ISSN: 1648-715X / eISSN: 1648-9179

VILNIUS TECH
Vilnius Gediminas Technical University

INTERNATIONAL JOURNAL OF STRATEGIC PROPERTY MANAGEMENT

2024
Volume 28
Issue 6
Pages 393–410
https://doi.org/10.3846/ijspm.2024.22686

# PROPERTY APPRAISAL VIA LENS OF PROPERTY REGISTRATION ABUNDANCE – REAL ESTATE MARKET ASYMMETRY ASSESSMENT

Marek WALACIK *

*Department of Real Estate and Urban Studies, University of Warmia and Mazury in Olsztyn, Olsztyn, Poland*

**Article History:**
- received 9 June 2024
- accepted 11 November 2024

**Abstract.** Information on transaction prices and the ones characterizing the property as the subject of the valuation are essential for a proper valuation process. The accuracy and completeness of the collected set of information directly affects the quality of the valuation process. When market participants operate on the basis of unequal sets of information, information asymmetry is revealed. This research investigates the effects of information asymmetry on property market from the perspective of property registration abundance and mass appraisal systems. It explores how disparities in information abundance and quality within property registration and appraisal processes can affect market fairness and transparency. Employing a mixed-methods approach, it analyses property transaction data and tries to investigate effects of information asymmetry. The findings indicate that enhanced transparency and data quality can significantly reduce valuation discrepancies and lead to a more equitable real estate market. The study concludes with recommendations aimed at justifying information asymmetry's negative effects, supporting for policies that promote information uniformity to improve the fairness and efficiency of property registration and mass appraisal practices.

**Keywords:** information asymmetry, property valuation, random forest, neural networks, multiple linear regression.

*Corresponding author. E-mail: *marek.walacik@uwm.edu.pl*

## 1. Introduction

The real estate market is deeply influenced by the dynamics of information flow and transparency. Effective property management and valuation depend on the accuracy, completeness, and symmetry of information available to all market participants. From that perspective the influence of information asymmetry on the real estate market becomes a critical area of study. This is particularly important considering the strategic role of property management in increasing asset value, ensuring sustainability, and utilizing technological advancements in data handling and analysis. This paper explores the diverse impacts of information asymmetry in real estate transactions and property valuations, especially from the perspective of property registration understood as formal process of recording the transfer of ownership or rights to a property from one party to another in a government-maintained registry or database. It focuses on disparities in the quality and abundance of information, understood as the combination of the volume of properties registered, the depth and completeness of the information collected during registration, and how they can bias market operations, leading to inefficiencies and inequities were subject of investigation. Given the complexities of modern real estate markets and the increasing application of digital technologies, the three main research questions (RQ) were articulated, that in author's opinion, enabled substantial contribution to the improvement of the real estate market information asymmetry significance comprehensibility via the lens of substantially structured and scientifically derived reasoning:

RQ1: How do registered property transaction attributes influence the degree of information asymmetry in real estate markets?

RQ2: To what extent do advanced property valuation algorithms, respond to varying levels of information symmetry in the dataset?

RQ3: What are the economic consequences of information asymmetry in property markets, particularly in terms of valuation accuracy and fairness?

The RQ1 addresses mutually inconsistent in real market analysis desire for, on one hand, extensive data collection resulting from rapid "application of digital technologies such as cloud computing, big data, and blockchain" (Wang et al., 2024), on the other hand, limited from behavioural point of view, human ability to make decisions on the basis of maximum seven variables – in this case property attributes (Ries & Trout, 1994). Dealing with the issue of the relation between quality and quantity of data, the question

aligns with Ludwig Mies van der Rohe's consideration if "less is more"? (Schulze & Windhorst, 2014) By focusing on the direct relationship between the quantity of transaction attributes and consequently the level of information asymmetry, the research tries to verity if straightforward increase in disclosed information can lead to greater clarity and less asymmetry, thereby improving the market's functionality and fairness for all participants.

The RQ2 is designed to evaluate how well modern, sophisticated property valuation algorithms adapt to the diverse and varying information structures on the basis of selected countries' property transaction registration solutions. This inquiry is crucial as it explores the capability of these algorithms to deliver consistent and accurate property valuations, addressing challenges posed by information asymmetry and the need for transparency and equity in property valuation. By focusing on the responsiveness of these advanced tools, the question aims to reveal their effectiveness in navigating the property appraisal complexities, thereby aiding stakeholders in making informed, reliable investment and policy decisions in the face of varying data availability and market conditions.

The RQ3 is crucial for exploring how disparities in information access and quality affect individuals in economic meaning. This inquiry seeks to reveal how such asymmetry can lead to poor decision-making, economic inefficiencies, consequently reduced trust in e.g. public institutions. By examining these impacts, the research aims to inform policies and interventions designed to enhance transparency, fairness, and well-being, ultimately supporting more equitable outcomes. This question is vital for understanding and mitigating the negative effects that arise when individuals or groups are disadvantaged by unequal access to important information.

Providing the following form of reasoning (research questions formulation), guided the research process in a way that enabled clear focus on articulated research problem with reference to the variety of classical theories and paradigms embedded in property valuation e.g. the highest and best use paradigm (Dotzour et al., 1990; Vandell, 1982; Vandell & Carter, 2000), the value paradigm (Sayce et al., 2006; Trinh, 2018), the location theory (Alonso, 1964), the sustainable development paradigm (Campbell, 2018; Ogryzek, 2023), the externalities theory (Batabyal, 2023; Cornes & Sandler, 1996; Tisdell, 1970). The highest and best use paradigm, alongside the value paradigm, were both critical to formulating the research assumptions due to their complementary roles in property valuation. The highest and best use paradigm emphasizes that a property's value is maximized when utilized in a legally permissible, physically possible, and financially feasible manner (Ragil Budi Perkasa et al., 2023; Ribera et al., 2020; Rymarzak et al., 2022; Utomo et al., 2018; Danastri Yuwono et al., 2023). This paradigm is essential for understanding how different factors, such as land use regulations and market conditions, directly affect the optimal use of a property, and thus its valuation. In conjunction, the value paradigm offers a broader economic perspective, which posits that property value is influenced by factors such as demand, scarcity, and transferability (d'Amato & Kauko, 2017; Sayce & Connellan, 2002). By integrating both paradigms, the research can better capture the multidimensional aspects of property analysis – where the interplay between legal, physical, and economic factors shapes the valuation process. These frameworks are crucial for examining how information asymmetry impacts perceived property value and optimal use. The abundance or scarcity of property information, as informed by these paradigms, influences stakeholders' ability to determine the highest and best use, thereby affecting property value and market dynamics. Therefore, incorporating both paradigms is necessary to provide a comprehensive understanding of property valuation in the context of evolving information transparency. For that reason the need for temporal consistency of property transactions was assumed so that the features do not differentiate the set of data being analyzed. Another research assumption concerning locational consistency was determined by the location theory, the theory highlights the need for spatial elements exploration and the externalities theory that assumes mutual indirect interdependencies between objects in economic space. The assumption derives from the need of decreasing property market analysis uncertainty and subjectivity described and investigated carefully by Renigier-Bilozor et al. (2019). Lastly, the sustainable development paradigm determined the need of environmental considerations inclusion. All the aforementioned assumptions constituted the foundation of the research that was executed according to particular steps presented in Figure 1.

The research expands the current state of knowledge of the phenomenon of information asymmetry with a completely new context, to which little scientific attention, has been paid so far. The execution of the presented research architecture enabled the following contribution to

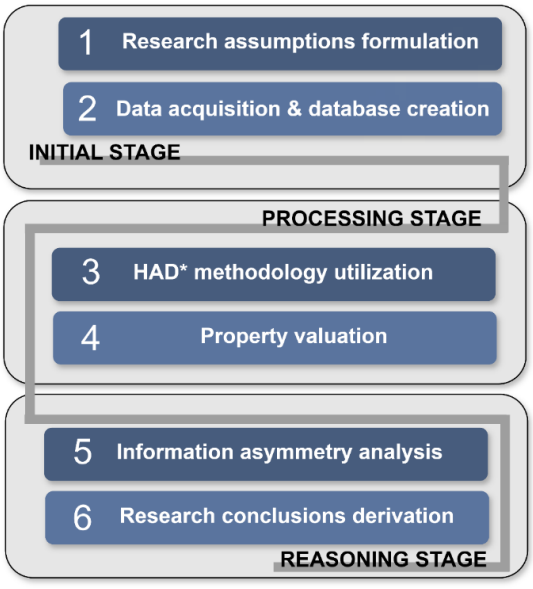



**Figure 1.** Research procedure (source: own elaboration)

the current state of art scientific literature dealing with the challenges of property market analysis and valuation enhancement comprehension of data utilization, evaluation of advanced valuation algorithms, integration of classical theories and paradigms. Additionally, by providing a comprehensive internationally based analysis of information asymmetry in a real estate context – a subject less explored in current literature – the paper expands the understanding of this phenomenon in a new and complex market setting.

Having provided a brief introduction to the scientific problem, the remaining part of the scientific discussion presented in the paper was structured in the following way: Section 2 delivers theoretical considerations for the research based on the current state of art literature review with special respect to the of information asymmetry phenomenon and the issue of property registration in property valuation. Section 3 presents the study area, the used sources and the justification for particular assumptions adoption. This section also describes the rational for selected methods, algorithms utilization e.g. modified HAD methodology. Section 4 presents the research results whereas Section 5 information asymmetry analysis. Section 6 presents conclusions, comparison of the results with recent state of art finding in the field, discussion and potential areas for future studies.

## 2. Information asymmetry in real estate market – literature review

In real estate market analysis, the quality and depth of information are essential for increasing market transparency, ensuring accurate property valuation, and facilitating efficient market operations (Ionascu et al., 2019). Comprehensive, timely, and accurate data allow investors, developers, and policymakers to mitigate market failures like adverse selection and moral hazard (Klein et al., 2016), which occur when one party in a transaction possesses more or better information than the other. As noted by (Ben-Shahar & Golan, 2019), "economists have long recognized the central role of information in the operation of markets". While there is a wealth of literature emphasizing the importance of information in real estate (Ambrose & Diop, 2021; Bergh et al., 2019; Gatzlaff & Tirtiroğlu, 1995; Brzezicka et al., 2022), what is often underexplored is the specific role that property transaction registration plays in reducing information asymmetry and promoting market transparency – Table 1. This gap is particularly notable given the growing reliance on data analytics and information systems in real estate (Huber et al., 2021; Jung et al., 2022). Although existing studies have addressed the broader implications of information asymmetry in market operations (Chau et al., 2007; Kurlat & Stroebel, 2014), few have directly examined how the formal registration of property transactions can mitigate these asymmetries.

