sciendo
www.degruyter.com/view/j/remav

# ALGORITHM-DRIVEN HEDONIC REAL ESTATE PRICING – AN EXPLAINABLE AI APPROACH

**Stanislav Tarasov[1], Bartłomiej Dessoulavy-Śliwiński[2*]**

[1] *Faculty of Economic Sciences, University of Warsaw, Długa 44/50, 00-241 Warsaw, Poland, e-mail: swtarasow@gmail.com, ORCID: 0000-0001-7018-9847*

[2] *Department of Management and Information Technology, Faculty of Economic Sciences, University of Warsaw, Długa 44/50, 00-241 Warsaw, Poland, e-mail: bsliwinski@wne.uw.edu.pl, ORCID: 0000-0002-4395-9815*

* *Corresponding author*

ARTICLE INFO

*Keywords:*

residential real estate, machine learning, explainable artificial intelligence, mass appraisal, automated valuation models

*JEL Classification:*

C14, C18, R39

ABSTRACT

Data-driven machine learning algorithms triggered a fundamental change in hedonic real estate pricing. However, their adaptive nonparametric structure makes inference and out-of-sample prediction challenging. This study introduces an explainable approach to interpreting machine learning predictions, which has not been done before in the local market context. Specifically, Random Forest and Extreme Gradient Boosting models are developed for residential real estate price prediction in Warsaw in 2021 on 10,827 property transactions. Model-agnostic Explainable Artificial Intelligence (XAI) methods are then used to investigate the black box decision making. The results show the practicability of applying XAI frameworks in the real estate market context to decode the rationale behind data-driven algorithms. Information about the relationships between input variables is extracted in greater detail. Accurate, reliable and transparent real estate valuation support tools can offer substantial advantages to participants in the real estate market, including banks, insurers, pension and sovereign wealth funds, as well public authorities and private individuals.

*Citation:*

Tarasov, S. & Dessoulavy-Śliwiński, B. (2025). Algorithm-driven hedonic real estate pricing - An explainable AI approach. *Real Estate Management and Valuation*, *33*(1), 22-34. https://doi.org/10.2478/remav-2025-0003

## 1. Introduction

Real estate is the world's biggest asset class, and residential real estate is a major asset for many individuals. Hence, property price prediction holds significant importance and increasingly attracts attention of researchers due to a variety of factors, including investment decisions, buy and sell transactions, tax considerations, or real estate investment trusts' activity (Arribas et al., 2016). Due to market and regulatory changes in recent years, real estate is appraised more frequently than ever before, and inaccurate valuation can lead to various complications affecting property owners, buyers, and financing providers. Hence, there is a strong demand for accurate, speedy, and dependable support tools for real estate valuation. As a result, automated valuation model (AVM) development has been steadily on the rise in the industry and in academic research.

On the one hand, the subject of predicting real estate prices is well recognized in the literature. Studies on this topic in the field of applied econometrics often employ conventional linear regression methods along with their spatial extensions, as well as ML-based models. Research has indicated varying results when comparing the performance of ML and spatial models, with some studies demonstrating superior results for the former (e.g., Rico-Juan & Taltavull de La Paz, 2021), and others presenting contrasting findings, in favor of the latter (e.g., Przekop, 2022). Li (2022), in their experimental study of extracting spatial effects from an ML model using a local interpretation method, shows that "locally interpreted machine learning models are good alternatives to spatial statistical models and perform better when complex spatial and non-spatial effects (e.g. non-linearities, interactions) coexist and are unknown". On the other hand, following the reasoning of Angrist and Pischke (2010), "empirical evidence on any given causal effect


eISSN: 2300-5289 

is always local, derived from a particular time, place, and research design." According to their line of thought, extending causal effects to new settings always entails speculation, and to address the distinctiveness of a particular research design, a constructive response is to develop additional evidence, gradually forming a better general understanding.

This study uses residential real estate transactional data with locational characteristics to develop an explainable machine learning solution for property price prediction in Warsaw, Poland. We consider a hedonic pricing approach estimated by decision-tree-based models – Random Forest and Extreme Gradient Boosting – which are well-suited for our research problem for a number of methodological advantages. The main research goal is to provide a practical, explainable perspective on how they arrive at price estimations, which can be particularly challenging for such models given their black-box nature. Upon achieving the optimal model specification, we apply an Explainable Artificial Intelligence (XAI) framework to get insights into the algorithmic decision-making process of black boxes, which, to the best of the authors' knowledge, has not been done before within the local market context. At the same time, we extract additional information about the associative relationships between variables in greater detail. This study attempts to find answers to the following questions:

1. What is the significance, strength and direction of the relationships between property characteristics, building characteristics, and property price per $m^2$ in advanced real estate valuation models?
2. How do advanced real estate valuation models account for the nonlinear relationship between geographic coordinates and property price per $m^2$, and what implications does this have for the significance of this relationship?

These questions, while to some extent addressed in prior research, are intended to validate the suitability and showcase the practical aspect of the methods that the authors have employed to achieve the initially posed research goal. Alternative machine learning models and spatial modeling remain out of scope of this study, as the focus of our research design is not on providing internal comparisons. New or improved data often encourages a fresh approach to long-standing research questions, especially with the reservoir of micro data and administrative records growing rapidly in many countries (Angrist & Pischke, 2010). Therefore, the study aims to allow the questions to drive the methods, rather than the methods to drive the questions.

## 2. Literature review

Hedonic pricing is well recognized in the real estate sector, with prior studies having established it as a standard method since the 1970s. It is theoretically rooted in the works of Lancaster (1966) and Rosen (1974). Lancaster's characteristics demand theory describes consumers as deriving their utility not directly from goods, but rather from the distinctive characteristics these goods possess. Consequently, the consumption of a good can be understood as the consumption of its composite attributes. Rosen (1974) further extends this theory to hedonic pricing. It treats a marketed good, e.g., a house, as a sum of individual goods (house characteristics or attributes) that cannot be sold separately on the market (Montero & Fernández-Avilés, 2014). However, this approach does not directly provide a universal functional form and list of explanatory variables to be considered for price modeling.

