INVESTMENT STRATEGY & RESEARCH

RESEARCH PAPER

Dr. Marcelo Cajias

24.03.2022

Optimising the value of institutional real estate portfolios with machine learning

A Hedonic Machine Learning Approach

PATRIZIA

Executive Summary

Machine learning understands residential portfolios

Artificial intelligence (AI) and especially machine learning (ML) methods increasingly offer valuable alternatives to answer questions in real estate research and practice.

We investigate whether ML methods are suitable of estimating residential rents by comparing a conventional hedonic model with four ML algorithms, namely Support Vector Regression (SVR), Random Forest Regression (RFR), Gradient Tree Boosting (GTB) and eXtreme Gradient Boosting (XGB). We find ML methods to model rental values more precisely than traditional linear regression.

We use these findings to estimate rental values for an institutionally managed portfolio and match them with their corresponding contract rents.

The results show that apartments tend to be underrented, with ML models indicating higher deviation of estimated and contract rents than linear Ordinary Least Squares (OLS) models.

Thus, our findings indicate that non-linear thinking reveals potential benefits when applying ML hedonic models in the area of residential markets and portfolios.

TABLE OF CONTENTS

INTRODUCTION

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

While these new methods are increasingly used outside of real estate, they have only been applied in limited use cases in the analysis of residential rents. This paper investigates whether hedonic machine learning (ML) methods are capable of providing new insights and applications in residential rental markets. The subject of rents and how market participants can use AI to assess and verify investment decisions has, however, not yet been investigated in depth. Consequently, literature on this topic is scarce even though new tools seem to have capabilities that may outperform conventional hedonic methods.

The aim of this study is to shed light on the application of algorithm-driven methods in rental markets. Most importantly, we assess the value that investment managers might obtain when managing a residential real estate portfolio based on ML methods as opposed to fundamental analysis via the Ordinary Least Squares. Consequently, we assess how accurate linear and algorithm-driven hedonic models predict rents based on a large data set and transfer these findings to an institutionally managed residential portfolio. Thus, we estimate rental values an investor could expect for the portfolio apartments in re-lettings scenarios. Further, we compare them to their corresponding contract rents to find out whether the different models would estimate a potential (or need) for rental adjustments.

HEDONIC MODELLING IN THE REAL ESTATE LITERATURE

The aim of hedonic modelling is to better understand the fundamental factors affecting property rents and prices. By expressing the rent of an apartment as the sum of its estimated individual characteristics, hedonic modelling can be used for inferential and predictive purposes. Traditionally, a hedonic model employs multiple linear regression to establish the relationship between the response and the corresponding hedonic characteristics (Rosen, 1974, A. C. Goodman, 1978). Depending on the spatial characteristics of the market under investigation and the data structure, a hedonic model needs to fulfil a minimum number of assumptions (see e.g. Sirmans et al., 2005 and Bourassa et al., 2007). However, several authors such as Lai et al. (2008), Bourassa et al. (2010) and Cajias (2018) have demonstrated the limited explanatory power of traditional hedonic models and shown that statistical developments such as the inclusion of spatial and non-linear effects lead to significant enhancements in model accuracy (see more: Fik et al., 2003; Lin et al., 2009; Banzhaf & Farooque, 2013).

Theoretical framework for connecting individual's utility function for real estate with the supply of local amenities

Source: PATRIZIA own depiction, Deriving the willingness to pay for real estate with hedonics.

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

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

Even though there is a growing body of literature on the topic, further investigation is needed due to various reasons:

