From Data to Insights: Unveiling Price Formation Processes in Direct Real Estate Markets with Machine Learning

Dissertation zur Erlangung des Grades eines
Doktors der Wirtschaftswissenschaft

eingereicht an der Fakultät für Wirtschaftswissenschaften
der Universität Regensburg

vorgelegt von:

JÜRGEN DEPPNER

Berichterstatter:	Prof. Dr. Wolfgang Schäfers
	Prof. Dr. Stephan Bone-Winkel
Tag der Disputation:	21. Juli 2023

From Data to Insights:

Unveiling Price Formation Processes in Direct Real Estate Markets with Machine Learning

Jürgen Deppner

Acknowledgments

The existence of this dissertation in its present form is a result of the support, inspiration, and encouragement provided by numerous individuals. I would therefore like to express my deepest appreciation to all those who have made this journey possible.

First and foremost, I am grateful to my first supervisor Prof. Dr. Wolfgang Schäfers, who has given me the opportunity to pursue a doctorate at the IRE|BS Institute and without whose persuasiveness I would have most probably not considered taking this step in the first place. His constant support, criticism, and motivation along the way have been instrumental in the success of this project. I would also like to express my sincere gratitude to Prof. Dr. Stephan Bone-Winkel, my second supervisor, for his supervision and support. He was always willing to listen and provide a sympathetic ear, while constantly reminding me not to lose sight of the practical implications of my research.

It goes without saying that the scientific articles forming the basis of this dissertation would not have been possible without my co-authors Benedict von Ahlefeldt-Dehn, PD Dr. Marcelo Cajias, Prof. Dr. Wolfgang Schäfers, and Prof. Eli Beracha (PhD). It has been a pleasure and an honor to work with them and I cannot overstate their invaluable contributions which certainly have enriched this dissertation.

I am also fortunate to have had the opportunity to work with an amazing group of colleagues and friends at the IRE|BS Institute, who have provided me with support and motivation, and have made this an exciting and unforgettable time. In particular, I would like to express my thankfulness to Daniel, Benedict, Nino, Moritz, and Bastian, as well as the many other colleagues and friends who have preceded and will follow.

Last but not least, I want to acknowledge my family and friends to whom I am deeply grateful for their unconditional support and unwavering belief in me. Above all, I would like to thank my parents Erika and Andreas for their encouragement and sacrifices that have been the driving force behind my academic pursuits. I owe them a debt of gratitude that can never be fully repaid. I also extend my thanks to my brother Markus, who has challenged and inspired me throughout my life. To my friends, especially Valentin, Matthias, and Tim, I am grateful for always having my back and providing a much-needed contrast to my professional life. Finally, I want to express my heartfelt appreciation to my partner Christina, whose relentless understanding and patience have sustained me through the challenges of this doctorate and whose emotional support has been invaluable to me.

I am humbled by the support and contributions of all these individuals, and I am deeply grateful for the role they have played in my professional and personal growth.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation and Background

Real estate is considered the largest asset class worldwide (Kok et al. 2017), constituting a substantial portion of both private and institutional net worth. This makes it not only a primary driver of economic activity but also one of the most significant stores of economic wealth. Implicit in this measurement of wealth is the market value of real estate, which can be defined as the most probable price to be expected in an arm's length transaction between informed and willing buyers and sellers in a competitive and open market (Schulz et al., 2014; Real Estate Lending and Appraisals, 2022).

Numerous stakeholders in the market rely on property valuations for various purposes such as accounting, monitoring, reporting, governance, and informed decision making. Banks and lenders require market values for loan underwriting, mortgage origination, equity withdrawal, refinancing, risk management, and accounting. Corporates and institutional investors depend on property appraisals for buy, hold, and sell decisions, financial reporting, performance measurement and monitoring, financing, loan covenant compliance, and portfolio transactions. Insurance companies further use market values to determine insurance premiums, assess risks, and evaluate claims and resale values in the event of a loss. For non-publicly traded real estate vehicles, market values are vital for reporting to investors, managing portfolio strategies, and executing transactions. In the private sector, homeowners or potential buyers have an interest to know the fair market value of a property when making the decision to relocate. In the public sector, government agencies utilize real estate market values to assess property taxes, determine land use policies, and inform planning decisions related to public infrastructure and services (Schulz et al., 2014; RICS, 2021). Although the purposes of property valuation are not limited to these examples, they illustrate the scale and significance of the real estate appraisal industry and stress the societal and economic importance of reliable market values.

In this context, accuracy, consistency, and timeliness of property appraisals have long been the primary concern of both industry and academia. This is a cost-intensive and time-consuming aim that requires considerable effort to achieve, as the value of real estate is determined by numerous factors, and the relationships between these factors are obscured and can vary significantly across markets and sectors. The high heterogeneity of properties and information asymmetries between market participants further complicates the identification of the relevant value drivers and their relationships. This issue is compounded by the fact that real estate is a bulky and relatively illiquid asset class with high transaction

costs, resulting in infrequent trades and a scarcity of comparable sales data. Consequently, market mechanisms and pricing processes mirror incomplete information and noisy signals which may further be distorted by the subjectivity of buyers and sellers and the individual circumstances of a transaction, as pointed out by Quan and Quigley (1991) and Dunse and Jones (1998). This does not only emphasize the complexity of property valuations but implies that there is arguably more than one opinion of market value resulting from varying judgements of price determinants.

As technological progress and increasing data availability have fostered digitization and automation during the past decades, computer-aided methods offer novel solutions to solve notorious problems. Particularly most recent advancements in the field of artificial intelligence (AI) and machine learning (ML) are catalyzing transformations across many sectors and innovate business processes that were once reserved for human intelligence. Prominent applications include prediction, object and image detection, voice recognition, and natural language processing (NLP). In healthcare, AI-powered systems are utilized for tasks such as diagnostics, drug discovery, and personalized medicine; in transportation, AI is employed for self-driving cars, traffic prediction, and logistics optimization; and in finance, it is utilized for fraud detection, risk management, and portfolio optimization. The latest generation of NLP models GPT-4 can generate human-like text in response to a given query and is likely to change many text-heavy business operations such as media and communication, marketing, e-commerce, human resource management, and customer service at a rapid pace. These are just a few examples that demonstrate how fast AI is evolving and becoming an increasingly vital tool for businesses and organizations to gain a competitive edge by creating added value, increasing efficiency, and reducing costs.

The increasing pressure of disruption spurred by AI applications is also becoming evident in the real estate appraisal industry, given the market's demand for more accurate, prompt, and cost-efficient property valuations. Automated valuation models (AVMs) that are based on hedonic models have been in use for a long time, providing ad-hoc property price valuations and allowing the analysis of price determinants (see Malpezzi, 2002; Mayer et al., 2019). However, the recent integration of AI and ML algorithms is currently marking an "inflection point" in their practicality (RICS, 2021), as these tools have achieved an unprecedented level of accuracy and precision. Yet, the widespread adoption of ML methods in the real estate valuation industry is impeded by certain constraints and limitations that need to be addressed before they can be implemented at scale.

First, these techniques have been criticized for ignoring "[…] the laws of economics […] as well as the limitations econometrics imposes on the models, leaving these systems free

to make inferences from a combination of data", as stated by Rico-Juan and Taltavull de La Paz (2021). This refers to the capacity of ML algorithms to eliminate noise from the data in an attempt to maximize predictive performance by detecting relationships between any given input and output, irrespective of whether these relationships are reasonable (Kok et al., 2017). Moreover, specific idiosyncrasies inherent in data structures can introduce biases into the models that may be replicated unnoticed (Lorenz et al., 2022). Incorporating economic and econometric theory is thus crucial to ensure the validity and reliability of data-driven predictions and inference.

Second, data insufficiency in direct real estate markets imposes limitations on the use of ML algorithms for automated valuation purposes, since these techniques rely on large amounts of training data to produce robust results. As transaction data is scarce, appraisal values and asking prices from online multiple listing systems (MLS) are the predominant data sources used in this domain. Such data may incorporate inherent biases as shown by Cannon and Cole (2011) and may not reflect market dynamics adequately (Downie and Robson, 2007). This compromises the interpretability and reliability of the produced results and entails the risk that AVMs may foster self-fulfilling prophecies and start leading the market rather than reflecting it (RICS, 2021). In addition, ML applications concentrate mostly on the housing sector, where data is less heterogeneous and easily accessible. In contrast, the nature of commercial real estate is much more unique and structured data is not yet readily available due to market intransparency. As of to date, little is known about the performance and reliability of ML-based AVMs using transaction data, neither in the housing sector, let alone in the commercial sector.

Third, machine learning algorithms are characterized as opaque "black boxes" due to their complexity and consequential inability to facilitate an inherent interpretability and explainability of the produced results (McCluskey et al., 2013; Mullainathan and Spiess, 2017; Adadi and Berrada, 2018). This impedes their acceptance and obstructs trust and confidence in the methods because the algorithms' decision making cannot be justified. Ensuring fairness, consistency, and integrity of market valuations over time is crucial for many applications such as lending, reporting, or taxation (Valier, 2020). To achieve this, both regulators and market participants require transparency and comprehensibility of the underlying methods.

Therefore, the primary objective of this dissertation is to contribute to the existing literature on econometric and data-driven real estate valuation models and demonstrate how price formation processes in direct real estate markets can be analyzed through the lens of machine learning algorithms. In this context, the thesis sets out to raise awareness of the

primary criticisms of ML methods for property valuation and shed light on their implications from an empirical perspective. In response, methodological frameworks are proposed to alleviate these concerns and translate the output of algorithmic approaches to property valuation into more meaningful and interpretable results. This discussion together with the proposed frameworks are intended to guide the development of data-driven methods in this domain and contribute to their practicality and marketability.

1.2 Course of Analysis and Research Questions

This section presents the course of analysis of the cumulative thesis and summarizes the objectives and research questions investigated in each article. The three papers are centered around the research field of property valuation and the identification of pricing mechanisms in direct real estate markets, discussing distinct aspects of the broader topic. More specifically, they set out to address the three major roadblocks to the practical use of ML methods in property valuation and pricing analysis outlined in section 1.1: *econometric limitations*, *data scarcity*, as well as *model interpretability and explainability*.

Paper 1: Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach

Paper 1 addresses *econometric limitations* and peculiarities of data structures that require consideration in the application of ML algorithms to direct real estate markets. More precisely, the objective of this article is to raise awareness of the implications of spatial dependence in housing markets for the workflow of non-parametric regression methods. The paper investigates the role of spatial autocorrelation on the model selection and model assessment of algorithmic hedonic models and proposes a technique named spatial cross-validation to mitigate spatial dependence structures in housing data. The central research questions can thus be stated as such:

- How does spatial autocorrelation generally impact algorithmic approaches for the estimation of property prices and rents?
- To what extent does spatial autocorrelation introduce bias in the model selection and model assessment of algorithmic hedonic models?
- How does spatial autocorrelation affect the predictive performance of algorithmic hedonic models in comparison to linear models?
- Is spatial cross-validation an adequate technique to account for spatial autocorrelation in house prices and rents?

Paper 2: Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach

Paper 2 addresses the issue of *data scarcity* and bridges the gap from the application of ML in the housing market to commercial real estate markets using transaction data. First, this article examines the accuracy and bias of market valuations in the U.S. commercial real estate market. In light of the discussed data constraints, the objective of this paper is

to explore the potential of ML techniques to provide a superior understanding of commercial real estate market dynamics compared to state-of-the-art valuation methods. In addition, the study sheds light on the determinants that are not adequately reflected in current appraisal practices. The research questions are as follows:

- How accurate and reliable are traditional commercial real estate appraisals?
- Do residuals of traditional appraisals exhibit structured variation that machine learning algorithms can exploit and further explain?
- If structural bias is existent in appraisals, which determinants of commercial real estate prices are not adequately reflected in current appraisal practices?
- Can machine learning algorithms close this gap and provide more dependable valuations in commercial real estate markets?

Paper 3: Increasing the Transparency of Pricing Dynamics in the U.S. Commercial Real Estate Market with Interpretable Machine Learning Algorithms

Lastly, Paper 3 addresses the concern over *model interpretability and explainability* in the context of data-driven valuation and pricing methods. The primary objective of this study is to demonstrate how ML can add to a deeper and more nuanced understanding of pricing mechanisms in institutional investment markets and how this understanding can guide the decision making of institutional investors. To achieve this, the paper proposes a comprehensive framework for the practical use of AVMs in commercial real estate that balances both precision and comprehensibility. More specifically, a model-agnostic interpretation technique named Shapley Additive Explanations (SHAP) is employed to investigate the value drivers of commercial real estate and their functional relationships with the market value. The central research questions can be summarized as follows:

- How can machine learning be effectively applied for commercial property valuation considering economic theory and the issues of data scarcity, market intransparency, and property heterogeneity in the sector?
- Can model-agnostic interpretation methods, in particular SHAP, alleviate the imbalance between the accuracy and interpretability of machine learning models?
- Do the inner workings of the applied machine learning algorithms follow an economic rationale?
- How can the proposed methodological framework add to the understanding and transparency of price formation processes in commercial real estate markets?

1.3 Co-Authors, Submissions and Conference Presentations

This section provides an overview of co-authors, journal submissions, publication status, conference presentations, as well as awards and funding for each of the three papers.

Paper 1: Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach

Authors:

Juergen Deppner, PD Dr. Marcelo Cajias

Submission Details:

Journal: The Journal of Real Estate Finance and Economics

Status: Accepted (06/17/2022) and published online ahead of print (07/13/2022)

Conference Presentations:

This paper was presented at:

- the 27th Annual Conference of the European Real Estate Society (ERES) online (2021)
- the 38th Annual Conference of the American Real Estate Society (ARES) in Bonita Springs, USA (2022)
- the 28th Annual Conference of the European Real Estate Society (ERES) in Milan, Italy (2022)
- the 4th Artificial Intelligence and Finance Workshop of the Center of Finance (CoF) at the University of Regensburg in Regensburg, Germany (2022)

Awards and Fundings:

This paper was awarded the "Manuscript Prize" in the category "Spatial Analytics/GIS Applications" and the "Doctoral Program Manuscript Prize" at the 38th Annual Conference of the American Real Estate Society

Paper 2: Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach

Authors:

Juergen Deppner, Benedict von Ahlefeldt-Dehn, Prof. Eli Beracha (PhD), Prof. Dr. Wolfgang Schaefers

Submission Details:

Journal: The Journal of Real Estate Finance and Economics

Status: Accepted (02/16/2023) and published online ahead of print (03/22/2023)

Conference Presentations:

This paper was presented at:

- the 38th Annual Conference of the American Real Estate Society (ARES) in Bonita Springs, USA (2022)
- the 28th Annual Conference of the European Real Estate Society (ERES) in Milan, Italy (2022)
- the Doctoral Seminar of the Center of Finance (CoF) at the University of Regensburg in Regensburg, Germany (2022)
- the 39th Annual Conference of the American Real Estate Society (ARES) in San Antonio, USA (2023)

Awards and Fundings:

This paper was awarded the "ALTUS Group Best Paper Award" at the 28th Annual Conference of the European Real Estate Society

Paper 3: Increasing the Transparency of Pricing Dynamics in the U.S. Commercial Real Estate Market with Interpretable Machine Learning Algorithms

Authors:

Benedict von Ahlefeldt-Dehn, Juergen Deppner, Prof. Eli Beracha (PhD), Prof. Dr. Wolfgang Schaefers

Submission Details:

Journal: The Journal of Portfolio Management

Status: Accepted (06/05/2023) and forthcoming in the 2023 Special Real Estate Issue

Conference Presentations:

This paper was presented at:

- the 2023 Doctoral Seminar of the Center of Finance (CoF) at the University of Regensburg in Regensburg, Germany (2023)
- the 2023 Real Estate Research Institute (RERI) Conference in Chicago, USA (2023)

Awards and Fundings:

This paper was funded with a research grant by the Real Estate Research Institute (RERI), a part of the Pension Real Estate Association (PREA)

1.4 References

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

Cannon, S. E., & Cole, R. A. (2011). How accurate are commercial real estate appraisals? Evidence from 25 years of NCREIF sales data. *The Journal of Portfolio Management, 35*(5), 68–88.

Downie, M. L., & Robson, G. (2007). Automated valuation models: An international perspective. *Council of Mortgage Lenders*, *CML Research*.

Dunse, N., & Jones, C. (1998). A hedonic price model of office rents. *Journal of Property Valuation and Investment*, *16*(3), 297–312.

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.

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2022). Interpretable machine learning for real estate market analysis. *Real Estate Economics*. Forthcoming.

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). Wiley.

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.

Quan, D. C., & Quigley, J. M. (1991). Price formation and the appraisal function in real estate markets. *The Journal of Real Estate Finance and Economics*, *4*, 127–146.

Real Estate Lending and Appraisals, (2022). 12 Code of Federal Regulations (C.F.R.) § 34.42.

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications*, *171*.

RICS (2021). Automated valuation models: Roadmap for RICS members and stakeholders. *The Royal Institution of Chartered Surveyors*.

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

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance*, *38*(3), 213–225.

2 Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach

2.1 Abstract

Data-driven machine learning algorithms have initiated a paradigm shift in hedonic house price and rent modeling through their ability to capture highly complex and non-monotonic relationships. Their superior accuracy compared to parametric model alternatives has been demonstrated repeatedly in the literature. However, the statistical independence of the data implicitly assumed by resampling-based error estimates is unlikely to hold in a real estate context as price-formation processes in property markets are inherently spatial, which leads to spatial dependence structures in the data. When performing conventional cross-validation techniques for model selection and model assessment, spatial dependence between training and test data may lead to undetected overfitting and overoptimistic perception of predictive power. This study sheds light on the bias in cross-validation errors of tree-based algorithms induced by spatial autocorrelation and proposes a bias-reduced spatial cross-validation strategy. The findings confirm that error estimates from non-spatial resampling methods are overly optimistic, whereas spatially conscious techniques are more dependable and can increase generalizability. As accurate and unbiased error estimates are crucial to automated valuation methods, our results prove helpful for applications including, but not limited to, mass appraisal, credit risk management, portfolio allocation and investment decision making.

Keywords: Hedonic modeling, Machine learning, Spatial autocorrelation, Spatial cross-validation, Mass appraisal, Automated valuation models

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

Real estate markets feature a spatial dimension that is pivotal to price and rent determination processes. The inherent spatial dependence in the economic value of assets cannot be ignored in hedonic models, as this would lead to spurious and biased results (Anselin, 1988; Can and Megbolugbe, 1997; Basu and Thibodeau, 1998). Guidance on how to account for spatial dependence in linear regression models is vast and remains the subject of many contributions to the hedonic and spatial econometric literature.

Moving from parametric hedonic regression techniques to the universe of non-parametric statistical learning methods, the literature has brought forth a growing body of evidence that machine learning algorithms can provide superior predictive performance for complex spatial regression problems, including various applications to house price estimation (e.g., Kok et al., 2017; Mullainathan and Spiess, 2017; Mayer et al., 2019; Hong et al., 2020; Pace and Hayunga, 2020; Bogin and Shui, 2020). To a great extent, the gains in explanatory power can be attributed to the flexibility of such models. This provides machine learning algorithms with the capability to exploit anisotropic and non-monotonic structures across space, which is of particular benefit when the spatial domain under investigation is a global one, as shown by Pace and Hayunga (2020). While this characteristic is a blessing when reproducing sample data, it can be a curse when predicting out-of-sample data since high flexibility is linked to overfitting, as demonstrated by Mullainathan and Spiess (2017) and Bogin and Shui (2020). Any kind of dependence structures in the data can exacerbate this problem, if not controlled for (Roberts et al., 2017). Thus, all the more surprising, little attention has been paid to the implications of spatial dependence in house prices and rents for the statistical validity of cross-validation (CV) errors, which are widely used to select and assess non-parametric models. For CV errors to be valid estimates of predictive performance, observations must be statistically independent of each other (Bishop, 1995; Brenning, 2005; Varma and Simon, 2006). This assumption is unlikely to hold in a real estate context (Bourassa et al., 2010) because "[…] error variance is not equal to zero but may be a function of spatial proximity among houses", as explained by Can and Megbolugbe (1997).

Two main problems arise when applying random resampling techniques to spatially dependent data. First, spatially structured variation in the residuals may be absorbed by non-causal regressors, consequently leading to the selection of overly complex and overfitted models that do not perform well with unseen data. Second, spatial autocorrelation between training and test observations provides the predictor with information that is assumed to be unavailable during model training, thus inflating

estimates of predictive accuracy. In turn, this may hide the first problem as CV errors appear to be legitimate (Brenning, 2012; Roberts et al., 2017). Using such models to predict unseen data can result in substantially lower accuracy than is approximated by CV. When furthermore applied in combination with model-agnostic interpretation techniques to draw inference on the relation between housing value and property features, spurious regression can result in the identification of meaningless relationships.

In response, researchers from geoscientific modeling fields have developed spatially conscious resampling methods to address these problems. However, the adequacy of such techniques for hedonic house price models cannot be blindly assumed since prediction goals may differ. To the best of our knowledge, no research has thus far accounted for spatial dependence in algorithmic hedonic models by applying spatial resampling techniques. We believe that a sound understanding of the implications arising from spatial dependence is of great importance when applying machine learning algorithms to hedonic regression problems. Hence, this study aims to investigate the role of spatial autocorrelation on resampling-based model selection and model assessment of algorithmic hedonic methods, thereby evaluating the efficacy of spatial CV in contrast to non-spatial (i.e., random) CV. By doing so, we demonstrate the pitfalls of resampling-based performance evaluation and intend to raise awareness of the importance of spatially conscious resampling techniques in hedonic house price modeling.

Based on a cross-section of apartment rents in Frankfurt, Germany, we train and evaluate tree-based algorithms using spatial as well as non-spatial CV. We subsequently forecast out-of-sample data to assess the bias in error estimates associated with spatial autocorrelation. The results are put into a broader perspective by benchmarking our machine learning algorithms against a non-spatial ordinary least squares (OLS) and a spatial autoregressive framework, allowing for a relative comparison of bias and predictive performance. Lastly, we analyze the residual spatial autocorrelation to detect signs of overfitting to spatial structures in the data.

To make informed decisions, the precise estimation of house prices and rents is imperative to parties in the real estate industry, such as investors, developers, lenders or regulators. Since CV is commonly used as an "out-of-sample experiment" (Mullainathan and Spiess, 2017) to assess the predictive accuracy of algorithmic hedonic models, a systematic bias in error estimates may have adverse effects on the allocation of both debt and equity (Kok et al., 2017). The results of this study prove helpful in increasing the reliability and generalizability of CV errors, thus containing valuable implications for mass appraisal

practices, credit risk management, portfolio allocation as well as investment decision making.

This paper is structured as follows: the section on "Hedonic Modeling of Spatially Structured House Prices and Rents" elaborates the problems of spatially structured data and their implications for hedonic analyses in the most commonly applied parametric as well as non-parametric regression frameworks, thereby providing an overview of the empirical literature on algorithmic hedonic approaches with a focus on applied resampling strategies. In the "Data and Methodology" section, the dataset is presented, followed by a description of the study design and the methodological approach. The empirical results are presented and discussed in the "Results" section and the final "Conclusion" section summarizes the findings of this study.

2.3 Hedonic Modeling of Spatially Structured House Prices and Rents

In his 1970 study on urban growth in the Detroit region, W. R. Tobler invoked his well-cited first law of geography, stating that the outcomes of nearby events correlate stronger than those of more distant events. Transferred to a housing context, this implies that the economic value of housing at any given location in geographic space depends, amongst other aspects, on the value of housing in neighboring locations. This deduction is well underpinned by spatial econometric as well as land economic theory for several reasons, such as spatial spillover effects (i.e., adjacency effects) and neighborhood effects (Can, 1992). Moreover, spatial clustering of house prices and rents may originate from a high correlation in the utility of the underlying houses derived from their structural characteristics (Basu and Thibodeau, 1998) and their fixed location in geographic space (Can and Megbolugbe, 1997; Osland, 2010), both of which determine the economic value of housing.

This leads to the conclusion that space is a fundamental factor that drives price formation processes in housing markets, subsequently resulting in two critical characteristics of housing market data: First, spatial autocorrelation, which is spatial dependence in price and rent determination processes; second, spatial heterogeneity, defined as the systematic variation in the behavior of price and rent formation processes across space (Anselin, 1988; Can and Megbolugbe, 1997). As stated by Osland (2010), one can assume that “[…] a mixture of these effects will be present in all housing market cross-section data”. This poses important methodological implications on both parametric and non-parametric hedonic regression frameworks, which will be discussed below.

2.3.1 Parametric Hedonic Models

The economic theory of hedonic pricing in a housing context dates to Rosen (1974), who implemented the derivation of implicit prices of hedonic characteristics using a least squares estimator. Due to their efficiency and ease of interpretability, least squares estimators have established themselves as the standard econometric approach to hedonic house price modeling. Likewise, the concept of hedonic price modeling has been successfully transferred and applied to the determination of apartment rents (e.g., Sirmans et al., 1989; Sirmans and Benjamin, 1991; Allen et al., 1995).

Implications of Spatial Dependence

Independent and identically distributed errors with a zero mean and constant variance are crucial Gauss-Markov assumptions to produce consistent and efficient estimates in a least squares context (Wooldridge, 2016). Spatial autocorrelation and spatial heterogeneity in the residuals violate these assumptions, resulting in unreliable confidence intervals and biased t-statistics, which lead to spurious statistical inference (Anselin, 1988; Basu and Thibodeau, 1998). Depending on the underlying spatial processes causing spatial effects, even point estimates might be biased and lead to erroneous results (Pace and LeSage, 2010).

Moreover, endogeneity is likely to occur due to omitted variable bias, measurement errors in the independent variables or feedback loops induced by adjacency effects. The explanatory power of spatial effects not explicitly reflected in the model specification is picked up by the error term or by covarying explanatory variables instead, leading to biased estimates and non-normality of the errors (LeSage and Pace, 2009). Even if spatial controls are included in the regression equation, the assumption of linearity in the functional form requires their relationship with the dependent variable to be constant across space. However, in the real world, such relationships are seldom linear and monotonic nor isotropic since slopes are likely to vary by distance and direction (Osland, 2010). Non-stationarity across space will persist as spatial heterogeneity in the residuals, violating the crucial OLS assumption of homogeneity in the errors.

It can be concluded that both theory, as well as empirical research, suggests that the Gauss-Markov assumptions underlying traditional OLS estimators cannot be naturally presumed in a real estate context (Can and Megbolugbe, 1997; Bourassa et al., 2010; Cajias and Ertl, 2018), resulting in biased and inconsistent least squares estimates as well as spurious inference.

Accounting for Spatial Dependence

Spatial autoregressive models are the typical statistical instruments to consider spatial effects in parametric frameworks. They control for spatial dependence by explicitly incorporating the underlying correlation structures as spatial lags in their functional form (see Cliff and Ord, 1973; Anselin, 1988; Cressie, 1993; Manski, 1993; Kelejian and Prucha, 1998; LeSage and Pace, 2009). As the necessity to account for spatial effects in linear models is well understood, spatial autoregressive, as well as other spatial modeling alternatives, are widely applied and discussed in a real estate context (e.g., Pace and Gilley, 1997; Case et al., 2004; Militino et al., 2004; Valente et al., 2005; Bourassa et al., 2007, 2010; Osland, 2010; Füss and Koller, 2016; Cajias and Ertl, 2018). Although such methods have been demonstrated to reduce residual spatial autocorrelation if applied carefully, the models continue to be linear, limiting their ability to capture highly complex and multi-dimensional relationships in the formation of house prices and apartment rents.

2.3.2 Non-Parametric Hedonic Models and Cross-Validation

As the real world can be more accurately described by logarithmic, exponential or step functions, the increasing availability of data together with technical progress in computational power has triggered the consideration of more flexible non-parametric machine learning methods for the problem of hedonic house price and rent modeling. In principle, such data-driven approaches do not rely on any a priori assumptions about the distributions of the errors, nor the functional form $f(x)$ that explains i house prices y_i using j regressors x_{ij}, but approximate the shape of $f(x)$ by fitting a spline to the data (James et al., 2013). However, it is to mention that the lack of a pre-defined additive functional form comes at the cost of inferential insights, as the prediction rules of the algorithms are opaque and cannot be directly interpreted due to their complexity. Moreover, their high flexibility makes them prone to overfitting, which is why modern statistical tools rely on resampling methods for model selection (i.e., selecting an appropriate level of regularization to approximate the shape of $f(x)$ during hyperparameter tuning) and for model assessment (i.e., assessing the test error rate of the selected model $\hat{f}(x)$ to evaluate its performance).

Resampling is typically performed using cross-validation, during which observations are randomly partitioned into mutually exclusive training and test subsets, whereby the predictor is fitted on the training data and evaluated on the respective test data (Stone, 1974; Snee, 1977). This concept can be thought of as creating "[…] an out-of-sample experiment inside the original sample", as described by Mullainathan and Spiess (2017).

In its most simple form, cross-validation randomly divides the data into two subsets, that is a training set and a validation (i.e., holdout) set based on a given percentage split. Subsequently, the model is fitted on the training sample, which is then used to predict the responses from the validation sample. This holdout strategy has been widely applied in the algorithmic hedonic house price literature. Worzala et al. (1995), Din et al. (2001), Peterson and Flanagan (2009) as well as Chiarazzo et al. (2014) use this technique for model assessment of artificial neural networks (ANNs), and Yoo et al. (2012), Kok et al. (2017) as well as Pérez-Rave et al. (2019) to validate different tree-based algorithms such as regression trees (RT), random forest regression (RFR), gradient tree boosting (GTB) and extreme gradient boosting (XGB). Lam et al. (2009), Antipov and Pokryshevskaya (2012), McCluskey et al. (2013) and Bogin and Shui (2020) benchmark different machine learning approaches, including support vector regression (SVR), shrinkage estimators (e.g., LASSO) as well as neural networks and tree-based methods using error estimates from a holdout sample. The applied split ratios vary between 60 to 80% for the training data and 40 to 20% for the test data, respectively. Such holdout strategies are computationally inexpensive and easy to implement. However, the test error rate may be heavily dependent on which observations are held out for validation and used for training, resulting in a potential bias in the error estimates (James et al., 2013).

To address this form of bias, k-fold cross-validation has been introduced to the statistical community (Lachenbruch and Mickey, 1968; Efron, 1983). During k-fold cross-validation, the data is partitioned into k mutually exclusive subsets of equal size. Subsequently, each of the k folds is once used as a test set and the remaining $k - 1$ folds are used to calibrate the model, consequently yielding k estimates of prediction error that are then averaged. This strategy attempts to generate more robust and reliable approximations of out-of-sample predictive performance. In a real estate hedonic context, k-fold cross-validation has gained in popularity during the past decade. Park and Bae (2015), Gu and Xu (2017), Čeh et al. (2018), Chin et al. (2020) as well as Pace and Hayunga (2020) apply k-fold cross-validation to evaluate the performance of tree-based methods. Applications to a broader spectrum of machine learning algorithms, including ANNs, SVR, k-nearest neighbors, shrinkage estimators as well as ensembles of regression trees using boosting and bagging techniques, can be found in Zurada et al. (2011), Mullainathan and Spiess (2017), Baldominos et al. (2018), Mayer et al. (2019), Hu et al. (2019), Ho et al. (2021), Cajias et al. (2021), as well as Rico-Juan and Taltavull de La Paz (2021). In all those studies, the applied number of folds is either five or ten, except for Mullainathan and Spiess (2017), who set k equal to eight.

Machine learning algorithms excel parametric models in the identification of complex non-linear relationships between the value of real estate and property characteristics, but they are also criticized for their black box character as their inner workings are opaque and comprehensibility as well as direct interpretation of the models are impeded by their complexity. Although recent developments allow insights into these opaque black boxes via model-agnostic interpretation techniques (see Rico-Juan and Taltavull de La Paz, 2021; Lorenz et al., 2022), this constitutes a limitation for the use of machine learning in both, academic research and practice. As stated by Rico-Juan and Taltavull de La Paz (2021), data-driven models might not be consistent with theoretical expectations and may thus have no economic meaning when identified relationships are spurious because "[…] the laws of economics (and the explanatory models that show causality) as well as the limitations econometrics imposes on the models [are ignored], leaving these systems free to make inferences from a combination of data". Spatial autocorrelation is one such econometric constraint that is typically ignored in machine learning applications to house price data.

Implications of Spatial Dependence

Although cross-validation proves to be decisive in reducing bias in error estimates, any kind of resampling technique is subject to one central assumption. By conducting an out-of-sample experiment that draws random observations from the data that are then used to approximate prediction errors, CV attempts to simulate unseen data. For cross-validation to yield unbiased prediction error estimates, statistical independence between training and test observations is required (Bishop, 1995; Brenning, 2005; Varma and Simon, 2006). Consequently, the meaningfulness of the resulting CV errors as a robustness test for out-of-sample predictive performance is highly reduced in spatial modeling fields where the independence assumption is violated (Le Rest et al., 2014). More specifically, spatial dependence structures in the data cause two main problems in the workflow of machine learning algorithms.

First, regressors are often covarying with unexplained spatial dependence structures in the residuals. During hyperparameter tuning (i.e., model selection), the model may overfit these spatial structures to non-causal, but covarying regressors in the attempt to optimize model performance, thereby reducing or completely absorbing unexplained structured covariation from the residuals. This may lead to a selection of overly complex models that can reproduce the training data but may not generalize well to unseen data (Can and Megbolugbe, 1997; Le Rest et al., 2014; Roberts et al., 2017; Meyer et al., 2019). Second, when the sample from the validation fold is drawn from the same dependence structure

as the training folds due to spatial proximity, the predictor may obtain information from the spatially autocorrelated test data that is assumed to be unavailable to the model during training. This unauthorized glimpse on the test data results in approximations of predictive power (i.e., model assessment) that may be overly optimistic and thus not representative for unseen data with different spatial structures (Picard and Cook, 1984; Hastie et al., 2009; Le Rest et al., 2014; Trachsel and Telford, 2016; Roberts et al., 2017; Schratz et al., 2019; Lovelace et al., 2019).

In a real estate context, such cases of poor out-of-sample predictive performance were, for instance, reported by Mayer et al. (2019) for their random forest. Bogin and Shui (2020) report significant overfitting using a random forest with a deviation of 22.1 percentage points in the R^2 compared to in-sample cross-validation errors. Mullainathan and Spiess (2017) demonstrated a similar bias in cross-validation errors for both bagging and boosting with 39.6 percentage points discrepancy in the R^2 of the random forest and 8.7 percentage points in the boosting trees.

This is problematic because, on the one hand, accuracy implied by cross-validation may lead to unjustified confidence in a model's predictive power that cannot be guaranteed when making predictions with unseen data. On the other hand, identified relationships may be spurious and can result in fallacious inferential conclusions when model-agnostic interpretation techniques are applied to interpret the pricing processes of the algorithms.