### 2.1. Addressing the issue of functional form

The most common econometric approach is utilizing multiple linear regression or extensions thereof to derive implicit prices. Econometric hedonic regression models have demonstrated their usefulness in that regard, being efficient in generating predictions and easy to interpret. However, several limitations of the underlying methods have been pointed out by researchers, such as imposed linearity and fixed parameters, which cannot be assumed to hold in reality (Osland, 2010). As a result, current literature on the use of AVMs in real estate places a strong focus on hedonic pricing with machine learning. These models follow a data-driven approach to finding the best model fit. Over the past few years, an extensive body of literature has demonstrated the potential of ML algorithms to accurately estimate prices of real estate properties in the residential sector. Given that the ML approach has entered the already well-established field of property valuation, some researchers direct their efforts toward comparing various modeling techniques (e.g., Sevgen & Tanrivermiş, 2024). Numerous comparative studies, which document the accuracy of a broader range of model alternatives, demonstrate that tree-based


eISSN: 2300-5289 | Received 2024-05-22 | Revised 2024-09-18 | Accepted 2024-11-21

models, particularly boosting and bagging algorithms, show superiority over other methods (e.g., Antipov & Pokryshevskaya, 2012; Bogin & Shui, 2020; Hu et al., 2019; Mayer et al., 2019; Mora-Garcia et al., 2022).

The main objective of a hedonic pricing model is to estimate the contribution of each characteristic or attribute to the property price (Montero & Fernández-Avilés, 2014). However, as the internal logic behind ML models' predictions remains vague, they often lack transparency and interpretability, with models based on multiple decision trees (e.g., Random Forest, XGBoost) or on artificial neural networks (ANN) being called black boxes (Valier, 2020). One of the possibilities to understand how predictions are achieved is to apply Explainable AI. XAI is a set of tools and frameworks enabling post-hoc decomposition and interpretation of local and global ML predictions. The objective is to enable human experts to understand the underlying logic behind AI-driven decisions, which holds particular significance in paving the way for ethical and responsible AI, as well as enhancing the transparency and verifiability of ML in decision support (Holzinger et al., 2022). Therefore, XAI applications are steadily on the rise within economic research, including real estate AVMs in international markets (e.g., Lorenz et al., 2023; Rico-Juan & Taltavull de La Paz, 2021). However, the application of XAI frameworks remains unexplored in the Polish real estate market, and this study seeks to address this gap.

### 2.2. Addressing the issue of explanatory variables

As evidenced in existing research, a large portion of property value comes from location. Alonso (1964), in his housing location theory, claims that agents compete for locations in the city and seek to maximize, within available budget constraints, utility derived from quick access to the city center. As later summarized by Straszhem (1987), residential location decisions are correlated with property prices and are a function of urban spatial structure, including city size, population density, business and housing locations. The new generations of theories have further enriched the classical theory fundamentals formulated by Alonso (1964). Behavioral housing models suggest that if factors other than accessibility (such as income, taxation, social composition of the neighborhood) impact the utility from housing location, the Alonso model may not work (Wheaton, 1977). Co-evolutionary models explore how the Alonso model interacts with neighborhood quality; for instance, Kauko (2006) considers both accessibility and "pleasantness" of housing location. From the lifecycle perspective, factors such as household activity demand and accessibility to amenities such as employment, education, shopping, or medical services may influence housing prices in urban areas (e.g., Niu & Liu, 2017). Recent studies have also affirmed the positive value of cultural amenities for housing markets when studied together with the effects of green spaces, public transportation, and proximity to universities (Borgoni et al., 2018). Therefore, our "hard" property and building characteristics are supplemented with additional information of spatial character. We include "soft" characteristics of the neighborhood (segmented by districts). We also consider "hard" geographic coordinates with the intention of replicating relation to the city points of interest and housing density with data-driven machine learning from crude coordinate data (e.g., Przekop, 2022).

## 3. Materials and methods

### 3.1. Information Sources and Dataset

Warsaw is the capital city of Poland, divided into 18 districts, with an area of 517.24 km$^2$ and population of 1,863,056 people (as of 2021), which makes it Poland's biggest and most significant real estate market. The estimated value of the residential real estate market in Poland at the end of 2021 was 5.6 trillion PLN, with Warsaw alone accounting for approximately 13.15% share (NBP, 2022). In this study we analyze the real estate transactions conducted in Warsaw in 2021. We gather and examine transactions for a single year to take the cyclical character of the real estate market out of the equation. Consequently, we do not consider time-varying macroeconomic indicators both for our modeling process as model factors (as considered by, e.g., Hong et al., 2020; Kok et al., 2017) or to control for structural differences in property prices across time (as considered by Deppner et al., 2023).

The core of the dataset consists of transactional data obtained from the Register of Property Prices (Rejestr Cen Nieruchomości, RCN) of Warsaw City Office. It contains transaction details on different types of properties and their attributes, as well as transaction description and price details. In this study we focus on deals originating from the secondary housing market with residential properties being the underlying assets, and the subject of the transactions being the transfer of property ownership and its related rights. Previous research has shown that the RCN data is frequently insufficient to justify its use as the only source of information for property value modeling (Zyga, 2019). Therefore, we incorporate additional information of spatial nature in our models. One way of including spatial information is feature engineering (Przekop, 2022). The use of economic and demographic data at the census tract level (or other statistical subdivision of a county or equivalent entity) is standard practice in real estate modeling (Kok et al., 2017). As part of the feature engineering process, we reinforce our core dataset with locational attributes, matching them with each property deal. These features are calculated based on the publicly available information for the same year as the transactional data from Statistical Office in Warsaw (2022) and Public Transport Authority in Warsaw (2022a, 2022b). This data provides a wide range of categorical attributes and reinforces the dataset by differentiating each property with more specific location qualities. However, it should be noted that while district-level location attributes could potentially capture global price differences across space, they might not be fully adequate in capturing the intricate pricing dynamics influenced by the spatial considerations of both buyers and sellers.

Another way of spatial information inclusion is adding coordinates alongside other attributes (Li, 2022; Przekop, 2022). Therefore, using Google Maps API we geocode 100% of the property addresses and add latitude and longitude as attributes into our dataset. Empirical evidence from existing research indicates that advanced models capable of effectively leveraging geographical coordinates often consider them as significant predictors. Pace and Hayunga (2020) in their comparison study of tree-based ML methods with spatiotemporal techniques find that the superior performance of boosting and bagging regression trees to a great extent stems from exploiting spatial structures in the residuals that cannot be captured with location dummies, such as administrative area labels.

