AUTOMATED EVALUATION MODELS in real estate market: A comparative analysis between linear regression and XGBoost

Silviu-Ionuț BĂBȚAN [1]ORCID ID: 0009-0004-3496-7794

Abstract: *The dynamic trend of the real estate market, as well as the rapid digitization process that has taken place recently, has created the need to identify techniques for automated property price valuation. This article aims to eliminate the problems of subjectivity and high resource consumption of traditional valuation by analyzing the practical implementation of automated valuation methods in the real estate market. Thus, the article performs a comparative analysis between classical prediction methods, such as linear regression, and complex prediction algorithms, such as XGBoost, which can be applied in real estate value prediction. By analyzing the accuracy of predictions on a sample of apartments in Cluj-Napoca, Romania, the study compares the two automatic valuation methods and draws the lines of their implementation. The results of this study can be used for the development of automatic valuation techniques at a global level.*

KEY WORDS

Real estate, mass appraisals, automated appraisal, valuation standards.

JEL Classification: C88 Other Computer Software; R31 Housing Supply and Markets; R32 Other Spatial Production and Pricing Analysis. R 30 General Real Estate Markets, Spatial Production Analysis, and Firm Location-

INTRODUCTION

In the context of the modern real estate market, which is characterized by dynamism and rapid fluctuations, real estate appraisal methods have evolved significantly to meet the new demands. Traditional valuation methods, which rely on physical inspection and subjective analysis by the appraiser, have significant limitations in terms of efficiency and cost (McCluskey & Anand, 1999). These methods are time and resource consuming, as they involve human appraisers, which can lead to discrepancies in final valuations influenced by subjective factors (Pagourtzi et al., 2003). Alternatively, the rising need in the real estate sector for swift and budget-friendly solutions is propelling Automated Valuation Models (AVMs) to become more prevalent, gradually surpassing traditional methods (Shiller, 2015).

AVMs offer clear advantages using advanced algorithms and the processing of large volumes of data in a short time. Unlike the traditional approach, which involves manual inspection of each property,

[1] Accounting and Audit Department, Faculty of Economics and Business Administration, Babeș-Bolyai University, Cluj-Napoca, Romania Email: silviu.babtan@econ.ubbcluj.ro
DOI: 10.29302/oeconomica.2024.2.3

automated models use artificial intelligence techniques to evaluate properties based on complex sets of economic and demographic variables, thus providing fast and consistent results (Rossini, 1999; Bourassa et al., 2010).

In addition, automated models eliminate much of the costs associated with traditional appraisals, which involve appraisers' physical labor and travel to the property location. By reducing direct human involvement in the process, AVMs help to lower costs and increase affordability for clients (Borst & McCluskey, 2008; Kummerow, 2000). Moreover, automated appraisals have the advantage of being more objective because they are based on standardized data and algorithms, thereby reducing the risk of human error or subjective influences (Nguyen & Cripps, 2001).

As the real estate market becomes increasingly complex, AVMs offer an innovative solution that significantly improves the efficiency of the valuation process and provides greater transparency and objectivity in establishing property values (Ho et al., 2012; Geltner et al., 2007). These models contribute to a faster and more accurate flow of information, dynamically adapting to the demands of a changing real estate market.

The degree of acceptance of the AVM around the world varies and the model is still blamed for certain weaknesses. AVM is very widespread in use in the US, a context compared to which specialists are calling for more transparency and trust in AVM models for the European space (Matysiak, 2017).

The study we propose has two objectives and deals with the issue of automated valuation methods (abbreviated in this paper as AVM), which in international parlance stands for AVM - Automated Valuation Model. First, the study aims to demonstrate the usefulness and practical applicability of automated valuation methods in the real estate market. And the second objective is to compare linear regression, a classical method used in the pediction process, with complex predictive methods such as XGBoost. Although specialized publications have promoted numerous such models, a new simulation - as we propose for the Romanian space - is beneficial. This is because it has been argued that the error induced by a given model, and hence its accuracy, differs with respect to market conditions, property types and countries. Therefore, the study aims to make a comparative analysis of the models' performance in the context of the Romanian real estate market. Therefore, this paper aims to analyze the automated valuation methods that can be applied to the real estate market and, depending on the performance, to present the most accurate valuation method. The sample used in the study includes one, two, three and four room apartments located in Cluj-Napoca.

We believe that, if it can be shown to be useful, with sufficient knowledge of its content to address some of its shortcomings, AVM would be of assistance to valuation practitioners. We are thinking in particular of practitioners involved in repetitive and standardizable valuations, such as credit guarantee valuations and tax valuations of residential properties, especially apartments.

1. LITERATURE REVIEW

The concept of the Automated Valuation Model (AVM) emerged as a trend in real estate appraisal towards the end of the 20th century. Previously, similar, if not identical, approaches were referred to as computer-assisted mass valuations. Today, professionals use the term automated valuation model to emphasize the high level of process automation, although these models generally still require some degree of human interaction.

Automated assessment models were first introduced in the United States in 1981. Following this, their development extended to the UK during the 1990s (Worzala et al., 1995).

Automated valuation models have been developed and improved for a variety of purposes, primarily to provide a quick and cost-effective alternative to a full assessment (Worzala et al., 1995).

Currently, AVMs are extensively used across the United States, Canada, Australia, the United Kingdom, and several EU member countries, including Denmark, Germany, and the Netherlands. Additionally, AVMs are seen as an effective tool for the European Central Bank's asset quality evaluations (Brucato, 2015), which may gradually encourage greater interest in other European Union countries.

The Automated Valuation Model integrates various algorithms, characteristics, and pricing data and ultimately provides an option for valuing real estate properties with a focus on residential (Galin et al., 2018).

These algorithms also encompass traditional multiple regression analysis (MRA), including techniques like hedonic analysis or ordinary least squares regression (OLS), which are commonly used and well-established methods for valuation technique, as noted by Kuburic et al. (2012). This approach emerged during the 1980s, aligning with the early expansion of information systems. The multiple linear regression model is applied to approximate real estate values, using numerous explanatory factors that aid in forecasting property value or price. Nonetheless, the MRA/OLS approach has faced criticism regarding its accuracy in property valuation, mainly due to challenges such as multicollinearity among independent variables and the inclusion of outliers within the sample, as noted by Worzala et al. (1995).

