Understanding Commercial Real Estate Markets with Machine Learning Methods

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft [Dr. rer. pol.]

eingereicht an der Fakultät für Wirtschaftswissenschaften der Universität Regensburg

vorgelegt von:

BENEDICT VON AHLEFELDT-DEHN

Berichterstatter: Prof. Dr. Wolfgang Schäfers

Prof. Dr. Tobias Just

Understanding Commercial Real Estate Markets with Machine Learning Methods

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

eingereicht an der Fakultät für Wirtschaftswissenschaften der Universität Regensburg

vorgelegt von:

Benedict von Ahlefeldt-Dehn

Berichterstatter: Prof. Dr. Wolfgang Schäfers

Prof. Dr. Tobias Just

Tag der Disputation: 20. Juli 2023

Understanding Commercial Real Estate Markets with Machine Learning Methods

Benedict von Ahlefeldt-Dehn

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation and Background

In times of extensive monetary policy, rapid technological advancement and the rise of digital assets such as cryptocurrencies and non-fungible tokens, global financial markets have chartered new courses changing portfolio and investment decisions for both private and institutional investors. Commercial real estate represents a small part of the investable universe, but accounts for about 15 per cent of global real estate value (Savills, 2021). Investors such as corporates, insurance companies, pension funds, sovereign wealth funds as well as private investors, including high net worth individuals and retail clients, seek to invest a significant share of their portfolio in commercial real estate markets as part of their investment strategies. With the impending digital transformation in the real estate industry we are drawn to new approaches and perspectives that are adapting the way we understand global commercial real estate markets. The use of modern econometric approaches such as machine learning and artificial intelligence in investment processes and property management will certainly be a step on a path yet to be travelled.

In its core however, commercial real estate is characterized by immobility and complex interrelationships. Across different property types and geographical location, commercial properties are considered heterogeneous. While residential real estate has a certain degree of homogeneity in terms of its structural characteristics, this is not the case for commercial real estate. Moreover, commercial property markets are illiquid due to their sheer size and capital commitment as well as high search and transaction costs. Additionally, low data availability leads to opaque markets and fosters information asymmetries across the sector. However, commercial real estate plays a significant role in investment and wealth creation as well as preservation. While retaining its risk reducing role in a portfolio real estate can compete with other alternative investments such as private equity, hedge funds and infrastructure (Clayton et al., 2009). Real estate maintains a favorable and safe investment opportunity due to strong income provision, inflation hedging characteristics and its significant diversification benefits (Gordon et al., 1998, Gilberto, 1999, Chua, 1999 and Mueller and Mueller, 2003).

Thus, production of reliable market value estimates is more important for real estate than for other asset classes. As real estate markets are decentralized the valuation process differs significantly from that in equity and bond markets. Transaction prices are the result of pairwise negotiations between buyers and sellers and not readily observable in a public market (Quan and Quigley, 1991). Accurate and timely valuations are important for

lending, taxation purposes, managing portfolios, and evaluating investment and divestment strategies. Various stakeholders, including authorities, banks and investors, therefore use econometric forecasting and valuation models (Shapiro et al., 2012). Setting up such models is a complex task that involves estimating the risk and return via several financial, physical, location-related, and structural parameters.

For the valuation of properties including the estimation of the aforementioned range of parameters multiple frameworks have been introduced in real estate research. On the one hand, structural and theoretical approaches are used to describe the relationships between supply and demand, while on the other hand statistical and purely data-based methods are used to find relationships between variables. In data-driven approaches the functional form ultimately determines the degree of flexibility and ability to capture complex relationships. Traditionally, valuation tasks and parameter forecasts have used linear models with economic rationales (Pagourtzi et al. 2003, Abidoye et al., 2019), but these capture the complexity and heterogeneity of the market only to a limited extent (Zurada et al., 2006). With the advent of digitalization, advances in computing power and increasing data availability in residential real estate markets, the application of data-driven frameworks with more flexible functional forms in the models gained track. Several studies have demonstrated the enormous potential of applying non-linear machine learning algorithms to the residential real estate market (Mullainathan and Spiess, 2017; Mayer et al., 2019; Bogin and Shui, 2020; Hong et al., 2020; Pace and Hayunga, 2020; Lorenz et al., 2022). However, commercial real estate markets have been continuously experiencing low data availability and thus not been in the focus of research. Only in recent years, data became more available in commercial real estate markets and the application of purely data-driven approaches has become possible.

Therefore, the objective of this dissertation is to explore the potential of non-linear machine learning algorithms to improve the accuracy and interpretability of commercial real estate markets. Foremost it will be discussed whether the application of machine learning methods leads to superior performance and greater transparency in both univariate and multivariate frameworks to rental forecasting and property valuation.

The first paper focuses on the univariate estimation and forecasting of rents with machine learning methods. Recent developments in time series forecasting using machine learning and deep learning methods offer an opportunity to update traditional univariate forecasting frameworks. Literature has shown that for univariate forecasting ensemble models produce best results. Hence, a hybrid methodology combining both ARIMA and a state-of-the-art deep neural network model is proposed to exploit the unique strengths of

both methods in linear and non-linear modelling. The methods are jointly applied to a unique dataset of 21 European office markets. The results show that the ensemble model tends to lower the error volatility and increase forecasting accuracy across the observed markets. The market heterogeneity allows and explicitly demands the usage of multiple approaches to compute adequate forecasts.

The second paper focuses on examining the accuracy and bias of market valuations in the U.S. commercial real estate sector and assesses the potential of machine learning algorithms to improve the appraisal accuracy and eliminate structural bias. A unique dataset from the National Council of Real Estate Investment Fiduciaries (NCREIF) between the years 1997 and 2021 covering apartment, industrial, office and retail properties enables the exploration of deviations between market valuations and subsequent transaction prices. Under consideration of 50 covariates it is found that the deviations between the appraised market values and the actual transacted prices exhibit structured information content that the employed machine learning algorithm can capture and further explain. Results show that, the appraisal accuracy can be increased and structural bias can be eliminated.

The third paper focuses on the application of a framework for the practical use of fully interpretable automated valuation models (AVMs) in commercial real estate. Market values are estimated with a deep neural network model based on the above mentioned dataset provided by NCREIF. The dataset is comprised of 400,370 quarterly property-level observations across four commercial property types apartment, industrial, office, and retail over a period of 30 years from 1991 to 2021. This dataset is complemented with a variety of macroeconomic and locational characteristics constituting a total of 32 financial, physical, location-related, and structural features. After evaluating the models' predictive performance, an advanced model-agnostic technique, named Shapley Additive Explanations (SHAP), is applied to mitigate the trade-off between accuracy and interpretability. SHAP provides full ex-post comprehensibility of the applied model. While achieving high accuracy and maintaining full interpretability of prediction rules it can be shown that the applied models are consistent with economic principles.

1.2 Research Questions

This section outlines the essential research questions that each of the three papers in this dissertation addresses, focusing on the commercial real estate markets as a unifying theme. The dissertation employs a methodological advancement as its framework. The first paper employs a univariate approach to examine the estimation and forecasting of commercial real estate markets, while the second and third paper employ multivariate approaches to explore the subject in greater depth. Consequently, the following research questions can be raised:

Paper 1: Forecasting Office Rents with Ensemble Models – the Case for European Real Estate Markets

- Are univariate methods eligible to be used in commercial real estate market forecasting problems?
- Do non-linear machine learning and deep learning methods add value to classical statistical methods such as autoregressive or exponential smoothing processes in the context of modelling prime office rents?
- Can ensemble models (as a combination of classical statistical methods and deep learning models) increase the forecasting accuracy and reduce volatility in a multiple market context?

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

- Can non-linear machine learning methods find structured information content in the appraisal errors of traditional valuation approaches?
- To what extent can tree-based ensemble learners effectively reduce appraisal errors with regard to accuracy and bias in commercial real estate appraisals?
- What are the determining factors that lead to deviations between traditionally appraised market values and the actual transaction prices?

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

- Can non-linear machine learning methods be effectively applied in an automated valuation modelling approach to commercial real estate markets?
- Can model-agnostic techniques, in particular, Shapley Additive Explanations, explain the influence of input features on the predictive response of machine learning models and thus mitigate the trade-off between accuracy and interpretability?
- To what extent do the applied models conform to economic principles and how do the proposed methods add to the understanding of pricing mechanisms in institutional real estate investment markets?

1.3 Co-Authors, Submissions and Conference Presentations

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

Paper 1: Forecasting Office Rents with Ensemble Models – the Case for European Real Estate Markets

---

Authors:
Benedict von Ahlefeldt-Dehn, Marcelo Cajias, Wolfgang Schäfers

Submission Details:
Journal: Journal of Property Investment and Finance
Current Status: accepted (08/03/2022) and published in Volume 41, Issue 2 (03/27/2023)

Conference Presentations:
- the 27th Annual Conference of the European Real Estate Society (ERES) in Kaiserslautern, Germany (2021)
- the 38th Annual Conference of the American Real Estate Society (ARES) in Bonita Springs, Florida, USA (2022)

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

---

Authors:
Jürgen Deppner, Benedict von Ahlefeldt-Dehn, Eli Beracha, Wolfgang Schäfers

Submission Details:
Journal: Journal of Real Estate Finance and Economics
Current Status: accepted (02/16/2023) and published online (03/22/2023)

Conference Presentations:
- the 38th Annual Conference of the American Real Estate Society (ARES) in Bonita Springs, Florida, USA (2022)
- Workshop "Artificial Intelligence and Finance" of the Center of Finance of the University of Regensburg (2022/2)
- the 28th Annual Conference of the European Real Estate Society (ERES) in Milan, Italy (2022)
- the 39th Annual Conference of the American Real Estate Society (ARES) in San Antonio, Texas, USA (2023)

Awards and Fundings:
This paper was awarded the 2022 "Altus Group Best Paper Award" of the 28th Annual Conference of the European Real Estate Society (ERES) in Milan, Italy.

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, Jürgen Deppner, Eli Beracha, Wolfgang Schäfers

Submission Details:
Journal: Journal of Portfolio Management (Special Real Estate Issue)
Current Status: accepted (06/05/2023)

Conference Presentations:
- Workshop "Artificial Intelligence and Finance" of the Center of Finance of the University of Regensburg (2023/1)
- the Research Conference 2023 of the Real Estate Research Institute (RERI) in Chicago, Illinois, USA (2023)
- Workshop "Artificial Intelligence and Finance" of the Center of Finance of the University of Regensburg (2023/2)

Awards and Fundings:
This paper was granted the 2022 research funding by the Real Estate Research Institute (RERI).

1.4 References

Abidoye, R.B., Junge, M., Lam, T.Y.M., Oyedokun, T.B. and Tipping, M.L. (2019). Property valuation methods in practice: evidence from Australia. *Property Management*, 37(5), 701-718.

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.

Chua, A., (1999). The role of international real estate in global mixed-asset investment portfolios. *The Journal of Real Estate Portfolio Management*, 5(2), 129-137.

Clayton, J., Ling, D. and Naranjo, A., (2009). Commercial real estate valuation: Fundamentals versus investor sentiment. *The Journal of Real Estate Finance and Economics*, 38(1), 5-37.

Gilberto, M., Foort, H., Hoesli, M., MacGregor, B., (1999). Optimal diversification within mixed-asset portfolios using a conditional heteroskedasticity approach: Evidence from the U.S. and U.K.. *Journal of Real Estate Portfolio Management, 1999*, 5(1), 31–45.

Gordon, J., Canter, T., Webb, J., (1998). The effect of international real estate securities on portfolio diversification. *Journal of Real Estate Portfolio Management,* 4(2), 83.

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.

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

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.

Mueller, A. and Mueller, G. (2003). Public and private real estate in a mixed-asset portfolio. *The Journal of Real Estate Portfolio Management*, 9(3), 193-204.

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.

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

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.

Savills, (2021). The total value of global real estate. *Savills World Research* https://www.savills.com/impacts/market-trends/the-total-value-of-global-real-estate.html, Accessed 24 April 2023

Shapiro, E., Mackmin, D. and Sams, G. (2012). Modern methods of valuation, 11th ed., Estates Gazette, London**.**

Zurada, J. M., Levitan, A. S., & Guan, J. (2006). Non-conventional approaches to property value assessment. *Journal of Applied Business Research*, 22(3).

2 Forecasting Office Rents with Ensemble Models – The Case for European Real Estate Markets

2.1 Abstract

Commercial real estate and office rental values, in particular, have long been the focus of research. Several forecasting frameworks for office rental values in multivariate and univariate fashions have been proposed. Recent developments in time series forecasting using machine learning and deep learning methods offer an opportunity to update traditional univariate forecasting frameworks. With the aim to extend research on univariate rent forecasting a hybrid methodology combining both ARIMA and a neural network model is proposed to exploit the unique strengths of both methods in linear and non-linear modelling. N-BEATS, a deep learning algorithm that has demonstrated state-of-the-art forecasting performance in major forecasting competitions, are explained. With the ARIMA model, it is jointly applied to the office rental dataset to produce forecasts for four-quarters ahead. When the approach is applied to a dataset of 21 major European office cities, the results show that the ensemble model can be an effective approach to improve the prediction accuracy achieved by each of the models used separately. Real estate forecasting is essential for assessing the value of managing portfolios and for evaluating investment strategies. The approach applied in this paper confirms the heterogeneity of real estate markets. The application of mixed modelling via linear and non-linear methods decreases the uncertainty of abrupt changes in rents. – To the best of the authors' knowledge, no such application of a hybrid model updating classical statistical forecasting with a deep learning neural network approach in the field of commercial real estate rent forecasting has been undertaken.

Keywords – Commercial real estate, Time series forecasting, Machine learning, Deep learning, Office rent, ARIMA, N-BEATS

Acknowledgments: The authors especially thank PATRIZIA SE for contributing to this study. All statements of opinion reflect the current estimations of the authors and do not necessarily reflect the opinion of PATRIZIA SE or its associated companies.

2.2 Introduction

Institutional investors such as corporates, insurance companies, sovereign wealth funds, pension funds and private investors, including high net worth individuals and retail clients, seek to invest a significant share of their portfolio in commercial real estate as part of their global investment strategies. Thus, the ability to forecast and assess market behaviour via the main value driver, the rent of a property, is part of every strategic investment allocation. Investors, mortgage underwriters, regulatory authorities, tax assessors and valuation firms base their work on rental forecasts.

Commercial real estate markets, particularly office markets, have been extensively researched since the 1960s, and multiple frameworks for office rent forecasting have been proposed. That is, in multivariate and univariate fashions tackling, on the one hand, a structural and theoretical approach and, on the other hand, a statistical and atheoretical viewpoint. Structural frameworks face the challenge of scarce data in commercial real estate markets. Univariate models are more flexible but cannot cope with market heterogeneity. Therefore, an ensemble approach combining multiple univariate approaches could solve the forecasting problem in a multiple market context. The objective of this study is focused on univariate approaches to forecasting and, in this context, to understand, describe and apply a combination of a classical statistical method and a deep learning approach. The practicability and functionality of the proposed structure are demonstrated via the application to a dataset of 21 major European office markets.

This study contributes to the existing literature in several ways. First, the history of commercial real estate rent estimation and forecasting is reviewed. Moreover, the state-of-the-art deep learning method (N-BEATS) is explained and proposed to complement and update univariate statistical forecasting models. The aim is to extend existing research on univariate rent forecasting. The literature has shown that office rental movements can be determined autoregressively to a high extent and ensemble models dominate univariate estimation. It is to be tested whether commercial real estate markets, office rents particularly can be estimated accurately from autoregressive processes by the use of ensemble models including a modern deep learning approach over a period of more than 20 years across 21 major European office markets. Therefore, it is to clarify how the proposed methods perform in the forecasting problem individually and in the combination of both approaches as an ensemble model. Comparisons are drawn to the benchmark. Generally, it is expected that the combination of the introduced methods proves to be an effective approach to improving the forecasting accuracy achieved by each of the models used separately.

The remainder of the study is divided into five sections. Firstly, existing frameworks for rental forecasting (structural and univariate) will be discussed and reviewed. The third and fourth sections provide an overview of the data that was used and outline the employed models. The results and forecasting accuracy in- and out-of-sample are discussed in section five. Section six is the conclusion.

2.3 Literature Review

The first section of the literature review focuses on the development of the determination of rental values in office markets with structural models that try to explain changes in the dependent variable by movements in other employed explanatory variables. In contrast, the second section of the review outlines univariate, ensemble forecasting and machine learning as well as deep learning approaches.

2.3.1 Structural Models

Researchers in the United States were pioneers in the task of explaining the adjustment process of commercial real estate rents. With historical data from San Francisco ranging from 1961 to 1983, Rosen (1984) demonstrates with a supply and demand model that office stock, construction starts, vacancy rate and office rent were key elements in estimating market behaviour. The author finds that changes in the observed rental markets are non-linearly related to the vacancy rate determined by the deviation of the observed vacancy rate to the equilibrium vacancy rate in the market. These findings also align with the results from studies by Hekman (1985) and Shilling et al. (1987). The authors estimate rent forecasting models for major metropolitan markets across the United States for periods between 1960 and 1983 and confirm that changes in rental rates are strongly related to changes in the observed vacancy rate and the general economic condition of the subject market. Wheaton and Torto (1988) follow the framework from previous findings and confirm that the change in commercial rental values can be directly related to the vacancy rate excessing a structural or equilibrium vacancy level. Frew and Jud (1988) extend the research using cross-sectional office market data from 1984 by adopting a hedonic approach employing property-specific attributes, which complements the included vacancy rate in estimating commercial office rents.

Gardiner and Henneberry (1989) develop a time series model with office data ranging from 1977 to 1984 to forecast a regional rent index. The authors state that regional economies strongly correlate with the national economy. Thus, the model includes both the regional gross domestic product and regional office stock as demand and supply proxies. Moreover, the authors conclude that models for the prediction of office rents should include variables

that can describe regional market dynamics. Giussani et al. (1993) investigate office rents across 13 major European cities from 1983 to 1991. The authors address data availability and provide further evidence that demand variables have a higher explanatory power on rents and are easier to obtain than supply-side proxies. Gross domestic product and unemployment rate are found to be statistically significant across all cities and had high explanatory power in the model. However, Giussani et al. (1993) show that the coefficients and the explanatory power of the two demand proxies vary significantly over time and section. Boon Foo and Higgins (2015) take another look at the complexities of commercial real estate markets and propose a single equation demand and supply model for Singapore's office property market. With data for the period from 1992 to 2005 the rent model could explain around 70% of variation. Once again, the authors can show that vacancy rates, construction costs, the prime lending rate and office sector employment play a decisive role in the model.