### 3.2. Data Preprocessing and Description

In summary, the following characteristics of the RCN information have been discovered:

- abundant missing data across several attributes, with more than two-thirds of observations lacking information,
- asymmetric distribution of quantitative variables. Due to real estate market heterogeneity, most distributions are skewed, hence methods used for the analysis should be robust to the violation of normality assumption, which is used in most traditional econometric methods,
- presence of rare characteristics, with some constituting less than five percent of observations. Therefore, analysis methods should be able to automatically consolidate some categories or disregard them if the impact is statistically negligible,
- presence of categorical variables with multiple levels, specifically the district features with 18 levels. Therefore, analysis methods should be able to effectively handle multi-level categorical variables without the need to dummy-code them, which can significantly increase the dimensionality of the dataset. Current implementations of RF and XGBoost directly accept categorical inputs and automate the encoding of categorical features,
- heterogeneity of the residential apartments market, e.g., high concentration of pre-war and newly developed high-end properties in the same district,
- heteroscedasticity, leading to increased variance in property values with larger property area. Therefore, the analysis methods should not make distributional assumptions,
- occurrence of non-typical transactions. Thus, the utilized methods should be robust to outliers.

Following data cleaning, the analyzed sample comprises 10,827 individual property deals with corresponding 56 variables. Leveraging geocoded coordinates, we generate Figure 1 by mapping the transactions onto the city map.

Table 1 presents a comprehensive overview of the utilized features, their respective units of measure, and encoding labels for binary and categorical variables. The selection of features is based on the literature review and the availability of data. The average property in the analyzed dataset is priced at 597,757.10 PLN, and has an area of 52.30 $m^2$ with 2.82 rooms, which results in the price of 11,429.39 PLN per $m^2$. The average property is located on the fourth floor of a total of seven floors in a 34-year-old building, which has one basement level underneath.

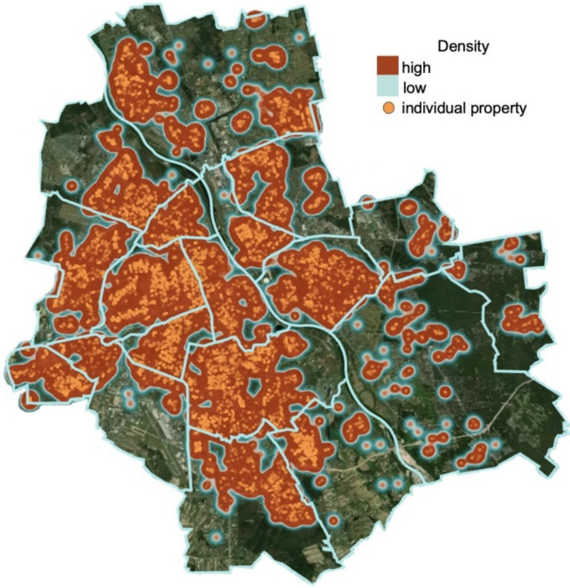



**Fig. 1**. Map of residential property sales in Warsaw, Poland in 2021. *Source*: own study.

**Table 1**

Variables of residential real estate

| Category | Feature | Feature Type | Description |
|---|---|---|---|
| Property characteristics | *Property price per sqm* | Numerical | Gross property price per square meter (m$^2$) in PLN |
| | *Usable property area* | Numerical | Usable property area in m$^2$ |
| | *Technical condition* | Categorical | Categorical feature defining property technical condition identified in RCN: normal = 0, in development = 1, renovation needed = 2, general renovation needed = 3 |
| | *Number of rooms* | Numerical | Number of rooms in a property |
| | *Property floor* | Numerical | Floor of a property in the building |
| Building characteristics | *Building floors* | Numerical | Number of floors in the building |
| | *Building underground floors* | Numerical | Number of underground non-residential floors in the building |
| | *Building age* | Numerical | Building age calculated as the difference between transaction and construction years |
| Locational characteristics | *District* | Categorical | Categorical feature defining districts encoding: Bemowo = 1, Białołęka = 2, Bielany = 3, Mokotów = 4, Ochota = 5, Praga-Północ = 6, Praga-Południe = 7, Rembertów = 8, Śródmieście = 9, Targówek = 10, Ursus = 11, Ursynów = 12, Wawer = 13, Wesoła = 14, Wilanów = 15, Włochy = 16, Wola = 17, Żoliborz = 18 |
| | *Urbanization ratios* | Categorical | Number of buildings per km$^2$ in a given district |
| | | Categorical | Percentage of developed and urbanized land from the total district area |
| | | Categorical | Percentage of residential-developed land from the total district area |
| | | Categorical | Percentage of industrial-developed land from the total district area |
| | | Categorical | Percentage of recreational-developed land from the total district area |
| | | Categorical | Percentage of communication-developed land from the total district area |
| | *Motorization ratio* | Categorical | Number of passenger cars registered per thousand population in a given district |
| | *Public transport ratios* | Categorical | Number of bus and tram stops per km$^2$ in a given district |
| | | Categorical | Number of metro stations in a given district |
| | *Housing development ratios* | Categorical | Number of residential buildings built in a given district |
| | | Categorical | Number of apartments put into use per thousand population in a given district |


eISSN: 2300-5289 

| | | |
|---|---|---|
| | Categorical | Average apartment usable area in $m^2$ in a given district |
| *Commercial infrastructure ratios* | Categorical | Number of stores per thousand population in a given district |
| | Categorical | Number of stores per $km^2$ in a given district |
| *Health infrastructure ratios* | Categorical | Number of hospitals per thousand population in a given district |
| | Categorical | Number of hospitals per $km^2$ in a given district |
| | Categorical | Number of pharmacies per thousand population in a given district |
| | Categorical | Number of pharmacies per $km^2$ in a given district |
| *Accommodation development ratio* | Categorical | Number of tourist accommodation facilities located in a given district |
| *Water-body adjacency indicator* | Binary | Binary feature defining if a district is adjacent to Wisła = 1, or not = 0 |
| *Greenery ratios* | Categorical | Percentage of park areas, green areas, street greenery and residential green areas from the total district area |
| | Categorical | Park areas, green areas, street greenery and residential green areas per person in $m^2$ in a given district |
| *Safety ratio* | Categorical | Percentage of reports to the city police from the total number of reports in a given district |
| *Population demographic ratios* | Categorical | Thousand population per km in a given district |
| | Categorical | Natural population growth per thousand population in a given district |
| | Categorical | Net migration for permanent residence per thousand population to / out of a given district |
| | Categorical | Percentage of people aged 0-14 from the total population in a given district |
| | Categorical | Percentage of people aged 15-64 from the total population in a given district |
| | Categorical | Percentage of people aged 65+ from the total population in a given district |
| *Unemployment ratio* | Categorical | Percentage of registered unemployed population from the labor force in a given district |
| *Education ratios* | Categorical | Kindergarten places per thousand kids in a given district |
| | Categorical | Preschools per thousand children in a given district |
| | Categorical | Middle schools per thousand pupils in a given district |
| | Categorical | Average middle school exam results in % in a given district |
| | Categorical | Average high school exam results in % in a given district |
| | Categorical | Average high school exam passing rate in % in a given district |
| *Cultural ratios* | Categorical | Number of museums in a given district |
| | Categorical | Number of theaters in a given district |
| | Categorical | Number of cinemas in a given district |
| | Categorical | Number of libraries in a given district |
| | Categorical | Number of cultural centers and clubs in a given district |
| *Sports ratio* | Categorical | Number of sports clubs and sections per thousand population in a given district |
| *Local government expenditure ratios* | Categorical | Revenues of a given district part of the city budget per capita in PLN |
| | Categorical | Expenditures of a given district part of the city budget per capita in PLN |
| *Business ratio* | Categorical | Number of REGON-registered entities per thousand population in a given district |
| *Latitude* | Numerical | Geographic coordinates of a property |
| *Longitude* | Numerical | |

