EKONOMIA I PRAWO. ECONOMICS AND LAW
Volume 23, Issue 4, December 2024
p-ISSN 1898-2255, e-ISSN 2392-1625
https://apcz.umk.pl/EiP

ORIGINAL ARTICLE
received 15.11.2024; accepted 12.12.2024; published 30.12.2024
Citation: Orzeszko, W., Fałdziński, M. & Siemińska, E. (2024). Real Estate Valuation Algorithm Using Nonparametric Regression Methods . *Ekonomia i Prawo. Economics and Law*, 23(4): 655–673.
doi:10.12775/EiP.2024.33

Real Estate Valuation Algorithm Using Nonparametric Regression Methods

WITOLD ORZESZKO

corresponding author
Department of Applied Informatics and Mathematics in Economics, Faculty of Economic Sciences and Management, Nicolaus Copernicus University in Torun, Gagarina 13a, 87–100 Torun
✉ vit@umk.pl
ORCID: https://orcid.org/0000-0001-7473-7775

MARCIN FAŁDZIŃSKI

Department of Econometrics and Statistics, Faculty of Economic Sciences and Management, Nicolaus Copernicus University in Torun, Gagarina 13a, 87–100 Torun
✉ marf@umk.pl
ORCID: https://orcid.org/0000-0002-6236-8500

EWA SIEMIŃSKA

Department of Investment and Real Estate, Faculty of Economic Sciences and Management, Nicolaus Copernicus University in Torun, Gagarina 13a, 87–100 Torun
✉ ewahsiem@umk.pl
ORCID: https://orcid.org/0000-0002-8885-0338

Abstract

Motivation: Real estate valuation is critical for various market participants and is influenced by numerous factors such as geographic location, property characteristics, and market trends. Traditional valuation methods often struggle to handle the complexity, scale, and

dynamic nature of real estate markets. Automated valuation models (AVMs) have emerged as a promising alternative, offering scalability, speed, and adaptability. However, their effectiveness depends heavily on the ability to manage diverse data and model nonlinear relationships in valuation.

Aim: This paper presents a novel algorithm for real estate valuation based on nonparametric regression methods, including local-linear kernel and Nadaraya-Watson estimators. The algorithm is designed to handle missing data, incorporate multiple influencing factors with adjustable weights, and adapt to dynamic market conditions. It aims to provide accurate property valuations by leveraging a systematic and flexible approach to segment data, select nearest neighbors, and update prices based on market trends.

Results: The proposed algorithm demonstrates robustness to data errors, flexibility in handling variables measured on different scales, and the capacity to model nonlinear relationships. It effectively estimates property values using a comprehensive eight-stage process, from data preparation to quality assessment. The proposed algorithm is designed to be scalable and adaptable, making it a promising tool for addressing the challenges of real estate valuation across diverse contexts.

Keywords: real estate valuation algorithm; AVM; nonparametric regression methods
JEL: R30; C14; R31, C53

1. Introduction

One of the key elements of any market, including the real estate market, is the valuation of its assets, which provides essential information to various market participants, such as buyers and sellers, owners and investors, tenants, developers and brokers, as well as banks, state and local authorities, and tax agencies.

The primary reasons for estimating the value of real estate include determining transaction prices in the real estate market, calculating compensation amounts, securing financial credit (as the basis for collateral), establishing rental rates, determining fees related to planning and development, and facilitating court proceedings and tax assessments. Various methods are used to determine the value of real estate, from traditional valuation methods to increasingly popular automated valuation models (AVMs). The aim of this paper is to present a novel automated valuation model (AVM). The proposed algorithm is based on nonparametric regression models, specifically local-linear kernel and Nadaraya-Watson estimators. The algorithm has several properties that make it a valuable tool for property valuation, including robustness to errors and missing data, scalability, the ability to incorporate various types of data, and the capacity to account for multiple factors influencing valuation, with adjustable weights assigned to each factor. Additionally, due to the use of nonparametric methods, the algorithm can capture nonlinearities in the data, does not require strict assumptions regarding

data and modeled dependencies, is robust to estimation errors, and is highly adaptable to dynamic changes in the real estate market. Furthermore, it is characterized by universality, as it can be implemented and adapted to the specific features of various real estate transaction databases.

The remainder of the paper is organized as follows: Section 2 addresses the institutional and legal conditions of the real estate market as well as the premises and essence of the valuation process. Section 3 presents the nature, properties, and construction of the proposed property valuation algorithm. Section 4 concludes the paper.

2. Institutional and Legal Frameworks of the Real Estate Market

The real estate market is a highly complex structure, which—depending on informational needs and context—can be examined from economic, technical (construction), legal, spatial, social, and other perspectives. Extensive literature, both international and Polish, provides a variety of definitions for the real estate market, including institutional perspectives that view it as a set of "...rules, conventions, and relationships that together create a system in which property is used, and rights to it are subject to transactions" (Załęczna, 2010, p. 67). This system is a multi-level structure: at the top are political, social, economic, and legal principles, followed below by the real estate market understood as a network of principles and interrelations tied to the functioning of real estate resources, and at the bottom are the organizations and institutions interacting and bringing about changes in the market (Keogh and D'Arcy, 1999, pp. 2401–2414). These changes are particularly significant as they result from socio-economic evolution and develop over extended periods.