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

DESCRIPTION OF DATA

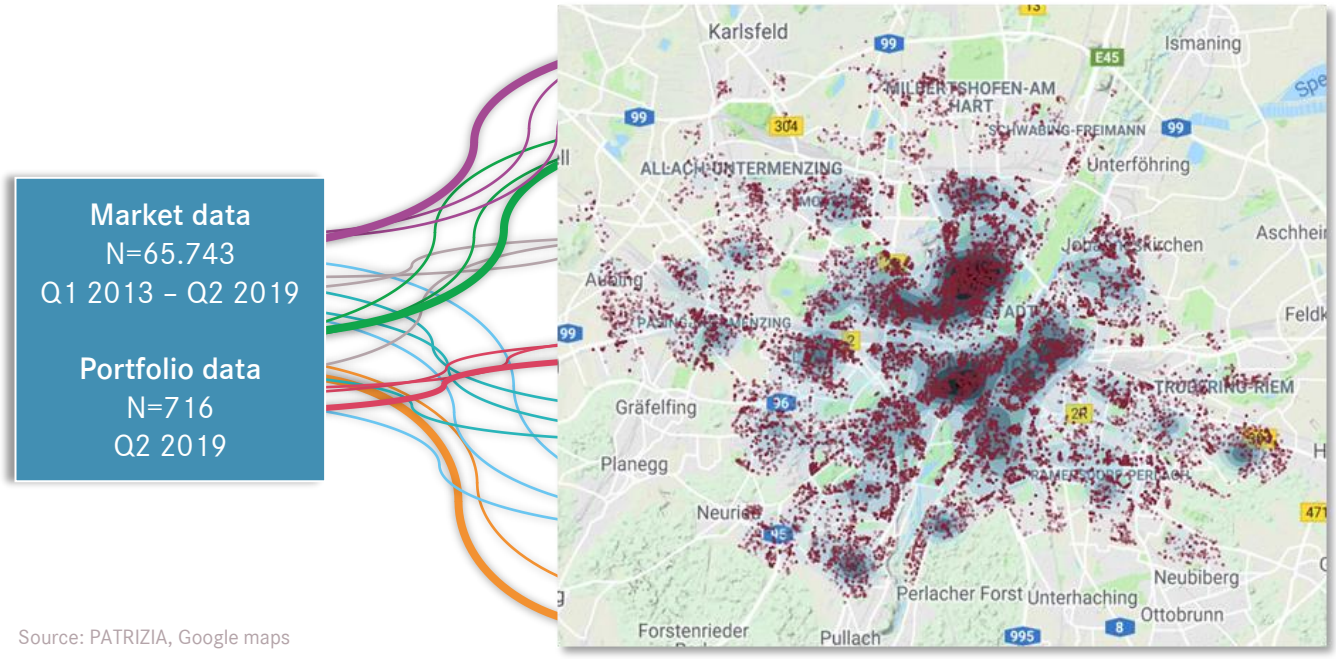
This study encompasses the residential real estate market in Munich, Germany. With approximately 1.5 million residents and an annual growth rate of about 0.75%, Munich is the third largest city in Germany. The city and its metropolitan areas have one of the most prospering economies in Germany, accommodating several globally active companies in sectors such as automotive, environmental techniques, information and communication, insurance, life sciences and medicine. Stable economic growth and good employment conditions have yielded a positive development of the residential market throughout the last decade.

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

Market Data from MLS

The use of asking data can be advantageous as it offers the possibility to capture and rapidly reveal market movements. Y. Chen et al. (2016) and Baldominos et al. (2018) argue that it is more appropriate for modelling timely dynamics of housing markets on a fine-scale level. Moreover, MLS asking data can overcome the challenges raised by the general lack of European housing contract data which is mentioned, for instance, by Rondinelli and Veronese (2011). It is actively used for empirical research by several authors such as Hanson and Hawley (2011), Rae (2014), Gröbel and Thomschke (2018), Pérez-Rave et al. (2019) and Gröbel (2019) for studies in Germany, the US and the UK. As Pérez-Rave et al. (2019) state, MLS data shows important characteristics of big data in terms of volume, variety and value. This enables researchers and market participants to overcome temporal delays and limited analyses on market developments that are associated with, for example, official statistics.

Description of data



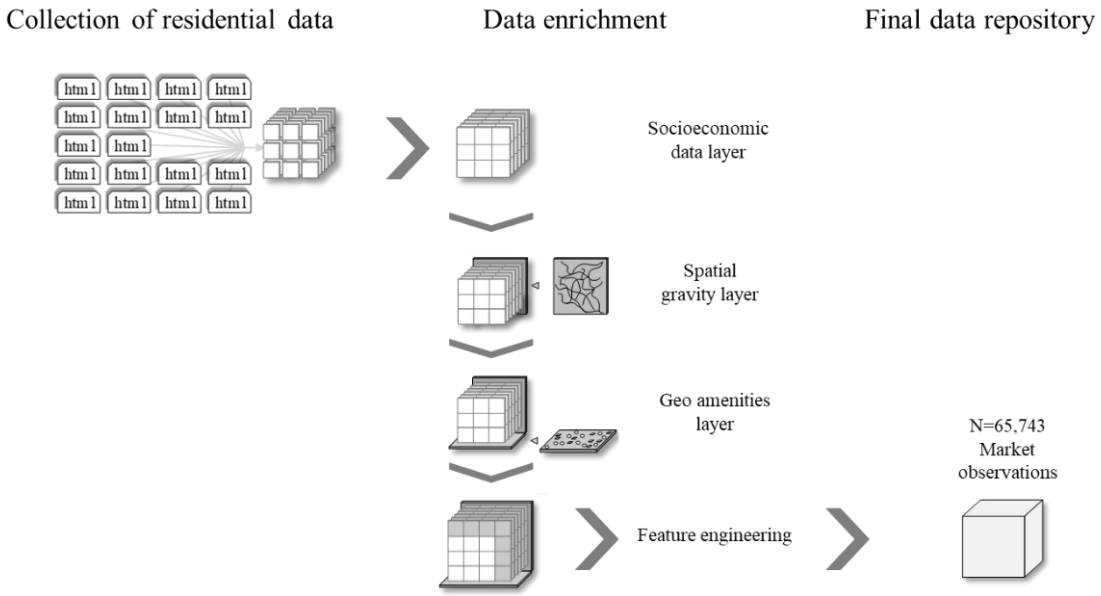
Source: PATRIZIA, Google maps

In this context, MLS are perceived as "one of the most significant feature of today's real estate industry" (Li & Yavas, 2015). Due to the characteristics of the Munich residential real estate market, we expect the asking rents to be a good approximation for market-conform rental values. Although asking data plays a significant role in housing markets (see e.g. Shimizu et al., 2016, Han & Strange, 2016), differences to transaction data can occur that need to be kept in mind.

