonic model? The advantages of the GAMLSS approach in real estate research. *Journal of European Real Estate Research*, *11*(2), 224–245.

Cajias, M., & Ertl, S. (2018). Spatial effects and non-linearity in hedonic modeling. *Journal of Property Investment & Finance*, *36*(1), 32–49.

Cajias, M., & Freudenreich, P. (2018). Exploring the determinants of liquidity with big data – market heterogeneity in German markets. *Journal of Property Investment & Finance*, *36*(1), 3–18.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 785–794.

Chen, Y., Liu, X., Li, X., Liu, Y., & Xu, X. (2016). Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning. *Applied Geography*, *75*, 200–212.

Chin, S., Kahn, M. E., & Moon, H. R. (2020). Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach. *Real Estate Economics*, *48*(3), 886–914.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Fik, T. J., Ling, D. C., & Mulligan, G. F. (2003). Modeling spatial variation in housing prices: a variable interaction approach. *Real Estate Economics*, *31*(4), 623–646.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Gallin, J. (2008). The long-run relationship between house prices and rents. *Real Estate Economics*, *36*(4), 635–658.

Genesove, D. (2003). The nominal rigidity of apartment rents. *Review of Economics and Statistics*, *85*(4), 844–853.

Goodman, A. C. (1978). Hedonic prices, price indices and housing markets. *Journal of Urban Economics*, *5*(4), 471–484.

Goodman, J. (2004). Determinants of operating costs of multifamily rental housing. *Journal of Housing Economics*, *13*(3), 226–244.

Gröbel, S. (2019). Analysis of spatial variance clustering in the hedonic modeling of housing prices. *Journal of Property Research*, *36*(1), 1–26.

Gröbel, S., & Thomschke, L. (2018). Hedonic pricing and the spatial structure of housing data–an application to Berlin. *Journal of Property Research*, *35*(3), 185–208.

Hamilton, T. L., & Johnson, E. B. (2018). Using Machine Learning and Google Street View to Estimate Visual Amenity Values. Working Paper, University of Richmond, University of Alabama.

Han, L., & Strange, W. C. (2016). What is the role of the asking price for a house? *Journal of Urban Economics*, *93*, 115–130.

Hanson, A., & Hawley, Z. (2011). Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of Urban Economics*, *70*(2-3), 99–114.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). Berlin: Springer.

Ho, W. K., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70.

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

IMF. (2018). *Global Financial Stability Report*. International Monetary Fund.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (6th ed.). New York: Springer.

James, V., Wu, S., Gelfand, A., & Sirmans, C. (2005). Apartment rent prediction using spatial modeling. *Journal of Real Estate Research*, *27*(1), 105–136.

Jud, G. D., Seaks, T. G., & Winkler, D. T. (1996). Time on the market: the impact of residential brokerage. *Journal of Real Estate Research*, *12*(2), 447–458.

Jud, G. D., & Winkler, D. T. (1994). What do real estate brokers do: an examination of excess returns in the housing market. *Journal of Housing Economics*, *3*(4), 283–295.

Kee, K., & Walt, N. (1996). Assessing the rental value of residential properties: an abductive learning networks approach. *Journal of Real Estate Research*, *12*(1), 63–77.

Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *The Journal of Portfolio Management*, *43*(6), 202–211.

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, *11*(1), 443–448.

Lai, T.-Y., Vandell, K., Wang, K., & Welke, G. (2008). Estimating Property Values by Replication: An Alternative to the Traditional Grid and Regression Methods. *Journal of Real Estate Research*, *30*(4), 441–460.

Lam, K. C., Yu, C. Y., & Lam, C. K. (2009). Support vector machine and entropy based decision support system for property valuation. *Journal of Property Research*, *26*(3), 213–233.

Li, L., & Yavas, A. (2015). The impact of a multiple listing service. *Real Estate Economics*, *43*(2), 471–506.

Lin, Z., Rosenblatt, E., & Yao, V. W. (2009). Spillover effects of foreclosures on neighborhood property values. *The Journal of Real Estate Finance and Economics*, *38*(4), 387–407.

Lindenthal, T. (2020). Beauty in the Eye of the Home-Owner: Aesthetic Zoning and Residential Property Values. *Real Estate Economics*, *48*(2), 530–555.

Lindenthal, T., & Johnson, E. B. (2020). Machine Learning, Architectural Styles and Property Values. Working Paper, University of Cambridge, University of Alabama.

Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research*, *12*(1), 134–150.

Pace, R. K., & Hayunga, D. (2020). Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests. *The Journal of Real Estate Finance and Economics*, *60*(1-2), 170–180.

Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*, *36*(1), 59–96.

R Core Team. (2020). *R: A language and environment for statistical computing*.

Rae, A. (2014). Online Housing Search and the Geography of Submarkets. *Housing Studies*, *30*(3), 453–472.

Rondinelli, C., & Veronese, G. (2011). Housing rent dynamics in Italy. *Economic Modelling*, *28*(1-2), 540–548.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *The Journal of Political Economy*, *82*(1), 34–55.

Schulz, R., Wersing, M., & Werwatz, A. (2014). Automated valuation modelling: a specification exercise. *Journal of Property Research*, *31*(2), 131–153.

Shimizu, C., Nishimura, K. G., & Watanabe, T. (2016). House prices at different stages of the buying/selling process. *Regional Science and Urban Economics*, *59*, 37–53.

Sirmans, S., Macpherson, D., & Ziets, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*(1), 1–44.

Sirmans, S., Sirmans, C., & Benjamin, J. (1989). Determining apartment rent: the value of amenities, services and external factors. *Journal of Real Estate Research*, *4*(2), 33–43.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Thomschke, L. (2015). Changes in the distribution of rental prices in Berlin. *Regional Science and Urban Economics*, *51*, 88–100.

Turnbull, G. K., & Dombrow, J. (2006). Spatial competition and shopping externalities: Evidence from the housing market. *The Journal of Real Estate Finance and Economics*, *32*(4), 391–408.

ULI. (2020). *Promoting Housing Affordability*. Urban Land Institute.

van Wezel, M., Kagie, M. M., & Potharst, R. R. (2005). *Boosting the accuracy of hedonic pricing models.*

Verbrugge, R., Dorfman, A., Johnson, W., Marsh III, F., Poole, R., & Shoemaker, O. (2017). Determinants of Differential Rent Changes: Mean Reversion versus the Usual Suspects. *Real Estate Economics*, *45*(3), 591–627.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.

Yao, Y., Zhang, J., Hong, Y., Liang, H., & He, J. (2018). Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Transactions in GIS*, *22*(2), 561–581.

Yavas, A., & Yang, S. (1995). The strategic role of listing price in marketing real estate: theory and evidence. *Real Estate Economics*, *23*(3), 347–368.

Yoo, S., Im, J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, *107*(3), 293–306.

Zhang, L., & Yi, Y. (2017). Quantile house price indices in Beijing. *Regional Science and Urban Economics*, *63*, 85–96.

Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research*, *33*(3), 349–387.

4 Peeking inside the Black Box: Interpretable Machine Learning and Hedonic Rental Estimation

4.1 Abstract

While Machine Learning (ML) excels at predictive tasks, its inferential capacity is limited due to the complex non-parametric structure. This paper aims to elucidate the analytical behavior of ML in real estate through Interpretable Machine Learning (IML). After estimating residential rents for Frankfurt am Main (Germany) with a hedonic ML approach, we apply a set of model-agnostic interpretation methods. Our results suggest that IML methods permit a peek into the 'black box' of algorithmic decision making by illustrating the relative importance of hedonic variables and their relationship with rental prices.

Keywords: Hedonic modeling, residential real estate, rental estimation, interpretable machine learning, black box

Acknowledgments: The authors especially thank PATRIZIA AG for contributing to this study. All statements of opinions are those of the authors and do not necessarily reflect the opinion of PATRIZIA AG or its associated companies.

4.2 Introduction

Possible applications of Artificial Intelligence (AI) and Machine Learning (ML) are manifold and are rapidly gaining importance across a number of domains. While most members of the general public interact with ML algorithms on a daily basis (e.g. personalized web ads, mail spam filter, etc.), there is also a growing number of discoveries and implementations in research. Recently, Deepmind and its interdisciplinary research team solved one of the biggest challenges in biology with their AI-based system AlphaFold to predict how proteins fold – a problem that has been investigated for nearly 50 years (Senior et al., 2020). Further high stake domains include arrival planning in emergency department and cancer diagnosis in healthcare (Ahmad et al., 2018) or recidivism forecasting in criminal justice (Berk & Bleich, 2013).

But how is it that these methods are only gradually coming to the fore? The high predictive performance marks ML as a promising extension for existing regression as well as classification tasks due to their ability to incorporate complex patterns and deal with large datasets. However, because the methods are often perceived as opaque, their so-called 'black box' character is repeatedly criticized. Certain use cases such as an AI-based decision support of credit applications may improve and accelerate business operations of banks, however the sole decision of whether a credit may be granted or denied lacks accountability and does not represent a satisfactory outcome for neither the applicant nor the creditor. Consequently, explaining the inner working of an ML model is important to justify and validate how a certain decision is made as well as to discover new insights (Adadi & Berrada, 2018).

A similar picture can be seen for the application of AI in the real estate industry. Because real estate represents one of the largest asset classes worldwide (Kok et al., 2017), an adequate estimation of real estate prices and rents are of crucial importance for investors, landlords and tenants. By treating the property as the sum of its individual characteristics, the hedonic price regression has established itself as the main approach for price and rent estimation. ML models have proven to be helpful in real estate hedonic modelling especially for predictive purposes. Nevertheless, their inferential capabilities are limited, since the aforementioned missing transparency hides the inner logic and decision making process (Mullainathan & Spiess, 2017). But how to overcome this obvious weakness? One possibility is to design models in such a way that their complexity is kept low from the beginning to ensure interpretability. An example comes from Lechner et al. (2020), who have created a deep learning algorithm that manages to control a car based on only a few artificial neurons. As a result, the decisions made by the algorithm are easy to understand

while maintaining robustness and functionality. Another possibility is to examine existing ML algorithms and their results with special analysis tools in order to establish interpretability. This is where this study picks up. The ML algorithm eXtreme Gradient Boosting (XGB) is used for a hedonic estimation of rents in the city Frankfurt am Main, Germany, and forms the basis for the application of Interpretable Machine Learning (IML) methods. Different model-agnostic tools such as feature importance and feature effects are applied to illustrate how hedonic characteristics contribute to the final prediction of the applied ML model. To the best of the authors' knowledge, this is the first real estate related study to use ex-post IML methods to justify machine-based decision-making on the one hand, and on the other hand, to gain further insights into the individual value of certain hedonic characteristics of an apartment.

4.3 Literature Review

For decades, hedonic models have formed the basis for empirically assessing prices and rents of properties based on their characteristics, such as amenities or location. A hedonic model estimates the effects of these characteristics by bundling them into a function and can thus determine the price of a property. The approach is commonly used because the concept offers many possible applications for a wide variety of problems.