The synthesis of both streams, the vacancy-based approach and the reduced form demand and supply models, led to the error correction approach which up to date dominates structural modelling of office rents. Hendershott et al. (2002a) describe error correction frameworks as capturing both short-term influences and long-run equilibria between two or more time series. Office market dynamics are found to be estimated more accurately when including both a time-varying equilibrium rent and vacancy as explanatory variables. In a study of 11 UK regions over 29 years of data Hendershott et al. (2002b) estimate error correction mechanism models for retail and office properties. The proposed long-run equilibrium relationships and the short run dynamic corrections are found to be formulated correctly and significant. Interestingly, the models for the London market are found to be more responsive to lagged rental change than the other observed regions. Mouzakis and Richards (2007) build upon the research carried out by Hendershottet al. (2002a) and propose an error correction mechanism model for a panel study of 12 European office markets for the period from earliest 1980 and 2001. The used models generate inferential insights to the markets demand and supply side behaviour in the short and long-run. Also, the authors can draw conclusions about the interdependencies within the markets and forecast short-term changes in real rents. The authors benchmark the model's performance against alternative models and can show superior performance for the theoretical model. Further studies by Brounen and Jennen (2009a, b), Hendershott et al. (2010), McCartney (2012) and Bruneau and Cherfouh (2015) confirm prior findings by testing the proposed methodological approach on representative US and European office rental datasets. Generally, the conclusions about theoretical interdependencies are strong and indicate that rental movements are determined by complex relationships in multiple

variables of both the demand and supply side. In both short-run adjustments and long-run equilibrium estimates the authors succeed to explain the subject market behaviour.

2.3.2 Univariate Time Series Models

In contrast to structural approaches taking different market characteristics into account, univariate time series frameworks allow model variables simply by their previous values and a stochastic error term. Most commonly applied univariate modelling techniques include autoregressive moving average models (ARMA) and exponential smoothing. McGough and Tsolacos (1995) chose a univariate forecasting framework to model office, retail and industrial rents in the United Kingdom from 1977 to 1993. The authors use quarterly data and are satisfied with the forecasting abilities of the employed ARIMA model for office rents; thus, it is mentioned that such models work best for only short-term forecast horizons. The results indicate that past changes in office rents influence current and future changes, and therefore, autoregressive structures can be adequately employed. Tse (1997) confirms previous findings by fitting quarterly Hong Kong office and industrial property prices from 1980 to 1995 into an ARIMA model. The author demonstrates that the model can be employed to compute short-term forecasts and indicate turning points in the market. Chaplin (1998) supports McGough and Tsolacos (1995) in his studies reviewing and clustering previous research that has been carried out. The author develops two general model forms, which on the one hand are derived from academic office rent models and on the other hand from practitioners' models. Both models include, among others, lagged rent as explanatory variables.

Although structural and univariate models can achieve adequate results, Chaplin (2000) raises concerns about the selection of models for forecasting purposes. Models are usually chosen ex post according to different selection criteria, that is, R^2, Akaike Information Criteria and Schwarz's Bayesian Information Criterion. Still, univariate approaches remain interesting and are compared to other structural and more theoretical approaches. Stevenson and McGarth (2003) compare an ARIMA, a Bayesian VAR, structural and simultaneous equation models for forecasting office rent. The multivariate models outperform the ARIMA model. However, the authors attribute this observation to the use of an ARIMA model in its simplest form and the comparatively long forecast horizon. In general, the employed methods individually show advantages and disadvantages. The authors find that the simultaneous equation model and the VAR produce accurate forecasts as both models have a strong theoretical background mimicking the market interplay of demand and supply. Nonetheless, the shortcoming of such models is very high data requirement. Hence, alternative approaches such as single equation models and

univariate frameworks should be considered. However, according to Zhang (2003), generally the best historical fit does not necessarily imply the best forecast ability due to sampling variation, model uncertainty and structural change. Therefore, the author suggests combining different methods to ease the ex post modelling selection process. Furthermore, real-world data are usually combinations of linear and non-linear patterns. Thus, in theory, the application of multiple methods that can capture a variety of patterns is useful. Zhang (2003) demonstrates that combining an ARIMA model and an Artificial Neural Network (ANN) improves forecasting accuracy significantly. The study of Stevenson (2007) on the office rent forecasting abilities of ARIMA models confirms the findings from prior research by McGough and Tsolacos (1995), Tse (1997) and Crawford and Fratantoni (2003) as the employed ARIMA models generally seem to produce good short-term forecasts. However, the author states that these models need to be viewed with caution. In general, Brooks and Tsolacos (2010) describe that the motivation in the use of pure time series models roots on the one hand in the tangibility of data for multivariate methods and, on the other hand, in the performance of the out-of-sample forecasting. Hence, time series models help anticipate market trends, but variations are high with alternative specifications and varying data sources. Nonetheless, this emphasises the need to combine multiple approaches and sheds light on the application of machine learning and deep learning methods in areas of research that have been accessed with traditional estimating and forecasting methods up to that date.

2.3.3 Machine Learning and Deep Learning Models

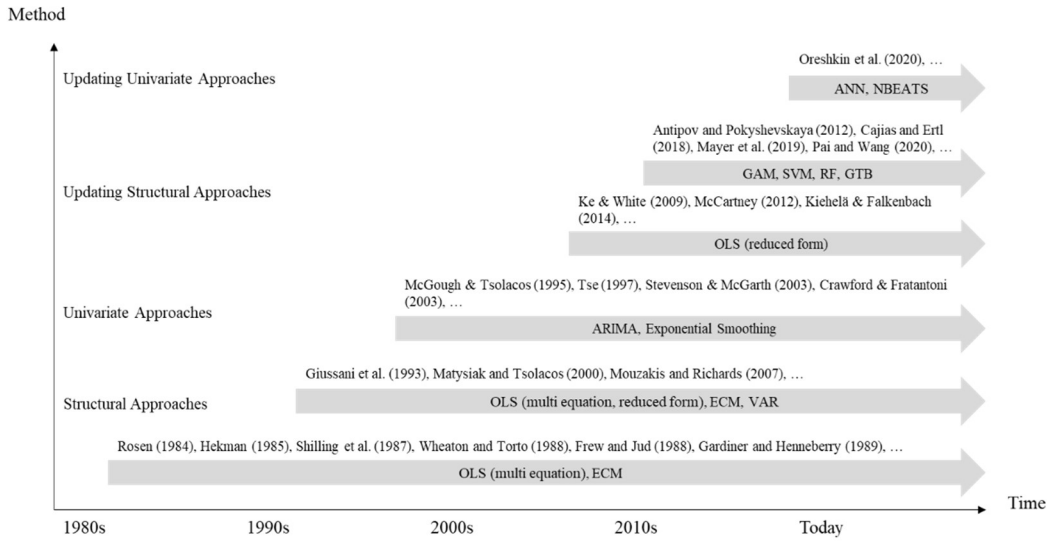
Meanwhile, many supervised machine learning and deep learning applications in real estate prove to be successful and beneficial. Notably, in hedonic valuation approaches, tree methods and neural networks yield superior results in comparison to traditional parametric modelling (see Dabrowski and Adamczyk, 2010; Antipov and Pokryshevskaya, 2012; Cajias and Ertl, 2018; Mayer et al., 2019; Pai and Wang, 2020). The authors collectively demonstrate that modern machine learning methods with non-parametric functions can pick up relations between features much more efficiently. However, in recent years the attention of researchers shifted to time series forecasting with classical statistical methods and machine learning methods. Oreshkin et al. (2019) focus on time series forecasting and find that this research area can be categorised into statistical modelling and machine learning approaches. The statistical modelling approaches are found to be mainly built on exponential smoothing and ARIMA models.

In contrast, the machine learning and deep learning approaches to time series forecasting were introduced later and are highly successful. That is, in forecasting competitions tree

methods (gradient and eXtreme gradient tree boosting) or more modern deep learning methods such as modified artificial or recurrent neural networks showed superior forecasting performance and dominated the field. Oreshkin et al. (2019) present a deep neural network architecture (neural basis expansion analysis for interpretable time series forecasting – N-BEATS) that demonstrates state-of-the-art performance on statistical benchmarks. In a practical application, Oreshkin et al. (2020) employ the developed N-BEATS to 35 electricity demand time series for European markets and find that the neural network outperforms all competitors (classical statistical methods, machine learning, hybrid approaches) regarding the forecast accuracy and bias. Similar results are extracted by Li et al. (2016) and Milunovich (2020), who focused on forecasting REIT and stock price as well as real house price indices. Both studies employ multiple methods to compare the forecasting performance of classical statistical forecasting algorithms and modern machine learning and deep learning approaches. Milunovich (2020) analyses the Australian real house price index between 1972 and 2017 at quarterly frequency with 47 different algorithms. The author applies various methods, including ARIMA models, support vector machines, decision tree methods (random forest and gradient boosting), an LSTM recurrent neural network and mean and median forecast combinations of different models. Furthermore, the author specifies most of the employed methods in a univariate (house price growth rate) as well as multivariate fashion. Results show that linear ARIMA models perform well in short-term forecasts. Additionally, deep learning models such as the LSTM recurrent neural network demonstrate good medium and long-term forecasting performance. These results are in line with the findings by Li et al. (2016) who confirm that long-term memory exists in most of the analysed time series (REITs and stocks) and note that neural network structures perform best to forecast in comparison to traditionally applied forecasting methods. Also, Milunovich (2020) describes the advantage of applying machine learning and deep learning models to mirror non-linear patterns and select the best models by cross-validating with training and test sets in comparison to relying on information criteria.

The development of different office rent estimation frameworks over time is displayed in Figure 2.1. Over the last years, new approaches to estimating and forecasting commercial real estate rents have been proposed, and structural and univariate methods have been updated. Existing literature, on the one hand, demonstrates that univariate methods can be employed as a viable alternative to structural approaches in office rent forecasting and, on the other hand, that more modern machine learning and deep learning approaches are used complementary to produce more accurate forecasts. Thus, a classical statistical method, the ARIMA model, and the state-of-the-art deep neural network model, N-BEATS,

Figure 2.1: Overview of Office Rent Estimation Frameworks over Time



are chosen to be tested in this study on its forecasting performance. Moreover, the use of ensemble models promises even more accurate and less volatile estimations. Thus, it is to prove how the ARIMA model and the N-BEATS model perform in the forecasting problem and furthermore, how an ensemble model behaves in comparison to the commonly applied ARIMA model as a benchmark.

2.4 Data Description

The data consists of quarterly office prime rental values for 21 major European office markets from 1991 to 2020 gathered from CBRE. This yields a maximum of 120 observations for the time series of the Paris office market. The office market prime rents are obtained in euro per square meter per year. The prime rent can be defined as an average rent of the top 3–5 percent of all lettings in the observed markets. However, to deal with non-stationarity issues, the quarterly rents are transformed to year on year growth rates. Furthermore, the literature shows that analyses are consistently focused on inflation-adjusted rent series. Therefore, all observed time series were deflated with the inflation rate of the respective country. Hence, all analyses and results are presented for real rental growth rates. After transformation, all observed time series are tested stationary with the Augmented-Dicky-Fuller (ADF) test. The null hypothesis of the ADF test assumes that the data is non-stationary. Hence rejecting the null hypothesis with a significant p-value below 10% confirms stationary data. This holds for all 21 time series after transformation. The stationarity-tests are displayed in Table 2.6 in appendix. The following table displays statistics for the real rental growth rates over the maximum period of each observed market. The mean, standard deviation and autocorrelation (first lag) are reported

accordingly (see Table 2.1). The analysis and estimation are carried out for two datasets. The first dataset is an unbalanced panel that includes each time series with the maximum number of available observations. The second dataset consists of a common period for all cities – this limits the number of observations for all cities to 72, but ensures comparability across all 21 cities and serves as a robustness test for the applied models.

Table 2.1: Summary Statistics of Data: Real Rental Growth, Unbalanced Panel

#	City	Start of Period	Mean	SD	ACF 1
1	Vienna	1993Q1	-0,01706	0,049576	0,915659
2	Helsinki	2001Q1	0,014285	0,054658	0,83431
3	Lille	2000Q1	0,009006	0,050265	0,670382
4	Lyon	2000Q1	0,014751	0,076394	0,641508
5	Marseille	2000Q1	0,021698	0,112136	0,656492
6	Paris Ile-de-France	1991Q1	-0,00422	0,109419	0,8873
7	Berlin	2001Q1	0,002829	0,090358	0,84532
8	Cologne	2001Q1	0,001625	0,063539	0,798317
9	Dusseldorf	2002Q1	-0,00141	0,03555	0,865217
10	Frankfurt am Main	1992Q1	-0,01373	0,09921	0,87577
11	Hamburg	2001Q1	-0,00138	0,060884	0,883223
12	Munich	2001Q1	-0,00136	0,050232	0,78568
13	Dublin	2003Q1	0,014933	0,157501	0,910203
14	Milan	1992Q2	-0,00175	0,116618	0,917141
15	Rome	2002Q1	-0,00198	0,076179	0,833584
16	Amsterdam	1992Q1	0,010476	0,059974	0,872425
17	Rotterdam	1997Q1	0,009621	0,041928	0,706746
18	The Hague	1997Q1	0,003417	0,051315	0,751937
19	Lisbon	1992Q1	-0,03421	0,094155	0,956627
20	Barcelona	1992Q1	-0,01668	0,13868	0,940067
21	Madrid	1992Q1	-0,00904	0,169209	0,951347

Notes: The start of period marks each beginning of the available time series for the observed market. The mean displays the mean real rental growth rate for the observed period. SD displays the standard deviation of the subject time series. ACF 1 displays the first lag autocorrelation of the growth rates.

2.5 Methodology

This chapter presents the employed methods and relevant error metrics to draw comparisons in the analysis. Furthermore, the methodology of building the ensemble model is outlined and described thoroughly.

2.5.1 Forecasting Methodology

In statistical modelling and particularly in time series forecasting, the most commonly applied approaches by researchers and practitioners are exponential smoothing, autoregression and moving average processes. In this study, the focus is on integrated autoregressive moving average models (ARIMA) as numerous applications in the field of real estate time series forecasting (McGough and Tsolacos, 1995; Tse, 1997; Stevenson and McGarth, 2003; Crawford and Fratantoni, 2003) have demonstrated its eligibility.

ARIMA modelling assumes the forecast of the variable is estimated by the movements of its past values and errors. Thus, as described in the study by McGough and Tsolacos (1995), the theoretical idea behind this approach is that past rental values contain information about future market behaviour. A model capturing past values and errors can generally be stated as:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \theta_0 + \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_p \varepsilon_{t-p} \quad (1)$$

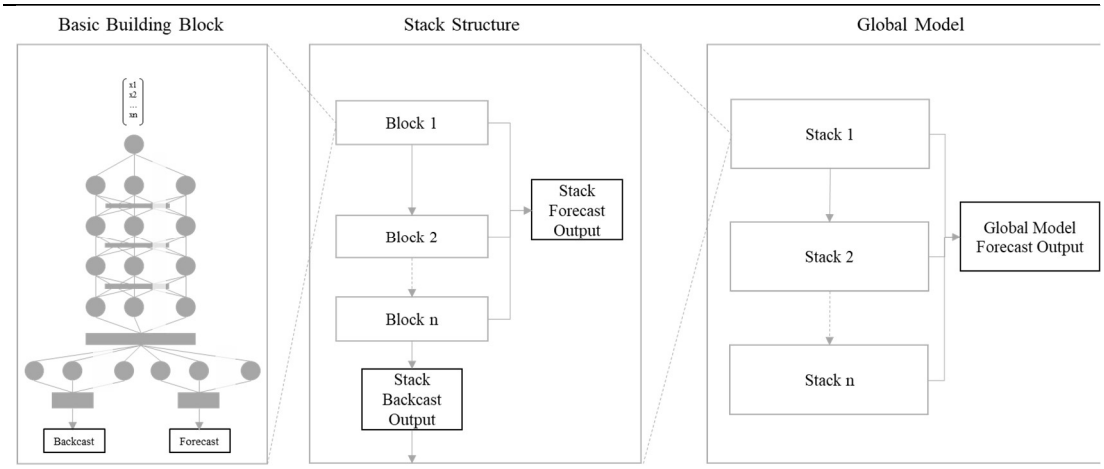
where y_t is the dependent variable office rent growth rate (which is the d^{th} difference of the variable office rent growth rate at time t – with d being the degree of integration). θ_0 is a constant and ε the error term which is assumed to have constant variance and zero mean as well as being independently and normally distributed. Furthermore, ρ is the number of lagged terms of y_t and q the number of lagged terms of ε. The forecast package (version 8.15) in R can be used with the auto.arima function to identify the three parameters by its automated correlation analysis. The function chooses the values of (p, d, q) based on the information criteria AIC, AICc or BIC by searching possible models for the constraints provided. This function follows the ARIMA modelling procedure developed initially by Box et al. (1977) and returns the best ARIMA model. However, Zhang (2003) mentions that observed patterns in real-world datasets often have non-linear properties that can be more adequately captured by applying non-linear models. Still, not all non-linear models are beneficial in the forecasting application (bilinear models, TAR models, ARCH models). Newly developed neural network structures show more flexibility in their time series modelling capabilities.

The suggested non-linear method in this study is a novel algorithm based on a deep neural network structure. Oreshkin et al. (2019) developed the neural basis expansion analysis for interpretable time series forecasting (N-BEATS). That is, a deep learning method explicitly developed to produce univariate time series point forecasting. The method is based on a deep neural network architecture with forward and backward residual connections and stacks of fully connected layers. Neural networks are algorithms designed to recognise patterns and estimate the relationship between a set of input and output signals. The basic architecture mimics the cell connections of a biological brain and uses artificial neurons to model the desired output. A basic artificial neuron's operations in the so-called building blocks can be stated by the following equation.

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right) \quad (2)$$

where x_i is the input values which are multiplied with the weights w_i. This product is then fed through the activation function f(x) and yield the output y(x). In general, neural networks are characterised by three main parts: the number of layers, the direction of travel for the information and the number of nodes within the layers. (Lantz, 2019) The stacking of multiple neural networks that are composed of multiple layers is referred to as deep learning in this context. The N-BEATS neural network architecture differs from existing neural network architectures that focus on sequence forecasting (such as LSTM recurrent neural networks) as, according to Oreshkin et al. (2020), the forecasting problem is treated as a non-linear multivariate regression problem. The inner working of the N-BEATS neural network will be described in the following passage. The neural network consists of several stacks, where each comprises multiple basic building blocks (see Figure 2.2).

Figure 2.2: Neural Network Architecture of N-BEATS Algorithm



The Basic Building Block: The basic building block is designed to make two estimations simultaneously. The first estimation takes a window of past values to compute a so-called backcast, which can be compared to an in-sample estimation. The second estimation computes the actual forecast for a given horizon. The architecture of a basic building block is based on the structure of a simple artificial neural network where the inputs are fed through a stack of four fully connected layers with rectified linear activation functions (RELU). The two outputs of a building block are the backcast and the forecast, which are then processed by the following blocks.

The Stack Structure: A stack consists of multiple basic building blocks that process the outputs of previous blocks in the principle of doubly residual stacking. The concept of doubly residual stacking comes from the sequential processing of each of the block's backcast residuals. Each block in a stack takes the previous block's backcast and a window of past values as its input signals. By subtracting the window of past values from the

previous block's backcast, a vector is generated which incorporates only those learnings (residuals) that are not learned by the previous block. The last block in the stack produces the stack backcast output, which is fed through to the following stack. At the same time, the forecast output of each block is summed up, yielding the aggregated stack forecast output.