This is a critical gap, as transparent registration systems provide a verifiable record of property attributes and transaction histories, which can significantly reduce the uncertainty faced by less-informed market participants. Moreover, although researchers like Garmaise and Moskowitz have pioneered methods to measure information asymmetry through property taxation and transaction history (Garmaise & Moskowitz, 2004), their work does not fully address how the digitalization and public availability of property transaction records influence market dynamics. This gap is particularly relevant in the context of mass appraisal and automated valuation models (Gdakowicz et al., 2019), where the availability of transaction data can significantly affect the accuracy and fairness of property valuations. This research aims to fill this gap by focusing specifically on the under-researched area of property transaction registration and its impact on information asymmetry in real estate markets. By investigating the ways in which registration practices either alleviate or exacerbate information imbalances, this study contributes new insights to both academic discussions and practical applications in property market transparency and valuation. In conclusion, while the literature has addressed various aspects of information asymmetry in real estate (Aizenman & Jinjarak, 2009; Li & Chau, 2024), the role of transaction registration remains scientifically neglected. This research seeks to deepen the academic understanding of this issue and provide actionable insights for stakeholders, contributing to more transparent and efficient real estate markets.

**Table 1.** Examples of selected studies on information asymmetry in real estate market, with special respect to the context of analysis, implemented methods and sources of asymmetry (source: own elaboration)

| Source of analysis | Utilized methodology | Real estate market context | Source of asymmetry |
|---|---|---|---|
| Ambrose and Shen (2023) | Bayesian learning model | Impact of fracking risk on house buyers | Lack of past experience |
| Ling et al. (2018) | Hedonic model | Impact of anchoring and search costs | Reference to other markets |
| Chau and Wong (2016) | ECM & SUR* | Decomposition of land and building value | Complex nature of property |
| Zhou et al. (2015) | Hedonic model | Investments of local/non-local buyers | Real estate market locality |
| Chinloy et al. (2013) | Hedonic model | Investments of local/non-local buyers | Real estate market locality |
| Wong et al. (2012) | Panel data analysis | Impact of warranties on house buyers | Perception of building quality |
| Levitt and Syverson (2008) | Hedonic model | Impact of brokers on property buyers | Data manipulation |
| Johnson et al. (2005) | Hedonic model | Impact of listings on property buyers | Kind of listing |

*Note:* * ECM – Error Correction Model; SUR – Seemingly Unrelated Regression.

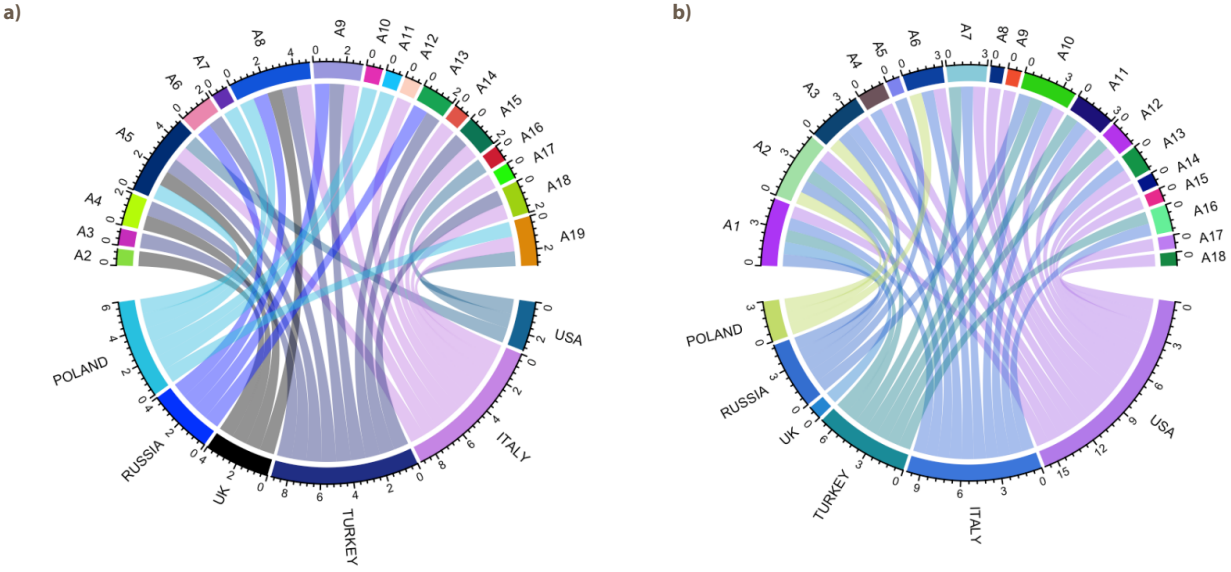
## 3. Material and methods

The implementation of the research according to the assumed procedure, presented in Figure 1, required, substantially justified selection of: case study reference, implemented methods, and the scope data being subject of investigation. Considering the fact that RQ1 is based on the scope of property transactions attributes registration, what had to be done before proper analysis was to identify with what level of detail real estate transaction data is recorded in selected countries. In order to answer this question, it was decided to conduct a survey among real estate professionals dealing directly with real estate price registers. That kind of method was justified by several compelling reasons. Firstly, the insights derived from the responses of industry professionals provide a grounded understanding of the current challenges and nuances within the real estate sector that may not be evident from secondary data alone. Moreover, collecting expert feedback ensured that the case study is relevant and tailored to real-world applications, improving its practical value to other professionals and policymakers in the field. Figure 2 presents the scope of data provided in either directly property transaction registers in selected countries (A) or data that can be derived from other public registers (B) – the figure was generated with the use of R-studio software and presents data provided by real estate professionals via on-line survey executed by the author. All the collected answers determined the scope of reference asymmetry metrics calculations.

Considering the fact that, according to the information acquired in the questionnaire, property transactions registration in US had the most detailed scope data collection (19 attributes), the US database structure formed the basis for further studies in the initial stage – "data acquisition and database creation" in selected case study area.

### 3.1. Study area

The research was conducted in the City of Raleigh, North Carolina, USA. The selection of Raleigh as the case study area was influenced by several strategic considerations, as well as the unique opportunity provided by the author's academic affiliation and participation in a research visit at the University of North Carolina[1]. This connection enabled the author access to local data, administrative cooperation, and substantive support, which significantly enriched the study. The choice of Raleigh was informed not only by its geographical and economic characteristics but also by the logistical feasibility and willingness of local government bodies to cooperate (Tax Office in Wake County Government), ensuring that the methodology could be tested thoroughly. Raleigh offers a diverse urban landscape with varying property types, sizes, and land uses, which presents a comprehensive testing ground for the research methodology. The city has undergone significant population growth, economic diversification, and urban planning reforms, making it an ideal candidate for studying the adaptability of the proposed methodology. While it is acknowledged that Commonwealth nations may have more




*Note:* A1 – Location, A2 – Parcel, A3 – Building built-up area, A4 – Building heated/usable area, A5 – Street type, A6 – Date of construction, A7 – Design style, A8 – Utilities, A9 – Effective building year, A10 – Number of stories, A11 – Foundation basement, A12 – Foundation basement [%], A13 – Exterior wall, A14 – Air conditioning, A15 – Bath, A16 – Floor finish, A17 – Interior finish, A18 – Heating, A19 – Grade factor.

**Figure 2.** Scope of data provided in: a) property transaction registers directly, b) other public registers

[1] The internship was supported by the Development Program at the University of Warmia and Mazury in Olsztyn (co-financed by the European Union under the European Social Fund) (POWR.03.05. 00-00-Z310/17).

transparent property registries, Raleigh's unique combination of demographic growth, economic diversity, and urban zoning complexity provides valuable insights into the challenges of property market transparency and registration in rapidly developing areas. These findings, while specific to Raleigh, provide a foundational understanding that can be adapted to other contexts with appropriate modifications, especially in regions with evolving property markets. Moreover, Raleigh's distinct mix of residential, commercial, and industrial zones and the available comprehensive data ensure that the methodology's applicability extends beyond North Carolina to other regions with similarly complex market structures.

### 3.2. Data acquisition and database creation

The identification of the case study area defined spatial extent of the data originated from public registers for real estate transactions (RET) which according to the database structural assumptions adopted by tax office was based on the (qualitative/quantitative) encoding of intrinsic variables presented with basic descriptive statistics in Appendix Table A1. The selected database structure was investigated with reference to the principles formed by RICS (Royal Institution of Chartered Surveyors, 2022) that included: recency, availability, security, privacy, ownership and ethics, provenance and lineage, assurance, consistency, collection methodology, scale and range. From the perspective of research methodology lack of uniformity in the following principles: availability and scalability occurred to determine the scope of analysis. In terms of the first principle the attributes floor finish and interior finish turned out to be unavailable (lack of data). The second principle excluded the possibility of including attribute "heating" in the analysis, since there was no differentiation in the database (all the transactions were heated with forced air). The necessity of reaching locational consistency derived from data analyzed via methodology based on location/externalities theory – HAD, required additionally taking advantage of the Open Street Map (OSM) data. Georeferencing property transactions and performing spatially based analysis focused on extrinsic variables for selecting homogenous geo-market areas enabled research execution. Spatial distribution of property transactions forming basis for further investigation has been presented in the proceeding chapter, while the extrinsic variables encoding in Appendix Table A2. In the study, a naive variable inclusion approach was adopted, where all variables were treated quantitatively without considering their qualitative context or real-world implications. This approach was chosen to maintain consistency across the dataset and ensure that the analysis remained purely data-driven, without introducing subjective adjustments.

Having analyzed the scope of available data on property transaction available in different countries (Figure 2), the further analysis was conducted on the dataset from, the selected case study area in the country of the most detailed (numerous) property transaction registers – USA. Based on the dataset the scope of information adequate for each country was subject of thorough investigation.

### 3.3. HAD methodology utilization

A comprehensive examination of the real estate market requires identifying sub-markets or particular areas where pricing dynamics occur and taking into account factors influencing property values that are presumed to be consistent. Without a clear delineation of these sub-markets, analysts and valuers risk oversimplifying complex market behaviors, so their identification plays a key role in understanding price dynamics and valuation factors reflecting property price formation components. A thorough identification process ensures that analyses are based on accurate, representative data, confirming the comparability of properties.

In real estate, homogeneity relates to areas or submarkets with uniform unit characteristics. This uniformity is crucial for conducting objective and accurate property analyses. The concept is closely linked to the aims of property market analyses for valuation purposes – a homogeneous area ensures that external factors influencing property value are consistent across the unit. This consistency prevents potential discrepancies that might occur when analyzing properties in diverse areas, that is why one can find a variety of methodological solutions utilized for that purpose – e.g. integrated clustering regression (Alenany et al., 2021), HO-MAR (Renigier-Biłozor et al., 2022), equilibrium models (Watkins, 2001), hedonic models (Watkins, 1999) or principle component analysis with cluster analysis (Keskin & Watkins, 2017).

To extract property attributes closely tied to structural and functional features (intrinsic variables) while minimizing distortion from spatial interactions and urban environmental factors, the modified methodology for homogenous area determination (HAD) was employed. This modified algorithm focuses on identifying uniform geo-market areas. Property evaluations are carried out within these homogeneous zones, which share similar locational characteristics. In line with the main objective of the methodology developed, which was to propose a solution to increase equity and fairness in property valuation procedures, a procedure based on 4 main stages, including unitization of investigated area, spatial data ETL, database model elaboration and spatial similarity model development was applied.

