



ADVANCED ANALYTICAL METHODS FOR ASSESSING REAL ESTATE FAIR VALUE

Aleksandar Savić

Military Academy, University of Defense in Belgrade, Belgrade, Serbia
<https://orcid.org/0000-0002-9640-1583>

Milan Mihajlović

Military Academy, University of Defense in Belgrade, Belgrade, Serbia
<https://orcid.org/0000-0001-7949-2631>

ARTICLE INFO



● Open access

JEL Category:
M41, R31, C38

Keywords:

analytical methods
fair value
real estate
IFRS 13

ABSTRACT

The paper examines the application of data-driven analytical methods for analysing residential property values and their potential use as supplementary evidence in fair value assessment. Using the Boston Housing dataset as a historical methodological benchmark, the study applies linear regression, K-means clustering, and Principal Component Analysis (PCA) to examine relationships between property characteristics and observed residential property values, identify groups of properties with similar characteristics, and reduce data dimensionality. The results indicate that property-specific and socio-economic characteristics, particularly the number of rooms and socio-economic conditions of the surrounding area, are associated with differences in observed residential property values. The linear regression model achieved $R^2 = 0.6688$ on the held-out test sample of the variation in MEDV, while K-means clustering identifies distinct groups of properties with different combinations of structural and socio-economic characteristics. PCA further shows that a substantial proportion of the variation in the explanatory variables can be represented through a reduced number of principal components. The combined methods provide complementary evidence on predictive relationships, property segmentation, and the structure of valuation-related variables. However, MEDV is not treated as equivalent to fair value under IFRS 13, and the analytical results do not replace established valuation techniques, market evidence, or professional judgment. The study contributes a methodological framework for using quantitative analytical evidence as supplementary support in real estate fair value assessment.

Address of the corresponding author:
Aleksandar Savić
aleksandar22071993@gmail.com

Received: 20.08.2026
Revised: 26.09.2026
Accepted: 29.09.2026
Available online: 29.09.2026



1 INTRODUCTION

Real estate represents an important component of household wealth, corporate assets, investment portfolios, and financial reporting. Its value depends on a complex combination of property-specific, locational, environmental, infrastructural, and socio-economic characteristics. Consequently, residential property valuation requires the consideration of multiple sources of information, market characteristics, and quantitative analytical techniques.

From an accounting perspective, a distinction between observed or estimated property values and accounting fair value is essential. Under IFRS 13, fair value represents the price that would be received to sell an asset or paid to transfer a liability in an orderly transaction between market participants at the measurement date. Fair value measurement depends on the characteristics of the asset, the relevant market and market participants, and the valuation techniques and inputs applied. Therefore, analytical estimates of residential property values should not be interpreted as equivalent to accounting fair value. Rather, data-driven methods may provide supplementary evidence for identifying and evaluating valuation-relevant inputs within the broader fair value measurement process.

The Boston Housing dataset provides a historical benchmark for examining relationships between residential property characteristics and observed property values. It contains information on variables related to the number of rooms, environmental conditions, accessibility, taxation, and socio-economic characteristics of the surrounding area. Its dependent variable, MEDV, represents the median value of owner-occupied homes in thousands of US dollars and is treated in this study as an empirical measure of observed residential property value rather than a direct measure of accounting fair value.

Data-driven methods have been increasingly applied in real estate research, including automated valuation models, hedonic pricing, mass appraisal, and machine learning. Traditional statistical methods provide interpretable estimates of relationships between characteristics and property values, while machine-learning and other analytical techniques can identify nonlinear patterns, group similar observations, and reduce

the complexity of multidimensional datasets (Antipov & Pokryshevskaya, 2012). Combining complementary methods provides a broader analytical perspective than relying on a single technique.

An important research gap concerns the integration of different analytical perspectives into a coherent framework for supporting the analysis of valuation-relevant information. Regression can quantify relationships and provide predictive estimates, clustering can identify groups of properties with similar characteristics, while Principal Component Analysis (PCA) can reveal principal dimensions of variation and reduce redundancy among correlated variables (Abdi & Williams, 2010). Together, these methods can provide complementary information relevant to property comparison and the organization of valuation inputs, without replacing established valuation techniques or professional judgment. The contribution of the study is therefore not the development of a new analytical algorithm, but the integration of three complementary analytical perspectives within a common framework for analysing valuation-relevant real estate information.