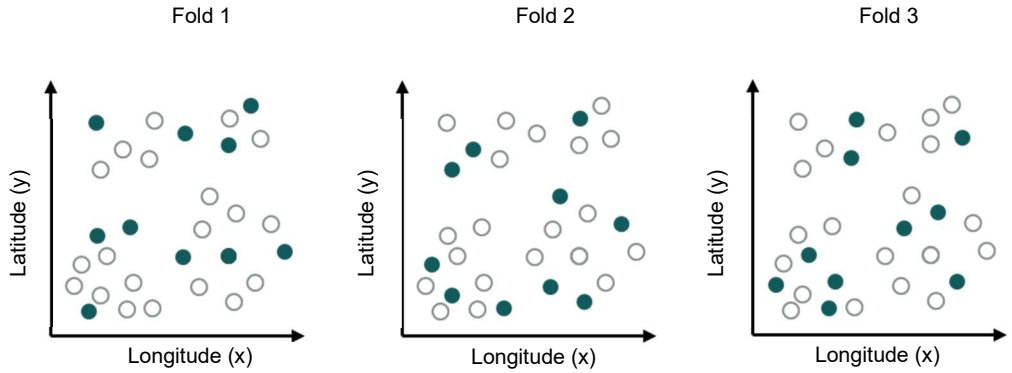
Accounting for Spatial Dependence

One possible approach to account for spatial dependence structures in the selection and assessment of non-parametric models is by using resampling techniques that split the data strategically by considering spatial proximity among observations rather than randomly. Spatial partitioning can be designed in many ways. However, the general concept is to increase independence between training and test data by clustering or blocking the individual folds across space or by removing training data within a specific distance band of each test point, such that performance is evaluated on more distant events that tend to be less correlated to the training sample (Tobler, 1970; Trachsel and Telford, 2016; Roberts et al., 2017). In a spatial context, such approaches have been introduced to the statistical community under many different terms. These include "spatial cross-validation" (Brenning, 2005, 2012), "spatial leave-one-out cross-validation" (Le Rest et al., 2014), "h-block cross-validation" (Trachsel and Telford, 2016), "spatial k-fold cross-validation" (Pohjankukka et al., 2017), "spatial buffering", "spatial blocking", "environmental blocking" (Valavi et al., 2018) and "leave-one-cluster-out cross-validation" (Meyer et al., 2019). Following the methodology and terminology of Brenning (2012), we will continue naming this concept

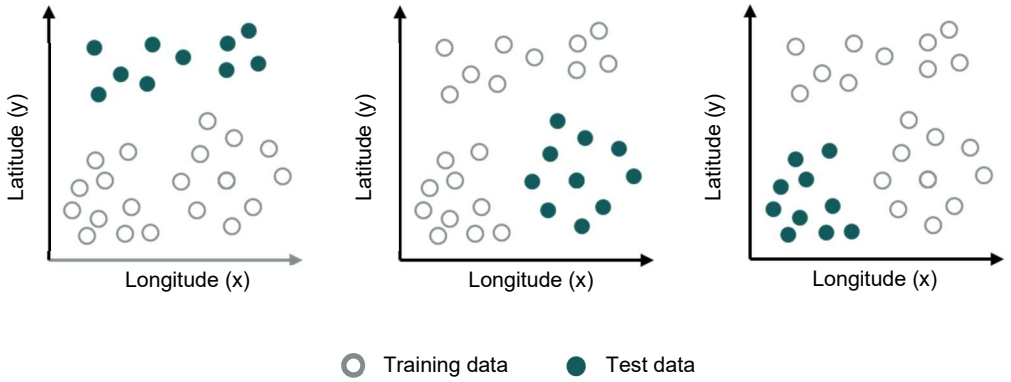
spatial cross-validation in the remainder of this study. The conceptual difference between random and spatial partitioning of folds during cross-validation is visualized in Figure 2.1 based on an example with three folds.

Figure 2.1: Partitioning of Folds

Panel A: Random Partitioning of Folds



Panel B: Spatial Partitioning of Folds



Notes: This figure depicts the conceptual difference between random partitioning and spatial partitioning of folds using *k*-means clustering during cross-validation.

Since "[…] the adequacy of non-spatial partitioning techniques for spatial datasets can be questioned" as stated by Schratz et al. (2019), spatial cross-validation methods are widely used in scientific fields such as climatology (Trachsel and Telford, 2016), ecology (Bahn and McGill, 2007; Schratz et al., 2019), remote sensing (Brenning, 2012; Meyer et al., 2019) and geosciences (Brenning, 2005). The need for spatial resampling has been stressed repeatedly in those fields, yet, its suitability and efficacy for real estate data has not been investigated thus far. Findings from other disciplines cannot be easily transferred to a real estate context since the objectives and circumstances under which predictive models are designed may differ. Despite compelling arguments to use spatial cross-validation when

modeling data in geographic space, there is good reason to be cautious when applying such methods to a real estate hedonic context.

Spatial partitioning may hide entire ranges or functional relationships of regressors during training, thereby introducing extrapolation to a model that is supposed to interpolate and consequently resulting in overly pessimistic estimates of prediction errors during model assessment (Snee, 1977; Roberts et al. 2017). However, the model can also be underfitted when the selected level of regularization is too high which may result in poor predictions (Kok et al., 2017). This dichotomy is particularly pronounced in real estate related regression tasks where high levels of spatial dependence tend to exist between observations but the prediction goal is usually dominated by the interpolation of existing properties within a delineated market. In cases where both the degree of spatial autocorrelation in the data and the extrapolation range are low, conventional cross-validation techniques that split the data randomly may be appropriate for performance optimization and evaluation. However, in situations where a model predicts outside the spatial domain of the training data and correlation structures between the residuals and non-causal regressors differ from the structures that were overfitted to non-causal regressors, random partitioning may yield unsatisfactory results (Bahn and McGill, 2007; Roberts et al., 2017).

As shown by Gröbel and Thomschke (2018) and Hong et al. (2020), prediction accuracy also depends on the spatial density of the sample locations. This is in line with Bahn and McGill (2007), who state that "[…] the sparser the existing coverage of sample locations for the dependent variable, the worse the spatial interpolation will perform". In other words, non-spatial CV may perform well in samples with a high spatial density but not so well if the distribution of observations across space is sparse, as this typically increases extrapolation. Furthermore, this implies that bias in prediction accuracy may be a function of distance from the city center since observations usually become sparser farther outside where housing structures are less dense and markets tend to be less active (Gröbel and Thomschke, 2018).

Although many studies on hedonic machine learning approaches exist, spatial cross-validation has so far not been applied to a real estate context, let alone to algorithmic hedonic house price and rent estimation problems. Reported cross-validation errors are almost consistently lower than errors of alternative parametric methods, such as least squares or spatial autoregressive frameworks. In particular tree-based ensemble learners, such as bagging (Breiman, 2001) and boosting (Friedman, 2001), have been shown to be most promising for house price estimation compared to alternative machine learning

methods (see Antipov and Pokryshevskaya, 2012; Kok et al., 2017; Mullainathan and Spiess, 2017; Baldominos et al., 2018; Mayer et al., 2019, Hu et al., 2019; Ho et al., 2021). Pace and Hayunga (2020) find evidence that the gains in explanatory power achieved by boosting and bagging algorithms are mainly attributable to the exploitation of spatial structures in the data. Consistent with the previously outlined logic presented by Bahn and McGill (2007) as well as Gröbel and Thomschke (2018), they find the error variance of bagging to increase the farther the model extrapolates to a global domain which could indicate that the model is overfitted to the spatial structures of more frequently observed houses in central districts. This notion is in line with Bogin and Shui (2020) who found a significant degree of overfitting in their random forest measured by a holdout strategy using appraisal records of homes in rural areas.

The extensive scientific debate about spatial dependence in real estate together with the concurrent, steadily growing corpus of literature on machine learning applications for house price and rent predictions motivates us to assess the sign and magnitude of potential bias associated with spatial dependence when using conventional CV methods for model selection and model assessment. Moreover, we investigate whether spatial cross-validation is an appropriate technique to account for spatial autocorrelation in apartment rents when using predictive machine learning algorithms, although the primary intention is to interpolate within a delineated spatial polygon.

2.4 Data and Methodology

We first train and cross-validate tree-based algorithms using a cross-section of apartment rents, thereby applying random as well as spatial partitioning during the cross-validation procedure for both, model selection and model assessment. With everything else remaining equal, there should be no substantial difference in the selected hyperparameters nor the cross-validation errors between spatial and non-spatial models if the assumption of spatial randomness was fulfilled. In a second step, we calculate the out-of-sample predictive performance of the models by estimating the data from a holdout sample one quarter ahead. We then analyze the difference between in-sample cross-validation errors and the true out-of-sample prediction errors to assess the bias associated with the respective partitioning techniques. A non-spatial linear model, as well as spatial autoregressive models, are used as points of reference. Third, we evaluate the deviation in bias when excluding spatial control variables from the model specification. Based on the hypotheses elaborated in section two, we would expect the bias to increase in non-spatial modeling frameworks when spatial information is absent due to overfitting unexplained spatial dependence structures in the data to covarying but non-causal regressors. This will

be more closely evaluated in a fourth step by analyzing the residual spatial autocorrelation in all model alternatives.

2.4.1 Data Description

Our sample consists of a pooled cross-section of apartment rents from the Frankfurt residential market spanning the period from January 2019 through March 2020. The data were sourced from German multiple listing systems (MLS) and are confined to apartment rentals excluding single, semi-detached and terraced houses, student apartments, senior living accommodations, furnished co-living spaces and shortstay apartments. Data cleaning was performed to account for duplicates, missing values and erroneous data points. The final sample comprises a total of 9,256 asking rents observed on a monthly scale, including the properties' most important structural attributes and equipment as well as their coordinates. A typical way to reflect differences in demand for locations in parametric models is to include district fixed effects by means of location dummies such as pre-defined submarkets (e.g., Bourassa et al., 2003, 2007) or attractiveness zones (e.g., Doszyń, 2020) that are specified by real estate experts. In non-parametric machine learning models, the inclusion of spatial coordinates (i.e., latitude and longitude) facilitates the identification of relevant submarkets based on spatial patterns in the data without the need to provide specific location zones. This allows a model to construct more local sub models for the identified areas (see Pace and Hayunga, 2020). The use of continuous coordinates is more efficient because it is computationally less expensive than a matrix of location dummies while at the same time, coordinates have a finer resolution, so the models are not forced into using pre-defined spatial polygons that limit their flexibility. Also, having too many dummy variables or too few observations per location zone may favor overfitting the models. Since our data only contain postcode areas that have very limited economic meaning, we refrain from the inclusion of location zones and include the observations' coordinates by means of latitude and longitude. Moreover, distances to nearby amenities were added using an Open Street Maps API to control for locational and neighborhood effects.

The building age was calculated relative to the year 2018, and values with a construction date before 1900 were trimmed to avoid disproportionate leverage of those observations. The entry date was transformed into a decimal number in years, and logarithmic transformations were used for the apartment rent and living area. The summary statistics of the features univariate distributions is presented in Table 2.1. The number of entries in each month is distributed uniformly throughout the sample period without a significant time trend in apartment rents, as shown in Table 2.2.

Table 2.1: Summary Statistics

Variable	N	Mean	Median	SD	Min	Max
Continuous						
Rent per month [Euro]	9,256	1,088.77	940.00	647.97	190.00	10,000.00
Living Area [sqm]	9,256	75.20	70.00	35.07	10.00	440.00
Age [years]	9,256	44.59	46.00	39.34	-2.00	118.00
Entry date [years]	9,256	0.65	0.67	0.35	0.08	1.25
Latitude	9,256	50.12	50.12	0.02	50.08	50.21
Longitude	9,256	8.66	8.66	0.05	8.49	8.78
Discrete						
Rooms	9,256	2.55	2.50	1.00	1.00	8.50
Floor	9,256	2.54	2.00	2.82	-0.50	39.00
Dummies [1=yes, 0=no]						
Bathtub	9,256	0.53	1.00	0.50	0.00	1.00
Refurbished	9,256	0.22	0.00	0.41	0.00	1.00
Built-in kitchen	9,256	0.71	1.00	0.45	0.00	1.00
Balcony	9,256	0.65	1.00	0.48	0.00	1.00
Parking	9,256	0.48	0.00	0.50	0.00	1.00
Elevator	9,256	0.50	1.00	0.50	0.00	1.00
Terrace	9,256	0.13	0.00	0.34	0.00	1.00
Distances						
NUTS centroid [km]	9,256	3.65	3.68	1.87	0.01	10.84
Bakery [km]	9,256	0.39	0.26	0.41	0.00	1.61
Bar [km]	9,256	0.73	0.52	0.64	0.00	2.54
Biergarten [km]	9,256	1.16	0.97	0.77	0.02	3.10
Café [km]	9,256	0.36	0.25	0.33	0.00	1.31
School [km]	9,256	0.31	0.28	0.17	0.02	0.75
Supermarket [km]	9,256	0.26	0.22	0.17	0.00	0.75
Bus station [km]	9,256	3.13	2.77	1.54	0.09	7.56

Notes: This table reports the univariate distributions of 9,256 asking rents of residential apartments listed between January 2019 and March 2020 in Frankfurt (Germany), and their observed characteristics after data cleaning. The entry date is represented as a decimal number in years, the building age is calculated relative to the year 2018 and is trimmed for buildings constructed before the year 1900, distances are calculated as the Euclidean distance to the apartment in kilometers, binary variables indicate whether a characteristic is included in the apartment (1) or not (0). N: number of observations, SD: standard deviation, Min: minimum value, Max: maximum value.

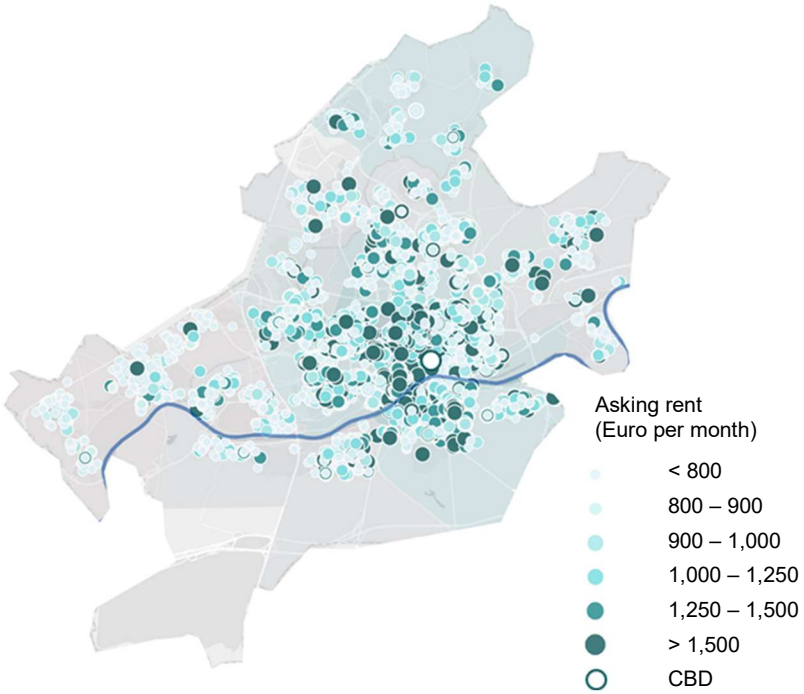
Table 2.2: Listings per Month

Entry Date	N	Mean Rent [Euro]	Mean Rent [Euro/sqm]
Jan-19	632.00	1,129.57	14.57
Feb-19	586.00	1,116.31	14.34
Mar-19	677.00	1,081.28	14.20
Apr-19	576.00	1,118.35	14.39
May-19	685.00	1,135.66	14.54
Jun-19	602.00	1,084.38	14.41
Jul-19	746.00	1,083.23	14.22
Aug-19	755.00	1,014.57	14.20
Sep-19	602.00	1,177.08	14.37
Oct-19	633.00	1,088.54	14.26
Nov-19	613.00	1,005.09	13.85
Dec-19	392.00	1,021.86	14.48
Jan-20	600.00	1,101.10	14.82
Feb-20	611.00	1,096.45	14.75
Mar-20	546.00	1,068.98	14.73

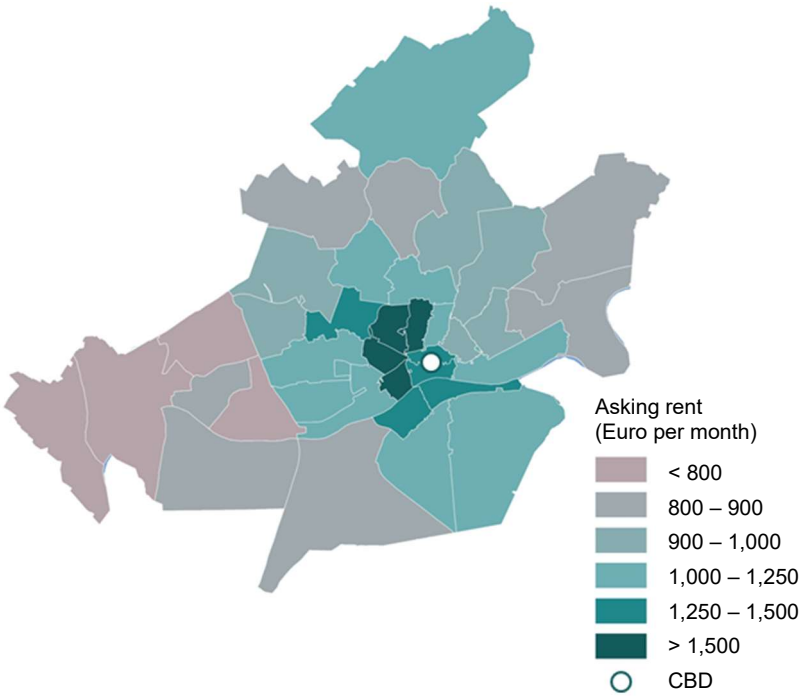
Notes: This table reports the occurrence of apartment listings throughout the sample period from January 2019 to March 2020 on a monthly scale with the respective mean absolute rent in Euro per month and the mean rent in Euro per square meter per month. N: number of observations.

Figure 2.2: Spatial Sample Distribution

Panel A: Spatial Distribution of Apartment Rents



Panel B: Mean Apartment Rents on ZIP-code Level

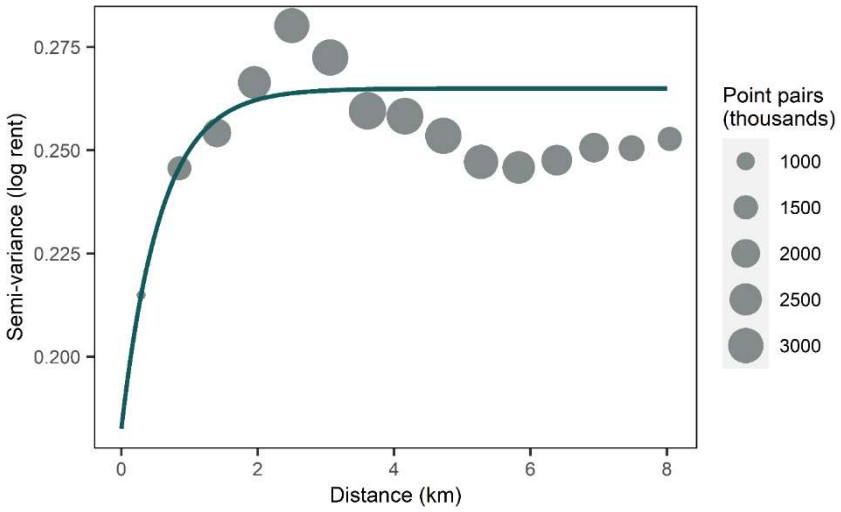


Notes: The upper map depicts the absolute monthly asking rent in Euro per month of each individual listing in our sample of 9,256 listings between January 2019 and March 2020 in Frankfurt. The bottom map shows the respective mean asking rents in Euro per month aggregated on a ZIP-code level.

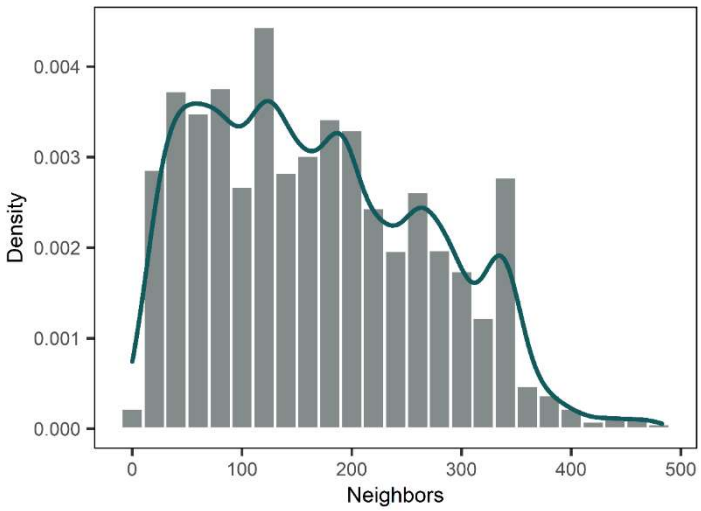
The spatial distribution of the data in Figure 2.2 Panel A does not seem to exhibit any distinct location bias, albeit the spatial density of observations increases toward the city center. As described by Gröbel and Thomschke (2018), this is not surprising as building structures are denser in central areas, which are moreover predominantly occupied by younger and more mobile tenants, resulting in higher fluctuation rates and subsequently more frequent rental offers compared to the outskirts. The average distances to the 1, 5, 10, 30, and 100-nearest neighbors amount to 0.02, 0.04, 0.06, 0.13 and 0.30 km respectively. Aggregated on a ZIP-code level, Figure 2.2 Panel B indicates that more expensive apartments tend to be clustered in the city center and along the north-south

Figure 2.3: Spatial Clustering of Apartment Rents

Panel A: Semi-variogram of the log Rent



Panel B: Distribution of Neighbors within the Spatial Autocorrelation Range



Notes: The empirical Matérn semi-variogram model suggest a spatial autocorrelation range of 0.58 kilometers, that is the distance up to which spatial autocorrelation persists in the data. The histogram presents the distribution of neighbors within the spatial autocorrelation range.

axis. More formally, spatial clustering of apartment rents is confirmed by the semi-variance of the log rent as depicted in Figure 2.3 Panel A. The empirical Matérn semi-variogram model suggests a spatial autocorrelation range of 0.58 km, which is the distance up to which spatial dependence between observations persists in the data (Cressie, 1993). In other words, an apartment in our sample has on average 168 neighbors that do not satisfy the assumption of independence. This number increases with spatial density and vice versa. The distribution of neighbors within the spatial autocorrelation range is presented in Figure 2.3 Panel B.

2.4.2 Methodological Approach

Parametric Models

We use an ordinary least squares (OLS) estimator as a non-spatial parametric benchmark model. Written in matrix notation, the multiple linear regression model follows a log-linear functional form of the relationship stated in equation (2.1)

$$Y = \alpha + X\beta + \varepsilon \tag{2.1}$$

with Y being the response vector with n observations of log-transformed apartment rents, α being a fixed intercept, X representing the regressor matrix with n rows and p columns, β being the corresponding n x 1 coefficient vector and ε being the random error term vector of length n.

Our modeling approach is based on the principle to avoid overfitting and bias in error estimates to isolate the bias originating from spatial dependence. We thus follow Harrell (2015) and Mayer et al. (2019) and exclude only regressors with almost no predictive power from the hedonic equation but, at the same time, refrain from the inclusion of interaction or quadratic terms to keep the models simple. We test the null hypothesis of spatial randomness in the OLS residuals by calculating the Moran's *I* statistic (Cliff and Ord, 1973) and account for potential spatial dependence using the spatial econometric toolbox.

Opposed to the spatial cross-validation technique (where spatially autocorrelated information is explicitly excluded from the model), the mechanism of spatial econometric models works the exact opposite way by explicitly mapping spatial interactions among neighboring observations as spatial lag terms in the functional form of the relationship. The spatial weight matrix W formally defines the spatial relationship between observations. To identify the source of the underlying processes causing spatial effects in the data, a model specification search is conducted following the general to specific approach advocated by LeSage and Pace (2009) and LeSage (2014) starting from the Spatial Durbin Model (SDM) in equation (2.2)

$$Y = \alpha + \rho WY + X\beta + WX\theta + \varepsilon \tag{2.2}$$

as well as the Spatial Durbin Error Model (SDEM) in equations (2.3) and (2.4).

$$Y = \alpha + X\beta + WX\theta + u \tag{2.3}$$

$$u = \lambda Wu + \varepsilon \tag{2.4}$$

Subsequently, we perform likelihood-ratio tests to challenge the relevance of the autoregressive coefficients ρ, θ, and λ from the SDM and the SDEM against more specific spatial model alternatives that include only one out of the two interactions respectively (Anselin, 1988; Anselin et al., 1996). As stated by LeSage and Pace (2009), this top-down approach has the advantage that the SDM still produces unbiased coefficient estimates even when the true data-generating process is a more specific model.

Non-Parametric Models

Among a wide variety of algorithmic hedonic methods evaluated in comparative studies, ensembles of regression trees using bagging and boosting techniques have consistently shown the most promising results concerning predictive power (see Antipov and Pokryshevskaya, 2012; Kok et al., 2017; Baldominos et al., 2018; Mayer et al., 2019; Hu et al., 2019; Ho et al., 2021; Bogin and Shui, 2020).

The idea behind a regression tree is to stratify the feature space into a set of M disjoint intervals $R_1, R_2, \ldots, R_M$, each of which is assigned a constant c_m as predicted value, being referred to as the leaf or terminal node of the tree (Breiman et al., 1984). Intervals are created by recursive binary partitioning at the nodes t_p, choosing a split-point s of a particular feature x_j in the process of solving a minimization problem that can be expressed as in equation (2.5)

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right] \tag{2.5}$$

under the conditions stated in equations (2.6) and (2.7)

$$R_1(j,s) = \{X|X_j \leq s\} \tag{2.6}$$

$$R_2(j,s) = \{X|X_j > s\} \tag{2.7}$$

following the notation of Hastie et al. (2009). Each observation is subsequently passed down the tree branches by making binary decisions at each split following the feature values until the data point has reached its final leaf. As single regression trees tend to

overfit easily and do not perform well on unseen data, we focus on ensembles of regression trees, more specifically the bagging-based random forest regression introduced by Breiman (2001) and the extreme gradient boosting algorithm developed by Chen and Guestrin (2016) which is an extension of the gradient tree boosting method dating to Friedman (2001).

The bagging algorithm grows a forest of many individual but slightly different trees b using bootstrapped training samples. Instead of pruning the trees, which is typically done to counteract the overfitting of individual regression trees, the trees in a forest are grown deeply, resulting in more terminal nodes with fewer observations being allocated to the same constant c_m as the predicted value. The minimum node size min_{node}, which is the smallest possible number of observations in each leaf, determines the depth of the tree. Deep trees can lead to overfitting, thus reacting very sensitively to changes in the training sample, whereas shallow trees may not pick up information from the data adequately, hence producing models which are underfitted (Kok et al., 2017). Consequently, the relatively deep trees in a forest have a high variance but low bias. The variance is then removed by averaging over the results of the b bootstrap trees to increase robustness (Breiman, 1996; James et al., 2013). Unlike conventional forests, a random forest considers only a randomly selected subset of m predictors from all available predictors p at each split, thereby introducing an additional source of variation into the model to counteract overfitting.

Boosting works somewhat similar, but unlike bagging, where trees are grown simultaneously and independently, boosted trees are grown sequentially using the residuals' information content from preceding trees to continue learning. At each sequence, a new regression tree is fitted to the residuals of the previous tree and is added into the fitted function, thereby iteratively updating the model (James et al., 2013). Model fit is improved with each iteration until the number of boosting rounds is exhausted. To avoid overfitting boosted trees, a shrinkage parameter η, also referred to as learning rate, is used to slow down the learning process by making the error corrections in each round more conservative. Like the random forest, the extreme gradient boosting algorithm is a more regularized alternative of the gradient boosting technique in the sense that it attempts to decorrelate the individual trees by using only a randomly subsampled portion $\frac{m}{p}$ of features in each round to increase the robustness of the boosting trees.

Generally speaking, model fit tends to increase with higher flexibility, such as a lower min_{node}, or a lower η. However, this is tied to the risk of overfitting the training data, subsequently resulting in poor generalizability of the models. Thus, a level of regularization

has to be chosen to limit the flexibility of a model. As manual choice of these hyperparameter combinations is arbitrary and unlikely to yield satisfactory results, model selection is typically conducted using data-driven optimization by iteratively testing different hyperparameter combinations within a pre-defined search space and evaluating their performance based on cross-validation errors in the attempt to minimize a given loss function. Eventually, the set of hyperparameters yielding the lowest cross-validation error rate (that is, the best performance in the out-of-sample experiment) is chosen as the optimal hyperparameter combination (James et al., 2013). This approach is called grid search optimization and is a widely used algorithm for automated hyperparameter tuning. For computational reasons, we restrict our search space to the three main tuning parameters of each model described above. For the random forest, these are the number of bootstrap trees b, the number of available features m at each split and the minimum node size min_{node} in each terminal node. Likewise, the number of boosting rounds n_{rounds}, the column subsample $\frac{m}{p}$ and the shrinkage parameter η are equally important to the boosting algorithm. As optimization criterion, we adopt the minimization of squared residuals from the least squares estimator.

2.4.3 Performance Evaluation

We measure predictive performance (i.e., the true error rate) of our models by predicting out-of-sample data from the first quarter of 2020, which is referred to as the "holdout sample" in the remainder of this study. The remaining "in-sample" data of 2019 is used to calibrate the models and approximate their out-of-sample predictive performance (i.e., the expected error rate). The resampling strategies used for the steps of model selection as well as model assessment are outlined below.

In the parametric world, model selection is performed manually by specifying a functional form of the estimator a priori rather than following a data-driven approach to maximize fit. Moreover, the flexibility of such models is usually restrained by linearity assumptions that make overfitting less of a problem. Thus, prediction errors of linear models are typically estimated by simple re-substitution of the data used for model fitting (Efron, 1983; Simon, 2007). We follow this standard statistical approach to approximate the predictive accuracy of the parametric benchmark models and calculate the true error rate by regressing the holdout data using the estimated parameters.

As elaborated earlier, data-driven approaches require resampling methods such as cross-validation to calculate fair estimates of predictive performance. To isolate the effects of spatial dependence and allow for a fair comparison between random and spatial

partitioning, the resampling strategy is designed with the principle to eliminate any bias resulting from sources other than spatial autocorrelation that could potentially distort our results. Thus, we perform k-fold cross-validation for model selection and model assessment to avoid that error estimates are biased by chance due to a specific training or validation set.

It is worth mentioning that the choice of k is associated with a bias-variance trade-off, that is, the bias becoming smaller with each additional fold whereas the variance of the error estimates increases at the same time due to a higher correlation of the training sets (Hastie et al., 2009). As suggested by theory and empirical research, a k of five or ten proves to be a reasonable compromise in this tradeoff, whereby a value of five is only recommended for very large datasets to ensure enough observations for model training (Breiman and Spector, 1992; Kohavi, 1995; Hastie et al., 2009; James et al., 2013). With the primary aim to isolate the effect of spatial autocorrelation, we accept a higher error variance in favor of lowering bias and therefore set k to ten, as was also done by Park and Bae (2015), Chin et al. (2020), Hu et al. (2019) as well as Rico-Juan and Taltavull de La Paz (2021).

Further following the logic that information flow between training and test observations leads to biased cross-validation errors, we apply a nested resampling strategy that strictly separates data used for model selection from data used for model assessment. This is important since assessing model performance on the same data used for model selection does not yield an unbiased estimate of prediction error but more of a re-substitution error (Varma and Simon, 2006). Thus, nested resampling consists of two resampling loops, that is the inner resampling loop for hyperparameter tuning (i.e., model selection), which is wrapped within the outer resampling loop for performance evaluation (i.e., model assessment) such that model selection and model assessment is repeatedly performed on mutually exclusive subsamples, thereby simulating independent data throughout the entire workflow of the algorithm (Simon, 2007).

Following this strategy, the resulting cross-validation errors should, at least in theory, provide an unbiased picture of out-of-sample predictive performance to be expected from the models if the assumption of spatial randomness was fulfilled, thus enabling us to disentangle the effects of spatial dependence by using spatial CV.

To implement spatial partitioning in the cross-validation procedure, we apply a k-means clustering algorithm as proposed by Brenning (2012). The k-means clustering method is a universal and commonly used technique to detect a specified number of k clusters among n observations based on a given set of features. In a first step, the algorithm randomly

chooses k centroids in the multi-dimensional feature space. The initial clustering is achieved by allocating each of n observation to the "nearest" centroid in the feature space (i.e., by minimizing the Euclidean distance from the feature values to the centroid). The positions of the cluster centroids are then adjusted by taking the mean feature values of each grouping and the clustering is repeated. The clusters are iteratively adjusted until the allocation doesn't change anymore, so the within-cluster sum of squares is minimized (James et al., 2013).

The goal of spatial cross-validation is to maximize the distance between training and test folds. In this context, a cluster refers to a fold whereby k denotes the number of equally sized folds to be partitioned and the point coordinates (latitude and longitude) represent the features. The feature space is a two-dimensional scatterplot as depicted in Figure 2.1. The algorithm arranges the folds in a way that minimizes the average distances within each fold and maximizes the average distance between the folds. This effectively decreases spatial autocorrelation between training and test data.

Using spatial and non-spatial partitioning, our nested resampling strategy provides four alternatives to calculate cross-validation errors which are:

(1) *non-spatial* model selection + *non-spatial* model assessment,
(2) *non-spatial* model selection + *spatial* model assessment,
(3) *spatial* model selection + *spatial* model assessment, and
(4) *spatial* model selection + *non-spatial* model assessment.

The first alternative is the conventional "off-the-shelf" approach typically applied in the hedonic literature, although nesting is not common yet. In contrast, the third option describes a pure spatial approach that should reduce spatial dependence between training and test observations to a minimum but may result in too pessimistic expectations of predictive performance. As the prediction goal in a housing context is not a pure spatial one, the second and fourth alternative could potentially provide a fair compromise in the trade-off between reduction of spatial autocorrelation and the extrapolation range introduced into the model.

To arrive at a final model that can predict the holdout data, all steps of the algorithm need to be executed once again, whereby the cross-validation in the outer loop is replaced by the holdout sample such that the full information from 2019 is used to train a model that predicts apartment rents from the first quarter of 2020 (Varma and Simon, 2006; Simon, 2007). Analogous to the nested resampling for the estimation of prediction error, optimal hyperparameters for the final prediction model are once again derived using spatial and non-spatial grid search CV resulting in two alternatives for the true error rate.

Based on the true error rates, we benchmark predictive performance and determine the bias in error estimates. Model accuracy and precision are assessed using the coefficient of determination (R^2), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean squared error (RMSE). To measure variation in the residuals, we calculate the interquartile range (IQR), the coefficient of dispersion (COD) and an error bucket that includes the proportion of predictions within 10% of the true value (PE10). The mean percentage error (MPE) is used as a measure of biasedness. Subsequently, the asymptotic properties of all estimators are evaluated by comparing the distributions of error estimates resulting from the respective resampling strategies to the distributions of the true prediction errors.

2.5 Results

This section first presents the final model specifications and evaluates differences in hyperparameters selected by the automated grid search CV. Second, the results of the error-based model assessment are reported to determine the bias for the respective resampling strategies, and the asymptotic behavior of the estimators is discussed. Third, we investigate the residual spatial autocorrelation in each of the estimated models to draw conclusions on whether differences in the selected levels of regularization and the related model performance are linked to overfitting spatial structures in the data.

2.5.1 Model Selection

All variables listed in Table 2.1 were kept in the final model specification of the linear model. Following the principle to avoid overfitting, only the squared term for the building age and no interaction terms were additionally included. The resulting model specification serves as a baseline for all subsequent model alternatives and is hereinafter referred to as model specification "A". We estimate an alternative model specification "B", which does not consider locational and neighborhood characteristics in the regressor matrix, to see how the results change in the absence of spatial controls. All remaining modeling decisions for the linear models outlined below are based on specification A and were adopted for specification B. If not stated otherwise, the presented results refer to specification A. The respective regression outputs of the OLS estimator are shown in Appendix Table 2.6.

The Moran's *I* statistic of the OLS residuals rejects the null hypothesis of spatial randomness in price formation processes at a close-to-zero level of significance. The likelihood-ratio (LR) tests of a restriction of the SDM confirms the common factor hypothesis and implies the presence of both endogenous as well as exogenous interaction effects, leading to the acceptance of the SDM. The relevance of the SDEM was further investigated as an

alternative spatial model. Again, the LR-tests reject a simplification of the SDEM with p-values close to zero. We subsequently consider both the SDM as well as the SDEM as spatially conscious linear model alternatives. As spatial density of observations tapers toward the outskirts, we follow Pace et al. (2000) for the specification of W and choose a κ-nearest neighbors (κ-nn) matrix where each observation has a fixed number of κ neighbors. After evaluating different values for κ between 10 and 100, we eventually set the number of neighbors to 30, as this yields fair error estimates without diluting spatial effects in the lag terms. Overall, results remain robust for different choices of κ as well as for distance-based matrices with different boundaries.

The optimal hyperparameters selected by the grid search CV after 100 evaluations are shown in Table 2.3. For the random forest, there are no structural differences in the number of trees b nor the number of features m considered at each split, although spatial tuning seems to favor slightly higher values of m. Notable deviations can be observed in the minimum node size min_{node} that determines the depth and, thus, the complexity of the individual trees in the forest. The non-spatial grid search CV consistently prefers a min_{node} between one and two, which is significantly lower compared to the spatial model that has on average a minimum node size of five. A min_{node} of one provides the trees with the flexibility to have virtually infinite vertical growth, allowing them to remove all noise from the data (Kok et al., 2017). Or as expressed by Mullainathan and Spiess (2017), a tree which grows one leaf for each observation in the data "[…] will have perfect fit, but of course this is really perfect overfit", consequently yielding unsatisfactory predictions for unseen data.