According to Sirmans et al. (2005), origins of the hedonic model do not go back to just one founding father. Whereas Court (1939) first used a hedonic procedure to determine automobile prices, Lancaster (1966) and Rosen (1974) paved the way for the application in real estate. Since then, a large body of literature has emerged dealing with issues surrounding the relationship between the price or rent of a property and its characteristics. Essays by Sheppard (1999), Malpezzi (2002) and Sirmans et al. (2005) provide an overview of the diversity, but also the complexity of the questions that arise within hedonic research. However, the starting point is, as so often, the underlying data set or the available features of a property. Dubin (1988) argues that building characteristics that usually determine prices in a hedonic model can be grouped into three categories: Structural, location and neighborhood variables. Can (1992) and Stamou et al. (2017) define them as follows: Structural variables describe the nature of an apartment, such as its size, the number of rooms or the age of the property. Location variables, on the other hand, such as distance to the central business district (CBD), define the geographic location. Neighborhood variables tie in here and illustrate the socio-economic environment such as household income or the physical make-up of the closer environment. Often, the location and

neighborhood variables are considered together, as sometimes the distinction is not evident (Can, 1992, Haider & Miller, 2000, Des Rosiers et al., 2011, Stamou et al., 2017). In the recent past, much of the focus of studies has been on the effect of these locational or neighborhood characteristics. Within this group, variables of interest come mainly from the environmental, infrastructure and social domains. With respect to features in the immediate environment of a property, Dumm et al. (2016), Rouwendal et al. (2017) and Jauregui et al. (2019) analyze the effect of proximity to water on price. Studies by Below et al. (2015) and Dumm et al. (2018) show the price impact of nearby subsurface conditions such as sinkholes or land erosion. Other issues such as the influence of distance to urban green spaces (Conway et al., 2010) or the presence of air pollution (Fernández-Avilés et al., 2012) also receive attention. Considering the group of neighboring infrastructural facilities and their impact on properties, different studies emerged. Hoen et al. (2015), Hoen and Atkinson-Palombo (2016) and Wyman and Mothorpe (2018) study the effects of nearby electric facilities on property prices, such as wind turbines and power lines. Availability of transportation facilities such as of a highway and rail transit are investigated by Chernobai et al. (2011), Li (2020) and Chin et al. (2020). According to Theisen and Emblem (2018) and Zheng et al. (2016), the possibility of an easy access to early childhood education and training in the form of nearby kindergarten or schools is also a price-determining factor of residential properties. There are even more exotic themes such as the influence of strip clubs (Brooks et al., 2020) or the proximity to food trucks (Freybote et al., 2017). Nevertheless, factors in the immediate social environment can also play a role. For example, Goodwin et al. (2020) find that the presence of home ownership associations has price-determining effects. Seo (2018) shows that the neighborhood condition is similarly price determining.

When it comes to the model design, the usual hedonic approach involves a parametric, semi- or non-parametric multiple regression analysis, which uses a pooled data set of properties and their individual features. Interestingly, the development of improved computational capabilities has recently allowed other methods such as ML to complement this estimation process. While the parametric hedonic price regression approach is largely applied for inferential purposes, its potential for predictive tasks is rather limited (Pérez-Rave et al., 2019). The scope of ML methods, however, is the other way around. While inference has hardly played a role so far due to the mostly opaque algorithms, the predictive qualities of these methods are much more pronounced. ML algorithms, like gradient tree boosting (GTB) (Friedman, 2001), random forest regression (RFR) (Breiman, 2001a) and support vector regression (SVR) (Smola & Schölkopf, 2004), are capable of artificially learning from the underlying data and continuously improving their predictive

performance. Hence, these algorithms have shown remarkable accuracy. In the real estate literature, various studies demonstrate the performance of ML algorithms and parametric hedonic models, including Lam et al. (2009) and Kontrimas and Verikas (2011) for SVR, Yoo et al. (2012), Antipov and Pokryshevskaya (2012) and Yao et al. (2018) for RFR and van Wezel et al. (2005) and Kok et al. (2017) for boosting methods such as GTB. Furthermore, Zurada et al. (2011), Mayer et al. (2019) and Ho et al. (2021) document the performance of different ML methods.

However, these methods are viewed critically due to their black box character (McCluskey et al., 2013), since the final result often delivers the raw prediction without letting one know how it came to the respective conclusion. As Mayer et al. (2019) state, the predictive accuracy is only achieved by reduced comprehensibility of the ML models due to its ability to artificially capture highly complex pattern within the underlying data. In consequence, researchers are mostly faced with the trade-off between what is predicted (prediction) and why the prediction took place (inference).

In general, many ML methods, such as SVR, RFR and GTB, provide model transparency since there is an understanding of how the underlying algorithm works and the algorithm can be described mathematically without further knowledge of the data – although the structure of ML methods is increasingly complex. Nevertheless, model interpretability in terms of identifying and understanding what factors impact the final predictions seems to be the bottleneck for an overall acceptance and implementation of ML methods, because sole measures like predictive accuracy are "an incomplete description of most real-world tasks" (Doshi-Velez & Kim, 2017).

In the real estate literature, first approaches have been made to combine predictive and inferential purposes within a ML context. Pérez-Rave et al. (2019) propose a variable selection approach called "incremental sample with resampling" tested on two data sets of property prices. They apply random forests to varying subsamples to predict the final property prices. Variables are identified as important, if the feature is used in the final prediction rule of the RFRs for 95% of the subsamples. The final inferential interpretation is based on a parametric hedonic model using only the ML-selected variables. Moreover, Pace and Hayunga (2020) analyze the informational content of residuals from linear, spatial hedonic regression and ML models. After applying regression trees, they find that spatial information is still present in the residuals of ML models. Although single trees are easy to understand and their decision rule can be illustrated graphically, they show limited predictive performance and tend to be unstable due to high sensitivity to changes in the data or tuning parameter.

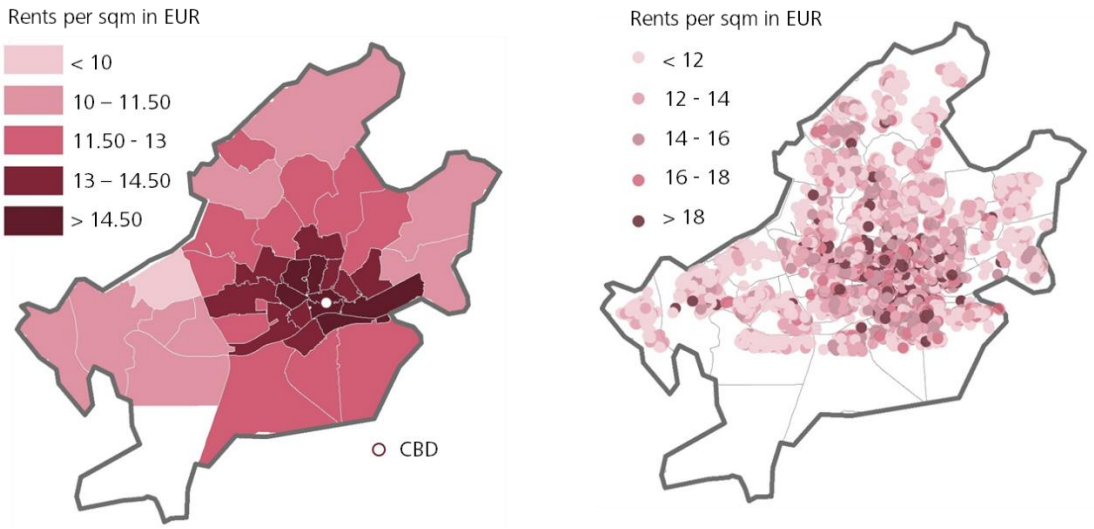
To conclude this section, this rather young field of research opens up the possibility to further engage with the interpretability of ML models and the impact of hedonic characteristics. In the following, we present the data set of our analysis and describe the methods we use to enable the interpretability of ML-based predictions. After that we discuss the results and summarize our findings in the conclusion.

4.4 Data

The sample for our analysis comprises 52,966 observations of residential rents in Frankfurt am Main, Germany. The country is the fourth largest economy worldwide and known as a safe haven for both domestic and cross-border real estate investments. With one of the lowest home ownership ratios of 51% being well below the European average, Germany is seen as a rental market rather than a homeowner market. Frankfurt represents the leading financial hub in continental Europe and is hosting the European Central Bank and the Frankfurt Stock Exchange amongst many important financial institutions. Its metropolitan region is home to more than 5.8 million inhabitants.

Rental data stems from Empirica Systeme, one of the largest German provider of real estate data, which comprises, amongst others, real estate listings of leading German Multiple Listing Systems (MLS). Data preparation and cleaning is performed to account for duplicates and erroneous data points. As the study focuses on the urban rental market in Frankfurt that is mainly determined by apartment rentals, we exclude single, semi-detached and terraced houses. We furthermore leave out student apartments, senior living accommodations, furnished co-living spaces, and short-stay apartments to control for highly specialized sub-markets that are expected to bias the overall rental market. Figure 4.1 provides two maps of the rental distribution in the data sample for Frankfurt. It highlights the average rent per sqm in every ZIP Code (left) and displays all observations gathered (right). Both maps indicate that the highest rents are found in the center, while lower rents tend to occur in the outskirts. There are no rental observations in the most southern part of Frankfurt due to highly forested areas and the airport of Frankfurt.

Figure 4.1: Distribution of Rents and Observations of the Frankfurt Data Sample



Notes: The left map shows average rents per sqm for each ZIP code. The right map depicts all observations. Both cover the Frankfurt city area from 2013 to 2019. The thin grey lines display the ZIP codes.

Besides the rent as target variable, the data contain information on structural characteristics in terms of living area, building age, floor and whether a kitchen, parking spot, balcony, terrace, bathtub and elevator is present or whether an apartment is refurbished. We add socio-economic data from Growth from Knowledge, Germany's largest market research institute. Since all rental data points are georeferenced, we are able to add a spatial gravity layer based on data from Eurostat, the German statistical office and Open Street Map to account for spatial information and therefore add several location variables. We include the distance to the CBD as well as to numerous important amenities. Proximity to bus and railway station account for public transport and accessibility. Bakery, supermarket, convenience and department store distances comprise the local supply. Bar, beergarden and café represent the access to hospitality. While distances to school and park allow insights on public amenities, proximity to car wash and traffic signal incorporate adverse effects mainly due to noise emissions.

MLS are frequently used in German rental markets from professional as well as from private landlords. Moreover, since neither landlords nor tenants are obliged to disclose contract information in Germany, listing data is the main source of information for both researchers and practitioners.[18] In addition, it should be noted that rental price formation in major German cities is generally dominated by the offering party since residential vacancy rates

[18] See e.g. Gröbel and Thomschke (2018) using German rental listing prices in research as well as well-established applications of listing data e.g. F+B Residential Index or Empirica Real Estate Index in practice.

in metropolitan areas are remarkably low.[19] A look at individual renting scenarios reveals that a landlord regularly receives inquiries in the double-digit range for an apartment that has been advertised. In consequence, the rental decision is not based on auction procedures but rather on timely application and best (personal and solvent) fit for the landlord. In the literature, Cajias and Freudenreich (2018) demonstrate that German residential markets are subject to low Time-on-Market and diminishing degrees of overpricing. As Gröbel (2019) suggests, asking data in Germany "reflect the currently prevailing overall market situation". Although we do not claim that rental listing precisely reflect the agreed contract rent, we expect the listing rents to be a useful framework for the ongoing analysis.

Table 4.1: Descriptive Statistics of the Data for Frankfurt am Main (2013 – 2019)

Variable	Unit	Mean	Median	Std.Dev
Rent	EUR/month	1,036.123	884	638.175
Living area	sqm	78.175	72	36.688
Floors	Integer	2.396	2	2.328
Age (relative to 2017)	Integer	49.377	48	39.701
Bathtub	Binary	0.564	1	0.496
Refurbished	Binary	0.242	0	0.428
Built-in kitchen	Binary	0.688	1	0.463
Balcony	Binary	0.633	1	0.482
Parking	Binary	0.487	0	0.500
Elevator	Binary	0.449	0	0.497
Terrace	Binary	0.136	0	0.342
Purchasing Power	EUR/HH/ZIP	50,390	49,993	5,798
CBD_distance	Km.	3.616	3.604	1.896
Bar_distance	Km.	0.722	0.511	0.636
Beergarden_distance	Km.	1.135	0.937	0.759
Cafe_distance	Km.	0.346	0.240	0.325
Bakery_distance	Km.	0.370	0.245	0.403
Convenience store_distance	Km.	0.849	0.589	0.748
Department store_distance	Km.	1.550	1.306	0.997
Supermarket_distance	Km.	0.252	0.223	0.167
Bus station_distance	Km.	3.062	2.667	1.566
Railway station_distance	Km.	0.835	0.581	0.685
Traffic signals_distance	Km.	0.186	0.157	0.135
Car wash_distance	Km.	1.266	1.234	0.584
Park_distance	Km.	0.266	0.236	0.158
School_distance	Km.	0.302	0.278	0.167

Notes: The table reports the summary statistics comprising data as of January 2013 to December 2019. Age is calculated as the difference of the building age to the year 2017. All distance variables are calculated as the distance to the specific dwelling in kilometers. Binary variables report whether the dwelling includes a certain characteristic (1) or not (0). Rent is presented as euro per month. Information on households (HH) is reported on ZIP level. SD: standard deviation, Min: minimum value, Max: maximum value.