In this framework, the hedonic pricing regression model was developed based on the idea that imperfections exist within the real estate market, thus questioning the MRA/OLS assumption of a perfectly balanced market. Property prices observed may not truly represent market-clearing values or offer an unbiased estimate of fair market value. As noted by Samaha and Kamakura (2008), who cite various studies, imperfections in the real estate market involve factors like wealth, gender, demographic variables, and property uniqueness, all of which influence market engagement. Glower et al. (1998) included in their model both list prices and stochastic errors, which represent a combination of the optimal price and any error by the seller when setting the listing price.Multiple regression has long been utilized as the conventional approach in the field. Typically, it involves modeling property attributes. Nonetheless, the traditional MRA method has notable limitations, which are widely recognized, such as the inability to effectively address variable interactions, nonlinearity, errors, and multicollinearity issues.

Newly emerging methods have become central to current AVM research, although their practical application remains inconclusive. One approach includes the use of Artificial Neural Networks (ANN), which, although met with some constructive skepticism (McGreal et al., 1998), demonstrates promise, as indicated by numerous studies (Peterson and Flanagan, 2009). Still, even proponents of ANN acknowledge aspects of its nature that may hinder its broader adoption (McCluskey et al., 2013).

To observe the way in which AVMs have been treated at the regulatory level, of standardization through specialized standards, we propose a comparative analysis of the main referential on the valuation of assets, which may be of interest also for the Romanian space.

A distinctive feature of an AVM is its ability to produce a market valuation using mathematical models. Therefore, AVMs should be developed by skilled market professionals, like appraisers, who apply statistical tools to analyze data and select the best market simulation for assessing location, market trends, and property characteristics from collected data. AVMs are designed to generate property value estimates at specific points in time, utilizing either historical or forecasted data, depending on client requirements.

The International Association of Assessing Officers (IAAO, 2003) defines an automated appraisal model as a mathematical software system that produces a market value estimate by analyzing location, market conditions, and property attributes based on collected data. The accuracy of an AVM depends on the quality of the data and the expertise of the individual who developed it (IAAO, 2003).

The Royal Institution of Chartered Surveyors' Automated Valuation Model Standards Working Group describes an AVM as employing one or more mathematical techniques to provide a value estimate for a specific property on a particular date. This estimate includes a measure of confidence in its accuracy and functions independently of human input once started (RICS, 2017).

It is important to note that these definitions of an AVM do not involve any direct participation of the valuer in determining the real estate property's value estimate.

1.1 Impact of using AVMs

AVMs have been credited as being valuable for so-called "mass appraisal", which involve large databases that provide consistent input data and generate in an automated, rapid and generalized manner the coordinates for the rapid valuation of standard assets comparable to those included in the databases. Specifically, in relation to the purpose of valuation, we believe that an empirically grounded version of the AVM would be useful for credit guarantees and real estate taxation.

Bogin and Shui (2018) emphasize the importance of AVMs from the perspective of financial institutions, especially banks. Focusing on the U.S. context, they note that appraisers often tend to inflate collateral valuations in favor of bank clients. Consequently, banks gain from the development of AVM models, which allow them to swiftly execute assessments and verify estimates given by the appraiser. In addition, the same authors emphasize, AVM can generate credit risk predictions, being superior in accuracy for this purpose to the hedonic models that have dominated appraisal research and practice over the last decade. Bellotti (2017) takes the same view on the role of AVMs in valuations for bank collateral, including through the possibility offered to update the value of bank collateral over the life of the loan. In addition, the study argues for the positive impact of AVMs that ensure the accuracy of valuations for real estate buyers and other investors, for the valuation of securities portfolios, or for checking valuations for misrepresentation or even fraud.

Undoubtedly, the primary use of AVMs is found within the mortgage sector, where they are utilized independently for risk assessments, evaluating current loans, refinancing, and conducting transaction appraisals during new home loan negotiations. Their adoption in mortgage lending is increasing, supported by systems that facilitate instant mortgage applications and related tools (Catt, 2007).

Many users, when they see an appraisal report, wonder how realistic the property price is and how much they can trust the appraiser's estimated value. This is also because of errors that can occur (e.g. not accounting for an impact variable) or the valuation method, which can lead to misrepresentation of the appraised property price. Another utility is to increase the credibility of the users of valuation reports (sellers, buyers, banks, judges) by helping appraisers to assert their skills and better choose their professional methodology (Gupta et al., 2020).

Specifically, in relation to the purpose of the evaluation, we believe that an empirically grounded version of the AVM would be useful for credit collateralization and taxation, in line with the view expressed by Kauko and d'Amato (2008).

Darrin Benhart from the US Office of the Comptroller of the Currency (OCC) expresses concerns regarding the choice of appraisal firms and their reliance on associated AVMs. He notes that while appraisal products (AVMs) comply with guidelines, they lack foundational elements necessary for bank lending when not paired with a traditional appraisal or analysis. This focus on speed and cost-effectiveness raises broader concerns about the current application of AVMs, with some experts even suggesting that AVMs significantly contributed to the events leading up to the 2007 financial crisis (Mooya, 2017).

Unsurprisingly, much of the debate revolves around the transparency issues associated with AVMs, with calls for more thorough testing and auditing. There is a strong emphasis on the need for modeling specifications to be clear and transparent, particularly if they are subject to legal scrutiny (Appraisal Institute of Canada, 2002). While increased regulation has not yet resulted in greater transparency, it has led to the emergence of new hybrid products that integrate AVMs with property condition reports and inspections. In these hybrid models, all pertinent information is directed to a licensed appraiser, who then exercises judgment to determine the most suitable valuation method (Dickstein, 2014).

Rating agencies, including Fitch and Standard & Poor's, commonly apply 'adjustments,' meaning they decrease estimated AVM values by a certain percentage when providing their default risk assessments. They express opposition to value estimates from rating professionals that are not adjusted in this manner (Downie & Robson, 2007). However, with the increasing utilization of AVMs in the U.S., the frequency of these adjustments has declined, attributed to the enhanced performance of AVM models and the assumption of improved overall risk awareness associated with AVM usage (Downie & Robson, 2008). Another possible explanation for the reduction in AVM-specific adjustments is the implementation of general adjustments following the mortgage crisis, irrespective of the valuation method employed (Gorton & Metrick, 2010).

Although the regulatory landscape may be viewed as restrictive, some AVM developers perceive an opportunity to design additional AVM tools aimed at aiding compliance with these regulations, including areas such as taxation, implicit loss modeling, bonds, and management of non-performing loans (Brucato, 2015).

These computer-driven quantitative methods offer the benefit of being systematic and efficient, thus diminishing the dependence on human labor in performing assessments (Tretton, 2007). By eliminating the human element, AVMs are believed to reduce potential errors stemming from subjective judgment.

In our view, AVMs could serve as a valuable tool for appraisers by simplifying the valuation process. AVMs are especially advantageous for residential properties, such as apartments, which are well-suited for such an automated system due to the presence of numerous similar properties within each urban area.