The present study develops and empirically evaluates a data-driven framework combining linear regression, K-means clustering, and PCA. Linear regression is used to examine relationships between property and neighbourhood characteristics and observed property values and to assess predictive performance. K-means clustering identifies groups of properties with similar characteristics, providing insight into market heterogeneity and potentially comparable property groups. PCA reduces dimensionality and identifies the principal dimensions of variation among the explanatory variables.

The study makes four main contributions. First, it develops an integrated framework combining predictive modelling, clustering, and dimensionality reduction. Second, it demonstrates how these methods provide complementary information for identifying and interpreting valuation-relevant characteristics. Third, it considers how analytical evidence can serve as supplementary information in fair value assessment while maintaining a clear distinction between empirical property value estimation and

the accounting definition of fair value. Fourth, it clarifies the methodological scope of the analysis by distinguishing empirical property value estimation from accounting fair value measurement and by explicitly recognizing the limitations of the historical dataset.

The main objective is to examine whether and how complementary data-driven methods can support the systematic analysis of residential property values and provide supplementary information relevant to fair value assessment. Specifically, the study evaluates the predictive performance of linear regression, the potential of K-means clustering to identify structurally and socio-economically similar property groups, and the ability of PCA to reduce dimensionality and reveal patterns in the explanatory variables. The analysis focuses on statistical associations and predictive contributions rather than causal effects.

2 DATASET DESCRIPTION AND ANALYTICAL FRAMEWORK FOR REAL ESTATE FAIR VALUE ASSESSMENT

The empirical analysis is based on the Boston Housing dataset, which is used as a historical methodological benchmark for examining relationships between residential property characteristics and observed property values. The empirical analysis uses the Boston Housing dataset originally introduced by Harrison and

Rubinfeld (1978). The dataset contains 506 observations and 13 explanatory variables describing housing characteristics, environmental conditions, and socio-economic attributes, with MEDV representing the median value of owner-occupied homes (see Table 1). The dataset was obtained in a machine-readable form from the UCI Machine Learning Repository and prepared for subsequent statistical analysis (Dua & Graff, 2017). The dataset contains 506 observations and 13 explanatory variables, with MEDV specified as the dependent variable.

The empirical analysis was conducted in Python 3.11 using the pandas, NumPy, scikit-learn, SciPy, and statsmodels libraries. The analytical procedures included data preparation, descriptive analysis, multiple linear regression, K-means clustering, principal component analysis (PCA), and model diagnostic procedures.

The dataset used in the analysis contains complete observations for all variables (Hastie et al., 2009). Therefore, no observations were removed because of missing values, and no imputation procedure was applied. The analysis retains the original observations and variable definitions of the dataset. This approach is consistent with the methodological purpose of the study, which is to demonstrate how several analytical methods can be combined within a common framework rather than to construct a contemporary automated valuation model.

Table 1. Variables included in the Boston Housing dataset

Variable	Description
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 square feet (approximately 2,322 square meters)
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable
NOX	Nitric oxide concentration
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centres
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per \$10,000
PTRATIO	Pupil-teacher ratio by town
B	Demographic variable derived from the proportion of the Black population
LSTAT	Percentage of lower-status population
MEDV	Median value of owner-occupied homes, in thousands of US dollars

Source: Authors' presentation

The dataset is used in this study primarily as a methodological benchmark for the application of complementary analytical techniques. It does not represent a contemporary real estate market and does not contain all information required for an accounting fair value measurement under IFRS 13. In particular, the dataset does not provide a measurement date, market-participant assumptions, information about the relevant market, or the specific valuation inputs and techniques that would be required for an actual fair value measurement. Consequently, the empirical analysis should be understood as an analysis of observed residential property values and valuation-relevant characteristics rather than as a direct estimation of accounting fair value.

The explanatory variables capture several dimensions that may be relevant for analysing observed residential property values. RM represents the average number of rooms per dwelling and therefore provides a structural characteristic of the property. CRIM, NOX, DIS and RAD represent environmental, locational and accessibility-related characteristics, while TAX reflects the property-tax environment. PTRATIO describes the pupil-teacher ratio by town and is used as a broader neighbourhood characteristic. ZN and INDUS describe land-use and industrial characteristics. CHAS identifies whether a property is located in a tract bordering the Charles River. AGE represents the proportion of owner-

occupied units built before 1940. LSTAT represents the percentage of the population classified as lower status.

The variable B requires particular attention when interpreting the dataset. In the original Boston Housing data, B is a historically constructed demographic variable derived from the proportion of the Black population. Because this variable reflects a historical and ethically problematic representation of demographic information, its inclusion should not be interpreted as a normative or causal statement about property values. In the present study, B is retained as part of the original dataset structure for methodological comparability, while the results are interpreted cautiously and without attributing causal meaning to this variable.