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

We access Value Marktdaten, one of the largest providers of real estate data in the German residential market. It uses web-scraping techniques for collecting, preparing and integrating real estate listings from more than 120 different MLS with full hedonic characteristics. Furthermore, we include socio-economic data from Growth from Knowledge (GfK), Germany's largest market research institute. We also add a gravity layer using data from Eurostat and the German statistical office to implicitly enable the models to account for spatial information. Finally, we complement each georeferenced residential data point by an amenities layer measuring the Euclidean proximity to important amenities. This information is gathered from Open Street Map (OSM) and Google via an API in R (R Core Team, 2016).

Extraction-Load-Transform (ELT) process for estimating hedonic market models



Source: PATRIZIA own depiction

This results in a dataset comprising eight structural characteristics (living area, age and whether the apartment has a bathtub, built-in kitchen, parking lot, terrace, balcony and an elevator), two socio-economic (number of households and households purchasing power in ZIP code area), and seven distance variables (proximity to bus station, park, school, subway, supermarket, neighbourhood centre and city centre). Rent, living area, distances as well as both socio-economic characteristics are incorporated using their log-transformation to account for the distribution. Quarter and year dummies are used to control for time effects.

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

Descriptive statistics of the MLS data

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

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

Portfolio Data

In addition to the obtained data through MLS, we use data of a managed residential real estate portfolio. The portfolio consists of 716 apartments located in Munich, comprising contract rents and the same explanatory variables as presented in the previous section. An average apartment in the portfolio contains 71.99 sqm and yields a rental income of 1,009.37 EUR/p.m. The distance to the city centre of 6.77 km is about 2 km further than the distance of an average apartment, but the distance to the centre of the related ZIP code is with 0.50 km 200 m shorter. Moreover, the distances to all important infrastructure facilities is on average closer compared to the apartments in the previous dataset. Purchasing power and number of households are about the same. We again consider additional hedonic characteristics and time controls as dummy variables.

Descriptive statistics of the portfolio

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

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

METHODOLOGY

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

Hedonic Modelling with Traditional and Machine Learning Methods

The analysis encompasses one linear and four ML models. We follow Zurada et al. (2011) and Chin et al. (2020) by choosing OLS as the base case for the comparison of several algorithm-driven hedonic models. OLS is a widespread variant for hedonic modelling and consequently a well-known and easy interpretable benchmark for performance analysis. SVR, RFR, GTB and eXtreme Gradient Boosting (XGB) represent the modern approaches that will be applied in our analysis. Except for XGB, all methods have been used for real estate related questions in areas such as valuation. XGB is a method developed in the last few years that shows computational advantages especially in large data sets. In the following, we discuss the basic structure of each hedonic method under investigation:

Ordinary Least Squares Regression – OLS

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

Machine Learning Methods

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

Overview of machine learning algorithms

Traditional method

Ordinary least squares

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i$$

Single algorithm-driven methods

Support vector machine

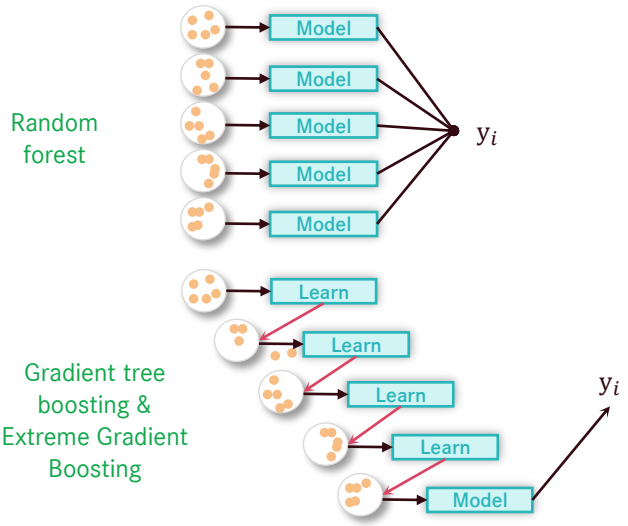
$$\min \left[\frac{1}{2} \|w\|^2 + c \sum_{i=1}^{l} (\xi_i + \xi_i^*) \right]$$

$$\text{subject to: } \begin{cases} y_i - \langle w, x \rangle - b \leq \varepsilon + \xi_i \\ -y_i + \langle w, x \rangle + b \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Regression trees

$$\min \left[\sum_{i:x_i \in t_1(i,s)} (y_i - \tilde{y}_{t_1})^2 + \sum_{i:x_i \in t_2(i,s)} (y_i - \tilde{y}_{t_2})^2 \right]$$

Multi-stage algorithm-driven methods



Source: PATRIZIA, Notes: This exhibit shows the machine learning methods. ε=estimation errors; ξ=error penalization; w=predicted value and c=weight of penalization. OLS = Ordinary least squares; SVR = Support vector regression; RFR = Random forest regression; GTB = Gradient tree boosting; XGB= Extreme gradient boosting.