Global Model: The stack backcast output serves as input for the subsequent stack and represents all learnings (residuals) not yet learned by the prior stacks. The sum of all stack forecast outputs is aggregated to a global forecast which is the overall output of the N-BEATS neural network. The structure described graphically and theoretically has been practically applied by Oreshkin et al. (2020) to a forecasting problem of mid-term electricity load. Mathematically this can be captured by the following notations. The forecast horizon of length H and a length T observed time series history is described by $[y_1, \dots, y_T] \in \mathbb{R}^T$ with the task to predict future values $y \in \mathbb{R}^H = [y_{T+1}, \dots, y_{T+H}]$ given past observations. Furthermore, the author describes the operation of the neural network with the following equations where $x \in \mathbb{R}^n$ is the input, *r/l* the subscripts for the blocks and layers, FC the fully connected layer with $W^{r,l}$ as the weights and $b^{r,l}$ as biases and RELU as the rectified linear unit activation function:

$$\begin{aligned} x^r &= RELU[x^{r-1} - \hat{x}^{r-1}] \\ h^{r,1} &= FC_{r,1}(x^r), \dots, h^{r,L} = FC_{r,L}(h^{r,L-1}) \\ FC_{r,L}(h^{r,L-1}) &\equiv RELU\left(W^{r,l}h^{r,l-1} + b^{r,l}\right) \end{aligned} \tag{3}$$

with B^r as the backcast and F^r as the forecast projection matrix of the dimensions (n*H and H)

$$\begin{aligned} \hat{x}^r &= B^r h^{r,L} \\ \hat{y}^r &= F^r h^{r,L} \end{aligned} \tag{4}$$

and the final forecast is the sum of forecasts of all the residual blocks:

$$\hat{y} = \sum_r \hat{y}^r \tag{5}$$

2.5.2 Ensemble Model Approach

Having outlined the individual workings and structure of the ARIMA and N-BEATS models, the novel ensemble model approach can be described. However, it first has to be outlined that the ARIMA and N-BEATS models can only be successfully applied in time series forecasting when it is assumed that the underlying time series make movements in the future that are equal to its patterns in the past. Only then the historical performance of

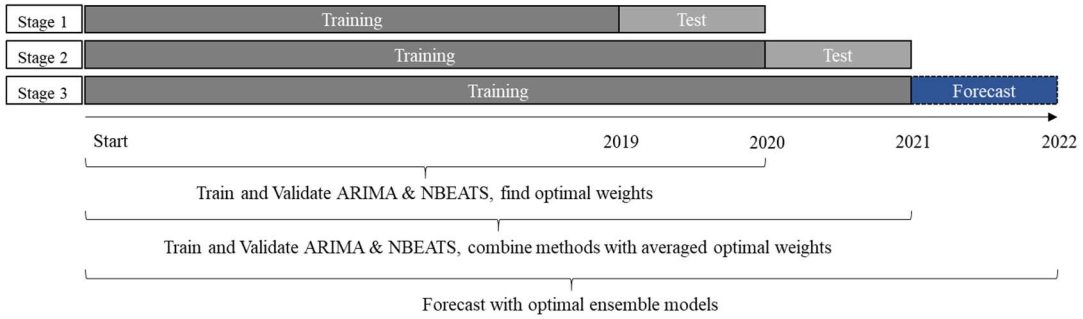
each model allows conclusions to be drawn on its future forecasting accuracy. Chaplin (1998) concluded from his study on office rent prediction frameworks that the best historically fitted model must not certainly produce the best forecast of office rents. Hence, the idea behind building an ensemble model from multiple methods is rooted in the combination of more than one historically good fitting model to be certain to tackle possible changes in patterns in future time series and thus provide a model that produces reliable forecasts.

The study of Atiya (2020) summarizes the lessons learned from the M4 forecasting competition. The winning algorithms were a hybrid of machine learning and statistical approaches and statistical and non-linear models. This is in line with historically best-performing approaches with respect to Armstrong (2001). Based on these developments and promising results, this study combines the classical statistical ARIMA model and the relatively novel N-BEATS neural network. However, there is more than one combination of possibilities. Gooijer and Hyndman (2006) summarise the main approaches of combining models and state that the simple average is the most commonly used approach. However, the authors find this approach insensitive to past information, such as the performance of methods in the underlying samples. Other combination methods considering the past behaviour of the applied models are described by determining the weights by some optimisation problem (that is, OLS, regime-switching models, STAR models, etc.). This study uses an ex-post approach as described by Armstrong (2001). The author suggests choosing weights according to the historic fit of the individual employed models. The following paragraph will outline the exact procedure of combining the two methods to one ensemble forecasting approach.

The final ensemble forecasting model is estimated in two stages. It follows a train, validation and test order and fits all employed models to different time windows of the obtained data set. In the first step, the ARIMA models are fitted individually to all 21 time series of the European office markets. Hence, 21 models are estimated for a time period (training) from the start of the observations to the end of the year 2018. This yields a total number of 72–112 (according to which market is estimated) observations for the training dataset. This is in line with requirements regarding the minimum number of 50 observations proposed by McGough and Tsolacos (1995) and Tse (1997). The models are then validated/ tested on unseen data between the first and last quarter of 2019 (four-quarters out of sample).

Additionally, the N-BEATS neural network is estimated for the same time windows (train and validation/test set) in a multi-task fashion, meaning that one model simultaneously fits

Figure 2.3: Modelling Infrastructure



networks for all cities. Oreshkin et al. (2019) demonstrate in their application and development of the algorithm that such deep learning models can explicitly be trained on multiple time series simultaneously dealing with the data in a multi-task fashion. Nonetheless, this does not mean that the estimated N-BEATS model computes the same forecast for all time series. Much more, 21 individual forecasts for each of the observed office markets are extracted. Optimal hyperparameters for the N-BEATS model are found via hyperparameter tuning in a grid search fashion.

Based on the two estimated models for each city, error metrics are calculated. The two models per city are then combined by minimising the mean absolute errors over the validation/test period. The optimisation problem is simplified by iterating over all possible combinations attributing weights between 1 and 99 to the two models per city successively. The optimal weights per city are then averaged over all cities to acquire one average optimal weighting per method. In a second step, the optimal averaged weightings are validated. Again, both methods are trained on the dataset (as depicted in Figure 2.3) from the start of the dataset to the end of 2019. The validation/test set is an out of sample period of four-quarters from the first to the last quarter of 2020. Hence, the performance of the combination of the individual models via the calculated optimal averaged weightings is tested and evaluated on unseen data.

2.5.3 Error Metrics

The following metrics are used to evaluate the performance of the applied methods. Hyndman and Koehler (2006) focus on assessing error measurements for univariate time series forecasts and propose the following metrics to be relevant in measuring the forecast performance and errors. Apart from the standard measures of forecast errors (mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE)), the authors introduce the mean absolute scaled error (MASE) as a suitable error

measurement method for all situations. This is described as a scaled error based on the in-sample mean absolute error. According to the authors, the MASE can be stated as:

$$MASE = \frac{\sum_{t=1}^{N}|e_t|}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|} \tag{6}$$

where e_t is the MAE divided by the MAE of the in-sample naïve forecast of the observed data (Y_i - Y_{i-1}). The optimisation of the N-BEATS neural network follows the MASE in the estimation process of the underlying time series. Moreover, as the data is in the form of growth rates, the standard MAPE error metric has its limitations as it cannot handle zero or close to zero values. This problem can be tackled by replacing the MAPE with a new metric, the mean arctan absolute percentage error (MAAPE) proposed by Kim and Kim (2016). The following formulas can express the MAPE and MAAPE:

$$\begin{aligned} MAPE &= \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \\ MAAPE &= \frac{1}{N}\sum_{i=1}^{N}\tan^{-1}\left(\frac{Y_i - \hat{Y}_i}{Y_i}\right) \end{aligned} \tag{7}$$

The authors describe the MAAPE as an error metric similar to the MAPE but overcoming the disadvantages laying in the problematic division by zero and preserving the advantages of easy interpretation. Table 2.7 defining all error metrics is attached to appendix.

2.6 Results

In this section, the results of this study about European office market forecasting are discussed. First, econometric findings are analysed. That is, the accuracy of the methods is assessed individually, and the methods are compared relative to one another. Both the individual forecasts and the ensemble model forecasts are depicted and interpreted in the following section. Furthermore, out-of-sample performance, the real forecast and the error distribution are reviewed and analysed in detail. Finally, a comparison to the benchmark model is drawn. The ensemble model developed in this study is called meta model in the subsequent passages.

Econometric Findings

The ARIMA model is the benchmark model as it has been the standard approach in univariate real estate rent forecasting problems due to its intuitive and simple application (Granger and Newbold, 1986). Concluding from the literature overview, the ARIMA models are expected to yield good short-term forecasts. The following results are discussed

for the case of estimating nominal rental growth rates. The performance of the models that have been estimated for real rental growth rates are discussed separately. Results show that in the majority of the observed office markets, the ARIMA models perform well, and yield a R^2 of around 74% on average (mean). Generally, the deep learning architecture with a doubly residual stacking principle leads to good forecasting performance. The out-of-sample error measurements demonstrate that neural network approaches can adequately describe univariate time series problems.

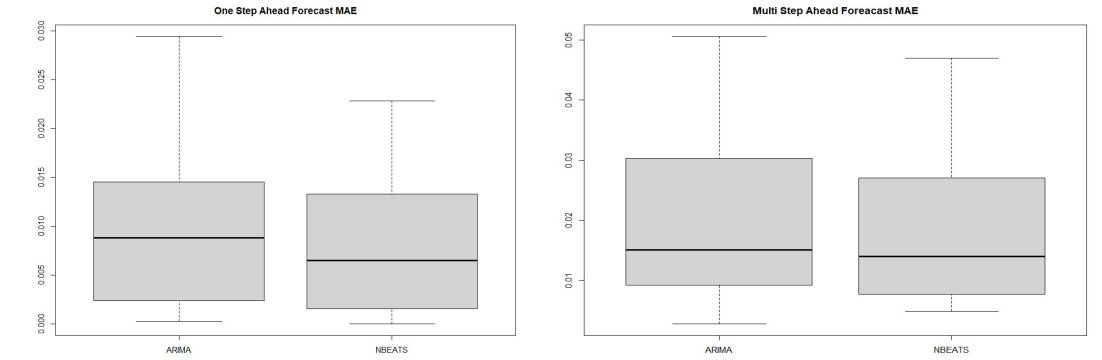
In comparison to the classical statistical approach, it is found to be more accurate. This holds for both periods (stage 1 and stage 2) in the modelling process. The individual errors of each market forecast are displayed in Tables 2.8, 2.9 and 2.10 in appendix. They indicate that some office markets can be forecasted better by the classical statistical approach and others by the deep learning framework. This suggests there is room for improvement by combining both methods in an ensemble model.

When comparing the one-step and multi-step forecasts of both employed methods, error metrics confirm results in the existing literature for the majority of the observed office markets. Milunovich (2020) forecasts growth rates and log prices of the Australian house price index. He finds that while simple ARIMA and VAR models outperform more complicated models in short-term forecasts, non-linear models such as deep learning specifications lead to accuracy in mid to long-term forecast problems. This relation of forecasting performance cannot directly be confirmed as displayed in Figure 2.4. However, this can be attributed to the fact that both forecasts (one and multi-step) are considered short-term forecasts.

The boxplots depict that the median MAE of the one-step/one-quarter ahead forecast of the ARIMA is higher than the median MAE of the N-BEATS. Also, the variance in the N-BEATS forecast is lower. The fitting of the ARIMA and N-BEATS models in stage 1 yields the averaged optimal weighting for both methods based on the historical deviation (sum of MAE of four-quarters out of sample forecast) of all 21 observed office markets. This yields an optimal allocation of 61%/39% for the ARIMA and N-BEATS models, respectively, to build the best meta model[1]. Stage 2 validates the meta model's

[1] The optimal weighting for the models on the balanced panel is 47%/53% for the ARIMA and N-BEATS models, respectively.

Figure 2.4: Mean Absolute Error Comparison of One-step Forecast to Multi-step Forecasts (Unbalanced Panel, Real Rental Growth)



performance by applying the averaged optimal weightings to the newly fitted ARIMA and N-BEATS models. Table 2.2 summarises the error-based model performance of all three approaches in stage 2 out-of- sample tests (see Figure 2.5).

Table 2.2: Real Rental Growth: Error-based Comparison of Model Performance (Unbalanced Panel)

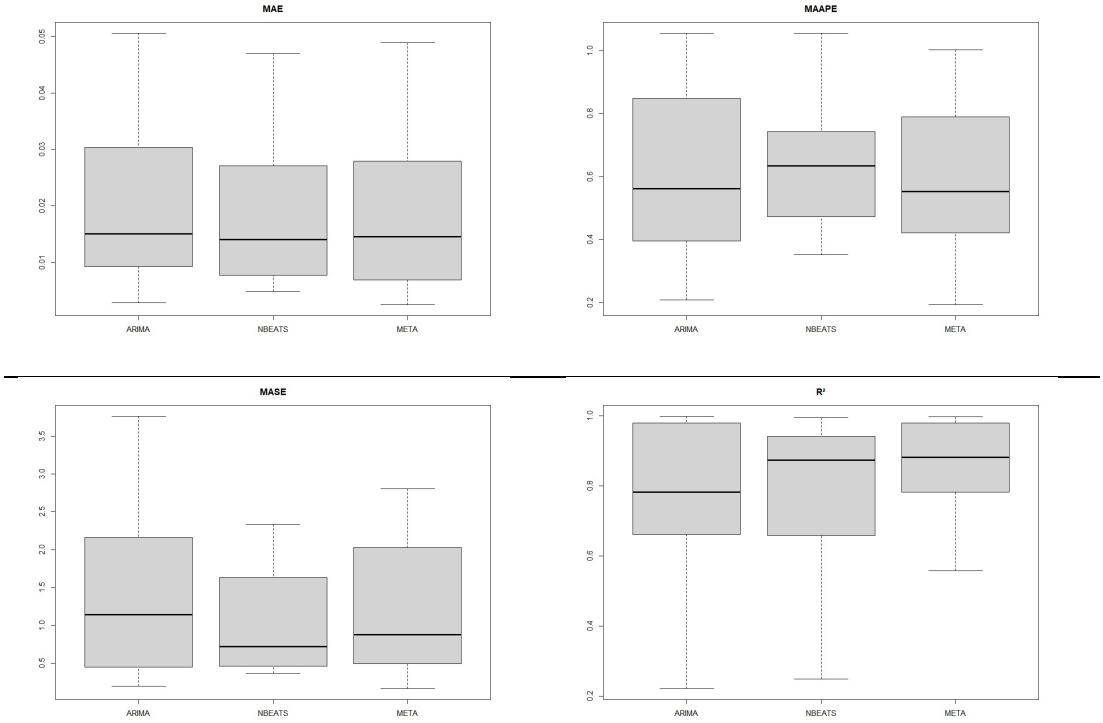
Mean	MAE	MASE	MAAPE	RMSE	R^2
ARIMA	0,0254	1,3911	62,5716	0,029	0,7399
N-BEATS	**0,0187**	**1,0493**	63,7475	**0,0213**	0,7673
Meta model	0,0207	1,133	**59,0782**	0,0236	**0,8263**

Median	MAE	MASE	MAAPE	RMSE	R^2
ARIMA	0.0151	1.1445	56.2573	0.0187	0.7836
N-BEATS	**0.0140**	**0.7230**	63.5159	**0.0148**	0.8741
Meta model	0.0145	0.8794	**55.34298**	0.0149	**0.8824**

Notes: The error-based comparison is drawn on the average (mean/median) over all 21 observed time series and thus includes all outliers. Bold font indicates the best results in each column. The MAE is the average (mean/ median) of the mean absolute error in percentage points of the year on year change of the prime office rents. The MASE is the average (mean/ median) of the mean absolute scaled error - a MASE > 1.0 implies that the out-of-sample forecast performs worse than a naïve in-sample forecast. The MAAPE is the average (mean/median) of the mean arctangent absolute percentage error of all observed forecasts in percentage points. The RMSE is the average (mean/median) of the root mean squared error of all observed forecasts. The R^2 is the average (mean/median) of the goodness of fit of all observed forecasts.

The measure for goodness of fit, R^2, shows that combining the ARIMA and N-BEATS models leads to better ensemble model forecasts than both individual models produce. The N-BEATS and the meta model significantly decrease the errors on average compared to the ARIMA benchmark model. The meta model is in terms of the error metrics MAE, MAAPE, MASE and RMSE close to the N-BEATS model. It successfully combines the advantages of both individual methods in the out-of-sample forecasts. The MAAPE indicates an average deviation of the predicted values to the actual values of around 59%. In real terms, this means that the forecasted office prime rent on average (median) deviates about 5.10 EUR p.a. from the actual value. The explanatory power is high for univariate time series models with an average (mean) R^2 of over 82% (median 88%) in the ensemble model. In the study by Mouzakis and Richards (2007) on forecasting 12 European office

Figure 2.5: Error-based Comparison of Model Performance (Unbalanced Panel, Real Rental Growth)



markets with ARIMA models, the authors achieved an explanatory power of around 43%. The combination of the individual models increases the goodness of fit by about 12% points compared to the benchmark. Furthermore, mean absolute errors can, on average (mean), be reduced by around 20% points in comparison to the benchmark model. To check for the robustness of the proposed approach the models are estimated as described in the data section for a balanced panel with one common observation period for all markets. The results are displayed in the following Table 2.3.

Table 2.3 displays the results and again the combination of the ARIMA and the N-BEATS models prove the adequacy in forecasting real rental growth rates. Figures 2.9, 2.10 and 2.11 in appendix show the error-based comparison in boxplots and the visualization of the mean absolute error by city in an ascending order and confirm the findings from the first analysis of the unbalanced panel. The three office markets of Munich, Dusseldorf and The Hague show the lowest MAE for the meta model (see Table 2.4) and are selected from the sample of the 21 markets explicitly to demonstrate the working of the meta model and are displayed in Figure 2.7 in appendix. The ARIMA generally forecasts more positive year on year growth rates, whereas the N-BEATS underpredicts the actual movements. Combining both methods with the optimal averaged weights leads to an optimally fitting meta model that reduces the mean absolute error of the four-quarters out-of-sample forecast substantially. However, the forecasts for office markets such as Barcelona,

Table 2.3: Real Rental Growth: Error-based Comparison of Model Performance (Balanced Panel)

Mean	MAE	MASE	MAAPE	RMSE	R^2
ARIMA	0,0227	1,2352	63,5318	0,0264	0,7384
N-BEATS	**0,0167**	**0,9483**	**60,1141**	**0,0197**	0,6739
Meta model	0,0175	0,9613	61,3233	0,0208	**0,8073**
Median	MAE	MASE	MAAPE	RMSE	R^2
ARIMA	**0,0145**	1,0439	69,3967	**0,017**	0,8983
N-BEATS	0,0181	**0,8969**	**55,7334**	0,0216	0,7936
Meta model	0,0174	0,901	56,6875	0,0198	**0,9399**

Notes: The error-based comparison is drawn on the average (mean/median) over all 21 observed time series and thus includes all outliers. Bold font indicates the best results in each column. The MAE is the average (mean/ median) of the mean absolute error in percentage points of the year on year change of the prime office rents. The MASE is the average (mean/ median) of the mean absolute scaled error - a MASE > 1.0 implies that the out-of-sample forecast performs worse than a naïve in-sample forecast. The MAAPE is the average (mean/median) of the mean arctangent absolute percentage error of all observed forecasts in percentage points. The RMSE is the average (mean/median) of the root mean squared error of all observed forecasts. The R^2 is the average (mean/median) of the goodness of fit of all observed forecasts.

Marseille or Lisbon perform worst in terms of error metrics. The ARIMA and N-BEATS methods cannot forecast the big jumps in office rental changes of the cities Barcelona and Lisbon and overpredict the market rental growth rate as depicted in Figure 2.8 in appendix.