A similar pattern can also be observed for the XGB. Again, there are no remarkable differences in the size of the column subsample $\frac{m}{p}$. However, the spatial instantiation of the resample call in the inner loop requires on average only 405 boosting rounds versus 593 boosting rounds for the non-spatial CV. Although the rate η at which the boosting algorithm learns at each round is more conservative in the nonspatial model, a higher number of n_{rounds} indicates excessive error corrections that may result in a model that overfits the residuals. The selection of less complex models compared to the non-spatial tuning persists for model specification B, although the higher complexity of the non-spatial random forest is now even more distinct. These findings corroborate our hypotheses derived from studies in other fields such as Le Rest et al. (2014), Roberts et al. (2017) as well as Meyer et al. (2019), who state that non-spatial partitioning during resampling is associated with the choice of overly complex models if the data exhibits spatial dependence.

Table 2.3: Optimal Hyperparameters selected by 10-fold Cross-Validation

Fold	b (1)	b (2)	b (3)	b (4)	m (1)	m (2)	m (3)	m (4)	min_{node} (1)	min_{node} (2)	min_{node} (3)	min_{node} (4)
	Panel A1: Random Forest including Spatial Controls											
1	650	500	400	400	9	9	12	14	2	2	8	7
2	300	350	550	250	9	9	10	10	1	1	5	8
3	300	500	200	500	9	9	10	10	1	1	6	6
4	300	600	500	450	7	9	9	10	2	1	2	1
5	500	450	300	650	7	7	10	10	1	2	6	5
6	350	500	450	600	9	9	9	12	2	2	7	6
7	600	300	500	500	9	7	12	9	1	1	1	7
8	650	650	650	550	7	9	9	10	1	1	3	7
9	550	650	600	400	9	9	10	9	1	3	4	6
10	550	500	650	500	9	10	9	10	2	1	2	3
	Panel B1: Random Forest excluding Spatial Controls											
1	550	650	200	500	5	6	5	6	2	1	9	10
2	250	250	450	550	5	5	5	6	1	1	9	8
3	600	450	500	250	5	5	6	5	2	1	10	8
4	500	550	200	600	5	5	5	6	2	1	8	10
5	600	650	500	500	5	5	5	6	1	2	10	8
6	600	350	450	300	5	5	6	6	2	2	9	9
7	450	650	350	350	5	6	6	6	1	1	9	9
8	400	200	450	600	5	5	6	5	1	1	10	9
9	450	450	600	600	5	5	5	6	1	1	8	10
10	450	550	500	300	6	6	6	6	2	1	9	9
Fold	n_{rounds} (1)	n_{rounds} (2)	n_{rounds} (3)	n_{rounds} (4)	m/p (1)	m/p (2)	m/p (3)	m/p (4)	η (1)	η (2)	η (3)	η (4)
	Panel A2: Extreme Gradient Boosting Trees including Spatial Controls											
1	600	550	250	400	72%	58%	65%	45%	0.06	0.07	0.04	0.05
2	650	600	550	250	38%	78%	65%	58%	0.06	0.1	0.02	0.06
3	650	550	200	300	58%	85%	52%	85%	0.1	0.07	0.08	0.05
4	550	600	500	350	78%	78%	58%	85%	0.08	0.08	0.02	0.02
5	600	600	200	550	72%	72%	78%	45%	0.07	0.08	0.04	0.02
6	600	600	450	450	65%	45%	45%	45%	0.07	0.09	0.06	0.03
7	650	550	550	300	72%	78%	45%	65%	0.08	0.09	0.09	0.04
8	500	650	450	350	52%	52%	78%	38%	0.08	0.09	0.02	0.07
9	650	650	500	650	58%	65%	78%	52%	0.05	0.06	0.02	0.02
10	500	550	400	450	78%	45%	72%	65%	0.07	0.08	0.03	0.03
	Panel B2: Extreme Gradient Boosting Trees excluding Spatial Controls											
1	500	650	500	600	72%	72%	65%	78%	0.07	0.06	0.01	0.01
2	500	400	300	200	78%	78%	72%	85%	0.05	0.07	0.02	0.03
3	650	650	250	550	58%	65%	72%	85%	0.04	0.04	0.02	0.01
4	650	650	550	250	72%	65%	58%	85%	0.05	0.06	0.01	0.02
5	650	600	550	600	65%	58%	72%	72%	0.04	0.07	0.01	0.01
6	550	600	200	500	65%	65%	85%	65%	0.05	0.05	0.03	0.02
7	400	600	600	550	78%	58%	72%	65%	0.09	0.08	0.01	0.01
8	450	350	500	500	52%	58%	72%	78%	0.07	0.07	0.01	0.01
9	450	650	200	300	78%	65%	72%	85%	0.06	0.04	0.03	0.02
10	450	450	500	600	65%	45%	65%	65%	0.07	0.06	0.01	0.01

Notes: This table reports the optimal hyperparameters for each fold of the inner loop in the four alternatives of the nested resampling procedure: *(1) non-spatial tuning + non-spatial validation*, *(2) non-spatial tuning + spatial validation*, *(3) spatial tuning + spatial validation*, *(4) spatial tuning + non-spatial validation* selected by an automated grid search using 10-fold cross-validation. b: number of bootstrap trees, m: number of features in the column subsample, min_{node}: minimum node size, n_{rounds}: number of boosting rounds, m/p: portion of all available features p in the column subsample, η: learning rate (eta).

2.5.2 Model Assessment

The subsequent section discusses how the differences in model selection affect model accuracy and whether increased complexity is indeed linked to overfitting and vice versa. Therefore, we analyze the bias and the asymptotic properties of our estimated models by comparing the true one quarter ahead prediction error to the expected error rates resulting from the respective resampling strategies outlined in the "Performance Evaluation" section. Bias is measured as the difference between the true error rate and the expected error rate. The aggregated performance measures are presented in Table 2.4.

The re-substitution errors from the linear models are significantly lower than the true error rates on all accounts, which, however, is not surprising (Efron, 1983). Whereas this overoptimism is smallest for the SDM, the error estimates of the SDEM are even more biased than those of the OLS model. This is mainly attributable to the relatively weaker predictive performance of the SDEM, since re-substitution errors of both spatial models are only marginally different from each other. Having said that, it is worth mentioning that spatial autoregressive models are primarily designed for statistical inference rather than out-of-sample predictions.

For the non-parametric models, one can see an improvement in all performance measures compared to the linear models, which is not surprising and in line with the literature (see Antipov and Pokryshevskaya, 2012; Yoo et al., 2012; Gu and Xu, 2017; Kok et al., 2017; Mullainathan and Spiess, 2017; Čeh et al., 2018; Bogin and Shui, 2020; Pace and Hayunga, 2020). With respect to predictive power, the XGB yields the most accurate results, closely followed by the RFR. Interestingly, performance measures do not seem noticeably affected by the resampling strategy in the inner loop for hyperparameter tuning despite the higher levels of regularization in the spatial models. Hence, predictive power is almost identical no matter whether spatial dependence has been accounted for during tuning or not.

Distinctive differences between spatial and non-spatial cross-validation can be observed for the outer resampling loop though. Compared to the true predictive performance, non-spatial CV errors are overly optimistic for both the RFR as well as the XGB. In contrast, spatial CV consistently yields overly pessimistic but more reliable approximations of prediction errors compared to non-spatial CV errors. It is, moreover, noteworthy that non-spatial hyperparameter tuning combined with spatial performance evaluation results in the most pessimistic cross-validation errors for all measures except the MAE and the MAPE of the random forest. The understatement of predictive accuracy is not surprising in this case since the model was trained with the objective to interpolate and subsequently validated by extrapolating to a new spatial domain, thus yielding non-optimal results.

Table 2.4: Error-based Performance Matrix

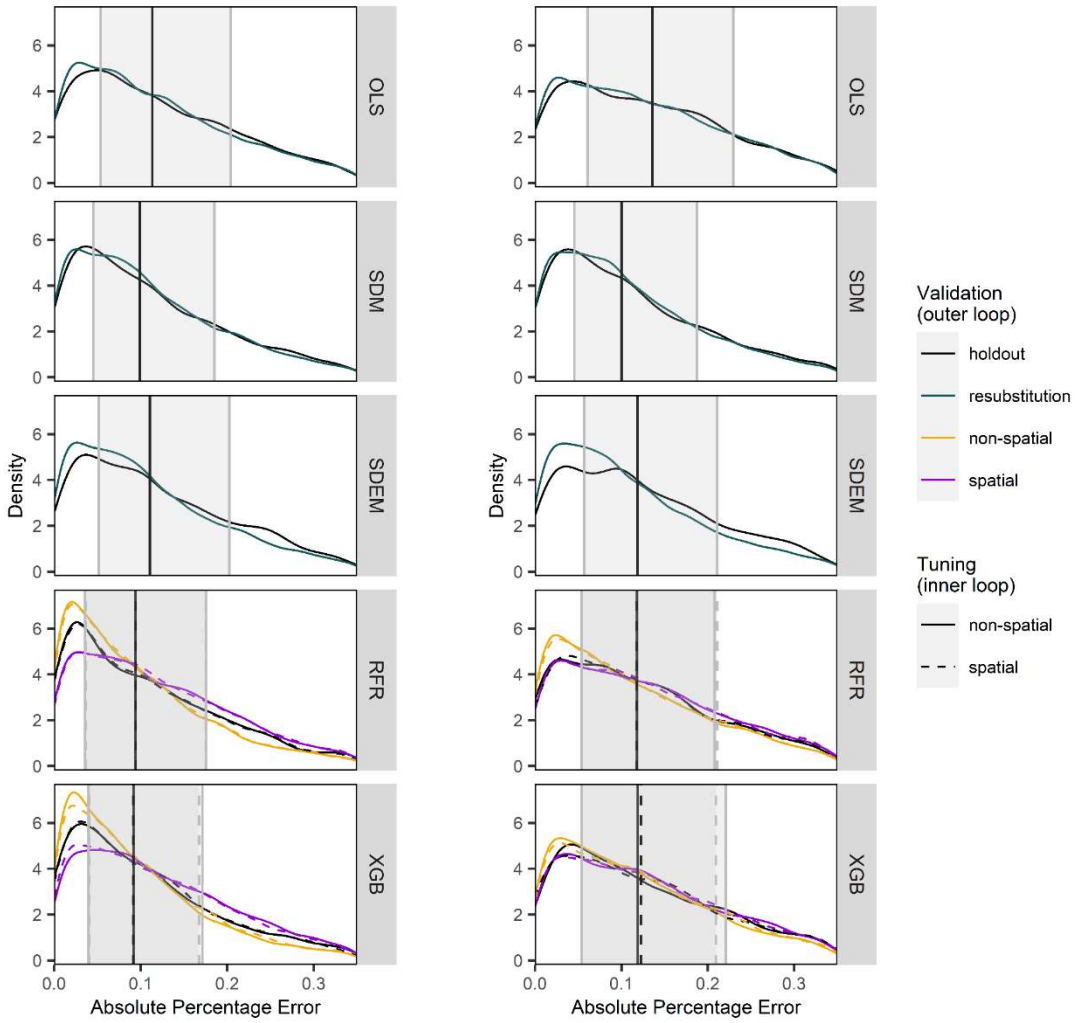
Method	Resampling Strategy	R^2	MAE	MAPE	MPE	RMSE	PE10	IQR	COD
		Panel A: Models including Spatial Controls							
OLS	*holdout*	78.81%	168.24	15.03%	-0.61%	292.14	44.45%	205.39	3.52%
	re-substitution	85.13%	149.79	13.91%	1.64%	251.01	47.21%	199.63	13.73%
SDM	*holdout*	81.62%	153.56	13.77%	-0.91%	272.09	50.54%	182.57	3.25%
	re-substitution	87.50%	134.81	12.64%	1.40%	230.13	51.81%	177.15	12.52%
SDEM	*holdout*	79.08%	166.51	14.70%	-1.57%	290.29	45.65%	198.80	3.45%
	re-substitution	87.46%	134.74	12.63%	1.40%	230.53	51.86%	175.95	12.51%
RFR	*holdout (non-spatial tuning)*	85.13%	143.48	12.79%	-0.33%	244.69	52.31%	169.58	3.14%
	holdout (spatial tuning)	85.14%	143.58	12.81%	-0.30%	244.64	52.48%	171.37	3.15%
	(1) non-spatial/non-spatial	89.42%	116.59	10.93%	1.45%	211.74	59.15%	141.57	11.12%
	(2) non-spatial/spatial	82.62%	157.30	14.35%	0.67%	271.40	45.83%	208.53	14.69%
	(3) spatial/spatial	83.07%	157.41	14.39%	0.58%	267.90	45.27%	208.00	14.69%
	(4) spatial/non-spatial	89.49%	117.93	11.08%	1.42%	211.07	58.65%	144.64	11.26%
XGB	*holdout (non-spatial tuning)*	85.20%	142.45	12.89%	0.71%	244.15	53.04%	165.43	3.15%
	holdout (spatial tuning)	85.21%	142.96	12.83%	0.16%	244.06	52.87%	171.23	3.11%
	(1) non-spatial/non-spatial	90.93%	112.66	10.53%	1.13%	196.08	60.71%	140.46	10.67%
	(2) non-spatial/spatial	83.90%	157.16	14.25%	-0.54%	261.20	44.86%	214.47	14.52%
	(3) spatial/spatial	84.54%	152.67	13.95%	0.00%	255.98	46.21%	202.73	14.19%
	(4) spatial/non-spatial	90.15%	117.11	10.97%	1.12%	204.33	58.27%	149.12	11.14%

Table 2.4 (continued)

		Panel B: Models excluding Spatial Controls							
OLS	*holdout*	74.52%	188.95	16.80%	-0.67%	320.33	38.13%	250.15	3.96%
	re-substitution	80.58%	173.36	15.99%	2.08%	286.89	40.23%	235.13	15.66%
SDM	*holdout*	81.56%	154.06	13.82%	-0.91%	272.49	49.91%	184.02	3.25%
	re-substitution	87.38%	135.79	12.72%	1.41%	231.27	51.77%	175.98	12.59%
SDEM	*holdout*	75.90%	177.71	15.41%	-2.07%	311.53	42.17%	218.24	3.67%
	re-substitution	87.08%	136.36	12.74%	1.42%	233.96	51.74%	175.61	12.60%
RFR	*holdout (non-spatial tuning)*	80.66%	171.26	15.14%	0.17%	279.07	44.05%	219.94	3.76%
	holdout (spatial tuning)	80.59%	172.98	15.30%	0.22%	279.60	43.54%	217.75	3.83%
	(1) non-spatial/non-spatial	85.80%	143.98	13.40%	1.67%	245.33	49.17%	185.16	13.68%
	(2) non-spatial/spatial	80.51%	173.25	15.85%	2.00%	287.38	40.70%	228.44	16.05%
	(3) spatial/spatial	80.29%	172.41	15.73%	1.96%	289.02	40.75%	226.66	15.94%
	(4) spatial/non-spatial	85.60%	146.48	13.62%	1.69%	247.07	48.66%	191.14	13.96%
XGB	*holdout (non-spatial tuning)*	80.33%	176.27	15.85%	1.65%	281.48	42.69%	223.19	3.83%
	holdout (spatial tuning)	79.47%	173.97	15.28%	-1.17%	287.56	43.26%	225.95	3.79%
	(1) non-spatial/non-spatial	85.61%	148.24	13.77%	1.63%	246.96	46.97%	194.32	13.96%
	(2) non-spatial/spatial	81.26%	172.33	15.78%	1.97%	281.83	40.39%	226.85	15.93%
	(3) spatial/spatial	79.75%	173.80	15.31%	-1.18%	292.96	40.78%	229.86	15.77%
	(4) spatial/non-spatial	84.24%	154.54	14.03%	-0.52%	258.41	45.02%	208.12	14.49%

Notes: This table reports the performance measures for the re-substitution errors in the linear models and the cross-validation errors in the non-parametric models *(1) non-spatial tuning + non-spatial validation, (2) non-spatial tuning + spatial validation, (3) spatial tuning + spatial validation, (4) spatial tuning + non-spatial validation* in comparison to the true predictive performance from the errors of the holdout sample. R^2: coefficient of determination, MAE: mean absolute error, MAPE: mean absolute percentage error, MPE: mean percentage error, RMSE: root mean squared error, PE10: error bucket of estimates within 10% of the true value, IQR: interquartile range, COD: coefficient of dispersion. Absolute values are reported in Euro per month.

Figure 2.4: Distribution of the Absolute Percentage Error



Notes: The density plots present the estimated distribution of the absolute percentage error resulting from the respective resampling strategies in comparison to the true out-of-sample distribution of the absolute percentage error from the holdout sample. The line type represents the resampling strategy used in the inner loop for model selection (non-parametric models only) and the line color represents the resampling strategy applied in the outer loop for model assessment. The true out-of-sample distribution is represented in black. The shaded areas depict the interquartile range, that is the area between the first quartile and the third quartile of the true absolute percentage error with the middle line representing the median.

The bias in non-spatial cross-validation errors is even more distinctive in Panel B of Table 2.4. In contrast, the spatially conscious cross-validation errors now closely resemble the true prediction errors, thereby reducing bias to a minimum. As already anticipated, the results indicate that over-optimism in non-spatial CV is likely to originate from spatial structures being overfitted to covarying but non-causal regressors during model training. This is particularly noticeable when locational and neighborhood controls are missing such that the spatial information content in the residuals is picked up by other attributes that are structured in space. Accounting for spatial dependence in the inner resampling loop had once again only a minor impact on both model accuracy and bias. Noteworthy, the non-parametric models are now outperformed by the SDM in terms of predictive accuracy,

which drops only marginally compared to specification A. This demonstrates the high robustness of the SDM, which is able to capture spatial effects through the spatial lag terms even in a scenario where locational control variables are not available. As stated by Doszyń (2020), machine learning methods require a very good data basis to take advantage of their flexibility, whereas less complex models are more robust in situations where extensive data is not available. The superiority of the SDM in the specification without spatial control variables moreover corroborates the findings of Pace and Hayunga (2020) who demonstrate that most of the improvement in accuracy achieved by machine learning models over parametric spatial models results from exploiting spatial structures in the data by creating spatially disaggregated models.

Figure 2.4 presents the asymptotic distribution of the absolute percentage error. The density plots reveal a higher variance and a lower kurtosis for prediction errors estimated by spatial CV opposed to the true error distribution. In comparison, nonspatial CV underestimates prediction errors, particularly in the lower tails of the distribution where errors are close to zero. For both the RFR and the XGB, non-spatial error estimates are centered around values that are considerably lower than the true means whereas spatial error estimates are more dispersed. This is affirmed by both the higher deviation of spatial error estimates from the median true error represented by the coefficient of dispersion as well as their larger spread illustrated by the interquartile range. These findings again confirm our expectations derived from literature in other spatial modeling fields (see Le Rest et al., 2014; Roberts et al., 2017; Schratz et al., 2019).

2.5.3 Residual Spatial Autocorrelation

Finally, we analyze the spatial autocorrelation found in the residuals of the models after calculating the Moran's *I* statistic to investigate whether overfitting is related to the exploitation of spatial dependence structures in the data. Since the relative magnitude of the Moran's *I* is only meaningful for identical spatial weight matrices, we also calculate the Z-scores as a standardized measure, which allows us to compare the spatial autocorrelation of in-sample and out-of-sample residuals (Anselin, 1995). The results are presented in Table 2.5.

The spatial linear models successfully reduce spatial autocorrelation in the in-sample residuals, although there is still spatial information content left, which is consistent with the findings of Pace and Hayunga (2020). The non-spatial cross-validation errors of the random forest exhibit spatial autocorrelation of roughly the same magnitude as the spatial linear models. Interestingly, spatially cross-validated errors show a significantly higher degree of spatial autocorrelation that even exceeds the Z-score of the simple OLS model.

Table 2.5: Residual Spatial Autocorrelation

		Panel A: Models including Spatial Controls			***Panel B: Models excluding Spatial Controls***		
Method	Resampling Strategy	Morans' *I*	Z-score	p-value	Morans' *I*	Z-score	p-value
OLS	*holdout*	0.17	55.44	0.00	0.29	95.36	0.00
	re-substitution	0.27	19.30	0.00	0.39	27.48	0.00
SDM	*holdout*	0.03	10.17	0.00	0.03	10.23	0.00
	re-substitution	0.16	11.60	0.00	0.17	11.76	0.00
SDEM	*holdout*	0.03	10.75	0.00	0.04	13.58	0.00
	re-substitution	0.25	17.89	0.00	0.34	23.59	0.00
RFR	*holdout (non-spatial tuning)*	0.17	11.75	0.00	0.33	23.45	0.00
	holdout (spatial tuning)	0.17	11.70	0.00	0.32	22.59	0.00
	(1) non-spatial/non-spatial	0.03	9.79	0.00	0.18	61.22	0.00
	(2) non-spatial/spatial	0.19	64.70	0.00	0.26	87.78	0.00
	(3) spatial/spatial	0.19	62.80	0.00	0.26	87.91	0.00
	(4) spatial/non-spatial	0.03	9.97	0.00	0.19	63.13	0.00
XGB	*holdout (non-spatial tuning)*	0.14	9.52	0.00	0.26	18.53	0.00
	holdout (spatial tuning)	0.14	10.10	0.00	0.33	23.26	0.00
	(1) non-spatial/non-spatial	-0.01	-1.90	0.06	0.15	51.03	0.00
	(2) non-spatial/spatial	0.16	54.67	0.00	0.23	75.07	0.00
	(3) spatial/spatial	0.16	53.59	0.00	0.28	92.21	0.00
	(4) spatial/non-spatial	0.01	2.55	0.01	0.21	71.22	0.00

Notes: This table reports the spatial autocorrelation found in the residuals of the models. A positive and significant Morans' *I* signals spatial clustering of similar values whereas a negative and significant Morans' *I* signals alternating values which indicates the presence of spatial outliers and/or spatial heterogeneity. The Z-score serves as a standardized value for comparison of the in-sample and out-of-sample statistics. It is calculated as the difference between the observed value of *I* and the expected value of *I* divided by the standard deviation of *I*, whereby the expected value of *I* is the theoretical mean defined as -1/(N-1), N being the number of observations.

The same applies to the boosted trees, although this method seems to understand spatial structures in the data slightly better. This may seem counterintuitive at first but knowing that non-spatial error estimates are biased downwards, the substantial differences in residual spatial autocorrelation between the spatial and the non-spatial cross-validation errors indicate that spatial partitioning in the outer resampling loop indeed prevents the models from exploiting unexplained spatially autocorrelated information from the test data during training. By and large, the outcomes are consistent for panel B but, unsurprisingly, the magnitude of spatial autocorrelation is in general higher as opposed to specification A, which further substantiates our hypotheses and underlines the importance of spatial cross-validation. Consistent with the results from the "Model Assessment" section, the SDM does have a superior understanding of spatial dependence structures in the data compared to all other models when spatial variables are not considered.

2.6 Conclusion

Recent literature has brought forth an increasing body of evidence that demonstrates a superior predictive performance of machine learning algorithms compared to parametric models for complex spatial regression problems involving the estimation of house prices

and rents. In non-parametric models, predictive performance is widely measured using resampling techniques such as cross-validation, which can be thought of as an out-of-sample experiment inside the original sample. This requires the statistical independence of the data to yield unbiased and meaningful prediction error estimates that can be used for model selection and model assessment. The inherent spatial dependence in house price and rent formation processes gives reason to question the validity of cross-validation errors in a hedonic context. Hence, this study investigates the adequacy of conventional k-fold cross-validation for the purpose of model selection and model assessment in an algorithmic hedonic context using tree-based boosting and bagging methods and proposes a spatially conscious alternative that attempts to reduce bias in cross-validation errors by accounting for the spatial proximity of observations.

Despite using a nested resampling strategy and applying column subsampling in our bagging and boosting algorithms to prevent overfitting, our results demonstrate that failing to account for spatial dependence during the cross-validation procedure still has two undesirable consequences. First, hyperparameter tuning using nonspatial grid search CV favors the selection of overly complex models that overfit spatial dependence structures in the training data, thereby compromising the models' generalizability. Second, performance estimates are artificially inflated through the exploitation of spatial dependence structures during model training, resulting in overly optimistic error estimates when compared to the true prediction errors.

In nested resampling approaches these two problems go hand in hand since the selection of overly complex models in the inner resampling loop is masked by overoptimistic accuracy measures during model assessment in the outer resampling loop. This can lead to spurious confidence in a model that overestimates predictive accuracy as nesting aims to simulate unseen data throughout the entire workflow of an algorithm, therefore suggesting unbiased error estimates (Varma and Simon, 2006). In contrast, spatial grid search CV prefers a higher level of regularization, thereby introducing extrapolation into the models, which results in error estimates that are slightly too pessimistic, yet closer to the true error rates.

An analysis of the residual spatial autocorrelation provides evidence that the spatially conscious CV technique hinders the algorithm from exploiting spatial dependence structures, thereby preventing overfitting. To see how the results vary with the extent to which spatial information is reflected in the feature space, we evaluate a second model alternative that does not consider spatial control variables. In this scenario, over-optimism in predictive accuracy is even more distinctive when spatial autocorrelation is not

accounted for, whereas the spatial CV procedure yields almost unbiased estimates of the true prediction error that converge asymptotically closer to the true error distribution.

Despite their flexibility and higher accuracy compared to traditional parametric methods, machine learning techniques are often criticized for their black box character that impedes direct model interpretation as well as for their high computational burden. To empirically illustrate where the costs and benefits of these methods lie, a least squares model as well as a linear spatial autoregressive framework are furthermore used as points of reference to assess predictive accuracy. Whereas the boosting algorithm performs best when spatial controls are reflected in the model, the spatial durbin model outperforms the non-parametric model alternatives in the absence of spatial information in the regressor matrix, which stresses the importance of considering parametric model alternatives besides non-parametric models.

We conclude that in a real estate hedonic context, state-of-the-art CV does not yield unbiased estimates of prediction error even when applying methods that intend to counteract overfitting. Resulting CV errors should rather be interpreted as an estimate of the lower bound of the true error rate. In contrast, spatial CV errors tend to be slightly too pessimistic but more reliable estimates of prediction errors. Likewise, the more conservative spatial CV errors can be regarded as an upper bound of prediction errors.

That being said, in scenarios where the study area is very small and clearly delineated so that spatial dependence structures do not vary significantly (i.e., on the submarket or ZIP-code level), spatial density of observations is high (i.e., CBD or city center), and spatial control variables are numerous, random partitioning of folds may yield fair estimates of predictive performance. However, for typical use cases (i.e., predictions on the city-level or above) where spatial dependence structures and spatial density vary continuously across space, spatial cross-validation should be preferred for model selection and model assessment, since we believe that, in general, the cost of a slightly too pessimistic perception of predictive accuracy is lower than having spurious confidence in a model's capability to predict unseen data. Overstatement of predictive accuracy may withhold appraisers, underwriters, lenders, as well as portfolio and investment managers from appropriately reflecting the uncertainties associated with appraised values in their decision making and risk management, potentially leading to adverse effects in capital allocation.

Future research in this field may apply model-agnostic interpretation techniques and analyze to what extent identified relationships are spurious when spatial dependence is not accounted for to shed light on the role of spatial autocorrelation on the decision-making of the algorithms.

2.7 Endnotes

1. As modeling choices may differ depending on whether analysis or prediction is the main objective of a study, we concentrate primarily on prediction and do not wish to draw any causal inference or conclusions of the market under investigation.

2. Although the use of asking rents can been criticized since they may deviate from actual contract rents, multiple listing systems (MLS) provide a valuable data source for statistical learning applications due to their high frequency of occurrence and timely availability and have been repeatedly used in the algorithmic hedonic literature (e.g., Chiarazzo et al., 2014; Park and Bae, 2015; Baldominos et al., 2018; Gröbel and Thomschke, 2018; Hu et al., 2019; Pérez-Rave et al., 2019; Pace and Hayunga, 2020; Rico-Juan and Taltavull de La Paz, 2021). Considering vacancy rates for dwellings in Frankfurt well below 1%, we can follow the rationale of Gröbel (2019) and assume that renters are price takers, such that there should be no notable differences between asking and contract rents. Besides that, deviations between asking and contract rents "[…] are not expected to lead to an error bias", especially when hedonic characteristics are controlled for, as stated by Cajias (2018). We hence do not see any reason to question the validity of asking rents, especially in view of the objective of our study.

3. The systematic variation of rent formation processes across space should not introduce bias into cross-validation errors since machine learning algorithms do not assume fixed hedonic pricing coefficients but have the flexibility to differentiate between spatially heterogeneous environments. In this study, spatial heterogeneity is, therefore, put aside and the focus lies on spatial autocorrelation only.

4. Evidently, housing data not only exhibit high levels of spatial but also temporal dependence when pooled across time (Pace et al., 2000). In this study, we leave temporal aspects aside for future research and focus solely on space.

5. One limitation of the selected k-means clustering cross-validation strategy (Brenning, 2012) is that full independence between training and test data can only be achieved if the distance between each pair of training and test observation exceeds the spatial autocorrelation range (Brenning, 2005; Le Rest et al., 2014). This is unlikely to be the case for all observations in our resampling instantiation, especially for data points located at the borders of the spatial clusters. Nonetheless, we believe that the number of observations in each test fold is large enough to counteract structural overfitting, such that the impact on aggregated results should be minor. Following the suggestion of Roberts et al. (2017), we refrain from further reducing the number of k folds in the cross-validation procedure since

this would withhold too much information during training and may introduce unnecessary extrapolation into the models.

6. All analyses and model estimations were executed using the open-source statistical programming language R under the version 4.0.4 (R Core Team, 2021). All machine learning algorithms and resampling techniques were employed using the *mlr3* framework implemented by Lang et al. (2019), which is an ecosystem that facilitates a standardized interface to many existing packages in the R environment. To obtain reproducible results and to ensure that the instantiation of the resampling calls do not vary between the different models, which could distort the results, all outputs were produced with the same random number generator using the *set.seed* function in R.

7. Estimations were executed on a standard 1.80GHz processor with four cores, eight logical processors and eight gigabytes of RAM using a 64-bit Windows operating system. After parallelization, the in-sample estimation of the random forest required between 11 and 20 hours for each of the four estimated resampling alternatives, whereby spatial tuning in the inner resampling loop reduced estimation time by up to 43%. The much more efficient extreme gradient boosting algorithm needed only about 3 to 3.5 hours respectively with the spatial tuner being slightly less time-consuming. During the one quarter ahead prediction, where cross-validation only needs to be performed for hyperparameter tuning in the inner loop, estimation time dropped to 3.5 hours for the bagging algorithm and to approximately 30 minutes for the boosting algorithm. The spatial linear models required less than 30 minutes each and the least squares estimator less than a second. Estimation time was significantly lower for the alternative model specification B for all models.

2.8 Appendix

Table 2.6: OLS Regression Output

Variable	Estimate	Std. Error	t-value	p-value	Significance
	Panel A: OLS Model including Spatial Controls				
(Intercept)	-74.91	5.92	-12.65	0.00	***
Continuous					
Living Area [log]	0.80	0.01	81.98	0.00	***
Age [years]	0.00	0.00	-20.11	0.00	***
Age²	0.00	0.00	19.59	0.00	***
Entry date [monthly]	-0.02	0.01	-2.12	0.03	*
Latitude	1.45	0.12	12.17	0.00	***
Longitude	0.67	0.06	11.68	0.00	***
Discrete					
Rooms	0.04	0.00	10.05	0.00	***
Floor	0.00	0.00	5.29	0.00	***
Binary [1=yes, 0=no]					
Bathtub	-0.04	0.00	-9.78	0.00	***
Refurbished	0.02	0.01	3.03	0.00	**
Built-in kitchen	0.11	0.01	21.65	0.00	***
Balcony	0.02	0.00	3.84	0.00	***
Parking	0.03	0.01	4.88	0.00	***
Elevator	0.04	0.01	7.64	0.00	***
Terrace	0.04	0.01	5.99	0.00	***
Distances					
NUTS centroid [km]	-0.02	0.00	-13.02	0.00	***
Bakery [km]	-0.01	0.01	-1.22	0.22	
Bar [km]	-0.02	0.00	-3.73	0.00	***
Biergarten [km]	-0.05	0.00	-14.71	0.00	***
Café [km]	-0.02	0.01	-2.11	0.03	*
School [km]	-0.03	0.01	-2.35	0.02	*
Supermarket [km]	0.02	0.01	1.20	0.23	
Bus station [km]	-0.04	0.00	-17.16	0.00	***
	Panel B: OLS model excluding Spatial Controls				
(Intercept)	3.12	0.04	83.32	0.00	***
Continuous					
Living Area [log]	0.84	0.01	76.62	0.00	***
Age [years]	0.00	0.00	-18.16	0.00	***
Age²	0.00	0.00	21.45	0.00	***
Entry date [monthly]	-0.02	0.01	-2.25	0.02	*
Discrete					
Rooms	0.03	0.00	6.32	0.00	***
Floor	0.01	0.00	6.43	0.00	***
Binary [1=yes, 0=no]					
Bathtub	-0.05	0.00	-10.78	0.00	***
Refurbished	0.03	0.01	4.53	0.00	***
Built-in kitchen	0.15	0.01	26.57	0.00	***
Balcony	0.02	0.01	2.94	0.00	**
Parking	0.01	0.01	1.52	0.13	
Elevator	0.10	0.01	16.78	0.00	***
Terrace	0.03	0.01	4.06	0.00	***

Notes: This table reports the ordinary least squares (OLS) regression outputs for the model including spatial controls in panel A and the model excluding spatial controls in panel B. The dependent variable is the log(rent), independent variables are listed in the left column accordingly. Significance codes: $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', $p < 0.1$ '.', $p > 0.1$ ' '. Std. Error: standard error. Panel A: F-statistic 2245.00 (p-value: 0.0000), AIC -4450.02, BIC -4276.96. Panel B: F-statistic 2968.00 (p-value: 0.0000), AIC -2592.44, BIC -2488.61.

2.9 References

Allen, M. T., Springer, T. M., & Waller, N. G. (1995). Implicit pricing across residential rental submarkets. *The Journal of Real Estate Finance and Economics*, *11*, 137–151.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Kluwer Academic Publishers.

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis*, *27*(2), 93–115.

Anselin, L., Bera, A. K., Florax, R., & Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, *26*(1), 77–104.

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.

Bahn, V., & McGill, J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, *16*(6), 733–742.

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.

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, *17*(1), 61–85.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

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), 139–160.

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Routledge.

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, *60*(3), 291–319.

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, *5*, 853–862.

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *2012 IEEE International Geoscience and Remote Sensing Symposium*, 5372–5375.

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: Will large data sets change our assumptions? *Journal of Property Investment & Finance*, *36*(1), 32–49.

Cajias, M., Willwersch, J., Lorenz, F., & Schaefers, W. (2021). Rental pricing of residential market and portfolio data – A hedonic machine learning approach. *Real Estate Finance, 38*(1), 1–17.

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

Can, A., & Megbolugbe, I. (1997). Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics*, *14*, 203–222.

Case, B., Clapp, J., Dubin, R., & Rodriguez, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. *The Journal of Real Estate Finance and Economics*, *29*(2), 167–191.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information*, *7*(5), 168–183.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chiarazzo, V., Caggiani, L., Marinelli, M., & Ottomanelli, M. (2014). A neural network based model for real estate price estimation considering environmental quality of property location. *Transportation Research Procedia*, *3*, 810–817.

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.

Cliff, A., & Ord, K. (1973). *Spatial autocorrelation*. Pion.

Cressie, N. A. C. (1993). *Statistics for spatial data* (Revised ed.). John Wiley & Sons, Inc.

Din, A., Hoesli, M., & Bender, A. (2001). Environmental variables and real estate prices. *Urban Studies*, *38*(11), 1989–2000.

Doszyń, M. (2020). Algorithm of real estate mass appraisal with inequality restricted least squares (IRLS) estimation. *Journal of European Real Estate Research*, *13*(2), 161–179.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, *78*(382), 316–331.