METHODOLOGY: Error-based Comparison of Model Performance

Following Zurada et al. (2011), Schulz et al. (2014) and Mayer et al. (2019), we use mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and coefficient of determination R^2 to conclude on the accuracy of the applied methods. We furthermore investigate the precision regarding over- or underestimation by applying the mean percentage error (MPE). While similar research give little attention to the dispersion of the errors within the prediction, we discuss error buckets (PE10 and PE20), coefficient of dispersion (COD) and inter-quartile-range (IQR) to assess the magnitude of the estimation errors. By looking at the accuracy, precision and dispersion, we aim to derive further insights on the differences between the applied ML methods.

Error-based measurements on the predictive performance

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

ECONOMETRIC RESULTS: Predictive Performance of Models

We find all ML methods to be more accurate in modelling rents than traditional OLS regression. While OLS provides on average highest absolute rental estimation errors (MAE), we find all ML methods to considerably increase the model accuracy, with RFR being most accurate. The graphical analysis regarding median and quantiles underpin the findings. To illustrate these results, we convert the MAE to EUR/sqm, dividing it by the size of an average apartment of 76.49 sqm. The estimation error decreases from 2.34 EUR/sqm (OLS) to 1.52 EUR/sqm (RFR). Regarding the RMSE, which differs from the MAE by penalizing extreme deviations, the results show a similar picture. Compared to OLS, all ML methods are more robust to extreme deviations. These findings complement the results of Bogin and Shui (2020).

Error-based comparison of model forecasting at market level

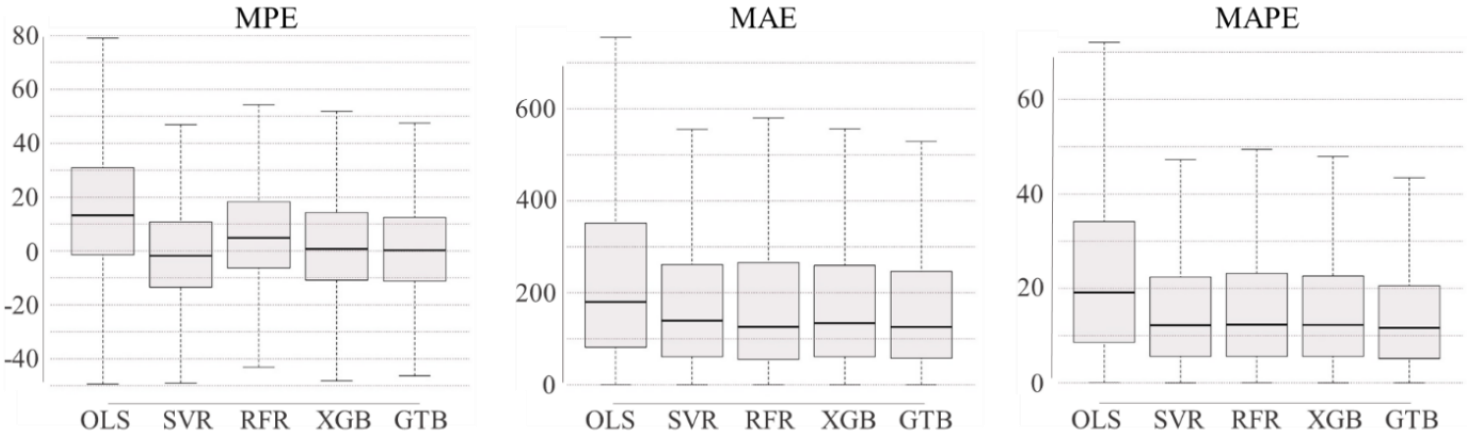
	Unit	OLS	SVR	GTB	XGB	RFR
MAE	EUR/p.m.	271.14	201.84	189.52	203.61	212.35
	EUR/sqm/p.m.	3.54	2.64	2.48	2.66	2.78
RMSE	EUR	418.86	303.76	292.20	320.88	366.93
MAPE	%	24.00	15.69	15.02	16.08	16.77
R^2	%	80.12%	84.39%	86.39%	84.60%	83.93%
ME	EUR	177.59	10.54	26.13	42.26	108.20
MPE	%	15.47	1.16	1.13	1.94	6.61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

Notes: This exhibit reports the error-based measurements on the predictive performance through MAE, RMSE, MAPE and R^2. ME and MPE indicate over- or underestimation. PE10, PE20, IQR and COD show the dispersion. All measures are out-of-sample (test set) and are based on the calculations presented in Exhibit 9. Absolute values are reported in euro per month. Relative values are reported in percent. Source: Own calculation.

While OLS shows an R^2 of 81.65%, GTB and RFR areable to explain approximately 90% of the deviation. Ho et al. (2020) find similar results for housing transactions. Wu et al. (2008) and Y. Chen et al. (2016) show SVR to be robust and also accurate in modelling property prices and rents. It is therefore not surprising that SVR works well in our setting (R^2 of 87.79%) and is similar to ensemble learners such as XGB and GTB. A look at the MAPE shows that traditional OLS misestimates the observed rents by 15.60% on average, while RFR improves model accuracy with an average misspecification of about 10%. These findings corroborate the results of Hu et al. (2019), who also show the tree-based bagging algorithm RFR to be most suitable for modelling property rents. Regarding transactions prices, Baldominos et al. (2018) likewise highlight ensembles of regression trees to perform best.