1.2 Presentation of technical and computational aspects of AVM. Linear regression and XGBoost

Automated Valuation Models are software applications that estimate the market values of real estate properties by analyzing market trends and characteristics of comparable properties, utilizing previously gathered and analyzed market data. The reliability of an AVM and the precision of its outputs hinge on both the amount and quality of the data employed in the valuation, as well as the expertise and training of the professionals who create the valuation model. The amount of data pertains to the sample size utilized in the evaluation (Oust et al., 2019).

These are software applications that employ various statistical and algorithmic techniques to analyze the relationship between the price and value of a residential property alongside its underlying characteristics. The goal is to generate an estimate of the property's market value. The methodologies employed vary among different AVM providers. In fact, some appraisers have access to multiple models, selecting the most suitable one for specific situations (Brucato, 2015).

The design and advancement of the information system necessitate the creation of a framework of rules that mimic the cognitive processes of human appraisers. Intelligent systems have leveraged this framework, allowing them to function similarly to human evaluators, thus facilitating property assessments (Kato, 2012).

The accuracy of an AVM depends on the quality of the data used and the skill of the individual who develops it. AVMs should be designed by appropriately qualified professionals in the market, such as appraisers and software engineers, who utilize statistical methods to analyze data and select the best simulation of market dynamics for evaluating location, market trends, and property characteristics based on previously collected information (Faishal et al., 2005).

Artificial intelligence methods, conversely, heavily rely on the effective selection of comparable properties to be used in the valuation process. This presents another possible source of error, as the availability of recent sales data is a statistical variable that can be influenced by its own sources of variation (Vandell, 1991).

Assessments generated by intelligent systems have the potential to substitute human evaluations. Within this framework, the accuracy of residential property valuations may also be affected by the users' expertise and the foundational standards that guide the valuation process (Bagnoli and Smith, 2009).

One of the primary methods used in quantitative estimation is linear regression. Although it is one of the oldest automated estimation techniques, it remains one of the most widely used and efficient techniques. This method, in addition to providing powerful and robust predictions, is available in most statistical software and is suitable for application in broad research domains (Wang, 2003). There are studies in the literature (Goundar S. et al. 2021, Sipos C. et al. 2008) that have demonstrated the usefulness and applicability of this method in the real estate market.

XGBoost, on the other hand, is a complex algorithm that builds on other techniques such as AdaBoost and decision trees (Carmona et al., 2019). The main advantage offered by this technique is the efficiency it provides, generating results with high accuracy in record time, up to 10 times faster than other prediction techniques (Shilong, 2021).

In contrast to other methods, in the XGBoost technique a sparsity-constrained mechanism is applied that handles even variables with missing values, removing them from the analysis and increasing model performance (Bentéjac et al., 2021).

Rossini and Kershaw (2008) highlight the competitive advantages of AVMs in terms of speed and cost-effectiveness, while carefully avoiding the idea that AVMs would intrinsically provide lower-quality evaluations compared to traditional methods. According to Rossini and Kershaw (2008), fast and economical methods do not necessarily equate to lower-quality results, despite the stringent stance some international regulatory authorities take toward AVMs based on the current selection available. The IAAO (2003), for instance, views AVMs comparably to professional appraisals, seeing no inherent shortcomings—emphasizing that AVM effectiveness depends, like any skill-based profession, regarding the quality of procedures and the expertise of the professionals engaged.

Mooya (2017) similarly concluded, after extensive study, that there are no theoretical or practical obstacles preventing AVMs from ultimately supplanting traditional valuation methods completely. However, although it is possible to develop high-quality AVMs that often exceed typical appraisal accuracy, many regulatory bodies remain cautious, limiting AVMs primarily to low-risk evaluations or as supplementary tools for certified appraisers (Appraisal Institute of Canada, 2002). Additionally, some organizations evaluate AVMs' standalone application potential but ultimately assign full responsibility to professional appraisers (Appraisal Institute of Canada, 2002).

3. Data and Methodology

3.1. Date

For the implementation of the AVM, information was collected on residential real estate properties, namely one-, two-, three- and four-bedroom apartments. These apartments are in Cluj-Napoca, Romania.

The real estate market in Cluj-Napoca ranks among the most vibrant cities in Romania. The average value of the properties in this city is more than 30% higher than the next city, the capital of the country - Bucharest.[2] All studies in the field emphasize with each new publication that the average sale price increases from one month to the next, even in times of crisis, such as Covid-19 (information for August 2021). This real estate market is very active due to the opportunities that this city offers its inhabitants[3] .

This influx of real estate purchases in Cluj-Napoca is largely because the incomes earned by residents are among the highest in Romania, Cluj County being the second highest in Romania, after Bucharest. Although statistics do not provide clear information only about Cluj-Napoca, which certainly exceeds the county average.[4]

The growing interest in this development pole stems from the demand of potential buyers from neighboring counties such as Alba, Bihor, Bistrița-Năsăud, Satu-Mare, Sălaj, Mureș and Maramureș. Thus, an important percentage of property buyers in Cluj-Napoca are residents of these counties, especially parents of students studying at the city's universities.[5]

The centralized database on which this research is based contains more than 27,828 pieces of information related to 773 apartments with 1, 2, 3, and 4 bedrooms.

[2] https://www.imobiliare.ro/indicele-imobiliare-ro/cluj napoca?cq_src=google_ads&cq_cmp=15381765798&cq_term=&cq_plac=&cq_net=x&cq_plt=gp&gclid=EAIaIQobChMIw5uSqrb49gIVCbTVCh0UyAFOEAAYASAAEgKw8fD_BwE

[3] http://ancpi.ro/index.php/presa-3/statistici

[4] http://statisticiromania.ro/clasamente

[5] https://actualdecluj.ro/cine-cumpara-apartamente-case-sau-terenuri-in-cluj/

The main source of data is the Argus platform[6] , for which information is collected from 700 apartments with 1, 2, 3 and 4 rooms, to which real estate agencies in Cluj-Napoca have access. Also, for a higher accuracy, information was collected directly from 73 evaluation reports, realized by the evaluation company Napoca Business S.R.L.

The Argus platform is designed exclusively for real estate agents, through which you can access all available real estate offers in real time. The online platform has a monthly usage fee and is most used by real estate agencies because it includes all the offers on the online market for real estate for sale.[7]

The study period covered in the current research is the period between the first semester of 2019 and the second semester of 2020, inclusive, and covers those offers for sale of apartments, which contain information about all the characteristics that will be presented below.