Marek Bryx similarly defines the real estate market as a system consisting of four distinct subsystems: real estate transactions, real estate investment, real estate financing, and real estate management. In this approach, it is necessary to define "...the elements of the system, the components of the system's environment, the essential relationships between elements, and the significant relationships between the system and its environment (inputs and outputs)" (Bryx, 2006, p. 87).

In this context, Ewa Kucharska-Stasiak highlights that institutional order – supported by efficient, trusted institutions and an evolving legal system – is essential for the effective functioning of the real estate market. The legal system, in particular, is often considered the key component that supports the development of the institutional framework of the real estate market and guarantees the protection of property rights (Kucharska-Stasiak, 2016, p. 55). In turn, Leszek Kalkowski, doyen of the Polish real estate market, very aptly

notes that the real object of transactions in this market, and therefore de facto "commodity," is not so much the real estate itself, but the rights to the real estate, including ownership, limited rights in rem and contractual rights such as rent and lease (Kałkowski, ed., 2003, p. 20).

Geoffrey Keogh and Eamonn D'Arcy, in their work on the real estate market, point to a certain relativity in assessing the market's effectiveness. They argue that this assessment depends, firstly, on the time frame under consideration, as the market must adapt to changing socio-economic conditions, and secondly, on the perspective of the evaluating entity (Keogh and D'Arcy, 1999, pp. 2401–2414). Thus, it is possible to have both positive and negative evaluations of the real estate market's functioning depending on whether the conditions are favorable for a given market participant (Siemińska, 2013, p. 28).

3. Essence of Real Estate Valuation

Real estate valuation is subject to specific procedures and regulations, which in Poland are governed by the Real Estate Management Act (Act, 1997) and the Regulation of the Minister of Development and Technology on the valuation of real estate (Act, 2023). Article 4 of the Act defines real estate valuation as "...a process by which the value of real estate is determined (...) as the subject of ownership rights and other property rights" (Real Estate Management Act, 1997).

The most recognizable and frequently encountered type of real estate value is market value, which is defined as "the estimated amount that, on the valuation date, could be obtained for the property in a sale transaction conducted under market conditions between a willing buyer and seller, who are both informed and acting prudently, and are not under compulsion" (Real Estate Management Act, 1997, Article 151, Section 1). Furthermore, Article 134, Section 2 specifies that "when determining the market value of real estate, particular consideration should be given to its type, location, usage, purpose, condition, and prevailing market prices" (Real Estate Management Act, 1997). Therefore, this assessment should be preceded by a market analysis, during which the property appraiser—"an individual with professional qualifications in property valuation, granted in accordance with the Act" (Real Estate Management Act, 1997, Article 174, Section 2))—gathers market information on pricing, rental rates, transaction conditions, and inspects the property being appraised. This latter point often causes practical confusion; while inspecting similar properties is a sound and advisable practice, it is not always feasible, as appraisers lack the legal authority to access any property deemed comparable to the one being valued. Consequently, the extent of the inspection or the decision to forgo it in particularly justified

cases should be recorded and justified in the valuation report (Valuation of real estate Act, 2023).

From the perspective of determining a property's market value, the most desirable situation is to obtain information on transaction prices and the characteristics of sold properties. However, this is only possible in some European countries where real estate prices are officially recorded (Rącka, 2024).

In addition to market value, the Real Estate Management Act allows for the determination of replacement value, used when properties, due to their type or current use or purpose, cannot or are not likely to be subject to market transactions, or when special regulations require it (Real Estate Management Act, 1997, Article 150, Section 3). In the real estate market, other types of value, such as cadastral, investment, banking-mortgage, fair value, forced-sale value, and "hope value," are also recognized (Kucharska-Stasiak, 2016; Trojanowski, 2019; ed. Dydenko, 2006; PFSRM, 2012; Konowalczuk, 2014; Malmon, 2022).

In the context of institutional and legal frameworks for the real estate market, the Real Estate Management Act identifies possible methods for determining property values, including comparative, income, or cost approaches, or a mixed approach that incorporates elements of the previous methods (Real Estate Management Act, 1997, Articles 152 and 153). The decision regarding the appropriate approach and the method and technique of property valuation is made by the property appraiser, considering the purpose of the valuation, the type and location of the property, its designation in the local zoning plan, condition, and available data on prices, income, and characteristics of similar properties.

In practice, a common issue arises in determining the purpose of a given property in the local zoning plan. According to estimates by the Institute of Geography and Spatial Planning of the Polish Academy of Sciences, about 31.7% of Poland's area is covered by local plans, though coverage is uneven. Certain regions have high coverage levels, between 66% and 73% (Lower Silesian, Małopolskie, and Śląskie voivodeships), while others do not reach even 10% (Kuyavian-Pomeranian, Lubuskie, and Podkarpackie voivodeships) (Śleszyński, 2023; Kowalewski, Markowski, and Śleszyński, 2018; PAP, 2022). In the absence of a local plan, the appraiser should establish the property's intended use based on a decision on development and land management conditions. In the absence of such a decision, the appraiser should base this on the property's current usage or the municipality's general planning framework.