As Fik et al. (2003) state, Freddie Mac early suggested that at least 50% of the predicted sale prices of residential properties should be within 10% of the true value. In common real estate valuation practice, the estimated value of a property is allowed to vary 10% to 20% from its market value. Transferring this to rents, all our models yield satisfactory results.

Graphical error-based comparison of model forecasting at market level



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

The median percentage deviation of all our ML methods is below 10%, as displayed in the boxplots. Therefore, we conclude ML algorithms to be capable of precisely modelling rents.

ECONOMETRIC RESULTS: Predictive Performance of Models

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

The positive MPEs indicate that all methods underestimate the observed rental value on average.

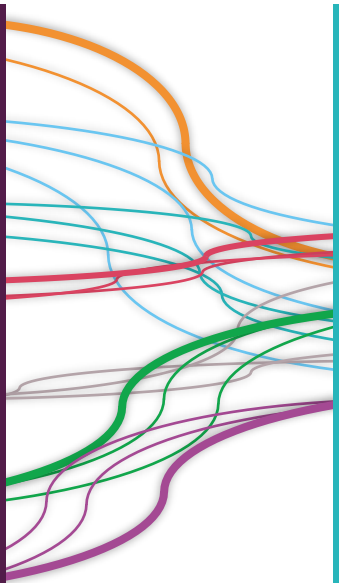
In addition, **the dispersion of the estimation** adds another possibility for investigation. The boxplots of MPE show a symmetric distribution of all methods, indicating no general bias for traditional as well as ML variants. PE10 calculates the percentage of observations with a deviation of less than 10%. This metric can also be referred to as '**hit rate**'. While OLS can estimate 40.65% of all observations within this range, algorithm-driven RFR models estimate 62.62% correctly. Within a deviation of +/-20%, we find all ML methods exceed 84%. The IQR draws a similar picture. While OLS estimates 50% of all observed values within a range of 1.68 EUR/sqm above or below the median, the ML models significantly decrease the range of deviations (+/-1.00 EUR/sqm). The COD also confirms these results.

ML methods are not only more accurate on average, but the error dispersion is also lower leading to a better predictive performance.

To verify the robustness of our results especially in terms of general applicability, we run all methods on an additional sample of rents from July 2019 to September 2019. The model specifications are the same as in the previous analysis. They consequently provide error-based measurements for a one-period-ahead out-of-sample forecast. Our findings are equivalent to the findings in the original dataset. An upward shift in all error-based measurements can be traced back to thriving residential real estate markets in German metropolitan areas – especially in Munich. Bogin and Shui (2020) find RFR to be prone to overfitting. We can corroborate their results. While RFR performs best when it comes to the original dataset, we now find all other ML methods to be more accurate in forecasting future rents. Regarding RMSE as well as PE10 and PE20, the results indicate that RFR seems to show some misspecification for high deviations. We suggest RFR fits extreme values generally well but fails to explain them within new sample of future rents as it shows the highest RMSE besides OLS, but good results for PE10 and PE20.

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

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



Altogether, a reasonable explanation for the better performance of ML methods can be given by the fact that they are able to capture non-linear and non-normal relationships (Pace & Hayunga, 2020; Bogin & Shui, 2020).

Because non-linearity is an important characteristic of real estate markets, the application of ML techniques provides more accurate estimates of residential rents.

ECONOMETRIC RESULTS: Rental prediction at portfolio level

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

The OLS displays the lowest absolute error. All ML methods show a considerably higher deviation within their estimation. While OLS only allows for an average estimation error of 2.20 EUR/sqm, tree-based methods RFR, GTB and XGB result in an average deviation of 2.34 to 2.75 EUR/sqm. RMSE and MAPE underpin these findings.

Interestingly, the performance of the ML methods is contrary to the previous findings. Hence, a look at the models' 'hit rate' reveals:

While tree-based methods can estimate about 63% of all observed rents within a deviation of +/-20%, OLS is able to model 68% accurately. For the portfolio data, we can consequently conclude that linear OLS leads to more accurate estimates.