As data collection methodology, among the more than 8,000 advertisements of apartments for sale available on the Argus platform, only those advertisements for which all the necessary information to complete the database was presented were selected. Thus, a listing was entered in our database if data for all 36 variables were available. Due to our desire to have a database as comprehensive as possible in terms of the characteristics influencing the price of a real estate property, more than 90% of the ads analyzed were not included in the present research. We should also mention that during the collection of this information we were confronted with some situations that led to a reduction in the number of selected apartments. Some apartments were identified in several advertisements, so we did not want to duplicate the information.

The variables considered in the research *Table 1* below, totaling 36 quantitative and factor type attributes. These variables are the most common variables in valuation reports produced by professional/licensed appraisers using the classical appraisal method. It should be noted that all these properties selected in the database are suitable and have all the necessary characteristics to be included in the database. These data collected from the Argus platform have also been checked on the web pages, so that the information is true.

We will list below the 36 characteristics related to the 773 apartments, and explain them in the table below, namely: neighborhood, area, total_price, type, number_rooms, number_bathrooms, number_bathrooms, unfurnished, floor, number_floors_block, finish, concentration, central_proper, windows_thermopan_panel, door_metal, balcony, parking_place, garage, elevator, thermal_insulation, distance_to_center, compulsory_education_institutions, higher_education_institutions, square, supermarket, shopping_center, common_transport_lines, financial_institutions, religious_institutions, business_center, green_area, relaxation_places, hotels, sports_spaces, culture, noise_pollution.

Variables can be characterized either quantitatively or qualitatively. Quantitative variables take numerical values. Examples include total_price, area, number_rooms and so on. Qualitative variables, on the other hand, take different values or categories. Examples of qualitative variables include places_relaxation, pollution_mica, culture, etc. We refer to problems with a quantitative answer as regression problems, and those involving a qualitative answer are often referred to as classification problems.

[6] www.argus.me
[7] https://www.mediapress.ro/argus-reo-imobiliare.php

Table 1. Description of variables

Variable	**Variable description**	**Variable type**
Neighborhood	The neighborhood variable represents the area where the analyzed apartment is located. Thus, the municipality of Cluj-Napoca was divided into 8 main neighborhoods: Mănăștur, Mărăști, Central, Iris, Gară, Bună_ziua, Gheorgheni, Grigorescu.	Categorical
Surface	The area is expressed in the number of usable square meters of the apartment.	Quantitative
Total_price	The sale value of the apartment, expressed in Euro.	Quantitative
Tip	It represents the type of construction: old (before what year?) or new.	Categorical
Number_rooms	The number of 1, 2, 3 or 4 rooms the apartment has.	Quantitative
Floor_number	Number of toilets in the apartment (1 or 2)	Quantitative
Decomandat	If the apartment is a fully furnished apartment, like: Yes or No	Categorical
Floor	Positioning of the apartment according to the floor on which it is located, such as: Ground floor, 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.	Quantitative
Total number of floors	Represents the number of floors in the block. The states that the variable can have are ground floor, 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.	Categorical
Finishing	Unfinished, semi-finished, finished and ultra-finished	Categorical
Concentration	Such as large, medium or large.	Categorical
Heating system	If the apartment has a central heating system, like: Yes or No.	Categorical
Thermal_glass	If the apartment has trmopan windows, like: Yes or No.	Categorical
Metal_door	If the apartment has a metal door, like: Yes or No.	Categorical

Variable	**Variable description**	**Variable type**
Balcony	If the apartment has a balcony, like: Yes or No.	Categorical
Parking_place	If the apartment has an assigned parking space or garage, such as: Yes or No.	Categorical
Box	If the apartment has a box/booth, such as: Yes or No.	Categorical
Lift	If the block where the apartment is located has an elevator, such as: Yes or No.	Categorical
Thermal_insulation	If the block where the apartment is located has thermal insulation, such as: Yes or No.	Categorical
Distance_center	Measured in km from city center.	Categorical
Compulsory_education_institutions	If the apartment has compulsory education institutions nearby, such as: Yes or No.	Categorical
Higher_education_institutions	If the apartment has nearby higher education institutions, such as: Yes or No.	Categorical
Market	If the apartment has a food market nearby, such as: Yes or No.	Categorical
Supermarket	If the apartment has a supermarket nearby, like: Yes or No.	Categorical
Shopping_Center	If the apartment has a shopping center nearby, such as: Yes or No.	Categorical
Common_transport_lines	If the apartment has nearby public transportation lines, such as: Yes or No.	Categorical
Financial_institutions	If the apartment has nearby financial institutions, such as: Yes or No.	Categorical
Healthcare_institutions	If the apartment has nearby health facilities, such as: Yes or No.	Categorical
Religious_institutions	If the apartment has nearby religious institutions, such as: Yes or No.	Categorical
Business_center	If the apartment has nearby business center, like: Yes or No.	Categorical

Variable	**Variable description**	**Variable type**
Green_Zone	If the apartment has nearby green area, like: Yes or No.	Categorical
Places_relaxation	If the apartment has nearby parking spaces, such as: Yes or No.	Categorical
Hotels	If the apartment has nearby hotels, like: Yes or No.	Categorical
Sports_spaces	If the apartment has sports facilities nearby, such as: Yes or No.	Categorical
Culture	If the apartment has cultural institutions nearby, such as: Yes or No.	Categorical
Chemical_pollution	If the apartment is in an area with low pollution	Categorical

For the *concentration* feature we used the Google Maps search engine, considering the number of properties per 500 square meters, and of course the number of floors of the properties (the more floors the properties have, the higher the concentration). The concentration of the buildings significantly influences the living experience through access to jobs, public transportation, or even noise levels.

Concentration can be calculated as follows:

Concentration = Total number of properties within a 500 meter radius/ Total built-up area (s.p.) within a 500m radius, where:

• Total number of properties: Represents the total number of individual buildings or units identified within a 500 meter radius.

• Total built-up area: Is the sum of all built-up areas of buildings (the footprint area multiplied by the number of floors).

We also calculated the *center_distance (km)* feature using the tools provided by this search engine.

For the features *compulsory_and_higher_education_institutions, markets, supermarket, shopping_center, financial_institutions, health_institutions, religious_institutions, business_center, green_areas, hotels, sports_areas, culture and noise_pollution* we used the same map generated by Google Maps, taking as reference approximately 500 linear meters in front of these features that the apartment evaluated has. The presence of supermarkets in the apartment's proximity is a key amenity that influences day-to-day life. They offer access to essential goods, satisfying immediate consumer needs.