*Source*: own study.

### 3.3. Machine Learning Framework

According to Antipov and Pokryshevskaya (2012), decision-tree-based algorithms successfully cope with the problems detected in our dataset, providing high accuracy and low variability, which makes them suitable for our research problem. These methods are also effective at automatically detecting and utilizing complex relationships between the features, including nonlinear relationships and interactions. Furthermore, they are robust to variable scaling and inherently efficient at handling categorical variables. Creating multiple decision trees and then obtaining the average prediction is called an ensemble of trees. One of the most popular ensembles is Random Forest (RF) proposed by Breiman (2001), which is a bagging-type ensemble consisting of bootstrapped samples of the training dataset. A decision tree is fully developed by selecting random explanatory variables at each node. This process is repeated numerous times, with each iteration involving a new bootstrapped sample. It ensures that each generated decision tree is independent of and uncorrelated with the others. The final estimation of the target variable is a weighted vote, or the average of the predictions of all trees in the collection. Another efficient approach is Extreme Gradient Boosting (XGBoost) introduced by Chen and Guestrin (2016), which is based on Gradient Boosting (GB) established by Friedman (2001). In GB, numerous decision trees built based on residual-like measures derived from the previous tree are constructed sequentially. It has three main components: a loss function to be optimized, a weak learner to predict, and an additive model to add weak learners to optimize the loss function (Yoshida et al., 2024). Eventually, the small trees are stacked together as a weighted sum of terms. Chen and Guestrin (2016) improved the algorithm by adding a regularization term to negate overfitting. XGBoost only considers a randomly chosen subset from the full array of available predictors at each split in the tree-growing process. This introduces an additional source of variation into the model, contributing to more broadly generalizable and robust estimations.

We select features for the final model configuration based on Feature Importance (FI), specifically feature weights, and Mutual Information (MI) while training the RF and XGBoost models, by removing the least important variables one at a time as long as it improves or has no impact on the model evaluation metrics. The prediction power of the models is evaluated with the mean absolute error (MAE), mean absolute percentage error (MAPE), and the root-mean-square error (RMSE)[1]. To further enhance the generalizability of the results, we assess the models' performance using k (5)-fold cross-validation (CV), where (k-1) folds are used for model training, and the hold-out fold is used for model testing. The CV process is then repeated k times, with each of the folds used exactly once as testing data. The results are then averaged to produce a single estimation.

### 3.4. Explainable AI Framework

As stated by Holzinger et al. (2022), the purpose of XAI is to "enable human experts to understand the underlying explanatory factors of why an AI decision has been made". After fully setting up our models, we utilize an XAI framework to shed light on the best models' estimation logic. Within XAI frameworks, the most popular approach which we apply in our research is Shapley Additive Explanations (SHAP), originally proposed by Lundberg and Lee (2017). These authors provide a unified approach to explaining model predictions based on Shapley (1953) values, which were initially introduced for fair payout distribution in cooperative games. Essentially, a Shapley value represents "a collaborator's average expected marginal contribution after accounting for all possible combinations" (Algaba et al., 2019), as given by Eq. 1

$$Shapley(X_j) = \sum_{S \subseteq N \setminus \{j\}} \frac{k!\,(p-k-1)!}{p!} \left(f(S \cup \{j\}) - f(S)\right) \tag{1}$$

where: $p$ – the total number of features, $N \setminus \{j\}$ – a set of all possible combinations of features excluding $X_j$, $S$ – a feature set in $N \setminus \{j\}$, $f(S)$ – the model prediction with features in $S$, $f(S \cup \{j\})$ – the model prediction with features in $S$ plus feature $X_j$.

SHAP is a model-agnostic post-hoc technique that quantifies the expected marginal contribution of each feature to all predictions made by the model, as described in Eq. 2

$$\hat{y}_i = shap_0 + shap(X_{1i}) + shap(X_{2i}) + \cdots + shap(X_{pi}) \tag{2}$$

where: $\hat{y}_i$ – the model prediction for observation $i$,

[1] The metrics chosen for this study most frequently appear in the research on the topic. Other metrics are often utilized as well – for more information on metrics for evaluating ML-based AVMs performance, see Steurer et al. (2021).

$shap_0$ – the expected value $E(\hat{y})$, $shap(X_{ji})$ – SHAP value of the $j^{th}$ feature for observation $i$ which represents the marginal contribution of that feature to the model prediction. Therefore, the sum of all SHAP values equals the difference between the actual prediction and the expected value. We also use SHAP Partial Dependence Plots (PDPs), which plot the feature value on the x-axis and the SHAP value of the same feature on the y-axis. This shows how the model depends on a given feature and is a richer extension of classical PDPs which are used for visualizing the marginal effect of a feature on the model prediction (Friedman, 2001).

## 4. Results and discussion

### 4.1. Model Evaluation and Selection

The metrics with RF and XGBoost best-tuned results in Table 2 show that both algorithms overfit to a certain degree, with predictions for withheld data showing an increase in the error metrics. XGBoost produces the lowest validation MAE of 1,009.91 PLN/m$^2$, RSME of 1,308.49 PLN/m$^2$, and MAPE of 8.99%.