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

Füss, R., & Koller, J. A. (2016). The role of spatial and temporal structure for residential rent predictions. *International Journal of Forecasting*, *32*(4), 1352–1368.

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.

Gu, G., & Xu, B. (2017). Housing market hedonic price study based on boosting regression tree. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *21*(6), 1040–1047.

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (2nd ed.). Springer.

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

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

Hong, J., Choi, H., & Kim, W. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategy Property Management, 24*(3), 140–152.

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.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to statistical learning: With applications in R*. Springer.

Kelejian, H. H., & Prucha, I. R. (1998). A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with spatial disturbances. *The Journal of Real Estate Finance and Economics*, *17*(1), 99–121.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International joint Conference on Artificial intelligence*, *2*, 1137–1143.

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.

Lachenbruch, P., & Mickey, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, *10*(1), 1–11.

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.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, *23*, 811–820.

LeSage, J. P. (2014). What regional scientists need to know about spatial econometrics. *The Review of Regional Studies*, *44*(1), 13–32.

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

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2022). Interpretable machine learning for real estate market analysis. *Real Estate Economics*. Forthcoming.

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with R*. CRC Press.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*(3), 531–542.

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.

Meyer, H., Reudenbach, C., Woellauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*.

Militino, A. F., Ugarte, M. D., & García-Reinaldos, L. (2004). Alternative models for describing spatial dependence among dwelling selling prices. *The Journal of Real Estate Finance and Economics*, *29*(2), 193–209.

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. & Gilley, O. W. (1997). Using the spatial configuration of the data to improve estimation. *The Journal of Real Estate Finance and Economics*, *14*(3), 333–340.

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.

Pace, R. K., & LeSage, J. P. (2010). Omitted variable biases of OLS and spatial lag models. In A. Páez, J. Le Gallo, R. N. Buliung, & S. Dall'erba (Eds.), *Progress in Spatial Analysis: Methods and Applications* (1st ed., pp. 17–28). Springer.

Pace, R. K., Barry, R., Gilley, O. W., & Sirmans, C. F. (2000). A method for spatial–temporal forecasting with an application to real estate prices. *International Journal of Forecasting*, *16*(2), 229–246.

Park, B., & and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928–2934.

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.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, *31*(2), 147–164.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, *31*(10), 2001–2019.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications*, *171*.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schroeder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. *Ecography*, *40*(8), 913–929.

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

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120.

Simon, R. (2007). Resampling strategies for model assessment and selection. In W. Dubitzky, M. Granzow, & D. P. Berrar (Eds.), *Fundamentals of data mining in genomics and proteomics* (1st ed., pp. 173–186). Springer.

Sirmans, G. S., & Benjamin, J. D. (1991). Determinants of market rent. *Journal of Real Estate Research*, *6*(3), 357–379.

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

Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics*, *19*(4), 415–428.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–147.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234–240.

Trachsel, M., & Telford, R. J. (2016). Technical note: Estimating unbiased transfer-function performances in spatially structured environments, *Climate of the Past*, *12*, 1215–1223.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *10*(2), 225–232.

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

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(91).

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Cengage Learning.

Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research*, *10*(2), 185–201.

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.

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.

3 Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach

3.1 Abstract

In this article, we examine the accuracy and bias of market valuations in the U.S. commercial real estate sector using properties included in the NCREIF Property Index (NPI) between 1997 and 2021 and assess the potential of machine learning algorithms (i.e., boosting trees) to shrink the deviations between market values and subsequent transaction prices. Under consideration of 50 covariates, we find that these deviations exhibit structured variation that boosting trees can capture and further explain, thereby increasing appraisal accuracy and eliminating structural bias. The understanding of the models is greatest for apartments and industrial properties, followed by office and retail buildings. This study is the first in the literature to extend the application of machine learning in the context of property pricing and valuation from residential use types and commercial multifamily to office, retail, and industrial assets. In addition, this article contributes to the existing literature by providing an indication of the room for improvement in state-of-the-art valuation practices in the U.S. commercial real estate sector that can be exploited by using the guidance of supervised machine learning methods. The contributions of this study are, thus, timely and important to many parties in the real estate sector, including authorities, banks, insurers and pension and sovereign wealth funds.

Keywords: Commercial real estate, Appraisal, Interpretable machine learning

Acknowledgments: The authors sincerely thank the National Council of Real Estate Investment Fiduciaries (NCREIF), and in particular Professor Jeffrey Fisher (PhD), for their support and provision of the data.

3.2 Introduction

Both institutional and private investors aim to diversify their portfolios with real estate. A significant share of this is accounted for by investments in commercial real estate sectors, which amount to around $32 trillion globally. The heterogeneity of commercial real estate contributes well to diversification, but it is also accompanied by characteristics such as illiquidity, opacity and unwieldiness that make it difficult to thoroughly understand market dynamics. Consequently, the valuation of commercial properties involves a great deal of effort that justifies an appraisal industry worth billions of dollars. Studies have repeatedly demonstrated that commercial property appraisals do not always adequately represent market dynamics and can differ significantly from actual sales prices (e.g., Cole et al., 1986; Webb, 1994; Matysiak and Wang, 1995; Fisher et al., 1999; Edelstein and Quan, 2006; Cannon and Cole, 2011). Despite the increasing complexity of pricing processes and more rapidly changing markets, the principal methods used by the valuation industry have largely remained unchanged for the past decades. However, this is slowly changing with an increasing availability of data and the emergence of artificial intelligence fostering the use of innovative technologies in the real estate sector.

In recent years, machine learning algorithms have been increasingly considered as a suitable method for the estimation of house prices and rents, with a large corpus of literature pointing to their high accuracy in the residential sector (e.g., Mullainathan and Spiess, 2017; Mayer et al., 2019; Bogin and Shui, 2020; Hong et al., 2020; Pace and Hayunga, 2020; Lorenz et al., 2022; and Deppner and Cajias, 2022). In the commercial sector, on the other hand, the scope of analysis has thus far been limited to multifamily assets and shows inconsistent results in terms of estimation accuracy (Kok et al., 2017). One prerequisite for machine learning methods to provide accurate and reliable property value estimates is the availability of substantial amounts of data with uniform property characteristics. While these criteria are largely met for residential real estate where property characteristics are considered relatively homogeneous, and data is widely accessible on multiple listing services, the nature of commercial real estate is more complex and heterogenous, and infrequent transactions and market opaqueness continue to hinder data availability. Despite the enormous potential for the sector, this poses a challenge for the application of data-driven valuation methods in commercial real estate and raises the question to what extent machine learning algorithms can provide significant improvement to the industry's state-of-the-art appraisal practices. To the best of our knowledge, there is no research in the current literature that investigates the usefulness of machine learning

algorithms for the valuation of commercial properties other than multifamily buildings (see Kok et al., 2017).

This article contributes to this field using 24 years of property-level transaction data of commercial real estate from the NCREIF Property Index (NPI) provided by the National Council of Real Estate Investment Fiduciaries (NCREIF). In a first step, we investigate the deviation between actual sales prices observed in the market and the appraised values before sale to assess the accuracy and bias associated with state-of-the-art valuation methods that were last examined by Cannon and Cole (2011). Given the findings of inaccuracy and structural bias of appraisals that the literature has reported over the past decades, we hypothesize that the observed deviations between sales prices and appraisal values exhibit structured information content that machine learning models can exploit to further explain and shrink these residuals, thereby providing a superior *ex-post* understanding of market dynamics. This is examined using a tree-based boosting algorithm, measuring how much of the variation in the residuals can be explained. While Pace and Hayunga (2020) follow a similar approach to benchmark machine learning methods against spatial hedonic tools in a residential context, no research empirically quantifies the potential of complementing traditional appraisal methods with data-driven machine learning techniques, neither in residential nor commercial sectors. Lastly, we apply model-agnostic permutation feature importance to reveal where improvements originate and point to price determinants that are not adequately reflected in current appraisal methods.

From a practical point of view, the application of machine learning can add to an enhanced *ex-ante* understanding of pricing processes that may support valuers in the industry and contribute to more dependable valuations in the future. By illustrating the potential and pointing to the shortcomings of these methods, we aim to provide guidance, stimulate the critical discussion, and motivate further research on machine learning approaches in the context of commercial real estate valuation.

3.3 Related Literature

The estimation of market values is the primary concern of most real estate appraisal assignments. According to federal financial institutions in the U.S., the market value is defined as:

> *"[…] the most probable price which a property should bring in a competitive and open market under all conditions requisite to a fair sale, the buyer and*

> *seller each acting prudently and knowledgeably, and assuming the price is not affected by undue stimulus"*[1] (Real Estate Lending and Appraisals, 2022).

However, the accurate and timely estimation of commercial property prices is a complex task, as direct real estate markets are characterized by high heterogeneity, illiquidity, and information asymmetries that are accompanied by high search and transaction costs. Over the past decades, many methods have been developed and refined to arrive at the most probable transaction price of a property in the market. Pagourtzi et al. (2003) distinguish between traditional (i.e., manual) and advanced (i.e., statistical) valuation approaches.

3.3.1 Traditional Valuation Methods

Traditional valuation models are characterized by a procedural approach (Mullainathan and Spiess, 2017) that follows pre-defined economic rules. These procedures can be thought of as "prediction rules" used to obtain appraised values of commercial real estate. The most common procedures in current appraisal practices are the *income approach*, the *sales-comparison approach,* and the *cost approach* as described by Fisher and Martin (2004) and Mooya (2016).

As the industry´s preferred approach to commercial property valuation, the *income approach* is based on the idea that the value of a property depends on the present value of its future cash flows, and is thus determined by two main factors: the net operating income and the capitalization rate. The latter incorporates all risks and upside potentials of the income-producing property. However, the correct assessment of the capitalization rate is not straightforward and depends on many assumptions. Hence, comparable transactions of similar properties observed in the market are often used as a point of reference. This is known as the *sales-comparison approach* and is based on the rationale that the value of a property should equal the value of a similar property with the same characteristics. Mooya (2016) finds this approach to be the most valid indicator of market conditions as new market valuations are based on recently transacted properties. However, comparable sales are scarce or outdated in very illiquid property sectors and markets. In such cases, the *cost approach* can be used following the principle that an informed investor would pay no more than for the substitute building as this would

[1] Implicit in this definition is the consummation of a sale as of a specified date and the passing of title from seller to buyer under conditions whereby:
(1) Buyer and seller are typically motivated;
(2) Both parties are well informed or well advised, and acting in what they consider their own best interests;
(3) A reasonable time is allowed for exposure in the open market;
(4) Payment is made in terms of cash in U.S. dollars or in terms of financial arrangements comparable thereto; and (5) The price represents the normal consideration for the property sold unaffected by special or creative financing or sales concessions granted by anyone associated with the sale.
12 C.F.R. § 34.42 (2022).

constitute an arbitrage opportunity. The market value of a property is thus derived from the cost of constructing a similar property including the land value and adjusting for physical and functional depreciation.

All these procedures have an economic justification and have served the industry well for decades; however, as prediction rules, they also suffer from certain limitations. For instance, the determination of the capitalization rate is subject to the discretionary scope and the assumptions (i.e., the assessment of risks and upside potentials, e.g., growth hypothesis versus risk hypothesis for vacant space in Beracha et al., 2019) of the individual executing them to arrive at a market value. In turn, capitalization rates derived from comparable sales may capture recent market dynamics but are inherently backwards looking such that appraisals may significantly lag. Furthermore, the availability of similar properties that have been sold recently is a limiting factor due to infrequent transactions and high heterogeneity. This requires adjustments, which again depend on subjective opinions of value, resulting in imprecise estimations. On the other hand, the cost approach can indicate a property's substitute value, but also allows a lot of room for subjectivity given the uniqueness of each property and the numerous assumptions to be made for adjustments and depreciation. Pagourtzi et al. (2003) note that "[…] price will be determined not by cost, but by the supply and demand characteristics of the occupational market" in case of scarcity, which is a typical characteristic of many real estate markets due to geographic constraints and building regulations. In addition, Matysiak and Wang (1995) raise the hypothesis that not all available data is considered at the time of valuation. While each of the approaches mentioned above is limited to a certain set of information, market intransparency may furthermore impose restrictions to the data that is available to individual appraisers.

Cole et al. (1986) are the first in the literature to document the differences between real estate appraisals and sales prices in the U.S. commercial real estate market. The authors examine properties sold out of the NCREIF Property Index (NPI) between 1978 and 1984 and find a mean absolute percentage difference of around 9% in that period of rising markets. In a similar study, Webb (1994) extends the sample of Cole et al. (1986) by updating the period from 1978 to 1992, thereby covering different price regimes of rising, stagnating, and falling markets. The author finds that the highest deviations occur during rising markets averaging 13%, declining to 10% during flat markets and 7% during falling markets. Fisher et al. (1999) update the studies of Cole et al. (1986) and Webb (1994) on the reliability of commercial real estate appraisals in the U.S. and show that from 1978 to 1998, manual appraisals of NPI properties across multiple asset types deviate on average between 9% and 12.5% from actual sales prices. This is in line with the findings of Cannon

and Cole (2011) who analyzed NPI sales data from 1984 to 2009 and observed deviations ranging between 11% and 13.5% over the entire sample period for the different asset sectors. The authors find appraisals to consistently lag actual sales prices, falling short of sales prices in bullish markets and remaining in excess of sales prices in bearish markets. With respect to mean percentage errors, the findings of Cannon and Cole (2011) confirm the hypothesis of Matysiak and Wang (1995), suggesting that appraisal errors do not solely arise due to the time differences but also due to a systematic valuation bias. Kok et al. (2017) take another look at appraisal errors in commercial real estate markets and propose the use of advanced statistical techniques to reduce the deviations found in the previous studies.

3.3.2 Advanced Valuation Methods

With an increasing data availability in real estate markets and the development of econometric and statistical techniques, researchers have started to tackle existing tasks empirically instead of procedurally (Mullainathan and Spiess, 2017). While a wide range of empirical methods exists in the current literature, we focus on the most discussed approaches for property valuation, that is *hedonic pricing* and *machine learning*.

The *hedonic pricing* model dates to Rosen (1974) who defines the value of a heterogenous good as the sum of the implicit prices of its objectively measurable characteristics. The most common econometric approach used to derive such implicit prices is multiple linear regression or extensions thereof. In commercial real estate markets, hedonic pricing models have been applied to disentangle price formation processes from an econometric point of view (e.g., Clapp, 1980; Brennan et al., 1984; Glascock et al., 1990; Mills, 1992; Malpezzi, 2002; Sirmans et al., 2005; Koppels and Soeter, 2006; Nappi-Choulet et al., 2007; Seo et al., 2019). Hedonic models have proven useful in understanding price determinants in real estate markets, but researchers have also pointed to the limitations of the underlying methods such as their imposed linearity and fixed parameters, which cannot be assumed to hold in reality (Dunse and Jones, 1998; Bourassa et al., 2010; Osland, 2010). Although these models are efficient in generating predictions and easy to interpret, their strong assumptions and need for manual specification carry the risk of bias, subjectivity, and inconsistency, which is to be eliminated in the first place.

In contrast to linear hedonic approaches, algorithmic *machine learning* models follow a purely data-driven approach and make use of stochastic rules to find the best possible model fit. Over the past decades, many algorithms such as artificial neural networks (Rumelhart et al., 1986), support vector regression (Smola and Schölkopf, 2004), and bagging and boosting algorithms (i.e., random forest regression by Breiman, 1996, 2001;

and gradient tree boosting by Friedman, 2001) that are based on ensembles of regression trees (Breiman et al., 1984) have been developed and refined. These algorithms can autonomously learn nonlinear relationships from the data without specifying them *a-priori* or making any implicit assumptions of the relationship between the property's price and its features. This means that the models consider all available information at the time of valuation and identify complex relationships based on patterns in the data. Since the training process of machine learning algorithms is computationally expensive compared to traditional econometric models, it took until this decade for technological progress to enable sufficient computational capacity for the widespread application of such techniques.

In recent years, a large corpus of literature has demonstrated the potential of machine learning algorithms to accurately estimate prices and rents of houses and apartments in the residential sector. This includes studies by McCluskey et al. (2013) for artificial neural networks, Lam et al. (2009), Kontrimas and Verikas (2011), and Pai and Wang (2020) for support vector regression, Levantesi and Piscopo (2020) for random forest regression and van Wezel et al. (2005) and Sing et al. (2021) for gradient tree boosting algorithms. In many comparative studies that document the accuracy of a broader range of model alternatives, tree-based methods and, in particular boosting and bagging algorithms, have shown superiority over other methods (e.g., Zurada et al., 2011; Antipov and Pokryshevskaya, 2012; Mullainathan and Spiess, 2017; Baldominos et al., 2018; Hu et al., 2019; Mayer et al., 2019; Bogin and Shui, 2020; Pace and Hayunga, 2020; Cajias et al., 2021; Rico-Juan and Taltavull de La Paz, 2022; Lorenz et al., 2022; and Deppner and Cajias, 2022).

In academia and the industry, however, high demands are placed not only on accuracy and consistency, but also on reliability and comprehensibility of the models. Hence, machine learning methods have been criticized for lacking an economic justification and having a black-box character (McCluskey et al., 2013; Mayer et al., 2019). Valier (2020) argues that although data-driven machine learning models might produce equivalent or even better results than traditional methods, too much variability comes with the flexibility of these methods as they rely entirely on the input data and can change quickly. This makes them "[…] difficult to use for public policies, where the evaluation process must guarantee fairness of treatment for all the cases concerned and maintain the same efficiency over time," as stated by Valier (2020). While Pérez- Rave et al. (2019) and Pace and Hayunga (2020) suggest to maintain interpretability by enhancing linear models with insights generated by machine learning techniques, Rico- Juan and Taltavull de La Paz (2022) and

Lorenz et al. (2022) apply model-agnostic interpretation techniques that allow *ex-post* interpretability of the models to circumvent this problem.

Besides their sensitivity to changes in the data, the methods can quickly overfit the training sample if applied without the necessary prudence and may thus not represent the true relationship between the dependent variable and its regressors. This is especially problematic when training data is scarce. For this reason, machine learning algorithms require a reasonable number of observations of previous transactions and attributes that adequately describe the respective properties to provide dependable and stable estimations of property values. Hence, research in this field has largely focused on the residential sector, where properties are considered relatively homogeneous, and data availability has increased exponentially over the last years with the transition from offline real estate offers to online multiple listing services. In turn, the high heterogeneity and data scarcity in commercial real estate markets imposes challenges for the application of machine learning techniques. Kok et al. (2017) are the first in the literature to apply machine learning methods to estimate prices of commercial multifamily properties. The authors benchmark tree-based boosting and bagging algorithms against a linear hedonic model across different model specifications and find mixed results in terms of their accuracy. While two different types of boosting provide error reduction in all cases tested, the bagging algorithm does not offer any significant improvement and is even outperformed by the ordinary least squares estimator in one case. To the best of our knowledge, there is no research on the predictive performance of machine learning methods for other property types in commercial real estate.

Although institutionally held multifamily properties are of residential use, the study of Kok et al. (2017) indicates that previous findings of the accuracy of machine learning algorithms in the residential sector cannot be easily transferred to a commercial real estate context, given the known limitations of these techniques and the peculiarities of the sector as discussed earlier. This raises the question to which extent algorithmic approaches can learn market dynamics in commercial real estate to generate insights into pricing processes that go beyond the understanding achieved with traditional valuation approaches, thus providing potential improvement to the state-of-the-art.

3.4 Data and Methodology

The principal dataset used for this study was provided by the National Council of Real Estate Investment Fiduciaries (NCREIF). It contains quarterly observations of all properties

included in the NCREIF Property Index[2] (NPI) on the asset level spanning 1Q 1978 through 1Q 2021. To be included in the NPI, a property must be

i. an operating apartment, hotel, industrial, office, or retail property,
ii. acquired, at least in part, by tax-exempt institutional investors and held in a fiduciary environment[3],
iii. accounted for in compliance with the NCREIF Market Value Accounting Policy[4],
iv. appraised – either internally or externally – at a minimum every quarter.

A qualifying property is included in the NPI upon purchase and removed again upon sale. The database contains all quarter-observations over that property's holding period, terminating with the sale quarter. For reasons of data scarcity in earlier years and in specific sectors, we limit the initial sample to 24 years from 1Q 1997 through 1Q 2021, including all asset sectors except for hotels. This is generally equivalent to the dataset in the study of Cannon and Cole (2011), with the time span shifted 12 years ahead.

3.4.1 Data Pre-processing

We filter all properties that had been sold during that period, excluding partial sales and transfers of ownership. This constitutes a sample of 12,956 individual assets for which we observe the net sale prices, the corresponding appraisal values and a series of structural, physical, financial, and spatial attributes recorded quarterly.

After examining the most recent appraisal values of the sold properties from the quarter before the sale, we find that the appraised value equals the net sale price in 6,091 cases, which corresponds to 47% of the entire sample. This is consistent with Cannon and Cole (2011) and indicates that the sale price for those properties was determined at least three months before a pending transaction. Since this price was used as the market value instead of an independent appraisal, we are forced to use the appraisal values of the second quarter before the sale to represent the properties' most recent market value. However, we still observe 587 properties where the market value equals the sale price and another 179 properties with missing data for that quarter, resulting in a reduced sample of 12,190 properties for which we have data on the sale prices and the market values. One possibility to account for the time lag between the appraisal date and the sale date is to roll back the sale prices as Cannon and Cole (2011) did for some properties in their sample. However, the authors find that overall, the unadjusted differences are, in fact, better measures of

[2] The NPI is a quarterly index tracking the performance of core institutional property markets in the U.S.
[3] This includes commingled real estate funds (open and closed-end), separate accounts, individual accounts, private REITs, REOCs, and joint-venture partnerships.
[4] For further details, refer to the NCREIF PREA Reporting Standards at www.reisus.org.

appraisal accuracy. This is no surprise as transaction prices are often determined three to six months before closing, known as due diligence lag. We subsequently do not adjust for the time lag between appraisal and sale date but control for moving markets in that period.

Missing and erroneous data points of the relevant variables are accounted for as follows. We remove observations with square footage and construction years reported as less than or equal to zero. Likewise, occupancy rates less than zero or higher than one were also regarded as erroneous data points. Furthermore, we omit observations with missing values for the square footage, the property subtype, the construction year, the occupancy rate, the appraisal type, the fund type, the metropolitan statistical area (MSA) code, the net operating income (NOI), and the capital expenditures (Capex), which represent the main explanatory variables collected from the raw, principal dataset. We further remove observations where the deviation between the sale price and the appraisal value two quarters before the sale is abnormally high, as this indicates a potential data error.[5] We also remove extreme outliers in the sale price, the building area and the sale price per square foot by cropping the upper and lower tails of the distributions.[6] After cleaning erroneous and missing data, the sample was reduced to 8,427 individual properties.

In addition, we enrich the initial data with a set of new variables. To better control for building quality, we calculate the building age as the difference between the year of sale and the construction date trimmed at 100 years[7] and the cumulative sum of a property's capital expenditures, that is the sum of all capital expenditures for building extensions and building improvements over the holding period.[8] Since we observe that NOIs tend to fluctuate materially in the quarters before sale, we also calculate the mean of the properties' annual NOIs over their holding period as a proxy for stabilized income. This measure incorporates different market cycles and is less prone to speculation, which may better capture a property's intrinsic value. As demonstrated repeatedly in the literature, the spatial dimension is an important driver of real estate prices. The dataset provides the location zones of a property on the ZIP code level. However, we cannot ensure enough observations for each ZIP code area in our sample, so we use the MSA level instead. That

[5] When we calculate the mean absolute percentage errors for the second quarter before sale, we observe market values that deviate from sale prices by up to 377%. We crop the distribution of percentage errors at the 99th percentile, thus allowing for deviations by up to 60%.

[6] After data cleaning, we observe sale prices per square foot between $0.8 and $915,501.1 indicating potential data errors. To keep data loss at a minimum, we crop the distributions at the lower 0.5th and the upper 99.5th percentiles.

[7] The sample includes 61 observations for which the building age takes values between 101 and 157 years, most of which are unique. We assign those observations the value 100, thus effectively creating a partition for buildings that are older than 100 years, so the trees cannot overfit single observations by using unique building ages.

[8] This excludes tenant improvements, lease commissions, and additional acquisition costs, which are incentives or fees that do not affect the quality of a property.

said, location dummies on the MSA level may capture global price differentials across space, but they are not adequate to efficiently reflect complex pricing behaviors driven by spatial considerations of buyers and sellers. To better assess appraisers' understanding of space, we geocode our sample observations using the property addresses. With the Google Places API, we managed to geocode 93%[9] of the addresses and retrieve the distances to relevant points of interest (POIs). This includes transport linkages and amenities that may produce spillover effects and thus cause positive or negative externalities to their neighborhood. For example, an office building might benefit from the proximity to a café, a gym or a laundry that serves white-collar workers, which translates into a location premium. Lastly, we omit MSA codes that include less than ten properties of the same asset class to counteract overfitting on the location dummies.

Our final sample contains 7,133 individual properties[10] that meet all the previously outlined criteria to be included in the study. Relative to the initial sample size this constitutes a

Table 3.1: Observations per Year

	All Types (N = 7,133)		**Apartment (N = 1,904)**		**Industrial (N = 2,337)**		**Office (N = 2,056)**		**Retail (N = 836)**	
Variable	n	Percent	n	Percent	n	Percent	n	Percent	n	Percent
Year										
... 1997	68	0.95%	17	0.89%	31	1.33%	9	0.44%	11	1.32%
... 1998	84	1.18%	12	0.63%	26	1.11%	31	1.51%	15	1.79%
... 1999	94	1.32%	18	0.95%	18	0.77%	31	1.51%	27	3.23%
... 2000	201	2.82%	51	2.68%	49	2.10%	74	3.60%	27	3.23%
... 2001	174	2.44%	53	2.78%	50	2.14%	42	2.04%	29	3.47%
... 2002	187	2.62%	49	2.57%	63	2.70%	51	2.48%	24	2.87%
... 2003	251	3.52%	60	3.15%	78	3.34%	80	3.89%	33	3.95%
... 2004	337	4.72%	74	3.89%	117	5.01%	107	5.20%	39	4.67%
... 2005	472	6.62%	109	5.72%	135	5.78%	132	6.42%	96	11.48%
... 2006	298	4.18%	75	3.94%	84	3.59%	115	5.59%	24	2.87%
... 2007	381	5.34%	91	4.78%	139	5.95%	124	6.03%	27	3.23%
... 2008	155	2.17%	42	2.21%	54	2.31%	53	2.58%	6	0.72%
... 2009	160	2.24%	57	2.99%	54	2.31%	40	1.95%	9	1.08%
... 2010	182	2.55%	66	3.47%	56	2.40%	40	1.95%	20	2.39%
... 2011	252	3.53%	68	3.57%	87	3.72%	50	2.43%	47	5.62%
... 2012	415	5.82%	112	5.88%	162	6.93%	100	4.86%	41	4.90%
... 2013	500	7.01%	149	7.83%	160	6.85%	122	5.93%	69	8.25%
... 2014	502	7.04%	112	5.88%	194	8.30%	137	6.66%	59	7.06%
... 2015	440	6.17%	130	6.83%	135	5.78%	126	6.13%	49	5.86%
... 2016	512	7.18%	154	8.09%	162	6.93%	146	7.10%	50	5.98%
... 2017	422	5.92%	126	6.62%	136	5.82%	123	5.98%	37	4.43%
... 2018	345	4.84%	119	6.25%	71	3.04%	140	6.81%	15	1.79%
... 2019	427	5.99%	90	4.73%	181	7.74%	110	5.35%	46	5.50%
... 2020	209	2.93%	60	3.15%	57	2.44%	59	2.87%	33	3.95%
... 2021	65	0.91%	10	0.53%	38	1.63%	14	0.68%	3	0.36%

Notes: This table presents the distribution of observations across the sample period from 1Q 1997 through 1Q 2021.

[9] The remaining 7% result mainly from missing or incomplete addresses.
[10] Of which 1,904 are apartments, 2,337 are industrial, 2,056 are office and 836 are retail.

heavy data loss, which again emphasizes the problem of data availability as mentioned earlier.[11] Table 3.1 provides an overview of the number of observations across the sample period.

We further follow Cannon and Cole (2011) in collecting macroeconomic data to control for structural differences in property prices across time. That includes the four-quarter percentage change in employment at the county-level sourced from the U.S. Bureau of Labor Statistics, the four-quarter percentage change in the gross domestic product (GDP) and the ten-year government bond yield sourced from the database of the Federal Reserve Bank of St. Louis, and the four-quarter percentage change in construction costs by region sourced from the U.S. Census Bureau. We further collect quarterly NPI data by property type, that is, the quarterly change in market value cap rates, vacancy rates, NOI growth rates and the quarterly number of sales of NPI properties. While all these variables capture the period between the sale date and the first quarter before sale, we also provide the lags of all macroeconomic and NPI index data for the period between the first and the second quarter prior to sale to control for the time lag between the appraisal and the sales date.

3.4.2 Appraisal Error

NCREIF follows the definition of market value as stated in the "Related Literature" section and adopted by the Appraisal Foundation as well as by the Appraisal Institute. According to this definition, the market value of a property represents the best estimate of a transaction price in the current market. Consequently, we assess the manual appraisals as predictions of sales prices by examining the mean absolute percentage error (MAPE) and the mean percentage error (MPE) as calculated in equations (3.1) and (3.2), respectively.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Sale\ Price_{i,t0} - Appraised\ Value_{i,t-2}}{Appraised\ Value_{i,t-2}}\right| \tag{3.1}$$

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{Sale\ Price_{i,t0} - Appraised\ Value_{i,t-2}}{Appraised\ Value_{i,t-2}} \tag{3.2}$$

The MAPE is used as a measure of accuracy, whereas the MPE can be understood as a measure of biasedness. That is, the appraised value is considered an unbiased predictor of sales prices, if the MPE is not significantly different from zero. This is examined using t-test statistics.

[11] In a similar study by Cannon and Cole (2011), the authors start with 9,439 properties for a period of 25 years and, after filtering, end up with a sample of 7,214 sales. The relative data loss is higher in our case, as we use substantially more covariates with missing entries that result in data leakage.

The vector of appraisal errors Y used as the dependent variable in our models is calculated as the difference between the vector of the log sale price per square foot (SP) and the vector of the log appraisal (market) value per square foot (MV). This is stated in equation (3.3), which corresponds to the log of the percentage appraisal error, however, keeping the signs.

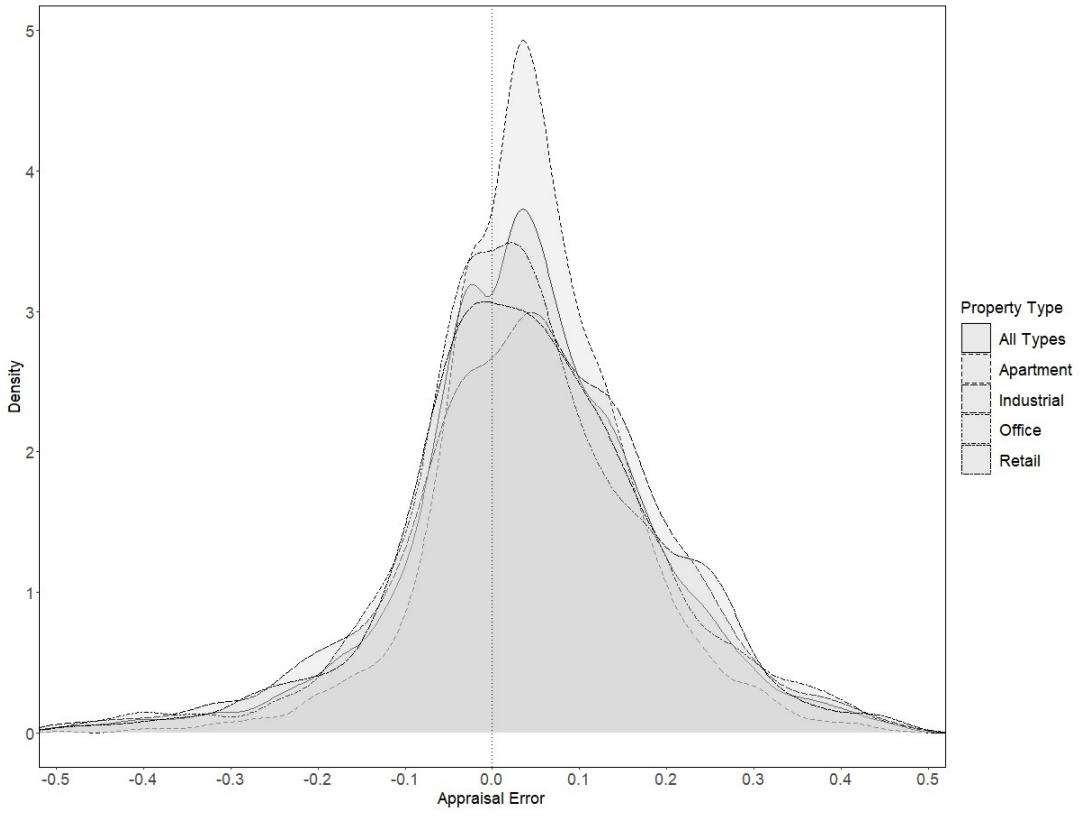
$$Y = [\, SP - MV \,] \tag{3.3}$$

$$SP = \log\left(\frac{Sale\ Price_{t0}}{SqFt}\right)$$

$$MV = \log\left(\frac{Appraised\ Value_{t-2}}{SqFt}\right)$$

Figure 3.1 depicts the distribution of the dependent variable for the different property types. We expect systematic differences between appraisal errors of the four property types, so we conduct an analysis of variance (ANOVA) test with the null hypothesis that there is no significant difference in the sample means of the respective groupings. The

Figure 3.1: Distribution of Appraisal Errors



Notes: The density plot shows the distribution of the raw residuals (appraisal errors) for all property types and for each property type individually. The dotted horizontal line marks the null point on the x-axis.

ANOVA test rejects the null at the 1% level of significance, indicating systematic differences in the sample distributions of the four asset sectors.

3.4.3 Explanatory Variables

Matysiak and Wang (1995) state that appraisal errors are generally rooted in two components. First, markets can change between the appraisal date and the sale date and second, a pure valuation error (i.e., bias) can be incorporated. The latter could be ruled out if the mean percentage error approaches zero, as positive and negative deviations should cancel out. If this is not the case, appraisal errors are unlikely to be entirely random, implying that some information content is left to be explained. To capture the two components from which deviations between appraised values and sales prices originate according to Matysiak and Wang (1995), we include a wide range of explanatory variables in our models.

The first component *a* refers to the time difference between the appraisal and transaction dates. That is, an appraisal error occurs due to a changing market environment during that period. To control for moving markets, we include the market indicators M_{t0} and M_{t-1} from the NPI data (i.e., the quarterly change in market value cap rates, vacancy rates, NOI growth rates and the quarterly number of sales of NPI properties as a proxy for market liquidity) for both quarters before sale as well as the continuous transaction year as temporal indicator T. However, a change in the value of a property could also result from a change in the property fundamentals. Although cash flows from the quarters before sale are backward-looking, and property values are inherently determined by future cash flows that can be estimated with existing lease contracts and maintenance plans, we control for the occurrence of unexpected events (such as rent defaults or repairs) by including the cash flows C_{t0}, C_{t-1} (that is the NOI and Capex) for both quarters before sale. The first component *a* of regressors can be specified in matrix notation as in equation (3.4).

$$X_a = [\, M_{t0}\, M_{t-1}\, T\, C_{t0}\;\; C_{t-1}\,] \tag{3.4}$$

The second component *b* refers to the pure valuation bias and can have various causes such as subjective opinions of value, varying risk appetite and assumptions of funds and individual appraisers or appraisal smoothing. To capture these effects, we include several structural (S), physical (P), financial (F), and locational (i.e., spatial) (L) property characteristics as well as economic (E) indicators for both quarters before sale, as specified in equation (3.5). This includes the fund type and the type of appraisal and the building occupancy for S, the property subtype, the building area, and the building age for P, the stabilized NOI and the cumulative sum of Capex for F, the MSA, latitude, longitude and

distances to 18 POIs for L, as well as the four-quarter percentage change in employment on the county-level, the four-quarter percentage change in the GDP, the 10-year government bond yield, and the four-quarter percentage change in construction costs by region in both quarters prior to sale, corresponding to E_{t0} and E_{t-1} respectively.