[19] According to CBRE, the vacancy rate for residential real estate in the city of Frankfurt am Main marks 0.4% of the stock. Moreover, Immobilienscout 24, the leading online listing platform for real estate in Germany, reports 198 clicks on average for an online apartment advertisement.

Table 4.1 shows the descriptive statistics. We find a mean asking rent of 1,036.12 EUR p.m. (euros per month). An average apartment is 78.175 sqm located on the 2nd floor in a property that was built in 1968. The apartment contains a bathtub, a built-in-kitchen, a balcony, but neither a parking slot nor an elevator. On average, it is 3.62 km away from the CBD, 350 meters to the next café and 250 meters to the closest supermarket. The bus and railways station are 3 km and 0.84 km away, whereas the next school is located 300 meters nearby. The mean household purchasing power amounts to 50,390 EUR p.m..[20]

4.5 Methodology

ML has proven its predictive power in the literature and is commonly used by real estate professionals to inform their decision making (RICS, 2017). We apply a tree-based approach to build the foundation for further analysis. As Pace and Hayunga (2020) state, a regression tree (RT) is easy-to-understand while still being capable of identifying complex pattern. That is because trees can capture non-linear relationships as well as interactions. In its core, a RT can be understood as nested if-else conditions. Tree-based models divide the data in distinct subsets and make a prediction for every subset (which usually is the average outcome of all observations in the specific subset). The division is made by several splitting steps, in which iteratively a feature variable is chosen and its feature space is split in a way that a certain criterion is affected most (e.g. the prediction error is reduced most) until a stopping point is reached.

Since single trees are prone to misspecification, ensembles are used to aggregate and combine the prediction rule of multiple trees. We choose XGB as an ensemble boosting method, which has shown to be capable of accurately predicting property prices and rents and at the same time yield robust estimation results.[21] Developed by Chen and Guestrin (2016), it is a promising approach for regression, as well as for classification, as it contains specific features that won it several Kaggle[22] competitions in the recent past. In its basic concept, boosting fits an initial tree, calculates the residuals of the initial prediction, and fits another tree on the residuals to stepwise reduce the prediction error and incrementally enhance the final prediction rule. To prevent overfitting cross-validation is applied.

[20] In Table 4.3 in Appendix 3, we provide a full set of correlation coefficients for all variables.

[21] In general, tree-based ensemble algorithms are based on two different approaches, namely boosting and bagging. See e.g. Hastie et al. (2009) for a more detailed introduction to the fundamentals of ML models.

[22] Kaggle is one of the leading online platforms for the data science community and regularly hosts data competitions. For further information see https://www.kaggle.com

Because the internal logic and consequently the rationale behind the individual predictions is rather hidden, the use of ML often lacks transparency. In consequence, a growing body of literature on IML[23] has evolved in recent years to further 'improve trust' in algorithmic decisions (See e.g. Adadi & Berrada, 2018; Carvalho et al., 2019; Arrieta et al., 2020 or Linardatos et al., 2021). In general, tree-based ML methods show some sort of algorithmic transparency, since their underlying concept and theory is comprehensible and mathematically described (James et al., 2013). Nevertheless, it is not evident, which feature[24] and to what extent it contributes to the prediction.

One possibility to understand how predictions are achieved in this context is to use **interpretable ML models.**[25] Like in parametric models, specific restrictions limit the complexity of the model and therefore allow inferential insights. RTs are a well-known example of interpretable ML models if e.g. the depth of the tree is limited. As Molnar (2020) states, short trees with a depth up to three splits are interpretable in a comprehensive way, since a maximum combination of three if-else-conditions as the decision rule is enough to explain how the model yield a certain prediction.

Limiting the models complexity often results in depriving ML much of its effect, since their flexible structure enables a strong predictive performance (Breiman, 2001b)[26]. Consequently, (post-hoc) model-agnostic **interpretation methods** have been developed, which separate the explanatory framework and the ML model, thus preserving its predictive capabilities. In contrast to interpretable models, the ML model remains a black box, with the separated interpretation methods aiming at extracting interpretable information post-hoc. Model-agnostic tools benefit from their flexibility because they do not depend on a specific ML method and can be applied to various learners (Ribeiro et al., 2016).

Interpretation methods differ on whether their focus is on feature importance or feature effects. The first one aims at evaluating which feature contributes the most to the prediction, whereas the second one sheds light on how a single feature contributes to the prediction. The methods are perceived as typical and useful tools to show the impact of features in ML models and explain the inner working on a global level (Hastie et al., 2009).

[23] In the context of IML, the term Explainable Artificial Intelligence (XAI) is often used synonymously.

[24] To describe the covariates, hedonic literature mainly refers to them as variables or characteristics, while research on IML generally uses the term features.

[25] Interpretable ML models are also referred to as transparent models, since they are considered to be understandable by itself.

[26] See e.g. Shmueli (2010) for further discussion on the trade-off between model accuracy and interpretability.

We use the FeatureEffect and FeatureImp functions both implemented in the iml package in R (R Core Team, 2020).

Feature importance (FI) measures the relevance of a single feature for the prediction. The importance of a feature is calculated by permutation of the observed feature values and its effect on the prediction error, keeping all other features constant. Based on the concept of Breiman (2001a) for random forests, Fisher et al. (2019) provides a model-agnostic framework for measuring the covariates contribution to the accuracy of an ML model called 'model reliance'.

Let X be the feature matrix, Y the dependent variable and f the ML model, with the prediction error e being measured by a loss function $L(Y, f(X))$. The feature importance is defined as the ratio of the model error after permutation to the original model error before switching features.

$$FI(f) = \frac{e_{perm}(f)}{e_{orig}(f)} \tag{10}$$

The permutated error is thereby calculated as the expected error of the ML model based on the permuted feature matrix X_{perm}.

$$e_{perm}(f) = \mathrm{E}L(Y, f(X_{perm})) \tag{11}$$

To visualize the most important features, every variable is ranked and plotted according to their FI. Alternatively, the FI score can also be calculated as the difference of both errors, although the ratio provides the advantage of higher comparability. We use the Mean Absolute Error (MAE) as loss function. By switching the feature values of all observations (e.g. an observation with 1 for a kitchen being present is switched to 0), FI calculates how much this change leads to an observable decrease in prediction accuracy. It can consequently identify whether the specific feature contributes to the overall prediction or whether its change does not perceptibly affect the outcome. Lastly, we average the importance measures over 100 repeated permutations. As Fisher et al. (2019) states, FI is a helpful tool to identify influential features and increase the transparency of black box models.

In addition to the individual importance, **feature effects** show how a single feature influences the predicted outcome of an ML model. After the training process, a ML model has learned a specific relationship between the covariates and the target variable that can be analyzed. Partial Dependence (PD) plots visualize the marginal effects of features on the model's prediction (Friedman, 2001). The plots are based on partial dependence functions which highlight the effect of one feature on the target variable when the average effects

of all other features are accounted for. PD plots reveal useful information e.g. whether the relationship can be explained linearly or in a more complex manner.

Let once again X_j be the vector of the *j* variables and *n* be the number of observations. The PD is the effect of features of a subset X_S by marginalizing over all other features in the complement subset X_C (Zhao & Hastie, 2021). Given the ML model f, the partial function f_{x_S} is defined as:

$$f_{x_S}(x_S) = E_{x_C}[f(x_S, x_C)] = \int f(x_S, x_C)\, d\mathbb{P}(x_C) \tag{12}$$

With $d\mathbb{P}(x)$ being the marginal distribution of X_C. Marginalizing over all other features leads to a function that is solely dependent on the features X_S to be analyzed. The partial function f_{x_S} is estimated using the Monte Carlo method to average over actual features values $x_C^{(i)}$ while keeping X_S constant:

$$f_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} f\left(x_S, x_C^{(i)}\right) \tag{13}$$

As shown in Greenwell (2017), all values of feature x_S (e.g. living area) are in a first step replaced with the particular feature value (e.g. of the first observations). The ML model predicts expected output values for the newly created dataset (where all observations have the same constant feature value x_S). Averaging over these predictions calculates the marginal effect at the particular feature value. This step is repeated *n* times to obtain a marginal effect for all observed feature values. Finally, the single feature values are plotted against the resulting f_{x_S}. For a linear hedonic model, e.g. based on ordinary least squares (OLS), a PD plot would show a straight line representing the specific estimated coefficient. As Zhao and Hastie (2021) state, PD plots are a valuable visualization tool to interpret how the prediction of ML models depend on specific features.

4.6 Econometric Results

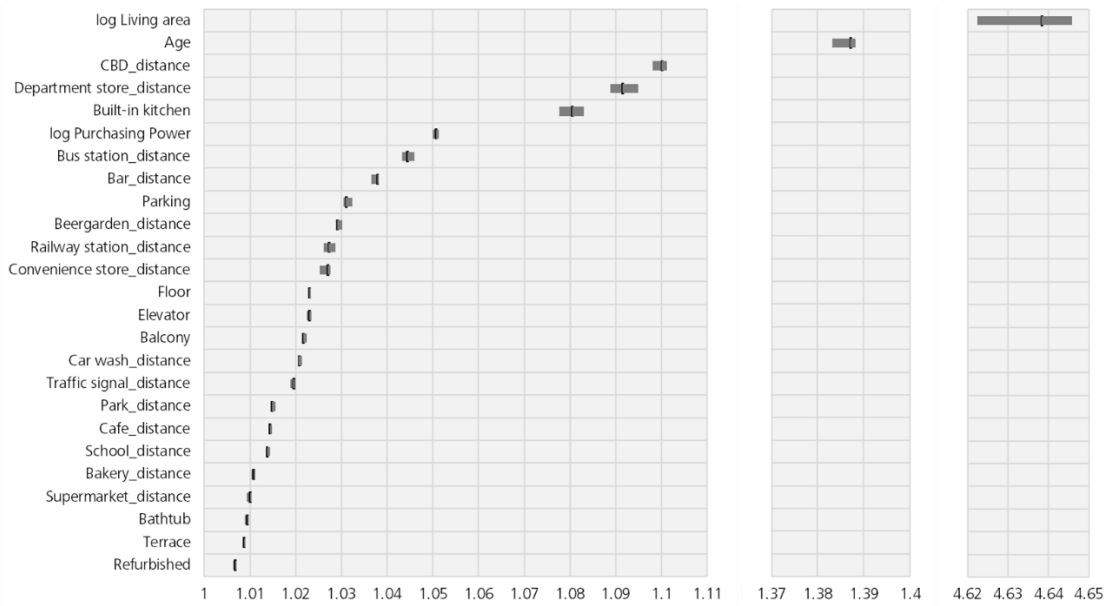
To set up a functional ML framework, we first train the XGB algorithm on our dataset of rental prices described in the data section. We apply random cross-validation with five folds and five repetitions. The tuning process takes 16 hours with 72 central processing units (CPUs) running simultaneously. The final XGB model is trained with η = 0.243, γ = 0.0431, λ = 28.99 and α = 22.64. The out of sample rental prediction with XGB yields to a R^2 of 92.50%. The mean absolute percentage error marks 11.13%. Moreover, 57.96% of all predictions deviate less than 10% from the observed values. The tuned XGB

algorithm subsequently allows a post-hoc analysis with a set of model-agnostic interpretation tools to identify feature importance and feature effects.[27]

4.6.1 Feature Importance of the Hedonic Characteristics

Figure 4.2 provides the relevance of all characteristics for the ML prediction based on FI. The features are individually ranked on the y-axis from most important at the top to least important at the bottom. The x-axis provides information of how much prediction accuracy changes when the feature values are permutated. Median values are plotted with the bar denoting the 5% and 95% quantiles. Feature importance ratios exceeding 1 indicate an observable impact on the overall prediction. Ratios that tend towards 1 imply a rather negligible influence of the features.

Figure 4.2: Feature Importance of the Hedonic Characteristics



Note: The figure displays the median values of the relative feature importance obtained with XGB. MAE is chosen as loss function. Variables are ranked based on their FI score. The bar denotes the 5% and 95% quantiles of the distribution of FI scores after 100 repetitions. A break in the horizontal axis is conducted to ease readability.