Table 2.4: MAE Selected Cities (Unbalanced Panel, Real Rental Growth)

MAE	Munich	Dusseldorf	The Hague
ARIMA	0.0028	0.0034	0.0053
N-BEATS	0.0071	0.0048	0.0053
Meta model	**0.0025**	**0.0034**	**0.0036**
Δ ARIMA – Meta model	-10.22%	-0.8%	-31.05%

Notes: The MAE of the three displayed cities measures the absolute deviation of the forecast to the actual year on year change of the observed office prime rents. The last line Δ ARIMA – Meta model is the error reduction of the meta model in comparison to the ARIMA benchmark model in percentage points. Bold font indicates the lowest MAE in each column.

In consequence, the meta model's deviation to the actual values is comparatively high. It was only in the Marseille time series that the N-BEATS model forecast the fast increase of rental growth rates adequately. Nonetheless, as the ARIMA massively over-predicted the changes, the meta model failed to produce accurate forecasts. Still, the meta model's forecast reduces errors by 9.14, 21.08 and 42.07% in the cities Barcelona, Marseille and Lisbon, respectively, in comparison to the benchmark model (see Table 2.5).

Generally, as Figure 2.6 depicts, the meta model tends to lower the volatility in terms of absolute errors in the observed office markets and generally smoothes out the forecasts when viewing and forecasting multiple office markets simultaneously. Different models lead to different forecasts, and thus conservative as well as optimistic forecasts can be combined. Results show that market heterogeneity allows and explicitly demands the usage of multiple approaches to compute adequate forecasts. This relation between the individual and combined models is displayed for the analysis with the unbalanced panel in Figure 2.6 and also holds true for the model estimation with one common observation period, as depicted in Figure 2.11 in appendix.

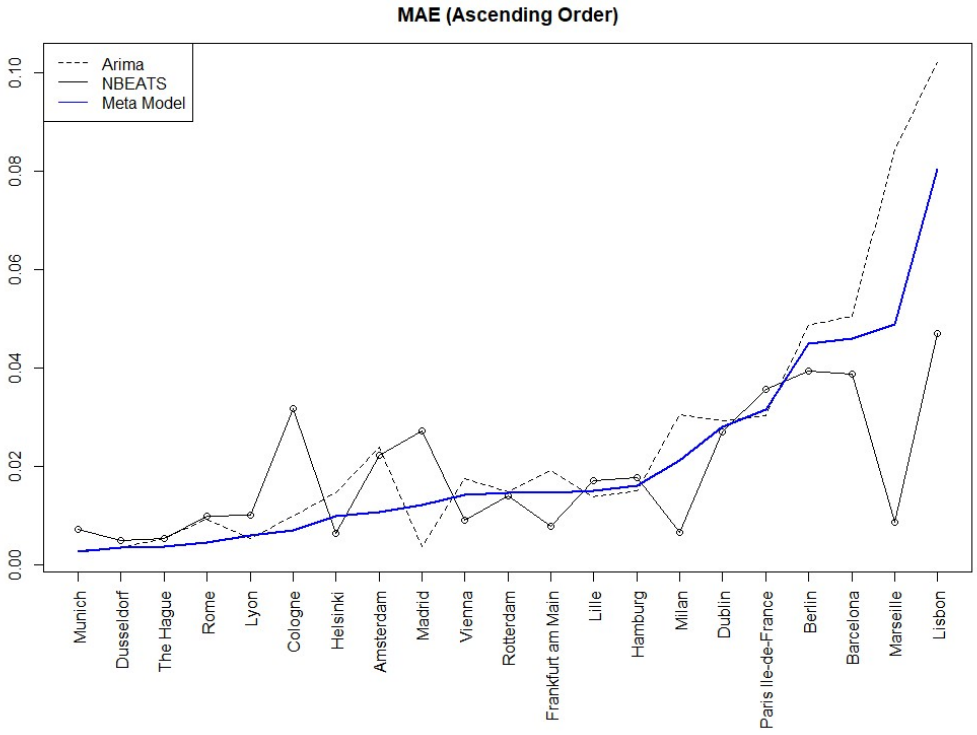
Table 2.5: MAE Selected Cities (2) (Unbalanced Panel, Real Rental Growth)

MAE	Barcelona	Marseille	Lisbon
ARIMA	0.0506	0.0843	0.1021
N-BEATS	**0.0387**	**0.0086**	**0.0469**
Meta model	0.0459	0.0489	0.0805
Δ ARIMA – Meta model	-9.14%	-42.07%	-21.08%

Notes: The MAE of the three displayed cities measures the absolute deviation of the forecast to the actual year on year change of the observed office prime rents. The last line Δ ARIMA – Meta model is the error reduction of the meta model compared to the ARIMA benchmark model in percentage points. Bold font indicates the lowest MAE in each column.

Real estate forecasting is essential for assessing the value of managing portfolios and for evaluating investment strategies. The approach applied in this paper confirms the heterogeneity of real estate markets and that one rule does not fit all. When applying mixed modelling of markets via linear and non-linear methods, the uncertainty of abrupt changes in rents decreases.

Figure 2.6: MAE by City in Ascending Order (Unbalanced Panel, Real Rental Growth)



2.7 Conclusion

This paper comprises an overview of commercial real estate rent forecasting frameworks and proposes an update on classical statistical univariate time series forecasting by combining an ARIMA model with a deep learning approach. Approaches in literature in recent years proposed to update classical forecasting frameworks with machine learning

and deep learning methods to take advantage of linear and non-linear estimation properties. Forecasting with modern machine learning and deep learning algorithms demonstrated superior results in many fields of application. The selected N-BEATS method proved to have state-of-the-art forecasting properties in numerous statistical forecasting competitions. In a hybrid fashion, the advantages of both the ARIMA model and the N-BEATS model are combined and significantly improve the forecasting performance in multiple out-of-sample forecasts. It is demonstrated that the combination of the classical statistical approach with a deep learning approach reduces the error rate in the observed time series point forecasts and significantly increases the explanatory power of the computed ensemble model. On average, over the 21 observed European office markets, the meta model outperforms both individual models. Hence, combining classical statistical forecasting methods and modern deep learning approaches yields more accurate and consistent forecasts. As a result, the study on forecasting European office market prime rents confirms heterogeneity of real estate markets. It also demonstrates that combining the forecast of different models can reduce uncertainty and is a good way to simultaneously approach office rent forecasting in multiple markets. Despite the simplicity of the variable structure and its comparably atheoretical characteristics, the proposed framework demonstrates superior properties in forecasting commercial real estate rents.

2.8 Appendix

Table 2.6: Augmented Dickey-Fuller Tests

#	City	P-Value [lag 0]	P-Value [lag 1]	P-Value [lag 2]	P-Value [lag 3]
1	Vienna	0,01	0,01	0,01	0,01
4	Helsinki	0,01	0,01	0,01	0,01
5	Lille	0,01	0,01	0,01	0,01
6	Lyon	0,01	0,01	0,01	0,01
7	Marseille	0,01	0,01	0,01	0,01
8	Paris Ile-de-France	0,01	0,01	0,01	0,01
9	Berlin	0,01628	0,01	0,01	0,01
10	Cologne	0,01	0,01	0,01	0,01
11	Dusseldorf	0,022642	0,01	0,01	0,01
12	Frankfurt am Main	0,01	0,013624	0,01	0,01
13	Hamburg	0,029201	0,01	0,01	0,01
14	Munich	0,01	0,01	0,01	0,01
15	Dublin	0,081857	0,090115	0,057765	0,01
16	Milan	0,027285	0,020568	0,01	0,01
17	Rome	0,010045	0,01	0,01	0,01
18	Amsterdam	0,01	0,01	0,01	0,01
19	Rotterdam	0,01	0,01	0,01	0,01
20	The Hague	0,01	0,01	0,01	0,01
22	Lisbon	0,144802	0,012201	0,01	0,01
23	Barcelona	0,061481	0,01	0,01	0,01
24	Madrid	0,088185	0,01	0,01	0,01

Notes: The null hypothesis of the Augmented Dickey-Fuller test is that the data are non-stationary. The null hypothesis is rejected if the p-value < 0.1 indicating that the used time series data is stationary and has no unit root. All p-values indicate that there is no unit root in the used data. The null hypothesis for Lisbon is accepted for a p-value < 0.15.

Table 2.7: Error Metrics Overview

Error Metric	Equation	Description
Mean Absolute Error	$MAE = \frac{\sum_{i=1}^{N}\lvert Y_i - \hat{Y}_i\rvert}{N}$	Average absolute deviation of predicted to actual values
Mean Absolute Scaled Error	$MASE = \frac{\sum_{t=1}^{N}\lvert e_t\rvert}{\frac{1}{n-1}\sum_{i=2}^{n}\lvert Y_i - Y_{i-1}\rvert}$	Average absolute deviation of predicted to actual values scaled by the in-sample deviation of actual values
Mean Absolute Percentage Error	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left\lvert\frac{Y_i - \hat{Y}_i}{Y_i}\right\rvert$	Average absolute deviation of predicted to actual values expressed as a ratio (in percent)
Mean Arctangent Absolute Percentage Error	$MAAPE = \frac{1}{N}\sum_{i=1}^{N}\tan^{-1}\left(\frac{Y_i - \hat{Y}_i}{Y_i}\right)$	Average absolute deviation of predicted to actual values expressed as an angle (in percent)
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}$	Average squared deviation of predicted to actual values, penalizes high deviations
Coefficient of Determination	$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$	Variation that can be explained by the model, goodness of fit

Table 2.8: ARIMA Model's Stage 2 Forecast Error Metrics (Unbalanced Panel, Real Rental Growth)

City	MAE	MASE	MAAPE	RMSE	RSQ
Vienna	0,0174	2,5231	56,2574	0,0187	0,4226
Helsinki	0,0145	0,9049	65,2406	0,0146	0,9995
Lille	0,0137	0,4529	39,5683	0,0236	0,2225
Lyon	0,0053	0,1962	21,6381	0,0072	0,9979
Marseille	0,0843	3,7629	105,4822	0,093	0,7307
Paris Ile-de-France	0,0303	1,9855	53,5203	0,0342	0,7836
Berlin	0,0486	2,7387	81,7631	0,0526	0,3788
Cologne	0,0098	0,2381	38,4055	0,0141	0,9956
Dusseldorf	0,0034	0,5005	43,6312	0,0042	0,6623
Frankfurt am Main	0,0191	1,1444	62,9327	0,0206	0,9599
Hamburg	0,0151	0,5521	35,435	0,0181	0,9858
Munich	0,0028	0,2604	37,1634	0,0034	0,9741
Dublin	0,0293	1,1647	42,3611	0,0443	0,6626
Milan	0,0304	1,7797	84,8693	0,033	0,7616
Rome	0,0093	0,4003	51,7786	0,0103	0,9973
Amsterdam	0,024	2,1625	100,1345	0,0291	0,9583
Rotterdam	0,0148	1,8354	97,582	0,0155	0,6728
The Hague	0,0053	0,6053	80,8757	0,0076	0,429
Lisbon	0,1021	3,5569	99,7833	0,1089	0,094
Barcelona	0,0506	2,2341	94,6278	0,0527	0,8684
Madrid	0,0036	0,2148	20,9531	0,0037	0,9805

Notes: The displayed error metrics are calculated on year on year growth rates of the original office prime rent series. The MAAPE is displayed in percentage points.

Table 2.9: N-BEATS Model's Stage 2 Forecast Error Metrics (Unbalanced Panel, Real Rental Growth)

City	MAE	MASE	MAAPE	RMSE	RSQ
Vienna	0,009	1,3016	35,2716	0,0099	0,8741
Helsinki	0,0063	0,3919	42,7174	0,0072	0,9387
Lille	0,0171	0,5662	74,2725	0,0197	0,2498
Lyon	0,01	0,3656	65,9632	0,0112	0,9108
Marseille	0,0086	0,3848	55,0491	0,0094	0,9779
Paris Ile-de-France	0,0356	2,3328	57,8914	0,0404	0,748
Berlin	0,0394	2,2216	73,945	0,0429	0,6566
Cologne	0,0318	0,7738	63,516	0,0348	0,9751
Dusseldorf	0,0049	0,723	63,4432	0,0053	0,4995
Frankfurt am Main	0,0078	0,4658	37,0157	0,0087	0,9568
Hamburg	0,0176	0,6456	36,7915	0,0186	0,9569
Munich	0,0071	0,6547	64,4032	0,0087	0,9424
Dublin	0,027	1,0747	47,3329	0,0391	0,0574
Milan	0,0066	0,3866	44,567	0,0094	0,925
Rome	0,0099	0,4273	50,5984	0,0117	0,9956
Amsterdam	0,0222	1,9984	105,5747	0,0225	0,9428
Rotterdam	0,014	1,7362	100,7256	0,0149	0,6597
The Hague	0,0053	0,6099	86,116	0,0064	0,693
Lisbon	0,047	1,6364	74,0775	0,0512	0,7262
Barcelona	0,0387	1,7111	80,8622	0,0433	0,7802
Madrid	0,0271	1,6268	78,5642	0,0318	0,6459

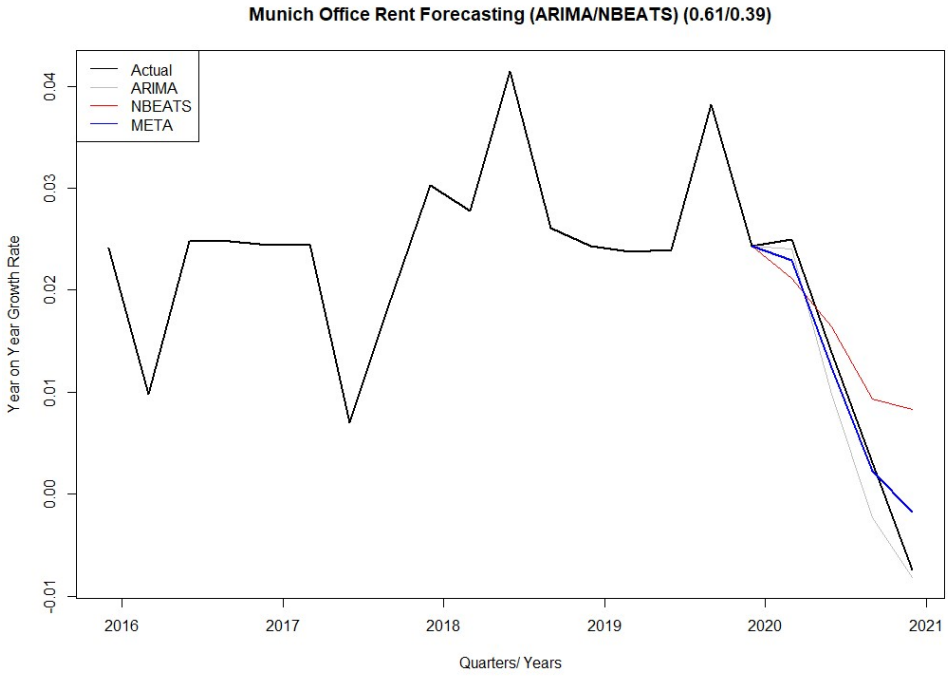
Notes: The displayed error metrics are calculated on year on year growth rates of the original office prime rent series. The MAAPE is displayed in percentage points.

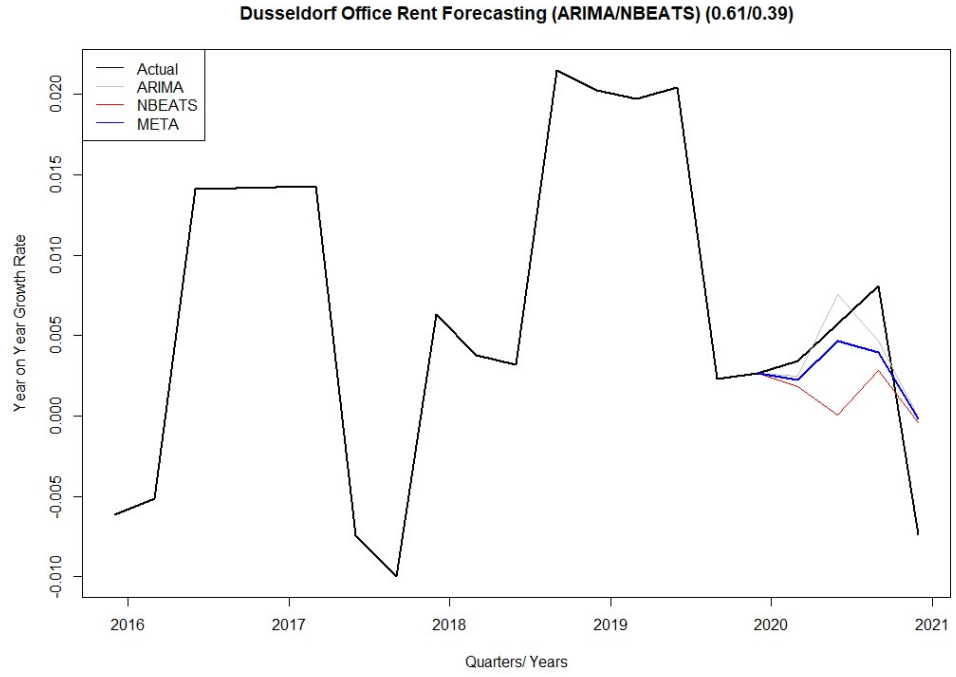
Table 2.10: Meta Model's Stage 2 Forecast Error Metrics (Unbalanced Panel, Real Rental Growth)

City	MAE	MASE	MAAPE	RMSE	RSQ
Vienna	0,0141	2,0462	49,0116	0,015	0,8276
Helsinki	0,0098	0,6088	53,8765	0,0104	0,9947
Lille	0,015	0,4971	59,1326	0,0214	0,2322
Lyon	0,0059	0,2179	41,6628	0,0067	0,9848
Marseille	0,0489	2,1798	94,1905	0,0545	0,8043
Paris Ile-de-France	0,0315	2,0588	51,7141	0,0364	0,7836
Berlin	0,045	2,5368	78,9275	0,0487	0,7583
Cologne	0,0069	0,1677	19,4218	0,0073	0,9977
Dusseldorf	0,0034	0,5047	43,9716	0,0042	0,8824
Frankfurt am Main	0,0147	0,8794	55,583	0,0156	0,9747
Hamburg	0,016	0,585	35,8988	0,018	0,9802
Munich	0,0025	0,2338	27,8759	0,0031	0,9705
Dublin	0,0279	1,1109	38,1426	0,0422	0,695
Milan	0,0211	1,2357	72,2685	0,0235	0,9233
Rome	0,0044	0,1894	42,2492	0,0055	0,9967
Amsterdam	0,0107	0,9658	80,0794	0,0127	0,8734
Rotterdam	0,0145	1,7967	100,3041	0,0149	0,8353
The Hague	0,0036	0,4173	59,7435	0,0062	0,5581
Lisbon	0,0805	2,807	91,418	0,086	0,3681
Barcelona	0,046	2,0299	89,8268	0,0487	0,9294
Madrid	0,012	0,7234	55,343	0,0136	0,9816

Notes: The displayed error metrics are calculated on year on year growth rates of the original office prime rent series. The MAAPE is displayed in percentage points.