Table 3.2: Descriptive Statistics of Numerical Variables

			All Types (N = 7,133)				
	Variable	Unit	Mean	Median	Sd	Min	Max
[T]	Year	[Years]	2010.74	2012.00	6.18	1997.00	2021.00
[P]	SqFt	[k]	273.43	203.29	283.02	2.25	5,995.50
	Building Age	[Years]	22.68	19.00	16.23	-1.00	100.00
[S]	Occupancy	[%]	0.91	0.95	0.15	0.00	1.00
[F]	CapEx Cumulative	[$/SqFt]	14.45	3.36	188.43	0.00	15,518.44
	Stabilized NOI	[$/SqFt]	8.21	6.70	5.75	0.01	45.54
[C_{t0}]	CapEx	[$/SqFt]	0.72	0.04	2.86	0.00	77.85
	NOI	[$/SqFt]	1.32	0.92	2.07	-53.10	46.73
[C_{t-1}]	CapEx (lag)	[$/SqFt]	0.76	0.16	2.45	0.00	58.59
	NOI (lag)	[$/SqFt]	2.35	1.83	2.16	-8.55	31.79
[L]	Longitude	[°]	-95.46	-93.27	17.19	-122.93	-70.49
	Latitude	[°]	36.69	37.38	5.21	25.60	47.94
	Bank	[km]	0.75	0.52	0.77	0.00	6.49
	Bar	[km]	0.73	0.51	0.69	0.00	5.86
	Cafe	[km]	0.59	0.42	0.59	0.00	5.18
	Convenience Store	[km]	0.66	0.53	0.54	0.00	5.91
	Department Store	[km]	1.92	1.39	1.87	0.00	8.68
	Doctor	[km]	0.37	0.23	0.44	0.00	6.65
	Gas Station	[km]	0.73	0.61	0.54	0.00	5.59
	Gym	[km]	0.62	0.43	0.62	0.00	5.85
	Laundry	[km]	0.71	0.53	0.65	0.00	5.92
	Lawyer	[km]	0.58	0.35	0.71	0.00	6.28
	Park	[km]	0.70	0.57	0.56	0.00	6.31
	Parking	[km]	0.82	0.56	0.88	0.00	8.48
	Pharmacy	[km]	0.71	0.51	0.68	0.00	6.48
	Restaurant	[km]	0.36	0.24	0.39	0.00	3.78
	School	[km]	0.43	0.32	0.40	0.00	4.20
	Shopping Mall	[km]	0.87	0.63	0.84	0.00	7.19
	Supermarket	[km]	1.37	1.02	1.30	0.00	8.66
	Public Transport	[km]	2.02	1.33	2.15	0.00	8.68
[E_{t0}]	GDP yoy	[%]	0.02	0.02	0.01	-0.09	0.05
	Bond Yield	[%]	0.03	0.03	0.01	0.01	0.07
	Construction Cost yoy	[%]	0.04	0.04	0.04	-0.10	0.20
	Employment yoy	[%]	0.02	0.02	0.03	-0.18	0.27
[E_{t-1}]	GDP yoy (lag)	[%]	0.02	0.02	0.02	-0.09	0.05
	Bond Yield (lag)	[%]	0.03	0.03	0.01	0.01	0.07
	Construction Cost yoy (lag)	[%]	0.04	0.04	0.04	-0.10	0.13
	Employment yoy (lag)	[%]	0.02	0.02	0.03	-0.20	0.26
[M_{t0}]	Cap Rate qoq	[%]	0.00	0.00	0.00	0.00	0.00
	Vacancy qoq	[%]	0.00	0.00	0.01	-0.03	0.03
	NOI Growth qoq	[%]	0.03	0.04	0.05	-0.33	0.18
	Sold Properties	[#]	617.46	665.00	178.90	182.00	907.00
[M_{t-1}]	Cap Rate qoq (lag)	[%]	0.00	0.00	0.00	0.00	0.00
	Vacancy qoq (lag)	[%]	0.00	0.00	0.01	-0.03	0.03
	NOI Growth qoq (lag)	[%]	0.03	0.04	0.05	-0.33	0.18
	Sold Properties (lag)	[#]	610.17	662.00	181.68	182.00	907.00

Notes: This table presents the summary statistics of numerical features.

Table 3.3: Descriptive Statistics of Categorical Variables

		All Types (N = 7,133)	
	Variable	n	Percent
[P]	**Property Type**		
	... Apartment	1,904	26.69%
	... Industrial	2,337	32.76%
	... Office	2,056	28.82%
	... Retail	836	11.72%
	Property Subtype		
	... Garden	1,295	18.16%
	... High-rise	455	6.38%
	... Low-rise	154	2.16%
	... Research and Development	120	1.68%
	... Flex Space	412	5.78%
	... Manufacturing	21	0.29%
	... Other	40	0.56%
	... Office Showroom	11	0.15%
	... Warehouse	1,733	24.30%
	... Central Business District	450	6.31%
	... Suburban	1,606	22.52%
	... Community Center	265	3.72%
	... Theme/Festival Center	1	0.01%
	... Fashion/Specialty Center	30	0.42%
	... Neighborhood Center	363	5.09%
	... Outlet Center	2	0.03%
	... Power Center	74	1.04%
	... Regional Mall	34	0.48%
	... Super-Regional Mall	22	0.31%
	... Single-Tenant	45	0.63%
[S]	**Appraisal**		
	... External	2,485	34.84%
	... Internal	3,079	43.17%
	... Other	1,569	21.99%
	Fund Type		
	... Closed-end Fund	1,370	19.21%
	... ODCE Fund	1,699	23.82%
	... Other	57	0.80%
	... Open-end Fund	1,060	14.86%
	... Single Client Account	2,947	41.32%

Notes: This table presents the summary statistics of categorical features.

The covariates included in component *b* can thus be summarized as in equation (3.5).

$$X_b = [\, S\ P\ F\ L\ E_{t0}\ \ E_{t-1}\,] \tag{3.5}$$

Our models incorporate 50 explanatory variables reflecting the main information used in the traditional appraisal methods discussed in the "Traditional Valuation Methods" section (i.e., the *income approach*, the *sales comparison approach*, and the *cost approach*). The input–output relationship is summarized in equation (3.6).

$$Y \sim [\, X_a\ X_b\,] \tag{3.6}$$

Table 3.2 provides a summary statistic of all numerical regressors, and Table 3.3 presents the distributions of the categorical features. It should be mentioned that, aside from the

components X_a and X_b following Matysiak and Wang (1995), appraisal values remain estimates and can rationally deviate from transaction prices for several reasons that are specific to the buyer or seller in the bargaining process and thus not foreseeable. However, we do not expect anything systematic in deviations of this kind, so we do not consider these random effects further.

3.4.4 Models

Non-parametric machine learning methods can identify interactions between the covariates without the need to specify them *a-priori*. Hence, these methods are not limited to any implicit assumptions of the relationship between X and Y and should be free of manual bias and specification error. To assess whether such methods can add to the understanding of pricing processes beyond the understanding achieved with traditional methods, we attempt to explain the information content in the appraisal errors Y using the extreme gradient boosting algorithm (i.e., boosting) by Chen and Guestrin (2016), which is an ensemble of regression trees.

The general concept of a regression tree as introduced by Breiman et al. (1984) is to divide the feature space into mutually exclusive intervals by creating binary decision rules for each feature that contributes to a reduction in the variation of the dependent variable. Such a decision rule is referred to as a split or node and can be thought of as a junction in the process of growing a branch of the tree. This splitting process is continued until the prediction error is minimized or a stopping criterion comes into effect. The resulting leaves of each branch are subsequently referred to as the terminal nodes of the regression tree, each representing a constant value as the final prediction rule. The entirety of these rules can be thought of as the regression tree model. To optimize model performance (i.e., select the optimal hyperparameters for model regularization), a tree model is iteratively trained (i.e., grown) using a training subsample and tested by passing the observations from the respective test subsample down the branches of the tree following the decision rules. Each observation is eventually assigned a terminal leaf corresponding to the final property price prediction.

However, individual trees' intuitiveness and flexibility are accompanied by the risk of quickly overfitting the training sample, thus imposing limitations on unseen data. A more dependable and robust approach is based on the idea of using many individual trees as building blocks of a larger prediction model, known as ensemble learner. The gradient boosting algorithm developed by Friedman (2001) is a prominent example of such ensemble learners. As demonstrated repeatedly in the literature, boosting achieves high accuracy and at the same time consistency for the prediction of property prices in the

residential sector, while being comparatively efficient from a computational perspective (e.g., Mayer et al., 2019; Deppner and Cajias, 2022; Lorenz et al., 2022).

In a boosting algorithm, a single regression tree is fitted as the base model and is then iteratively updated by sequentially growing new regression trees on the residuals of the preceding tree to continue learning and thereby "boosting" model accuracy. The final boosting model consists of an additive expansion of regression trees. The extreme gradient boosting algorithm by Chen and Guestrin (2016) only considers a randomly selected subset from all available predictors at each split in the tree-growing process and is thus a more regularized alternative of the gradient boosting algorithm by Friedman (2001). This introduces an additional source of variation into the model to provide more generalizable and robust estimations.

To further ensure the generalizability of the results, the performance of our models is evaluated using k-fold cross-validation. Cross-validation is a resampling technique used to counteract overfitting by partitioning the dataset into k mutually exclusive folds of the same size. The model is trained k times on $k-1$ folds and tested on the k^{th} fold, respectively, such that the model performance is entirely evaluated on unseen data without losing any observations.

By taking the appraisal error as our dependent variable, the manual appraisals from the NPI can be thought of as the base model in our boosting algorithm. Following Pace and Hayunga (2020), we use the standard deviation to measure the total variation in our dependent variable, that is, the manual appraisal error as specified in equation (3.3), as $\sigma_{Appraisal}$ and the unexplained residual variation of our boosting estimator as $\sigma_{Boosting}$, shown in equations (3.7) through (3.9).

$$\sigma_{Appraisal} = \sqrt{\frac{\sum_{i=1}^{n}|y-\bar{y}|^2}{n}} \tag{3.7}$$

$$\sigma_{Boosting} = \sqrt{\frac{\sum_{i=1}^{n}|\varepsilon-\bar{\varepsilon}|^2}{n}} \tag{3.8}$$

$$\varepsilon = y - \hat{y} \tag{3.9}$$

Our null hypothesis can thus be stated as:

> ***H*₀**: "*The difference between manual appraisals and sales prices cannot be explained by the existing covariates.*"

This is the case when the condition in equation (3.10) is fulfilled.

$$H_0: \frac{\sigma_{Appraisal}}{\sigma_{Boosting}} \leq 1 \tag{3.10}$$

In other words, this means that deviations between appraisals and sales prices follow a random process, and the improvement provided by machine learning algorithms over existing valuation approaches is not significantly different from zero. In contrast, the alternative hypothesis implies there is structured information content in the deviations between appraisals and sales prices, which machine learning models can exploit to explain these residuals further. This would provide an improvement in the understanding of pricing processes that goes beyond the understanding achieved with current appraisal methods:

> ***H*₁**: "*The difference between manual appraisals and sales prices can be explained by the existing covariates*."

Following the rationale of Pace and Hayunga (2020), the H_0 is rejected when the ratio of the total variation to the residual variation exceeds the value of 1, satisfying the condition in equation (3.11).

$$H_1: \frac{\sigma_{Appraisal}}{\sigma_{Boosting}} > 1 \tag{3.11}$$

Considering the results of the ANOVA test, which indicates systematic differences in appraisal errors across property types, we estimate separate models for each of the four asset sectors. Additionally, we calculate one global model for all property types, including the property type as an additional explanatory variable. In total, this results in five models.

After testing our hypotheses, we apply model-agnostic permutation feature importance (Fisher et al., 2019) to all models where the null hypothesis is rejected to examine the structure in appraisal errors. This method yields insights into the decision tree building process of the models so that the features are ranked according to their relative influence in reducing the variation between sales prices and market values and, thus, their contribution to shrinking the appraisal error.

3.5 Empirical Results

This section features the empirical results of our analyses. First, we present the descriptive statistics of the deviation between sales prices and appraisal values of commercial real estate from the NPI. We then examine the variation in these appraisal errors using extreme

gradient boosting trees. With respect to our research objectives, we analyze whether appraisal errors contain structured information that tree-based ensemble learners can exploit to further reduce appraisal errors. Subsequently, we discuss the features' relative importance to infer where the shrinkage in appraisal errors originates.

3.5.1 Descriptive Statistics

Following Cannon and Cole (2011), we investigate the accuracy and bias in appraisal values as estimates of sales prices. Table 3.4 provides a summary of the absolute percentage appraisal errors in our sample population and a disaggregated overview for each year and property type. Overall, the MAPE in our sample is 11.1% across all property types and years. This is smaller than the 13.2% reported by Cannon and Cole (2011) for the period between 1984 and 2009 but roughly the same magnitude. On average, accuracy is highest for apartments with an error of 8.6% and lowest for industrial sites with an error of 12.5%. The t-statistic tests the null hypothesis that the MAPE is not significantly different from zero in the respective groupings. The null can be rejected across all years, property types and for the aggregated sample, indicating inaccurate appraisals. We also do not find any evidence that the MAPE has significantly narrowed over the past decade compared to previous years when disregarding the large deviations that occurred during the great financial crisis in 2009.

Subsequently, we examine the signed percentage errors as a metric for bias, which is presented in Table 3.5. Matysiak and Wang (1995) and Cannon and Cole (2011) state that, on average, positive and negative deviations should cancel out, so appraisals are considered unbiased if the null hypothesis of the t-statistic, that is, the MPE is not significantly different from zero, is accepted. We find this to be the case for some individual years, particularly during flat market phases such as in 2001 and 2002 after the burst of the Dot-com bubble, in 2012 in the aftermath of the great financial crisis, between 2016 and 2017 when capital appreciation in U.S. commercial real estate markets was cooling off, and from 2020 through 2021, when the Covid-19 pandemic caused uncertainty in commercial markets, dampening growth. However, the null hypothesis is rejected for all years in which markets were either in rising or falling regimes. We find that the MPE averages 4.97% during rising markets, indicating a structural underestimation of property prices, whereas this metric turns negative at 12.95% during the sharp downturn between 2008 and 2009, the only period of falling markets in our sample, indicating overestimation of prices. This provides evidence that appraisal values tend to lag sales prices in moving markets and strongly corroborates the findings by Cannon and Cole (2011) and previous studies showing that market cycles have an impact on the reliability of appraisals.

Table 3.4: Absolute Percentage Error between Sale Prices and Manual Appraisals

	All Types (N = 7,133)				Apartment (N = 1,904)				Industrial (N = 2,337)				Office (N = 2,056)				Retail (N = 836)			
Year	MdAPE	MAPE	t-Stat		MdAPE	MAPE	t-Stat		MdAPE	MAPE	t-Stat		MdAPE	MAPE	t-Stat		MdAPE	MAPE	t-Stat	
1997	8.02%	9.38%	11.56	***	6.14%	7.04%	5.36	***	8.10%	10.38%	7.78	***	13.25%	13.44%	6.80	***	5.21%	6.87%	4.54	***
1998	11.15%	13.87%	8.96	***	14.14%	11.46%	5.95	***	9.89%	15.21%	3.58	***	15.27%	14.86%	7.94	***	8.78%	11.42%	5.34	***
1999	9.01%	10.12%	14.45	***	6.43%	7.39%	6.65	***	13.54%	11.60%	7.57	***	8.13%	9.71%	7.40	***	13.05%	11.44%	8.20	***
2000	7.65%	9.95%	16.99	***	9.87%	10.63%	11.17	***	5.25%	8.51%	6.87	***	9.13%	10.57%	10.11	***	6.51%	9.60%	5.97	***
2001	6.52%	9.69%	12.29	***	6.70%	8.54%	10.10	***	7.56%	9.91%	6.91	***	6.30%	10.07%	5.35	***	5.02%	10.88%	4.19	***
2002	7.63%	10.87%	12.64	***	7.12%	10.19%	5.62	***	9.30%	12.65%	8.28	***	7.29%	10.18%	6.02	***	8.42%	9.08%	5.98	***
2003	7.14%	9.17%	19.29	***	6.27%	7.94%	9.23	***	6.90%	8.58%	11.31	***	6.48%	9.99%	9.66	***	9.75%	10.83%	10.62	***
2004	8.98%	11.49%	20.21	***	7.85%	9.66%	11.14	***	9.67%	13.28%	10.54	***	8.77%	10.95%	13.08	***	9.59%	11.05%	8.72	***
2005	15.74%	15.97%	29.25	***	9.71%	13.09%	13.90	***	20.20%	17.21%	23.10	***	13.92%	16.32%	10.93	***	17.63%	17.01%	21.11	***
2006	11.43%	13.11%	23.63	***	10.54%	12.36%	12.53	***	12.99%	13.88%	11.77	***	10.78%	13.58%	14.89	***	10.53%	10.52%	7.97	***
2007	10.57%	12.28%	26.97	***	9.28%	11.22%	13.35	***	11.12%	12.14%	18.48	***	11.87%	14.09%	14.78	***	5.21%	8.25%	6.47	***
2008	7.26%	11.96%	8.34	***	7.44%	10.75%	8.29	***	5.91%	7.86%	8.10	***	8.75%	17.75%	4.63	***	5.16%	6.22%	3.49	**
2009	17.32%	22.77%	14.13	***	13.43%	17.51%	9.27	***	20.34%	26.12%	8.10	***	19.19%	27.61%	7.70	***	9.36%	14.49%	3.67	***
2010	9.20%	11.62%	16.70	***	7.87%	10.30%	10.46	***	10.64%	12.90%	9.28	***	9.10%	11.30%	7.64	***	12.62%	12.99%	5.44	***
2011	7.91%	10.53%	18.79	***	6.86%	8.04%	11.84	***	8.98%	10.24%	12.49	***	7.69%	11.32%	8.44	***	8.96%	13.80%	7.49	***
2012	7.02%	10.63%	16.50	***	6.28%	7.74%	11.81	***	7.57%	12.05%	10.16	***	8.08%	12.61%	7.86	***	7.39%	8.05%	7.63	***
2013	7.10%	9.52%	22.82	***	4.08%	5.57%	15.08	***	8.32%	10.90%	15.33	***	9.23%	11.38%	11.25	***	9.38%	11.59%	8.31	***
2014	7.70%	10.57%	19.32	***	5.66%	6.96%	14.28	***	12.62%	12.94%	17.39	***	5.22%	9.69%	7.50	***	5.65%	11.66%	5.12	***
2015	8.47%	11.51%	19.61	***	7.46%	8.59%	15.59	***	10.77%	15.43%	11.29	***	6.64%	10.74%	9.30	***	8.68%	10.45%	8.78	***
2016	6.72%	10.74%	18.53	***	4.78%	7.13%	14.83	***	8.04%	12.43%	14.09	***	6.11%	13.03%	7.97	***	8.12%	9.71%	10.10	***
2017	5.62%	9.09%	15.64	***	4.25%	6.28%	13.99	***	7.50%	11.24%	10.37	***	5.59%	8.37%	11.85	***	4.79%	13.12%	3.00	***
2018	5.96%	8.46%	18.46	***	5.01%	7.21%	12.13	***	7.84%	11.61%	8.57	***	5.73%	7.86%	11.83	***	5.08%	9.09%	3.65	***
2019	6.08%	8.66%	21.56	***	5.57%	6.04%	13.81	***	8.47%	10.61%	16.34	***	5.59%	7.16%	10.23	***	5.29%	9.70%	5.43	***
2020	6.30%	9.29%	14.28	***	3.43%	5.26%	8.79	***	9.75%	10.35%	10.85	***	6.43%	10.66%	6.97	***	7.17%	12.35%	5.80	***
2021	10.32%	13.56%	8.19	***	6.71%	8.45%	4.83	***	14.92%	15.13%	9.43	***	8.40%	14.27%	2.31	**	10.32%	7.26%	2.24	**
All	7.99%	11.12%	81.72	***	6.47%	8.62%	49.70	***	9.73%	12.50%	51.31	***	7.69%	11.71%	39.01	***	8.71%	11.52%	28.86	***

Notes: This table presents the median absolute percentage appraisal error (MdAPE) and the mean absolute percentage appraisal error (MAPE) as a measure of accuracy. The t-statistic tests the null hypothesis that the MAPE is not significantly different from zero, i.e., appraisals are accurate. Significance codes indicate that the MAPE is statistically different from zero at the respective level of confidence: $p < 0.01$ '***', $p < 0.05$ '**', $p < 0.1$ '*'.

Table 3.5: Signed Percentage Error between Sale Prices and Manual Appraisals

	All Types (N = 7,133)				Apartment (N = 1,904)				Industrial (N = 2,337)				Office (N = 2,056)				Retail (N = 836)			
Year	MdPE	MPE	t-Stat		MdPE	MPE	t-Stat		MdPE	MPE	t-Stat		MdPE	MPE	t-Stat		MdPE	MPE	t-Stat	
1997	7.22%	6.80%	6.00	***	6.14%	5.97%	3.70	***	7.87%	7.26%	3.82	***	12.36%	10.50%	2.94	**	4.87%	3.73%	1.57	
1998	8.71%	7.60%	3.79	***	14.14%	10.27%	4.17	***	8.46%	3.15%	0.61		15.27%	12.79%	5.50	***	4.75%	2.46%	0.67	
1999	5.46%	5.37%	4.74	***	5.80%	4.62%	2.59	**	6.83%	4.22%	1.39		4.74%	4.83%	2.39	**	10.83%	7.25%	3.25	***
2000	2.69%	2.04%	2.26	**	5.77%	5.74%	3.63	***	1.09%	-1.58%	-0.92		2.09%	1.61%	1.00		4.81%	2.81%	1.16	
2001	2.79%	0.47%	0.43		6.00%	7.04%	6.53	***	3.47%	-0.49%	-0.24		-2.97%	-4.49%	-1.91	*	2.14%	-2.73%	-0.84	
2002	3.21%	0.93%	0.79		4.69%	3.91%	1.73	*	1.23%	-1.89%	-0.86		0.79%	0.25%	0.11		4.56%	3.68%	1.60	
2003	4.24%	3.18%	4.40	***	4.75%	4.10%	3.32	***	3.46%	2.56%	2.13	**	1.31%	1.30%	0.85		8.11%	7.51%	4.38	***
2004	5.23%	4.66%	5.77	***	3.94%	4.58%	3.47	***	3.90%	2.94%	1.69	*	5.83%	5.73%	4.65	***	6.18%	6.99%	3.72	***
2005	14.76%	12.19%	16.84	***	8.95%	10.75%	9.08	***	19.44%	14.05%	12.36	***	10.93%	8.56%	4.45	***	17.06%	16.19%	16.76	***
2006	9.80%	9.26%	11.98	***	9.79%	10.89%	9.09	***	9.96%	6.65%	3.73	***	9.86%	10.62%	8.79	***	8.68%	6.80%	3.19	***
2007	8.47%	7.85%	11.82	***	7.42%	6.82%	5.41	***	9.28%	8.44%	8.52	***	8.67%	8.28%	5.90	***	4.55%	6.36%	3.88	***
2008	-1.85%	-5.44%	-3.26	***	-3.94%	-2.33%	-1.12		-2.99%	-4.89%	-3.80	***	0.85%	-8.87%	-2.02	**	-2.68%	-1.99%	-0.62	
2009	-16.87%	-20.46%	-11.40	***	-13.20%	-14.94%	-6.64	***	-20.34%	-24.66%	-7.19	***	-19.19%	-24.47%	-5.93	***	-7.67%	-12.41%	-2.61	**
2010	4.50%	2.55%	2.34	**	6.24%	7.01%	5.16	***	-1.35%	-1.14%	-0.51		5.93%	2.20%	0.95		3.08%	-1.11%	-0.29	
2011	4.28%	3.21%	3.80	***	5.72%	4.09%	3.77	***	4.61%	3.14%	2.35	**	3.75%	3.56%	1.74	*	-1.05%	1.73%	0.63	
2012	1.41%	-0.68%	-0.82		2.21%	1.06%	1.08		2.02%	-0.29%	-0.19		-1.38%	-3.34%	-1.65	*	-2.13%	-0.45%	-0.27	
2013	3.42%	2.96%	5.09	***	2.69%	3.26%	6.23	***	4.86%	4.07%	3.80	***	3.01%	1.74%	1.21		3.53%	1.90%	0.96	
2014	5.28%	4.45%	6.41	***	4.50%	4.11%	5.69	***	10.75%	6.97%	6.44	***	2.19%	1.83%	1.20		3.78%	2.91%	1.07	
2015	4.57%	2.89%	3.65	***	6.72%	6.36%	8.48	***	4.75%	-0.31%	-0.16		3.17%	1.60%	1.07		5.29%	5.81%	3.36	***
2016	0.75%	-1.14%	-1.53		3.35%	3.99%	5.89	***	-3.19%	-3.47%	-2.69	***	-1.65%	-5.06%	-2.64	***	2.36%	2.00%	1.20	
2017	1.83%	0.38%	0.52		3.35%	3.59%	5.59	***	1.78%	-0.23%	-0.16		-0.84%	-0.10%	-0.10		-0.65%	-6.74%	-1.42	
2018	2.76%	1.87%	2.93	***	3.94%	3.70%	4.49	***	4.78%	2.70%	1.41		1.25%	0.57%	0.61		-2.58%	-4.43%	-1.36	
2019	2.58%	2.62%	4.62	***	1.11%	0.59%	0.76		6.10%	6.95%	7.87	***	1.60%	1.33%	1.37		-3.74%	-7.37%	-3.66	***
2020	0.53%	-0.39%	-0.42		1.36%	1.71%	1.93	*	4.08%	4.89%	3.16	***	-2.90%	-3.26%	-1.61		-5.36%	-8.17%	-3.04	***
2021	3.21%	0.87%	0.37		6.42%	4.67%	1.59		5.49%	3.29%	1.13		-1.65%	-6.64%	-0.93		-10.32%	-7.26%	-2.24	
All	3.78%	2.71%	14.51	***	4.09%	4.07%	16.57	***	4.60%	2.62%	7.46	***	2.53%	1.47%	3.72	***	3.76%	2.90%	5.23	***

Notes: This table presents the median percentage appraisal error (MdPE) and the mean percentage appraisal error (MPE) as a measure of biasedness. The t-statistic tests the null hypothesis that the MPE is not significantly different from zero, i.e., appraisals are unbiased. Significance codes indicate that the MPE is statistically different from zero at the respective level of confidence: $p < 0.01$ '***', $p < 0.05$ '**', $p < 0.1$ '*'.

3.5.2 Residual Standard Deviation

After confirming the findings of inaccuracy and structural bias made by Cannon and Cole (2011) for our sample period, we investigate the variation in the respective appraisal errors (i.e., residuals). The results of the analysis were obtained by applying the extreme gradient boosting algorithm (i.e., boosting) separately for each property type and to the aggregated dataset. The models were repeatedly cross-validated by ten mutually exclusive folds to avoid overfitting, such that each of the folds was used once as a test sample. The hyperparameters of the boosting estimators were optimized via the root mean square error using a grid search procedure. All error measures are reported as 10-fold cross-validation errors, thus representing out-of-sample estimations. The results are displayed in Table 3.6. By analogy to the study of Pace and Hayunga (2020), the last two columns depict the ratio of the standard deviation from the dependent variable (i.e., total variation of appraisal errors) to the residuals resulting from the machine learning estimations (i.e., unexplained variation of appraisal errors). The ratio exceeds 1 for any case where the appraisal errors can be further explained by the applied boosting procedure.

Table 3.6: Residual Standard Deviation

	$\sigma_{Appraisal}$	$\sigma_{Boosting}$	$R^2_{Boosting}$	$\frac{\sigma_{Appraisal}}{\sigma_{Boosting}}$
All Types	0.15	0.13	0.26	1.17
Apartment	0.11	0.09	0.31	1.20
Industrial	0.16	0.14	0.28	1.18
Office	0.16	0.14	0.25	1.16
Retail	0.15	0.13	0.22	1.14

Notes: This table benchmarks the residual variation of manual appraisals against the residual variation of the boosting algorithm, whereby σ is the standard deviation of the respective residuals. A performance improvement occurs whenever the ratio of $\sigma_{Appraisal}$ over $\sigma_{Boosting}$ exceeds the value 1.

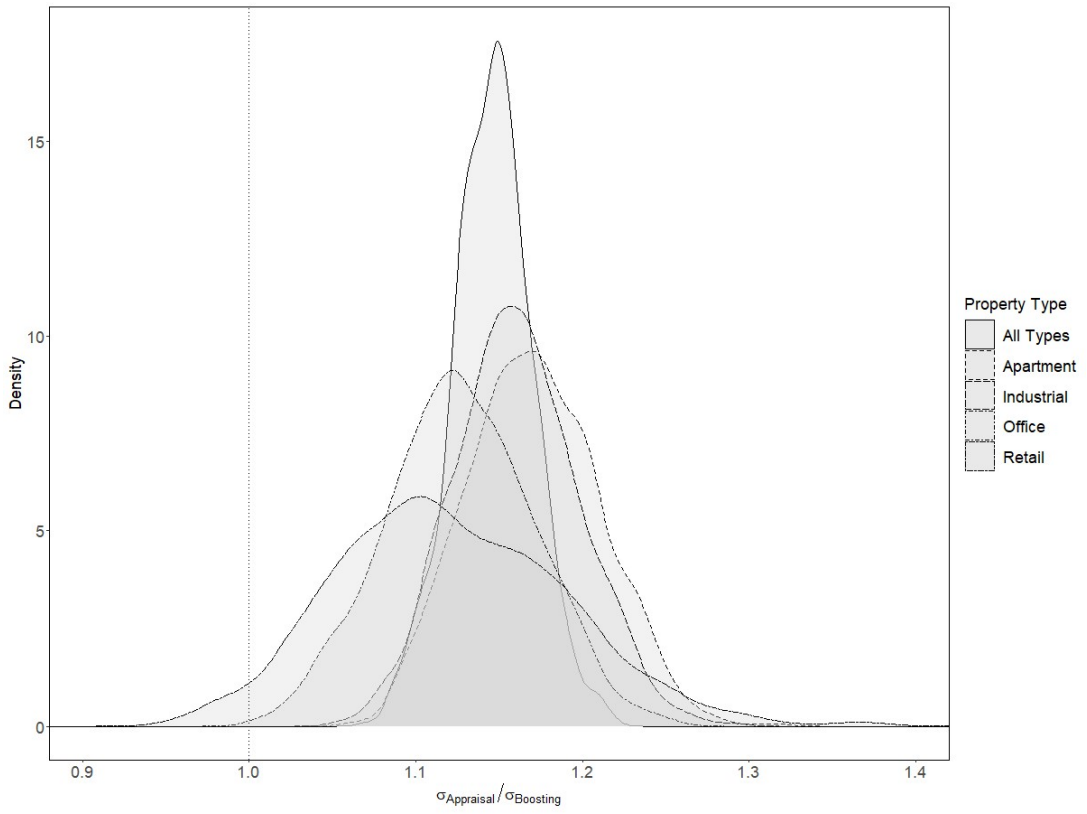
We find the results in Table 3.6 to be unequivocal in all four asset classes, as a reduction in the variation of appraisal errors (i.e., residual variation) can be achieved in all cases. The boosting algorithms yield considerable improvements, with coefficients taking values well above 1. The reduction in the residual variation is highest for apartments with 20.5% and lowest for retail properties with approximately 14.2%. By implication, such a reduction signals that the appraisal error is systematic to some extent rather than purely random.

To formally test our hypothesis and rule out that improvements occur by pure chance, we apply bootstrapping to create confidence intervals for the shrinkage of the residual variation in our dependent variable. This is achieved by generating 1,000 random bootstrap samples and repeatedly training and testing the models on each sample. Figure 3.2 presents the bootstrap distribution of the model performance for all five models. Based on the bootstrap confidence intervals, the null hypothesis stated in equation (3.10) can be

rejected at a 5% level of significance for the retail model and at a 1% level of significance for all other models.

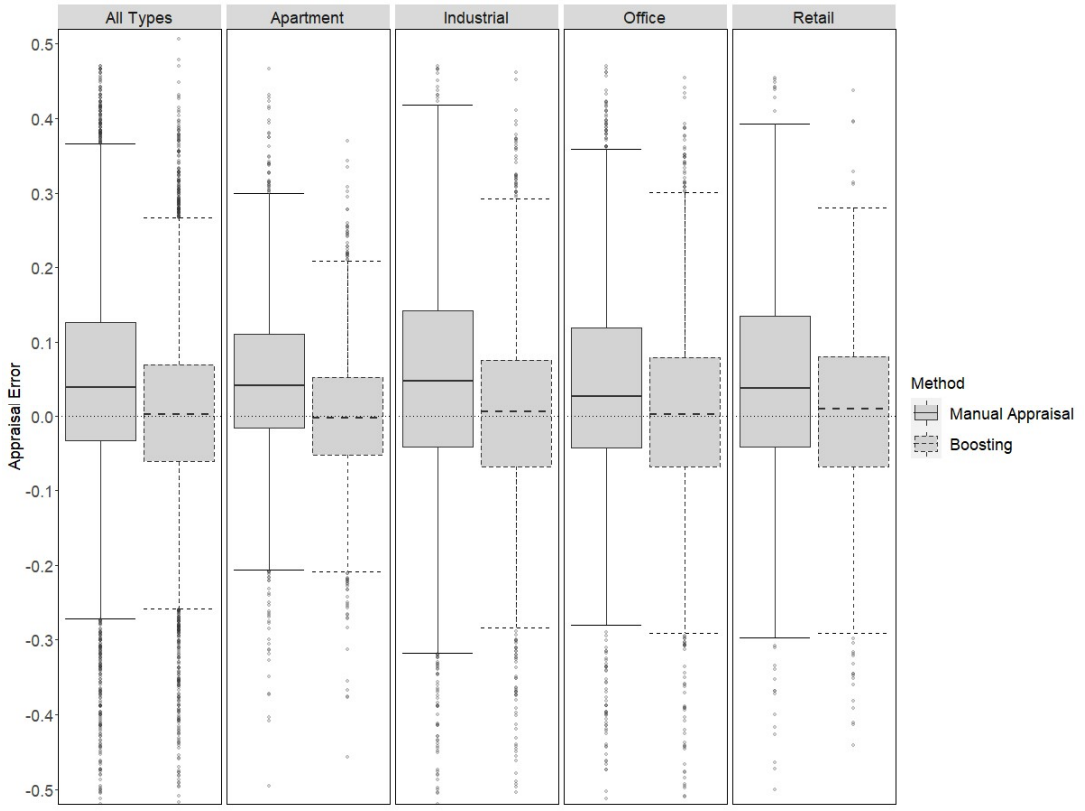
Figure 3.3 depicts the distributions of the residuals by asset class. Matysiak and Wang (1995) and Cannon and Cole (2011) show appraisal errors to be biased in their samples. That is, the mean of the error distribution was positive or negative and not around zero. This can also be observed in Figure 3.3 for the median appraisal errors, which are considerably above the horizontal null point line in all asset classes, indicating that most properties are overvalued. In contrast, all machine learning models produce residuals close to zero. This indicates that the estimated models are not biased and produce reliable responses. Furthermore, the 25th and 75th percentiles of the boxplots show that the dispersion of the residuals from boosting is smaller than the original appraisal errors for all property types.

Figure 3.2: Bootstrap Distribution of Model Performance



Notes: The density plot shows the bootstrap distribution of the model performance for all five models using 1,000 random bootstrap samples. A performance improvement occurs whenever the ratio $\frac{\sigma_{Appraisal}}{\sigma_{Boosting}} > 1$, as indicated by the dotted horizontal line. The area to the right of the dotted line can be interpreted as the confidence interval for which the null hypothesis $\frac{\sigma_{Appraisal}}{\sigma_{Boosting}} \leq 1$ can be rejected. The null hypothesis can be rejected at a 5% level of significance for all models and at a 1% level of significance for all models except for the retail model. The respective ratios measured by 10-fold cross-validation are presented in Table 3.6.

Figure 3.3: Comparison of Residual Variation



Notes: The boxplots show the distribution of the raw appraisal errors (solid line) in comparison to the boosted appraisal errors (dashed line). The box of each boxplot represents 50% of the data within the 25th and 75th percentile. The bold line within the box indicates the median of each distribution. The whiskers indicate the 1.5 interquartile range (IQR). The dotted horizontal line marks the null point on the y-axis.