Methodology for identifying homogeneous areas was developed under International Association of Assessing Officers Research Grant. Details of the methodology developed were included and published in the presentation "The original methodology for homogenous area determination (HAD) for the purpose of property taxation procedures' fairness and equity increase" during GIS/Valuation Technologies Conference (Walacik & Janowski, 2024). Defining homogeneity is challenging due to the numerous of factors influencing property value. Implication of the modified HAD methodology for homogeneous area determination was therefore carried out as follows:

STAGE 1: The objective of this stage was to establish a structured framework for spatial analysis within a defined area, utilizing a regular grid of nodes with assumed parameters.

The parameters that were a subject of optimization include:

- the *A* distance between the grid nodes,
- the *B* edge length in case of hexagon and squares and *C* the radius length in case of circles,
- the computing time *H* with highest amount of 3 hours (measured for ordinary computing device).

$$E(A,B,C,H) = f_1(A,B,C,H) + f_2(A,B,C,H) + f_3(A,B,C,H) + f_4(A,B,C,H). \quad (1)$$

STAGE 2: The extraction of spatial (extrinsic) data involves obtaining information from external sources to augment or complement existing datasets. For the purposes of this study, data transformation was achieved using two entropy measurement solutions. The transformation process, grounded in the entropy function, employed the following mathematical framework.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2\left(p(x_i)\right). \quad (2)$$

STAGE 3: Several critical assumptions were made during the design of the database model, forming the basis for its overall structure and functionality: data integrity, normalization, and consistency.

STAGE 4: Cauterization with k-mean method.

## 3.4. Property valuation

Investigating the phenomenon of asymmetry in the property market resulting from unequal access to information required a property valuation to examine the discrepancies in the value of individual properties and their transaction prices. The valuation was based on databases prepared for this purpose (see Chapter 3.2). For this purpose, the author used three methods capable of solving regression problems, i.e. the Multiple Linear Regression (MLR) – the classic and well-known method, and methods from the machine learning group: Random Forest (RF) and Neural Network (NN MLP). The selection of MLR, RF, and NN MLP models was motivated by the need to compare the predictive performance of both traditional and advanced machine learning approaches. MLR provides a baseline by capturing linear relationships between property attributes and price, while RF and NN MLP allow for the modeling of more complex, nonlinear relationships often present in real estate data. RF was selected for its robustness and interpretability, and NN MLP for its ability to capture deep, multidimensional patterns in the datasets.

### 3.4.1. NN MLP model

Neural Networks (NN) can be applied in almost any situation where there is a relationship or set of relationships between the dependent and independent variables, even if these are very complex and not expressible in a classical way, through correlations or differences between groups of objects. Among the most commonly solved tasks using Neural Networks, regression tasks stand out, the aim of which is to forecast the value (usually continuous) of a specific variable, such as the price of a property. In this case, a single numerical variable is required at the output of the network. NNs, as networks based on the backward error propagation algorithm, can approximate any functional relationship between a set of independent and dependent variables (Dennis & Schnabel, 1996; Fletcher, 2000).

In general, the following stages can be identified in the data learning process:

1. Transmitting information to the input layer, through the hidden layers, to the output layer – current weight values are used when calculating the output values.
2. Calculating errors for the neurons of the output layer (by comparing the values calculated by the network with the assumed output values).
3. Modifying the weights of the output layer neurons.
4. Transmitting error information to the neurons of the previous layer (hidden) – the error information calculated for the output neurons is transmitted through the same connections as the information used to calculate the output values – only the direction of transmission is reversed. The error information is multiplied by the weight coefficients.
5. Training the neurons of the hidden layer.

### 3.4.2. RF model

Random Forest is an advanced implementation of the bagging algorithm that uses a tree model as the base model. In random forests, each tree in the ensemble is built from samples drawn with replacement (e.g., bootstrap samples) from the training set (Breiman, 2001; Hong & Kim, 2022). When splitting a node during the tree creation, the selected split is no longer the best among all predictors. Instead, the best split from a random subset of predictors is chosen. Due to this randomness, the bias of the forest usually increases slightly (compared to the bias of a non-random tree), but as a result of averaging, its variance also decreases–usually to an extent that more than compensates for the increase in bias (Ho, 1995, 1998).

In regression analysis, every tree is built from a randomly chosen subset of the training data, and the final output is determined by taking the average of all the predictions made by these trees (Buodd & Derås, 2020). When outlining the training methodology for a regression problem (Walacik & Chmielewska, 2024a, 2024b), the procedure can be described in the following manner:

1. Selecting subsets of data with replacement (bootstrap samples): suppose *D* represents the original training dataset comprising *N* feature-response pairs, where the size of the dataset is *N*:

$$D = \left\{\left(X_1,Y_1\right),\left(X_2,Y_2\right),\ldots,\left(X_3,Y_3\right)\right\}. \quad (3)$$

where: $X_1$ is the feature vector; $Y_1$ is the corresponding value of the target parameter.

The model randomizes $B$ subsets of data with returns from $D$. Let $D_b$ be the $b_{\text{th}}$ random subset of data, where $b$ = 1, 2, ..., $B$.

2. Construction of decision trees: for each $b_{\text{th}}$ subset of data $D_b$, a decision tree $h_b(X)$ is built using the classification and regression trees method;

3. Averaging of forecasts: Upon constructing $B$ decision trees, predictions for new observations are obtained by averaging the outcomes from each individual tree. The ultimate prediction for a given observation $X$ is computed as follows:

$$\hat{Y}(X) = \frac{1}{B}\sum_{b=1}^{B} h_b(X). \quad (4)$$

where: $\hat{Y}(X)$ – the forecast value of the target parameter for observation $X$; $h_b(X)$ denotes the forecast of the $b^{\text{th}}$ tree for the same observation.

#### 3.4.3. MLR model

One of the most widely utilized techniques is multiple linear regression (MLR), which is categorized under linear additive models (Meszek & Dziadosz, 2011). Key motivations for adopting these models include their straightforwardness and interpretability. By design, MLRs facilitate the examination of interconnections among variables and offer a mechanism for forecasting future values of a phenomenon. The general aim of multiple regression is to quantify the relationship between multiple independent (explanatory) variables and the dependent (criterion, explanatory) variable.

The multiple regression model is expressed by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon. \quad (5)$$

where: $y$ is the dependent variable–what you are trying to predict or explain; $x_1$, $x_2$, ..., $x_n$ are the independent variables, that are believed to influence the dependent variable; $\beta_0$ is the $y$-intercept of the regression line; it represents the predicted value of $y$ when all the independent variables are equal to zero; $\beta_1$, $\beta_2$, ..., $\beta_n$ are the regression coefficients corresponding to the independent variables. Each coefficient represents the change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other independent variables are held constant; $\varepsilon$ is the error term, which accounts for the variability in $y$ that cannot be explained by the independent variables. It is assumed to be a normally distributed random variable with a mean of zero.

## 4. Results

### 4.1. HAD methodology utilization

To extract property attributes closely linked to structural and functional features (intrinsic variables) and reduce distortion caused by spatial interactions and urban environmental factors, an adapted HAD methodology was utilized After preparing the assumed variables, the modified HAD algorithm was utilized to select homogenous geo-market areas. The application of these procedures resulted in the identification of 9 homogenous geo-market areas. The extracted homogenous geo-market areas, primarily composed of units belonging to the same group, were transformed into continuous areas using a grouping function. One of the selected geo-market areas (Figure 3) was used to scope property transaction analysis and apply the random NN MLP, MLR and RF models.

The study area was located in the northern part of the Raleigh city. During the study period (2021–2023), 256 residential and freehold secondary market property transactions took place there. Given the methodologies employed in this study, such as Multiple Linear Regression (MLR), Random Forest (RF), and Neural Network (NN MLP) models, it was essential to make an assumption that

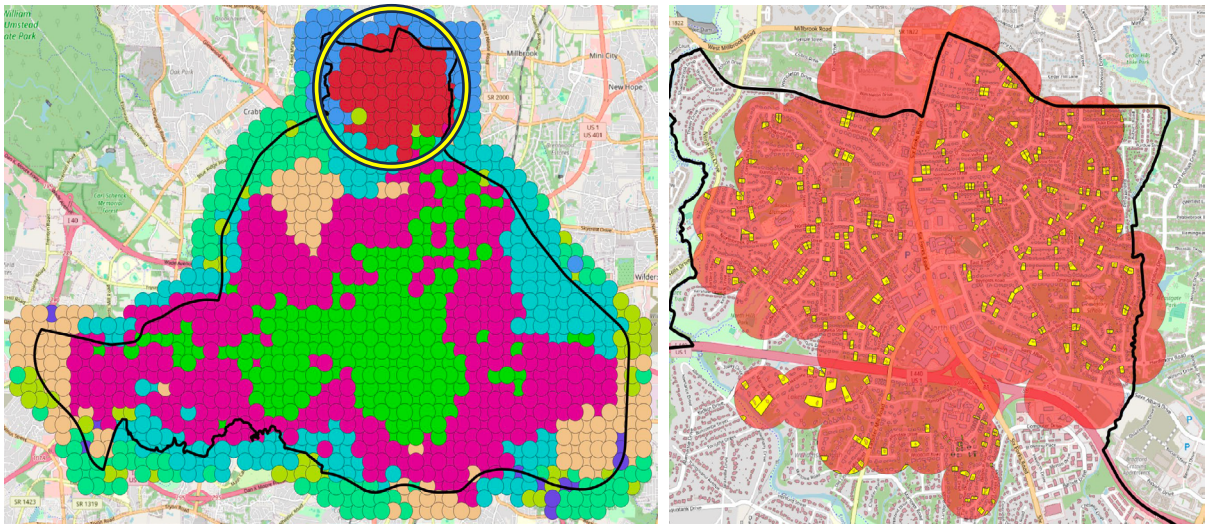


**Figure 3.** The selected geo – market area (source: own elaboration and Walacik and Janowski, 2024)

the data concerning property transactions had to be complete and fully descriptive. From the selected set of 256 property transactions, the author selected 244 that were fully described in terms of the assumed attributes. It was noted that an emerging problem was the completeness of the data, so 12 transactions had to be removed from the central register.

## 4.2. Property valuation

### 4.2.1. NN MLP model

The Neural Network model (NN MLP) was trained using a random sampling method to divide the dataset into training, testing, and validation subgroups (70%, 15%, 15%). Optimal network parameters, such as the number of neurons in the hidden layers and the loss function, were selected to balance model complexity with predictive performance. These parameters were chosen to maximize prediction accuracy while avoiding overfitting. The selected network structure (e.g., 19-14-1 for DB – USA) and backpropagation learning algorithm demonstrated high generalization capability, especially for markets with better data transparency, such as the USA dataset. In the analysis, a feedforward multilayer perceptron (MLP) was used with the following structure: input layer – hidden layer – output layer. As a rule, the choice of an appropriate loss function depends on the type of problem being solved and the nature of the output variable. In the problem analysed, optimum results were achieved using, for both databases, the Sum of Squares (SOS) loss function. The backpropagation learning algorithm was used, in which weight modification occurs after the presentation of each element of the training set (rather than cumulatively after the presentation of all elements comprising the training set). The loss metrics charts were presented in Figure 4.

The obtained optimal model's metrics and architectures were presented in Table 2.