The aforementioned Real Estate Valuation Ordinance specifies:

1) types of real estate valuation methods and techniques;
2) ways of determining property value;
3) methods for assessing the value of improvements and damages on real estate;

4) procedures, format, and content for valuation reports" (Valuation of real estate Act, 2023, § 1). Additionally, §3 of the Regulation states that "the appropriate real estate market is determined by indicating its type, area, and research period, in line with the subject, scope, purpose, method of valuation, and availability of data" (Valuation of real estate Act, 2023). This practically means that it is the responsibility of the property appraiser to select the appropriate information used in the valuation process, including identifying "comparable properties" for further analysis, as well as other sources of information about the property being appraised.

The importance of accurate and up-to-date information on the real estate market for various stakeholders is highlighted by a provision in the Polish Financial Supervision Authority's (KNF) Recommendation J, which states that "from the perspective of a bank's operational security, a prudent policy for financing real estate transactions, especially the assessment of collateral value, is crucial. Given the high share of mortgage-secured credit exposures in banks' credit portfolios and the economic significance of these exposures, incorrect or overly lenient adoption of collateral values may increase systemic risk. To mitigate potential crisis impacts, banks should possess the most comprehensive knowledge possible about the real estate market. (…) Databases should primarily serve to verify collateral values and update them throughout the credit term. Information in the databases should enable analysis of local real estate markets, particularly identifying market changes and risks associated with properties securing credit exposures" (KNF, 2023).

The collateral value assessment of real estate is defined in Recommendation S by the Polish Financial Supervision Authority on best practices for managing mortgage-secured credit exposures as "the bank's estimate of the amount obtainable from the market sale of the property securing the credit exposure, updated as of the loan issuance or the next valuation point, based on statistical methods or market analysis" (KNF, 2019). Notably, the KNF explicitly states in Recommendation S that "the bank's assessment of collateral value should not be equated with a property valuation performed by an appraiser, as the objectives and, consequently, the processes (rules and procedures) differ" (KNF, 2019). This distinction indicates that the valuation process and the collateral value assessment process differ in rationale, purpose, and nature.

Given this distinction, it is essential to note the range of methods used to determine these values, from traditional valuation methods to increasingly popular automated valuation models (AVMs). The main advantages of AVMs include their speed, lower cost, immunity to appraiser biases or stakeholder influence on valuation results, ease of use, and applicability for large property groups, such as for tax purposes or in mortgage credit procedures (Bełej, Figurska, 2017; RICS, June 2021; TEGoVA, 2020). However, the most significant limitations of AVMs are that they do not capture the

property characteristics that appraisers identify during property inspections, and they overlook specific property traits in favor of the most standard, typical ones for which information is more abundant. AVM providers require large amounts of reliable, detailed data on property features and transaction prices to accurately model the market (RICS, June 2021; Kucharska-Stasiak, 2018; Sing et al., 2022; Matysiak, 2017, 2018; Peyman et al., 2024; University of Oxford Research, 2022).

Acknowledging the advantages of using big data and artificial intelligence in the modern world, international property appraiser organizations, such as The European Group of Valuers Associations (TEGoVA) and the Royal Institution of Chartered Surveyors (RICS), and national organization – The Polish Federation of Valuers Associations (PFSRM) recommend using AVMs as tools to support appraisers and enhance their skills in the valuation process (RICS, June 2021; TEGoVA, 2020; PFSRM, 2017). According to University of Oxford experts, AVMs will continuously improve with advancements in AI, offering broad applicability in practice and becoming indispensable tools for the modern real estate sector and the economy as a whole (University of Oxford Research, 2022).

4. Construction of the Algorithm

The aim of this algorithm is to estimate the current market value of a property based on characteristics provided by the user, using historical transaction data from the real estate market. The valuation of a single property can be influenced by numerous factors, including geographic location, year of construction, living area, technical condition, standard of the property, construction technology, property area, surroundings, number of rooms, floor level, and more. An added challenge in predicting individual property prices is the variability of historical transactions over time, necessitating the comparison of past property prices by accounting for market price dynamics. This is complex, as the real estate market is subject to dynamic and often irregular fluctuations and trends.

The algorithmic approach, unlike expert-based methods, leverages extensive data resources, naturally linked to data quality and diversity issues. Due to the complexity and characteristics of this problem, standard prediction methods may not be suitable or may yield high estimation errors. To provide an accurate property valuation, all the above factors must be incorporated into the valuation algorithm. As a result, the algorithm should be universal, flexible, and capable of adapting to numerous factors that may affect property value.

The algorithm is based on a database of transactions related to a specific real estate segment, to which the property being evaluated belongs. Information in the database is segmented under the assumption that the algorithm

will be optimized and executed separately for each segment. These segments should be as homogeneous as possible, grouping similar properties together without excessive segmentation, as this would reduce the sample size in each segment, thereby limiting data resources for valuation. The database is first cleaned to remove erroneous data (according to the GIGO principle, "garbage in, garbage out").