If we turn our attention to Table 2 above, we can describe certain characteristics, namely:

- *The reporting date* is May 2020, the month in which all this information was sifted and processed;
- from the analysis we can conclude that certain *neighborhoods*, such as Mărăști and Gheorgheni, are the most common in the advertisements, while Grigorescu is the least common;
- *total_price* is directly related to the *useful floor area of* the building;

- the most common *surface area* for 1-bedroom apartments is between 25 and 40 sq.m., for 2-bedroom apartments between 45 and 60 sq.m., for 3-bedroom apartments between 55 and 70 sq.m., and for 4-bedroom apartments between 65 and 90 sq.m.;
- more than 60% of the apartments are part of old buildings built before 2000, even 1990;
- *the subdivision* <<decomandat>> has a proportion of 70%;
- 65% of the buildings to which the identified apartments belong are *4-storey*;
- Over 90% of these buildings are located in areas with a high *concentration of* properties;
- over 90% have their own *central heating;*
- only 20% of apartments have a *parking space*;
- *The average distance from the city center* is 6 km and more than 95% of the buildings have *public transport* facilities within walking distance (10 - 15 minutes);
- over 60% have a *green area/park/recreation area* nearby; and the *level of pollution* (noise, dust, etc.) is average.

The amount of information initially collected was considerably higher, but redundant or incomplete information was removed. More specifically, if an ad was missing one of the 36 characteristics mentioned in the previous section, it was removed from the database. To provide a more relevant and detailed picture of the real estate market in Cluj-Napoca, a single database was centralized using both sources. The inclusion of data from both sources is crucial to create a diverse, accurate and relevant dataset for the context studied. This aggregation of results allows both to capture the dynamics of the market and to provide an objective assessment.

3.2. Methodology

In the subsequent sections, we will outline the methodology employed in the case study, presenting both the data processing stage and the application stage of the automatic property evaluation methods we propose.

The next step in our analysis was to process the dataset so that categorical variables with yes and no states were transformed in numeric format into dummy variables with values 1 and 0. This step was necessary to provide numeric inputs to the automatic evaluation methods. The majority of the statistical methods, such as learning algorithms or linear multiple regression, require, as mandatory, numerical inputs. Subsequently, using the code language, all categorical variables in the database were transformed into factor variables.

After the process of data collection and processing, the database was divided into two data sets: the learning set, which comprises 80% of the information in the initial database, and the test set, which comprises the remaining 20% of the information in the database. The splitting of the information was done randomly (random split), directly in the RStudio computer program, without selection criteria, to ensure a balanced distribution of the data. A data validation technique was also applied for each method.

Linear regression and XGBoost have been applied in turn on the database processed according to the above-mentioned steps to predict the real estate prices. The program used to run the predictions is Rstudio, version 4.1.1. The aim behind the application of these methods is to identify the best performing technique. Therefore, after applying each automatic evaluation method listed above, the results obtained were compared to identify the most accurate and appropriate method. The performance comparison metric used in the study is the Root Mean Squre Error (RMSE). The use of this metric allowed the determination of the model that provides the highest accuracy of real estate property values in the context of the market in Cluj-Napoca, Romania.

Starting from the data processing step and splitting the dataset, data validation was applied using the 5-fold cross-validation method within both evaluation methods.

Linear regression is based on the following function *(*Wooldridge, 2019):

$Y=\beta_0 +\beta_1 \times X +\beta_{12} \times X_{2+\cdots+} \beta n \times Xn+ \epsilon$, *where: (1)*

Y- the dependent variable, i.e. Price,

β_0 *-the constant term, i.e. the base value of Y when the other variables are 0,*

β_{1-n} *-coefficients of the independent variables,*

X_{1-n} *-independent variables that influence the values of Y,*

ϵ*- the error term in the model.*

The regression function started from an initial model containing the variables Neighborhood, Area, Distance to Center and Parking Space. To obtain the best possible final model, an iterative selection method was applied to the initial model, which involves adding one variable at a time and evaluating the resulting models according to the F statistic. Thus, a single variable is added to the initial model, it is compared which of the created models generates the best performing model, and according to this criterion, the variable is incorporated into the original model. Finally, the final model was selected that generates the highest F statistic result. This iterative method allows the creation of the most robust model possible, including only those variables that contribute to increase the accuracy of the price assessment.

The complex XGBoost prediction algorithm is based on decision trees, and the formula used in the method (Chen et al.,2016) is as follows:

$f(x) = \sum_{i=1}^{k} fi(x)$, *where: (2)*

$f(x)$*- the final model,*

k- number of decision trees,

$fi(x)$*- prediction made by the first tree.*

For the XGBoost application, a grid of hyperparameters was created to determine the best combination of parameters for model efficiency (Table 2). 2,916 model training trials were performed with this step, at the end of which the combination of parameters generating the best result evaluated by the RMSE metric was displayed.

Table 2. Hyperparameters used in model training

Parameters	***Values***
nrounds	100, 200, 300, 400, 500, 600
max_depth	1,2,3,4,5,6
eta	0.01, 0.1, 0.3
range	0, 0.1, 1
colsample_bytree	0.5, 0.75, 1
min_child_weight	1
subsample	0.5, 0.75, 1

Source: Author's own processing

Subsequently, after the identification of the best performing combinations of parameters takes place, the process of training the final model takes place. Finally, the final model is used to predict the real estate value of the apartments included in the study sample.

The intervals for parameters were chosen to stabilize the model's efficiency and performance. For instance, nrounds: 100-600, the number of trees, includes simple models with boosting rounds to fit the data better. On the other hand, max_depth: 1-6, the maximum depth, controls the tree complexity to prevent overfitting to deeper trees that record more detailed patterns. We decided to exclude the 6 value to avoid overfitting and guarantee computational feasibility.

The other hyperparameter values were also designed to optimize the learning step and decrease the risk of overfitting. Eta: 0.01-0.3, the learning rate, harmonizes accurate learning with faster convergence and gamma (0-1), aiding in avoiding unnecessary complexity. Colsample_bytree: 0.5-1, the proportion of features, guarantees the model considers feature subsets to increase diversity while keeping the flexibility to include all features for precision. These intervals were established considering the standard practices and tailored to the specifics of our dataset in order to ensure efficient optimization.

Results and Discussions

Following the proposed analysis, significant results were obtained using both the classical linear regression method and the complex XGBoost algorithm. Although the two methods are based on different reasoning, both could be successfully applied to automatically evaluate the price of one, two, three and four-room apartments in Cluj-Napoca included in the database.

The table below summarizes the results obtained from running the regression, highlighting both the selected variables and their importance, as well as the performance characteristics of the model.