It is not surprising, that living area and age are seen to have by far the biggest impact on rental prediction. Their median values highlight that randomly permuting living area and age individually 100 times, increases the model error by a factor of 4.64 and 1.39, while keeping all other variables constant. Furthermore, distance to the CBD and to a department store are of high importance and associated with an increase in MAE of 1.10 and 1.09. We expect both variables to be a suitable proxy for a good location.[28] Moreover,

[27] To ensure basic hedonic functionality of a hedonic rent estimation, we apply linear, spatial and non-linear methods in advance. The corresponding methodology and the results are presented and discussed in Appendix 1 (methodology) and 2 (results). All variables show expected signs and do not contradict findings from related literature.

[28] In major German cities, department stores are usually located either close to the city center or in highly frequented and therefore good shopping locations.

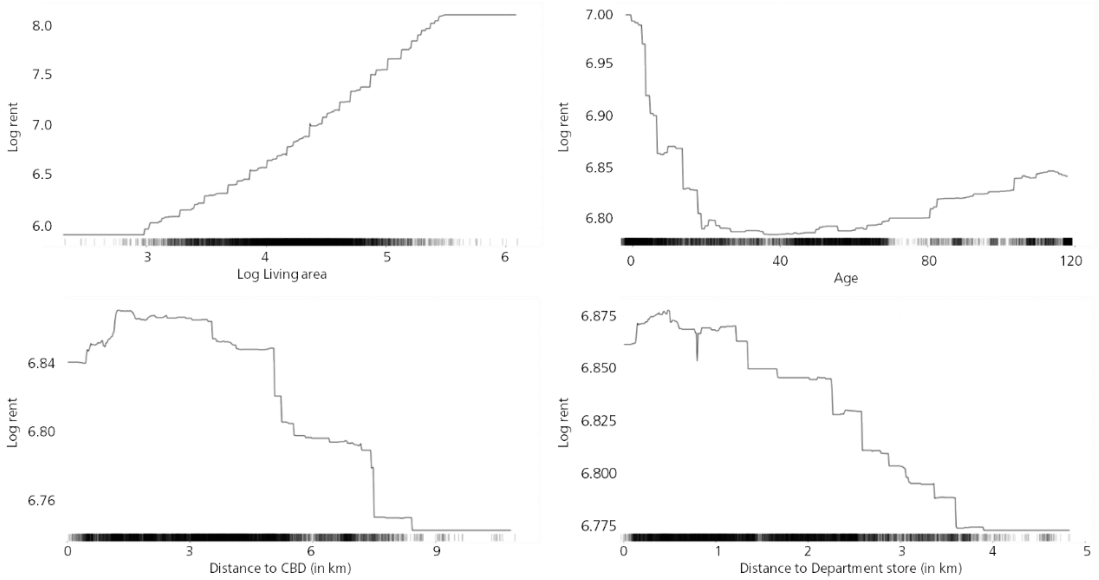
the presence of a built in kitchen is also heavy influential. The purchasing power per household is followed by the distances to the bus station and the next bar and beergarden.[29] The existence of a parking spot complements the ten most influential variables. We will not discuss the remaining variables in detail since their contribution seems rather marginal. The small distribution of FI for all variables demonstrated by the 5% and 95% quantile indicates that the results are stable over all repetitions. To summarize, feature importance ranks how relevant a variable is for the predictive task as it provides which variables are more or less influential for an ML model. One can thus obtain a first impression whether an algorithmic hedonic model delivers reliable results that are based on a plausible understanding of the economic context. However, FI does not provide any information about the sign. To clarify e.g. whether a small or large distance is decisive, we investigate feature effects in a next step.

4.6.2 Feature Effects of the Hedonic Characteristics

PD plots enable an analysis of how a certain feature influences the rental prediction and which relationships between residential rents and property characteristics has been traced by the algorithm. While the X-Axis provides information on the independent variable with the stacked black lines indicating the amount of observations, the Y-Axis shows the respective rent level. Since marginal effects are calculated and averaged for every feature value, PD plots require high computational power. Thus, we plot the partial dependence for the year 2019, whose generation took eight hours of computing time.

[29] Beergardens are perceived as important hospitality institutions in Germany and thus the result is not surprising.

Figure 4.3: PD Plots - Living area, Age, Distances CBD & Department store in 2019



Note: The figure displays the partial dependence of the most important feature regarding two structural characteristics and distance to CBD and department store. The vertical axis denotes the feature values of log rent level while the horizontal axis represents the covariates feature values. Stacked black lines display the number of observations.

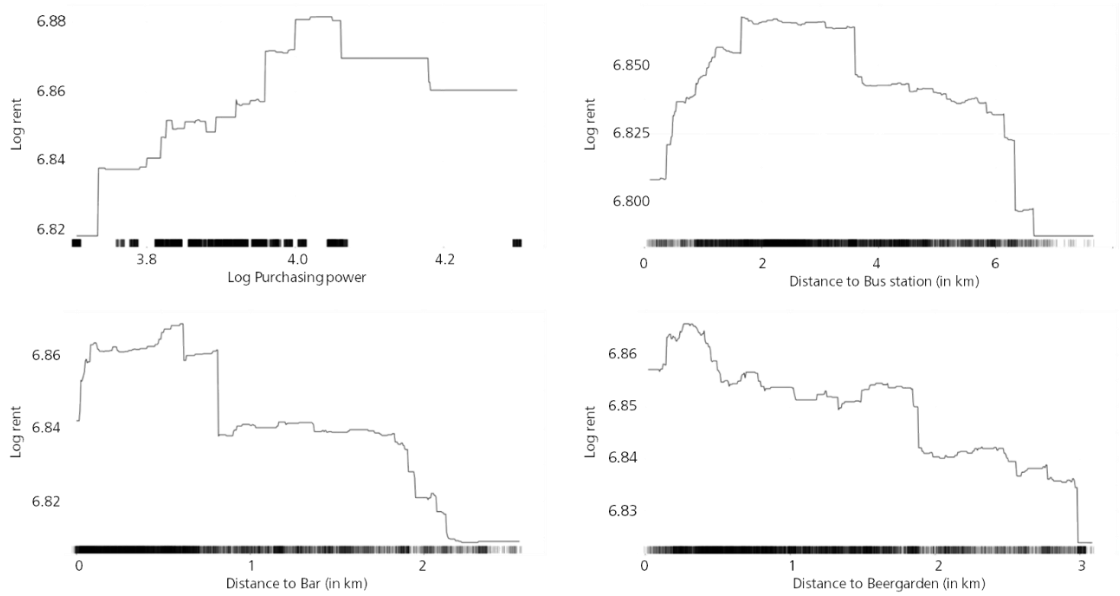
Figure 4.3 demonstrates how rental prices are associated with the four most influential characteristics living area, age and distance to CBD and department store. We start with the most important feature living area, which is incorporated as the natural logarithm. Since the PD plot highlights a linear relationship, the commonly applied log-log transformation can be confirmed as a good approximation of the positive relationship between living area and rent. Recent hedonic literature on property prices provides similar findings for the positive relationship (e.g. Dumm et al., 2016, Dumm et al., 2018 or Stamou et al., 2017). Age is perceived to be more complex, though intuitive. We find rental values to decrease with greater age until a building year of 1990-2000. While newly build apartments obtain highest rents, depreciation, changes in living preference as well as increasing requirements on energy-efficient construction most likely result in a steep decline in rental values. This is followed by an indifference of rental values up to 1940th. Frankfurt was heavily bombed in World War II, with emergence constructions of social housing provided by the government in the following decades. Therefore, historical pre-war buildings face higher rents. Consequently, building age displays a u-shaped relationship, as e.g. incorporated in Mayer et al. (2019).

Distance to CBD is perceived to be highly influential. In general, we find rental prices to decline with greater distance to the city center. Hedonic literature suggests similar conclusions since authors such as Osland (2010) or Zheng et al. (2016) also find a negative relationship between property prices and distance to the city center. However, the opposite effect is visible for close proximity. We expect tenants to appreciate separation from very urban areas. A graphical turning point can be found at about 1.5 km, followed

by moderate decline in rental prices. Interestingly, apartments close to the CBD face comparable rental values than the ones in 5 km distance. A steep decrease in rent levels can be seen beyond 5 and 7.5 km.

Regarding local supply, department stores are rather linearly and negatively associated with rental values. The proximity to shopping facilities results in increasing rents. We do not find an equivalent distance variable in the hedonic literature, however, Dubé and Legros (2016) show a positive price effect for properties not more than 1 km away from a shopping center. Interesting to note, the distance to department store drops sharply at about 1.5 and 2.5 km. This could indicate a critical distance for consumer goods. However, FI identifies supermarket as the least important distance variable. We assume that a high density of supermarkets in urban areas ensure local supply for everyday goods and therefore result in a negligible influence on rental values. In contrast, we assume different circumstances in rural communities. With minor influence due to the limited appearance of department stores, we expect the importance of supermarket to be more pronounced in non-metropolitan areas. Furthermore, FI ranks the presence of a built-in kitchen as important. Gröbel and Thomschke (2018) find a significant positive relationship between built-in kitchens and rents in Berlin (Germany). However, due to its binary nature, the visualization with PD plots is limited.

Figure 4.4: PD Plots - Purchasing power, Distances Bus station, Bar & Beergarden



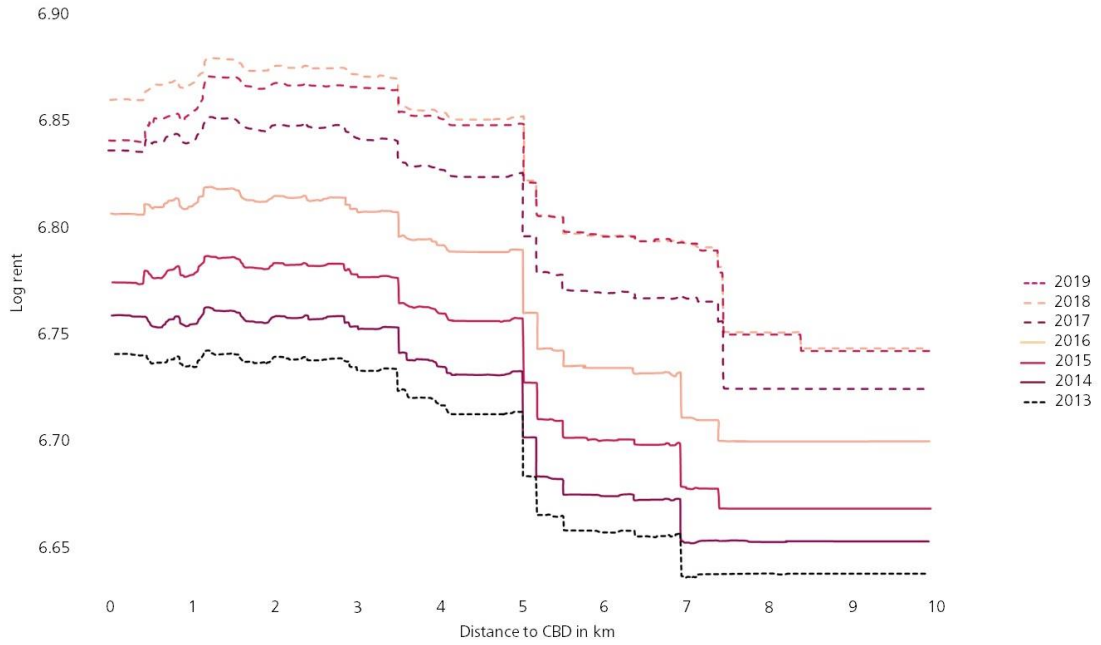
Notes: The figure displays the partial dependence of the most important feature regarding two structural characteristics and distance to CBD and department store. The vertical axis denotes the feature values of log rent level while the horizontal axis represents the covariates feature values. Stacked black lines display the number of observations.