Let N_0 denote the total number of transactions (i.e., the number of records in the database), and denote the number of explanatory variables (i.e., property attributes). Each transaction in the database can be viewed as a pair $(\overline{x}_i, y_i)$, where:

- i is the transaction number $i = 1,2, ..., N_0$,
- y_i is the price of the property per square meter,
- $\overline{x}_i$ is the vector of explanatory variables.

Each vector $\overline{x}_i$ has components, i.e., $\overline{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})$, where for each j (where j = 1,2, ..., K) x_{ij} represents the value of the j-th explanatory variable in the i-th vector. Depending on the nature of a given variable, each x_{ij} may be a numerical value (e.g., property area) or a categorical label (e.g., province).

In statistics, variables can appear on different measurement scales: nominal, ordinal, interval, and ratio. The nominal scale offers the least precision, used for variables taking values (so-called labels) without any inherent ordering (e.g., location). The ordinal scale allows ordering observations by the values of the examined variable (e.g., property standard). The interval scale not only preserves order but also allows measuring the distance between values (e.g., year of construction). The ratio scale has a natural zero point, enabling comparison of magnitudes (e.g., property area, property price). Nominal and ordinal scales are considered "weak", while interval and ratio scales are considered "strong" (see Sobczyk, 2024).

The proposed property valuation algorithm is universal in nature and can be easily adapted for use with any real estate transaction database.

The proposed property valuation algorithm comprises eight key stages:

1. Creation of dynamic variables,
2. Variable classification,
3. Selection of valid records,
4. Variable normalization,
5. Selection of nearest neighbors,
6. Price update for nearest neighbors,
7. Property valuation,
8. Quality assessment of valuation.

Step 1: This stage involves creating certain new variables (if needed), with values not available in the database, as they need to be dynamically generated based on user-specified property characteristics. An example could be the variable "Distance," which indicates the geographic distance between properties in the database and the property being valued.

Step 2: Based on expertise, each explanatory variable is assigned to one of three categories:

- "Redundant" variables, which do not influence property prices and are excluded from further steps in the algorithm.
- "Necessary" variables, where differences in values of these renders properties to be considered dissimilar regardless of similarities in other variables (these act as filters for records).
- "Useful" variables, which shape property prices but are not classified as "necessary."

This classification affects each variable's role and utilization in subsequent steps of the algorithm, and is determined based on expert knowledge.

Step 3: At this stage, records are filtered based on specified validity criteria. Transactions may be excluded from further analysis if they are too outdated, lack essential information, or have certain (undesirable) values for a given variable. It may also be necessary to remove records with outliers for certain variables. Additionally, records that do not have identical values for "necessary" variables, as in the case of the property being valued, should be discarded. For example, if geographic location is considered a "necessary" variable, only transactions involving properties with the same location as the valued property will be used in further stages. The term "valid records" refers to records that meet all assumed selection criteria, for the remainder of the algorithm.

Step 4: Normalization is performed for "valid records" to transform variables (which naturally have values across different ranges) into a standard range of $\langle 0, 1 \rangle$. Each "useful" variable on an ordinal or stronger scale is normalized. The most common normalization method used is the unit transformation:

$$x'_{ij} = \frac{x_{ij} - min_j}{R_j}, \tag{1}$$

where x_{ij} is the value of the j-th "useful" variable in the i-th record, x'_{ij} is the normalized value for x_{ij}, min_j is the minimum value of the j-th variable, and R_j is the range of this variable, defined as $R_j = max_j - min_j$ (see, e.g., Walesiak 2011). Normalization is required for correctly conducting the next algorithm step – selecting the nearest neighbors.

Step 5: This step aims to select, from among the "valid records," properties most similar in characteristics to the user-specified property. Let $\overline{x}'_0$ represent the vector of normalized explanatory variables for the property being evaluated. To determine if property $\overline{x}'_i$ can be considered among the nearest neighbors for $\overline{x}'_0$ a measure of similarity between them must be established. A suitable distance metric needs to be used to account for the variable types

that make up the components of these vectors (see, e.g., Walesiak 2011). Since the explanatory variables are measured on different scales, a universal distance measure capable of calculating distances regardless of scale is required. The most commonly used example of such a measure is Gower's distance (1971) and its generalization proposed by Cox and Cox (2000).

Gower's distance, denoted as $d(\overline{x}_i', \overline{x}_0')$ measures the distance between entire vectors, $\overline{x}_i'$ and $\overline{x}_0'$ but first requires calculating distances between the individual components of these vectors. These component distances are denoted as $d_j(\overline{x}_i', \overline{x}_0')$, where $j = 1,2, ..., K$. For each j, the value $d_j(\overline{x}_i', \overline{x}_0')$ expresses the similarity between properties $\overline{x}_i$, and $\overline{x}_0$ and in terms of the j-th attribute. The distances $d_j(\overline{x}_i', \overline{x}_0')$ are calculated sequentially for each j-th variable and for each "valid record" $\overline{x}_i'$.

For a variable measured on an ordinal or strong scale, the distance between properties $\overline{x}_i'$ and $\overline{x}_0'$ is determined using the formula:

$$d_j(\bar{x}_i', \bar{x}_0') = \left| x_{ij}' - x_{0j}' \right|, \tag{2}$$

where x_{ij}' and x_{0j}' represent the values of the j-th variable in vectors $\overline{x}_i'$ and $\overline{x}_0'$, respectively.