The models developed in this paper provide precise and stable estimations, with significantly lower overfitting and higher robustness than, e.g., neural networks developed in a comparison study of ANNs and spatial regression models on similar data in Przekop (2022). The MAPE of 8.99% obtained in this study can be interpreted as highly accurate forecasting in line with the criteria established by Lewis (1982, p. 40).

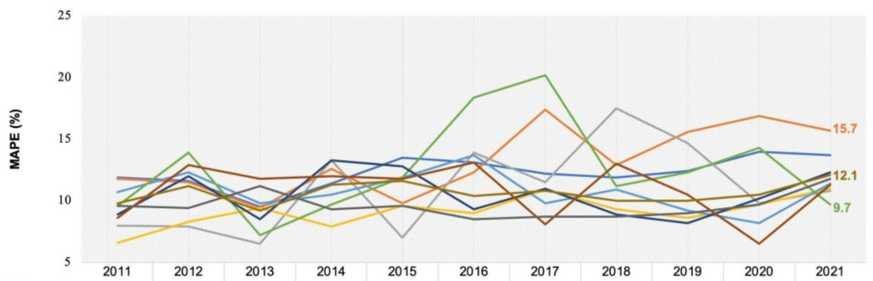
**Table 2**

RF and XGBoost performance metrics

| Metric | Random Forest | | XGBoost | |
|---|---|---|---|---|
| | Train result | Validation result | Train result | Validation result |
| *MAE* | 798.18 | 1,030.99 | 858.58 | 1,009.91 |
| *MAPE* | 0.0710 | 0.0918 | 0.0765 | 0.0899 |
| *RSME* | 1,034.23 | 1,328.74 | 1,113.37 | 1,308.49 |

*Source*: own study.

However, according to Michael Gilliland (2010), it is irresponsible to set arbitrary performance targets (such as MAPE<10% - Excellent, MAPE<20% - Good) without the context of the data, as the criteria that denote model excellence may significantly vary based on the local context. To put this issue into perspective, MSCI (2022) reports a global appraisal MAPE of 12.1% in 2021, as shown in Figure 2, with local values varying from 9.7% in Switzerland to 15.7% in Germany. Therefore, the models demonstrate their effectiveness as valuable support tools for expert real estate valuation systems in the local context.




| | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| France | 11.9 | 11.6 | 9.5 | 11.4 | 13.5 | 13.1 | 12.2 | 11.9 | 12.4 | 14 | 13.7 |
| Germany | 11.8 | 11.5 | 9.4 | 12.6 | 9.8 | 12.3 | 17.4 | 12.9 | 15.6 | 16.9 | 15.7 |
| Italy | 8 | 7.9 | 6.5 | 13.3 | 7 | 13.9 | 11.5 | 17.5 | 14.7 | 9.8 | 10.8 |
| Netherlands | 6.6 | 8.3 | 9.4 | 7.9 | 9.6 | 9 | 10.9 | 9.3 | 8.6 | 9.7 | 11 |
| Sweden | 10.7 | 12.3 | 9.8 | 10.5 | 11.9 | 13.7 | 9.8 | 10.9 | 9.2 | 8.2 | 11.4 |
| Switzerland | 9.6 | 13.9 | 7.2 | 9.7 | 11.9 | 18.4 | 20.2 | 11.2 | 12.3 | 14.3 | 9.7 |
| UK | 8.9 | 12 | 8.5 | 13.3 | 12.8 | 9.3 | 11 | 8.9 | 8.2 | 10.2 | 12.3 |
| Japan | 8.6 | 12.9 | 11.8 | 12 | 11.8 | 13.1 | 8.1 | 13 | 10.5 | 6.5 | 11.3 |
| USA | 9.6 | 9.4 | 11.2 | 9.3 | 9.6 | 8.5 | 8.7 | 8.7 | 9 | 9.7 | 11.8 |
| Global | 9.8 | 11.2 | 9.2 | 11.3 | 11.6 | 10.4 | 10.8 | 10 | 10 | 10.5 | 12.1 |



**Fig. 2**. Real estate appraisal MAPE (%). *Source*: own study, based on MSCI. (2022). Private Real Estate: Valuation and Sale Price Comparison 2021.

### 4.2. Model Interpretation

At this point, we apply SHAP to the best-performing model, specifically XGBoost, to extract the details about the model predictions in the validation set and relationships between the variables. The focus has been placed on the impact of eight selected features out of the 55 initially introduced into the model due to the lack of significance that the remaining 47 features had on the model. The remaining eight features with a significant influence are put under more scrutiny to showcase the underlying explanatory factors of the model's decision making. SHAP allows


eISSN: 2300-5289 | Received 2024-05-22 | Revised 2024-09-18 | Accepted 2024-11-21

for both local and global decomposition of predictions. Using a waterfall plot to visualize the SHAP values for a single observation as shown in Figure 3, we can investigate the model's logic on the local prediction level. The waterfall plot starts with the expected value of the model output at the bottom, E[f(X)] equal to 11,477.04 PLN/m$^2$, and then each row shows how the positive or negative contribution of each feature moves the value to the model's output for a given prediction, f(x) of 12,148.60 PLN/m$^2$. This specific property is a residence with a usable area of 51.64 m$^2$ and three rooms. The apartment is located on the 4$^{th}$ floor in a four-story residential building constructed in 1959. Evidently in this case, the fact that the apartment complex is relatively old has a strong negative impact on the price. Additionally, we can infer that, given the number of floors in the building, the apartment being situated on the highest floor has a positive impact on the predicted price. The property is located in Praga-Południe, with its exact location having the most significant positive effect on the final prediction, since latitude and longitude pinpoint the property's location to be Saska Kępa. Nonetheless, there is a unique waterfall plot for every observation in the dataset, so it is problematic to extend these conclusions about the features' importance or impact on the whole model.