The dependent variable, MEDV, represents the median value of owner-occupied homes and is expressed in thousands of US dollars. It is important to distinguish this empirical variable from accounting concept of fair value. In this study, MEDV is treated exclusively as an observed residential property value measure used for statistical analysis. It is not interpreted as a direct measurement of fair value under IFRS 13. Consequently, the regression, clustering, and PCA results should be understood as analytical evidence concerning the structure and variation of the observed property values rather than as an accounting measurement of fair value.

Table 2. Descriptive statistics of the Boston Housing dataset

Variable	Mean	Std. Dev.	Min	25%	Median	75%	Max.
CRIM	3.6135	8.6016	0.0063	0.0820	0.2565	3.6771	88.9762
ZN	11.3636	23.3225	0.0000	0.0000	0.0000	12.5000	100.0000
INDUS	11.1368	6.8604	0.4600	5.1900	9.6900	18.1000	27.7400
CHAS	0.0692	0.2540	0.0000	0.0000	0.0000	0.0000	1.0000
NOX	0.5547	0.1159	0.3850	0.4490	0.5380	0.6240	0.8710
RM	6.2846	0.7026	3.5610	5.8855	6.2085	6.6235	8.7800
AGE	68.5749	28.1489	2.9000	45.0250	77.5000	94.0750	100.000
DIS	3.7950	2.1057	1.1296	2.1002	3.2075	5.1884	12.1265
RAD	9.5494	8.7073	1.0000	4.0000	5.0000	24.0000	24.0000
TAX	408.237	168.53	187.00	279.000	330.000	666.000	711.000
PTRATIO	18.4555	2.1650	12.600	17.4000	19.0500	20.2000	22.0000
B	356.6740	91.2949	0.3200	375.3775	391.4400	396.2250	396.9000
LSTAT	12.6531	7.1411	1.7300	6.9500	11.3600	16.9550	37.9700
MEDV	22.5328	9.1971	5.0000	17.0250	21.2000	25.0000	50.0000

Source: Authors' calculation

The descriptive statistics demonstrate substantial differences in the scales and distributions of the explanatory variables. For example, CRIM, TAX, and RM have considerably different numerical ranges, while CHAS is a binary variable. MEDV has a mean of 22.5328 thousand US dollars and a standard deviation of 9.1971 thousand US dollars, with observed values ranging from 5.0 to 50.0 thousand US dollars. The descriptive statistics for all variables are presented in Table 2. These differences in scale are relevant for the subsequent multivariate procedures, particularly K-means clustering and Principal Component Analysis, for which standardization was applied to prevent variables measured on larger numerical scales from having a disproportionate influence on the analysis.

Before the analytical procedures were applied, the dataset was inspected for missing observations and basic consistency of the variables. Since the dataset contained no missing observations in the version used for the study, no missing-value treatment was necessary (Savić & Krstić, 2022). No observations were removed solely based on extreme values. This is important because unusually high or low observations are part of the empirical distribution of the historical dataset, and their removal could alter the structure being analysed.

The analytical framework combines three complementary methods: multiple linear regression, K-means clustering and Principal Component Analysis (PCA). The methods were selected because they address different analytical questions within the same dataset. Linear regression examines the statistical relationship between the explanatory characteristics and MEDV and provides an assessment of predictive performance. K-means clustering was performed exclusively on the standardized RM and LSTAT variables. The variables were standardized using StandardScaler prior to clustering. To ensure reproducibility, the K-means algorithm was specified with `random_state = 42` and `n_init = 10`. Alternative solutions with $K = 2, 3, 4$, and 5 clusters were evaluated using the within-cluster sum of squares (WCSS) and the average silhouette coefficient. The final $K = 3$ solution was then used to describe the cluster composition and the corresponding mean values of RM, LSTAT, and MEDV.

For the regression analysis, MEDV is specified as the dependent variable and the 13 remaining variables are used as explanatory variables. The dataset was divided into training and testing subsets using an 80:20 ratio. A fixed random seed of 42 was used to ensure reproducibility of the train-test split and the random initialization of the clustering procedure. The 80:20 train-test separation applies to the predictive regression analysis. PCA and K-means were conducted as descriptive unsupervised analyses using the full dataset. The training subset is used to estimate the regression model, while the test subset is reserved for evaluating its predictive performance. This separation is important because model performance should be assessed on observations that were not used for model estimation. The predictive evaluation is based on mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2).