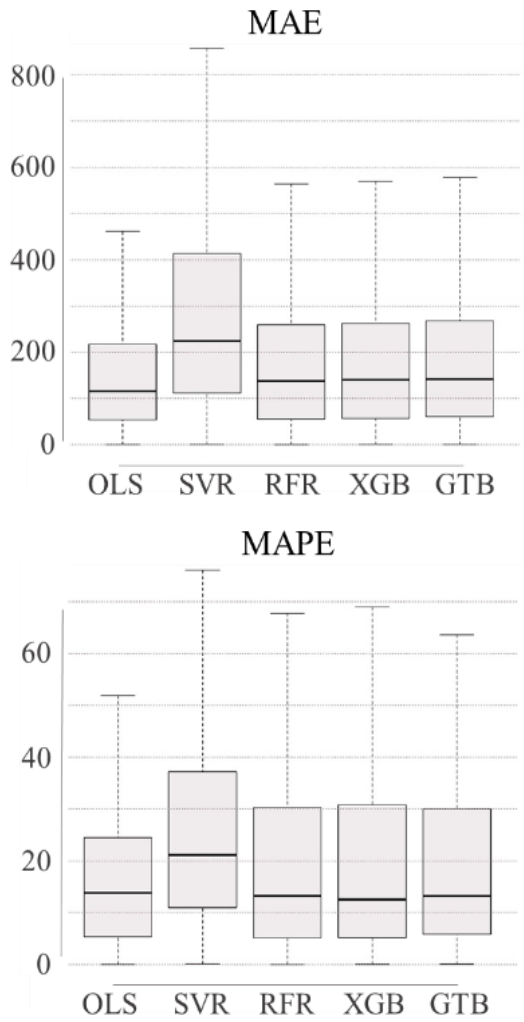
Furthermore, it is noticeable that SVR shows the highest deviation of portfolio rents from estimated rents, with an MAE of 3.73 EUR/sqm, which requires a deeper discussion. SVR is very sensitive to the choice of support vectors and tends to neglect the informational content of observations within the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

Error-based comparison of model performance at portfolio level

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

Notes: This exhibit reports the model accuracy through MAE, RMSE and MAPE. PE20 shows the dispersion. All measures are based on the calculations presented in Exhibit 9 in the Appendix. Absolute values are reported in euro per month. Relative values are reported in percent. Source: Own calculation.

Graphical error-based comparison of model performance at portfolio level



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

ECONOMETRIC RESULTS: Rental prediction at portfolio level

Average potential for rental increases

Method	As % of contract rents (MPE)			As rent in EUR/sqm (ME/sqm)		
	All	Q5 & Q95	Q10 & Q90	All	Q5 & Q95	Q10 & Q90
OLS	-4.95% *	-4.85%*	-4.75%*	-0.87*	-1.02*	-1.09*
GTB	-14.81% ***	-14.13%***	-13.64%***	-2.29***	-2.32***	-2.34***
XGB	-14.56%***	-13.99%***	-13.59%***	-2.21***	-2.30***	-2.36***
RFR	-12.54%***	-12.10%***	-11.91%***	-1.67***	-1.82***	-1.92***

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

To assess to which extent this rental potential exists and consequently whether portfolio apartments are under- or overrented, we calculate the relative difference of estimated rents to contract rents. All models indicate that contract rents are below estimated rents. While OLS indicates portfolio apartments to be underrented by 4.95% (0.87 EUR/sqm) on average, algorithm-driven hedonic models signal contract rents to be 12.54% (1.67 EUR/sqm) (RFR) to 14.81% (2.29 EUR/sqm) (GTB) below estimated rents. Our results are robust even if we exclude the highest and lowest 5%-quantile and 10%-quantile, respectively. The fact that all models show underrented situations is intuitive, especially in metropolitan areas in Germany, since rental growth in the residential real estate market exceeds inflation and hence, contract rents lag behind.

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

A low error measurement and low average deviation indicates that estimated rents are to a large extent in line with observed contract rents. Because estimated rents represent a rental value a landlord could expect in re-lettings, OLS with the lowest error measures would indicate a low potential for rental adjustments.

In contrast, ML models show considerably higher deviations. Because these models have confirmed a higher predictive performance, we assume that estimates from ML models more accurately reflect the potential rental value in re-letting.

An investment manager that "thinks OLS" would underestimate possible rental changes in upcoming re-letting negotiations. The underestimation results from the high concordance between OLS and contract rents.

The results indicate that ML methods can identify higher rental potentials that can be used by investment managers to manage portfolios more accurately.

An investment manager using OLS underestimates the rental-lift potential in his portfolio. By 'thinking linear', he assumes that contract rents are in line with estimated rents to a high extent. In contrast, our study reveals that ML methods show the potential for rental increases to be 2x to 3x times higher.

Key takeaways

In this study we investigate the predictive performance of traditional and algorithm-driven hedonic models and the added value an application of those methods can provide for an investment manager.

Because non-linearity is an important characteristic of real estate markets, the application of ML techniques provides more accurate estimates of residential rents.

Both traditional linear and ML methods perform well in explaining residential rents. However, algorithm-driven models are more accurate.

Linear models misestimates observed market rents by 15.6% (2.34 EUR/sqm), whilst tree-based machine learning (ML) models show the highest accuracy by reducing the absolute estimation error to 10.16% (1.52 EUR/sqm).

Based on an institutional residential portfolio, linear models indicate that contract rents are only 4.95% below estimated rents, whilst ML models identify potential for rental increases that is two to three times higher.

An investment manager that “thinks linearly” would underestimate possible rental changes in upcoming re-letting negotiations.