In DB – USA, both training and test error rates decrease sharply at first and then level off, indicating rapid improvement and good generalization without overfitting. In DB – PL, the close convergence of training and test errors suggests good generalization and minimal overfitting. In Turkey, both errors converge to similar values, indicating

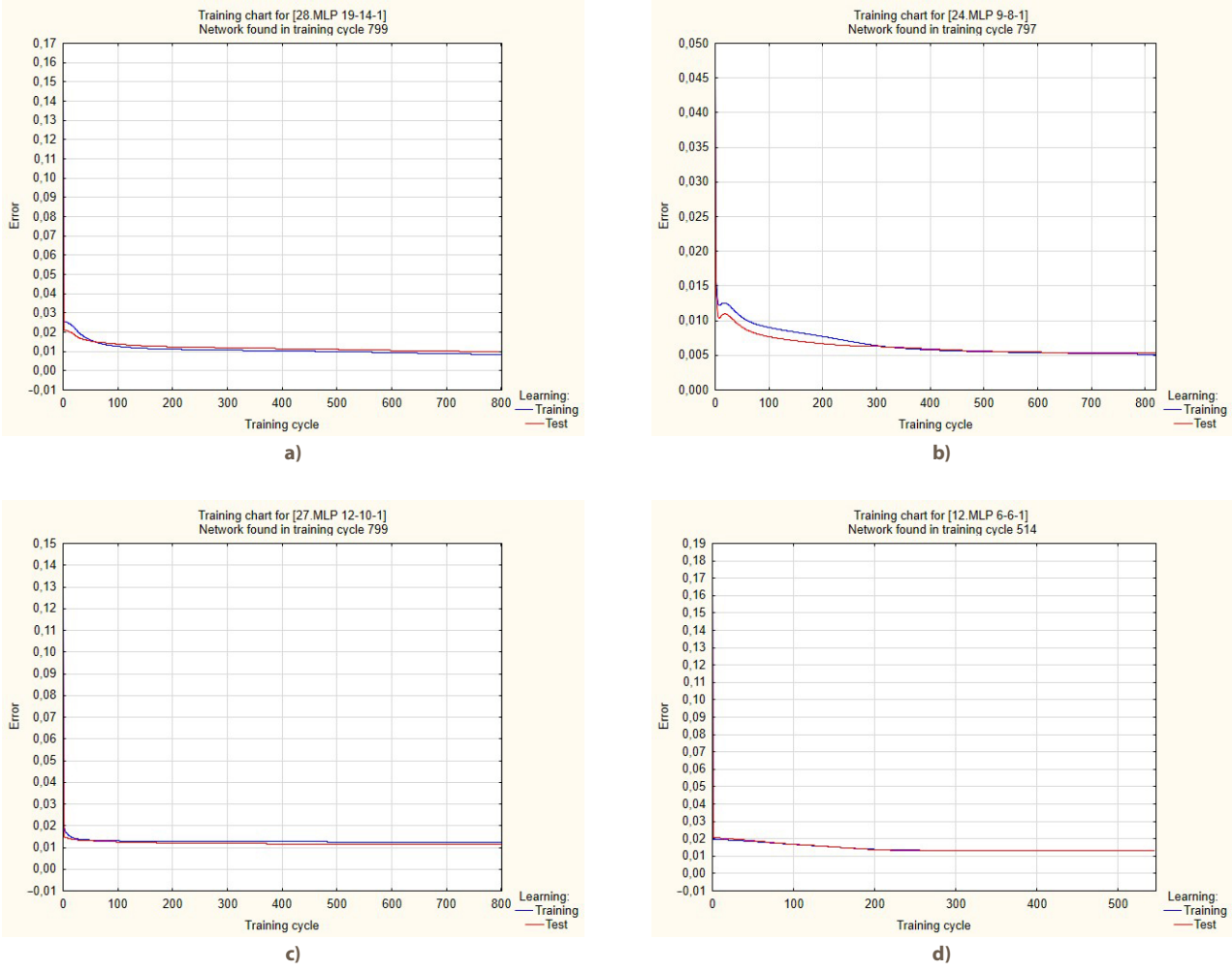



**Figure 4.** Loss metrics chart of the model (training and test sets) versus the number of training epochs for: a) DB – USA, b) DB – PL, c) DB – TUR, and d) DB – UK (source: own elaboration with the use of Statistica software)

**Table 2.** Neural Network Architecture and metrics for DB – USA/PL/TUR/UK (source: own elaboration using Statistica software)

| | NN | Quality (training) | Quality (test) | Quality (validation) | Learning algorithm | Loss function |
|---|---|---|---|---|---|---|
| DB – USA | MLP 19-14-1 | 0.6581 | 0.6019 | 0.5245 | Backprop 799 | SOS |
| DB – PL | MLP 9-8-1 | 0.7633 | 0.7595 | 0.7567 | Backprop 797 | SOS |
| DB – TUR | MLP 12-10-1 | 0.5567 | 0.5567 | 0.4681 | Backprop 799 | SOS |
| DB – UK | MLP 6-6-1 | 0.4545 | 0.4413 | 0.4203 | Backprop 514 | SOS |

effective generalization and avoidance of overfitting. Similarly, in DB – UK, training and test errors converge and run parallel, showing good generalization and effective avoidance of overfitting.

### 4.2.2. RF model

The Random Forest model used parameters optimized through an iterative process, such as the number of trees, maximum depth, and minimum samples split. These parameters were chosen to balance model complexity and prevent overfitting. For example, using 100 trees in the USA dataset improved prediction accuracy while maintaining a manageable model size. The chosen parameters minimized the bias-variance trade-off, with the number of trees ensuring stable results and the depth controlling the complexity. This resulted in better generalization of the model across all datasets, particularly in handling data asymmetry, as seen in the USA dataset's lower error rates.

The summary of database processing using RF is presented in Figure 5.

The training graphs for the Random Forest model illustrate its performance in terms of mean square error as the number of trees increases. Initially, the training error for DB – USA drops sharply and stabilizes around 20 trees, indicating good learning and generalization without overfitting. For DB – PL, the training error decreases significantly

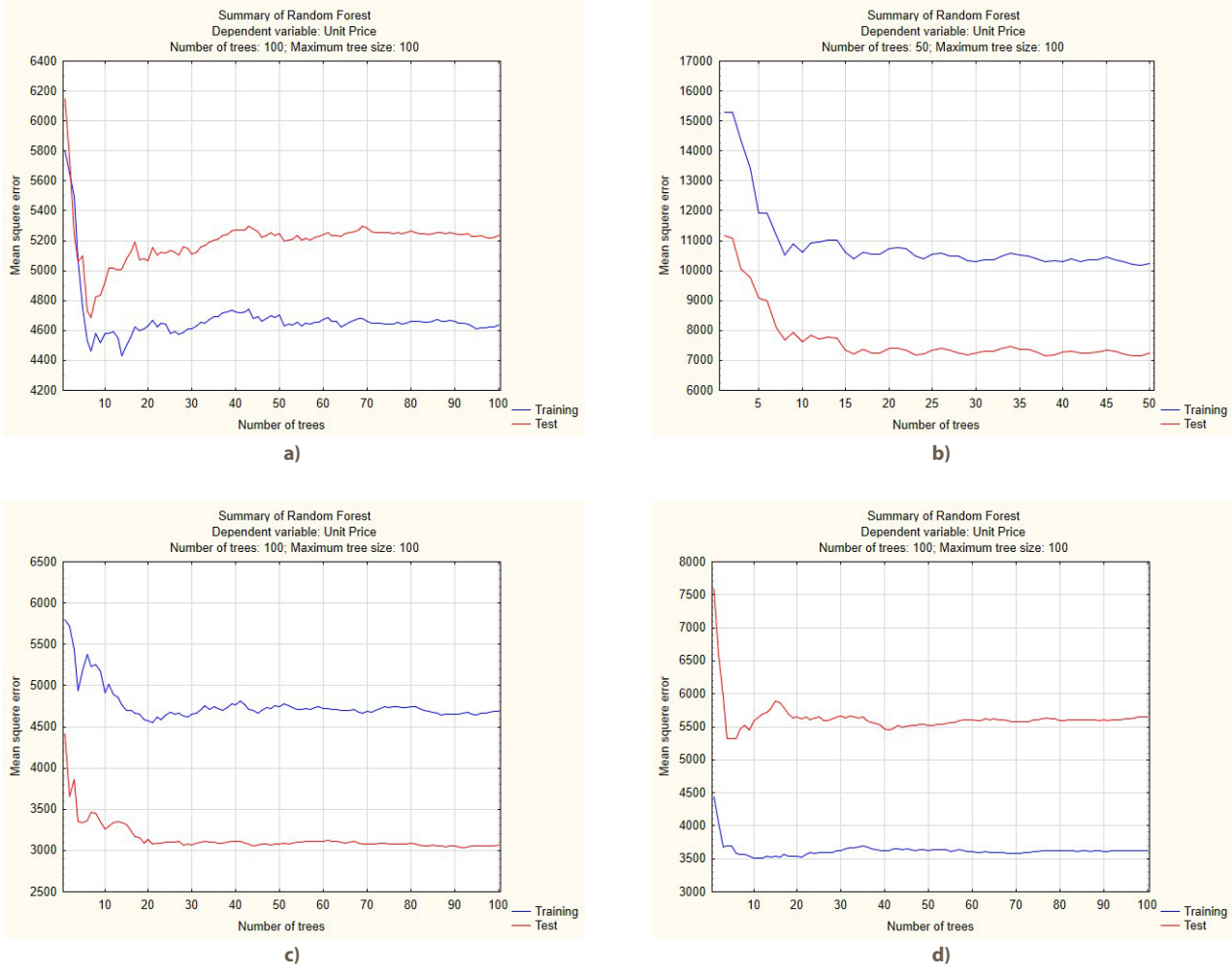



**Figure 5.** The training graph of the random forest model for: a) DB – USA, b) DB – PL, c) DB – TUR, and d) DB – UK (source: own elaboration with the use of Statistica software)

**Table 3.** Random Forest risk assessment results (source: own elaboration using Statistica software)

| RF | DB – PL | | DB – USA | | DB – TUR | | DB – UK | |
|---|---|---|---|---|---|---|---|---|
| | Risk assessment | Standard error | Risk assessment | Standard error | Risk assessment | Standard error | Risk assessment | Standard error |
| Training | 10243.45 | 1947.05 | 4636.35 | 837.81 | 4691.22 | 792.10 | 3629.82 | 479.02 |
| Test | 7236.80 | 1726.43 | 5231.88 | 814.53 | 3068.84 | 464.47 | 5646.47 | 1567.62 |

and stabilizes around 50 trees, suggesting effective learning and balanced performance. In the DB – TUR and DB – UK models, the training error drops rapidly before stabilizing, with the test error showing similar trends, indicating rapid pattern capture and effective generalization. The obtained risk assessments for both training and test datasets, along with their standard errors, are presented in Table 3.