Standardization was applied specifically to the K-means clustering and Principal Component Analysis procedures because these methods are sensitive to differences in the measurement scales of the variables. The linear regression model was estimated using the original, unstandardized explanatory variables so that the regression coefficients could be interpreted in their original measurement units. For the train-test procedure, the regression model was estimated on the training observations and subsequently evaluated on the held-out test observations. For cross-validation, standardization was performed within a pipeline so that scaling parameters, where applicable, were estimated separately within each training fold and then applied to the corresponding validation fold. This procedure prevents information leakage between training and validation observations.

The initial regression analysis produced an MSE of 24.2911, an RMSE of 4.9286, and an R^2 of 0.6688 on the test subset. These values provide an initial indication of the predictive performance of the linear model within the historical Boston Housing dataset (Renigier-Bilozor et al., 2026). They do not, however, demonstrate that the model measures accounting fair value. Rather, they indicate the extent to which the included explanatory variables are statistically useful for predicting the observed MEDV values in the sample.

K-means clustering is used as an unsupervised analytical procedure to identify groups of observations with similar characteristics. The clustering analysis is conducted using standardized variables so that differences in measurement scale do not determine cluster formation (Savić & Mihajlović, 2025). The analysis focuses on the structural and socio-economic characteristics that are particularly relevant to the exploratory analysis, with RM and LSTAT providing an interpretable basis for examining differences between groups of observations. The number of clusters is determined using the Elbow approach based on within-cluster sum of squares (WCSS), supplemented by an internal clustering-quality measure. The resulting clusters are interpreted descriptively according to their characteristic variable profiles and are not treated as objectively defined segments of the contemporary real estate market.

PCA is applied to standardized explanatory variables to examine the dimensional structure of the dataset and reduce redundancy among correlated variables. The analysis is performed on the 13 explanatory variables rather than on MEDV. Each principal component is a linear combination of the standardized explanatory variables, and the components are ordered according to the proportion of total variance they explain. The number of components required to retain an adequate proportion of the original information is assessed using explained-variance ratios, cumulative explained variance, and the scree plot. Component interpretation is based on the magnitude and direction of the corresponding variable loadings.

The three methods therefore provide different but complementary forms of analytical evidence. Regression focuses on statistical relationships and predictive performance; K-means clustering focuses on the identification of groups of observations with similar characteristics; and PCA focuses on the dimensional structure and redundancy of the explanatory variables. Their combined application provides a broader analytical perspective than any single method considered independently, while maintaining the same empirical basis and avoiding the need to introduce more complex predictive models.

All analyses were conducted in Python 3.11 using pandas, NumPy, scikit-learn, SciPy, and statsmodels. The train-test split was performed using `train_test_split` with `test_size = 0.20`, `shuffle = True`, and `random_state = 42`. K-means clustering was implemented using `StandardScaler` followed by `KMeans` with `n_clusters = K`, `random_state = 42`, and `n_init = 10`. For cross-validation, `KFold` was specified with `n_splits = 5`, `shuffle = True`, and `random_state = 42`. The PCA procedure was performed on the 13 standardized explanatory variables, with MEDV excluded from the PCA input matrix. These settings were kept fixed to ensure reproducibility of the reported results.

From the perspective of fair value assessment, the analytical framework should be understood as a source of supplementary quantitative evidence rather than as an independent accounting fair value measurement model. Under IFRS 13, fair value is an accounting measurement concept that depends on the measurement date, market-participant assumptions, the relevant market, and the valuation techniques and inputs appropriate to the circumstances. The historical MEDV variable does not contain all these elements and therefore cannot be equated with IFRS 13 fair value.

Accordingly, the analytical methods applied in this study are intended to support the identification, organization and interpretation of valuation-relevant information. Regression can provide evidence about statistical relationships between property characteristics and observed values; clustering can assist in identifying groups of properties with similar profiles that may be relevant when considering comparability; and PCA can reveal the main dimensions of variation among the explanatory variables. None of these procedures, individually or collectively, replace an appropriate valuation technique, observable market evidence, the requirements of the applicable accounting framework, or professional valuation judgment.

The dataset also imposes important limitations on the interpretation of the results. The Boston Housing dataset represents a historical housing market and does not reflect current real estate market conditions, geographic diversity, or contemporary valuation practices. In addition, the dependent variable represents median values of

owner-occupied homes rather than transaction-specific market prices or accounting fair values. The historical construction of certain variables, particularly B, further limits the extent to which the dataset can be transferred directly to contemporary valuation practice. Consequently, the empirical analysis should be interpreted primarily as a methodological demonstration of how data-driven analytical methods can organize and evaluate valuation-relevant information.