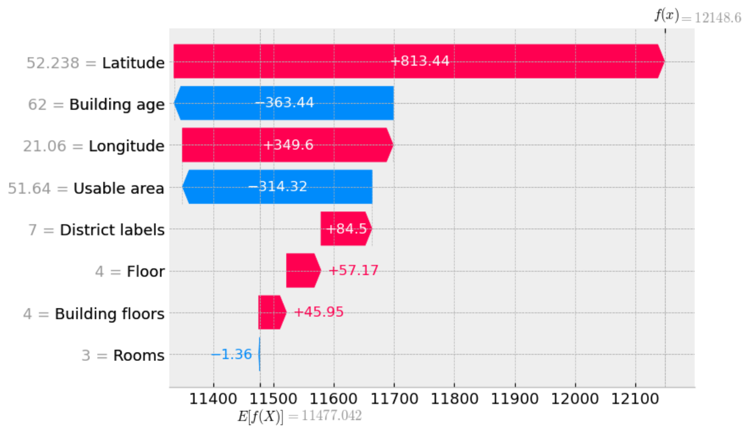



**Fig. 3**. The waterfall plot of the model SHAP values for a single observation. *Source*: own study.

Consequently, to facilitate global interpretability we use layered violin and heatmap plots. On the left side, the violin plot offers a compact representation of the distribution and density of SHAP values, providing insight into the range, variability, skewness and symmetry of the SHAP values distribution for each feature. On the y-axis, individual violin plots are stacked by the sum of absolute SHAP values per feature, making it suitable for use as a feature importance score (Molnar, 2020). The x-axis shows whether the value effect is associated with a higher or lower prediction. Color scale shows whether variable values are high (red) or low (blue) for the observations. On the right side, a matrix of SHAP values passed to the heatmap function creates a plot with the observations on the x-axis, features on the y-axis, and the SHAP values encoded on the color scale. The order of observations is based on hierarchical clustering by their explanation similarity. The model output is shown above the heatmap matrix and is centered around the expected value, while the global importance of each feature is presented as a bar chart on the right side of the plot.

Figure 4 reveals the overall feature importance in the model's predictions on the global level, with building age, geographic coordinates and usable property area ranking high, and district labels, building and property floors, and number of rooms ranking relatively low. Based on the violin plot, we can observe that higher building age accounts for lower predicted price per square meter, which could reflect factors such as depreciation, maintenance costs, or changing preferences for newer constructions. The influence of latitude and longitude exhibits a less straightforward pattern due to the expected interaction effects, thereby complicating interpretation. At the same time, it appears that smaller apartment areas tend to yield higher predicted prices per square meter, suggesting a potential premium for properties with limited space. The remaining features have lower impact and are evenly split, which indicates that, while generally remaining important, their interaction is dependent on the other variables. The heatmap shows that some observations that have similar predictions for similar reasons are grouped together. There are noticeable clusters of instances with a high and low impact on the model output (shown by color intensity) influenced largely by longitude, latitude, or building age, as well as high impact influenced by usable area. Comparing the positions of these clusters with the model output function above the heatmap shows that, in many instances, certain variables have a predominant impact on the predicted price, as marked on the plot, while still interacting with other features in the same or opposite direction.

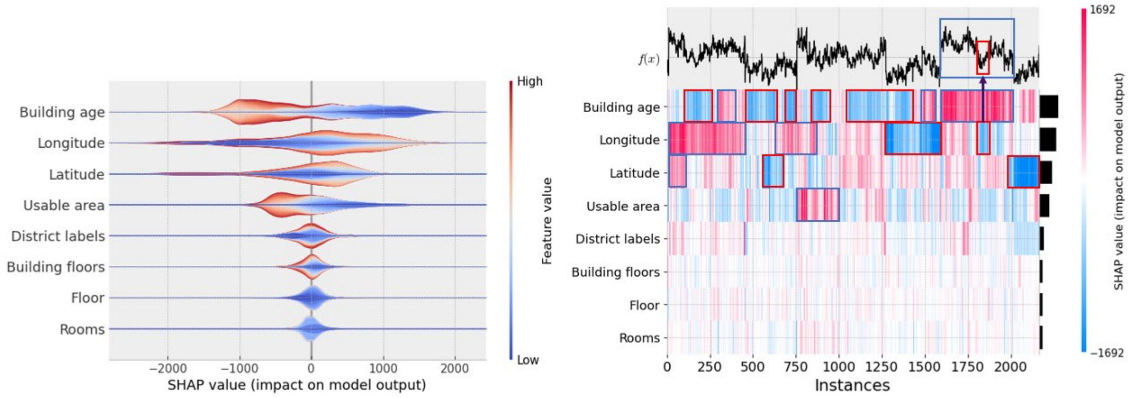



**Fig. 4**. SHAP layered violin and heatmap plots. *Source*: own study.

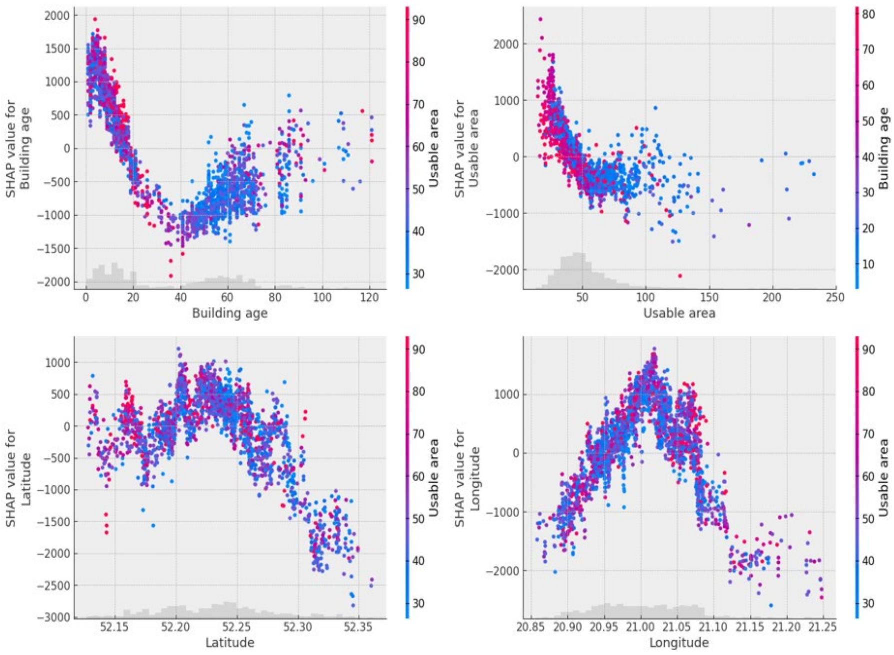



**Fig. 5**. SHAP Partial dependence plots of significant variables against their SHAP values using the Usable Area and Building Age color scale. *Source*: own study.