REFERENCES

- **Antipov, E. A., & Pokryshevskaya, E. B. (2012).** Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772–1778.
- **Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018).** Identifying Real Estate Opportunities Using Machine Learning. Applied Sciences, 8(11), 2321.
- **Banzhaf, H. S., & Farooque, O. (2013).** Interjurisdictional housing prices and spatial amenities: Which measures of housing prices reflect local public goods? Regional Science and Urban Economics, 43(4), 635–648.
- **Bogin, A. N., & Shui, J. (2020).** Appraisal Accuracy and Automated Valuation Models in Rural Areas. The Journal of Real Estate Finance and Economics, 60(1-2), 40–52.
- **Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007).** Spatial dependence, housing submarkets, and house price prediction. The Journal of Real Estate Finance and Economics, 35(2), 143–160.
- **Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010).** Predicting house prices with spatial dependence: a comparison of alternative methods. Journal of Real Estate Research, 32(2), 139–159.
- **Breiman, L. (2001).** Random forests. Machine Learning, 45(1), 5–32.
- **Breiman, L., Friedman, J., Stone, C. J., & & Olshen, R. A. (1984).** Classification and regression trees. CRC press.
- **Cajias, M. (2018).** Is there room for another hedonic model? The advantages of the GAMLSS approach in real estate research. Journal of European Real Estate Research, 11(2), 224–245.
- **Cajias, M., & Ertl, S. (2018).** Spatial effects and non-linearity in hedonic modeling. Journal of Property Investment & Finance, 36(1), 32–49.
- **Cajias, M., & Freudenreich, P. (2018).** Exploring the determinants of liquidity with big data – market heterogeneity in German markets. Journal of Property Investment & Finance, 36(1), 3–18.
- **Chen, T., & Guestrin, C. (2016).** Xgboost: A scalable tree boosting system. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 785–794.
- **Chen, Y., Liu, X., Li, X., Liu, Y., & Xu, X. (2016).** Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning. Applied Geography, 75, 200–212.
- **Chin, S., Kahn, M. E., & Moon, H. R. (2020).** Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach. Real Estate Economics, 48(3), 886–914.
- **Cortes, C., & Vapnik, V. (1995).** Support-vector networks. Machine Learning, 20(3), 273–297.
- **Fik, T. J., Ling, D. C., & Mulligan, G. F. (2003).** Modeling spatial variation in housing prices: a variable interaction approach. Real Estate Economics, 31(4), 623–646.
- **Friedman, J. H. (2001).** Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.
- **Gallin, J. (2008).** The long-run relationship between house prices and rents. Real Estate Economics, 36(4), 635–658.
- **Genesove, D. (2003).** The nominal rigidity of apartment rents. Review of Economics and Statistics, 85(4), 844–853.
- **Goodman, A. C. (1978).** Hedonic prices, price indices and housing markets. Journal of Urban Economics, 5(4), 471–484.
- **Goodman, J. (2004).** Determinants of operating costs of multifamily rental housing. Journal of Housing Economics, 13(3), 226–244.
- **Gröbel, S. (2019).** Analysis of spatial variance clustering in the hedonic modeling of housing prices. Journal of Property Research, 36(1), 1–26.
- **Gröbel, S., & Thomschke, L. (2018).** Hedonic pricing and the spatial structure of housing data– an application to Berlin. Journal of Property Research, 35(3), 185–208.
- **Hamilton, T. L., & Johnson, E. B. (2018).** Using Machine Learning and Google Street View to Estimate Visual Amenity Values. Working Paper, University of Richmond, University of Alabama.
- **Han, L., & Strange, W. C. (2016).** What is the role of the asking price for a house? Journal of Urban Economics, 93, 115–130.
- **Hanson, A., & Hawley, Z. (2011).** Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. Journal of Urban Economics, 70(2-3), 99–114.
- **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** The elements of statistical learning: data mining, inference, and prediction (2nd ed.). Berlin: Springer.
- **Ho, W. K., Tang, B.-S., & Wong, S. W. (2020).** Predicting property prices with machine learning algorithms. Journal of Property Research.
- **Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019).** Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. Land Use Policy, 82, 657–673.
- **IMF. (2018).** Global Financial Stability Report. International Monetary Fund.
- **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013).** An introduction to statistical learning (6th ed.). New York: Springer.
- **James, V., Wu, S., Gelfand, A., & Sirmans, C. (2005).** Apartment rent prediction using spatial modeling. Journal of Real Estate Research, 27(1), 105–136.
- **Jud, G. D., Seaks, T. G., & Winkler, D. T. (1996).** Time on the market: the impact of residential brokerage. Journal of Real Estate Research, 12(2), 447–458.
- **Jud, G. D., & Winkler, D. T. (1994).** What do real estate brokers do: an examination of excess returns in the housing market. Journal of Housing Economics, 3(4), 283–295.
- **Kee, K., & Walt, N. (1996).** Assessing the rental value of residential properties: an abductive learning networks approach. Journal of Real Estate Research, 12(1), 63–77.
- **Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017).** Big Data in Real Estate? From Manual Appraisal to Automated Valuation. The Journal of Portfolio Management, 43(6), 202–211.