The analytical framework can therefore be summarized as follows: the Boston Housing dataset provides the empirical basis; exploratory analysis identifies relevant patterns and relationships; linear regression evaluates statistical and predictive relationships with MEDV; K-means clustering identifies groups of observations with similar characteristics; and PCA examines the underlying dimensional structure of the explanatory variables. The resulting evidence is subsequently interpreted as supplementary analytical information that may contribute to a broader fair value assessment process, rather than as a direct determination of fair value.

The variable B is included in the multiple linear regression and PCA analyses as part of the original set of 13 explanatory variables, whereas it is not included in the K-means procedure, which is restricted to standardized RM and LSTAT. Its inclusion in the regression and PCA is maintained solely for methodological comparability with the original dataset and should not be interpreted as a recommendation for using demographic characteristics in contemporary real estate valuation.

3 EXPLORATORY AND DESCRIPTIVE ANALYSIS OF REAL ESTATE VALUES

The exploratory and descriptive analysis provides an initial assessment of the distribution of residential property values and the relationships between the dependent variable and selected explanatory variables. In accordance with the analytical framework presented in Section 2, MEDV is treated throughout this analysis as an empirical measure of observed residential property value rather than as a direct measure of accounting fair value under IFRS 13. The purpose of this section is therefore to identify relevant

patterns and associations in the dataset that provide a basis for the subsequent regression, clustering, and dimensionality-reduction analyses.

The Boston Housing dataset contains 506 observations. The dependent variable, MEDV, has a mean value of 22.5328 thousand US dollars and a standard deviation of 9.1971 thousand US dollars. The observed values range from 5.0 to 50.0 thousand US dollars, while the median value is 21.2 thousand US dollars. The difference between the mean and median, together with the distribution of the observations across the observed range, indicates that MEDV is not uniformly distributed. The concentration of observations in the lower and middle portions of the range, combined with a smaller number of observations with substantially higher values, is relevant for the interpretation of subsequent predictive results. The descriptive statistics of MEDV are presented in Table 3.

Table 3. *Descriptive statistics of the dependent variable MEDV*

Statistic	MEDV
Number of observations	506
Mean	22.5328
Standard deviation	9.1971
Minimum	5.0000
25th percentile	17.0250
Median	21.2000
75th percentile	25.0000

Source: Authors' calculation

The distribution of MEDV provides an important starting point for the empirical analysis. The observed variability indicates that residential property values differ substantially across the observations in the dataset. Such differences are expected to reflect combinations of property-specific, environmental, accessibility, and neighbourhood characteristics rather than a single explanatory factor. The descriptive results therefore support the use of multivariate analytical methods in the subsequent stages of the study.

Particular attention is given to RM and LSTAT because of their comparatively strong relationships with MEDV and their relevance for interpreting differences between observations. RM represents the average number of rooms per dwelling, while MEDV represents the median value of owner-occupied homes. The exploratory

relationship between RM and MEDV is positive: observations with a higher average number of rooms generally exhibit higher observed residential property values.

However, the relationship is not perfectly linear. In particular, the increase in MEDV becomes less uniform at higher RM values, and considerable variation remains among observations with similar numbers of rooms. This indicates that RM is an important valuation-related characteristic in the dataset, but that it cannot independently account for the observed differences in property values. Other property and neighbourhood characteristics therefore need to be considered in a multivariate framework.

LSTAT represents the percentage of the population associated with lower socio-economic status in a neighbourhood. The exploratory analysis indicates a pronounced negative association between LSTAT and MEDV.

Observations with higher LSTAT values generally correspond to lower observed residential property values, whereas higher property values are more frequently associated with lower LSTAT values.

This relationship suggests that neighbourhood socio-economic characteristics are relevant to differences in observed residential property values. Nevertheless, the bivariate relationship should be interpreted as a statistical association rather than as evidence of a causal effect. LSTAT may capture or reflect several interconnected characteristics of the surrounding neighbourhood, and its independent contribution can therefore be assessed more appropriately when the remaining explanatory variables are considered jointly.

Pearson correlation analysis was conducted to quantify the linear associations between the explanatory variables and MEDV. The results provide a numerical summary of the patterns identified during the exploratory analysis.