The next most important characteristics displayed in Figure 4.4 are, according to FI, purchasing power and distance to bus station, bar and beergarden. We find socio-demographic information to show a rather linear relationship. Neighborhoods with high

purchasing power are associated with more expensive apartments and thus the variable is perceived as a characteristic of a good residential area. A steep increase in rental values for high wealth districts could reflect the segment of high-rise apartments in residential towers. While the construction of high-rise buildings is restricted in most German cities, Frankfurt has early incorporated tower buildings in urban planning. These do not only represent the highest price segment in the residential market of Frankfurt but have shown to be driver of residential prices and rents in the last years.

Interesting to note, the distance to bar, beergarden and bus station have shown to affect the overall prediction the most out of all hospitality and public transport features. All three variables show a non-linear relationship with residential rents. We find the distance to a bar to be positively associated with rental values up to approx. 700 meters. While a bar in close proximity would result in lower rents, the access to hospitality leads to an increase in rental values only from a certain distance. We expect tenants to face a trade-off between accessibility and negative externalities such as noise. The same relationship holds for the variable bus station. A location further away from a central bus hub is linked to higher rental values up to approx. 1.7 km. Since central hubs are related to mostly high urban density and traffic, we assume that tenants appreciate locational separation. The plot reveals the relationship to be quite constant until 3.5 km, followed by declining rental prices. The accessibility to central hubs through different means of transport seems to overlay negative effect of a larger distance. However, after 3.5 km, we find this effect to become visible and apartments that are poorly located in terms of transport face discounts for low accessibility. The presence of a parking spot complements the ten most influential variables, yet as a binary variable it is not displayed as a PD plot.

Figure 4.5: PD Plots - Rent and Distance CBD from 2013 to 2019

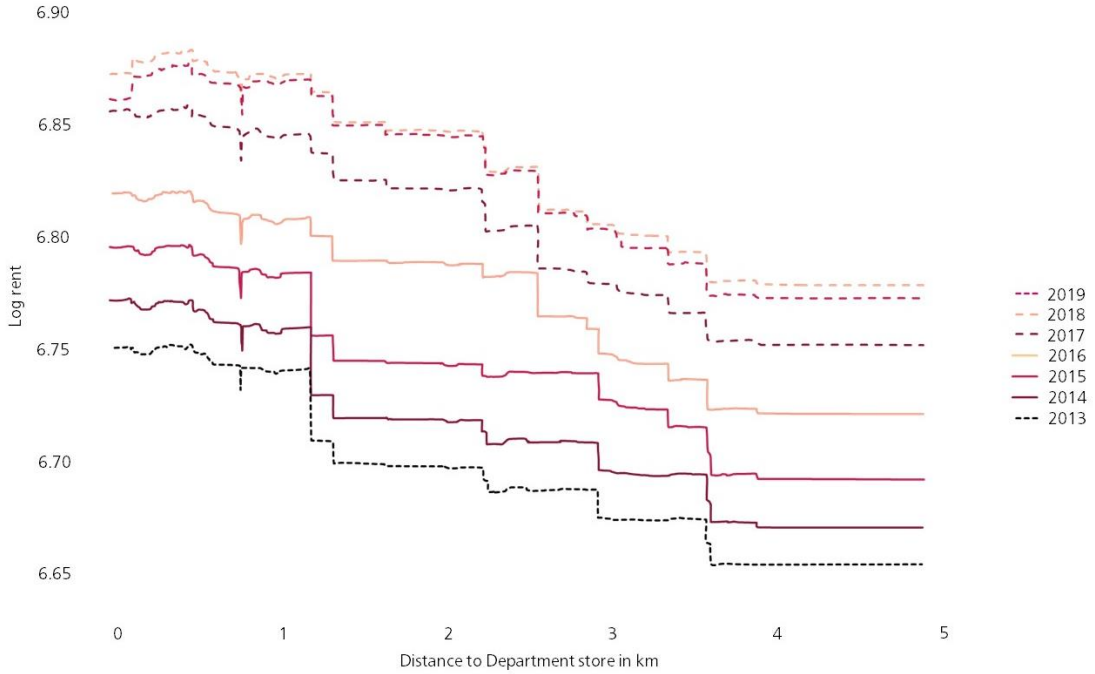


Notes: The figure displays the partial dependence of important variables over different periods. The vertical axis denotes the feature values of the log rent level while the horizontal axis denotes the covariates feature values.

Adding a temporal dimension to our analysis by displaying feature effects on a yearly basis enables us in a last step to illustrate temporal dynamics of the effects of hedonic characteristics. We demonstrate the latter by analyzing the distance to the CBD (Figure 4.5) and the distance to a department store (Figure 4.6).

At first, Figure 4.5 shows a negative relationship between rents and the distance to CBD across time. A continuous upwards shift for all feature values indicates increasing rent levels during the observed period. Only the graph of the year 2019 behaves differently, since it moves below 2018 for closer proximity and analogous from 5 km distance onwards. This development could be attributed to a declining preference for downtown locations in combination with overall stable rent levels in recent years. Although the course of all lines is quite similar, we find some differences. First, a drop in rental prices at a distance of 5 km is less pronounced for 2017, 2018 and 2019 than for previous years. This possibly indicates that residential locations further away from the center experienced rent increases due to a growing preference for sub-urban areas during the last years. Second, another major decline can be recognized at 7 km for 2013 to 2016. In the following years 2017 to 2019, however, this is only noticeable at a distance of approx. 7.5 km, but the downturn is considerably stronger. Both changes indicate that residential locations in medium distance to the center (5 to 7.5 km) experienced stronger rent increases compared to central as well as periphery location. We would assume that high demand in central locations results in a preference shift towards apartments further away from the CBD.

Figure 4.6: PD Plots - Rent and Distance Department store from 2013 to 2019



Notes: The figure displays the partial dependence of important variables over different periods. The vertical axis denotes the feature values of the log rent level while the horizontal axis denotes the covariates feature values.

In Figure 4.6, a negative relationship between rents and the distance to a department store is displayed, yet a similar pattern for the graphs can be seen in terms of comparable upwards shift of rents throughout all periods and 2019 being slightly below 2018. A first major decline is visible at approximately 1.2 km, with the years 2013, 2014 and 2015 experiencing a stronger decrease. From 2.6 km distance, the picture is the other way around. Whereas rents fell rapidly from 2016 to 2019, the downturn was not as strong as in previous years. The findings indicate that while locations between 1.2 km and 2.8 km gained popularity, locations in close proximity as well as further away remained more or less stable. Figure 4.7 in Appendix 4 provides an additional and centered PD plot for the feature distance to department store. Centered PD plots aid and underpin the interpretation of the differences in PDs throughout the years.

Ultimately, the feature effects technique yields greater transparency of how the different inputs contribute to the final estimation of the ML model. By visualizing the individual relations between the variables and the rent to be estimated, this method demonstrates which (economic) rational the algorithm has learned from the data and accordingly integrated into its internal calculations.

4.7 Conclusion

This paper sheds light on how Machine Learning (ML) based decision making in hedonic modelling can be made more transparent. We visualize and investigate the relationship between residential rents and a set of hedonic variables which was learned by a ML model. Based on a residential dataset of more than 52k apartments in Frankfurt am Main, Germany, we apply the eXtreme Gradient Boosting algorithm (XGB) for rental prediction. Model-agnostic Interpretable Machine Learning (IML) methods are subsequently used to examine feature importance and feature effects. Feature importance (FI) reveals that living area, age and the distance to CBD and a department store influence the overall rental prediction the most. In contrast, the least important features are several structural dummy variables and the distance to a supermarket and a bakery.

We plot the partial dependences (PD) for the influential variables that were detected in the preceding analysis to highlight feature effects. Although the relationship of rental values and the distance to CBD and department store is mainly linear, major declines at specific proximity values indicate that critical distances to the center as well as to local supply exist. Furthermore, there seems to be a difference in rent level to the wealthiest neighborhoods. Interestingly, we find that close proximity to hospitality and public transport is associated with rental discounts. In addition, the inspection of PD plots on a yearly basis reveals that especially apartments in a medium distance to the city center face considerable higher rent increases over the years. We assume both an increasing preference for less urban areas as well as peaking rent in the center to be possible reasons.

To conclude, interpretation methods can reveal the rationale behind the ML models estimation by demonstrating what relationship the algorithm detects in the underlying data. Peeking inside the black box enables researchers to reenact how a ML model arrived at its prediction and will help to gain new insights, ease practical applications and enhance reliability in algorithmic decisions.

The insights gained by these methods are relevant not only for research but also for practice in the private as well as public sector. Since real estate professionals commonly use ML to inform their decision making (RICS, 2017), model-agnostic methods provide a useful framework to effectively handle AI-based results. Whereas the advantages of these methods have already been discussed in detail, difficulties and limitations must also be pointed out. First of all, there are challenges in terms of computing power. Whereas parametric or semi-parametric methods are usually able to estimate hedonic models within seconds, ML-based methods such as XGB take considerably longer. This also applies to the application of IML. Furthermore, it should be noted that data availability is of course

essential for hedonic models. Even with ML-based models, an omitted variable bias can drastically reduce the informative value and thus the applicability. Admittedly, the data set of this study is quite extensive, but there are of course other additional apartment features imaginable that could influence the meaning of the results.

IML is a rapidly evolving field with new methods and applications being continuously proposed. Although this research area has achieved a degree of stability (Molnar et al., 2020), it is still in its infancy and faces several challenges to overcome. On the one hand, there is a need to define what interpretability means to then evaluate how black box models can be made more interpretable. On the other hand, the sensitivity of interpretation methods is of high importance, since not only these methods, but also the ML techniques are dynamically developing. To further improve 'trust' in algorithmic decisions ongoing research is necessary. We expect IML methods to be a valuable addition to the hedonic practice, both because it contributes to the transparency of ML models and because it provides insights on potentially unknown relationships in real estate hedonic modelling.

4.8 Appendix

Appendix 1

We apply different hedonic methods that have been used regularly in the literature. First, we deploy a hedonic OLS modelling approach to estimate the effects of property characteristics on rental prices. Linear hedonic regression represents the standard approach in modelling real estate prices and rents and is frequently used in housing studies (Mayer et al., 2019). The hedonic regression describes the rent Y as the sum of the predicted values of its characteristics X_j:

$$Y = \beta_0 + \sum_{j=1}^{J} X_j \beta_j + \varepsilon \tag{14}$$

In accordance to the real estate literature, a semi-log functional form with log-transformation of the dependent variable is conducted. Property characteristics include structural, socio-economic neighborhood and locational features. Proximity variables account for the spatial distance to public amenities and transport. Further spatial effects are modelled via spatial expansion by incorporating the coordinates in terms of longitude and latitude (Bitter et al., 2007, Chrostek & Kopczewska, 2013, Pace & Hayunga, 2020). Furthermore, temporal dummies are included for the specific month and year.

Many authors argue that property prices and rents may contain two key figures, namely spatial autocorrelation and spatial heterogeneity, that can require the spatial extension of hedonic models (LeSage, 1999). Since the occurrence of spatial effects can lead to misspecifications and biased results in the OLS framework (Anselin, 1988), we additionally apply a spatial autoregressive regression (SAR) with the following functional form:

$$Y = X\beta + \rho WY + \varepsilon \tag{15}$$

ρWY denotes a spatial lag of the target variable Y, with W being the spatial weight matrix that specifies the spatial structure, and ρ representing the spatial lag parameter.

However, linear models are subject to various restrictions due to their functional parametric form that can yield to misspecifications (Mason & Quigley, 1996; Pace, 1998). Because relationships in housing markets appear often to be non-linear, hedonic modelling can require the incorporation of more flexible functional forms to account for nonlinearity (Bontemps et al., 2008; Brunauer et al., 2013). Hence, a semi-parametric generalized additive model (GAM) is further considered.

$$Y = \beta_o + \sum_{j=1}^{J} X_j \beta_j + \sum_{p=1}^{P} f_p(X_p) + \varepsilon \tag{16}$$

GAM relaxes the linearity assumption by replacing the parametric linear relationship with non-parametric smoothers (e.g. splines, near neighbor and kernel smoothers). The linear equation is expanded by p smooth functions f_p in order to identify latent non-linear effects.

The results of the aforementioned methods are presented in Appendix 2. The coefficients provide expected signs and confirm a good model fit by showing acceptable R^2.