We also see a relationship between the homogeneity of asset classes and the performance improvement. Relatively homogenous property types (i.e., apartments, industrial) benefit more from machine learning than relatively heterogenous asset classes (i.e., retail, office). The same applies to the sample size, as data-driven techniques require homogenous and large samples to learn patterns from the data.

To test whether the reduction in the residual variation can also reduce bias in the actual appraisals, we infer hypothetical appraisal values from the estimated percentage appraisal errors by multiplying these by the original appraisal values. In analogy to the descriptive statistics of the manual appraisal errors in the "Appraisal Error" section, Tables 3.7 and 3.8 present the adjusted appraisal values obtained by the boosting algorithms. Overall, the MAPE presented in Table 3.7 is reduced for all asset classes. In the aggregated models, a reduction from 11.12% to 9.25% is achieved. The highest absolute reduction in the MAPE was achieved for industrial properties with 2.48 percentage points (i.e., 19.85%) by the boosting model. The highest relative reduction in the MAPE was achieved for apartments with 20.91% (i.e., 1.80 percentage points). The lowest absolute and relative improvement

Table 3.7: Absolute Percentage Error between Sale Prices and Boosted Appraisals

	All Types (N = 7,133)				Apartment (N = 1,904)				Industrial (N = 2,337)				Office (N = 2,056)				Retail (N = 836)			
Year	MdAPE	MAPE	t-Stat		MdAPE	MAPE	t-Stat		MdAPE	MAPE	t-Stat		MdAPE	MAPE	t-Stat		MdAPE	MAPE	t-Stat	
1997	5.88%	7.63%	9.07	***	5.37%	5.76%	5.67	***	7.06%	8.03%	5.79	***	8.83%	11.03%	4.07	***	6.41%	5.71%	4.32	***
1998	8.54%	11.24%	6.94	***	6.50%	7.09%	5.97	***	9.05%	14.30%	2.99	***	9.17%	10.46%	8.99	***	7.51%	9.53%	4.72	***
1999	6.86%	8.55%	13.47	***	5.17%	6.41%	4.81	***	10.87%	11.88%	8.36	***	7.24%	8.25%	7.75	***	6.86%	7.94%	6.46	***
2000	6.43%	8.69%	15.34	***	6.51%	7.43%	8.97	***	5.60%	8.49%	6.67	***	8.47%	10.45%	9.54	***	7.94%	10.13%	5.75	***
2001	5.77%	8.62%	11.86	***	5.05%	6.22%	8.60	***	7.00%	9.30%	6.81	***	5.70%	8.77%	5.21	***	5.83%	11.90%	4.62	***
2002	7.13%	10.29%	11.34	***	6.69%	8.89%	4.41	***	9.00%	12.36%	7.89	***	5.94%	8.67%	5.97	***	5.94%	8.47%	6.56	***
2003	6.07%	8.24%	17.01	***	4.80%	6.93%	7.25	***	5.98%	7.78%	8.72	***	7.23%	9.10%	9.59	***	7.52%	8.04%	7.68	***
2004	7.64%	10.39%	18.16	***	7.62%	7.95%	11.45	***	8.78%	12.60%	9.57	***	8.25%	9.47%	13.28	***	9.10%	9.87%	10.90	***
2005	8.83%	11.33%	19.01	***	7.58%	9.51%	11.46	***	7.82%	10.67%	13.30	***	9.25%	12.64%	10.87	***	8.41%	9.41%	13.14	***
2006	8.06%	10.20%	21.02	***	7.17%	8.59%	11.40	***	8.01%	10.00%	11.49	***	9.32%	11.10%	13.34	***	7.75%	8.96%	6.87	***
2007	6.93%	9.03%	21.85	***	6.49%	8.44%	11.88	***	7.28%	8.47%	13.41	***	8.66%	11.30%	11.85	***	7.18%	8.04%	7.75	***
2008	6.19%	10.00%	8.14	***	8.16%	9.05%	7.78	***	4.15%	6.27%	7.82	***	9.33%	16.24%	4.80	***	8.83%	8.07%	5.00	***
2009	10.16%	12.82%	13.56	***	8.57%	9.98%	10.04	***	10.45%	14.62%	7.19	***	9.53%	13.01%	6.41	***	4.97%	11.64%	3.24	***
2010	8.21%	10.12%	15.29	***	7.30%	8.85%	9.69	***	9.02%	10.46%	8.52	***	8.01%	10.04%	7.87	***	10.28%	13.18%	6.37	***
2011	6.65%	9.12%	16.26	***	4.27%	6.77%	9.10	***	6.11%	8.00%	9.88	***	8.75%	11.35%	8.29	***	6.75%	11.62%	6.71	***
2012	6.32%	9.36%	17.82	***	5.35%	6.75%	11.30	***	6.10%	10.28%	10.13	***	7.76%	10.80%	9.77	***	6.94%	7.64%	7.96	***
2013	6.18%	8.89%	21.72	***	2.82%	4.87%	12.25	***	7.68%	9.43%	14.58	***	8.10%	10.79%	11.25	***	8.27%	12.16%	7.31	***
2014	5.88%	8.69%	16.04	***	5.52%	6.31%	14.90	***	5.77%	9.16%	12.56	***	6.33%	9.55%	7.51	***	6.92%	11.16%	4.76	***
2015	6.76%	9.26%	19.35	***	5.28%	6.52%	14.07	***	8.99%	12.23%	10.66	***	6.52%	9.74%	10.15	***	7.50%	8.43%	8.57	***
2016	6.33%	9.21%	20.62	***	4.17%	5.26%	14.38	***	7.98%	10.88%	12.81	***	6.82%	11.74%	9.80	***	7.32%	8.90%	9.29	***
2017	5.58%	8.82%	14.08	***	3.74%	5.18%	13.14	***	8.03%	11.97%	9.39	***	6.25%	8.64%	11.36	***	3.04%	12.04%	2.84	***
2018	5.96%	8.02%	17.42	***	4.45%	5.77%	10.82	***	7.91%	10.80%	7.49	***	6.00%	7.61%	11.98	***	6.87%	10.44%	3.44	***
2019	4.91%	6.92%	20.00	***	4.78%	5.63%	11.87	***	4.38%	7.22%	12.50	***	5.54%	7.09%	11.22	***	5.24%	9.48%	6.02	***
2020	5.40%	7.90%	14.15	***	4.36%	5.42%	8.83	***	5.19%	7.22%	8.55	***	8.28%	10.76%	7.89	***	7.72%	10.80%	6.58	***
2021	10.40%	12.27%	8.39	***	3.51%	7.15%	2.97	**	13.39%	13.26%	10.17	***	9.30%	16.09%	2.86	***	7.26%	5.34%	2.25	
All	6.58%	9.25%	75.61	***	5.24%	6.82%	46.47	***	7.27%	10.02%	44.60	***	7.49%	10.27%	41.11	***	7.48%	9.89%	25.83	***

Notes: This table presents the boosting-adjusted median absolute percentage appraisal error (MdAPE) and the mean absolute percentage appraisal error (MAPE) as a measure of accuracy. The t-statistic tests the null hypothesis that the MAPE is not significantly different from zero, i.e., appraisals are accurate. Significance codes indicate that the MAPE is statistically different from zero at the respective level of confidence: $p < 0.01$ '***', $p < 0.05$ '**', $p < 0.1$ '*'.

Table 3.8: Signed Percentage Error between Sale Prices and Boosted Appraisals

	All Types (N = 7,133)				Apartment (N = 1,904)				Industrial (N = 2,337)				Office (N = 2,056)				Retail (N = 836)			
Year	MdPE	MPE	t-Stat		MdPE	MPE	t-Stat		MdPE	MPE	t-Stat		MdPE	MPE	t-Stat		MdPE	MPE	t-Stat	
1997	0.29%	0.31%	0.25		-1.34%	-0.17%	-0.10		0.66%	1.14%	0.57		5.73%	3.70%	0.81		0.21%	0.54%	0.24	
1998	3.00%	-0.18%	-0.09		1.51%	1.23%	0.51		4.48%	-1.33%	-0.24		4.07%	2.97%	1.37		2.47%	-1.95%	-0.61	
1999	-1.07%	-0.58%	-0.53		-1.38%	-2.13%	-1.08		4.12%	-0.36%	-0.11		0.49%	0.15%	0.08		0.35%	-0.82%	-0.42	
2000	-0.35%	-1.18%	-1.42		-2.17%	-1.47%	-1.11		-1.04%	-2.50%	-1.44		-0.72%	-1.80%	-1.11		0.06%	-1.58%	-0.60	
2001	0.66%	-1.35%	-1.39		0.49%	0.24%	0.22		3.28%	-1.88%	-1.00		-1.93%	-3.03%	-1.43		1.47%	-3.11%	-0.92	
2002	0.29%	-1.20%	-1.02		-1.26%	-2.39%	-1.01		1.03%	-1.63%	-0.74		-0.38%	-0.78%	-0.41		1.39%	1.30%	0.60	
2003	0.47%	-0.21%	-0.30		-0.03%	-0.60%	-0.46		-0.86%	-0.95%	-0.76		-1.09%	-0.88%	-0.63		1.16%	-0.12%	-0.07	
2004	0.05%	-0.99%	-1.23		-1.22%	-0.78%	-0.68		-0.92%	-2.70%	-1.55		0.39%	-0.62%	-0.54		2.70%	0.69%	0.38	
2005	1.22%	-1.00%	-1.27		-1.20%	-0.35%	-0.28		3.40%	0.58%	0.48		1.52%	-0.72%	-0.45		1.87%	-0.13%	-0.11	
2006	0.25%	-0.44%	-0.57		0.44%	0.36%	0.28		1.43%	0.55%	0.39		-0.22%	0.20%	0.15		-0.10%	-0.02%	-0.01	
2007	-0.14%	-0.47%	-0.75		0.04%	-0.28%	-0.25		1.36%	-0.09%	-0.10		0.94%	0.19%	0.13		0.10%	1.14%	0.61	
2008	-0.24%	-2.07%	-1.42		-0.81%	-1.50%	-0.82		-1.21%	-1.74%	-1.51		1.85%	-5.00%	-1.25		-2.75%	-2.68%	-0.71	
2009	3.25%	-1.51%	-1.09		-0.98%	-1.10%	-0.67		1.87%	-3.89%	-1.38		0.11%	-3.73%	-1.31		-2.45%	-4.87%	-0.94	
2010	0.71%	0.28%	0.28		2.55%	1.80%	1.28		-1.15%	-0.24%	-0.13		3.16%	2.47%	1.23		6.48%	0.55%	0.15	
2011	0.88%	-0.46%	-0.57		1.34%	-0.81%	-0.73		1.00%	-0.34%	-0.29		-1.41%	-0.86%	-0.40		0.52%	-0.57%	-0.24	
2012	0.22%	-0.89%	-1.28		0.28%	-1.07%	-1.23		-0.64%	-1.91%	-1.48		0.94%	-1.24%	-0.80		-1.47%	-0.13%	-0.08	
2013	0.10%	-0.68%	-1.19		0.48%	-0.46%	-0.82		-0.59%	-0.82%	-0.83		2.01%	-0.74%	-0.54		0.65%	-0.63%	-0.28	
2014	0.33%	-0.81%	-1.22		-0.93%	-0.49%	-0.67		0.99%	-0.51%	-0.52		0.55%	-0.81%	-0.53		0.37%	-2.02%	-0.73	
2015	-0.04%	-0.67%	-1.03		-0.43%	-0.21%	-0.28		-1.44%	-2.59%	-1.68	*	0.38%	-0.82%	-0.63		1.86%	1.86%	1.21	
2016	-0.08%	-1.01%	-1.68	*	-0.24%	0.04%	0.07		-0.84%	-1.43%	-1.19		-0.54%	-3.07%	-2.02	**	1.11%	-0.05%	-0.03	
2017	-0.13%	-1.30%	-1.72	*	-0.34%	-0.25%	-0.40		-0.71%	-2.53%	-1.56		-1.41%	-1.11%	-1.02		1.13%	-6.07%	-1.33	
2018	-0.73%	-1.17%	-1.87	*	-0.36%	-0.97%	-1.30		0.28%	-0.88%	-0.46		-0.71%	-1.35%	-1.50		-1.93%	-5.99%	-1.58	
2019	-0.24%	-0.47%	-0.98		-0.04%	-1.16%	-1.54		0.65%	0.59%	0.75		-0.32%	-0.59%	-0.64		0.71%	-3.39%	-1.65	
2020	-0.05%	-0.44%	-0.56		-1.48%	-0.47%	-0.51		0.51%	0.61%	0.47		-1.48%	-1.35%	-0.69		-2.49%	-3.22%	-1.31	
2021	-0.13%	-0.99%	-0.47		1.96%	1.29%	0.38		-0.43%	-0.39%	-0.15		2.66%	-3.95%	-0.56		-0.63%	-0.50%	-0.11	
All	0.17%	-0.81%	-4.94	***	-0.27%	-0.49%	-2.28	**	0.46%	-1.01%	-3.30	***	0.09%	-1.03%	-3.06	***	0.85%	-1.00%	-1.96	**

Notes: This table presents the boosting-adjusted median percentage appraisal error (MdPE) and the mean percentage appraisal error (MPE) as a measure of biasedness. The t-statistic tests the null hypothesis that the MPE is not significantly different from zero, i.e., appraisals are unbiased. Significance codes indicate that the MPE is statistically different from zero at the respective level of confidence: $p < 0.01$ '***', $p < 0.05$ '**', $p < 0.1$ '*'.

can be observed for office buildings. However, this is still 1.44 percentage points absolute and above 12.32% relative. These figures confirm the findings of a significant reduction in the residual variation (see Table 3.6) and support the hypothesis that machine learning algorithms can exploit the structured covariance found in the residuals to further shrink appraisal errors.

Compared to Table 3.5, the mean percentage errors in Table 3.8 reveal that the bias in appraisal values could be successfully eliminated in most of the years and asset sectors. The acceptance of the null hypothesis that the MPE is not significantly different from zero for all the years except for the period between 2016 and 2018, in which the null could only be rejected at the 10% confidence level, confirms that manual appraisal errors are systematic. It also further supports previous findings in that the boosting estimator provides unbiased estimates, although the mean percentage errors are negative for all years except for 1997 and 2010, indicating a slight overestimation of the inferred appraisal values.

Overall, we find that boosting can provide material improvements in increasing accuracy and reducing structural bias in commercial appraisal values. However, it should also be mentioned that machine learning methods are no crystal ball that can accurately predict downturns such as during the great financial crisis without previously learning the effects of varying economic conditions under transitioning market regimes. Moreover, external shocks such as pandemics, wars, or any sort of crises are difficult to train since they occur infrequently and can take on various forms.

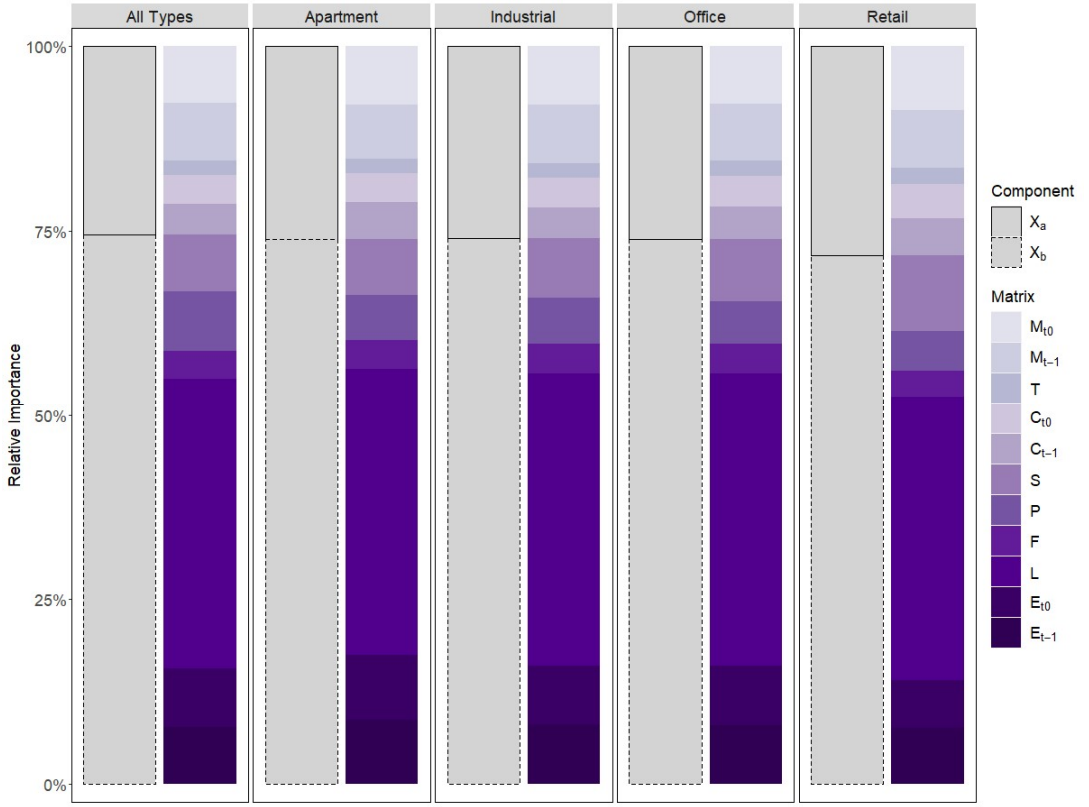
3.5.3 Permutation Feature Importance

To draw conclusions about which features contribute most to the shrinkage of the residual variation, we apply the model-agnostic permutation feature importance by Fisher et al. (2019). Figure 3.4 provides a summary of the feature groupings introduced in the "Explanatory Variables" section, decomposed according to their relative importance in shrinking the appraisal error. Features that repeatedly appear at early splitting points of the individual regression trees or show up more often in the tree-growing process have a high importance score. Identifying these features provides insights into factors that are not adequately reflected in current appraisal practices. This can offer constructive criticism to improve the state-of-the-art (Pace and Hayunga, 2020).

The bar chart in Figure 3.4 shows that both components *a* and *b* have an evident influence on appraisal errors, with component *b* dominating by about three-quarters. This indicates

that the improvement achieved by the boosting algorithm is not solely due to the time lag between appraisal and sale, but results to a great extent from valuation bias.

Figure 3.4: Relative Permutation Feature Importance



Notes: The bar chart shows the relative permutation feature importance of both components X_a and X_b (indicated by the linetype) and the various feature clusters described in section *3.4.3* (indicated by the color) for each of the five models. The relative importance on the y-axis indicates the relative contribution of each component and cluster to the reduction of the prediction error. The order of groupings is arbitrary.

Overall, location (L) appears to be the most relevant cluster for explaining appraisal errors, accounting for nearly 40% across all models. To a great extent, this is driven by the spatial coordinates. When a regression tree splits on the latitude and longitude, it effectively identifies new submarkets for which it generates individual models, indicating that spatial considerations on the micro-level are not appropriately reflected in appraisal values. This is consistent with Pace and Hayunga (2020), who find that the performance improvement of boosting and bagging regression trees compared to linear hedonic models results to a great extent from exploiting spatial structures in the residuals that cannot be captured with location dummies, such as ZIP code or MSA code areas. However, this seems to be different for industrial properties, as the resolution of MSAs appears to exploit spatial structures in the residuals better than the coordinates, implying that locational factors on the macro-level are overlooked in this sector.

With respect to component *a*, we find Capex in the second quarter before the sale to be the feature with the highest average impact on appraisal errors across all models. This is surprising, as the appraiser should know Capex measures before they occur. However, Beracha et al. (2019) find that in instances, appraisals are updated by simply adding Capex to the market values. This is known as a stale appraisal and may not adequately reflect the true intrinsic value of a building improvement.

For component *b*, the building occupancy is on average the most important feature driving appraisal errors. As described by Beracha et al. (2019), the relation between vacant space and commercial real estate value depends on the optionality of vacant space, which can be based on either a growth hypothesis (i.e., assuming higher future NOI growth from the potential of leasing up vacant space) or a risk hypothesis (i.e., assuming idiosyncratic weaknesses and higher uncertainty in future NOI growth due to vacant space). Differences between valuations and sales prices can occur depending on whether appraisers and investors see vacant space as an upside potential related to rental growth or as a downside potential associated with uncertainty. Consistent with our findings on the systematic overvaluation of appraisals in the "Descriptive Statistics" section, Beracha et al. (2019) demonstrate that, on average, the option value of vacant space is overvalued, as buyers may incorporate more risks than sellers aiming to achieve a higher sale price.

Based on Cannon and Cole (2011), we also control for appraisal type and fund type. The authors expect internal appraisals to be less accurate than external appraisals and properties owned by open-end funds to be more accurate than those owned by closed-end funds or separate accounts. This is because internal appraisers tend to be less objective and more likely to smooth appraisals and open-end funds rely on higher appraisal accuracy as investors can trade in and out based on the appraised values, thus allowing informed investors to gain excess returns if the deviation between appraised values and market values is too high (Cannon and Cole, 2011). The authors confirm that appraisal errors are smaller for properties held in open-end funds than properties owned by closed-end funds and separate accounts. However, they find no evidence that external appraisals from an independent third party are significantly lower than internal appraisals. These findings are consistent to our feature importance, as the fund type has a moderate average influence in explaining appraisal errors, while the appraisal type is, on average, the least important feature across all models, implying no significant impact on the predictions of the models.

3.6 Conclusion

Accurate and timely valuations are important to stakeholders in the real estate sector, including authorities, banks, insurers as well as pension and sovereign wealth funds. They

form the basis for informed decisions on financing, developing portfolio strategies and undertaking transactions, as well as for reporting to boards, investors, and tax offices. However, research has shown that, over the past 40 years, commercial real estate appraisals have had a consistent tendency of structural bias and inaccuracy, while lagging true market dynamics (Cole et al., 1986; Webb, 1994; Matysiak and Wang, 1995; Fisher et al., 1999; Cannon and Cole, 2011). While traditional appraisal methods used in the commercial sector have by and large remained the same for decades, statistical learning methods have become increasingly popular. These methods have demonstrated their potential to accurately capture quickly changing market dynamics and complex pricing processes in the residential property sector. However, the transfer of such data-driven valuation methods to commercial real estate faces significant challenges such as data scarcity, heterogeneity, and opaqueness of the models. This poses the question of whether machine learning algorithms can provide material improvement to state-of-the-art appraisal practices in commercial real estate with respect to accuracy and bias of valuations.

Using property-level transaction data from 7,133 properties included in the NCREIF Property Index (NPI) between 1997 and 2021 across the United States, we analyze whether deviations between appraisal values and subsequent transaction prices in the four major commercial real estate sectors (apartment, industrial, office, and retail) contain structured variation that can be further explained by advanced machine learning methods. We find that extreme gradient boosting trees can substantially decrease the variation in appraisal errors across all four property types, thereby increasing accuracy and eliminating structural bias in appraisal values. Improvements are greatest for apartments and industrial properties, followed by office and retail buildings. To clarify where the improvements originate, we employ model-agnostic permutation feature importance and show the features' relative importance in explaining appraisal errors. We find that especially spatial and structural covariates have a dominant influence on appraisal errors, while only one-fourth of the explained variation can be attributed to the time lag between the appraisal and sale date.

The results of our study indicate that current appraisal practices leave room for improvement, which machine learning methods can exploit to provide additional guidance for commercial real estate valuation. The use of such algorithms can make valuations more efficient and objective while being less susceptible to subjectivity and receptive to a wider range of information. Moreover, these methods offer regulatory bodies and central banks the opportunity to quickly analyze and forecast real estate price developments to detect

early signs of price bubbles, stress-test the banking system's stability in shock scenarios or assess the impact of interest rate decisions and rent controls.

Despite their potential for many areas in the industry, machine learning algorithms also encounter limitations that should be carefully considered before their use, as they are not a panacea for all problems in the sector. While algorithms can reduce bias and increase objectivity, they are still developed and trained by humans and thus, remain subject to bias to some extent. In this context, data availability is currently one of the most critical problems for the use of machine learning in the commercial real estate sector, since the complex architectures of the models require substantial amounts of representative training data to produce unbiased and reliable results. Moreover, it should be mentioned that, although the methods can produce accurate predictions of property values by finding patterns between input and output data, they do not consider the laws of economics and thus, cannot justify the rationale behind these patterns or determine causality in the relation between input and output data. This issue is amplified by the lack of inherent interpretability of these models, as they are opaque black boxes that do not provide inference. Although this can be partly circumvented with model-agnostic interpretation techniques, these methods have their very own limitations and pitfalls, and high computational expense can be another limiting factor for their practical implementation.

That said, algorithms can excel humans in quickly learning relationships from large amounts of data, but they have no economic justification and cannot consider aspects that require reasoning. If applied prudently, these methods can add to an enhanced *ex-ante* understanding of pricing processes that may support valuers in the industry and contribute to more dependable and efficient valuations in the future. Yet, we do not believe that machine learning algorithms can substitute the profession of appraisers any time soon due to the restrictions mentioned above as well as regulatory and ethical challenges.

Having demonstrated the potential of machine learning for many areas of the industry, while at the same time raising awareness for the limitations of these techniques, we hope to stimulate further research that contributes to the development of algorithmic approaches in this field. Such research may, for instance, address the exact relations between features and property prices to offer further guidance for the appraisal industry.

3.7 Endnotes

1. Estimations were executed on a standard 1.80 GHz processor with four cores, eight logical processors and eight gigabytes of RAM using a 64-bit Windows operating system. Hyperparameter tuning for optimization of the boosting models required between 25 and 64 hours for each of the four property types, running in parallel. The model including all four property types required 116.5 hours of computation time. Hyperparameter tuning was performed via a grid search procedure with 1,000 evaluations and 10-fold cross-validation. The training and testing of the optimized boosting models via 10-fold cross-validation took between 1.5 and 3.8 minutes for each of the four property types and 7 minutes for the aggregated model.

2. We have considered and tested a random forest regression (i.e., bagging) next to the extreme gradient boosting algorithm (i.e., boosting) and found no material difference in the explanatory power between the boosting and bagging estimators (referring to Table 3.6, $\sigma_{Bagging}$ was on par with $\sigma_{Boosting}$ up to the second decimal place for all models and up to the third decimal place for all models except for office with a deviation of 0.001). However, computation time for bagging was up to twice as long as that for boosting. For reasons of brevity, the results for the bagging estimator were not reported in the paper.

3.8 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.

Beracha, E., Downs, D., & MacKinnon, G. (2019). Investment strategy, vacancy and cap rates. *Real Estate Research Institute*, *Working Paper.*

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. (2010). Predicting house prices with spatial dependence: A comparison of alternative methods. *Journal of Real Estate Research*, *32*(2), 139–160.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Routledge.

Brennan, T. P., Cannaday, R. E., & Colwell, P. F. (1984). Office rent in the Chicago CBD, *Real Estate Economics*, *12*(3), 243–260.

Cajias, M., Willwersch, J., Lorenz, F., & Schaefers, W. (2021). Rental pricing of residential market and portfolio data – A hedonic machine learning approach. *Real Estate Finance, 38*(1), 1–17.

Cannon, S. E., & Cole, R. A. (2011). How accurate are commercial real estate appraisals? Evidence from 25 years of NCREIF sales data. *The Journal of Portfolio Management, 35*(5), 68-88.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Clapp, J. M. (1980). The intrametropolitan location of office activities. *Journal of Regional Science*, *20*(3), 387–399.

Cole, R., Guilkey, D., & Miles, M. (1986). Toward an assessment of the reliability of commercial appraisals. *The Appraisal Journal*, *54*(3), 422–432.

Deppner, J., & Cajias, M. (2022). Accounting for spatial autocorrelation in algorithm-driven hedonic models: A spatial cross-validation approach. *The Journal of Real Estate Finance and Economics,* Forthcoming.

Dunse, N., & Jones, C. (1998). A hedonic price model of office rents. *Journal of Property Valuation and Investment*, *16*(3), 297–312.

Edelstein, R. H., & Quan, D. C. (2006). How does appraisal smoothing bias real estate returns measurement? *The Journal of Real Estate Finance and Economics, 32*(1), 41–60.

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.

Fisher, J. D., & Martin, R. S. (2004). *Income property valuation* (2. ed.). Dearborn Real Estate Education, III.

Fisher, J., Miles, M., & Webb, B. (1999). How reliable are commercial real estate appraisals? Another look. *Real Estate Finance*, Fall 1999, 9–15.

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

Glascock, J. L., Jahanian, S., & Sirmans, C. F. (1990). An analysis of office market rents: Some empirical evidence. *Real Estate Economics*, *18*(1), 105–119.

Hong, J., Choi, H., & Kim, W. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategy Property Management, 24*(3), 140–152.

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.

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.

Koppels, P., & Soeter, J. (2006). The marginal value of office property features in a metropolitan market. *6th International Postgraduate Research Conference*, 553–565.

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.

Levantesi, S., & Piscopo, G. (2020). The importance of economic variables on London real estate market: A random forest approach. *Risks*, *8*(4), 1–17.

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2022). Interpretable machine learning for real estate market analysis. *Real Estate Economics*. Forthcoming.

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). Wiley.

Matysiak, G. A., & Wang, P. (1995). Commercial property market prices and valuations: Analysing the correspondence. *Journal of Property Research*, *12*(3), 181–202.

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.

Mills, E. S. (1992). Office rent determinants in the Chicago area. *Real Estate Economics*, *20*(2), 273–287.

Mooya, M. M. (2016). *Real Estate Valuation Theory: A Critical Appraisal*. Springer.

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

Nappi-Choulet, I., Maleyre, I., & Maury, T.-P. (2007). A hedonic model of office prices in Paris and its immediate suburbs. *Journal of Property Research*, *24*(3), 241–263.

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., & 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.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: Review of valuation methods. *Journal of Property Investment & Finance*, *21*(4), 383–401.

Pai, P.-F., & Wang, W.-C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, *10*(17), 5832.

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.

Real Estate Lending and Appraisals, (2022). 12 Code of Federal Regulations (C.F.R.) § 34.42.

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications*, *171*.

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. Rumelhart, J. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (Vol. 1, pp. 318–362). MIT Press.

Seo, K., Salon, D., Kuby, M. & Golub, A. (2019). Hedonic modelling of commercial property values: Distance decay from the links and nodes of rail and highway infrastructure. *Transportation*, *46*(3), 859–882.

Sing, T. F., Yang, J. J., & Yu, S. M. (2021). Boosted tree ensembles for artificial intelligence based automated valuation models (AI-AVM). *The Journal of Real Estate Finance and Economics*.

Sirmans, S., Macpherson, D., & Zietz, 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.

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance*, *38*(3), 213–225.

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

Webb, B. (1994). On the reliability of commercial appraisals: An analysis of properties sold from the Russell-NCREIF Index (1978–1992). *Real Estate Finance*, *11*, 62–65.

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 Increasing the Transparency of Pricing Dynamics in the U.S. Commercial Real Estate Market with Interpretable Machine Learning Algorithms

4.1 Abstract

Machine learning (ML) algorithms have sown an unprecedented accuracy in estimating house prices, given an abundance of data recorded by multiple listing services and the development of increasingly sophisticated methods. Despite the potential that such methods offer for automated valuation models (AVMs), their adoption in the institutional sector is progressing slowly. In contrast to the residential sector, little is known about the usefulness of data-driven methods in commercial real estate markets as the availability of structured data is limited due to market intransparency and property heterogeneity. Moreover, practitioners and regulators are reluctant to rely on these techniques as their mechanisms are black boxes in the sense that an inherent comprehensibility of their predictions is impeded by the complexity of their architectures. The objective of this study is to propose a holistic framework for the practical use of AMVs in a commercial real estate context that considers both accuracy and interpretability. We train a deep neural network (DNN) on a unique sample of more than 400,000 property-quarter observations from the NCREIF Property Index and perform model-agnostic analysis using "Shapley Additive exPlanations" (SHAP) to provide *ex-post* comprehensibility of the algorithm's prediction rules. In doing so, we furthermore assess to which extent the inner workings of the DNN follow an economic rationale and set out how the proposed methods can add to the understanding of pricing processes in institutional investment markets. By addressing the caveats and illustrating the potential of ML in the field of commercial real estate, this article represents another important pillar in the practical use of AVMs.

Keywords: Automated valuation models, Commercial real estate, Interpretable machine learning

Acknowledgments: This article received generous support and funding from the Real Estate Research Institute (RERI), a part of the Pension Real Estate Association (PREA). The National Council of Real Estate Investment Fiduciaries (NCREIF) kindly provided the data. The authors sincerely thank Jeffrey Fisher for data access, and the RERI mentors James Chung, Heidi Learner, Mark Roberts and Timothy Savage for their helpful comments.

4.2 Background

Estimating real estate prices and identifying relevant price determinants remains complex due to the inherent heterogeneity of properties and the diversity of factors that influence their values. As stated by Quan and Quigley (1991), market mechanisms are obfuscated by "[…] a noisy signal, reflecting incomplete information as well as the conditions of sale," given that real estate markets are illiquid, opaque and individual agents in the market are only infrequently engaged in transactions. Appraisers must extract meaningful information (i.e., the signal) from irrelevant data (i.e., the noise) using their expert knowledge about the market, based on their experience observing past transactions. Consequently, pricing processes must be disentangled based on limited information and subjective judgments of price determinants that a valuer considers relevant, resulting in imprecise and biased valuations (Dunse and Jones, 1998; Cannon and Cole, 2011).

This gave rise to hedonic pricing models introduced by Rosen (1974) as the prevalent framework to analyze the mechanisms behind property pricing more objectively from an econometric point of view. Parametric hedonic models, such as those proposed by Mills (1992), Sirmans and Guidry (1993), or Lockwood and Rutherford (1996), utilize linear regression methods to estimate property prices based on intrinsic property characteristics (e.g., location, size, amenities). Literature has demonstrated the hedonic models' efficiency and ease of interpretability in revealing relevant property price determinants.

However, these models are built on strict assumptions which are unlikely to hold and require a fixed additive functional form between the property value and the explanatory variables that needs to be specified *a-priori.* This entails a high risk of misspecification. As Dunse and Jones (1998) explained, hedonic prices may vary across space and time and can thus not be assumed to be constant. Other concerns refer mainly to the non-linearity of pricing processes that cannot be adequately captured with linear models. Studies by Grether and Mieszkowski (1974), Do and Grudnitski (1993), and Goodman and Thibodeau (1995) identify significant non-linearities between property prices and the building age as well as the square footage, demonstrating that complex relationships between property prices and features cannot be reduced to a single, invariant beta coefficient.

As data becomes more readily available and artificial intelligence (AI) continues to advance, industry and academia have witnessed a shift towards more adaptable machine learning (ML) techniques for determining property values. This shift has become evident in automated valuation models (AVMs), which have gained importance in the sector, particularly in residential real estate, given the increased flexibility in the underlying models. In the literature, ML-based AVMs have repeatedly demonstrated unprecedented accuracy

in their predictions. They also do not require judgment concerning the model's functional form as they are designed to autonomously find complex non-linear relationships in the data to optimize model fit.

However, the adoption of ML in industry, and particularly in the institutional sector, is facing critical issues. First, ML techniques rely on large amounts of data to produce reliable and consistent results, as demonstrated by Worzala et al. (1995). In contrast to the residential domain, data availability is still limited in the commercial sector, which is particularly problematic due to the high heterogeneity of commercial property types (Deppner et al., 2023). Second, the models are criticized for lacking an economic justification and do not foresee any form of intrinsic interpretability (e.g., Din et al., 2001; McCluskey et al., 2013; Valier, 2020). This refers to the fact that these models are purely data-driven, allowing them to make predictions from any combination of data (Rico-Juan and Taltavull de La Paz, 2021), while their complex and opaque architectures impede understanding of how the algorithm arrived at a particular valuation, and how the input factors have affected the outcome. This hampers the comprehensibility of the models and prohibits drawing inferences on price determinants, making it difficult for practitioners to trust and rely on AVMs, particularly given that regulators and authorities demand transparency in estimating market values.