Figure 5 presents SHAP PDPs. Examining the plot for building age in the upper left corner, we observe that the individual effect of this feature on the predicted price is nonlinear. The SHAP values show the cut-off point for the positive impact of building age to be approximately 20 years, with higher age generally starting to have negative effect on the price. However, the negative trend reverses at age 40, possibly indicating a gradual shift in perception or appreciation of the architectural styles accumulated in the city throughout the years. The upper right plot shows the nonlinear impact of usable area on the predicted price per square meter, which is typically higher for smaller apartment areas under 50 m$^2$ and stays stable for bigger areas. This trend could reflect factors such as location desirability, demand for compact living spaces, or higher-end features in smaller units. The color coding of data points represents building age, where blue denotes younger buildings, and red signifies older buildings. While the distribution of these colored-coded points is predominantly heterogeneous, a greater concentration of red points continues to be observed on the left-hand side, while a comparatively higher number of blue points is located on the right-hand side. This can be attributed to historical urban development patterns, where smaller living spaces were more prevalent. Additionally, renovations or conversions of older buildings may have subdivided larger units into smaller ones to meet evolving housing needs and market demands. The influence of geographic coordinates on the predicted price per m$^2$, as illustrated in the lower plots, also follows a nonlinear pattern. The turned v shape of the plots indicates that location closer to the city center has the most substantial positive impact on the predicted price. The peak occurs at approximately (52.23, 21.02), which corresponds to a location within one kilometer walking distance of the Palace of Culture and Science. Moving towards the city periphery, the impact on the predicted price becomes notably negative. As Deppner et al. (2023) conclude, when a regression tree splits on the latitude and longitude, it effectively identifies new submarkets for which it generates individual sub-models, accounting for spatial considerations on the micro-level. Hence, this could confirm the model's correct understanding of the collective contribution to increased demand (and subsequently higher prices) in central areas of a variety of factors such as proximity to employment centers, amenities, transportation hubs, and the overall desirability of central locations in source coordinate data. Although, the ability to further examine this understanding with higher granularity and the same efficiency remains uncertain, as district-level variables representing these factors did not demonstrate the expected level of significance at the model specification stage.

The whole empirical exercise demonstrates how the presented methods can capture and visualize pattern changes hidden in the data to produce accurate estimations. Economic researchers have long pursued methods to identify what drives performance of market mechanisms in specific contexts. Therefore, knowledge that could help develop an explainable algorithm allowing to precisely model hidden patterns and provide insight into its decision-making process would be a significant step forward in socio-economic analysis. This kind of model can eventually be deployed on production systems and become a valuable part of a real estate valuation expert system to complement a human expert's knowledge.

## 5. Conclusions

The real estate industry is progressively integrating Big Data and Artificial Intelligence, thereby offering new services and improving existing industry processes. The purpose of this article is to provide an explainable perspective on the workings of bagging and boosting tree-based algorithms performing hedonic pricing of residential real estate. A large database with 10,827 transactions in Warsaw, Poland is created, with information obtained from various sources, such as local real estate registers, cadastral information, socio-demographic and economic statistics, and coordinate information.

The best-performing model in this study estimating property price per m$^2$ is XGBoost with MAE of 1,009.91 PLN/m$^2$, RSME of 1,308.49 PLN/m$^2$, and MAPE of 8.99%. Model-agnostic XAI methods are then used to summarize feature importance and effects. In our empirical analysis, building age, geographic coordinates and usable property area have the greatest influence. The inspection of SHAP violin and heatmap plots shows that the model understands interactions between the most significant predictors, while SHAP partial dependence plots illustrate how the model depends on a given feature and their individual nonlinear impact on the predicted target variable. This showcases the practicability of applying XAI frameworks in the real estate market context to decode the rationale behind data-driven algorithms, thus achieving the research goal of our study. This adds to the growing body of research on this developing topic, strengthening the claim to external validity of obtained results. In order to be ready for broader application, this work can be extended in several ways.

Firstly, the research can be extended across space. As demonstrated in prior research on the topic, real estate prices data frequently exhibits spatial autocorrelation (Basu & Thibodeau, 1998; Cellmer, 2013) and heterogeneity (Wu et al., 2020), so further research is needed with ensemble-based learning approaches aimed to be used with spatially heterogeneous data. Extension of tree-based models for spatially dependent data is a developing field which includes, e.g., the spatial RF based on higher-order spatial statistics of Talebi et al. (2022) or the RF-GLS of Saha et al. (2023). Following the findings of Li (2022), spatial effects could also be retrieved by mapping SHAP values of interaction effects from coordinates as spatial features and checked for consistency with estimations in spatial statistical models at the parameter level for real estate valuation. Moreover, Deppner and Cajias (2022) propose a spatially conscious alternative for conventional k-fold cross-validation accounting for spatial dependence, which could change the approach to model selection in the context of real estate research.

Secondly, the study can be extended across time. Yet it is crucial to bear in mind that selecting an appropriate machine learning model for time-series data analysis is yet another task. Suitable models include, for instance, certain types of neural networks; however, this would pose a challenge in terms of AI explainability, since these complex neural networks have more limited explanatory techniques than those used in the current paper.

Finally, data availability is one of the current challenges. Data-driven machine learning models require extensive volumes of representative training and validation observations to produce unbiased and dependable results. A missing feature or observations excluded due to data entry errors can introduce bias, diminishing the overall informativeness of the system. Therefore, we encourage both public authorities and private institutions to improve data quality and support database maintenance.

## References

Algaba, E., Fragnelli, V., & Sánchez-Soriano, J. (Eds.). (2019). Handbook of the Shapley value. CRC Press. https://doi.org/10.1201/9781351241410

Alonso, W. (1964). *Location and land use: Toward a general theory of land rent.* Harvard University Press., https://doi.org/10.4159/harvard.9780674730854

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *The Journal of Economic Perspectives, 24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

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. https://doi.org/10.1016/j.eswa.2011.08.077

Arribas, I., García, F., Guijarro, F., Oliver, J., & Tamošiūnienė, R. (2016). Mass appraisal of residential real estate using multilevel modelling. *International Journal of Strategic Property Management, 20*(1), 77–87. https://doi.org/10.3846/1648715X.2015.1134702

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics, 17*, 61–85. https://doi.org/10.1023/A:1007703229507

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. https://doi.org/10.1007/s11146-019-09712-0 https://doi.org/10.1007/s11146-019-09727-7.