Table 3. Linear regression results

Predictor	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	*26395.75*	*5276.6*	*5.002*	*7.44E-07*	***
Surface	*1363.19*	*32.74*	*41.643*	*2.00E-16*	***
CartierCentral	*7415.23*	*4637.44*	*1.599*	*0.110346*	
CartierGara	*-10831.8*	*3330.23*	*-3.253*	*0.001208*	**
CartierGheorgheni	*68.9*	*2954.76*	*0.023*	*0.981405*	
CartierGrigorescu	*3334.15*	*4686.73*	*0.711*	*0.47711*	
CartierIris	*-11115*	*3774.14*	*-2.945*	*0.003353*	**
CartierManastur	*-2149.25*	*2540.42*	*-0.846*	*0.397875*	
CartierMarasti	*-3133.58*	*2542.98*	*-1.232*	*0.218336*	
Distance_center	*-2130.97*	*544.35*	*-3.915*	*0.000101*	***
Parking_place1	*6400.76*	*1675.41*	*3.82*	*0.000147*	***
Culture1	*4616.82*	*1586.82*	*2.909*	*0.003754*	**
Boxa1	*4887.35*	*1817.66*	*2.689*	*0.007369*	**
Number_bai	*5209.16*	*2366.8*	*2.201*	*0.028119*	*
Supermarket1	*-5690.49*	*2649.21*	*-2.148*	*0.032112*	*
Thermal_insulation1	*3250.13*	*1507.94*	*2.155*	*0.03153*	*
Concetratiemedie	*-9092.97*	*8416*	*-1.08*	*0.280379*	
Concetratiemica	*-6252.8*	*1936.87*	*-3.228*	*0.001313*	**

Residual standard error: 17780 on 603 degrees of freedom
Multiple R-squared: 0.8133
Adjusted R-squared: 0.8081
F-statistic: 154.5 on 17 and 603 DF, p-value: < 2.2e-16
Mean Absolute Error (MAE) for the final model: 13350.3

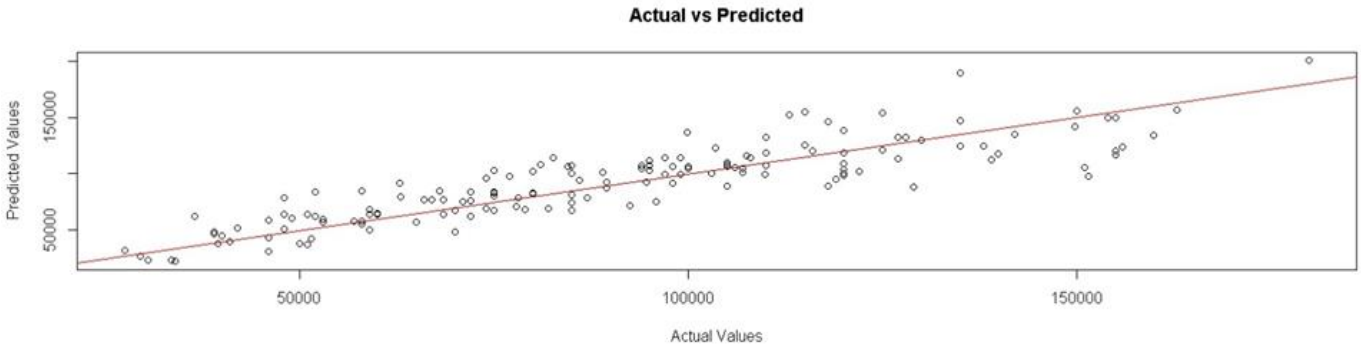
* Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' ' ' 1

From the above table out of the 32 variables initially included in the analysis, only 17 of them led to an increase in model performance and were included in the final model. Notable by an increased significance are the variables: Area, Distance to Center and Parking Space. Although the variables Average Concentration and Neighborhood with the states Mănăștur, Mărăști, Gheorgheni and Grigorescu are not statistically significant, they were kept in the model because they are multi-categorized variables and omitting them from the model would create inconsistency and an incorrect interpretation of the results.

At the same time, the model performance results show that there is a significant predictive capability of the data. The residual standard error of the model is 17780, which indicates that, on average, the estimation of the price of an apartment deviate by about 17780 from its real value.

The value of the degrees of freedom is 603, showing that the database has enough observations to identify trends in the model. Also, the F statistic has a value of 154.5, thus suggesting the overall high statistical significance of the price estimation model.

Linear regression: Estimated value vs. real value of prices



Source: Author's own processing generated with RStudio

The scatter plot presented above was generated using the code language in the RStudio program and captures the difference between the actual and estimated values for the real estate prices in Cluj-Napoca. As can be seen, there is a strong correlation between the values on the OX-axis and the values on the OY-axis, thus demonstrating that the model is successful in capturing the general trend in real estate prices.

The red-dashed line on the graph illustrates the alignment between the actual and estimated values, and the small distance between it and the points on the graph represent the errors in the model. We can therefore state that the high accuracy of the model is also observed from the visual analysis by a comparatively consistent distribution of the data points with respect to the red line.

In terms of XGBoost run results, the best combination of hyperparameters was achieved in trial number 888 out of 2916 runs. The combination of the maximum tree depth of 6, the learning rate of 0.01

with 600 iterations, the proportion of observations used of 0.5 and the proportion of 0.75 randomly selected variables yielded the lowest RMSE result of 17135.19.

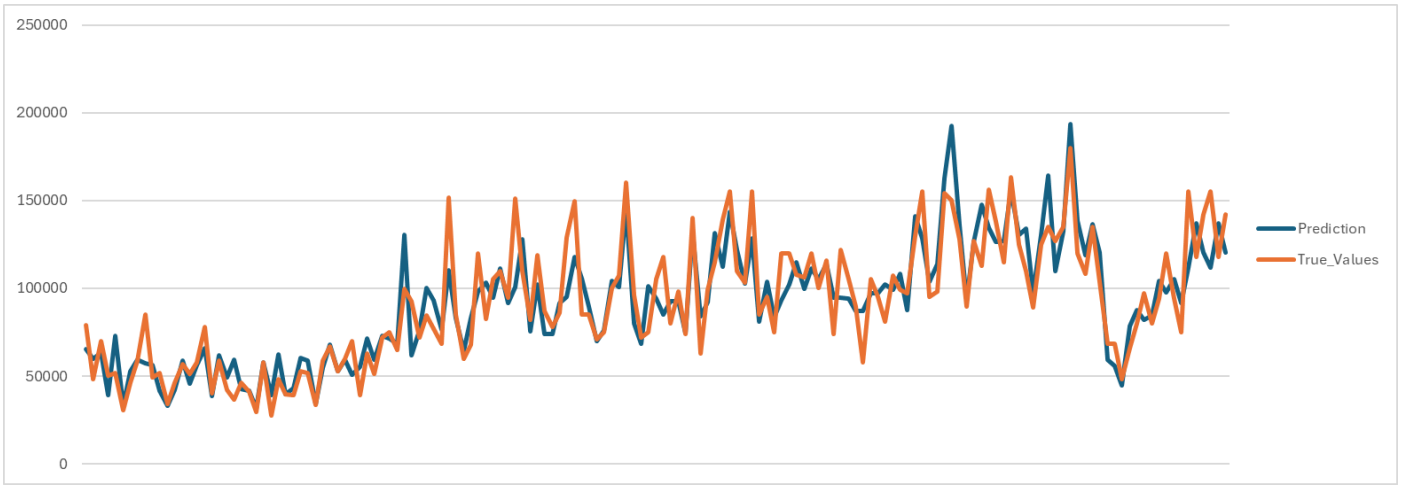
Table 4. Selected hyperpameters and model performance with XGBoost