Appendix 2

Table 4.2: Results of the OLS, GAM and SAR Estimation

	Dependent variable: log Rent per month								
	OLS			GAM			SAR		
log Living area	0.939	***	(0.002)	0.900	***		0.928	***	(0,008)
Floors	0.002	***	(0.0004)	0.003	***	(0.0003)	0.003	***	(0,002)
Age (relative to 2017)	-0.0002	***	(0.00003)	s 8.000	***		-0.000	***	(0,0001)
Bathtub	-0.032	***	(0.002)	-0.016	***	(0.001)	-0.032	***	(0,006)
Refurbished	-0.015	***	(0.002)	0.005	***	(0.002)	-0.013	***	(0,007)
Built-in kitchen	0.084	***	(0.002)	0.077	***	(0.002)	0.077	***	(0,007)
Balcony	0.011	***	(0.002)	0.025	***	(0.002)	0.012	***	(0,007)
Parking	0.053	***	(0.002)	0.032	***	(0.002)	0.048	***	(0,008)
Elevator	0.053	***	(0.002)	0.020	***	(0.002)	0.048	***	(0,009)
Terrace	0.041	***	(0.002)	0.020	***	(0.002)	0.041	***	(0,009)
log Purchasing Power	0.406	***	(0.011)	0.069	***	(0.002)	0.313	***	(0,040)
CBD_distance	-0.019	***	(0.001)	s 8.692	***		-0.014	***	(0,002)
Bar_distance	-0.031	***	(0.002)	s 8.579	***		-0.024	***	(0,008)
Beergarden_distance	-0.020	***	(0.002)	s 8.631	***		-0.015	***	(0,005)
Cafe_distance	-0.014	***	(0.003)	s 8.700	***		-0.011	***	(0,010)
Bakery_distance	-0.011	***	(0.003)	s 8.842	***		-0.016	***	(0,009)
Convenience store_distance	-0.036	***	(0.002)	s 8.144	***		-0.035	***	(0,007)
Department store_distance	-0.006	***	(0.001)	s 8.580	***		-0.008	***	(0,005)
Supermarket_distance	-0.018	***	(0.006)	s 6.487	***		-0.029	***	(0,020)
Bus station_distance	-0.028	***	(0.001)	s 8.794	***		-0.017	***	(0,004)
Railway station_distance	-0.020	***	(0.002)	s 8.757	***		-0.020	***	(0,007)
Traffic signals_distance	0.086	***	(0.007)	s 8.243	***		0.075	***	(0,024)
Car wash_distance	0.012	***	(0.002)	s 8.763	***		0.007	***	(0.006)
Park_distance	-0.024	***	(0.006)	s 8.343	***		-0.019	***	(0.020)
School_distance	-0.003	***	(0.005)	s 8.412	***		0.008	***	(0,006)
Constant	-34.043	***	(3.087)	2.405	***	(0.100)	-22.860	***	(11,359)
rho							0.131	***	
time controls	Yes			Yes			Yes		
locational controls	Yes			Yes			Yes		
observations	52,966			52,966			52,966		
R^2	0.880						0.885		
adjusted R^2	0.880			0.898					
UBRE				0.028					

Notes: *p<0.1; **p<0.05; ***p<0.01, standard errors are displayed in parentheses. The GAM column reports the estimated degrees of freedom of the smooth terms (s) as well as their joint significance. Time controls (year and month) as well as location controls (apartment coordinates) are included in all models.

Appendix 3

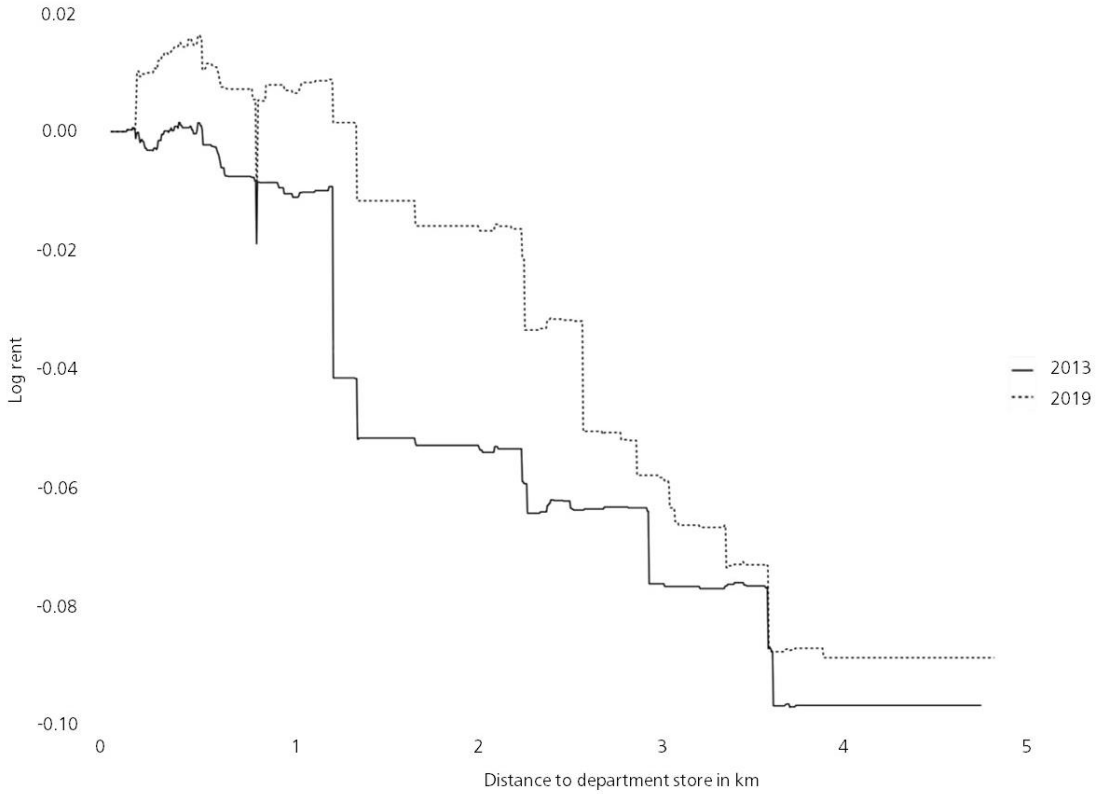
Table 4.3: Correlation matrix

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.
1.	Log rent p.m.	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.	Log living area	0,89	1	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0,90	0	0	0	0	0,21	0,92	0	0	0
3.	Floor	0,07	0,02	1	0	0,05	0	0	0	0	0	0	0	0	0	0,23	0	0	0	0	0	0	0	0	0	0	0
4.	Age	-0,14	-0,10	-0,09	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5.	Bathtub	0,12	0,18	-0,01	-0,06	1	0,74	0	0	0	0	0	0,22	0	0,02	0	0	0,77	0	0	0,01	0,51	0	0,01	0,62	0	0,05
6.	Refurbished	-0,09	-0,08	-0,04	0,23	0,00	1	1,00	0	0	0	0	0	0	0,91	0	0	0	0	0	0	0	0	0	0	0,46	0
7.	Bulit-in-kitchen	0,32	0,21	0,05	-0,13	0,02	0,00	1	0	0	0	0	0,01	0	0	0,01	0	0	0	0	0	0	0	0	0	0	0
8.	Balcony	0,19	0,19	0,10	-0,27	0,12	-0,06	0,06	1	0	0	0	0,85	0	0	0	0	0	0	0	0	0,09	0	0,02	0	0	0
9.	Parking	0,37	0,33	0,05	-0,57	0,08	-0,12	0,24	0,21	1	0	0	0	0	0,24	0	0	0	0	0	0	0,04	0	0	0	0	0
10.	Elevator	0,24	0,12	0,28	-0,56	0,03	-0,17	0,18	0,23	0,45	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11.	Terrace	0,21	0,21	-0,17	-0,20	0,05	-0,06	0,09	-0,09	0,21	0,11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0,07	0
12.	Purchasing power	0,10	0,11	-0,10	0,04	0,01	0,05	0,01	0,00	0,01	-0,06	0,05	1	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0
13.	CBD_distance	-0,22	-0,07	-0,11	-0,04	0,01	0,02	-0,12	-0,03	-0,02	-0,18	0,03	0,42	1	0	0	0	0	0	0	0	0	0	0	0	0	0
14.	Bar_distance	-0,20	-0,04	-0,13	-0,08	0,01	0,00	-0,16	0,02	-0,01	-0,19	0,04	0,19	0,42	1	0	0	0	0	0	0	0	0	0	0	0	0
15.	Biergarten_distance	-0,06	0,01	-0,01	-0,28	0,03	-0,11	-0,01	0,09	0,18	0,13	0,06	-0,06	0,07	0,05	1	0	0	0	0	0	0	0	0	0	0	0
16.	Cafe_distance	-0,15	-0,03	-0,10	-0,14	0,03	-0,03	-0,15	0,07	0,04	-0,10	0,06	0,18	0,35	0,53	0,20	1	0	0	0	0	0	0	0	0	0	0
17.	Bakery_distance	-0,09	0,00	-0,07	-0,12	0,00	-0,02	-0,08	0,04	0,06	-0,06	0,06	0,22	0,33	0,37	0,11	0,31	1	0	0	0	0	0	0	0	0	0
18.	Convenience store_distance	-0,11	0,03	-0,11	-0,22	0,02	-0,05	-0,07	0,06	0,13	-0,02	0,09	0,45	0,54	0,48	0,38	0,42	0,42	1	0	0	0	0	0	0	0	0
19.	Department store_distance	-0,18	-0,02	-0,09	-0,21	0,05	-0,05	-0,11	0,07	0,11	-0,07	0,06	0,17	0,43	0,47	0,46	0,43	0,22	0,58	1	0	0	0	0	0	0	0
20.	Supermarket_distance	-0,05	0,03	-0,10	-0,08	0,01	-0,03	-0,07	0,03	0,03	-0,11	0,06	0,27	0,28	0,37	0,16	0,39	0,31	0,34	0,22	1	0	0	0	0	0	0
21.	Bus station_distance	-0,25	-0,09	-0,15	-0,08	0,00	0,02	-0,15	0,01	-0,01	-0,18	0,03	0,29	0,62	0,57	0,07	0,40	0,48	0,59	0,48	0,29	1	0	0	0	0	0
22.	Railway station_distance	-0,14	0,01	-0,12	-0,21	0,03	-0,05	-0,09	0,07	0,11	-0,06	0,07	0,41	0,59	0,40	0,45	0,43	0,35	0,65	0,64	0,34	0,52	1	0	0	0	0
23.	Traffic signals_distance	-0,08	0,00	-0,09	-0,03	0,01	0,02	-0,08	0,01	-0,01	-0,14	0,04	0,20	0,38	0,45	0,10	0,37	0,26	0,42	0,30	0,37	0,33	0,35	1	0	0	0
24.	Car wash_distance	0,08	0,08	-0,01	-0,05	0,00	-0,02	0,03	0,03	0,04	0,06	0,03	0,16	-0,13	0,31	0,16	0,09	0,03	0,13	-0,08	-0,03	-0,05	0,02	0,09	1	0	0
25.	Park_distance	-0,11	-0,03	-0,04	-0,06	0,03	0,00	-0,08	0,03	0,01	-0,08	0,01	0,01	0,24	0,32	-0,02	0,23	0,35	0,25	0,23	0,14	0,30	0,18	0,24	-0,08	1	0
26.	School_distance	-0,06	-0,02	0,03	-0,11	0,01	-0,03	-0,04	0,05	0,04	0,02	0,02	-0,08	0,05	0,19	0,16	0,24	0,17	0,20	0,24	0,30	0,10	0,16	0,25	0,01	0,15	1

Notes: Pearson correlation coefficients are displayed below the diagonal and p-values above

Appendix 4

Figure 4.7: Centered PD Plots - Distance Department store



Notes: The figure displays the centered partial dependence of important variables over different periods. The vertical axis denotes the centered feature values of the log rent level while the horizontal axis denotes the covariates feature values.

4.9 References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 559–560.

Anselin, L. (1988). *Spatial econometrics: methods and models*. Springer Science & Business Media.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Below, S., Beracha, E., & Skiba, H. (2015). Land erosion and coastal home values. *Journal of Real Estate Research*, *37*(4), 499–536.