Borgoni, R., Michelangeli, A., & Pontarollo, N. (2018). The value of culture to urban housing markets. *Regional Studies, 52*(12), 1672–1683. https://doi.org/10.1080/00343404.2018.1444271

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Cellmer, R. (2013). Use of spatial autocorrelation to build regression models of transaction prices. *Real Estate Management and Valuation, 21*(4), 65–74. https://doi.org/10.2478/remav-2013-0038

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785 https://doi.org/10.5194/acp-20-8063-2020.

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, 68*, 235–273. https://doi.org/10.1007/s11146-022-09915-y

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*, 1–38. https://doi.org/10.1007/s11146-023-09944-1 PMID:38625136

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gilliland, M. (ed.). (2010). *The Business forecasting deal: exposing myths, eliminating bad practices, providing practical solutions*. John Wiley & Sons., https://doi.org/10.1002/9781119199885

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022). Explainable AI methods – A brief overview. In A. Holzinger, R. Goebel, R. Fond, T. Moon, K. R. Müller, & W. Samek (Eds.), *xxAI – Beyond explainable AI* (pp. 13–38). Springer., https://doi.org/10.1007/978-3-031-04083-2_2

Hong, J., Choi, H., & Kim, W. S. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management, 24*(3), 140–152. https://doi.org/10.3846/ijspm.2020.11544

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. https://doi.org/10.1016/j.landusepol.2018.12.030

Kauko, T. (2006). What makes a location attractive for the housing consumer? Preliminary findings from metropolitan Helsinki and Randstad Holland using the analytical hierarchy process. *Journal of Housing and the Built Environment, 21*, 159–176. https://doi.org/10.1007/s10901-006-9040-y

Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *Journal of Portfolio Management, 43*(6), 202–211. https://doi.org/10.3905/jpm.2017.43.6.202

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*(2), 132–157. https://doi.org/10.1086/259131

Lewis, C. D. (1982). *Industrial and business forecasting methods*. Butterworths.

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems, 96*, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2023). Interpretable machine learning for real estate market analysis. *Real Estate Economics, 51*(5), 1178–1208. https://doi.org/10.1111/1540-6229.12397

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30. ISBN: 978-1-5108-6096-4.

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. https://doi.org/10.1108/JERER-08-2018-0035

Molnar, C. (2020). *Interpretable Machine learning. A guide for making black box models explainable*. Lean Publishing.

Montero, J. M., & Fernández-Avilés, G. (2014). Hedonic Price Model. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 2834–2837). Springer., https://doi.org/10.1007/978-94-007-0753-5_1279

Mora-Garcia, R. T., Cespedes-Lopez, M. F., & Perez-Sanchez, V. R. (2022). Housing price prediction using machine learning algorithms in COVID-19 times. *Land (Basel), 11*(11), 2100. https://doi.org/10.3390/land11112100

MSCI. (2022). *Private real estate: Valuation and sale price comparison 2021*.

NBP. (2022). *Raport o sytuacji na rynku nieruchomości mieszkaniowych i komercyjnych w Polsce w 2021 r.* [Report on the situation in the residential and commercial real estate market in Poland in 2021].

Niu, F., & Liu, W. (2017). Modeling urban housing price: The perspective of household activity demand. *Journal of Geographical Sciences, 27*, 619–630. https://doi.org/10.1007/s11442-017-1396-2

Osland, L. (2010). An application of spatial econometrics in relation to hedonic house price modeling. *Journal of Real Estate Research, 32*(3), 289–320. https://doi.org/10.1080/10835547.2010.12091282

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*, 170–180. https://doi.org/10.1007/s11146-019-09724-w

Przekop, D. (2022). Artificial neural networks vs spatial regression approach in property valuation. *Central European Journal of Economic Modelling and Econometrics, 14*, 199–223. https://doi.org/10.24425/cejeme.2022.142630

Rico-Juan, J. R., & Taltavull de La Paz, P. T. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications, 171*, 114590. https://doi.org/10.1016/j.eswa.2021.114590

Rosen, S. (1974). Hedonic Prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55. https://doi.org/10.1086/260169

Saha, A., Basu, S., & Datta, A. (2023). Random forests for spatially dependent data. *Journal of the American Statistical Association, 118*(541), 665–683. https://doi.org/10.1080/01621459.2021.1950003

Sevgen, S. C., & Tanrivermiş, Y. (2024). Comparison of machine learning algorithms for mass appraisal of real estate data. *Real Estate Management and Valuation, 32*(2), 100–111. https://doi.org/10.2478/remav-2024-0019

Shapley, L. (1953). 17. A Value for n-Person Games. In H. Kuhn & A. Tucker (Eds.), *Contributions to the theory of games* (Vol. AM-28, pp. 307–318). Princeton University Press., https://doi.org/10.1515/9781400881970-018

Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research, 38*(2), 99–129. https://doi.org/10.1080/09599916.2020.1858937

Straszhem, M. (1987). The theory of urban residential location. In Handbook of Regional and Urban Economics (Vol. 2, pp. 717–757). Elsevier., https://doi.org/10.1016/S1574-0080(87)80004-4

Talebi, H., Peeters, L. J., Otto, A., & Tolosana-Delgado, R. (2022). A truly spatial random forests algorithm for geoscience data analysis and modelling. *Mathematical Geosciences, 54*, 1–22. https://doi.org/10.1007/s11004-021-09946-w

Statistical Office in Warsaw. (2022). *Panorama dzielnic Warszawy w 2021 r.* [Panorama of Warsaw districts in 2021].

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance, 38*(3), 213–225. https://doi.org/10.1108/JPIF-12-2019-0157

Wheaton, W. C. (1977). Income and urban residence: An analysis of consumer demand for location. *The American Economic Review, 67*(4), 620–631. https://www.jstor.org/stable/1813394

Wu, Y., Wei, Y. D., & Li, H. (2020). Analyzing spatial heterogeneity of housing prices using large datasets. *Applied Spatial Analysis and Policy, 13*, 223–256. https://doi.org/10.1007/s12061-019-09301-x

Yoshida, T., Murakami, D., & Seya, H. (2024). Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset. *The Journal of Real Estate Finance and Economics, 69*, 1–28. https://doi.org/10.1007/s11146-022-09929-6

Public Transport Authority in Warsaw. (2022). *Informator statystyczny 2021* [Statistical guide 2021].

Public Transport Authority in Warsaw. (2022). *Informator statystyczny nr XII (333)* [Statistical guide no. XII (333)].

Zyga, J. (2019). Data selection as the basis for better value modelling. *Real Estate Management and Valuation, 27*(1), 25–34. https://doi.org/10.2478/remav-2019-0003