Figure 2.7: Visualisation of the Error Reduction via the Combination of Methods in the Meta model (Unbalanced Panel, Real Rental Growth)




Dusseldorf Office Rent Forecasting (ARIMA/NBEATS) (0.61/0.39)
Actual
ARIMA
NBEATS
META
Year on Year Growth Rate
0.020
0.015
0.010
0.005
0.000
-0.005
-0.010
2016
2017
2018
2019
2020
2021
Quarters/ Years

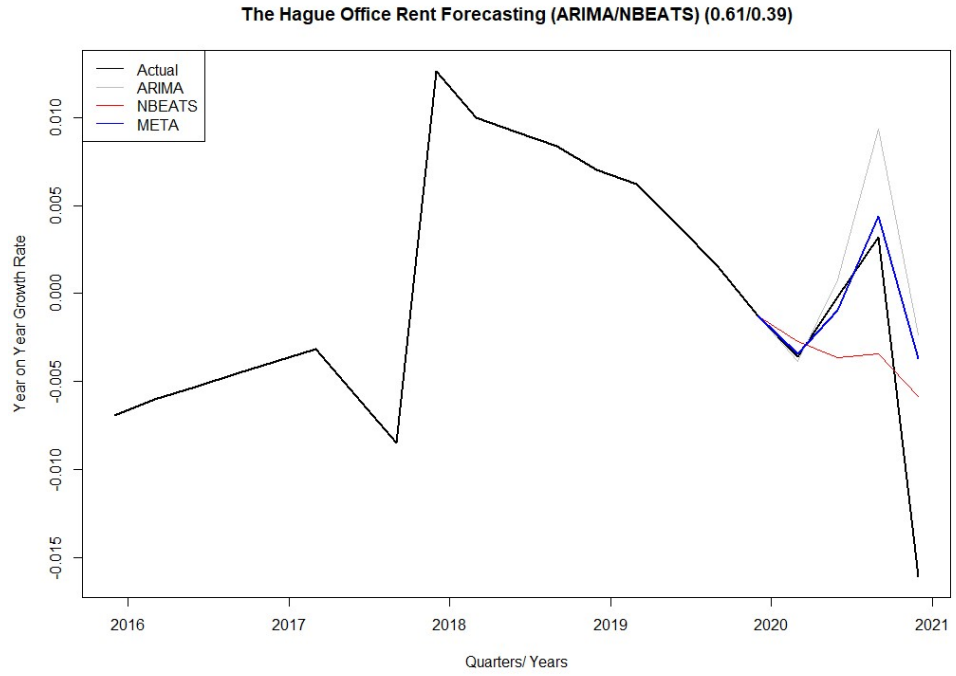
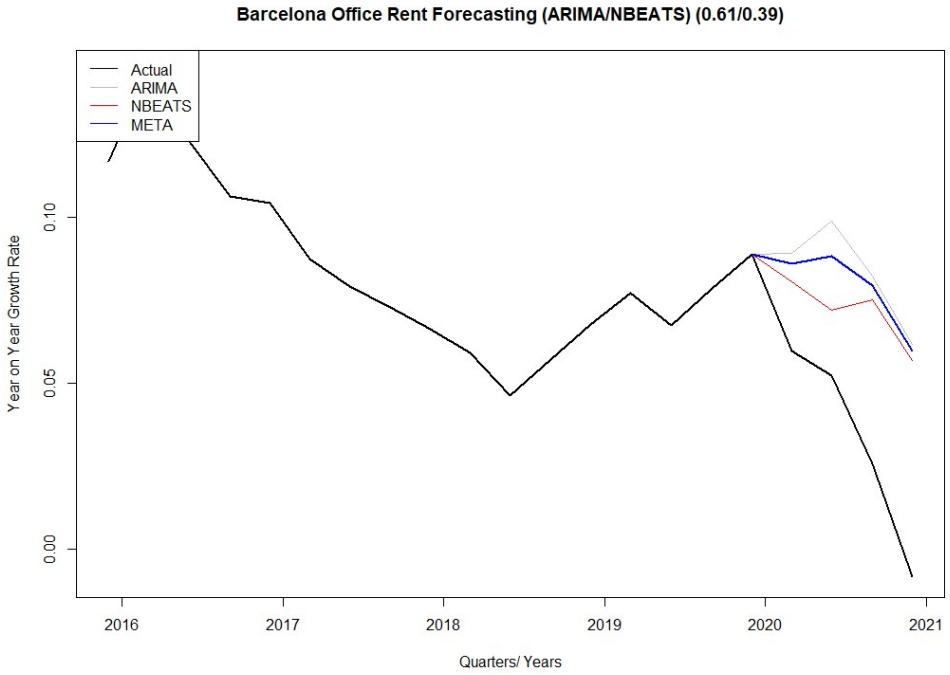
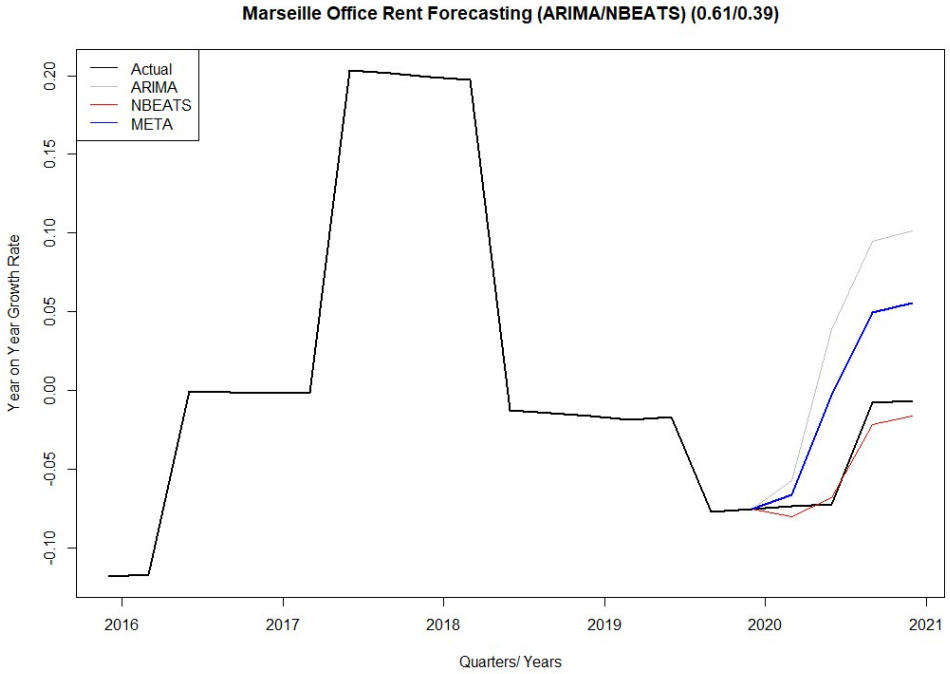

The Hague Office Rent Forecasting (ARIMA/NBEATS) (0.61/0.39)
Actual
ARIMA
NBEATS
META
Year on Year Growth Rate
0.010
0.005
0.000
-0.005
-0.010
-0.015
2016
2017
2018
2019
2020
2021
Quarters/ Years

Figure 2.8: Visualisation of the Modelling Failure via the Combination of Methods in the Meta model (Unbalanced Panel, Real Rental Growth)





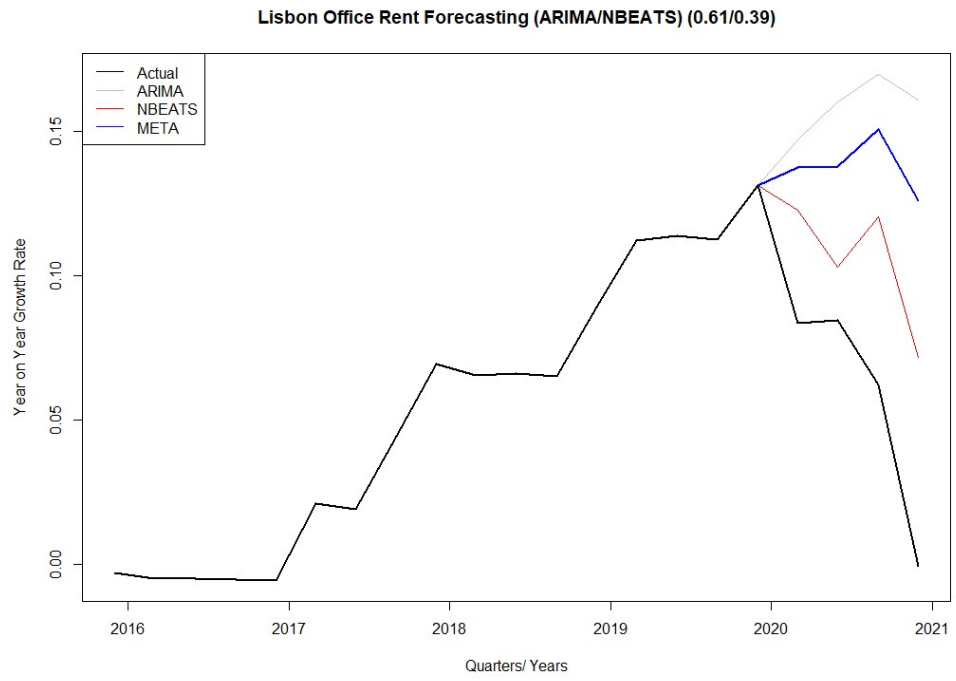


Figure 2.9: Mean Absolute Error Comparison of One Step Forecast to Multi-Step Forecasts (balanced panel, real rental growth)

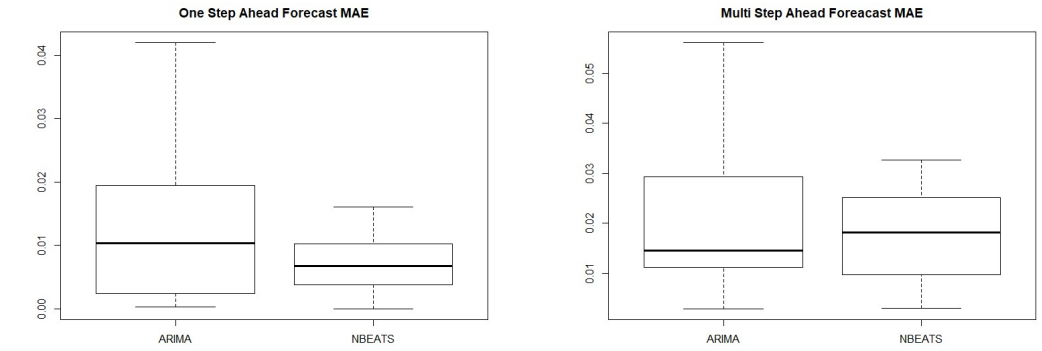


Figure 2.10: Error-based comparison of model performance (Balanced Panel, Real Rental Growth)

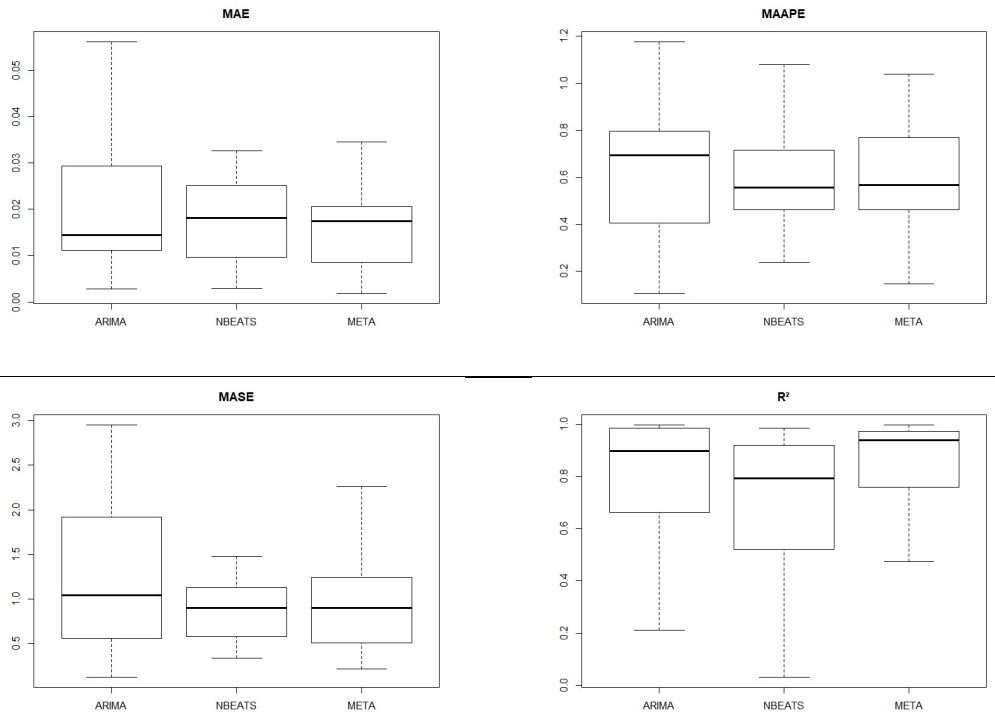
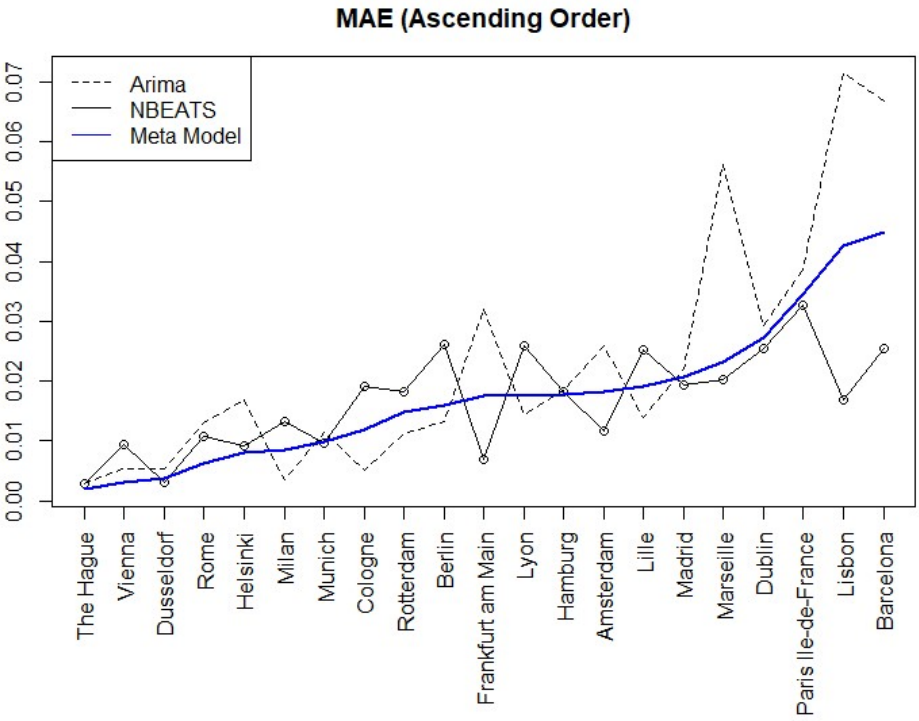


Figure 2.11: Mean Absolute Error by City in Ascending Order (balanced panel, real rental growth)



2.9 References

Antipov, E.A. and 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.

Armstrong, J.S. (2001). Combining forecasts. *Principles of forecasting: A Handbook for Researchers and Practitioners*, 30.

Atiya, A.F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, 36(1), 197–200.

Boon Foo, N.G. and Higgins, D. (2015). Modelling the commercial property market: an empirical study of the singapore office market. *Pacific Rim Property Research Journal*, 13(2), 176-193.

Box, G.E.P., Geurts, M. and Jenkins, G.M. (1977). Time series analysis: Forecasting and control. *Journal of Marketing Research*, 14(2), 269.

Brooks, C. and Tsolacos, S. (2010). Real estate modelling and forecasting. *Cambridge University Press*, Cambridge.

Brounen, D. and Jennen, M. (2009a). Asymmetric properties of office rent adjustment. *Journal of Real Estate Finance and Economics*, 39(3), 336-358.

Brounen, D. and Jennen, M. (2009b). Local office rent dynamics: A tale of ten cities. *Journal of Real Estate Finance and Economics*, 39(4), 385-402.

Bruneau, C. and Cherfouh, S. (2015). Long-run equilibrium for the greater Paris office market and short-run adjustments. *Journal of Property Research*, 32(4), 301-323.

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

Chaplin, R. (1998). An ex post comparative evaluation of office rent prediction models. *Journal of Property Valuation and Investment*, 16(1), 21–37.

Chaplin, R. (2000). Predicting real estate rents: walking backwards into the future. *Journal of Property Investment and Finance*, 18(3), 352–370.

Crawford, G.W. and Fratantoni, M.C. (2003). Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices. *Real Estate Economics*, 31(2), 223–243.

Dąbrowski, J. and Adamczyk, T. (2010). Application of GAM additive non-linear models to estimate real estate market value. *Geomatics and Environmental Engineering*, 4, 55–62.

Frew, J. and Jud, D. (1988). The vacancy rate and rent levels in the commercial office market. *Journal of Real Estate Research*, 3(1), 1–8.

Gardiner, C. and Henneberry, J. (1989). The development of a simple regional office rent prediction model. *Journal of Valuation*, 7(1), 36–52.

Giussani, B., Hsia, M. and Tsolacos, S. (1993). A comparative analysis of the major determinants of office rental values in europe. *Journal of Property Valuation and Investment*, 11(2), 157–173.

Gooijer, J.G. de and Hyndman, R.J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3), 443–473.

Granger, C. and Newbold, P. (1986). Forecasting economic time series, 2nd ed., Academic Press.

Hekman, J.S. (1985). Rental price adjustment and investment in the office market, *Real Estate Economics*, 13(1), 32–47.

Hendershott, P., Lizieri, C. and MacGregor, B. (2010). Asymmetric adjustment in the city of London office market. *Journal of Real Estate Finance and Economics*, 41(1), 80-101.

Hendershott, P., MacGregor, B. and Tse, R. (2002a). Estimation of the rental adjustment process. *Real Estate Economics*, 3(2), 165-183.

Hendershott, P., MacGregor, B. and White, M. (2002b). Explaining real commercial rents using an error correction model with panel data. *Journal of Real Estate Finance and Economics*, 24(1/2), 59-87.

Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.

Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3), 669–679.

Lantz, B. (2019). Machine learning with R: Expert techniques for predictive modeling, 3rd Edition, Packt Publishing Ltd.

Li, R.Y., Fong, S., Chong, K.W. (2016). Forecasting REITs and stock indices: Group method of data handling neural network approach. *Pacific Rim Property Research Journal*, 23(2), 123-160.

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

McCartney, J. (2012). Short and long-run rent adjustment in the Dublin office market. *Journal of Property Research*, 29(3), 201-226.

McGough, T. and Tsolacos, S. (1995). Forecasting commercial rental values using ARIMA models. *Journal of Property Valuation and Investment*, 13(5), 6–22.

Milunovich, G. (2020). Forecasting Australia's real house price index: A comparison of time series and machine learning methods. *Journal of Forecasting*, 39(7), 1098–1118.

Mouzakis, F. and Richards, D. (2007). Panel data modelling of prime office rents: A study of 12 major european markets. *Journal of Property Research*, 24(1), 31–53.

Oreshkin, B.N., Carpov, D., Chapados, N. and Bengio, Y. (2019). N-Beats: Neural basis expansion analysis for interpretable time series forecasting. available at: https://arxiv.org/pdf/1905.10437.