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, *12*, 513–544.

Bitter, C., Mulligan, G. F., & Dall'erba, S. (2007). Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems*, *9*(1), 7–27.

Bontemps, C., Simioni, M., & Surry, Y. (2008). Semiparametric hedonic price models: assessing the effects of agricultural nonpoint source pollution. *Journal of Applied Econometrics*, *23*(6), 825–842.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Brooks, T. J., Humphreys, B. R., & Nowak, A. (2020). Strip Clubs, "Secondary Effects" and Residential Property Prices. *Real Estate Economics*, *48*(3), 850–885.

Brunauer, W., Lang, S., & Umlauf, N. (2013). Modelling house prices using multilevel structured additive regression. *Statistical Modelling*, *13*(2), 95–123.

Cajias, M., & Freudenreich, P. (2018). Exploring the determinants of liquidity with big data – market heterogeneity in German markets. *Journal of Property Investment & Finance*, *36*(1), 3–18.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, *22*(3), 453–474.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 785–794.

Chernobai, E., Reibel, M., & Carney, M. (2011). Nonlinear Spatial and Temporal Effects of Highway Construction on House Prices. *The Journal of Real Estate Finance and Economics*, *42*(3), 348–370.

Chin, S., Kahn, M. E., & Moon, H. R. (2020). Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach. *Real Estate Economics*, *48*(3), 886–914.

Chrostek, K., & Kopczewska, K. (2013). Spatial prediction models for real estate market analysis. *Ekonomia*, *35*(0).

Conway, D., Li, C. Q., Wolch, J., Kahle, C., & Jerrett, M. (2010). A spatial autocorrelation approach for examining the effects of urban greenspace on residential property values. *The Journal of Real Estate Finance and Economics*, *41*(2), 150–169.

Court, A. T. (1939). Hedonic price indexes with automotive examples. In *The Dynamics of Automobile Demand* (pp. 99–117).

Des Rosiers, F., Dubé, J., & Thériault, M. (2011). Do peer effects shape property values? *Journal of Property Investment & Finance*, *29*(4/5), 510–528.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *Working Paper. ArXiv:1702.08608*.

Dubé, J., & Legros, D. (2016). A Spatiotemporal Solution for the Simultaneous Sale Price and Time-on-the-Market Problem. *Real Estate Economics*, *44*(4), 846–877.

Dubin, R. A. (1988). Estimation of regression coefficients in the presence of spatially autocorrelated error terms. *Review of Economics and Statistics*, 466–474.

Dumm, R. E., Sirmans, G. S., & Smersh, G. T. (2016). Price variation in waterfront properties over the economic cycle. *Journal of Real Estate Research*, *38*(1), 1–26.

Dumm, R. E., Sirmans, G. S., & Smersh, G. T. (2018). Sinkholes and Residential Property Prices: Presence, Proximity, and Density. *Journal of Real Estate Research*, *40*(1), 41–68.

Fernández-Avilés, G., Minguez, R., & Montero, J.-M. (2012). Geostatistical air pollution indexes in spatial hedonic models: the case of Madrid, Spain. *Journal of Real Estate Research*, *34*(2), 243–274.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Freybote, J., Fang, Y., & Gebhardt, M. (2017). The impact of temporary uses on property prices: the example of food trucks. *Journal of Property Research*, *34*(1), 19–35.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Goodwin, K. R., La Roche, C. R., & Waller, B. D. (2020). Restrictions versus amenities: the differential impact of home owners associations on property marketability. *Journal of Property Research*, *37*(3), 238–253.

Greenwell, B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *R Journal*, *9*(1), 421–436.

Gröbel, S. (2019). Analysis of spatial variance clustering in the hedonic modeling of housing prices. *Journal of Property Research*, *36*(1), 1–26.

Gröbel, S., & Thomschke, L. (2018). Hedonic pricing and the spatial structure of housing data–an application to Berlin. *Journal of Property Research*, *35*(3), 185–208.

Haider, M., & Miller, E. J. (2000). Effects of transportation infrastructure and location on residential real estate values: application of spatial autoregressive techniques. *Transportation Research Record*, *1722*(1), 1–8.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). Berlin: Springer.

Ho, W. K., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70.

Hoen, B., & Atkinson-Palombo, C. (2016). Wind Turbines, Amenities and Disamenitites: Astudy of Home Value Impacts in Densely Populated Massachusetts. *Journal of Real Estate Research*, *38*(4), 473–504.

Hoen, B., Brown, J. P., Jackson, T., Thayer, M. A., Wiser, R., & Cappers, P. (2015). Spatial hedonic analysis of the effects of US wind energy facilities on surrounding property values. *The Journal of Real Estate Finance and Economics*, *51*(1), 22–51.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (6th ed.). New York: Springer.

Jauregui, A., Allen, M. T., & Weeks, H. S. (2019). A spatial analysis of the impact of float distance on the values of canal-front houses. *Journal of Real Estate Research*, *41*(2), 285–318.

Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *The Journal of Portfolio Management*, *43*(6), 202–211.

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, *11*(1), 443–448.

Lam, K. C., Yu, C. Y., & Lam, C. K. (2009). Support vector machine and entropy based decision support system for property valuation. *Journal of Property Research*, *26*(3), 213–233.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, *74*(2), 132–157.

Lechner, M., Hasani, R., Amini, A., Henzinger, T. A., Rus, D., & Grosu, R. (2020). Neural circuit policies enabling auditable autonomy. *Nature Machine Intelligence*, *2*(10), 642–652.

LeSage, J. (1999). The theory and practice of spatial econometrics. *University of Toledo. Toledo, Ohio*, *28*(11).

Li, T. (2020). The Value of Access to Rail Transit in a Congested City: Evidence from Housing Prices in Beijing. *Real Estate Economics*, *48*(2), 556–598.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18.

Malpezzi, S. (2002). Hedonic Pricing Models: A Selective and Applied Review. In T. O'Sullivan & K. Gibb (Eds.), *Housing Economics and Public Policy* (pp. 67–89). Blackwell Science Ltd.

Mason, C., & Quigley, J. M. (1996). Non-parametric hedonic housing prices. *Housing Studies*, *11*(3), 373–385.

Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research*, *12*(1), 134–150.

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, *30*(4), 239–265.

Molnar, C. (2020). *Interpretable machine learning*.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning--A Brief History, State-of-the-Art and Challenges. *Working Paper ArXiv:2010.09337*.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Osland, L. (2010). An application of spatial econometrics in relation to hedonic house price modeling. *Journal of Real Estate Research*, *32*(3), 289–320.

Pace, R. K. (1998). Appraisal using generalized additive models. *Journal of Real Estate Research*, *15*(1), 77–99.

Pace, R. K., & Hayunga, D. (2020). Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests. *The Journal of Real Estate Finance and Economics*, *60*(1-2), 170–180.

Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*, *36*(1), 59–96.

R Core Team. (2020). *R: A language and environment for statistical computing*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-Agnostic Interpretability of Machine Learning. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

RICS. (2017). *The Future of Valuations*. Royal Institute of Chartered Surveyor.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *The Journal of Political Economy*, *82*(1), 34–55.

Rouwendal, J., Levkovich, O., & van Marwijk, R. (2017). Estimating the Value of Proximity to Water, When Ceteris Really Is Paribus. *Real Estate Economics*, *45*(4), 829–860.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., &

Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710.

Seo, W. (2018). Does neighborhood condition create a discount effect on house list prices? Evidence from physical disorder. *Journal of Real Estate Research*, *40*(1), 69–88.

Sheppard, S. (1999). Hedonic analysis of housing markets. In P. Cheshire & E. S. Mills (Eds.), *Applied Urban Economics: Vol. 3. Handbook of Regional and Urban Economics* (pp. 1595–1635).

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Sirmans, S., Macpherson, D., & Ziets, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*(1), 1–44.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Stamou, M., Mimis, A., & Rovolis, A. (2017). House price determinants in Athens: a spatial econometric approach. *Journal of Property Research*, *34*(4), 269–284.

Theisen, T., & Emblem, A. W. (2018). House prices and proximity to kindergarten – costs of distance and external effects? *Journal of Property Research*, *35*(4), 321–343.

van Wezel, M., Kagie, M. M., & Potharst, R. R. (2005). *Boosting the accuracy of hedonic pricing models.*

Wyman, D., & Mothorpe, C. (2018). The pricing of power lines: A geospatial approach to measuring residential property values. *Journal of Real Estate Research*, *40*(1), 121–154.

Yao, Y., Zhang, J., Hong, Y., Liang, H., & He, J. (2018). Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Transactions in GIS*, *22*(2), 561–581.

Yoo, S., Im, J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, *107*(3), 293–306.

Zhao, Q., & Hastie, T. (2021). Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, *39*(1), 272–281.

Zheng, S., Hu, W., & Wang, R. (2016). How much is a good school worth in Beijing? Identifying price premium with paired resale and rental data. *The Journal of Real Estate Finance and Economics*, *53*(2), 184–199.

Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research*, *33*(3), 349–387.

5 Conclusion

5.1 Executive Summary

This part of the thesis briefly summarizes the content of all three papers. It discusses the objectives of each study, data and methodologies used, as well as results and implications for science and practice.

Paper 1: Do Cross-Border Investors Benchmark Commercial Real Estate Markets? Evidence from Relative Yields and Risk Premia for a European Investment Horizon

Problems and Objective

Direct cross-border investments in commercial real estate have increased steadily over the past two decades. Accordingly, related research and market participants have demonstrated an increased interest in understanding the determinants of capital flows across national borders. Some authors describe the attractiveness of a target investment market as consequence of its socio-economic and institutional framework (Fuerst et al., 2015; Devaney et al., 2017a). Others find additional factors such as economic growth, demographics, urbanization or political stability of a particular country (see e.g. Lieser & Groh, 2011; Lieser & Groh, 2014; Salem & Baum, 2016; Devaney et al., 2017b). In consequence, literature identifies several absolute determinants, which attract cross-border capital.

The present article introduces a new approach to explaining inflowing cross-border capital into real estate markets, namely relative attractiveness for a target investment horizon. As opposed to previous studies, it sheds light on whether cross-border investors benchmark investment opportunities against each other. More precisely, the study investigates if the relative attractiveness in form of relative yields or relative risk premia determines the capital allocation of investors. Additionally, the analysis not only concentrates on the possible reveal of a linear relationship between the relative attractiveness and cross-border investment, but also applies non-linear estimation to gain further insights.

Methodology and Data

Aside from a classic linear estimation technique namely pooled OLS estimation, the study aims at contributing to the existing body of literature by analyzing non-linear patterns. Inspired by the real estate literature on hedonic pricing models (see e.g. Cajias & Ertl, 2018),

the present paper uses a generalized additive mixed model (GAMM). Given that classic prime European investment markets represent rather homogenous, geographically densely located competing investment hubs, European real estate markets provide a well-suited laboratory for the investigation at hand. The analyzed data sample contains 28 European cities across 18 countries, with quarterly observations of transaction volumes for office properties from Q1 2008 to Q3 2018. The data was obtained from various data providers. The dependent variable covers quarterly aggregated cross-border transaction volumes of office buildings provided by Real Capital Analytics. The target variables namely relative yields and relative risk premia are constructed based on the relative return measure of MSCI (2019). The remaining macroeconomic control variables stem from the OECD and the world economic forum. Real estate related controls are derived from CoStar.

Results and their Contribution to Science and Practice

The study expands the existing knowledge on cross-border investment in a twofold manner. First, it shows that not only absolute determinants influence investors when choosing among different investment destinations, but also relative attractiveness imposed by a benchmarking behavior of relative yields and risk premia can determine capital flows. Moreover, classic linear estimation serves as a useful tool to show the aforementioned relationship. To gain deeper insights, the application of non-linear estimation techniques helps to obtain more information on how investors react when looking for investment alternatives.

The results show that while relative yields do not have a significant impact, relative risk premiums do influence investors' investment decisions when choosing among foreign countries for capital allocation in direct real estate. Moreover, the results also indicate that it takes six months for foreign money to flow into a market, after its risk premium exceeded the European average for the first time. Finally, there is statistical evidence for a non-linear relationship between foreign capital and relative risk premia, which, on closer examination, shows that foreign investors prefer markets with particularly high relative risk premiums.