The current state of research suggests three ways to address this. The first is to reduce the complexity of the applied models to such an extent that their interpretability is preserved. However, this makes the models more sensitive to changes in the data and increases the tendency of overfitting, resulting in poor out-of-sample performance (Kok et al., 2017; Pace and Hayunga, 2020; Lorenz et al., 2022). Second, ML can be used to provide constructive criticism, such as in the variable selection, model specification (e.g., Yoo et al., 2012; Perez-Rave et al., 2019), or model selection (e.g., Pace and Hayunga, 2020), which can help to improve upon traditional models. However, this means giving up the flexibility and accuracy of ML models for the sake of interpretability. The third alternative is to apply model-agnostic interpretation techniques that can decipher the black box of ML models, thus enabling *ex-post* interpretability while maintaining accuracy and precision, as shown by Levantesi and Piscopo (2020), Rico-Juan and Taltavull de La Paz (2021), Lorenz et al. (2022) as well as Potrawa and Teterava (2022).

This study aims to expand upon this discussion by proposing a novel and comprehensive framework for utilizing AVMs in commercial real estate that balances both precision and comprehensibility. To achieve this, we train four deep neural networks (DNNs) on a large data sample comprising over 400,000 property-quarter observations from the asset sectors

apartment, industrial, office and retail. We then apply model-agnostic analysis using "Shapley Additive exPlanations" (SHAP) to provide clear insight into the prediction rules of the algorithms. In doing so, we further assess to which extent the inner workings of the DNNs follow economic principles. We also set out how the proposed methods can add to a deeper and more nuanced understanding of pricing mechanisms in institutional investment markets by revealing non-linear and three-dimensional relationships in the value drivers of commercial real estate.

The study's contributions are relevant and timely for academia and practice for several reasons. While we do not believe that AVMs have developed to the point where they can substitute manual appraisers in the foreseeable future, the underlying technology still exhibits high disruptive potential. It is likely to reshape the multi-billion-dollar valuation industry in the future (Kok et al., 2017). Especially in the commercial domain, where valuations are more complex and need to be executed frequently, these techniques can generate valuable insights to support data-driven decision-making and thus leverage efficiency in both markets and business processes by increasing the speed and scale of valuations, reducing the cost of transactions and, ultimately, increasing transparency in pricing processes. Market participants that incorporate such technologies into their business processes earlier than their competitors will be able to streamline their processes and gain a competitive edge.

4.3 Data

The National Council of Real Estate Investment Fiduciaries (NCREIF) provided the data for this study. The principal study data comprises quarterly, property-level observations of all properties included in the NCREIF Property Index (NPI) from the first quarter of 1978 to the first quarter of 2021. The NPI is the oldest and most widely followed commercial real estate investment index in the United States. It covers institutionally owned commercial real estate properties across the asset sectors apartment, hotel, industrial, office and retail. The properties included in the index fluctuate over time as properties enter the database upon purchase and leave the database upon sale. This constitutes an initial unbalanced sample of 648,098 property-quarter observations across 30,254 individual properties, for which we record the corresponding market values, a series of structural and physical attributes, and cash flows. Due to limited data availability, we excluded non-operating properties and hotels from the initial sample.

We account for missing and erroneous data as follows. Observations with market values, square footage and construction years reported as less than or equal to zero are regarded as data errors and are dropped. Likewise, observations with occupancy rates taking values

below zero or higher than one are removed. Furthermore, observations with missing values for the square footage, the construction year, the occupancy rate, the net operating income (NOI), the capital expenditures (CapEx), and the property subtype were omitted, as these represent the main explanatory variables from the raw NCREIF dataset. After scaling market values, NOI, and CapEx by the property's square footage, we note that the remaining errors and anomalies in the data seem concentrated at the tails of the market values per square foot distribution. For this reason, we follow Calainho et al. (2022) and cut off the lower and upper percentile of the distribution for each property type.

We subsequently enrich the cleaned data with a set of new variables. First, we calculate the building age as the difference between the valuation date and the construction date, as well as the cumulative sum of a property's capital expenditures scaled by square footage as a proxy for building quality. We also note that NOIs can fluctuate materially over the holding period and in individual quarters. Since the average property in our sample has a five-year holding period, we use the eight-quarter moving average of the properties' NOIs as a proxy for stabilized income.

Table 4.1: Clustering of POIs

Category	POI	Source
Public Transport	Bus Station	Google
	Subway Station	Google
	Light Rail Station	Google
	Train Station	Google
	Public Transport	OSM
Negative Externalities	Prison	OSM
	Graveyard	OSM
	Gas Station	Google, OSM
Food Establishments	Restaurant	Google, OSM
	Cafe	Google, OSM
Healthcare Provider	Pharmacy	Google, OSM
	Doctor	Google
Retail Stores	Shopping Mall	Google, OSM
	Department Store	Google, OSM
Food Stores	Supermarket	Google, OSM
	Convenience Store	Google, OSM
Nightlife Venue	Bar	Google, OSM
	Nightclub	Google, OSM
Educational Institutions	Kindergarten	OSM
	School	Google, OSM
Cultural Institutions	Museum	OSM
	Attraction	OSM
Service Establishments	Bank	Google, OSM
	Post Office	Google, OSM
Fitness	Gym	Google, OSM
	Fitness Centre	OSM
Park	Park	Google, OSM

Notes: This table presents the POI categories and how they were clustered. Google corresponds to the POIs sourced from the Google Places API and OSM corresponds to POIs sourced from Open Street Maps.

As demonstrated repeatedly in the literature, location is an important determinant of real estate values. We geocode our sample using the property addresses to retrieve the distances to relevant points of interest (POIs). Around 12.1% of the addresses could not be geocoded because of missing or incomplete addresses, so we omitted those observations. For the remaining properties, we source a set of relevant POIs that are expected to cause either a premium or a discount to their surrounding area. For optimal data coverage, we use both Google Places and Open Street Maps (OSM) to retrieve the data and calculate the shortest distance from each property to the respective POIs. We subsequently cluster POIs that are similar into categories. This helps avoid missing data and reduce the dimensionality of the regressor matrix, making the models more interpretable and more efficient. Table 4.1 provides a summary of the POI clusters.

In addition, we collect macroeconomic data to control for market cycles and varying economic conditions. This includes the ten-year government bond yield as well as the four-quarter percentage change in the gross domestic product (GDP) at the state level retrieved

Table 4.2: Descriptive Statistics of Numerical Variables
All Property Types (N = 402,490)

Variable	Unit	Mean	Sd	Min	1st Q.	Median	3rd Q.	Max
Market Value	[$/SqFt]	189.54	198.54	18.57	71.60	125.40	229.63	2,634.53
SqFt	[k]	283.08	371.09	1.50	109.50	200.64	341.25	22,119.56
Building Age	[Years]	20.77	16.78	0.00	10.00	17.00	27.00	156.00
Occupancy	[%]	0.92	0.12	0.00	0.90	0.96	1.00	1.00
NOI	[$/SqFt]	2.62	2.45	-48.58	1.13	1.90	3.43	73.74
NOI Stabilized	[$/SqFt]	2.60	2.28	-19.69	1.14	1.89	3.39	56.26
CapEx	[$/SqFt]	0.77	2.91	0.00	0.00	0.14	0.59	311.02
CapEx Cumulative Sum	[$/SqFt]	13.20	40.51	0.00	0.41	3.34	11.65	1,802.37
Longitude	[°]	-96.14	17.66	-158.12	-117.53	-93.24	-80.36	-68.75
Latitude	[°]	36.85	5.27	19.63	33.58	37.48	40.72	61.56
Public Transport	[km]	1.70	2.00	0.00	0.32	1.06	2.29	12.99
Negative Externalities	[km]	0.76	0.59	0.00	0.36	0.62	1.00	7.95
Food Establishments	[km]	0.36	0.44	0.00	0.07	0.22	0.50	7.20
Healthcare Provider	[km]	0.42	0.65	0.00	0.08	0.22	0.51	11.93
Retail Stores	[km]	0.92	1.05	0.00	0.24	0.61	1.23	12.93
Food Stores	[km]	0.61	0.55	0.00	0.21	0.46	0.84	8.45
Nightlife Venue	[km]	0.78	0.95	0.00	0.20	0.51	1.06	12.36
Educational Institutions	[km]	0.49	0.52	0.00	0.17	0.35	0.63	8.25
Cultural Institutions	[km]	2.12	1.96	0.00	0.77	1.65	2.84	12.96
Service Establishments	[km]	0.70	0.74	0.00	0.18	0.47	1.00	8.16
Fitness	[km]	0.69	0.84	0.00	0.19	0.44	0.90	12.85
Park	[km]	0.79	0.84	0.00	0.30	0.59	1.00	12.85
GDP yoy	[%]	0.02	0.03	-0.11	0.01	0.02	0.04	0.22
Gov. Bond Yield	[%]	0.03	0.02	0.01	0.02	0.03	0.04	0.08
Construction Cost yoy	[%]	0.03	0.05	-0.10	0.01	0.04	0.05	0.20
Employment yoy	[%]	0.01	0.03	-0.50	0.00	0.01	0.03	1.10
Market Cap Rate qoq	[%]	0.06	0.01	0.04	0.05	0.06	0.07	0.10
Market Vacancy qoq	[%]	0.08	0.03	0.03	0.06	0.07	0.10	0.17
Market NOI Growth qoq	[%]	0.01	0.03	-0.32	-0.01	0.01	0.02	0.14

Notes: This table presents the summary statistics of numerical features.

Table 4.3: Descriptive Statistics of Categorical Variables
All Property Types (N = 402,490)

Variable	n	Percent
Property Type		
... Apartment	88,442	21.97%
... Industrial	151,109	37.54%
... Office	99,271	24.66%
... Retail	63,668	15.82%
Property Subtype		
... Garden	55,566	13.81%
... High-rise	26,889	6.68%
... Low-rise	5,987	1.49%
... Research and Development	6,049	1.50%
... Flex Space	17,054	4.24%
... Manufacturing	729	0.18%
... Other	2,328	0.58%
... Office Showroom	440	0.11%
... Warehouse	124,509	30.93%
... Central Business District	23,114	5.74%
... Suburban	76,157	18.92%
... Community Center	17,757	4.41%
... Theme/Festival Center	167	0.04%
... Fashion/Specialty Center	2,951	0.73%
... Neighborhood Center	23,511	5.84%
... Outlet Center	113	0.03%
... Power Center	6,776	1.68%
... Regional Mall	4,843	1.20%
... Super-Regional Mall	4,319	1.07%
... Single-Tenant	3,231	0.80%
Market Cycle		
... 1991Q1-1994Q1 (Gulf Crisis)	6,324	1.57%
... 1994Q2-2001Q3	47,506	11.80%
... 2001Q4-2002Q2 (Dotcom Crisis)	8,310	2.06%
... 2002Q3-2008Q1	80,138	19.91%
... 2008Q2-2010Q1 (Subprime Crisis)	35,742	8.88%
... 2010Q2-2020Q1	201,418	50.04%
... 2020Q2 (Covid-19 Pandemic)	5,565	1.38%
... 2020Q3-2021Q1	17,487	4.34%

Notes: This table presents the summary statistics of categorical features.

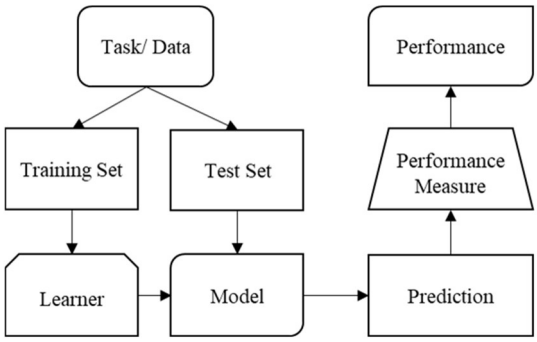
from the database of the Federal Reserve Bank of St. Louis, the four-quarter percentage change in construction costs by region retrieved from the U.S. Census Bureau, and the four-quarter percentage change in employment at the county-level retrieved from the U.S. Bureau of Labor Statistics. We also collect quarterly real estate market data by property type from NCREIF: market value cap rates, market vacancy rates and market rental growth rates. Furthermore, we include a dummy indicator for different market cycles during the sample period to better control for shocks and the effect of cyclical movements in the overall market. Market cycles are defined as periods of consecutive positive (i.e., rising markets) or negative (i.e., falling markets) quarterly capital appreciation returns derived from the NCREIF Property Index (NPI).

In the last step, we exclude CBSA codes with fewer than ten properties of the same property type to prevent overfitting. The final study sample consists of 402,490 quarterly market value observations across 18,286 individual properties and is balanced across 30 explanatory variables that are presented in the summary statistics in Table 4.2 and Table 4.3. Missing and erroneous data seem concentrated in the early years of the initial sample, as the final study data ranges from the first quarter of 1991 to the first quarter of 2021, covering 30 years.

4.4 Methodology

The basic workflow behind machine learning algorithms is illustrated in Figure 4.1 following Lang et al. (2019). A supervised ML model works by learning patterns from the data and improving on past experiences (i.e., model errors). This process starts by dividing the data into a training and a test subsample. The starting point of each ML model is training a selected algorithm (i.e., learner) on the subjective training sample. Such algorithms learn patterns from the training data to create prediction rules. Based on previous model errors, these rules are assessed and refined in an iterative process. Once the out-of-sample performance of the model is sufficient, it can be applied to an independent test dataset (i.e., unseen or future data) to make predictions.

Figure 4.1: General Overview of the Machine Learning Process



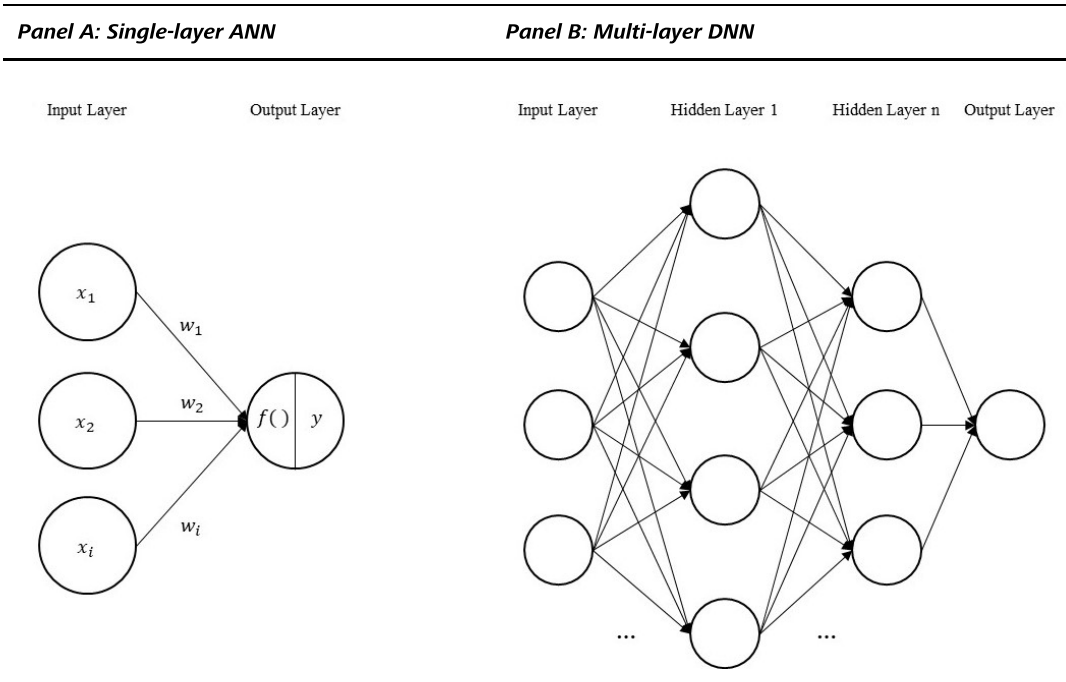
Notes: This figure depicts the basic workflow of a machine learning algorithm.

To understand pricing processes in commercial real estate markets, it is crucial that the selected models (i.e., learners) and the resulting prediction rules adequately capture relationships in the data but are still generalizable enough to predict well out-of-sample. Studies that compare different learners show that particularly artificial neural networks (ANNs) produce robust and accurate predictions when applied in combination with sufficient data (e.g., Peterson and Flanagan, 2009; Zurada et al., 2011; Antipov and Pokryshevskaya, 2012; Baldominos et al., 2018; Mayer et al., 2019; Hu et al., 2019).

4.4.1 Machine Learning Approach – Artificial Neural Networks

An ANN imitates the structure and function of the human brain. It is created of many artificial neurons, called nodes, that are interconnected in layers to process information and learn from experience. In many ways, this corresponds to how the human brain learns from experience and adapts its expectations. When new information is processed, the actual outcome of an event is compared with the expected (i.e., predicted) outcome, which is fed by knowledge and experience. An error signal is generated in case of discrepancies between the expected and the actual outcome. The brain adjusts the strength of the connections between its neurons (i.e., synapses) to better represent the new information. The stronger a synapse develops, the more likely it is that connected neurons will fire in response to an incoming signal released by other neurons. Eventually, our final predictions and expectations result from how stimulations are translated to chemical signals and propagated through the network of neurons in our brain. In this way, the adjustment of the connections marks the learning process such that previous errors are mitigated, and the structure is constantly adapted to new information.

Figure 4.2: Structure of Neural Networks



Notes: This figure depicts the conceptual structure of a neural network.

Analogously, an ANN learns by adjusting the weights of the connections between each node in an iterative process. The optimal model fit is found by minimizing a loss function that measures the distance from the actual to the predicted values, thus improving the accuracy of the network's prediction.

In its simplest form, an ANN consists of only one input and one output layer (i.e., single-layer ANN) and uses a linear activation function f, as depicted in Figure 4.2, Panel A. This type of network can be compared to a linear regression. The bias b and the weights w_i of the input values x_i represent the intercept and the beta coefficients in an ordinary least squares (OLS) regression model and formulate the prediction y as exhibited in equation (4.1).

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{4.1}$$

The more complex the input, the more sophisticated the structure becomes to adequately process the information. This is achieved by adding more hidden layers with multiple nodes and choosing other than linear activation functions in the model. This will introduce interaction effects and non-linearity to the model and is referred to as a deep neural network (DNN), as depicted in Figure 4.2, Panel B.

4.4.2 Model Agnostic Analysis – Shapley Additive Explanations

Interpretable machine learning (IML) methods are model-agnostic techniques for explaining and interpreting opaque ML models to achieve *ex-post* transparency. This facilitates understanding of how and why the model produces a specific outcome. One such technique is named "Shapley Additive exPlanations" (SHAP), introduced by Lundberg and Lee (2017). It is conceptually based on Shapley values, a method used in coalitional game theory to determine the marginal contributions of each player to the outcome of a collaborative game (Shapley, 1953). Transferred to an ML context, Shapley values can be thought of as the average marginal contribution of a feature (i.e., "player") in an ML model (i.e., "game") on its prediction (i.e., "outcome"), as described by Molnar (2020). Shapley values are derived by repeatedly simulating different combinations of input features (i.e., "coalitions") and assessing how changes to the coalitions correspond to the final model predictions. This is done for each possible coalition in the model, so that a feature's impact on the model prediction is eventually calculated as the average marginal contribution to the overall model score.

4.4.3 Model Estimation

We estimate a separate DNN for each property type due to the peculiarities of the different sectors. The process of model estimation can generally be divided into two parts. The first involves data transformation, training, and optimization of the model. The second involves out-of-sample performance testing to ensure the generalizability of the results.

First, the initial sample is split into three subsets: 60% training data, 20% validation data and another 20% test data. Subsequently, all numerical explanatory variables are z-score standardized. Each model is trained as a sequential feedforward DNN with a variable number of hidden layers and neurons. Bayesian optimization is used to determine the best combination of hyperparameters such as the number of layers, neurons, dropout and learning rate. Subsequently, the model with the best hyperparameter combination is trained on the whole training set (i.e., training and validation data aggregated), and out-of-sample performance is assessed on the remaining 20% test subsample. To evaluate the performance of the DNN in the application context, we estimate a linear regression model as a point of reference. The estimation and performance evaluation of the DNN is then complemented using SHAP. This facilitates the interpretability and comprehensibility of the model's prediction rules.

4.4.4 Performance Evaluation

Model performance is assessed using the mean absolute percentage error (MAPE), the mean percentage error (MPE), the mean absolute error (MAE), the mean squared error (MSE), the root mean squared error (RMSE) and the coefficient of determination (R^2). The error buckets (PE10) and (PE20) show the proportion of absolute percentage errors below 10% and 20%, respectively. MAPE and MAE are direct measures of accuracy (i.e., absolute distance). MSE and RMSE are used to assess the models' performance for exceedingly high values in the test data as high errors are penalized more (i.e., squared distance). MPE measures the biasedness of the model (i.e., whether the model's predictions generally tend to be higher or lower than the actual values), and R^2 is utilized to measure overall model fit. Lastly, the error buckets show how reliable the models are in relation to certain error thresholds (i.e., errors between 10% to 20% is commonly considered a tolerable range in valuation practices).

4.5 Empirical Results

This section features the empirical results of the analysis. First, model performance in estimating market values is assessed. Concerning the research objective, we discuss the results from the model-agnostic analysis with SHAP and draw conclusions on the features' functional relationships with the dependent variable.

4.5.1 Model Performance

Table 4.4 depicts the out-of-sample performance metrics of the DNN and the OLS, respectively. The DNN is highly accurate in estimating market values per square foot, with

the MAPE between 9.29% and 10.98% and the corresponding MAE between 7.56 and 25.54 dollars per square foot. The MSE and RMSE show that the apartment, office, and retail models generally produce higher errors that are penalized stronger than in the industrial model, as market values are generally lower in this sector. Across all property types, over 85% of the market value predictions of the DNN are estimated within a MAPE of 20%. In the OLS estimation, only 55% of predictions fall within this range. The OLS generally shows a considerably lower model fit than the DNN.

Table 4.4: Model Performance Metrics

Method	R^2	MAPE	MPE	MAE	MSE	RMSE	PE10	PE20
Unit	[%]	[%]	[%]	[$/SqFt]	[$/SqFt]	[$/SqFt]	[%]	[%]
	Panel A: Apartment							
OLS	0.77	0.26	0.04	43.61	7,959.58	89.22	0.31	0.55
DNN	0.97	0.09	-0.03	18.88	1,177.55	34.32	0.65	0.91
	Panel B: Industrial							
OLS	0.73	0.24	0.06	17.53	659.82	25.69	0.30	0.56
DNN	0.95	0.11	0.04	7.56	128.04	11.32	0.62	0.87
	Panel C: Office							
OLS	0.76	0.32	0.07	64.99	9,351.87	96.71	0.26	0.48
DNN	0.96	0.11	-0.03	25.54	1,490.37	38.61	0.58	0.87
	Panel D: Retail							
OLS	0.81	0.30	0.07	62.19	15,125.86	122.99	0.31	0.54
DNN	0.97	0.10	0.03	22.94	2,139.41	46.25	0.67	0.88

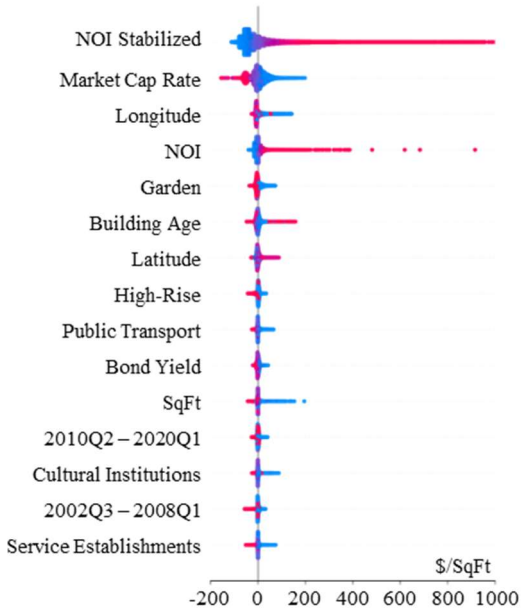
Notes: This table reports the performance measures of the linear models and the DNN. R^2: coefficient of determination, MAPE: mean absolute percentage error, MPE: mean percentage error, MAE: mean absolute error, MSE: mean squared error, RMSE: root mean squared error, PE10 and PE20: error bucket of estimates within 10% and 20% of the true value respectively. Absolute values are reported in dollars per square foot.

4.5.2 Global Model Interpretability

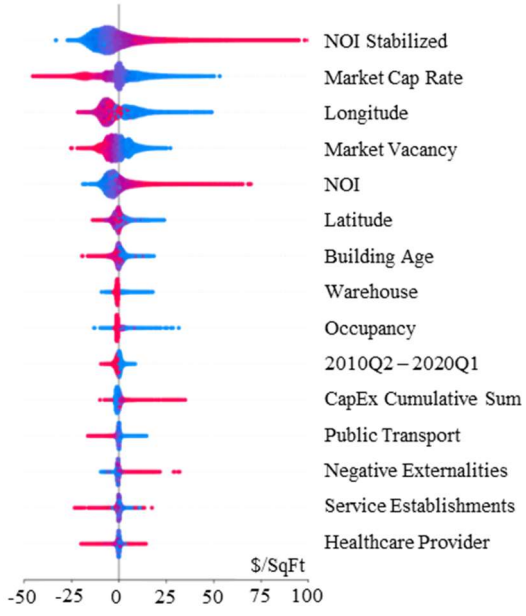
In traditional property valuation, market values of income-generating properties are determined with the income approach, which consists of two primary elements, rental income and the capitalization rate. However, alternative methods such as the sales comparison approach and the cost approach consider various other factors, including locational, physical, financial, and macroeconomic characteristics (see Pagourtzi et al., 2003) that are not necessarily reflected in the income approach. Our research focuses on a data-driven methodology grounded in economic theory. We use a comprehensive set of physical and structural property attributes, neighborhood characteristics, macroeconomic and real estate market indicators, and cash flows to capture all relevant price-determining attributes.

Figure 4.3: SHAP Summary Plot (Top 15 Features)

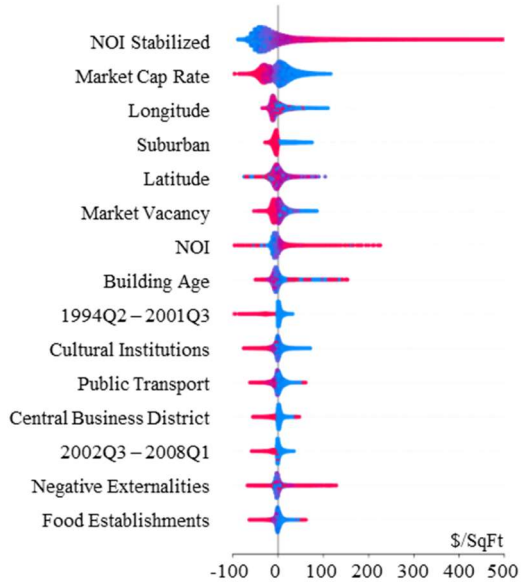
Panel A: Apartment



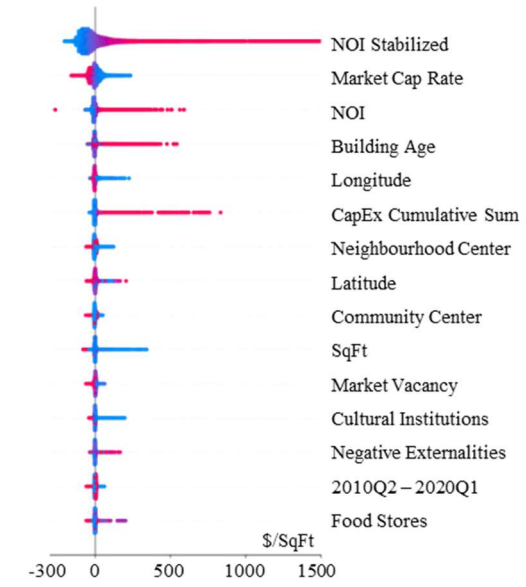
Panel B: Industrial

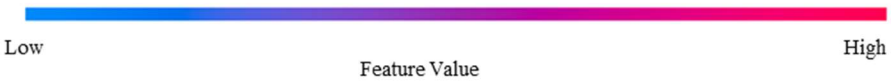


Panel C: Office



Panel D: Retail





Notes: This figure depicts the feature importance for the top 15 (i.e., most important) features. The x-axis depicts the feature impact on the estimated market value in dollars per square foot. The y-axis shows the top 15 features by property type. Color indicates whether the contribution of a feature to the final prediction is positive or negative.

To review the relations of employed features in our models, we analyze the features' marginal influences that are presented in Figure 4.3. In the respective summary plots, three dimensions can be explored, with the features arranged in a specific order that reflects their relative importance in the model predictions. The stabilized net operating income appears to be the most crucial feature for all sectors. The plot also illustrates the characteristics of the features in the second and third dimensions by indicating whether the contribution of a feature to the final prediction is positive or negative and which value the feature takes (as indicated by color).

We use the SHAP summary plot to identify the critical value drivers and relate them to their economic meaning to bridge the gap between economic theory and the data-driven machine learning approach. It is important to note that our models do not incorporate inferential assumptions that can determine causal relationships. That is, the significance of the features is determined solely by the statistical relationships that the model identifies. Ideally, the statistical relationships determined by the model are consistent with economic principles and thus contribute to understanding price formation process in commercial property markets. As Lorenz et al. (2022) discuss, a feature importance plot can be utilized to evaluate the relevance of variables for a given predictive task. This method allows insight into the reliability of an algorithmic hedonic model and its ability to capture a plausible understanding of the economic context.

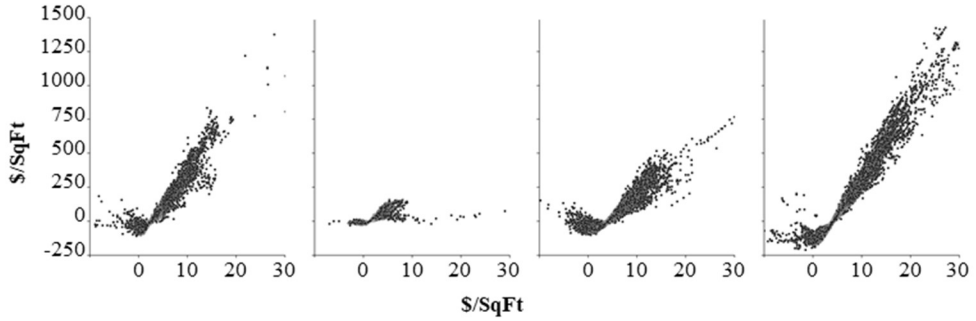
In line with economic theory, Figure 4.3 depicts the stabilized NOI and the market capitalization rate as the most crucial feature in the prediction process of the model across all property types. Furthermore, the location expressed by the geo-coordinates, the physical condition proxied with building age, and the current NOI appear to be equally important across all asset sectors and strongly influence the model predictions. Moreover, it becomes clear that each property sector has individual value drivers, such as the presence of a garden in the case of apartment properties or the location of an office building in the central business district (CBD). As alluded to previously, SHAP can be used to draw conclusions about the functional relationship between explanatory variables and the dependent variable. This is particularly beneficial in real estate valuation, where understanding pricing processes is paramount. Figure 4.4 shows the relationships of four explanatory variables with SHAP partial dependence plots.

Figure 4.4, Panel A depicts the dependence plots of stabilized NOI and its impact on the market value prediction. A positive linear relationship for values greater than zero can be observed across all asset sectors, as expected market values increase with an increasing stabilized NOI. A negative stabilized NOI shows a non-linear pattern that will be interpreted

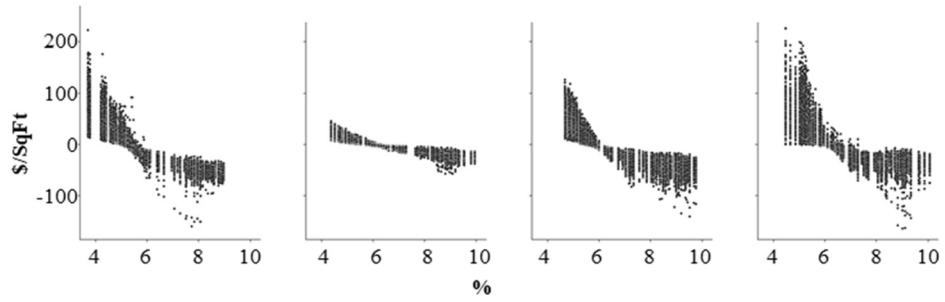
Figure 4.4: SHAP Partial Dependence (1)

Apartment	Industrial	Office	Retail

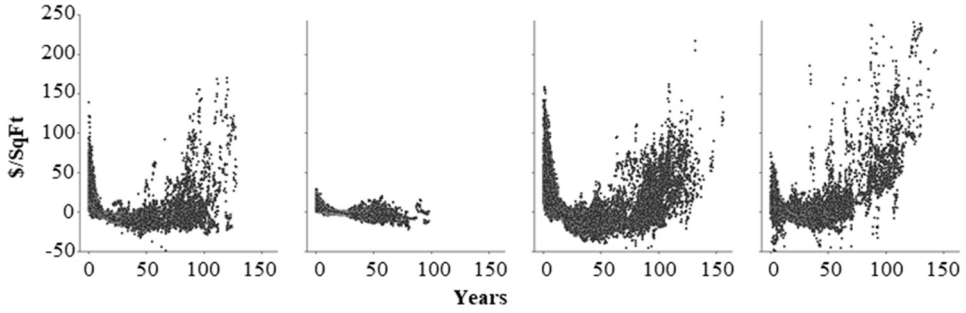
Panel A: NOI Stabilized



Panel B: Market Cap Rate



Panel C: Building Age

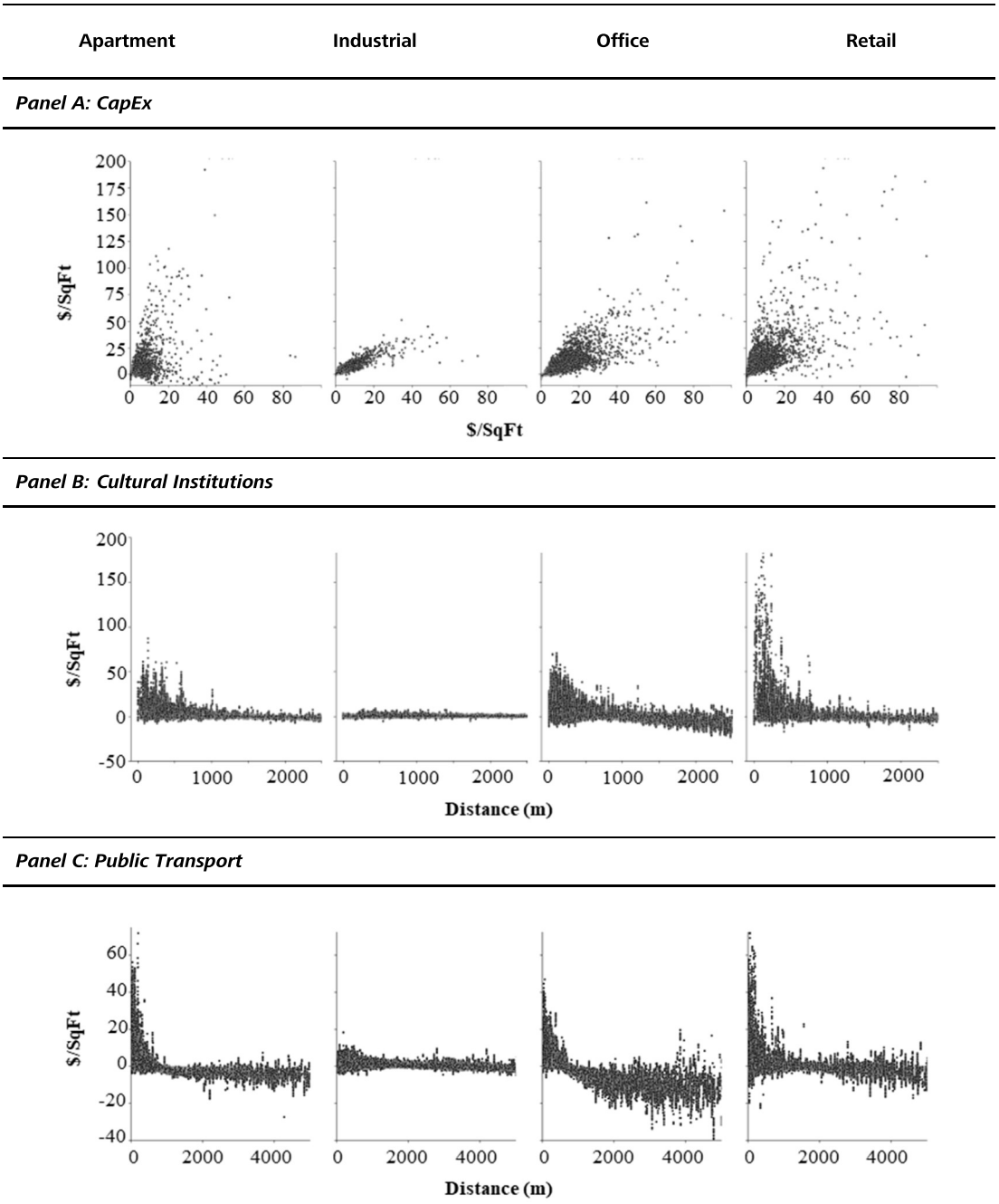


Notes: This figure depicts the partial dependence of selected features. The x-axis shows the values of the respective features. The y-axis shows the model impact on the estimated market value in dollars per square foot.

with further analysis below. The second most important feature in the prediction of market values is the market capitalization rate. Figure 4.4, Panel B depicts the relation of this feature to the impact on the market value, and it takes the expected relationship in all four property types. As the capitalization rate is a proxy of risk and return in the real estate market, market values generally decrease with increasing capitalization rates. Notably, the plot for industrial properties deviates from the other property types, but this is due to the mean value of industrial properties in the sample being significantly smaller. Concerning a property's physical condition, we focus on the impact of property age. Lorenz et al. (2022)

show that, in line with economic theory, the age of an apartment exhibits a U-shaped pattern; that is, the newest and oldest buildings generate the highest rents. In Figure 4.4, Panel C, we observe that this is also the case for the apartment sample and the office and retail properties. This U-shape seems to be less pronounced for industrial properties. The plot of industrial properties generally shows a lower building age, which can be attributed to the nature of heavy industry use and the limited usability by third parties.