The Random Forest risk assessment results presented in Table 2 show varying levels of prediction accuracy and error across different datasets (DB – PL, DB – USA, DB – TUR, DB – UK). For the training datasets, the risk assessment is lowest for DB – USA (4636.35) and highest for DB – PL (10243.45), indicating better model performance and lower prediction error for the USA dataset. The test datasets reveal a similar trend, with the lowest risk assessment again for DB – USA (5231.88) and the highest for DB – UK (5646.47). Standard errors are also lower for the USA dataset, both in training (837.81) and testing (814.53), suggesting more reliable and stable predictions compared to other countries. Overall, the Random Forest model performs best on the USA dataset, demonstrating the lowest prediction risk and error, while the performance is less optimal for Poland and the UK.

#### 4.2.3. MLR model

The metrics of the multiple regression analysis for each database are presented in Table 4.

For the DB – USA and DB – PL the *F*-statistic is significantly large and the *p*-value is less than 0.0001, indicating that the overall model is statistically significant. The moderate *R*-squared value (0.36) suggests that while the model explains some variability in the dependent variable, there is still unexplained variability. The MLR models for the Turkey and UK datasets show a moderate explanatory ($R^2$ of 0.2933 and 0.2072).

MLR allow for the study of interrelationships between factors and provide a tool for predicting the future values

**Table 4.** MLR metrics (source: own elaboration using Statistica software)

| *N* = 244 | DB – USA | | DB – PL | | DB –TUR | | DB – UK | |
|---|---|---|---|---|---|---|---|---|
| | b | p | b | p | b | p | b | p |
| intercept | –1726.49 | 0.1573 | –3834.49 | 0.0000 | –4435.46 | 0.0000 | 339.54 | 0.0000 |
| parcel | 0.00 | 0.0000 | 0.00 | 0.0000 | 0.00 | 0.0000 | 0.00 | 0.0000 |
| heated area | –0.05 | 0.0000 | x | x | –0.03 | 0.0004 | –0.04 | 0.0000 |
| built-up area | x | x | –0.07 | 0.0000 | x | x | x | x |
| street type | 0.00 | 0.9996 | –1.84 | 0.5725 | 0.19 | 0.9398 | –0.32 | 0.9016 |
| year built | –0.04 | 0.9275 | 2.23 | 0.0000 | 1.01 | 0.0099 | x | x |
| design style | –14.79 | 0.0066 | –10.01 | 0.0447 | –13.08 | 0.0148 | x | x |
| utilities | 2.24 | 0.9408 | –15.32 | 0.6963 | –0.27 | 0.9928 | 16.65 | 0.5967 |
| effective year | 1.11 | 0.0232 | x | x | 1.48 | 0.0000 | x | x |
| remodeled year | 0.00 | 0.5488 | x | x | x | x | x | x |
| story height | –5.21 | 0.1743 | 9.58 | 0.0525 | –5.43 | 0.1653 | x | x |
| foundation basement | 10.40 | 0.0977 | –9.91 | 0.2027 | 13.38 | 0.0348 | x | x |
| foundation basement percent | –0.03 | 0.9162 | x | x | –0.08 | 0.7807 | x | x |
| exterior wall | –1.67 | 0.7724 | x | x | 0.00 | 0.9993 | x | x |
| air | –9.44 | 0.7793 | x | x | x | x | x | x |
| bath | –7.30 | 0.1972 | x | x | –7.89 | 0.1533 | x | x |
| bath fixtures | –1.73 | 0.5069 | x | x | x | x | x | x |
| built ins | 5.41 | 0.5719 | x | x | x | x | x | x |
| grade | –1.28 | 0.1589 | –3.33 | 0.0022 | | | –2.31 | 0.0063 |
| assessed grade difference | 1.17 | 0.0000 | x | x | x | x | 1.20 | 0.0000 |
| accrued assessed condition | 0.04 | 0.9138 | x | x | x | x | x | x |

of a phenomenon. In the practice of real estate valuation, the use of MLRs is hampered by the method's assumptions about the linear nature of the data. Moreover, MLR tends to encounter issues such as overfitting when dealing with numerous qualitative variables, leading to a higher number of estimated parameters. These findings indicate that multiple regression, i.e. due to the assumption of a linear relationship between the independent variables and the dependent variable, may not correspond to the specifics of the real estate market and consequently distort valuation results.

## 5. Models' validation – information asymmetry analysis

When forecasting economic phenomena, which are, among other things, the result of decisions made by participants in the real estate market, prediction error is inevitable. Conducting a comparative analysis of the results obtained using selected forecasting models can allow for the assessment and selection of optimal solutions for a given problem–both in terms of choosing the analytical tool and illustrating the impact of the richness of the applied database on the quality of the prediction. In regression analysis and forecasting, accuracy metrics are crucial for evaluating the performance of predictive models. Scalar accuracy metrics quantify the average agreement between individual pairs of predictions and actual observations (Morley et al., 2018; Murphy, 1993). Quantitative assessment of modeling and forecasting of continuous quantities uses a variety of approaches. To determine the impact of the richness of real estate databases on the asymmetry of the real estate market, several classic criteria for evaluating prediction results were applied. The following measures of prediction accuracy were used for this purpose:

- The Mean Error (ME) measures the average of the errors in a set of predictions, without considering the direction of the errors (positive or negative). It is the sum of the residuals (predicted value minus actual value) divided by the number of observations (Equation (6)):

$$ME = \left( \frac{1}{n} \sum_{i=1}^{n} A_i - P_i \right), \tag{6}$$

where: $A_i$ – property price; $P_i$ – predicted property value; $n$ – number of properties.

- Mean Squared Error (MSE) is a measure of errors between paired observations expressing the same phenomenon. Comparing two paired sets of data, MAE is the average vertical distance between each point and the identity line. MAE is a linear score which means that all the individual differences are weighted equally in the average (Equation (7)):

$$MAE = \left( \frac{1}{n} \sum_{i=1}^{n} \left| A_i - P_i \right| \right). \tag{7}$$

- Mean Absolute Percentage Error (MAPE) measures the average magnitude of the errors in percentage terms. It is calculated as the average of the absolute values of the errors divided by the actual values, multiplied by 100 to express it as a percentage (Equation (8)):

$$MAPE = \left( \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - P_i}{A_i} \right| \right) \cdot 100\%. \tag{8}$$

- The Mean Squared Error (MSE) calculates the average of the squares of the errors. MSE is more sensitive to larger errors due to squaring each term, which penalizes larger errors more than smaller ones (Equation (9)):

$$MSE = \left( \frac{1}{n} \sum_{i=1}^{n} \left( A_i - P_i \right)^2 \right). \tag{9}$$

- Root Mean Squared Error (RMSE) is the square root of the mean squared error, which adjusts the scale of the errors to be compatible with the scale of the targets (Equation (10)):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( A_i - P_i \right)^2}. \tag{10}$$

- Root Mean Squared Percentage Error (RMSPE) is similar to RMSE but normalized to the scale of the actual values, expressed in percentage terms. It provides an estimation of the error size relative to the actual value (Equation (11)):

$$RMSPE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \frac{A_i - P_i}{A_i} \cdot 100\% \right)^2}. \tag{11}$$

- Coefficient of Dispersion (COD) is a statistical measure that indicate the uniformity or variability of individual property assessments relative to the median assessment ratio. It reflects how much individual property assessments deviate from the median of all assessed properties (Equation (12)):

$$COD = \frac{\sum_{i=1}^{n} \left| \frac{A_i}{P_i} - \frac{A_{\left(\frac{n+1}{2}\right)}}{P_{\left(\frac{n+1}{2}\right)}} \right|}{n \cdot \frac{A_{\left(\frac{n+1}{2}\right)}}{P_{\left(\frac{n+1}{2}\right)}}} \cdot 100. \tag{12}$$

- Price-Related Differential (PRD) provides a simple gauge of price-related bias. It indicates whether lower- or higher-valued properties are systematically over- or under-assessed compared to each other (Equation (13)):

$$PRD = \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{P_i}}{\frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} P_i}}. \tag{13}$$

**Table 5.** Accuracy metrics (source: own elaboration)

| Accuracy metrics | NN MLP | | | | MLR | | | | RF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PL | USA | TUR | UK | PL | USA | TUR | UK | PL | USA | TUR | UK |
| ME | –20.63 | 6.87 | –15.74 | –3.48 | 0.00 | 0.00 | 0.00 | 0.00 | –5.94 | –3.36 | 1.49 | 2.48 |
| MAE | 62.79 | 43.63 | 63.13 | 59.80 | 62.56 | 45.52 | 47.64 | 52.33 | 72.35 | 52.29 | 47.98 | 47.36 |
| RMSE | 80.03 | 60.18 | 85.54 | 79.82 | 84.52 | 61.60 | 64.62 | 68.44 | 96.41 | 69.42 | 64.80 | 65.32 |
| MAPE | 18.09% | 10.37% | 15.84% | 14.71% | 17.89% | 11.02% | 11.58% | 12.77% | 22.34% | 12.90% | 11.67% | 11.56% |
| RMSPE | 23.39% | 14.49% | 23.16% | 20.90% | 24.21% | 15.53% | 16.34% | 17.62% | 31.62% | 18.20% | 16.49% | 16.70% |
| MSE | 6405.39 | 3621.39 | 7316.52 | 6372.00 | 7143.47 | 3794.06 | 4175.50 | 4684.48 | 9294.63 | 4819.40 | 4199.18 | 4266.22 |
| COD | 1.04 | 1.02 | 1.03 | 1.03 | 1.04 | 1.02 | 1.02 | 1.02 | 1.08 | 1.03 | 1.03 | 1.03 |
| PRD | 15.24% | 10.41% | 15.11% | 14.46% | 17.40% | 10.99% | 11.50% | 12.86% | 21.17% | 12.57% | 11.59% | 11.54% |

Value determination prediction accuracy metrics for all databases were used to identify and assess discrepancies between property values and their transaction prices. The Table 5 presents the results of the adopted analysis of accuracy metrics four databases based on the results of the processing of the Neural Network method.

The DB – USA shows a slight overestimation, whereas the PL, TUR, and UK datasets show underestimation, with the PL dataset having the highest bias. The USA dataset has the lowest MAE, indicating the most accurate predictions, while the TUR dataset has the highest MAE, indicating less accurate predictions. The results for DB – PL and DB – UK are slightly lower than for DB – TUR. A similar trend is found for MAPE, RMSPE, and MSE values. These results provide clear evidence in support of RQ2, demonstrating that the RF model adapts well to varying levels of information symmetry. The superior performance in the USA dataset, characterized by the lowest error metrics, suggests that advanced algorithms like RF are highly effective in environments with comprehensive property registration systems. In contrast, the poorer performance in the TUR dataset highlights the challenges faced when applying these models to less transparent property markets. This underscores the theoretical framework that links market transparency to reduced information asymmetry (Kurlat & Stroebel, 2014). Additionally, the USA dataset has a relatively low COD value of 102, reflecting uniformity in the property assessments, whereas the PRD value of 104.1% indicates minimal price-related bias.

In the case of MLR analysis, all datasets have an ME of 0.00, indicating no systematic bias in the predictions for any dataset. The highest MAE, RMSE, MAPE, RMSPE, and MSE values occurred for DB – PL, while the lowest values occurred for DB – USA. DB – TUR and DB – UK had values similar to the USA, but slightly higher, indicating their comparable but still slightly lower ability to predict prices. For the COD metric, all datasets exhibit relatively low values, reflecting a consistent prediction spread, with PRD values indicating minimal bias, except for DB – PL, which shows the highest PRD (174.0%), suggesting potential regressivity in the predictions.