For a variable measured on a nominal scale, the distance between $\overline{x}_i'$ and $\overline{x}_0'$ is determined as follows:

$$d_j(\bar{x}_i', \bar{x}_0') = \begin{cases} 1; \text{ if the value of the } j\text{-th variable in } \bar{x}_i' \text{ and } \bar{x}_0' \text{ differ} \\ 0; \text{ if the value of the } j\text{-th variable in } \bar{x}_i' \text{ and } \bar{x}_0' \text{ are identical} \end{cases} \tag{3}$$

Once $d_j(\overline{x}_i', \overline{x}_0')$ has been calculated for each component, the Gower distance between the entire vectors $\overline{x}_i'$ and $\overline{x}_0'$ can be determined using the formula:

$$d(\bar{x}_i', \bar{x}_0') = \frac{\sum_{j=1}^{K} \delta_j(\bar{x}_i', \bar{x}_0') v_j d_j(\bar{x}_i', \bar{x}_0')}{\sum_{j=1}^{K} \delta_j(\bar{x}_i', \bar{x}_0') v_j}, \tag{4}$$

where $\delta_j(\overline{x}_i', \overline{x}_0')$ takes the value 1 if measurement on variable is possible for both $\overline{x}_i'$, $\overline{x}_0'$ and 0 otherwise. The values v_j are pre-defined weights, determined by expert judgment, reflecting the importance of the j-th attribute.

From the normalized "valid" records, those with a Gower $d(\overline{x}_i', \overline{x}_0')$ distance below a specified threshold are selected. Among these, records with extreme values for the explained variable (i.e., property price) are further excluded. Properties identified through this process are referred to as the "nearest neighbors" for the remaining steps in the algorithm.

Step 6: Let N denote the number of nearest neighbors $\overline{x}_i'$ identified for the user-specified property $\overline{x}_0'$ and let y_i be the price of property $\overline{x}_i'$ (for i = 1,2, ..., N). This stage of the algorithm aims to update the prices y_i to the current valuation date. To do this, a trend model is estimated, which is then used to calculate updated prices.

Initially, the time aggregation level of the analyzed transactions must be determined. For this stage, it is considered sufficient to use monthly data, meaning that each transaction $\overline{x}_i'$ has a time distance t_i (expressed in months) from the current month. Thus, for each i, the value t_i is a natural number starting from 0.

Let y_i^* denote the updated value of the price y_i (for i = 1,2, ..., N). In the algorithm, a trend function $f(t)$ is estimated, where the argument is the time variable $t \in \{t_1, t_2, ..., t_N\}$ For each t, the function $f(t)$ represents the (theoretical) price of a property (with specified characteristics) at time t.

Various modeling approaches can be used to determine the trend function. The simplest method is to construct a linear trend model, which assumes a linear mechanism for price changes over time. This assumption is inherently restrictive, and in practice, the linear model is often treated as an approximation of the true price dynamics, which are usually nonlinear.

An alternative approach is nonparametric regression, which approximates the unknown true relationship with functions that are sufficiently flexible to improve approximation accuracy as the sample size increases (Härdle et al. 1997). The primary advantage of nonparametric regression is its universality, achieved by avoiding restrictive assumptions about the analytical form of the model. This approach allows the "data to speak for themselves", resulting in models that can fit the analyzed data well (Orzeszko, 2018a; Orzeszko, Bejger, 2018). This flexibility is particularly valuable for modeling nonlinear relationships, as confirmed by simulation studies (e.g., Vilar-Fernández, Cao, 2007; Orzeszko, 2018b).

One of the most important methods in nonparametric regression is kernel smoothing. The most widely used kernel smoother is the Nadaraya-Watson estimator (referred to as N-W hereafter; Nadaraya, 1964; Watson, 1964). However, literature indicates that the Nadaraya–Watson method suffers from boundary bias, which makes it unsuitable for trend-based forecasting (see Fan, Gijbels, 1996). Therefore, the algorithm proposes using an alternative kernel-based method called the local-linear kernel estimator (LLKE).

LLKE combines local linear approximation with kernel smoothing. This method estimates the regression function by "locally" fitting a first-degree polynomial to the data through weighted least squares, where the weights are defined by the kernel function (see Stone, 1977; Fan, Gijbels, 1992; Orzeszko, Bejger, 2018). For each $t \in \{t_1, t_2, ..., t_N\}$, the value $f(t)$ in the LLKE method is calculated as:

$$f(t) = \sum_{i=1}^{N} w_i(t) y_i. \tag{5}$$

This means that $f(t)$ is a weighted average of the prices y_i. The core of the method lies in calculating the weights $w_i(t)$, which are computed separately for each using the formula:

$$w_i(t) = \frac{k\left(\frac{t-t_i}{h}\right)\left(\hat{s}_2(t) - (t-t_i)\hat{s}_1(t)\right)}{\sum_{i=1}^{N} k\left(\frac{t-t_i}{h}\right)\left(\hat{s}_2(t) - (t-t_i)\hat{s}_1(t)\right)}, \tag{6}$$

where

$$\hat{s}_r(t) = \sum_{i=1}^{N} k\left(\frac{t-t_i}{h}\right)(t-t_i)^r, \qquad \text{for } r = 1, 2, \tag{7}$$

where k is the kernel function, and h is the bandwidth. The kernel is a non-negative, symmetric function that satisfies $\int_{-\infty}^{+\infty} k(x)dx = 1$ and $\int_{-\infty}^{+\infty} xk(x)dx = 0$ and it is expected to have a global maximum at $x = 0$ (Orzeszko, Bejger, 2018). A popular choice for the kernel function is the Gaussian kernel, defined as:

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{8}$$

Next, for the estimated trend model, residuals e_i (for $i = 1,2, \ldots, N$) are calculated as follows:

$$e_i = y_i - \hat{y}_i, \tag{9}$$

where $\hat{y}_i$ represents the theoretical values from the model, i.e.:

$$\hat{y}_i = f(t_i). \tag{10}$$

The final updated prices y_i^* are computed as:

$$y_i^* = f(0) + e_i \tag{11}$$

for each i.

Step 7. This stage plays a crucial role in the proposed algorithm, as it uses the previously completed steps to estimate the price of the user-specified property $\overline{x}'_0$. This estimated price (denoted $\hat{y}_0$) is calculated as:

$$\hat{y}_0 = f(\bar{x}'_0), \tag{12}$$

where f is a regression function determined on the basis of the nearest neighbors $\overline{x}'_i$ of the vector $\overline{x}'_0$ and their updated prices y_i^*. The Nadaraya-Watson estimator is proposed to estimate the regression function f. In this method, the value $f(\overline{x}'_0)$ is calculated as:

$$f(\bar{x}'_0) = \sum_{i=1}^{N} w_i(\bar{x}'_0) y_i^*, \tag{13}$$

where the $w_i(\overline{x}'_0)$ weights depend on the considered $\overline{x}'_0$ and are calculated using:

$$w_i(\bar{x}'_0) = \frac{k\left(\frac{d(\bar{x}'_0, \bar{x}'_i)}{h}\right)}{\sum_{i=1}^{N} k\left(\frac{d(\bar{x}'_0, \bar{x}'_i)}{h}\right)}, \tag{14}$$

where k is the kernel function, h is the bandwidth, and $d(\overline{x}'_0, \overline{x}'_i)$ is the Gower distance between vectors $\overline{x}'_0$ and $\overline{x}'_i$. This means that the estimated price $\hat{y}_0$ is a weighted average of the updated prices y_i^* of the nearest neighbors of the property $\overline{x}'_0$ with weights increasing as the similarity (measured by the Gower distance) between each property and $\overline{x}'_0$ increases.

Step 8: The purpose of this stage is to assess the quality of the estimated price $\hat{y}_0$. This evaluation consists of two aspects:

(a) Measuring the fit of the N-W model to the data on which it was built,

(b) Estimating the ex-ante prediction error of the N-W model.

Part (a): To measure the fit of the N-W model, the model's residuals and their standard deviation are calculated. The residuals ε_i (for i = 1,2, ..., N) are computed as follows:

$$\varepsilon_i = y_i^* - \hat{y}_i^*, \tag{15}$$

where $\hat{y}_i^*$ is the theoretical value for the updated price y_i^*, calculated using the formula:

$$\hat{y}_i^* = f(\bar{x}_i'). \tag{16}$$

To obtain $\hat{y}_i^*$ (for i = 1,2, ..., N), formula (13) is used, along with formula (14), substituting $\bar{x}_i'$ in place of $\bar{x}_0'$.

The standard deviation of the model residuals, S_ε, is calculated using:

$$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})^2}{N}}, \tag{17}$$

where $\bar{\varepsilon} = \frac{\sum_{i=1}^{N}\varepsilon_i}{N}$.

The standard deviation provides a measure of the model's accuracy by indicating how closely the model-predicted prices $\hat{y}_i^*$ match the observed (updated) prices y_i^*.

To evaluate the N-W model's fit, the following metric is proposed:

$$\left(1 - \frac{S_\varepsilon}{\bar{y}^*}\right) \cdot 100\%, \tag{18}$$

where $\bar{y}^*$ is the mean of the updated prices, calculated as:

$$\overline{y^*} = \frac{\sum_{i=1}^{N} y_i^*}{N}\ . \tag{19}$$

The closer this metric is to 100%, the better the fit of the constructed model.

Part (b): The absolute ex-ante prediction error (denoted σ_{pred}) is calculated as:

$$\sigma_{pred} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i^* - f_{(-i)}(\bar{x}_i')\right)^2}{N}}, \tag{20}$$

where $f_{(-i)}(\bar{x}_i')$ (for i = 1,2, ..., N) is computed similarly to $f(\bar{x}_0')$ as defined in formula (13). However, in this case, $\bar{x}_i'$ replaces $\bar{x}_0'$, and all calculations are performed on a dataset that excludes the i-th transaction. Specifically, for each $\bar{x}_i'$, to determine $f_{(-i)}(\bar{x}_i')$, the weights must be recalculated with $\bar{x}_i'$ substituted for $\bar{x}_0'$ in formula (14).