Oreshkin, B.N., Dudek, G., Pełka, P. and Turkina, E. (2020). N-Beats: Neural network for mid-term electricity load forecasting. available at: https://arxiv.org/pdf/2009.11961.

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

Rosen, K.T. (1984). Toward a model of the office building sector. *Real Estate Economics*, 12(3), 261–269.

Shilling, J.D., Sirmans, C.F. and Corgel, J.B. (1987). Price adjustment process for rental office space. *Journal of Urban Economics*, 22(1), 90–100.

Stevenson, S. (2007). A comparison of the forecasting ability of ARIMA models. *Journal of Property Investment and Finance*, 25(3), 223–240.

Stevenson, S. and McGarth, O. (2003). A comparison of alternative rental forecasting models: empirical tests on the London office market. *Journal of Property Research*, 20(3), 235–260.

Tse, R.Y. (1997). An application of the ARIMA model to real-estate prices in Hong Kong. *Journal of Property Finance*, 8(2), 152–163.

Wheaton, W.C. and Torto, R.G. (1988). Vacancy rates and the future of office rents. *Real Estate Economics*, 16(4), 430–436.

Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159–175.

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 Jeffery Fisher 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; Fisher et al., 1999; Matysiak & Wang, 1995; Edelstein & Quan, 2006; Cannon & 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 & Spiess, 2017; Mayer et al., 2019; Bogin & Shui, 2020; Hong et al., 2020; Pace & Hayunga, 2020; Lorenz et al., 2022; and Deppner & 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"[2] (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 & 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

[2] 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 & 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 & 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 & 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 non-linear 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 & Pokryshevskaya, 2012; Mullainathan & Spiess, 2017; Baldominos et al., 2018; Hu et al., 2019; Mayer et al., 2019; Bogin & Shui, 2020; Pace & Hayunga, 2020; Cajias et al., 2021; Rico-Juan and Taltavull de La Paz, 2022; Lorenz et al., 2022; and Deppner & 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 (Mayer et al., 2019; McCluskey et al., 2013). 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[3] (NPI) on the asset level spanning 1Q 1978 through 1Q 2021. To be included in the NPI, a property must be

[3] The NPI is a quarterly index tracking the performance of core institutional property markets in the U.S.

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,[4]
iii. accounted for in compliance with the NCREIF Market Value Accounting Policy,[5]
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 appraisal accuracy. This is no surprise as transaction prices are often determined three to

[4] This includes commingled real estate funds (open and closed-end), separate accounts, individual accounts, private REITs, REOCs, and joint-venture partnerships.
[5] For further details, refer to the NCREIF PREA Reporting Standards at www.reisus.org.

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[6]. 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.[7] 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[8] 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.[9] 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 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

[6] 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%.

[7] 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.

[8] 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.

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

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%[10] 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[11] that meet all the previously outlined criteria to be included in the study. Relative to the initial sample size this constitutes a heavy data loss, which again emphasizes the problem of data availability as mentioned earlier.[12] 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.

[10] The remaining 7% result mainly from missing or incomplete addresses.

[11] Of which 1,904 are apartments, 2,337 are industrial, 2,056 are office and 836 are retail.

[12] 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.

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.

3.4.2 Appraisal Error

NCREIF follows the definition of market value as stated in section 3.3 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 Equation 8 and 9, respectively.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Sale\ Price_{i,t0} - Appraised\ Value_{i,t-2}}{Appraised\ Value_{i,t-2}}\right| \tag{8}$$

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{Sale\ Price_{i,t0} - Appraised\ Value_{i,t-2}}{Appraised\ Value_{i,t-2}} \tag{9}$$

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. 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 10, which corresponds to the log of the percentage appraisal error, however, keeping the signs.

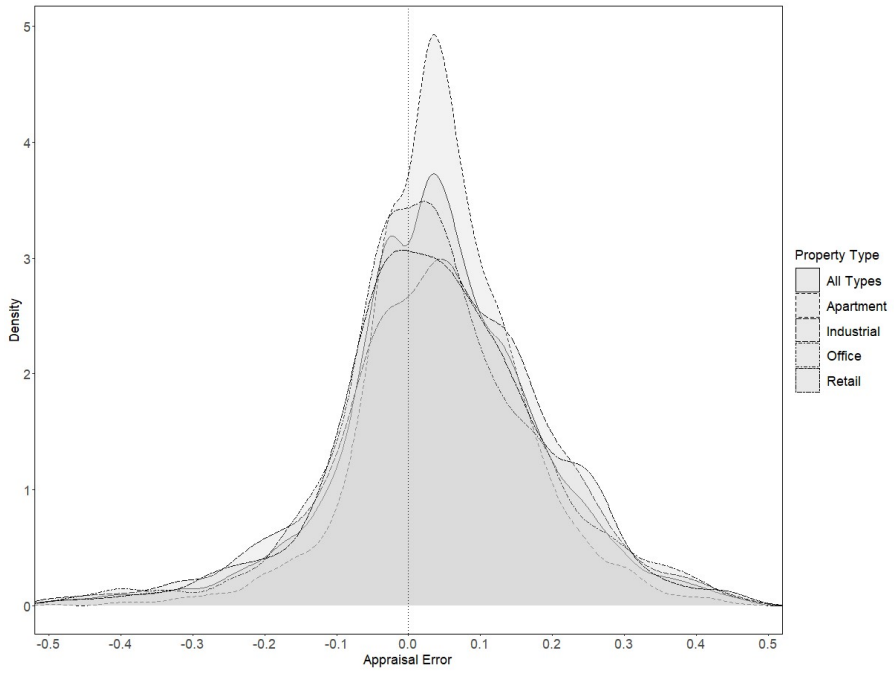
$$\mathrm{Y} = [\, SP - MV \,] \tag{10}$$

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

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

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

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.

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 11.

$$X_a = [\, M_{t0}\, M_{t-1}\, T\, C_{t0}\;\; C_{t-1}\,] \tag{11}$$

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 12. 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. The covariates included in component *b* can thus be summarized as in Equation 12.

$$X_b = [\, S\ P\ F\ L\ E_{t0}\ \ E_{t-1}\,] \tag{12}$$

Our models incorporate 50 explanatory variables reflecting the main information used in the traditional appraisal methods discussed in section 3.3.1 (i.e., income approach, sales comparison approach, cost approach). The input-output relationship is summarized in Equation 13.

$$Y \sim [\, X_a\ X_b\,] \tag{13}$$

Table 3.2 provides a summary statistic of all numeric 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 transactions 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.

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	0.00	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 Categorial 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.

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[i] (e.g., Mayer et al., 2019; Lorenz et al., 2022; Deppner and Cajias, 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 kth 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 10 as $\sigma_{Appraisal}$ and the unexplained residual variation of our boosting estimator as $\sigma_{Boosting}$, shown in Equations 13 through 15.

$$\sigma_{Appraisal} = \sqrt{\frac{\sum_{i=1}^{n}|y - \bar{y}|^2}{n}} \tag{13}$$

$$\sigma_{Boosting} = \sqrt{\frac{\sum_{i=1}^{n}|\varepsilon - \bar{\varepsilon}|^2}{n}} \tag{14}$$

$$\varepsilon = y - \hat{y} \tag{15}$$

Our null hypothesis can thus be stated as:

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

This is the case when the condition in Equation is fulfilled.

$$H_0: \frac{\sigma_{Appraisal}}{\sigma_{Boosting}} \leq 1 \tag{16}$$

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 17.

$$H_1: \frac{\sigma_{Appraisal}}{\sigma_{Boosting}} > 1 \tag{17}$$

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 real estate appraisals.

Table 3.4: Absolute Percentage Error between Sales Price and Manual Appraisal Value

	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 Sales Price and Manual Appraisal Value

	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	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 ten-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.

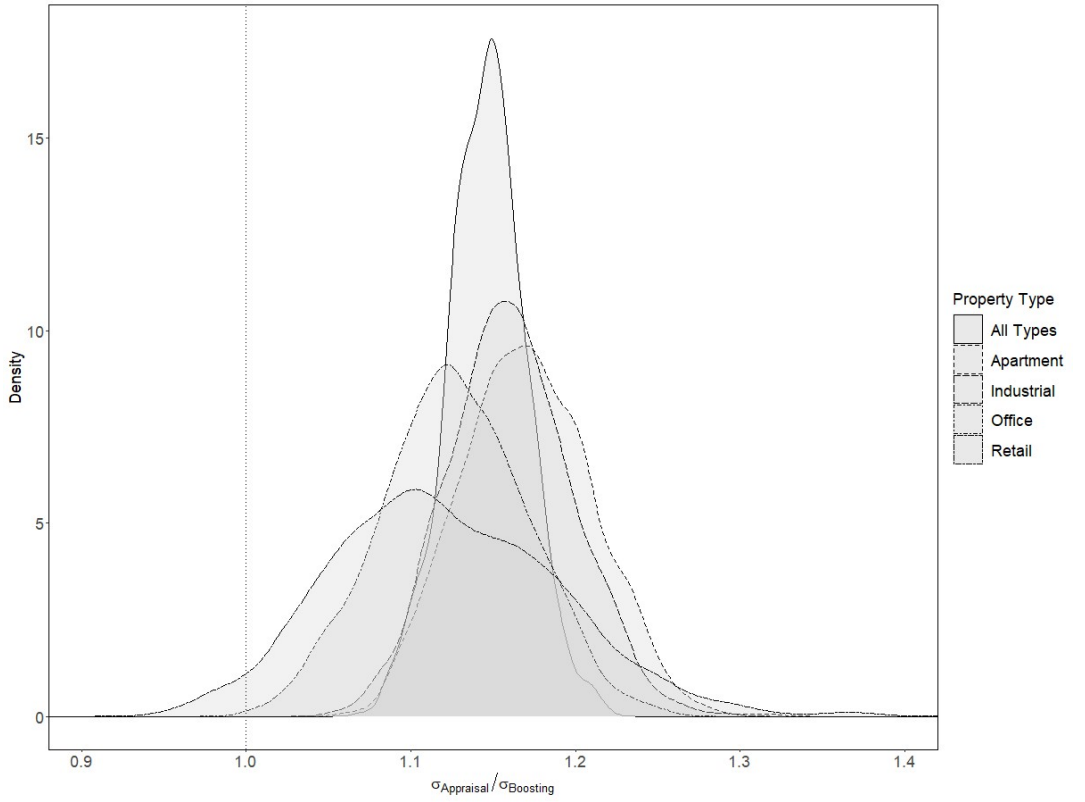
Table 3.6: Residual Standard Deviation

	$\sigma_{Appraisal}$	$\sigma_{Boosting}$	$R^2_{Boosting}$	$\frac{\sigma_{Appraisal}}{\sigma_{Boosting}}$
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.ii 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 10 can be

Figure 3.2: Bootstrap Distribution of Model Performance



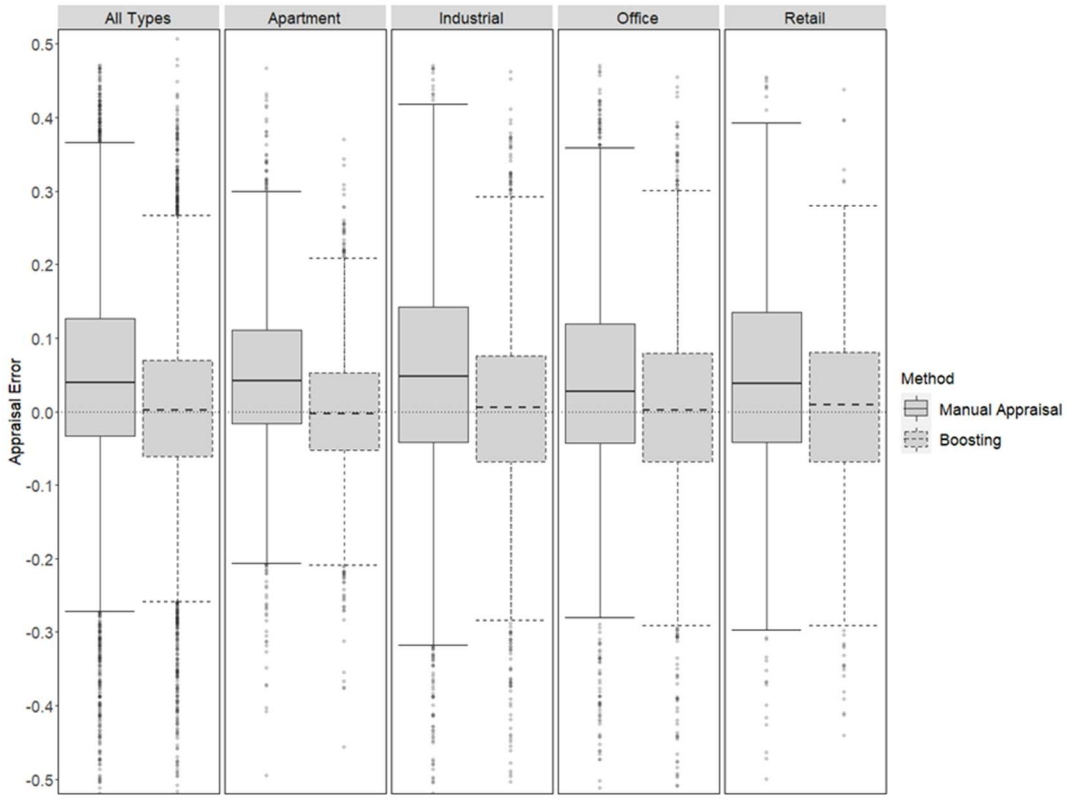
Notes: The density plot shows the bootstrap distribution of the model performance for all five models using 1,000 random bootstrap samples. A performance improvement occurs whenever the ratio $\frac{\sigma_{Appraisal}}{\sigma_{Boosting}} > 1$, as indicated by the dotted horizontal line. The area to the 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.

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.

We also see a relationship between the homogeneity of asset classes and the performance improvement. Relatively homogenous property types (i.e., apartments, industrial) benefit

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.

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 section 3.4.1, Table 3.7 and Table 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 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 3.7 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.

Table 3.7: Absolute Percentage Error between Sales Price and Boosting Appraisal Value

	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 Sales Price and Boosting Appraisal Value

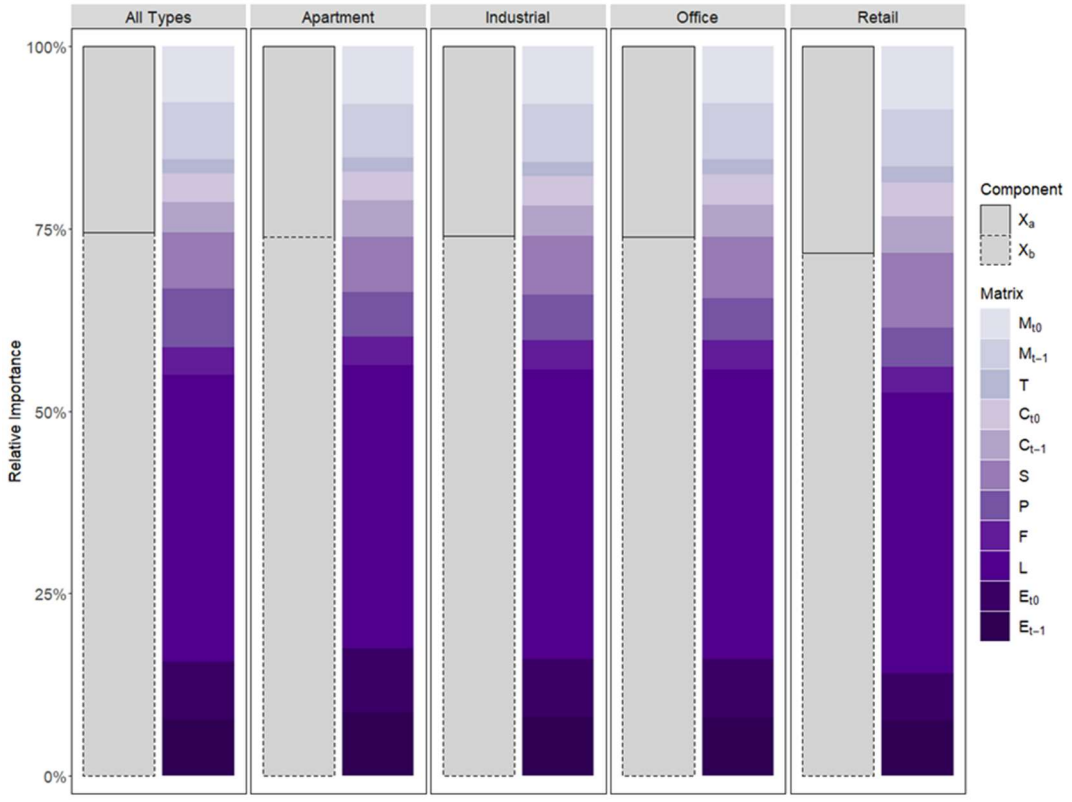
	All Types (N = 7,133)				Apartment (N = 1,904)				Industrial (N = 2,337)				Office (N = 2,056)				Retail (N = 836)			
Year	**MdAPE**	**MAPE**	**t-Stat.**		**MdAPE**	**MAPE**	**t-Stat.**		**MdAPE**	**MAPE**	**t-Stat.**		**MdAPE**	**MAPE**	**t-Stat.**		**MdAPE**	**MAPE**	**t-Stat.**	
1997	0.29%	0.31%	0.25		-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:
$p < 0.01$ '***', $p < 0.05$ '**', $p < 0.1$ '*'.

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 3.4.3, 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).

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.

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.

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 section 3.5.1, Beracha et al. (2019) demonstrate that, on average, the option value of vacant space is overvalued, which is not surprising 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 closed-end funds or separate account properties. 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 and 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.

Endnotes

[i.] 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. 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.

[ii.] 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 ($\sigma_{Bagging}$ 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.

3.7 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., 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. https://www.reri.org/research/files/2018_Beracha-Downs-MacKinnon.pdf.* Accessed 17 June 2022.

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. *The 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. *Journal of 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. *The 22nd ACM SIGKDD International Conference*.

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, Chicago, Ill.

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. *Journal of 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. *Journal of Real Estate Economics,* Forthcoming.

Malpezzi, S. (2002). Hedonic pricing models: A selective and applied review. In O'Sullivan, T. and Gibb, K. (Eds.), *Housing Economics and Public Policy*, Wiley, Oxford, UK, 67–89.

Matysiak, G. A., & Wang, P. (1995). Commercial property market prices and valuations: Analysing the correspondence. *Property Investment Research Centre, Department of Property Valuation and Management, City University Business School, London*.

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. *Journal of Real Estate Economics*, *20*(2), 273–287.

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

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. *The 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., Correa-Morales, J., & 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*, 59–96.

R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing.*

Real Estate Lending and Appraisals, (2022). 12 Code of Federal Regulations (C.F.R.) § 34.42. https://www.ecfr.gov/current/title-12/chapter-I/part-34.

Rico-Juan, J. R., and Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonic 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, 65,* 649-674.

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*. http://hdl.handle.net/1765/7145. Accessed 18 April 2022

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*, 349–388.