For the PL and USA datasets, the RF model shows slight underestimation, whereas for TUR and UK, it shows slight overestimation. The TUR and UK datasets have the lowest MAE, indicating better performance and smaller errors in property value predictions compared to other databases. The results for DB – USA are only slightly larger than TUR and UK, and can therefore be considered comparable. The results for RMSE, MAPE, RMSPE, and MSE metrics follow a similar trend. The TUR and UK datasets exhibit the best performance across most metrics (MAE, RMSE, MAPE, RMSPE, MSE), suggesting that the RF model's predictions are more accurate for these datasets. Additionally, the COD for TUR and UK are low (103), reflecting uniformity, while the PRD values for TUR (115.9%) and UK (115.4%) suggest slight under-assessment of lower-valued properties.

Each metric provides a different lens to assess the performance of the regression model. The metrics indicated above are used to monitor prediction accuracy. In order to explore the problem deeply, the author decided to use additional differentiation criteria:

- PREDICTION ACCURACY: The Price – to – Value (PTV) metric indicates the average ratio of the actual transaction prices to the estimated property values. It is one of the best-known indicators of financial analysis. It is often used to roughly assess the desirability of investing in financial stocks. The implication of the PTV indicator in the problem under analysis provides knowledge of the extent to which the value of the property "matches" the transaction price. PTV close to 1 suggests that the model's estimated values are close to the actual market prices (Equation (14)).

$$PTV = \frac{A_i}{P_i}. \tag{14}$$

- VARIABILITY: The Coefficient of Price Variability (CPV) measures the dispersion of the price-to-value ratios. Lower CPV indicates less variability and more consistent performance of the model (Equation (15)):

$$CPV = \frac{St.Dev.\ (A_i - P_i)}{\frac{1}{n}\sum_{i=1}^{n} A_i - P_i} \cdot 100\%. \tag{15}$$

- EQUALITY: The Gini Index measures inequality in the distribution of PTV ratios. A lower Gini Index indicates more equitable performance. The Gini coefficient is a coefficient of inequality, usually used in the context

of income. The Gini Index measures inequality in the distribution of PTV ratios. A lower Gini Index indicates more equitable performance. For the problem under consideration, the equation for the Gini coefficient has the following form (Equation (16)):

$$G = \frac{n+1-2\left(\frac{\sum_{i=1}^{n}(n-i+1)(A_i-P_i)}{\sum_{i=1}^{n}(A_i-P_i)}\right)}{n}. \quad (16)$$

- QUARTILE ANALYSIS: Interquartile Range (IQR) is a measure of data dispersion that describes the range of the middle 50% of values in a data set. IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). It is particularly useful for identifying outliers and understanding variability in the data, as it is not sensitive to extreme values. First quartile (denoted as Q1 or Q25) – 25% of observations are located below this value, while 75% are above it. The first quartile divides the observations in a 25% to 75% ratio, meaning that 25% of observations are lower than or equal to the value of the first quartile, and 75% of observations are equal to or greater than the value of the first quartile. Third quartile (Q3 or Q75) – three-quarters of the observations are located below this value, while one-quarter are above it. The third quartile divides the observations in a 75% to 25% ratio, meaning that 75% of observations are lower than or equal to the value of the third quartile, and 25% of observations are equal to or greater than the value of the third quartile.

$$IQR = Q3 - Q1. \quad (17)$$

The results for the measures adopted are shown in the Table 6.

The Table 6 presents asymmetry metrics for NN MLP, MLR, and RF models across four countries. For NN MLP, the USA shows the highest Mean PTV (1.0165) and lowest CPV (0.1409), indicating high prediction accuracy and low variability. Poland has the lowest Mean PTV (0.9484) and highest IQR (0.2265), indicating wider data spread and lower prediction accuracy. For MLR, the USA, Turkey, and the UK show nearly perfect Mean PTV values (0.9999) with low variability, while Poland shows higher variability and slightly lower prediction accuracy. For RF, Turkey and the UK have the highest Mean PTV values, with Turkey showing the lowest variability. Overall, the USA consistently exhibits high prediction accuracy and low variability across models, while Poland shows higher variability and lower accuracy.

In order to analyze and interpret the relationship between real estate transaction prices and their values obtained using different predictive methods, classic scatter plots were developed for each of the analyzed databases. The introduced graphical solutions allow for examining both the relationship between two variables and the distribution of each of these variables. Data visualization can help by delivering data in the most efficient way possible. The analysis of scatter plots and regression functions (Figure 6) indicated the predictive capabilities of individual methods in light of the richness of the databases.

The main difference between the scatter plot for Poland and the plots for the USA, Turkey, and the UK arises from the use of a different reference unit, which in Poland was the building area (while in the other analyzed countries, the measure of usable area – unavailable in public records in Poland – was used). As a result, Poland exhibits the greatest price dispersion compared to other countries, a relatively low level of prediction fit, and clear bimodality as indicated by the density plot. The plots presenting results for the USA, Turkey, and the UK show greater comparability due to the use of heated area as the reference unit for transaction prices. The highest fit is shown by the results of the USA data set analysis, due to the lowest dispersion of values relative to real estate transaction prices. Neural Networks (NN) and Multiple Linear Regression (MLR) exhibit a similar level of model fit to the data. Lower variability and higher predictability in the USA real estate market suggest better market transparency and more reliable property valuations compared to Turkey, the United Kingdom, and Poland. These results demonstrate clear differences in model performance across datasets with varying levels of transparency. The following section explores how these findings relate to the theoretical framework of information asymmetry and property market dynamics, answering RQ2 and RQ3.

**Table 6.** Asymmetry metrics – additional differentiation criteria (source: own elaboration)

| Assymetry metrics | NN MLP | | | | MLR | | | | RF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PL | USA | TUR | UK | PL | USA | TUR | UK | PL | USA | TUR | UK |
| Mean PTV | 0.9484 | 1.0165 | 0.9723 | 0.9978 | 0.9791 | 0.9999 | 0.9999 | 0.9999 | 0.9714 | 0.9898 | 1.0003 | 1.0030 |
| CPV | 0.1859 | 0.1409 | 0.1881 | 0.1857 | 0.3466 | 0.1419 | 0.1496 | 0.1585 | 0.2518 | 0.1596 | 0.1502 | 0.1512 |
| Gini Index | 0.1089 | 0.0751 | 0.1047 | 0.1011 | 0.1366 | 0.0776 | 0.0814 | 0.0876 | 0.1404 | 0.0877 | 0.0815 | 0.0815 |
| Q1 | 0.8242 | 0.9444 | 0.8577 | 0.8801 | 0.8550 | 0.9157 | 0.9061 | 0.9044 | 0.7992 | 0.8896 | 0.9078 | 0.9161 |
| Q2 | 0.9175 | 1.0083 | 0.9719 | 0.9855 | 0.9755 | 0.9976 | 0.9940 | 0.9905 | 0.9610 | 0.9819 | 0.9941 | 0.9987 |
| Q3 | 1.0507 | 1.0912 | 1.0655 | 1.1028 | 1.1187 | 1.0807 | 1.0814 | 1.0968 | 1.1130 | 1.0842 | 1.0842 | 1.0813 |
| IQR | 0.2265 | 0.1468 | 0.2078 | 0.2227 | 0.2637 | 0.1651 | 0.1753 | 0.1924 | 0.3138 | 0.1947 | 0.1763 | 0.1653 |

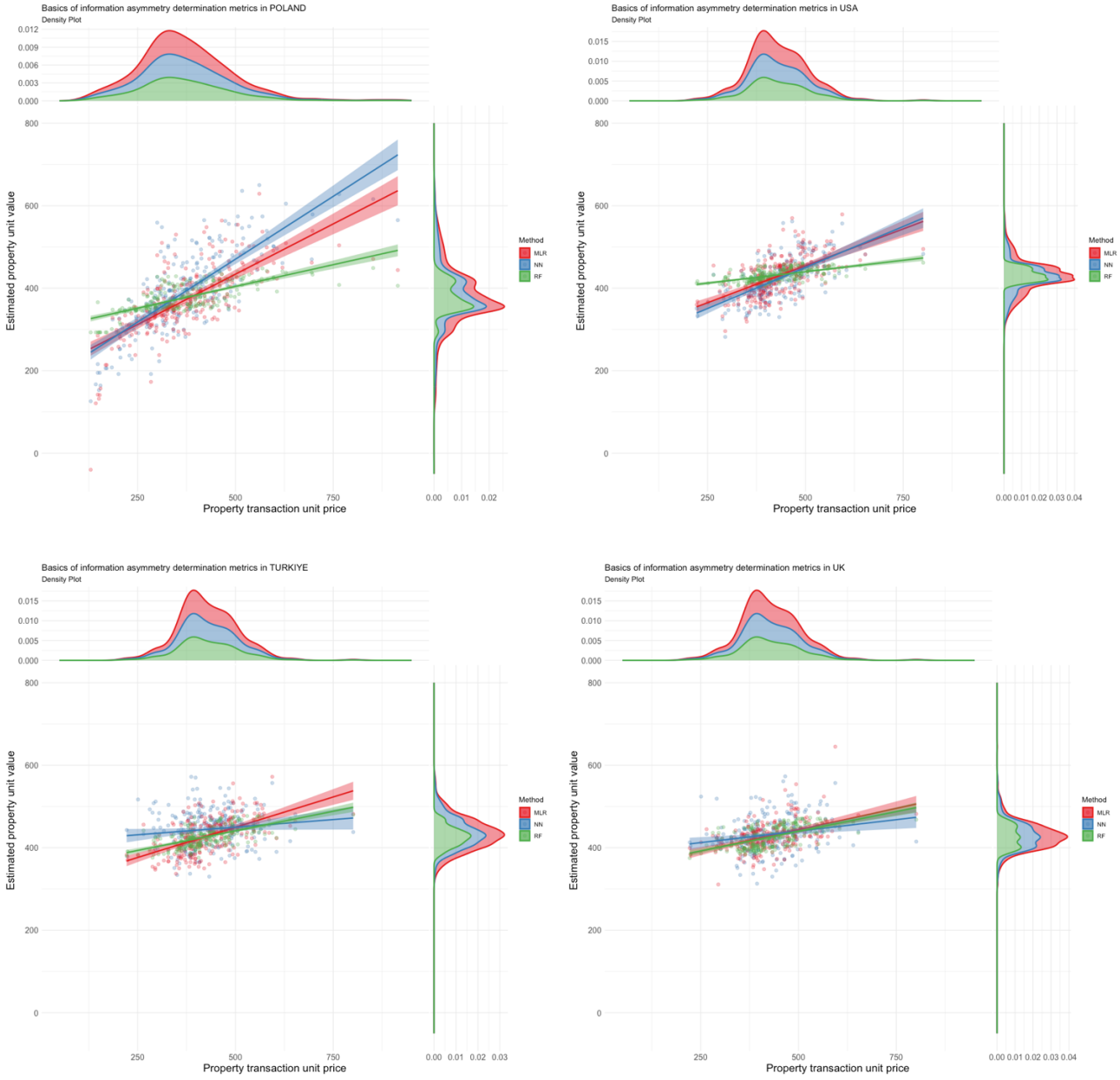



**Figure 6.** Real estate price and value scatter chart (source: own elaboration)

## 6. Conclusions and discussion

This research provides substantial insights into the effects of information asymmetry in real estate markets, particularly from the perspective of property registration abundance and appraisal theory and practice. The study's findings emphasize that increased transparency and improved data quality can significantly reduce information asymmetry and strengthen the market equitability (answer to RQ1). This aligns with the theoretical frameworks of the highest and best use paradigm and externalities theory, supporting the importance of comprehensive, accurate data in real estate valuation. For that reason the paper substantially contributes in several ways:

- by focusing on property registration systems, the study improves understanding of how information asymmetry affects market fairness and efficiency,
- it effectively integrates various classical theories and paradigms, such as location theory and the sustainable development paradigm, to analyze the complexities of real estate markets,
- it utilizes advanced analytical techniques for property valuation and provides a modern approach to addressing the challenges posed by information asymmetry.