- **Kontrimas, V., & Verikas, A. (2011).** The mass appraisal of the real estate by computational intelligence. Applied Soft Computing, 11(1), 443–448.
- **Lai, T.-Y., Vandell, K., Wang, K., & Welke, G. (2008).** Estimating Property Values by Replication: An Alternative to the Traditional Grid and Regression Methods. Journal of Real Estate Research, 30(4), 441–460.
- **Lam, K. C., Yu, C. Y., & Lam, C. K. (2009).** Support vector machine and entropy based decision support system for property valuation. Journal of Property Research, 26(3), 213–233.
- **Li, L., & Yavas, A. (2015).** The impact of a multiple listing service. Real Estate Economics, 43(2), 471–506.
- **Lin, Z., Rosenblatt, E., & Yao, V. W. (2009).** Spillover effects of foreclosures on neighborhood property values. The Journal of Real Estate Finance and Economics, 38(4), 387–407.
- **Lindenthal, T. (2020).** Beauty in the Eye of the Home-Owner: Aesthetic Zoning and Residential Property Values. Real Estate Economics, 48(2), 530–555.
- **Lindenthal, T., & Johnson, E. B. (2020).** Machine Learning, Architectural Styles and Property Values. Working Paper, University of Cambridge, University of Alabama.
- **Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019).** Estimation and updating methods for hedonic valuation. Journal of European Real Estate Research, 12(1), 134–150.
- **Pace, R. K., & Hayunga, D. (2020).** Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests. The Journal of Real Estate Finance and Economics, 60(1-2), 170–180.
- **Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019).** A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. Journal of Property Research, 36(1), 59–96.
- **R Core Team. (2016).** R: A language and environment for statistical computing.
- **Rae, A. (2014).** Online Housing Search and the Geography of Submarkets. Housing Studies, 30(3), 453–472.
- **Rondinelli, C., & Veronese, G. (2011).** Housing rent dynamics in Italy. Economic Modelling, 28(1-2), 540–548.
- **Rosen, S. (1974).** Hedonic prices and implicit markets: product differentiation in pure competition. Journal of Political Economy, 82(1), 34–55.
- **Schulz, R., Wersing, M., & Werwatz, A. (2014).** Automated valuation modelling: a specification exercise. Journal of Property Research, 31(2), 131–153.
- **Shimizu, C., Nishimura, K. G., & Watanabe, T. (2016).** House prices at different stages of the buying/selling process. Regional Science and Urban Economics, 59, 37–53.
- **Sirmans, S., Macpherson, D., & Zietz, E. (2005).** The composition of hedonic pricing models. Journal of Real Estate Literature, 13(1), 1–44.
- **Sirmans, S., Sirmans, C., & Benjamin, J. (1989).** Determining apartment rent: the value of amenities, services and external factors. Journal of Real Estate Research, 4(2), 33–43.
- **Smola, A. J., & Schölkopf, B. (2004).** A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222.
- **Thomschke, L. (2015).** Changes in the distribution of rental prices in Berlin. Regional Science and Urban Economics, 51, 88–100.
- **Turnbull, G. K., & Dombrow, J. (2006).** Spatial competition and shopping externalities: Evidence from the housing market. The Journal of Real Estate Finance and Economics, 32(4), 391–408.
- **ULI. (2020).** Promoting Housing Affordability. Urban Land Institute.
- **van Wezel, M., Kagie, M. M., & Potharst, R. R. (2005).** Boosting the accuracy of hedonic pricing models.
- **Verbrugge, R., Dorfman, A., Johnson, W., Marsh III, F., Poole, R., & Shoemaker, O. (2017).** Determinants of Differential Rent Changes: Mean Reversion versus the Usual Suspects. Real Estate Economics, 45(3), 591–627.
- **Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Philip, S. Y. (2008).** Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1–37.
- **Yao, Y., Zhang, J., Hong, Y., Liang, H., & He, J. (2018).** Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. Transactions in GIS, 22(2), 561–581.
- **Yavas, A., & Yang, S. (1995).** The strategic role of listing price in marketing real estate: theory and evidence. Real Estate Economics, 23(3), 347–368.
- **Yoo, S., Im, J., & Wagner, J. E. (2012).** Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. Landscape and Urban Planning, 107(3), 293–306.
- **Zhang, L., & Yi, Y. (2017).** Quantile house price indices in Beijing. Regional Science and Urban Economics, 63, 85–96.
- **Zurada, J., Levitan, A., & Guan, J. (2011).** A comparison of regression and artificial intelligence methods in a mass appraisal context. Journal of Real Estate Research, 33(3), 349–387.

Dr. Marcelo Cajias

Head of Data Intelligence

Marcelo heads the Data Intelligence section at PATRIZIA. In his role he is responsible for the global portfolio of analytical solutions and dashboards that support strategic investment decisions by means of observed and unobserved machine learning forecast models for various asset classes. Marcelo studied business administration at the University of Regensburg in Germany, majoring in statistics, econometrics and real estate economics. He received his doctorate for his thesis on the economic impact of sustainability on listed real estate companies. As a dedicated researcher, his work has been published in various international journals and he was awarded the RICS Best Paper Award and the German Real Estate Research Prize.

A similar version of this article can be found in: Real Estate Finance, Summer 2021, Volume 38, Number 1, pages 3–19.

PATRIZIA