eta	max_depth	range	colsample_bytree	min_child_weight	subsample	nrounds	RMSE	R^2	MFA	RMSESD	R^2 SD	SD
0.01	6	0.1	0.75	1	0.5	600	17135.19	0.8303	11403.3	2471.49	0.01916	828.3514

Source: Author's own processing

As can be seen from the above table, the resulting performance using the XGBoost algorithm is high. The R^2 metric has a value of 0.83 which indicates that the model explains about 83% of the variance. Also, the Mean Absolute Prediction Error (MAE) is at the level of 11403 units and the Standard Deviation (SD) is about 2471. The standard deviation value indicates that the prediction of prices varies from their mean by about 2471.The mean model prediction error (RMSE) is 17135.19, which, relative to the scale of the dependent variable data, indicates that the model is successful in generating good price estimates.

Graph 2. XGBoost: Estimated value vs. real value of prices



Source: Author's own processing

Figure 2 illustrates the actual vs. predicted value using the XGBoost price algorithm for 152 properties included in the database. For these apartments, the predicted price varies on average by 13.37% from the actual price. It is important to note that almost half of the predicted values have a rate of less than 10% variation from the actual price. At the same time, for almost a quarter of the apartments, the predicted value varied from the actual value by no more than 5 percent.

From the comparison of the performance of the two automatic price evaluation methods we can conclude that the XGBoost method is more suitable for the Romanian real estate context. Although both

applied methods had a high accuracy, both in terms of RMSE and MAE, the XGBoost algorithm managed to register a higher performance. While the linear regression estimates yielded an RMSE of 17780 and an MAE of approximately 13350, the XGBoost algorithm yielded an RMSE of 17135 and an MAE of approximately 11403. Although at first sight the difference in the performance of the two methods does not appear major, when applied in practice, any more accurate estimate influences the outcome of the valuation reports or even market conditions.

4. Conclusions

By examining the global use of AVMs, gathering insights from researchers, and noting regulatory interest in this model, our aim was to evaluate the practicality of its implementation in the Romanian real estate sector. To support the possibility of practical implementation of AVMs, this paper practically approached the estimation of real estate prices for 152 apartments in Cluj-Napoca.

The results of the analysis confirm that both classical automated valuation methods, such as linear regression, and complex algorithms, such as XGBoost, can be successfully applied. The comparative case study revealed that, for the dynamic market in Cluj-Napoca, complex valuation methods perform better than classical valuation methods.

In the past two decades, property valuation has transitioned from conventional approaches relying on comparable data to automated valuation models. This study, therefore, enables the implementation of an AVM for the Municipality of Cluj-Napoca, which can also be extended to other areas.

The constraints of automated valuation models are widely recognized and understood. These include a restricted capacity to factor in external influences, a limited scope for assessing property condition, insufficient data availability in specific regions, and an inability to verify the existence of a property. However, the results obtained in the study show that the AVMs can estimate the prices of apartments in Cluj-Napoca without major differences between the estimated and the actual price.

The present study was subject to certain limitations and difficulties. Firstly, there were difficulties in obtaining confidential information and values. There also there seems to be hesitation in sharing information due to GDPR policies.

The study revealed numerous similarities, as well as several concerns and uncertainties regarding AVM systems—many of which cannot be universally addressed. However, AVMs should also be developed in our country, as some regions, such as Cluj Napoca, can derive significant economic benefits from their implementation.

References

American Society of Real Estate Counsellors (2011), *Statistical Primer for Real Estate Problem Solving,* Boston.

ANEVAR (2018), *Standardele de evaluare a bunurilor (SEV 2018)*, Ed. ANEVAR, București.

Appraisal Institute of Canada (2002), *Automated Valuation Models*.

Asociația Națională a Evaluatorilor din România – ANEVAR, 2015: http://site2.anevar.ro/sites/default/files/page-files/anexa_2_gev_520_2015.pdf

Bagnoli, C. și Smith, H. (2009), *The theory of fuzz logic and its application to real estate valuation*. Jurnalul Real Estate Research 2009.

Bellotti A. (2017), Reliable region predictions for automated valuation models, *Annals of Mathematics and Artificial Intelligence*, 81, 71-84.

Bentéjac C., Csörgő A., & Martínez-Muñoz G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967

Bogin, A. și Shui, J. (2018) *Appraisal Accuracy, Automated Valuation Models, And Credit Modeling in Rural Areas*, Federal Housing Finance Agency.

Bogin, A. & Shui, J. 2019, *Appraisal Accuracy and Automated Valuation Models in Rural Areas.*

Bogin A.M., Shui J. (2018), *Appraisal accuracy, automated valuation models and credit modeling in rural areas*, Federal Housing Finance Agency Staff working paper series, working paper 18-03, April, SUA.

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence: a comparison of alternative methods. *Real Estate Economics*.

Borst, R. A., & McCluskey, W. J. (2008). The impact of location in mass real estate appraisal: a case study in the United States. *International Journal of Housing Markets and Analysis*.

Brucato, L. 2015, *Automated Valuation Models (AVMs),* Use & Usage.

Carmona P., Climent F., & Momparler A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 61, 304-323

Catt, D. (2007) „Not just a valuation tool", Motgage Finance Gazette, available at: http://www.mortgagefinancegazette.com/guide-to-avms/not-just-a-valuation-tool (accesat la data de 21 martie 2021).

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 785-794. doi:10.1145/2939672.2939785.

Cohen, J. P., Coughlin, C. C., & Lopez, D. A. (2015). The boom and bust in US housing prices from various geographic perspectives. *The Annals of Regional Science*.

Dickstein, J. (2014) „Home-equity lending regulations in the *new normal*." Mortgage Bnaking 74(7), pp. 89-90.