The relative ex-ante prediction error (denoted V_{pred}) is then calculated as:

$$V_{pred} = \frac{\Sigma_{i=1}^{N} \left| \frac{y_i^* - f_{(-i)}(\bar{x}_i')}{y_i^*} \right|}{N} \cdot 100\%' \ . \tag{21}$$

The lower the calculated prediction errors, the greater the model's predictive power, or its ability to generate accurate predictions.

Compared to other automatic valuation systems, the advantages of the presented algorithm include:

a) Robustness to errors and missing data;
b) Flexibility in adapting to the nature of the data, i.e., the ability to account for variables on different measurement scales;
c) The ability to consider multiple factors affecting the valuation, while simultaneously allowing them to be weighted differently;
d) Use of methods that account for nonlinearity in the data;
e) Utilization of nonparametric methods, which do not require as restrictive assumptions about the data's properties as classical econometric models, nor assumptions about the nature of the modeled relationships;
f) Scalability, meaning the ability to apply the algorithm to databases of various sizes;
g) Use of methods and tools robust to computational errors (e.g., issues such as algorithm convergence failure, insufficient degrees of freedom, or matrix singularity).

At the same time, it is worth noting that the nonparametric forecasting methods used in the algorithm, despite their numerous advantages, also have their drawbacks. Most importantly, compared to classical (linear) econometric models, they are characterized by longer computation time and poorer interpretability and explainability. However, the literature on the subject highlights the many advantages of this approach, which has led to growing interest from both the scientific community and the business sector.

4. Conclusion

Real estate valuation involves considering many factors that contribute to the final result. The factors describing properties are of a very diverse nature and are measured on different scales. Furthermore, relationships within the data can be nonlinear, and price changes in the real estate market are dynamic and difficult to predict. Therefore, the problem of real estate valuation is not trivial and requires solutions that can handle the various challenges of modeling and forecasting such data. This article presents a real estate valuation algorithm using nonparametric regression methods, specifically the

local-linear kernel and Nadaraya-Watson estimators. The proposed algorithm consists of eight stages, which aim to: create dynamic variables, classify variables, select "valid" records, normalize variables, choose the nearest neighbors, update the prices of selected neighbors, estimate the value of the property being appraised, and assess the quality of the valuation. The presented algorithm is characterized by universality and the ability to account for many aspects necessary for accurate real estate valuation. Its main advantages include errors and missing data robustness, the ability to account for nonlinearity in the data, the ability to use a wide and diverse set of property characteristics, and the potential for use in various market conditions. Additionally, the algorithm is universal because it can be easily adapted to the specifics of a particular real estate transaction database.

References

Bełej, M., Figurska, M. (2017). Teoretyczne aspekty stosowania automatycznych modeli wyceny. *Biuletyn Stowarzyszenia Rzeczoznawców Majątkowych Województwa Wielkopolskiego* (1–2).

Bryx, M. (2006). *Rynek nieruchomości. System i funkcjonowanie.* Warszawa: POLTEXT.

Cox, T.F., Cox, M.A.A. (2000). A general weighted two-way dissimilarity coefficient, Journal of Classification, 17, 101–121.

Dydenko, J. (2006). *Szacowanie nieruchomości.* Dom Wydawniczy ABC.

Fan, J., Gijbels, I. (1992). Variable Bandwidth and Local Linear Regression Smoothers. Annals of Statistics, 20, 2008–2036.

Fan, J., Gijbels, I. (1996), Local Polynomial Modeling and Its Applications, London: Chapman and Hall.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties, „Biometrics", 27, 857–871.

Gusta, A. (2023). Instytucjonalne uwarunkowania determinujące proces wyceny nieruchomości w Polsce. *praca doktorska*. Łódź.

Härdle W., Lütkepohl H., Chen R. (1997). A Review of Nonparametric Time Series Analysis, International Statistical Review, 65, 49–72.

Kałkowski, L. (2003). *Rynek nieruchomości w Polsce.* Warszawa: TWIGGER.

Keogh, G., D'Arcy, E. (1999). Property Market Efficiency: An Institutional Economics Perspective, Property Market Efficiency: An Institutional Economics Perspective. *Urban Studies, 36*, 2401–2414.

KNF (2019). *Rekomendacja S dotycząca dobrych praktyk w zakresie zarządzania ekspozycjami kredytowymi zabezpieczonymi hipotecznie.* https://www.knf.gov.pl/dla_rynku/regulacje_i_praktyka/rekomendacje_i_wytyczne/rekomendacje_dla_bankow.

KNF (2023). *Rekomendacja J dotycząca zasad gromadzenia i przetwarzania przez banki danych o rynku nieruchomości.* https://www.knf.gov.pl/dla_rynku/regulacje_i_praktyka/rekomendacje_i_wytyczne/rekomendacje_dla_bankow.

Konowalczuk, J. (2014). *Wycena nieruchomości do celow kredytowych.* Warszawa: POLTEXT.