Aside from its theoretical importance, the results also have implications for the industry. Equity investors such as portfolio holders can anticipate capital flows into the market at an early stage and thus prepare exit strategies. Conversely, debt investors can expect loan repayments but also new financing business. Hence, understanding and knowing these mechanisms will lead to enhanced decision-making of market participants in the future.

Paper 2: Rental Pricing of Residential Market and Portfolio Data – A Hedonic Machine Learning Approach

Problems and Objective

Since decades, researchers as well as practitioners in the real estate industry apply hedonic modeling to derive property prices and rents (see e.g. Rosen, 1974 for one of the earliest implementation of hedonic models in real estate). The concept relies on the detection of individual characteristics, which either increase or decrease the value of a property. To do so, among other assumptions, the econometrician expects normally distributed residuals and therefore presupposes a parametric or semi-parametric relationship between property prices or rents as the dependent variable and the covariates also known as the hedonic characteristics (Bourassa et al., 2003; Sirmans et al., 2005).

Although traditional models can provide insights on real estate markets, they may lack in terms of prediction accuracy due to prediction errors induced by restrictions on various model-related assumptions. This is where new advances of statistical modeling and computational power can be beneficial. Artificial intelligence (AI) and machine learning (ML) combine progressive methods that especially excel in processing Big Data and when variable relationships are not known (James et al., 2013). Generally, the aim of ML is to autonomously understand and maximize the explanatory power of a model whereas traditional hedonic models rather focus on identifying interrelationships while relying on predefined assumptions of the econometrician (Mayer et al., 2019). In the light of predicting property prices and rents, this offers an entirely new way of analyzing and estimating real estate markets at asset level. Thus, the paper aims to answer questions concerning the application of ML methods in real estate markets. The objective is twofold: First, the study discusses and compares the performance of traditional hedonic models with ML approaches investigating rents as the dependent variable. Second, it confronts a market analysis with a real residential portfolio by means of ML methods to assess whether and how investors decide capital allocation based on their market understanding.

Methodology and Data

The paper applies traditional hedonic models via ordinary least squares and compares their performance with several ML techniques, namely Random Forest Regression (Breiman, 2001), Support Vector Regression (Smola & Schölkopf, 2004), Gradient Boosting (Friedman, 2001), and XG Boost (Chen & Guestrin, 2016). All methods investigate prediction performance at market and portfolio level. The study at hand analyzes the Munich market from January 2013 to June 2019, which represents one of the most active

residential markets in Germany. The dataset is comprised of different data sources. Asking rents of 65,743 apartments stem from Empirica, which collects data from 120 multiple listing systems in Germany. For the portfolio data, transaction rents of 716 apartments are gathered from a listed German asset manager. The remaining hedonic, socio-economic, microeconomic and geographic covariates are derived from other sources including GfK and Open Street Map.

Results and their Contribution to Science and Practice

The aim of this study is to shed light on whether ML methods are able to outperform traditional hedonic regression techniques considering predictive performance of residential property rents. It becomes apparent that in fact, tree-based Random Forest Regression and boosting methods such as Gradient Tree Boosting and XG Boost show considerably higher accuracy than OLS methods. Turning to the portfolio however, OLS results seem to be most accurate. One possible explanation is that investors rather rely on linear methods to derive at their rent level in tenant negotiations. Thus, using ML methods can uncover rental potential when engaging in rental scenarios. The findings show that the application can support transparency at market, portfolio and asset level since it enhances the understanding of the prevailing rental levels. Academics and practitioners are able to benefit by gaining deeper insights on how ML methods mount up to a better predictive performance. These tools excel in handling opaque real estate markets due to their capabilities of uncovering relations and pattern in the data.

Paper 3: Peeking inside the Black Box: Interpretable Machine Learning and Hedonic Rental Estimation

Problems and Objective

While AI and ML applications are becoming increasingly important in many areas of practice and research, critical voices are also gradually being raised. The ability of these applications to understand large data sets and make precise predictions is due to complex algorithms that run in the background. Unlike parametric methods such as OLS, which deliver interpretable coefficients, these methods lack transparency since estimation and decision-making processes of ML algorithms are hardly or not at all comprehensible (McCluskey et al., 2013). That is why researchers often speak of the 'black box' nature of these methods. To counteract this, there have been attempts to create a kind of algorithmic transparency, which is referred to as eXplainable Artificial Intelligence (XAI) or Interpretable Machine Learning (IML). Methods of this category are able to elucidate internal processes of AI and ML algorithms. Thus, they help to better justify results, to control and improve estimation procedures, and also to gain new insights (Adadi & Berrada, 2018).

In the field of real estate, ML methods are increasingly used for the prediction of prices and rents (as shown in Paper 2). This study investigates the applicability of IML in real estate. First, residential rents are estimated using a hedonic ML technique. Second, a number of so-called interpretable model-agnostic methods are applied to trace how the ML algorithm worked, and further, to detect economic relationships that became visible by applying such ex-post methods.

Methodology and Data

The methodological basis of this study forms the XG Boost algorithm (Chen & Guestrin, 2016), which estimates rents using a hedonic approach. The algorithm is known for showing very strong predictive performance and is therefore often used in the data science community. After that, two IML methods, feature importance and feature effects, are applied to study the algorithm's estimation. The former method is able to create an importance ranking, highlighting the most important hedonic characteristics of an apartment that lead to the rental prediction. The latter method is capable of an even deeper analysis. It quantifies the economic relationships between a hedonic variable and rent, allowing conclusions about the respective influence of the variable on the prediction. The data sample consists of 52,966 apartments in Frankfurt am Main (Germany) covering the period from 2013 to 2019. Asking rents as well as structural variables such as the size,

age and floor are obtained from Empirica. Locational variables such as the distance to CBD, public transport or hospitality amenities are calculated by means of geographical coordinates through Open Street Map. Purchasing power of the respective ZIP Code is provided by GfK.

Results and their Contribution to Science and Practice

The application of IML techniques reveals the following insights. Feature importance, the first IML Method investigated, highlights apartment size, age, distances to CBD and the next department store as most influential in determining a residential rent. In contrast, the least important characteristics comprise distances to supermarket and bakery as well as the existence of a bathtub or terrace. The analysis of feature effects, the second IML method, adds up to this. While a short distance to acquire convenience goods has a positive impact on rents, a close proximity of public transportation or hospitality induces rental discounts. It further reveals rental increases for apartments in medium distance to the CBD. Not only can economic insights be achieved, but the results are also able to gain a deeper understanding of which variables drive algorithmic decision-making and to what extent. Aside from the obvious advantage of greater transparency for the most valuable characteristics an apartment has to offer, this knowledge enables further trust and acceptance of ML methods in real estate. Hence, science and practice can benefit from this study in two ways. On the one hand, it should encourage the use of ML methods in the real estate research and industry by allowing to inspect basic relationships within the data that impacts algorithmic processes. On the other hand, it provides valuable insights at asset level to market participants, helping them to better understand the determinants of residential real estate.

5.2 Final Remarks

"[…] *As much information as possible available at any point in time"* will likely remain a utopia when assessing real estate markets. Thus, no fully transparent markets can be expected in the near future when following the definition of Schulte et al. (2005), p.91. The heterogeneous nature of real estate will always exacerbate the disclosure of all-comprehensive information. Furthermore, real estate is often a very private matter. Owners and occupiers cannot be forced to reveal all relevant details due to, i.e., legal protection. However, when analyzing the definition of Schulte et al. (2005) further, the question of *"*[…] *how the market mechanisms and the variables behind these mechanisms work* […]*"* can be addressed with today's technologies and help to improve transparency, as this thesis shows.

More precisely, this dissertation supports the progress towards more transparency by highlighting both application opportunities and economic insights to examine and evaluate market mechanisms and variables. All three studies underline data importance as the mandatory basis for the quality of assessing real estate investments. The more data available, the better the starting point. With the use of various types of data, such as transaction, asking, macro- and microeconomic or geographic data, this dissertation provides crucial groundwork for answering different questions at market and asset level. The application of linear, non-linear and complex statistical learning methods complements the research and transparency process. However, such methods should always only be considered as purposeful tools for certain tasks. Paper 1 uses linear and non-linear methods to investigate international market selection. Methods of AI are undoubtedly on the rise, yet econometric problems exist today and in the future, especially in the area of inference, that in turn can be perfectly solved using traditional statistical methods. Hence, it is not a matter of identifying the one dominant method, but rather of finding the one that can best solve the specific issue at hand. Paper 2 underlines this further by focusing on the predictive performance of rents. In the area of prediction, ML methods are known to perform exceptionally well. The paper demonstrates that linear methods are less precise in estimating residential rents. However, as Paper 3 shows, it is not only important to simply apply these methods, but also to understand how they work. Especially in the area of ML, it will become more and more important not only to trust blindly but also to question critically. This will ensure correct functioning and also be a way to clarify further economic issues by means of these methodologies.

Considering how future research can create further added value, in the area of Paper 1, it is conceivable to differentiate between origin and professional background of the investors

to investigate whether these are characteristics affecting benchmarking behavior. Extending the analysis to other asset classes may also yield interesting results. The same applies to Paper 2. While the proof of the suitability of ML methods on residential rents has been provided, the extent to which this also applies to the rents of other asset classes, such as office or logistics, still needs to be examined. Moreover, it would also be interesting to learn how these methods perform in falling markets. Consequently, one may be able to make a time- and asset-independent recommendation for this type of methodology. With respect to Paper 3, one could investigate whether the application of IML methods leads to similar results compared to other algorithms. Thus, algorithmic decision-making could be compared and underlying economic relationships further examined. In addition, IML also offers great application potential for other areas of research. In this context, the field of valuation represents a particularly interesting starting point. Since AI methods are already regularly used in this field, the potential for further research is substantial.

Other tasks of the future include advanced data collection, its processing and quality, as well as to improve existing statistical methods in terms of interpretability but also computational complexity. As this dissertation shows, a good interplay of these two factors is necessary for markets to become more transparent and thus for better investment decisions to be made. This dissertation is equally important for theory and practice. Partly to identify existing problems, but especially to offer answers and possibilities and thus to set the course towards transparent real estate markets.

5.3 References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Bourassa, S. C., Hoesli, M., & Peng, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics*, *12*(1), 12–28.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Cajias, M., & Ertl, S. (2018). Spatial effects and non-linearity in hedonic modeling. *Journal of Property Investment & Finance*, *36*(1), 32–49.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 785–794.

Devaney, S., McAllister, P., & Nanda, A. (2017a). Determinants of transaction activity in commercial real estate markets: evidence from European and Asia-Pacific countries. *Journal of Property Research*, *34*(4), 251–268.

Devaney, S., McAllister, P., & Nanda, A. (2017b). Which factors determine transaction activity across U.S. metropolitan office markets? *The Journal of Portfolio Management*, *43*(6), 90–104.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Fuerst, F., Milcheva, S., & Baum, A. (2015). Cross-border capital flows into real estate. *Real Estate Finance*, *31*(3), 103–122.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 103). Springer New York.

Lieser, K., & Groh, A. P. (2011). The attractiveness of 66 countries for institutional real estate investments. *Journal of Real Estate Portfolio Management*, *17*(3), 191–211.

Lieser, K., & Groh, A. P. (2014). The determinants of international commercial real estate investment. *The Journal of Real Estate Finance and Economics*, *48*(4), 611–659.

Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research*, *12*(1), 134–150.

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, *30*(4), 239–265.

MSCI. (2019). *MSCI Property Indexes Methodology: Index construction objectives, guiding principles and methodology for the MSCI Property Indexes.*

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *The Journal of Political Economy*, *82*(1), 34–55.

Salem, M., & Baum, A. (2016). Determinants of foreign direct real estate investment in selected MENA countries. *Journal of Property Investment & Finance*, *34*(2), 116–142.

Schulte, K.-W., Rottke, N., & Pitschke, C. (2005). Transparency in the German real estate market. *Journal of Property Investment & Finance*, *23*(1), 90–108.

Sirmans, S., Macpherson, D., & Ziets, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*(1), 1–44.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.