Crosby, N., Lizieri, C., & McAllister, P. (2003). *Means, motive and opportunity? Disentangling client influence on performance measurement appraisals*. RICS Research.

Downie, M. & Robson, G. (2008), *Automated Valuation Models: An International Perspective*, Council of Mortgage Lenders, London.

Faishal, I. Fook, J. C. and Kheng, H. E. (2004), *Automated valuation model: an application to the public housing resale market in Singapore*.

Galin, J. H. Molloy, R. Nielsen, E. Smith, P. & Sommer, K. (2018), *Measuring Aggregate Housing Wealth: New Insights from an Automated Valuation Model*, U.S. Federal Reserve Board's Finance & Economic Discussion Series, No. l., pp. 1–42.

Ho, D. T., Nguyen, T. T., Nguyen, Q. T., & Le, H. T. (2012). Using automated valuation models (AVM) for real estate appraisal in Vietnam. *Vietnam Journal of Real Estate*.

Goundar S., Maharaj K., Kumar A., & Bhardwaj A. (2021). Property valuation using linear regression and random forest algorithm. International Journal of System Dynamics Applications (IJSDA), 10(4), 1-16

Gloudemans, R. J. (1999). *Mass Appraisal of Real Property*. International Association of Assessing Officers.

Glower, M., Haurin, D. R. și Hendershott, P. H. (1998), *Selling Time and Selling Price: The Influence of Seller Motivation*, Real Estate Economics 26, pp. 719-740.

Gorton, G. & Metrick, A., 2010. Haircuts. Federal Reserve Bank of St. Louis Review.

Gupta, R., Segnon, M., Lesame, K. Și Wohar, M. E. (2020) ,,High-Frequency Valatility Forecasting of US Housing Makets. Jurnalul Real Estate Finance Econ.

Ho, D. T., Nguyen, T. T., Nguyen, Q. T., & Le, H. T. (2012). Using automated valuation models (AVM) for real estate appraisal in Vietnam. *Vietnam Journal of Real Estate*.

I.A.A.O. 2003 *Standard on Automated Valuation Models*, International Association of Assessing Officers, MO, Kansas City.

International Valuation Standards Council (IVSC) (2017), *International Valuation Standards – Standarde Internaţionale de Evaluare*, https://www.ivsc.org/standards/international-valuation-standards, consulted August 2020.

Kato, T. (2012) ,,Prediction in the lognormal regression model with spatial error dependence" J. Hous. Econ. 2012, 21, 66-76. Applied Sciences.

Kauko, T & d'Amato, M. (2008,) *Mass Appraisal Methods, An international perspective for property valuers*, Real Estate Issues. West Sussex: John Wiley & Sons.

Kuburic M., Tomic H., Mastelic Ivic S. (2012), *Use of multicriteria valuation of spatial units in a system of mass real estate valuation*, KiG, 11(17): 58-74.

Kummerow, M. (2000). The influence of current market value on appraisal accuracy. *Journal of Property Research*.

Matisiak, G. (2017) *Automated Valuation Models (AVMs): a brave new world?* Wraclaw Conference in Finance.

Matysiak, G. A. and Wang, P. (1995), *Commercial property market prices and valuations: analyzing the correspondence*, Journal of Property Research, Vol.12, No.3, pp.181-202.

McCluskey, W. J., & Anand, S. (1999). The application of intelligent hybrid techniques for the mass appraisal of residential properties. *Journal of Property Valuation and Investment*.

McCluskey, W. J., McCord, M., Davis P. T., Haran, M. și McIlhaton, D. (2013) *Prediction Accuracy in Mass Appraisal: A Comparison of Modern Approaches*, Jurnalul Property Research, 30, No. 4, pp. 239-265.

McGreal, S., Adair, A., McBurney, D. and Patterson, D. (1998), Neural networks: the prediction of residential values, Jurnalul Property Valuation and Investment, Vol. 16 Nr. 1, pp. 57-70.

Moore, J. M. (2006), *Performance Comparison of Automated Valuation Models*, Journal of Property Tax Assessment & Administration, [s. l.], v. 3, n. 1, pp. 43–59.

Mooya, M. (2017) „Automated Valuation Models and Economic Theory". Springer International Publishing 2017.

Myers, D. (1998), *Housing Market Research: A Time for a Change, Urban Land, Quebec*.

Nguyen, T. H., & Cripps, A. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research*.

Oust, A., Hansen, S. N., Pettrem, T. R. (2019) *Combining Property Price Prediction from Repeat Sales and Spatially Enhanced Hedonic Regressions*.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*.

Peterson S. Flanagan A.B. (2009), *Neural network hedonic pricing models in mass real estate appraisal*, Journal of Real Estate Research, 31 (2), 147-164.

Power, M. (1994), *The Audit Explosion,* London, Demos.RICS (2017), *Automated Valuation Models (AVMs)*, The Royal Institution of Chartered Surveyors, London.

Rossini, P. (1999). Accuracy issues for automated and artificial intelligence based valuation models. *First Annual Pacific-Rim Real Estate Society Conference*.

Rossini, P. și Kershaw, P. (2008), *Automated Valuation Model Accuracy: Some Empirical Testing*. Publicată în a 14-a Conferință Pacific Rim Real Estate Society.

Samaha, S. A. și Kamakura, W. A. (2008), *Assessing the Market Value of Real Estate Property with a Geographically Weighted Stochastic Frontier Mode*, Jurnalul Real Estate Economics, Vol. 36, pp. 717-751.

Shiller, R. J. (2015). Irrational exuberance. Princeton University Press.

Shilong Z. (2021). Machine learning model for sales forecasting by using XGBoost. In 2021 IEEE international conference on consumer electronics and computer engineering, pp. 480-483

Standard on Automated Valuation Models (2018), *Journal of Property Tax Assessment & Administrati*on, V. 15, No. 2, p. 67–101.

Tretton, David. (2007), *Where is the world of property valuation for taxation purposes going?*, Journal of Property Investment & Finance.

Vandell, K. D., Lane, J. S. (1991), *The Economics of Arhitecture and Urban Design: Some Preliminary Findings*. Jurnalul Real Estate Economy, Ediția 17.

Wang S. C., (2003). Artificial neural network. Interdisciplinary computing in java programming, 81-100

d *Econometrics: A Modern Approach* (7th ed.). Cengage Learning.

Worzala, E. Lenk, M. & Silva, A. (1995), *An exploration of neural networks and its application to real estate valuation,* The Journal of Real Estate Research, Vol.10, No.2, pp .185-201.

Zhang, R., Do, Q., Geng, J., Liu, B. Și Huang, Y. (2015) „An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study." Habitat Int., 46, pp. 196-205.