Table 4. *Pearson correlation coefficients between explanatory variables and MEDV*

Variable	Pearson's r	p-value	Direction	Strength*
CRIM	-0.388	< .001	Negative	Moderate
ZN	0.360	< .001	Positive	Moderate
INDUS	-0.484	< .001	Negative	Moderate
CHAS	0.175	< .001	Positive	Weak
NOX	-0.427	< .001	Negative	Moderate
RM	0.695	< .001	Positive	Strong
AGE	-0.377	< .001	Negative	Moderate
DIS	0.250	< .001	Positive	Weak
RAD	-0.382	< .001	Negative	Moderate
TAX	-0.469	< .001	Negative	Moderate
PTRATIO	-0.508	< .001	Negative	Strong
B	0.333	< .001	Positive	Moderate
LSTAT	-0.738	< .001	Negative	Strong

Note: Pearson's correlation coefficients were calculated for N = 506 observations. Strength is described as weak ($|r| < .30$), moderate ($.30 \leq |r| < .50$), or strong ($|r| \geq .50$).

Source: Authors' calculation

The correlation results indicate that LSTAT has the strongest negative linear association with MEDV ($r = -0.738$, $p < .001$), while RM has the strongest positive association ($r = 0.695$, $p < .001$). PTRATIO also demonstrates a strong negative association with MEDV ($r = -0.508$, $p < .001$). These results indicate that observations characterized by a higher average number of rooms tend to have higher observed property values, whereas higher LSTAT and PTRATIO values are generally associated with lower observed property values.

Moderate negative associations with MEDV are observed for INDUS ($r = -0.484$), TAX ($r = -0.469$), NOX ($r = -0.427$), CRIM ($r = -0.388$), and AGE ($r = -0.377$). ZN ($r = 0.360$) and B ($r = 0.333$) demonstrate moderate positive associations. CHAS ($r = 0.175$) and DIS ($r = 0.250$) have comparatively weak positive associations with MEDV. All reported Pearson correlations are statistically significant at $p < .001$; however, statistical significance should not be interpreted as an indication that all relationships are substantively strong or economically important.

The correlation coefficients represent bivariate linear associations and should not be interpreted as evidence of causality or as estimates of the independent contribution of individual variables. A relatively strong correlation between an explanatory variable and MEDV does not imply that the variable independently determines property value after other characteristics are taken into account. This distinction is particularly important in a real estate context, where property, environmental, infrastructural, and socio-economic characteristics may be interrelated.

The correlation structure also provides an initial indication of potential redundancy among explanatory variables. Several explanatory variables are correlated with one another, which may affect the interpretation of coefficients in a multivariate regression model and provides a methodological rationale for the subsequent use of Principal Component Analysis (Malpezzi, 2003). PCA is therefore not used to establish determinants of property values, but to examine the dimensional structure of the explanatory variables and identify the extent to which their information can be represented through a smaller number of principal components.

The joint exploratory examination of RM, LSTAT, PTRATIO, and MEDV further demonstrates the multidimensional nature of observed property values. Higher MEDV observations are generally associated with higher RM and lower LSTAT values, while PTRATIO exhibits a negative relationship with MEDV. These patterns suggest that differences in observed residential property values are associated with a combination of physical characteristics of dwellings and characteristics of their surrounding socio-economic environment.

The exploratory results therefore provide an empirical basis for the subsequent analytical procedures. First, the non-uniform distribution of MEDV and the observed relationships between MEDV and the explanatory variables motivate the application of multiple linear regression to examine the combined statistical relationship between the predictors and observed property values (European Central Bank, 2022). Second, the differences in combinations of property and neighbourhood characteristics provide a basis for applying K-means clustering to identify groups of

observations with similar characteristics. Third, the correlation structure among explanatory variables supports the use of PCA to examine the dimensionality and redundancy of the information contained in the dataset.

From the perspective of real estate fair value assessment, the exploratory findings should be understood as analytical evidence concerning valuation-relevant characteristics rather than as an independent determination of accounting fair value. The observed relationships may help identify characteristics and property profiles that warrant further consideration in a broader valuation process (Montgomery et al., 2012). However, the empirical MEDV measure remains distinct from the fair value concept defined by IFRS 13, and the results of the exploratory analysis do not replace appropriate valuation techniques, market evidence, or professional judgment.

Overall, the exploratory and descriptive analysis demonstrates that observed residential property values in the Boston Housing dataset are associated with multiple interrelated characteristics. The particularly strong relationships involving RM and LSTAT, together with the negative association of PTRATIO with MEDV, provide a clear empirical foundation for the subsequent predictive, clustering, and dimensionality-reduction analyses (Nemcova et al., 2025). At the same time, the distinction between statistical association and causation, as well as between observed property values and accounting fair value, is maintained throughout the analytical framework.