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 that provide the analytical core of automated valuation models (AVMs) have demonstrated thus far unprecedented accuracy in estimating property prices. However, these techniques also face criticism as their mechanisms are considered black-boxes in the sense that an inherent comprehensibility of their predictions is impeded by the complexity of their architectures. For the practical application of such techniques, it is essential for professionals to have the ability to comprehend and interpret the predictions by these models. Moreover, research in this field has predominantly focused on the residential sector, while applications to the commercial domain remain scarce given limited data availability. The main contribution of this article is thus twofold: First, we extend the application of AVMs to commercial real estate markets, including the sectors industrial, office, and retail by training a deep neural network (DNN) on a unique sample of more than 400,000 property-quarter observations from the NCREIF Property Index (NPI). Second, we propose an advanced model-agnostic methodology, Shapley Additive Explanations (SHAP), to mitigate the trade-off between accuracy and interpretability in ML models and provide ex-post comprehensibility of the algorithm's prediction rules. In doing so, we furthermore aim to assess to which extent the prediction rules of the applied DNN follow an economic rationale and whether the proposed methods can add to the understanding of pricing processes in institutional investment markets by revealing non-linear and three-dimensional relationships in pricing dynamics of commercial real estate. The resulting implications of this study can support the decision making of appraisers and executives. In the long term, these techniques moreover have the potential to leverage efficiency in both markets and business processes by increasing the speed and scale of valuations, reducing transaction cost, and ultimately increasing transparency in pricing processes.

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 post hoc 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.

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 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

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

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.

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
Stabilized NOI	[$/SqFt]	2.60	2.28	-19.69	1.14	1.89	3.39	56.26
CapEx	[$/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

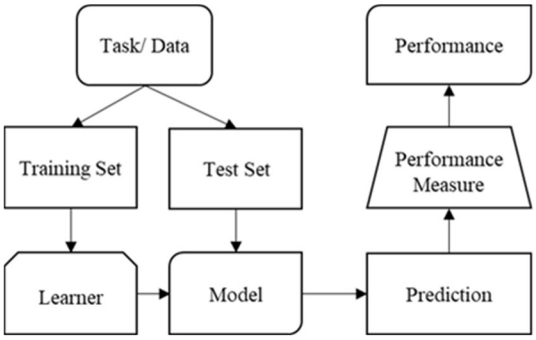
Table 4.3: Descriptive Statistics of Categorical Variables

	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%

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

Figure 4.1: General Overview of the Machine Learning Process



Source: Own illustration based on Lang et al. (2019)

the out-of-sample performance of the model is regarded as sufficient, it can be applied to an independent test dataset (i.e., unseen or future data) to make predictions.

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).

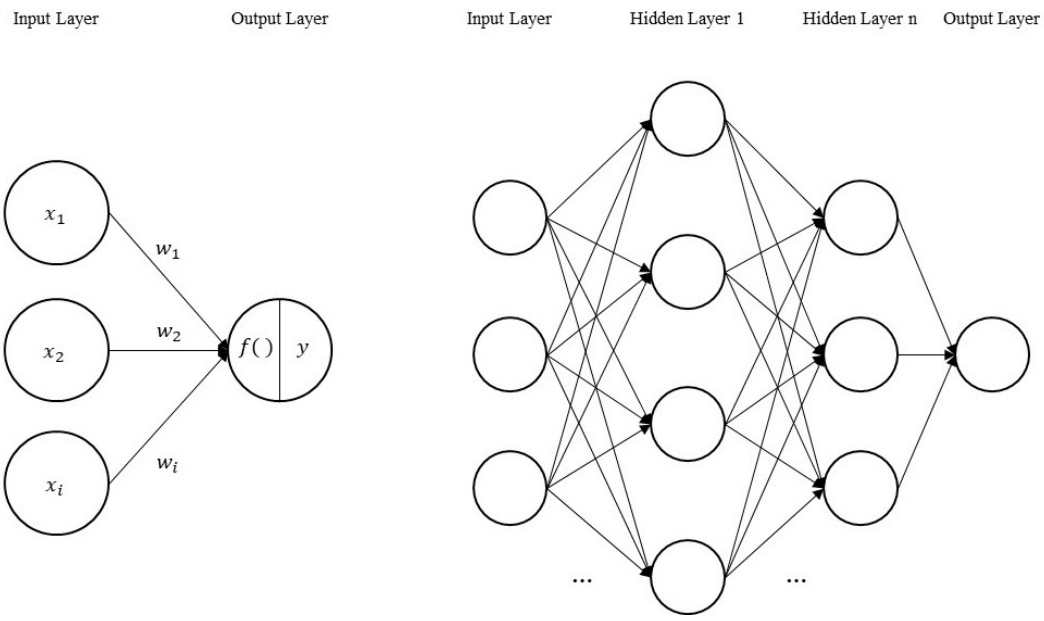
4.4.1 Machine Learning Approach – Artificial Neural Networks

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. Analogously, an ANN learns by adjusting the weights of the

Figure 4.2: Structure of Neural Networks

Panel A: Single-layer ANN **Panel B: Multi-layer DNN**



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 and formulate the prediction *y* as exhibited in Equation 18 below.

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{18}$$

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 more 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
ANN	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
ANN	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
ANN	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
ANN	0.97	0.10	0.03	22.94	2,139.41	46.25	0.67	0.88

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.

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 (i.e., illustrated 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

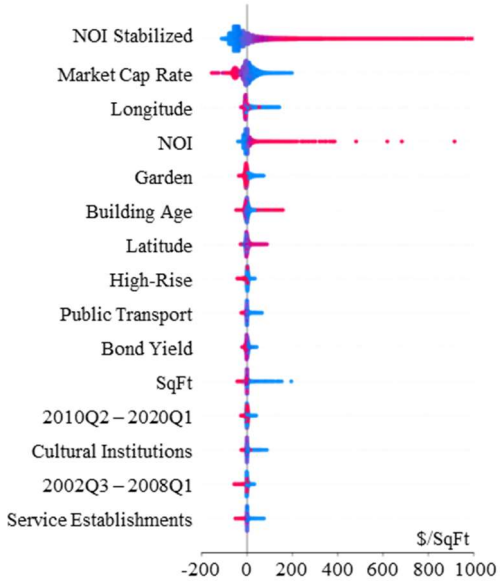
machine learning approach. It is important to note that our models do not incorporate inferential assumptions that can determine causal relationships. That is, the significance of the features is determined solely by the statistical relationships that the model identifies. Ideally, the statistical relationships 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.

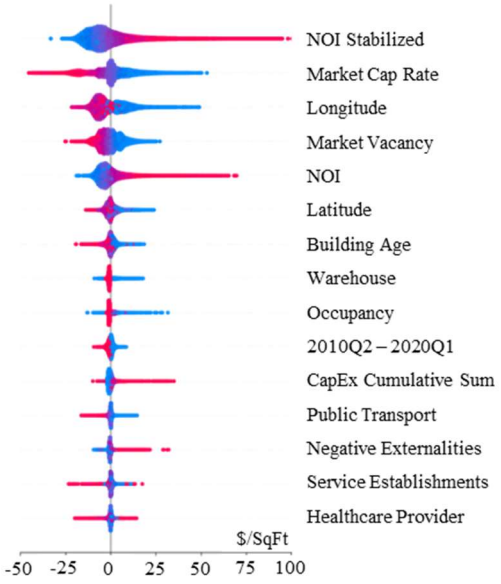
Figure 4.4, Panel A depicts the dependence plots of stabilized NOI and its impact on the market value prediction. A positive linear relationship for values greater than zero can be observed across all asset sectors, as expected market values increase with an increasing stabilized NOI. A negative stabilized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

Figure 4.3: SHAP Summary Plot (Top 15 Features)

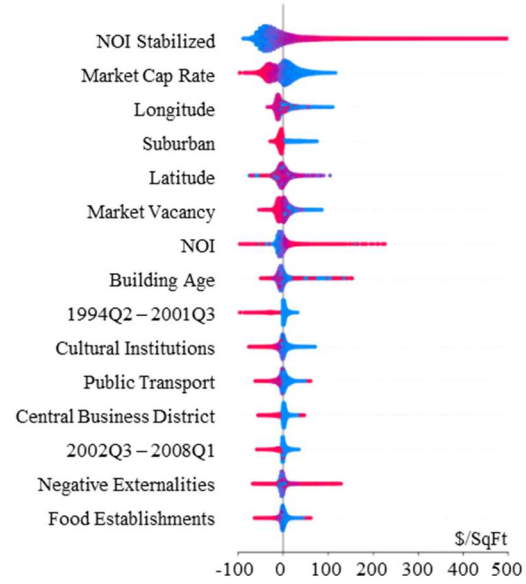
Panel A: Apartment



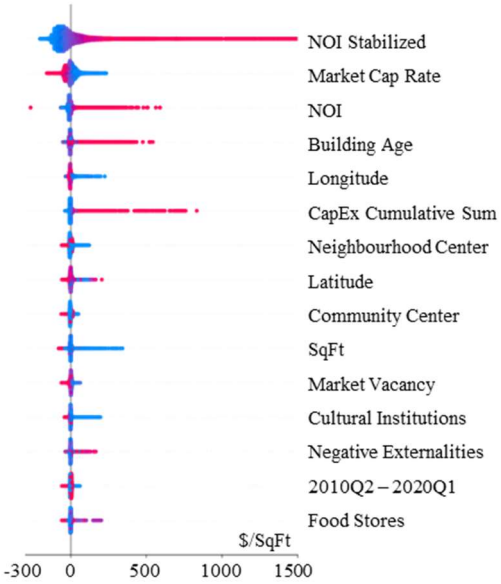
Panel B: Industrial

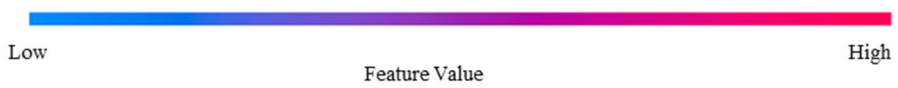


Panel C: Office



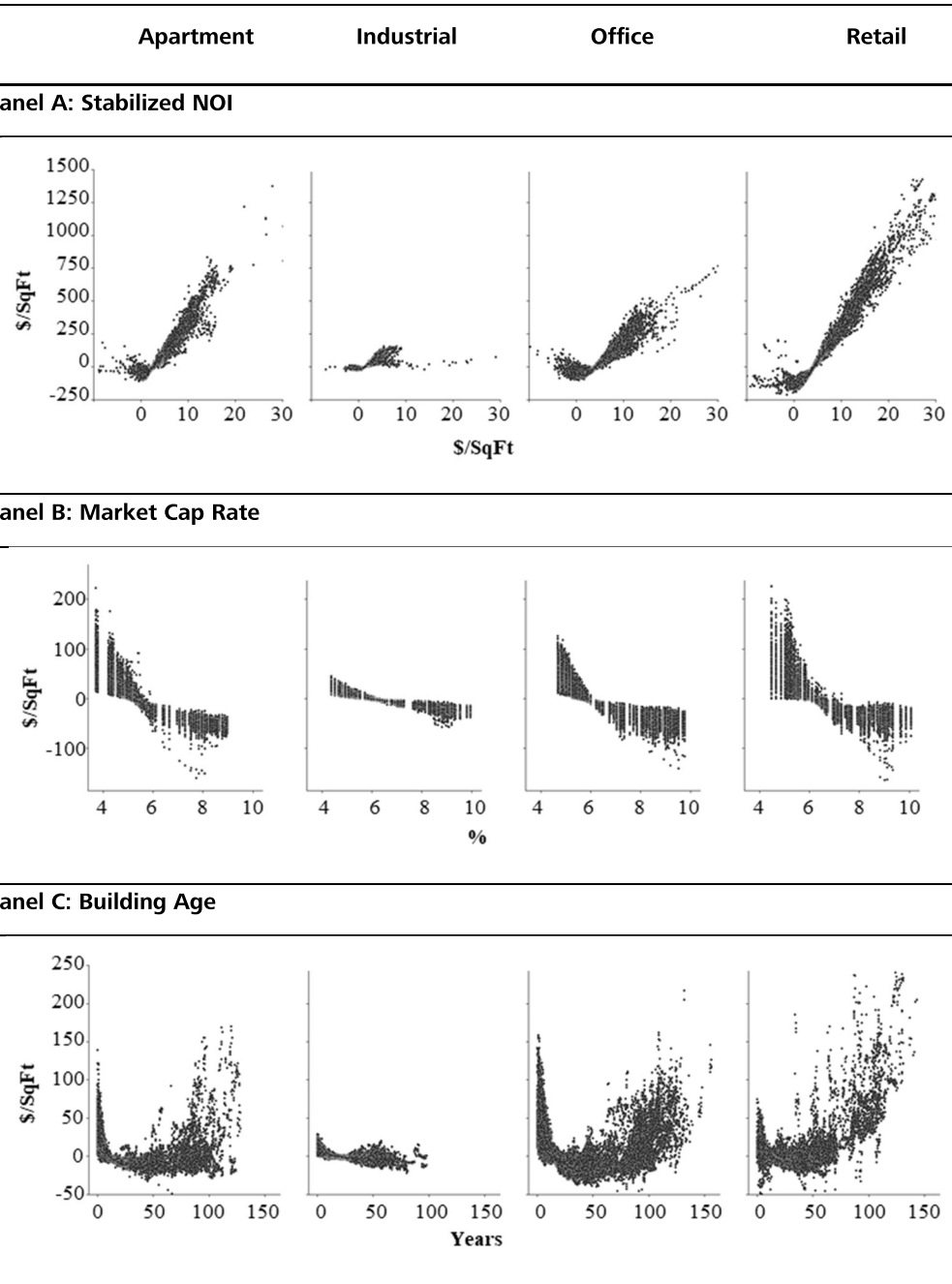
Panel D: Retail





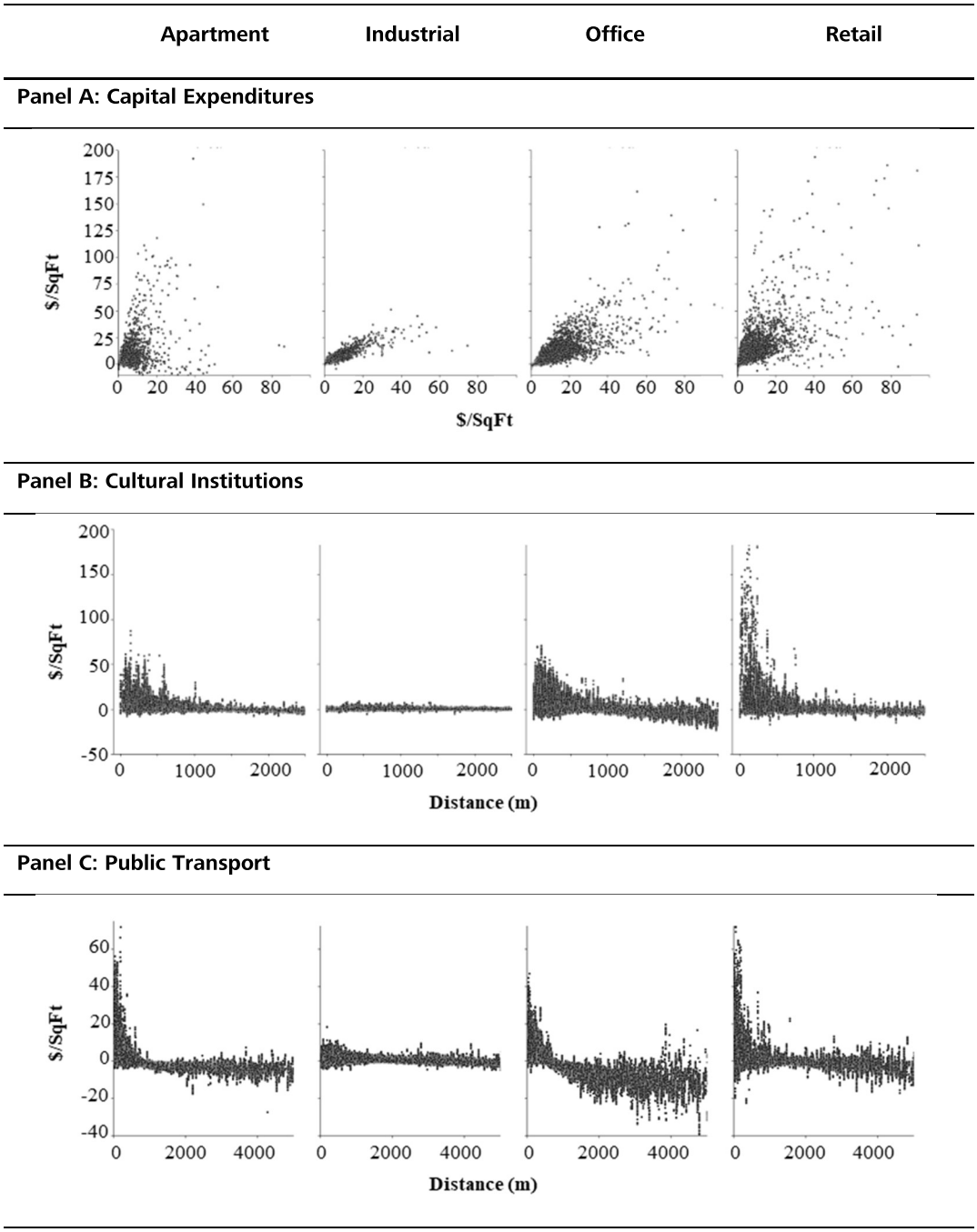
Panel C, we observe that this is also the case for the apartment sample and the office and retail properties. This U-shape seems to be less pronounced for industrial properties. The plot of industrial properties generally shows a lower building age, which can be attributed

Figure 4.4: SHAP Partial Dependence



to the nature of heavy industry use and the limited usability by third parties. 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 model's impact on the 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

Figure 4.5: SHAP Partial Dependence (2)



the model's prediction of the market value, Interestingly, retail properties near 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 value. 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 further 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.

Figure 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.