Figure 4.5. SHAP Partial Dependence (2)



Notes: This figure depicts the partial dependence of selected features. The x-axis shows the values of the respective features. The y-axis shows the model impact on the estimated market value in dollars per square foot.

While Figure 4.4 shows features with similar impacts across the four property types, Figure 4.5 depicts features that behave differently concerning market values across the property types. Figure 4.5, Panel A illustrates the relationship between CapEx and its impact on the expected market value. Generally, CapEx increases market values, whereby the marginal effect varies across property types. A dollar of CapEx per square foot appears to have the most decisive impact on the market value per square foot for apartment properties. In contrast, industrial properties exhibit the lowest marginal effect.

Figure 4.5, Panel B depicts the impact of proximity to a cultural institution (i.e., museum, entertainment facilities or attractions) on the model's prediction of the market value, Interestingly, retail properties close to cultural institutions experience a higher premium than all other property types. This could be related to increased pedestrian flows generated by cultural institutions, which drive market values of retail properties. In contrast, the proximity to cultural institutions does not affect industrial properties' market values.

Figure 4.5, Panel C shows the impact of a property's proximity to public transport on the market value. Whereas the impact seems low for industrial properties, retail, apartment and office properties show strong relations to this POI. Interestingly, retail and apartment properties experience a positive impact on the market values when near public transport but barely see negative impacts when public transport is located farther away. However, in the office sector, public transport seems particularly interesting as larger distances are related to negative impacts on the predictions. Hence, there seems to be a sweet spot up to which the presence of POIs matters.

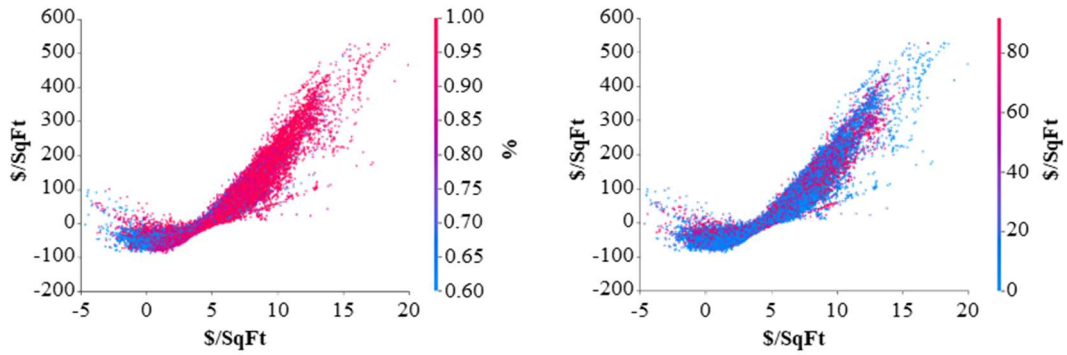
Figures 4.4 and 4.5 present multiple instances where a feature can take values that result in both a positive and negative model impact. The factors contributing to such attributions can be examined more closely with the interaction effects for the respective variable. For example, the stabilized NOI in Figure 4.4, Panel A shows negative values leading to both positive and negative model impacts. We expect such behavior to be related to structural characteristics of the related properties and thus analyze the interaction effects of the stabilized NOI with both capital expenditures and occupancy, illustrated in Figure 4.6.

Panel A of Figure 4.6 displays the interaction effect between occupancy and stabilized NOI, while Panel B shows the interaction effect between cumulative CapEx and stabilized NOI. The blue color on the graphs indicates low interaction feature values, while the red color indicates high interaction feature values. We observe that in cases where negative NOI contributes negatively to the model prediction and thus leads to the expectation of lower market values, both occupancy and CapEx tend to be low, indicating high vacancy and potentially lower building quality compared to other properties. On the other hand,

observations with negative NOI that contribute to the model's prediction positively are characterized by higher occupancy and high CapEx that increase the quality of a building and, thus, its value.

Figure 4.6: SHAP Partial Dependence with Interaction Effects (Financial)

Panel A: NOI Stabilized vs. Occupancy	*Panel B: NOI Stabilized vs. CapEx Cumulative Sum*

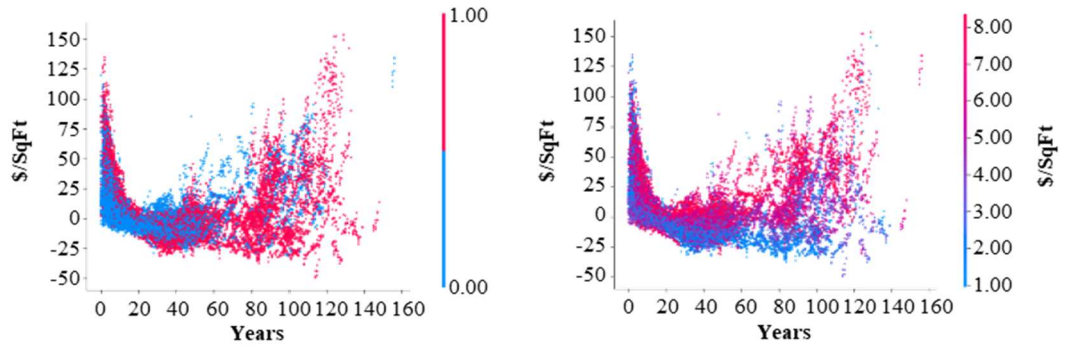


Notes: This figure depicts interaction effects in the partial dependence of the stabilized NOI. The x-axis shows the stabilized NOI in dollars per square foot. The left y-axis shows the model impact on the estimated market value in dollars per square foot. The right y-axis shows the respective feature value of the interaction feature.

In Figure 4.7 we analyze the observed U-shaped pattern in the building age by inspecting interaction effects with both location (Panel A) and income (Panel B). In suburban areas, the building age generally shows a negative relationship, as seen in Figure 4.7, Panel A. That is, older properties in suburban areas tend to have lower market values. From Figure 4.7, Panel B, we can deduce that properties for which high building ages are positively related to market value and high NOIs tend to be clustered in CBDs. Osland (2010) summarizes the main rationale behind early land economic theories and concludes that overcoming space in any form is costly and, therefore, needs to be economized. Thus, the

Figure 4.7: SHAP Partial Dependence with Interaction Effects (Structural)

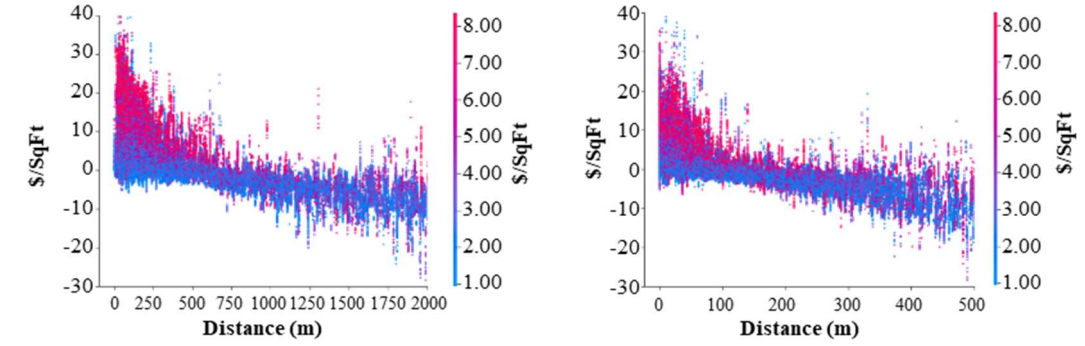
Panel A: Building Age vs. CBD	*Panel B: Building Age vs. NOI Stabilized*



Notes: This figure depicts interaction effects in the partial dependence of the building age. The x-axis shows the building age in years. The left y-axis shows the model impact on the estimated market value in dollars per square foot. The right y-axis shows the respective feature value of the interaction feature.

highest centrality in the CBD of a city creates high demand that generally leads to high market values. Of course, the centrality of a property cannot only be described by its location in the CBD or a suburban area. It can also be formulated as the sum of multiple characteristics that define the location of a property. Can (1992) mentions neighborhood effects that refer to characteristics that drive demand for real estate in a specific location (i.e., neighborhood) and should materialize in the price function.

Figure 4.8: SHAP Partial Dependence with Interaction Effects (POIs)



Notes: This figure depicts interaction effects in the partial dependence of selected POIs. The x-axis shows the distance to the respective POI in meters. The left y-axis shows the model impact on the estimated market value in dollars per square foot. The right y-axis shows the respective feature value of the interaction feature.

Such trends are not only seen for the market value but generally for the price level when observing the interaction effect of the stabilized NOI and the proximity to public transport or food establishments. This is demonstrated in Figure 4.8 – the larger the distances to public transport or food establishments, the lower the stabilized NOI that is paid for a property. Notably, the turning points for the positive effects on the models diverge between the two POIs. Figure 4.8, Panel A shows that public transport links located within approximately 750 meters of a property show a positive impact. In comparison, food establishments only show positive neighborhood characteristics within a radius of approximately 150 meters, as depicted in Figure 4.8, Panel B.

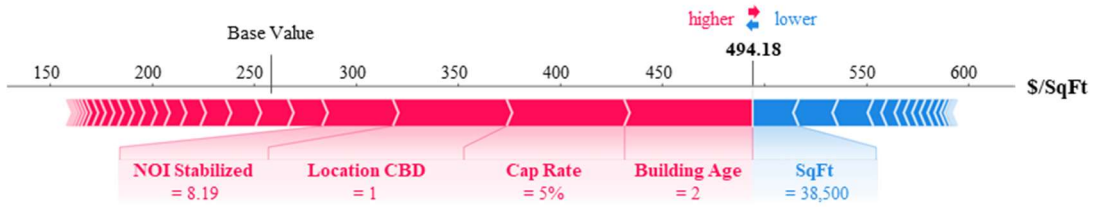
4.5.3 Local Model Interpretability

Shapley values are calculated for each observation individually, which offers the possibility to draw inference on both a global (i.e., aggregated) and a local (i.e., disaggregated) level. That is, each dot on the SHAP summary and partial dependence plots shown earlier represents a single prediction and can be explained locally on the property level. SHAP force plots visualize the decomposition of a specific prediction into the respective features. This makes each single market value estimate comprehensible and transparent. The sum

of all feature contributions represent the difference between the actual prediction and the mean prediction (base value) in the sample. It is important to note that feature effects can behave differently for different observations due the imposed non-linearity.

Figure 4.9 shows the composition of a market value prediction for an office property in Boston, Massachusetts. The expected market value for this property is 494.18 $/SqFt. The mean prediction (base value) of office market values in the sample is 258.53 $/SqFt. It can be considered the "best guess" for the market value without knowing anything about the specific property. The features that mainly drive the expectation from the base value of 258.53 $/SqFt to the predicted value of 494.18 $/SqFt are the stabilized NOI, location, market cap rate and building age.

Figure 4.9: SHAP Force Plot



Notes: This figure depicts the SHAP Force Plot, which presents the local decomposition of a selected prediction into the respective model features. The x-axis shows the estimated market value in dollars per square foot given the contributions of the individual features.

The property is newly built (building age = 2 years), located in the CBD and has a stabilized NOI of 5.23 $/SqFt, well above the sample average of 2.50 $/SqFt, thus increasing the prediction relative to the base value. The positive contribution of the stabilized NOI to the prediction increases the expected value by 149.58 $/SqFt. Additionally, the building age contributes 46.85 $/SqFt, its CBD location 38.65 $/SqFt and the market value cap rate of 5% in the quarter of observation contributes 54.99 $/SqFt. In sum, these four features contribute 290.07 $/SqFt to the expected value. Starting from the base value of 258.53 $/SqFt, this leads to an expected value of 548.59 $/SqFt. However, the negative contributions have been left aside so far. In this example, the property's square footage of 38,500 square foot reduces the expected market value as it is smaller than the average office building of 281,403 square foot. As highlighted in blue color, the size of the property pushes against the other features, thus reducing the expected market value by 30.63 $/SqFt. In sum, the remaining features add up to a negative 23.78 $/SqFt, leading to the final predicted market value of 494.18 $/SqFt.

4.6 Summary and Discussion

The objective of this study was to introduce an effective and comprehensive framework for the practical utilization of ML-based automated valuation models (AMVs) in the domain of commercial real estate that seeks to strike a balance between the accuracy and interpretability of the estimation method without compromising either one. To illustrate this, we trained a deep neural network (DNN) using a unique sample of more than 400,000 property-quarter observations from the NCREIF Property Index (NPI). We then applied a model-agnostic "Shapley Additive exPlanations" (SHAP) to shed light on the algorithm's prediction rules, offering *ex-post* interpretability. It could disentangle value drivers on an aggregated global level and a disaggregated local level for each property individually.

The used methodological framework achieves high accuracy in estimating commercial real estate market values across all four asset sectors. SHAP demonstrates that the inner workings of data-driven techniques are generally consistent with economic theory and mainly follow the traditional income approach by using the net operating income and market capitalization rates as the key explanatory features. Moreover, the location expressed by the geo-coordinates, the distance to points of interest and the properties' physical condition proxied with CapEx and building age strongly influenced the models' predictions. Deviations in the feature importance across property types were observed, predominantly in sector-specific characteristics. Furthermore, non-linear and three-dimensional relationships between market values and features were revealed and confirmed previous findings in the literature. For instance, it could be shown that the relation between market value and building age follows a U-shaped function, which can be explained by the bid-rent curve, as older buildings tend to be concentrated in city centers and CBDs, as well as a sample selection bias as good-quality buildings prevail while outdated or stranded assets leave the market to make room for new developments. On the local level of interpretation, SHAP furthermore showed that the effect of individual features could differ significantly across properties due to non-stationarity across space and time. This is one of the main advantages of machine learning techniques compared to linear hedonic models, as the latter reduces feature effects to a single, fixed beta coefficient that does not differentiate complex interactions between regressors.

In summary, our study demonstrates that machine learning algorithms can obtain both estimation accuracy and interpretability while following economic logic and being consistent with the current understanding of pricing processes in the literature. Moreover, these techniques can add to the existing knowledge by providing a deeper and more nuanced understanding of pricing processes in institutional investment markets.

That said, the findings of this study should be interpreted in light of certain limitations within both data and methods. Although the NPI is the most widely used commercial real estate price index in the United States, it is appraisal-based. Cannon and Cole (2011) as well as Deppner et al. (2023) find evidence that appraisal values tend to lag market dynamics and can be subject to bias. Moreover, the NPI is derived from a relatively small data sample of prime institutional properties. The findings may thus not be generalizable to all types of commercial real estate properties or investors.

While our main objective is to illustrate the potential of ML in increasing the understanding of pricing mechanisms in commercial real estate by providing valuable insights into price formation processes, a more comprehensive sample of transaction data is required to derive fully undistorted and generalizable results that are free of appraisal bias. This could be achieved by limiting the used NCREIF sample to sales data in conjunction with other data sources such as CoStar, CompStak or Real Capital Analytics. However, this is challenging as different data sources record different property characteristics. Merging these sources to increase the length of the data matrix comes at the cost of reducing its' width (i.e., property characteristics) or having to impute missing data.

The issue of data availability is linked to the limitations of machine learning techniques, which should be considered carefully next to the choice of data sources to ensure that the results are dependable and free from bias. As with any data-driven approach, ML methods are sensitive to the input data, which may exacerbate the issue of robustness and generalizability. More robust, universal, and reliable results can be expected with increased training data.

Despite their powerful applications, ML methods are not a panacea that can solve all real-world problems. However, if applied prudently, they could provide an answer to several problems and may become an indispensable tool for many tasks. With immense amounts of data being recorded every day and the development of quantum computing, machine-learning applications are about to experience a steep improvement in scale and efficiency. However, with these advances taking at least another five to ten years, applying interpretable AVMs in the commercial real estate sector is a milestone on a path yet to be traveled. By pointing to the caveats and illustrating the potential of these methods, our contribution represents a further step along this path and will hopefully motivate further research in this field.

4.7 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.

Calainho, F. D., van de Minne, A., & Francke, M. K. (2022). A machine learning approach to price indices: Applications in commercial real estate. *The Journal of Real Estate Finance and Economics,* Forthcoming.

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

Cannon, S. E., & Cole, R. A. (2011). How accurate are commercial real estate appraisals? Evidence from 25 years of NCREIF sales data. *The Journal of Portfolio Management, 35*(5), 68-88.

Deppner, J., von Ahlefeldt-Dehn, B., Beracha, E., & Schaefers, W. (2023). Boosting the accuracy of commercial real estate appraisals: An interpretable machine learning approach. *The Journal of Real Estate Finance and Economics,* Forthcoming.

Din, A., Hoesli, M., & Bender, A. (2001). Environmental variables and real estate prices. *Urban Studies*, *38*(11), 1989–2000.

Do, A. Q., & Grudnitski, G. (1993). A neural network analysis of the effect of age on housing values. *Journal of Real Estate Research*, *8*(2), 253–64.

Dunse, N., & Jones, C. (1998). A hedonic price model of office rents. *Journal of Property Valuation and Investment*, *16*(3), 297–312.

Goodman, A. C., & Thibodeau, T. G. (1995). Age-related heteroskedasticity in hedonic house price equations. *Journal of Housing Research*, *6*(1), 25–42.

Grether, D. M. & Mieszkowski, P. (1974). Determinants of real values. *Journal of Urban Economics*, *1*(2), 127–145.

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.

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.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903.

Levantesi, S., & Piscopo, G. (2020). The importance of economic variables on London real estate market: A random forest approach. *Risks*, *8*(4), 1–17.

Lockwood, L. J., & Rutherford, R. C. (1996). Determinants of industrial property value. *Real Estate Economics*, *24*(2), 257–272.

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2022). Interpretable machine learning for real estate market analysis. *Real Estate Economics*. Forthcoming.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS)*.

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.

Mills, E. S. (1992). Office rent determinants in the Chicago area. *Real Estate Economics*, *20*(2), 273–287.

Molnar, C. (2020). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Leanpub.

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., & 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.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: Review of valuation methods. *Journal of Property Investment & Finance*, *21*(4), 383–401.

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.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, *31*(2), 147–164.

Potrawa, T., & Tetereva, A. (2022). How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market. *Journal of Business Research*, *144*, 50–65.

Quan, D. C., & Quigley, J. M. (1991). Price formation and the appraisal function in real estate markets. *The Journal of Real Estate Finance and Economics*, *4*, 127–146.

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications*, *171.*

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

Shapley, L. S. (1953). A value for n-person games. In H. Kuhn, & A. Tucker (Eds.), *Contributions to the theory of games* (Vol. II, pp. 307–317). Princeton University Press.

Sirmans, C. F., & Guidry, K. A. (1993). The determinants of shopping centre rents. *Journal of Real Estate Research*, *8*(1), 107–15.

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance*, *38*(3), 213–225.

Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research*, *10*(2), 185–201.

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.

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 section provides a concise overview of the three scientific articles that comprise the cumulative thesis. It summarizes the objectives of each study, the data and methodologies employed, as well as the main outcomes and their implications for both the scientific community and practical applications.

Paper 1: Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach

Problems and Objective

In the past decade, machine learning (ML) techniques have seen increasing application for hedonic house price regression problems. In this context, literature has brought forth a growing body of evidence highlighting the superior predictive performance of ML compared to traditional statistical models (e.g., Mayer et al., 2019; Pace and Hayunga, 2020; Bogin and Shui, 2020). The measurement of predictive performance in ML methods is commonly based on resampling techniques such as cross-validation (CV) that implicitly assume statistical independence of the data (Bishop, 1995; Brenning, 2005; Varma and Simon, 2006). However, direct real estate markets feature a spatial dimension that causes inherent spatial dependence structures in the underlying price determination processes (Anselin, 1988; Can and Megbolugbe, 1997; Basu and Thibodeau, 1998). Applying such methods to direct real estate data without accounting for these dependence structures may lead to undetected overfitting and over-optimistic perception of predictive power (Roberts et al., 2017; Lovelace et al., 2019; Schratz et al., 2019). The meaningfulness and statistical validity of the resulting performance measures are therefore compromised.

The objective of this research paper is to investigate the role of spatial autocorrelation on the model selection and accuracy assessment of algorithmic regression methods and to assess the adequacy of conventional cross-validation errors in the context of hedonic house price modeling. In addition, this study proposes a spatial cross-validation strategy that can account for spatial dependence and reduce bias in error estimates.

Data and Methodology

The sample used for this study comprises a pooled cross-section of 9,256 asking rents from the Frankfurt residential market spanning the period from January 2019 through March

2020 on a monthly scale. The data stem from Empirica and were originally sourced from German multiple listing systems.

Tree-based algorithms are trained on a subsample and evaluated using spatial as well as non-spatial CV. Subsequently, out-of-sample data is forecasted to assess the bias in error estimates associated with spatial autocorrelation. The results are put into a broader perspective by benchmarking the applied ML algorithms against a non-spatial ordinary least squares (OLS) and a spatial autoregressive framework, allowing for a relative comparison of bias and predictive performance. Lastly, the residual spatial autocorrelation is analyzed to detect signs of overfitting to spatial structures in the data.

Results and their Contribution to Science and Practice

This study is the first in the literature to shed light on the bias in cross-validation errors of algorithmic hedonic approaches induced by spatial autocorrelation. To address this issue it proposes a spatial cross-validation strategy that reduces the bias. The findings confirm that error estimates from non-spatial resampling methods are overly optimistic, whereas spatially conscious techniques are more dependable and can increase generalizability. The results prove useful for increasing the robustness of algorithmic approaches to hedonic regression problems.

The precise estimation of property prices and rents is imperative to inform the decisions of many parties in the real estate industry, such as investors, developers, lenders or regulators. Since CV is commonly used as an "out-of-sample experiment" (Mullainathan and Spiess, 2017) to assess the predictive accuracy of algorithmic hedonic models, a systematic bias in error estimates may have adverse effects on the allocation of both debt and equity (Kok et al., 2017). This study helps increase the reliability and generalizability of algorithmic hedonic models, thus containing valuable implications for mass appraisal practices, credit risk management, portfolio allocation as well as investment decision making.

Paper 2: Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach

Problems and Objective

Commercial real estate markets are characterized by a high degree of heterogeneity, intransparency, and illiquidity, that complicate a thorough understanding of market dynamics and pricing mechanisms. State-of-the-art commercial real estate appraisals are therefore based on the experience and knowledge of experts, but they remain "[…] a subjective opinion of value [which] is based on an assessment of influences that a valuer

considers relevant […]", as stated by Dunse and Jones (1998). Consequently, commercial real estate appraisals have been subject to criticism for smoothing out market fluctuations, lagging behind actual market dynamics, incorporating systematic biases, and frequently deviating from actual transaction prices, as pointed out by Matysiak and Wang (1995), Geltner et al. (2003), Cannon and Cole (2011), and Kok et al. (2017).

In theory, statistical regression models can address most of the issues concerning subjectivity, structural bias, and outdated valuations, provided that they are specified correctly. In this context, so-called "intelligent" statistical learning methods have been extensively discussed in the residential real estate sector and have demonstrated remarkable results in accurately estimating prices and rents of houses and apartments (e.g., Mayer et al., 2019; Bogin and Shui, 2020; Pace and Hayunga, 2020; Pai and Wang, 2020; Ho et al., 2021). However, bridging the gap from their application in the housing sector to the commercial real estate sector seems to be more intricate, given the specifics of the markets. The aforementioned heterogeneity, intransparency, and illiquidity of commercial real estate markets hamper the availability of structured data. This, in turn, restricts the application of machine learning (ML) algorithms for valuation purposes, as these techniques rely on substantial amounts of information to learn relationships and generate reliable results.

The objective of this article is to extend the application of ML for property valuation and pricing analysis to a commercial real estate context. In addition, this contribution aims to explore the capacity of data-driven ML algorithms to adequately capture price formation processes in commercial property markets and provide a superior understanding of market dynamics that goes beyond traditional valuation methods. The study also examines structural bias in appraisals and points to the determinants that are not adequately reflected in current appraisal practices.

Data and Methodology

The principal dataset was provided by the National Council of Real Estate Investment Fiduciaries (NCREIF) and contains quarterly appraisal values and transaction prices of commercial properties across the United States that are included in the NCREIF property index (NPI). The sample spans the period from 1Q 1997 through 1Q 2021 and contains a series of financial and physical characteristics of the individual properties. These are enriched with locational attributes from Google Places, real estate market data from the NPI, and macroeconomic variables from the Federal Reserve Bank of St. Louis, the U.S. Census Bureau, and the U.S. Bureau of Labor Statistics. First, the deviation between actual sales prices observed in the market and the pre-sale appraised values of the properties in

the sample is examined. Second, an extreme gradient boosting (XGB) algorithm is applied to investigate the information content found in the residuals between appraised values and transaction prices with the aim to assess how much of the variation in these residuals can be explained. Finally, model-agnostic permutation feature importance is employed to shed light on the determinants that were not adequately reflected in appraisals.

Results and their Contribution to Science and Practice

This study breaks new ground in the literature by expanding the scope of ML applications in the context of property appraisals to commercial real estate markets. The findings show that the applied extreme gradient boosting trees could significantly decrease the variation in appraisal errors of commercial properties, thereby increasing accuracy and eliminating structural bias in appraisal values. The greatest improvements were observed for apartments and industrial properties, followed by office and retail buildings. This order coincides with both decreasing homogeneity and smaller sample size of the respective property types. The study also identified spatial and structural covariates as the primary factors influencing appraisal errors.

The results suggest that the application of machine learning methods has the potential to improve current appraisal practices, leading to more efficient and objective valuations that can consider a broader range of evidence. Analyzing transaction data through the lens of interpretable machine learning algorithms can furthermore add to a superior *ex-ante* understanding of pricing processes that may support practitioners in their decision making.

Paper 3: Increasing the Transparency of Pricing Dynamics in the U.S. Commercial Real Estate Market with Interpretable Machine Learning Algorithms

Problems and Objective

The advent of machine learning has brought new approaches to property valuation to the fore. ML-supported automated valuation models (AVMs) show promising results in terms of accuracy but lack inherent interpretability. This precludes their use in an institutional context as well as in regulatory and government applications. The aim of this study is to propose an integrated framework for the practical use of AMVs in a commercial real estate context that achieves high levels of precision and full *ex-post* interpretability of the models' prediction rules. Based on this, the article further aims to assess the consistency of the applied models with economic principles and showcases how the proposed methods can add to the understanding of pricing mechanisms in institutional real estate investment markets. By pointing to the caveats and illustrating the potential of these methods, this

contribution is intended to advance the application of AVMs in the commercial real estate sector and motivate further research in this field.

Data and Methodology

The principal dataset used in this study was provided by the National Council of Real Estate Investment Fiduciaries (NCREIF) and comprises quarterly property-level observations of appraisal values and property characteristics across four commercial property types (i.e., apartment, industrial, office, and retail) observed over a period of 30 years from Q1 1991 through Q1 2021. Furthermore, real estate market data from the NCREIF Property Index (NPI), macroeconomic data from the Federal Reserve Bank of St. Louis, the U.S. Census Bureau, and the U.S. Bureau of Labor Statistics, as well as spatial data from Open Street Maps and Google Places were added.

First, a deep neural network (DNN) was trained and calibrated for each property type individually. Second, an advanced model-agnostic methodology named Shapley Additive Explanations (SHAP) was applied to mitigate the trade-off between accuracy and interpretability and provide *ex-post* comprehensibility of the algorithms' prediction rules. Third, non-linear relationships as well as three-dimensional interaction effects were analyzed. In addition, a linear multiple regression analysis was conducted to serve as a point of reference.

Results and their Contribution to Science and Practice

This study is the first in the literature to extend the application of data-driven AVMs, let alone model-agnostic interpretation techniques such as SHAP, to commercial property types. The proposed methodological framework demonstrates high accuracy in the estimation of market values across all four asset sectors. Furthermore, significant non-linear and three-dimensional relationships in price determinants could be revealed. In summary, the relevant price determinants and the identified relationships follow an economic rationale and are in line with both hedonic literature and traditional valuation methods. Deviations across sectors are observed predominantly in sector specific features.

Since comprehensibility and interpretability are essential for the acceptance and operationalization of ML-driven AVMs in the industry, this study provides a valuable contribution to enhancing their practicality and marketability. In the long term, the proposed methods have the potential to leverage efficiency in both real estate markets and business processes by increasing the speed and scale of valuations, reducing costs, and ultimately increasing transparency of real estate pricing processes.

5.2 Final Remarks

The Royal Institution of Chartered Surveyors (RICS), a professional organization dedicated to promoting and regulating international standards of property valuation, states that

> "*[the] valuation of property has long been characteri[z]ed as both an art and a science: an art because of the need to make value judgments concerning the intangible features that attract certain buyers; a science because it is possible to establish trends and analy[z]e how these are interpreted by buyers and sellers, including the value placed on particular property characteristics*" (RICS, 2021).

The appraisal profession is assigned with the task to strike the balance between the art and the science of property valuation by extracting relevant signals and trends from irrelevant noise in the data. This exercise is typically performed by considering a small set of previously transacted comparable properties that possess similar attributes to the property being valued and adjusting for property-specific differences and intangible characteristics (Kok et al., 2017). The distinction between noise and signal in the underlying data is based on a valuer's subjective judgements and discretion that are grounded by their expert knowledge gained from observing past transactions in the market (Quan and Quigley, 1991; Dunse and Jones, 1998). The heterogeneity, intransparency, and illiquidity of real estate markets further complicate this process and hamper a thorough understanding of market dynamics and pricing mechanisms. This makes property valuation a challenging, time-consuming, and costly business.

While explaining and predicting the noise that is inherent in comparable sales will most likely remain an art for itself that involves a great deal of subjectivity and allows a wide margin of discretion, this dissertation has demonstrated that science can help to identify patterns and extract meaningful insights (i.e., the signals and trends) from large and complex data in an objective and structured manner. This ability becomes increasingly relevant in a world with enormous amounts of information being recorded every day and the evolution of computer-aided scientific techniques that enable efficient analysis of these data. In this context, statistical models, and in particular machine learning algorithms, exhibit significant potential in accurately modeling real estate markets, adequately capturing price formation processes, and estimating property prices precisely and promptly on a large scale.

The systematic analysis of real estate data using statistical learning models can help to reduce discretionary scope and mitigate the issues arising from subjective involvement in property appraisals, while simultaneously increasing efficiency and reducing costs. By

automating some of the more tedious and time-consuming aspects of the valuation process, such as data collection and analysis, statistical approaches can accelerate decision-making processes and offer increased accuracy and objectivity of valuations. This promotes transparency and confidence in the valuation process while benefiting real estate owners, including public and private pension plans, insurance companies, banks, and their respective stakeholders and customers, who ultimately have to bear the costs associated with property appraisals (Kok et al., 2017).

That said, caution is required when drawing conclusions based on a machine's output. ML algorithms may be perceived as intelligent and have accomplished impressive results in detecting complex patterns and relationships to solve real-life problems such as extracting the signal from noisy transaction prices, but they have no intellect that allows them to think, let alone to truly understand the problems they are solving. That is why artificial intelligence still relies on the supervision of human agents to motivate analyses and to make sense of the data.

This thesis has addressed the most critical limitations of data science techniques and in particular data-driven machine learning algorithms in the context of property valuation and pricing. In doing so, it has contributed to the complementation of these methods with the required economic and methodological frameworks to translate the produced outputs into more meaningful results. Nonetheless, this work does not claim to offer an all-encompassing solution to the challenges of using data-driven methods for the analysis of direct real estate markets. Alternative approaches and aspects need to be considered to further enhance the understanding and transparency of property pricing mechanisms. Moreover, it is important to note that the results presented in each of the articles are based on rather specific and limited data samples. Despite the aim to make the findings as generalizable as possible, the sensitivity of data-driven techniques to changes in training data means that the reported results may not be universally applicable.

With respect to Paper 1, it may be worthwhile to extend the analysis beyond spatial dependence structures and consider the temporal dimension as well to capture spatio-temporal dependencies in the data. This can provide a more comprehensive solution to the problem at hand, particularly if the methods are applied in a prediction context. Regarding Paper 2, the analysis could be expanded to consider more varied and extensive data, allowing the results to be more generalizable to other property and investor types beyond the institutional prime sector. To achieve this, the collection and integration of unstructured data from different sources will be necessary to mitigate the issue of data scarcity. The same applies to Paper 3. In addition, further research is needed in the field of

interpretable machine learning to increase the consistency and reliability of the produced results. Although model-agnostic interpretation techniques are often portrayed as a panacea that provides an appealing answer to the complexity and opaqueness of data-driven algorithms, it is important to recognize that these methods come with their very own limitations and pitfalls. One such limitation is the presence of collinearity and dependencies between input features, which can cause bias and compromise inferential analyses (Molnar et al., 2022). This closes the circle to Paper 1, given the presence of spatial dependence in real estate markets. Proper consideration of the hidden assumptions in both machine learning and model-agnostic interpretation techniques is thus of crucial importance to ensure a reliable and robust output.

Having demonstrated the potential and pointed to the limitations of using statistical learning methods to examine price formation processes in direct real estate markets, this thesis aims to provide guidance, stimulate critical discourse, and motivate further research in this field.

5.3 References

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Kluwer Academic Publishers.

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, *17*(1), 61–85.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

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.

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, *5*, 853–862.

Can, A., & Megbolugbe, I. (1997). Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics*, *14*, 203–222.

Cannon, S. E., & Cole, R. A. (2011). How accurate are commercial real estate appraisals? Evidence from 25 years of NCREIF sales data. *The Journal of Portfolio Management, 35*(5), 68–88.

Dunse, N., & Jones, C. (1998). A hedonic price model of office rents. *Journal of Property Valuation and Investment*, *16*(3), 297–312.

Geltner, D., MacGregor, B. D., & Schwann, G. M. (2003). Appraisal Smoothing and Price Discovery in Real Estate Markets. *Urban Studies*, *40*(5/6), 1047–1064.

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

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.

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with R*. CRC Press.

Matysiak, G. A., & Wang, P. (1995). Commercial property market prices and valuations: Analysing the correspondence. *Journal of Property Research*, *12*(3), 181–202.

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., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. and Samek, (Eds.), *xxAI – Beyond Explainable AI, Lecture Notes in Artificial Intelligence (LNAI)* (13200, pp. 39–69). Springer.

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

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.

Pai, P.-F., & Wang, W.-C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, *10*(17), 5832.

Quan, D. C., & Quigley, J. M. (1991). Price formation and the appraisal function in real estate markets. *The Journal of Real Estate Finance and Economics*, *4*, 127–146.

RICS (2021). Automated valuation models: Roadmap for RICS members and stakeholders. *The Royal Institution of Chartered Surveyors*.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schroeder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. *Ecography*, *40*(8), 913–929.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(91).