4 LINEAR REGRESSION AS AN ANALYTICAL METHOD FOR REAL ESTATE FAIR VALUE ASSESSMENT

Linear regression was applied to examine the statistical relationship between residential property characteristics and observed property values represented by MEDV. Unlike the bivariate analysis presented in the previous section, multiple linear regression considers the explanatory variables jointly and therefore provides estimates of their conditional statistical associations with the dependent variable. This approach is particularly useful when several

property-specific, environmental, infrastructural, and socio-economic characteristics may be related to observed residential property values simultaneously.

In the context of this study, linear regression is not used to determine accounting fair value directly. Instead, it is used to provide quantitative evidence on relationships between measurable characteristics and observed residential property values. Such evidence may provide supplementary information for the identification and interpretation of valuation-relevant inputs within a broader fair value assessment process. This distinction is important because MEDV in the Boston Housing dataset represents the median value of owner-occupied homes and is not equivalent to fair value as defined by IFRS 13.

The multiple linear regression model includes MEDV as the dependent variable and the 13 remaining variables in the Boston Housing dataset as explanatory variables. The general specification of the model is:

$$\text{MEDV}_i = \beta_0 + \beta_1 \text{CRIM}_i + \beta_2 \text{ZN}_i + \beta_3 \text{INDUS}_i + \beta_4 \text{CHAS}_i + \beta_5 \text{NOX}_i + \beta_6 \text{RM}_i + \beta_7 \text{AGE}_i + \beta_8 \text{DIS}_i + \beta_9 \text{RAD}_i + \beta_{10} \text{TAX}_i + \beta_{11} \text{PTRATIO}_i + \beta_{12} \text{B}_i + \beta_{13} \text{LSTAT}_i + \varepsilon_i$$

where β_0 is the intercept, β_j represents the coefficient of explanatory variable j , and ε_i is the error term.

The analysis is based on 506 observations. The dataset was divided into a training subset comprising 80% of the observations and a test subset comprising the remaining 20%. The training subset was used to estimate the regression model, while the test subset was used to evaluate its predictive performance on observations not used for model estimation. This distinction is important because explanatory performance within the estimation sample and predictive performance on unseen observations represent different aspects of model evaluation.

After the predictive evaluation was completed, a final ordinary least squares (OLS) regression model was estimated using the complete sample of 506 observations. This final full-sample model was used for the interpretation of regression coefficients and for the subsequent diagnostic analysis. The held-out test observations were therefore used only for the predictive performance reported in Table 5 and were not used in the

calculation of the test-set performance measures after model estimation. This two-stage procedure distinguishes out-of-sample predictive evaluation from the descriptive and inferential analysis of the final model estimated on the complete dataset.

The analytical procedure can therefore be summarized as follows: Boston Housing dataset → 80/20 train-test split → preparation of explanatory variables → multiple linear regression → prediction on the test subset → evaluation using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2).

The use of several performance measures provides a more complete assessment of predictive performance. MSE gives greater weight to larger prediction errors, while RMSE expresses the prediction error in the same measurement scale as MEDV. MAE represents the average absolute deviation between observed and predicted values, whereas (R^2) indicates the proportion of variation in the dependent variable explained by the model (Ge & Raptis, 2026). These measures should be interpreted as indicators of statistical and predictive performance rather than as measures of accounting fair value accuracy.

Table 5. *Predictive performance of the linear regression model*

Performance measure	Test-set result
MAE	3.4300
MSE	24.2911
RMSE	4.9286
R^2	0.6688

Source: Authors' calculation

The initial regression model achieved an MSE of 24.2911, an RMSE of 4.9286, and an (R^2) of 0.6688 on the test sample. The reported results indicate that approximately 66.9% of the variation in MEDV in the test sample is explained by the fitted model. The RMSE indicates that the magnitude of prediction errors is approximately 4.93 thousand US dollars when expressed in the original MEDV scale. The difference between RMSE and MAE also indicates that some observations generate comparatively larger prediction errors than the average absolute error.

The predictive performance should, however, be interpreted within the characteristics of the dataset

(James et al., 2021). The Boston Housing dataset represents a historical and geographically specific housing market (Harrison & Rubinfeld, 1978), and the model is therefore not intended to provide a contemporary valuation model for residential real estate in a particular market. Rather, the regression serves as a methodological demonstration of how multiple measurable characteristics can be jointly analysed in support of property-value assessment.