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

Panel A: Stabilized NOI vs. Occupancy **Panel B: Stabilized NOI vs. Cumulative Capex**

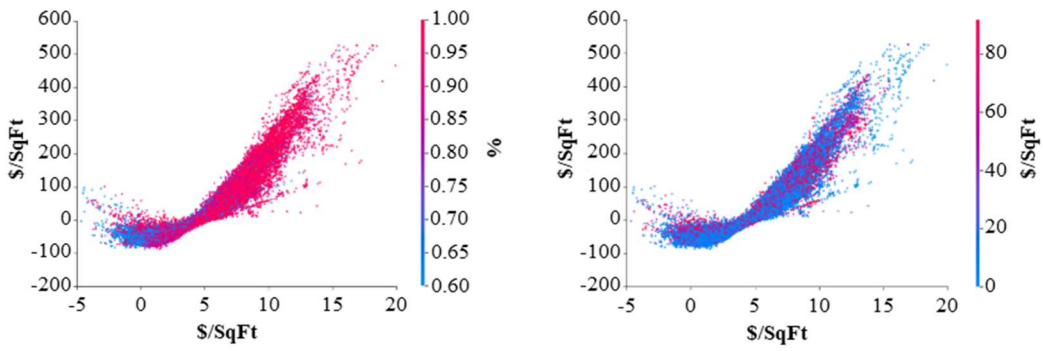
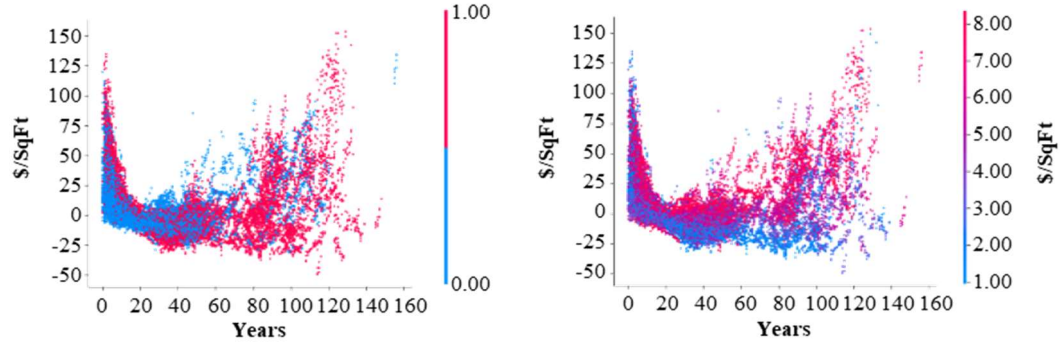


Figure 4.6, Panel A displays the interaction effect between occupancy and stabilized NOI, while Figure 4.6, Panel B shows the interaction effect between cumulative CapEx and stabilized NOI. The blue color on the graphs indicates low interaction feature values, while the red color indicates high interaction 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 4.7 analyzes the observed U-shaped pattern in the building age by inspecting interaction effects with both location (Panel A) and income

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

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

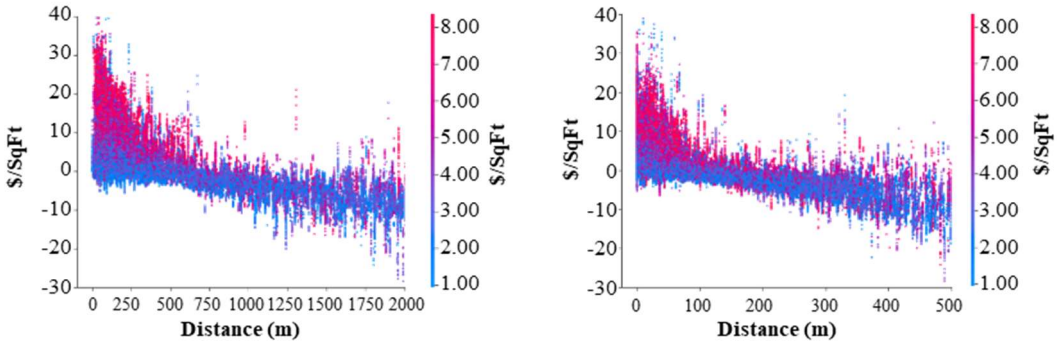


(Panel B). In suburban areas, the building age generally shows a negative relationship, as seen in Figure 4.7, Panel A. That is, older properties in suburban areas tend to have lower market values. From Figure 4.7, Panel B, we can deduce that properties for which high building ages are positively related 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 highest centrality in the CBD of a city creates high demand that generally leads to high 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. 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 that 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.

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

Panel A: Stabilized NOI vs. Public Transport	Panel B: Stabilized NOI vs. Food Establishments



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 contributions of all features are shown as 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: SHAP Force Plot

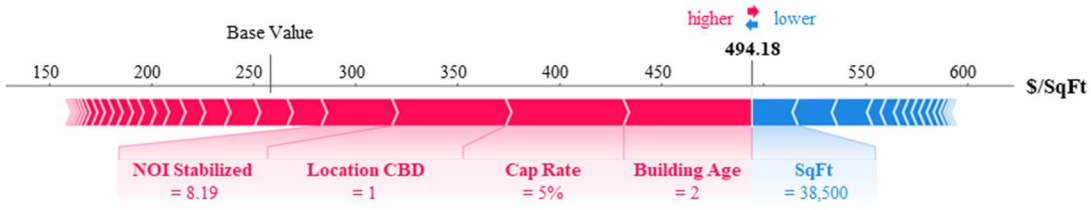


Figure 4.9 shows the composition of a market value prediction for an office property in Boston, Massachusetts. The prediction for this office property's market value is estimated to be 494.18 $/SqFt. The mean prediction (base value) of office market values in the sample is 258.53 $/SqFt. It can be considered the "best guess" to predict the market value without knowing anything about the specific property. The features that mainly drive the prediction

from the base value of 258.53 $/SqFt to the estimated 494.18 $/SqFt are the stabilized NOI, location, market cap rate and building age. In this example, the property's square footage reduces the prediction as it contributes negatively. 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 base 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 total, these four features contribute 290.07 $/SqFt to the base value of 258.53 $/SqFt, leading to a predicted value of 548.59 $/SqFt. However, this is not the predicted 494.18 $/SqFt as the negative contributions have been left aside so far. As highlighted in blue color, the size of the property has a negative impact and pushes against the other features, thus reducing the final prediction. The property size (38,500 square feet) is smaller than the sample average of 277,124 square feet resulting in a negative impact 30.63 $/SqFt. In sum, all other features add up to a negative 23.78 $/SqFt and lead to an expected 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 travelled. 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., 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*.

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. *Journal of 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. *The 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., Correa-Morales, J., & 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*, 59–96.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *The 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., & La Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonic 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 Kuhn, H. and Tucker, A. (Eds.), *Contributions to the theory of games, Vol. II*, 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 and Finance*, *38*(3), 213–225.

Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *The 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*, 349–388.

5 Conclusion

5.1 Executive Summary

This chapter summarizes the three individual contributions. First, the problems and objectives are discussed. Secondly, the data used and the methodology are outlined, and finally the results and implications for science and practice are presented.

Paper 1: Forecasting Office Rents with Ensemble Models – The Case for European Real Estate Markets

---

Problems and Objective

Commercial real estate markets, particularly office markets, have been extensively researched since the 1960s, and multiple frameworks for office rent forecasting have been proposed. That is, in multivariate and univariate fashions tackling, on the one hand, a structural and theoretical approach and, on the other hand, a statistical and atheoretical viewpoint. Structural frameworks face the challenge of scarce data in commercial real estate markets. Univariate models are more flexible but cannot cope with market heterogeneity. Therefore, an ensemble approach combining multiple univariate approaches could solve the forecasting problem in a multiple market context. The objective of this study is focused on univariate approaches to forecasting and, in this context, to understand, describe and apply a combination of a classical statistical method and a novel deep learning approach. The practicability and functionality of the proposed structure are demonstrated via the application to a dataset of 21 major European office markets.

Methodology and Data

The research paper aims at contributing to the existing body of literature by assessing the value of the application of machine learning and deep learning methods in the univariate estimations of office prime rents. In statistical modelling and particularly in time series forecasting, the most commonly applied approaches by researchers and practitioners are exponential smoothing, auto-regression and moving average processes. In this study, the focus is on integrated auto-regressive moving average models (ARIMA) as numerous applications in the field of real estate time series forecasting (McGough and Tsolacos, 1995; Tse, 1997; Stevenson and McGarth, 2003; Crawford and Fratantoni, 2003) have demonstrated its eligibility. ARIMA modelling assumes the forecast of the variable is estimated by the movements of its past values and errors. Thus, as described in the study

by McGough and Tsolacos (1995), the theoretical idea behind this approach is that past rental values contain information about future market behaviour. However, these methods solely rely on linear processes in the underlying data and cannot capture non-linear patterns that are likely to be found in real world data. Thus, ARIMA modelling shall be complemented by the suggested deep neural network (N-BEATS). That is, a non-linear method based on a deep neural network structure. Oreshkin et al. (2019) developed the neural basis expansion analysis for interpretable time series forecasting (N-BEATS) which is a deep learning method explicitly developed to produce univariate time series point forecasting. The method is based on a deep neural network architecture with forward and backward residual connections and stacks of fully connected layers. The N-BEATS neural network architecture differs from existing neural network architectures that focus on sequence forecasting (such as LSTM recurrent neural networks) as, according to Oreshkin et al. (2020), the forecasting problem is much more treated as a non-linear multivariate regression problem. Consequently, a hybrid methodology is proposed to exploit the unique strengths of both methods and is applied to the office rental dataset to produce forecasts for four quarters ahead. The dataset consists of quarterly office prime rental values for 21 major European office markets from 1985 to 2020 gathered from CBRE. This yields a maximum of 147 observations for the time series of the London Central market. The office market prime rents are reported in Euro per square meter per year and can be defined as an average rent of the top 3-5 percent of all lettings in the observed markets.

Results and their Contribution to Science and Practice

This paper comprises an overview of commercial real estate rent forecasting frameworks and proposes an update on classical statistical univariate time series modelling. Forecasting with modern machine learning and deep learning algorithms demonstrated superior results in many fields of application. The selected N-BEATS method proved to have state-of-the-art forecasting properties in numerous statistical forecasting competitions. In a hybrid fashion, the advantages of both the ARIMA model and the N-BEATS model are combined and significantly improve the forecasting performance in multiple out-of-sample forecasts. It is demonstrated that the combination of the classical statistical approach with a deep learning approach reduces the error rate in the observed time series point forecasts and significantly increases the explanatory power of the computed ensemble model. On average, over the 21 observed European office markets, the meta model outperforms both individual models. Hence, combining classical statistical forecasting methods and modern deep learning approaches yields more accurate and consistent forecasts. As a result, the study on forecasting European office market prime rents confirms heterogeneity of real

estate markets. It also demonstrates that combining the forecast of different models can reduce uncertainty of abrupt changes in rents and is a good way to simultaneously approach office rent forecasting in multiple markets. In practice this can be used to improve forecasting models that are essential for assessing the value of portfolios and for evaluating investment strategies more thoroughly.

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

Problems and Objective

Accurate and up-to-date valuations of properties in commercial real estate play a significant role for numerous market players including investors, lenders, tax authorities and policy makers. However, the manual appraisal of properties is a time-consuming and resource-intensive process and has been criticized for its inaccuracy and subjectivity (Dunse and Jones, 1998; Fisher et al., 1999; Kok et al., 2017). Cannon and Cole (2011) analyse the deviations of manual appraisals to actual sale prices of US commercial properties and find that these are largely systematic. In the meantime, hedonic price models have been extensively discussed in the literature, not only as an alternative to manual valuation approaches, but also to identify the determinants of property prices and subsequently disentangle the price formation processes from an econometric point of view, naturally excluding subjectivity, structural biases or time lags (Rosen, 1974; Mills, 1992; Dunse and Jones, 1998; Malpezzi, 2002; Sirmans et al., 2005). Yet, the timely and accurate estimation of property values remains a challenging task, as price formation processes in real estate markets are highly complex and are likely to be determined by non-linear and non-monotonic relationships that differ across markets. Furthermore, van Wezel et al. (2005) state that such traditional hedonic models suffer from the risk of misspecification of their functional form and are based on a set of unrealistic assumptions.

Hence, the aim of this research paper is to assess whether data-driven machine learning algorithms can effectively use structured information from deviations between manual appraisals and sale prices.

Methodology and Data

Highly flexible machine learning methods tackle the problem of misspecification in the functional form and unrealistic assumptions. The application in the residential real estate sector has shown great success in recent years in accurately predicting prices of houses and apartments (Mayer et al., 2019; Bogin and Shui, 2020; Pace and Hayunga, 2020; Pai and Wang, 2020). This gave rise to the development of highly flexible machine learning techniques for the estimation of commercial property values that learn the relationship between the response and the regressors autonomously without the need for any a-priori specifications of their functional form, merely assuming a sufficiently large and representative sample. However, academic research in the field of commercial property markets faces major challenges due to heterogeneous characteristics and scarce data.

Following Cannon and Cole (2011), the deviation of commercial real estate appraisals to sale prices in the NCREIF property database is calculated. Taking advantage of non-parametric machine learning methods such as Regression Tree Methods (Random Forest Regression (Breiman, 2001), Gradient Boosting (Friedman, 2001) and XGBoost (Chen and Guestrin, 2016)) the information content found in the subject residuals is investigated on structured variation. With the aim to explain where the deviations originate from model-agnostic interpretation methods, in particular, permutation feature importance, are employed. For this purpose, the U.S. NCREIF property database is complemented by data on financial, physical, locational and macroeconomic attributes.

Results and their Contribution to Science and Practice

The study's findings reveal that advanced machine learning techniques offer valuable insights beyond traditional valuation methods. By analyzing 50 different features, we discover that boosting tree models can significantly enhance the accuracy of commercial appraisal values while also reducing structural bias. This improvement is achieved by explaining the deviations between market values and transaction prices, providing a more nuanced understanding of the underlying factors influencing commercial property values. Through the use of permutation feature importance, we identify spatial and structural covariates as the most influential factors in contributing to appraisal errors, while time lags between appraisal and sale dates accounted for only a fraction of the variation. However, it's worth noting that in times of economic uncertainty, such as during a pandemic or war, machine learning models may struggle to adjust to infrequent events, limiting their effectiveness. Nonetheless, this study highlights the potential for supervised machine learning methods to enhance valuation practices in the US commercial real estate sector.

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

Problems and Objective

The advent of machine learning has brought new approaches to property valuation to the fore. Automated valuation models (AVMs) show promising results in terms of accuracy, but lack of inherent interpretability, which 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 post-hoc interpretability of the models' prediction rules. Based on this, the study 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.

Methodology and Data

The cleaned principal dataset used in this study comprises 400,370 quarterly property-level observations across four commercial property types (i.e., apartment, industrial, office, and retail) and 18,868 individual properties observed over a period of 30 years from Q1 1991 through Q1 2021, provided by the National Council of Real Estate Investment Fiduciaries (NCREIF). In addition, the dataset was enriched with 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. This constitutes a total of 32 features used to estimate the properties market values.

First, we calibrate and train a deep neural network (DNN) separately for each property type using mutually exclusive training, validation, and test splits to ensure generalizability of the models. Second, after evaluating the models' predictive performance, we apply an advanced model-agnostic methodology, Shapley Additive Explanations (SHAP), to mitigate the trade-off between accuracy and interpretability and provide ex-post comprehensibility of the algorithm's prediction rules. Third, non-linear relationships as well as three-dimensional interaction effects are analyzed. In addition, a linear multiple regression analysis is conducted to serve as a point of reference.

Results and their Contribution to Science and Practice

The proposed methodological framework demonstrates high accuracy in the estimation of market values across all four asset sectors. In line with traditional valuation methods, the SHAP analysis shows that market values across all sectors are mainly driven by the net operating income and the market capitalization rates. Moreover, the location expressed by the geo-coordinates and the distance to points of interest as well as the properties' physical condition proxied with building age have a strong influence on the models' predictions. Deviations across sectors are observed predominantly in sector specific characteristics. Furthermore, non-linear and three-dimensional relationships are revealed. In summary, the observed relationships between the four asset sectors follow an economic logic and confirm the results of previous studies.

Comprehensibility and interpretability are essential for the practical application of AVMs. Hence, the proposed methods have the potential to leverage efficiency in both markets and business processes in the long term by increasing the speed and scale of valuations, reducing transaction cost, and ultimately increasing transparency in pricing processes. By pointing to the caveats and illustrating the potential of the methods, we aim to take the application of AVMs in the commercial real estate sector one step further and hope to motivate further research in this field.

5.2 Final Remarks

In a world of constant change and digital evolution, data is key to our society and plays a prominent role in all industries. The real estate industry is undergoing a profound transformation driven by algorithms that foremost enhance the human ability to quickly learn complex relationships from large amounts of data. This dissertation is a step towards the unrestricted application of advanced machine learning approaches and the use of artificial intelligence for understanding commercial real estate markets. Pointing out the caveats and illustrating the potential, it is shown that the application of non-linear machine learning models yields improvements in direct comparison to traditionally applied methodological frameworks. However, it was found that the effectiveness of these approaches depends on a single parameter, namely the quality and quantity of the available data.

The three research papers have explored the use of cutting-edge machine learning models to improve forecasting accuracy for prime office rents, increase accuracy and reduce bias in appraisals and extract valuable insights from appraisal errors by opening up black box models. The practicability of the application of advanced machine learning methods in the real estate industry is demonstrated. While all approaches are purely data-driven it is exciting to see that the models follow economic principles consistent with real estate literature. This promotes the acceptance of such models in the valuation industry.

On the one hand, the work has promising implications for the future of the industry and research. On the other hand, limitations of the research have been recognized. In univariate forecasting limitations become apparent with structure breaks in the time series. Univariate models cannot account for external shocks, which limits their forecasting ability. Adding exogenous factors can make the forecast more accurate, but predicting those factors is challenging. The findings of the studies are based on limited datasets, with papers 2 and 3 focusing exclusively on prime institutional properties. It is important to note that the current research on commercial real estate markets is limited to a small set of property characteristics. Despite a wide range of developed markets being covered, structural property characteristics have received little attention in the past. This is primarily due to the time-consuming and laborious nature of collecting such information by hand.

Technological advancements have opened up new possibilities for information collection in this sector. Property characteristics that are related to the shell of a building can be collected through image evaluation. The combination of big data analytics and the aggregation of multiple data sources could provide valuable insights. Lease data with information on lease terms and rent payments, environmental data on factors such as air

quality, water and energy consumption, social media data with customer sentiment and trends and public records such as parking usage or pedestrian flows are all data sources that would complement future commercial real estate market analysis. The potential of automated frameworks that integrate multiple approaches and the above data sources such as sentiment analysis, location analysis and valuation models to gain a deeper understanding of commercial real estate markets is not yet fully explored. When data collection is extended three-dimensionally, encompassing time, markets and the building itself to provide full transparency and enable comprehensive analytics, machine learning applications are expected to have enormous potential to revolutionize the industry.

In conclusion this dissertation represents a significant step forward in advancing our understanding of commercial real estate markets. By demonstrating the potential of machine learning and artificial intelligence and providing a roadmap for future research, this work contributes to real estate literature and holds practical implications for industry professionals. While this work marks a milestone on a path yet to be travelled, it certainly serves as a catalyst for further research in this field.

5.3 References

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.

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

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. The 22nd ACM SIGKDD International Conference.

Crawford, G.W. and Fratantoni, M.C. (2003). Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices. *Real Estate Economics*, 31(2), 223-243.

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

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.

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.

Malpezzi, S. (2002). Hedonic pricing models: A selective and applied review. In O'Sullivan, T. and Gibb, K. (Eds.), Housing Economics and Public Policy, Wiley, Oxford, UK, 67–89.

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.

McGough, T. and Tsolacos, S. (1995). Forecasting commercial rental values using ARIMA models. *Journal of Property Valuation and Investment*, 13(5), 6-22.

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

Oreshkin, B.N., Carpov, D., Chapados, N. and Bengio, Y. (2019). N-beats: neural basis expansion analysis for interpretable time series forecasting, available at: https://arxiv.org/pdf/1905.10437

Oreshkin, B.N., Dudek, G., Pełka, P. and Turkina, E. (2020). N-Beats: neural network for mid-term electricity load forecasting. available at: https://arxiv.org/-pdf/2009.11961.

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.

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

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

Stevenson, S. and McGarth, O. (2003). A comparison of alternative rental forecasting models: empirical tests on the London office market, *Journal of Property Research*, 20(3), 235-260.

Tse, R.Y. (1997). An application of the ARIMA model to real-estate prices in Hong Kong. *Journal of Property Finance*, 8(2), 152-163.

van Wezel, M., Kagie, M. M., & Potharst, R. R. (2005). Boosting the accuracy of hedonic pricing models. Econometric Institute, Erasmus University Rotterdam. http://hdl.handle.net/1765/7145. Accessed 18 April 2022