This research advances the theoretical discourse surrounding information asymmetry in property markets by providing empirical evidence that machine learning

models can significantly improve valuation accuracy, even under conditions of incomplete or asymmetric data. The superior performance of the Random Forest algorithm, particularly in the U.S. dataset, emphasizes the critical role that transparent and complete property registration plays in market efficiency. In contrast, the challenges faced in less transparent markets, such as those in Poland and Turkey, highlight the limitations of current property data systems. These findings contribute to RQ2 and RQ3 by illustrating how advanced algorithms can bridge some of the gaps caused by information asymmetry but also underscore the need for more robust data governance frameworks to fully realize the potential of these tools. This study thus positions itself as a key piece of evidence in the ongoing debate about the role of technology and data transparency in modernizing property markets. The practical implications of the research are critical for policymakers and investors in the real estate sector. The study supports policies promoting information uniformity, which could lead to more informed decision-making in the real estate markets by increasing their comparability. Real estate professionals, including investors and developers, can use the insights from the study to improve their strategies mitigating poor decision making, economic inefficiencies, reduced trust in market and inequitable outcomes – potential consequences that typical citizens might face (answer to RQ3).

While the research provides valuable insights, it has several limitations. The primary data was collected from selected region, which may not represent other geographic contexts with different real estate dynamics. Some attributes like heating in property registrations were uniformly recorded, which could distort the analysis and limit the applicability of the findings across different settings.

To build on the findings of this study, further research could expand geographical scope. Future studies could include diverse geographic areas to validate the findings across different real estate markets. Investigating other variables, such as environmental impact factors or different types of property data, could provide deeper insights into the valuation processes. Long-term studies could examine the effects of policy changes and technological advancements on information symmetry in real estate markets. From the perspective of previous author's studies within the scientific problem of real estate market delineation, the research indicated the need of defining a structured pattern or procedure tailored for the purpose of property valuation. Additionally the utilization of HAD methodology indicated further elements of the solution improvement especially, with reference to the methodological transparency and possibility of the results substantial verification. In conclusion, the study's results provide strong evidence supporting RQ1 and RQ2 by showing that property registration abundance and data transparency play a critical role in reducing information asymmetry. The superior performance of advanced models such as RF in the USA dataset illustrates the potential of these algorithms to improve property valuation accuracy in transparent markets. These findings also emphasize the importance of addressing data asymmetry, as seen in the weaker results for PL and TUR, which highlight the need for policy interventions aimed at improving data completeness and quality in less transparent markets.

## Funding

This research was funded by the Development Program at the University of Warmia and Mazury in Olsztyn (internship in the School of Government at the University of North Carolina, USA, co-financed by the European Union under the European Social Fund) (POWR. 03.05. 00-00-Z310/17).

## Disclosure statement

I declare that I do not have any competing financial, professional, or personal interests from other.

## References

Aizenman, J., & Jinjarak, Y. (2009). Current account patterns and national real estate markets. *Journal of Urban Economics*, *66*(2), 75–89. https://doi.org/10.1016/j.jue.2009.05.002

Alenany, E., Lekham, L. A., & Lu, S. (2021). Integrated clustering regression for real estate valuation. *Real Estate Finance*, 1–36. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3835967

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

Ambrose, B. W., & Diop, M. (2021). Information asymmetry, regulations and equilibrium outcomes: Theory and evidence from the housing rental market. *Real Estate Economics*, *49*(S1), 74–110. https://doi.org/10.1111/1540-6229.12262

Ambrose, B. W., & Shen, L. (2023). Past experiences and investment decisions: Evidence from real estate markets. *The Journal of Real Estate Finance and Economics*, *66*(2), 300–326. https://doi.org/10.1007/s11146-021-09844-2

Batabyal, A. A. (2023). The theory of externalities. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4643096

Ben-Shahar, D., & Golan, R. (2019). Improved information shock and price dispersion: A natural experiment in the housing market. *Journal of Urban Economics*, *112*, 70–84. https://doi.org/10.1016/J.JUE.2019.05.008

Bergh, D. D., Ketchen, D. J., Orlandi, I., Heugens, P. P. M. A. R., & Boyd, B. K. (2019). Information asymmetry in management research: Past accomplishments and future opportunities. *Journal of Management*, *45*(1), 122–158. https://doi.org/10.1177/0149206318798026

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brzezicka, J., Łaszek, J., Olszewski, K., & Wisniewski, R. (2022). The missing asymmetry in the Polish house price cycle: An analysis of the behaviour of house prices in 17 major cities. *Journal of Housing and the Built Environment*, *37*(2), 1029–1056. https://doi.org/10.1007/s10901-021-09861-w

Buodd, M. F., & Derås, E. J. (2020). *Machine learning for property valuation: An empirical study of how property price predictions can improve property tax estimations in Norway* [Master thesis, Norwegian School of Economics]. NHH Brage.

Campbell, S. (2018). Green cities, growing cities, just cities? Urban planning and the contradictions of sustainable development. In J. Stein (Ed.), *Classic readings in urban planning* (pp. 308–326). Routledge. https://doi.org/10.4324/9781351179522-25

Chau, K. W., Wong, S. K., & Yiu, C. Y. (2007). Housing quality in the forward contracts market. *Journal of Real Estate Finance and Economics*, *34*(3), 313–325. https://doi.org/10.1007/s11146-007-9018-x

Chau, K. W., & Wong, S. K. (2016). Information asymmetry and the rent and vacancy rate dynamics in the office market. *The Journal of Real Estate Finance and Economics*, *53*, 162–183. https://doi.org/10.1007/s11146-015-9510-7

Chinloy, P., Hardin III, W., & Wu, Z. (2013). Price, place, people, and local experience. *Journal of Real Estate Research*, *35*(4), 477–506. https://doi.org/10.1080/10835547.2013.12091376

Cornes, R., & Sandler, T. (1996). *The theory of externalities, public goods, and club goods*. Cambridge University Press. https://doi.org/10.1017/cbo9781139174312

d'Amato, M., & Kauko, T. (2017). Appraisal methods and the non-agency mortgage crisis. In M. d'Amato & T. Kauko (Eds.), *Advances in automated valuation modeling: AVM after the non-agency mortgage crisis* (pp. 23–32). Springer International Publishing. https://doi.org/10.1007/978-3-319-49746-4_2

Danastri Yuwono, A., Khairunnisa, A., Ikasanti, M., Adilah, N., & Widiyani, W. (2023). Pendekatan highest and best use dalam pelestarian bangunan bersejarah. *Arsitektura: Jurnal Ilmiah Arsitektur dan Lingkungan Binaan*, *21*(2), 247–260. https://doi.org/10.20961/arst.v21i2.77861

Dennis, J. E., & Schnabel, R. B. (1996). Secant methods for unconstrained minimization. In *Numerical methods for unconstrained optimization and nonlinear equations* (pp. 194–215). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611971200.ch9

Dotzour, M., Grissom, T., Liu, C., & Pearson, T. (1990). Highest and best use: The evolving paradigm. *Journal of Real Estate Research*, *5*(1), 17–32. https://doi.org/10.1080/10835547.1990.12090599

Fletcher, R. (2000). *Practical methods of optimization*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118723203

Garmaise, M. J., & Moskowitz, T. J. (2004). Confronting information asymmetries: Evidence from real estate markets. *The Review of Financial Studies*, *17*(2), 405–437. https://doi.org/10.1093/RFS/HHG037

Gatzlaff, D., & Tirtiroğlu, D. (1995). Real estate market efficiency: Issues and evidence. *Journal of Real Estate Literature*, *3*(2), 157–189. https://doi.org/10.1080/10835547.1995.12090046

Gdakowicz, A., Putek-Szeląg, E., & Kuźmiński, W. (2019). Information asymmetry and mass appraisal. *Metody Ilościowe w Badaniach Ekonomicznych*, *20*(3), 149–166. https://doi.org/10.22630/MIBE.2019.20.3.15

Ho, T. K. (1995). Random decision forests. In *Proceedings of the International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Hong, J., & Kim, W. (2022). Combination of machine learning-based automatic valuation models for residential properties in South Korea. *International Journal of Strategic Property Management*, *26*(5), 362–384. https://doi.org/10.3846/IJSPM.2022.17909

Huber, R., D'Onofrio, C., Devaraju, A., Klump, J., Loescher, H. W., Kindermann, S., Guru, S., Grant, M., Morris, B., Wyborn, L., Evans, B., Goldfarb, D., Genazzio, M. A., Ren, X., Magagna, B., Thiemann, H., & Stocker, M. (2021). Integrating data and analysis technologies within leading environmental research infrastructures: Challenges and approaches. *Ecological Informatics*, *61*, Article 101245. https://doi.org/10.1016/J.ECOINF.2021.101245

Ionascu, E., Mironiuc, M., & Anghel, I. (2019). Transparency of real estate markets: Conceptual and empirical evidence. *Audit Financiar*, *17*(154), 306–326. https://doi.org/10.20869/AUDITF/2019/154/013

Johnson, K. H., Springer, T. H., & Brockman, C. M. (2005). Price effects of non-traditionally broker-marketed properties. *The Journal of Real Estate Finance and Economics*, *31*, 331–343. https://doi.org/10.1007/s11146-005-2793-3

Jung, J., Kim, J., & Jin, C. (2022). Does machine learning prediction dampen the information asymmetry for non-local investors? *International Journal of Strategic Property Management*, *26*(5), 345–361. https://doi.org/10.3846/IJSPM.2022.17590

Keskin, B., & Watkins, C. (2017). Defining spatial housing submarkets: Exploring the case for expert delineated boundaries. *Urban Studies*, *54*(6), 1446–1462. https://doi.org/10.1177/0042098015620351

Klein, T. J., Lambertz, C., & Stahl, K. O. (2016). Market transparency, adverse selection, and moral hazard. *Journal of Political Economy*, *124*(6), 1677–1713. https://doi.org/10.1086/688875

Kurlat, P., & Stroebel, J. (2014). *Testing for information asymmetries in real estate markets* (Working Paper No. 19875). National Bureau of Economic Research. https://doi.org/10.3386/w19875

Levitt, S. D., & Syverson, C. (2008). Market distortions when agents are better informed: The value of information in real estate transactions. *The Review of Economics and Statistics*, *90*(4), 599–611. https://doi.org/10.1162/rest.90.4.599

Li, L., & Chau, K. W. (2024). Information asymmetry with heterogeneous buyers and sellers in the housing market. *Journal of Real Estate Finance and Economics*, *68*(1), 138–159. https://doi.org/10.1007/s11146-023-09939-y

Ling, D. C., Naranjo, A., & Petrova, M. T. (2018). Search costs, behavioral biases, and information intermediary effects. *The Journal of Real Estate Finance and Economics*, *57*, 114–151. https://doi.org/10.1007/s11146-016-9582-z

Meszek, W., & Dziadosz, A. (2011). Wpływ nieefektywności rynku nieruchomości na dokładność opisu wartości nieruchomości za pomocą liniowych modeli regresji wielorakiej. *Budownictwo i Inżynieria Środowiska*, *2*(4), 589–594.

Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather*, *16*(1), 69–88. https://doi.org/10.1002/2017SW001669

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*(2), 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

Ogryzek, M. (2023). The sustainable development paradigm. *Geomatics and Environmental Engineering*, *17*(1), 5–18. https://doi.org/10.7494/geom.2023.17.1.5

Ragil Budi Perkasa, A., Utomo, C., & Budi Santoso, E. (2023). A review of research methods on highest and best use for toll rest area. *Materials Today: Proceedings*, *85*, 19–23. https://doi.org/10.1016/j.matpr.2023.05.247

Renigier-Bilozor, M., Janowski, A., & Walacik, M. (2019). Geoscience methods in real estate market analyses subjectivity decrease. *Geosciences*, *9*(3), Article 130. https://doi.org/10.3390/geosciences9030130

Renigier-Biłozor, M., Janowski, A., Walacik, M., & Chmielewska, A. (2022). Modern challenges of property market analysis-

homogeneous areas determination. *Land Use Policy*, *119*, Article 106209. https://doi.org/10.1016/j.landusepol.2022.106209

Ribera, F., Nesticò, A., Cucco, P., & Maselli, G. (2020). A multicriteria approach to identify the highest and best use for historical buildings. *Journal of Cultural Heritage*, *41*, 166–177. https://doi.org/10.1016/j.culher.2019.06.004

Ries, A., & Trout, J. (1994). *The 22 immutable laws of marketing: Violate them at your own risk!* https://archive.org/details/22immutablelawso00alri/page/n1/mode/2up

Royal Institution of Chartered Surveyors. (2022). *Automated valuation models (AVMs): Implications for the profession and their clients*. Royal Institution of Chartered Surveyors (RICS).

Rymarzak, M., Siemińska, E., & Sakierski, K. (2022). Reflecting sustainability in the analysis of highest and best use: Evidence from Polish municipalities. *Real Estate Management and Valuation*, *30*(4), 103–115. https://doi.org/10.2478/remav-2022-0032

Sayce, S., & Connellan, O. (2002). From existing use to value in use: Time for a paradigm shift? *Property Management*, *20*(4), 228–251. https://doi.org/10.1108/02637470210444268

Sayce, S., Smith, J., Cooper, R., & Venmore-Rowland, P. (2006). *Real estate appraisal: From value to worth*. Wiley-Blackwell.

Schulze, F., & Windhorst, E. (2014). *Mies van der Rohe: A critical biography*. The University of Chicago Press.

Tisdell, C. (1970). On the theory of externalities. *Economic Record*, *46*(1), 14–25. https://doi.org/10.1111/j.1475-4932.1970.tb02462.x

Trinh, T. H. (2018). Towards a paradigm on the value. *Cogent Economics & Finance*, *6*(1), Article 1429094. https://doi.org/10.1080/23322039.2018.1429094

Utomo, C., Rahmawati, Y., & Krestawan, I. (2018). Development of urban market spatial for highest and best use of land productivity and sustainability. *Planning Malaysia*, *16*(1), 163–172. https://doi.org/10.21837/PM.V16I5.420

Vandell, K. D. (1982). Toward analytically precise definitions of market value and highest and best use. *Appraisal Journal*, *50*(2), 253–268.

Vandell, K. D., & Carter, C. C. (2000). Graaskamp's concept of highest and best use. In J. R. DeLisle & E. M. Worzala (Eds.), *Essays in honor of James A. Graaskamp: Ten years after* (pp. 307–319). Springer. https://doi.org/10.1007/978-1-4615-1703-0_15

Walacik, M., & Chmielewska, A. (2024a). Energy performance in residential buildings as a property market efficiency driver. *Energies*, *17*(10), Article 2310. https://doi.org/10.3390/EN17102310

Walacik, M., & Chmielewska, A. (2024b). Real estate industry sustainable solution (environmental, social, and governance) significance Assessment-AI-Powered algorithm implementation. *Sustainability*, *16*, Article 1079. https://doi.org/10.3390/SU16031079

Walacik, M., & Janowski, A. (2024). *The original methodology for homogeneous area determination (HAD) for the purpose of property taxation procedures' fairness and equity increase* [Unpublished presentation form GIS/Valuation Technologies Conference 2024].

Wang, F., Gai, Y., & Zhang, H. (2024). Blockchain user digital identity big data and information security process protection based on network trust. *Journal of King Saud University – Computer and Information Sciences*, *36*(4), Article 102031. https://doi.org/10.1016/j.jksuci.2024.102031

Watkins, C. (1999). Property valuation and the structure of urban housing markets. *Journal of Property Investment & Finance*, *17*(2), 157–175. https://doi.org/10.1108/14635789910258543

Watkins, C. (2001). The definition and identification of housing submarkets. *Environment and Planning A: Economy and Space*, *33*(12), 2235–2253. https://doi.org/10.1068/a34162

Wong, S. K., Yiu, C. Y., & Chau, K. W. (2012). Liquidity and information asymmetry in the real estate market. *The Journal of Real Estate Finance and Economics*, *45*, 49–62. https://doi.org/10.1007/s11146-011-9326-z

Zhou, X., Gibler, K., & Zahirovic-Herbert, V. (2015). Asymmetric buyer information influence on price in a homogeneous housing market. *Urban Studies*, *52*(5), 891–905. https://doi.org/10.1177/0042098014529464

## Appendix

**Table A1.** Encoding of intrinsic variables presented with basic descriptive statistics (source: own elaboration)

| No | Property attribute | Coding | Basic descriptive statistics | | | |
|---|---|---|---|---|---|---|
| | | | Min | Max | Mean | St. Dev. |
| 1 | Location | address | descriptive variable | | | |
| 2 | Parcel | square feet | 1534 | 89846 | 15640 | 8391 |
| 3 | Building built-up area | square feet | 936 | 9121 | 2582 | 1132 |
| 4 | Building heated/usable area | square feet | 1120 | 5877 | 2516 | 972 |
| 5 | Street type | 1-cir, 2-ct, 3-dr, 4-ln, 5-pl, 6-rd, 7-st, 8-way | 1 | 8 | 4 | 2 |
| 6 | Date of construction | year | 1951 | 2021 | 1978 | 22 |
| 7 | Design style | 1-conventional, 2-ranch, 3-split foyer, 4-split level, 5-townhouse | 1 | 5 | 2 | 1 |
| 8 | Utilities | 1-all, 2-e, 3-w | 1 | 3 | 1 | 0 |
| 9 | Effective building year | year | 1955 | 2021 | 1992 | 19 |
| 10 | Remodeled year | year | 0 | 2021 | 215 | 623 |
| 11 | Number of stories | numeric | 1 | 7 | 2 | 2 |
| 12 | Foundation basement | 0-full basement, 1-% basement, 2-pier foundation, 3-no basement | 0 | 3 | 2 | 1 |

*End of Table A1*

| No | Property attribute | Coding | Basic descriptive statistics | | | |
|---|---|---|---|---|---|---|
| | | | Min | Max | Mean | St. Dev. |
| 13 | Foundation basement | percentage | 0 | 95 | 14 | 24 |
| 14 | Exterior wall | 1-frame, 1-brick, 3-brick & frame, 4-alum vinyl siding | 0 | 4 | 2 | 1 |
| 15 | Air conditioning | 1-separate, 2-no air conditioning | 0 | 2 | 1 | 0 |
| 16 | Bath | 1-one bath, 2-one and a half bath, 3-two baths, 4-two and a half bath, 5-three baths, 6-three and a half bath | 1 | 6 | 4 | 1 |
| 17 | Bath Fixtures | numeric | 0 | 10 | 0 | 2 |
| 18 | Built ins | 0-no fireplace, 1 one fireplace, 2-multiple fireplace | 0 | 2 | 1 | 1 |
| 19 | Floor finish | lack of data | | | | |
| 20 | Interior finish | lack of data | | | | |
| 21 | Heating | 1-forced air | 1 | 1 | 1 | 1 |
| 22 | Grade factor | 1-A, 2-A+05, 3-A+10, 4-A+20, 5-A+25, 6-A-5, 7-A-10, 8-AA, 9-AA+05, 10-AA+10, 11-AA+15, 12-AA+20, 13-AA+25, 14-AA+30, 15-AA+50, 16-AA-5, 17-AA-10, 18-AA-15, 19-B, 20-B-05, 21-B-105, 22-C+10 | 1 | 22 | 15 | 6 |
| 23 | Assessed condition | numeric | 3 | 97 | 67 | 24 |

**Table A2.** Extrinsic variables encoding (source: own elaboration)

| Factors | Layer name (number of objects) | Kind | Object classes | Extraction |
|---|---|---|---|---|
| Environmental conditions | Waterways (568) | lines | stream, drain | Distance (Euclidean & PGRouting) |
| | Water (109) | polygons | reservoir, riverbank, water, wetland | Distance (Euclidean & PGRouting) |
| Communication | Transport (829) | points | bus stops, helipad, railway halt, railway station, taxi | Euclidean & PGRouting/ number in units |
| | Railways (197) | lines | railway | Euclidean & PGRouting/ length in units |
| | Traffic (2680) | polygons | parking, parking multistorey, parking underground | Euclidean & PGRouting/area in zone |
| | Roads (30683) | lines | bridleway, cycleway, footway, motorway, path, pedestrian, primary, residential, secondary, service, tertiary, track, trunk, unclassified | Euclidean & PGRouting/ length in units |
| Facilities/services | Sacred objects (70+51) | Points /polygons | Christian, Catholic, Lutheran, Methodist, Anglican, Muslim | Euclidean & PGRouting/ number in units |
| | Points of interest (108) | points | arts center, bakery, bank, bar, beauty shop, bookshop, café, chemist, cinema, clinic, community center, dentist | Euclidean & PGRouting/ number in units |
| | Buildings (44984) | polygons | apartments, boathouse, college, commercial, dormitory, fire station, garage, government, hospital, hotel, house, industrial, museum, office, prison, public, residential, retail, school, theatre, train station, university, warehouse (...) | Number in units/area in units |
| Aesthetics* | – | – | – | – |
| Social and economic background* | – | – | – | – |

*Note:* * The factors were not represented by objects.