Kowalewski, A., Markowski, T. i Śleszyński, P. (2018). *Studia nad chaosem przestrzennym* (Tom III). Warszawa: PAN, Komitet Przestrzennego Zagospodarowania Kraju.

Kucharska-Stasiak, E. (2014). Wpływ współczesnych koncepcji wyceny na metodykę wyceny nieruchomości w Polsce. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania, 1*(36), s. 101.

Kucharska-Stasiak, E. (2016). *Ekonomiczny wymiar nieruchomości.* Warszawa: PWN.

Kucharska-Stasiak, E. (2018). Dysfunkcje na rynku nieruchomości w warunkach. *Bank i Kredyt* (49(5)), strony 493–514.

Malmon, M. (2022, June). The future of the profession and valuation standards. *European Valuer*(26).

Matysiak, G. (2017). Automated Valuation Models (AVMs). *Conference in Finance.* Wrocław.

Matysiak, G. (2018). *TEGoVA Wiosenne Walne Zgromadzenie, Ateny, Grecja, 18–20 październik 2018.* Pobrano z lokalizacji TEGoVA www.tegova.org: www.pfsrm.pl.

Nadaraya, E. A. (1964). On Estimating Regression. Theory of Probability and its Applications, 9, 141–142.

Orzeszko, W. (2018a), Prognozowanie indeksu WIG za pomocą jądrowych estymatorów funkcji regresji, Bank i Kredyt, 49 (3), 253–288.

Orzeszko, W. (2018b), Wybrane aspekty nieparametrycznego prognozowania nieliniowych szeregów czasowych, Przegląd Statystyczny, 65 (1), 5–22.

Orzeszko, W., Bejger, S. (2018), Nonparametric prediction of indices from the Central European stock exchanges, 36th International Conference on Mathematical Methods in Economics – Conference Proceedings, 366–371.

PAP. (2022). *https://samorzad.pap.pl/kategoria/aktualnosci/pan-obliczyla-wskaznik-pokrycia-planistycznego-dla-kazdej-z-2477-gmin-tabela*. Pobrano z lokalizacji https://samorzad.pap.pl/kategoria/aktualnosci/pan-obliczyla-wskaznik-pokrycia-planistycznego-dla-kazdej-z-2477-gmin-tabela.

Peyman, J., Davood, S., Abbas, R., Tuan, N. (2024). Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors. *Cities*(151).

PFSRM. (2012). *Standardy zawodowe Polskiej Federacji Stowarzyszeń Rzeczoznawców Majątkowych.* Pobrano z lokalizacji Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych: https://pfsrm.pl/aktualnosci/item/14-standardy-do-pobrania.
PFSRM, (2017). https://pfsrm.pl/aktualnosci/item/480-nowy-europejski-standard-dot-automatycznych-modeli-wyceny.
Rącka, I. (2024). Valuing in nontransparent markets. *European Valuer Journal* (32).
Real Estate Management, [Act of August 21, 1997] (Dz.U. 2024, vol. 1145.) (Poland).
RICS. (June 2021). *Automated Valuation Models Roadmap for RICS members and stakeholders.* Royal Institution of Chartered Surveyors (RICS). Pobrano z lokalizacji www.rics.org.
Siemińska, E. (2013). *Ryzyka inwestowania i finansowania na rynku nieruchomości w kontekście etyki i społecznej odpowiedzialności.* Toruń: Uniwersytet Mikołaja Kopernika.
Sing, T., Jingye Yang, J., Ming Yu, S. (2022). Boosted Tree Ensembles for Artificial Intelligence Based. *The Journal of Real Estate Finance and Economics*(vol. 65, issue 4, No 5), 649–674.
Sobczyk, M. (2024). Statystyka, Wydawnictwo Naukowe PWN, Warszawa
Stone, C. J. (1977). Consistent Nonparametric Regression. Annals of Statistics, 5, 595–620.
Śleszyński, P. (2023). Zagospodarowanie przestrzenne. *https://www.youtube.com/watch?v=jbqOSJQgheY*.
TEGoVA. (2020). *Europejskie Standardy Wyceny* .
Trojanowski, D. (2019). *Dylematy wyceny nieruchomości komercyjnych w Polsce.* Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.
University of Oxford Research. (2022). *The future of automated.* Pobrano z lokalizacji www.sbs.oxford.edu
Valuation of real estate (Act of September 5, 2023) (Dz.U. 2023 vol. 1832) (Poland).
Vilar-Fernández, J.M., Cao, R. (2007). Nonparametric Forecasting in Time Series—A Comparative Study. Communications in Statistics – Simulation and Computation, 36(2), 311–334.
Walesiak, M. (2016), Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Watson, G. S. (1964). Smooth Regression Analysis. Sankhya: The Indian Journal of Statistics (Series A), 26, 359–372.
Załęczna, M. (2010). *Instytucjonalne uwarunkowania rozwoju rynku nieruchomości w Polsce na tle doświadczeń państw zachodnich.* Łódź: Uniwersytet Łódzki.

Acknowledgements

Author contributions: authors have given an approval to the final version of the article. Author's total contribution to the manuscript: Marcin Fałdziński (35%); Witold Orzeszko (35%); Ewa Siemińska (30%).