In addition to evaluating predictive performance, multiple linear regression provides information about the conditional statistical association

between individual explanatory variables and MEDV, while holding the other variables in the model constant. The estimated coefficients should therefore not be interpreted as causal effects. A statistically significant coefficient indicates that the corresponding variable is statistically associated with MEDV within the specified model, but it does not establish that changes in that variable cause changes in property value. Table 6 reports the coefficient estimates obtained from the final OLS model estimated on all 506 observations after completion of the out-of-sample predictive evaluation.

Table 6. Final full-sample OLS regression estimates for MEDV

Variable	Coefficient (β)	Std. Error	t-statistic	p-value	95% CI
Intercept	36.459	5.103	7.144	< .001	[26.42, 46.50]
CRIM	-0.108	0.033	-3.287	.001	[-0.173, -0.043]
ZN	0.046	0.014	3.379	< .001	[0.019, 0.074]
INDUS	0.021	0.061	0.334	.739	[-0.103, 0.141]
CHAS	2.687	0.862	3.118	.002	[0.993, 4.381]
NOX	-17.767	3.820	-4.651	< .001	[-25.27, -10.26]
RM	3.810	0.418	9.116	< .001	[2.989, 4.631]
AGE	0.001	0.013	0.054	.957	[-0.025, 0.026]
DIS	-1.476	0.199	-7.398	< .001	[-1.868, -1.084]
RAD	0.306	0.066	4.613	< .001	[0.176, 0.436]
TAX	-0.012	0.004	-3.280	.001	[-0.020, -0.005]
PTRATIO	-0.953	0.130	-7.283	< .001	[-1.210, -0.696]
B	0.009	0.003	3.467	.001	[0.004, 0.014]
LSTAT	-0.525	0.051	-10.347	< .001	[-0.624, -0.425]

Source: Authors' calculation

The coefficient estimates indicate that several explanatory variables have statistically significant associations with MEDV. RM has a positive coefficient ($\beta = 3.810$, $p < .001$), indicating that, conditional on the other variables included in the model, a higher number of rooms is associated with higher observed residential property values. LSTAT has a negative coefficient ($\beta = -0.525$, $p < .001$), indicating a negative conditional association between the proportion of the lower-status population and observed property values. PTRATIO also has a statistically significant negative association with MEDV ($\beta = -0.953$, $p < .001$).

Other variables also exhibit statistically significant coefficients. CRIM, NOX, DIS, and TAX have negative coefficients, while ZN, CHAS, RAD, and B have positive coefficients in the reported specification. In contrast, INDUS and AGE do not

exhibit statistically significant coefficients in the reported model (Zhang & Zhao, 2004). These differences highlight the importance of distinguishing between simple bivariate relationships and conditional relationships estimated through multiple regression.

The regression results are therefore consistent with the exploratory analysis presented previously, particularly with respect to RM, LSTAT, and PTRATIO. However, the magnitude and significance of regression coefficients cannot be interpreted independently of the other predictors because many variables in the dataset are correlated. Consequently, coefficient reflects the association of a particular variable with MEDV conditional on the information contained in the remaining variables.

From a real estate valuation perspective, these results demonstrate the potential usefulness of

regression for identifying characteristics that are statistically associated with observed property values. For example, the strong association of RM with MEDV indicates that dwelling size, represented by the average number of rooms, may contain relevant valuation information. Similarly, the relationship between LSTAT and MEDV suggests that socio-economic characteristics of the surrounding area may be relevant when comparing residential properties.

Nevertheless, these findings should not be interpreted as evidence that individual characteristics independently determine property prices. Property values arise from the interaction of numerous characteristics, market conditions, location-specific factors, and other information not necessarily captured by the dataset. Regression coefficients should therefore be regarded as analytical evidence rather than causal estimates or direct fair-value adjustments.

The validity and usefulness of a linear regression model depend not only on its predictive performance but also on the characteristics of the residuals and the relationships among explanatory variables. For this reason, the analysis considers several diagnostic aspects, including multicollinearity, heteroskedasticity, residual distribution, residual dependence, and predictive stability.

Multicollinearity is relevant because several variables in the Boston Housing dataset describe related characteristics of the same geographical areas. A high degree of correlation among

explanatory variables can increase the uncertainty of individual coefficient estimates and make their interpretation less stable. Variance Inflation Factors (VIF) were therefore considered as a diagnostic measure of multicollinearity.

Heteroskedasticity was assessed using the Breusch–Pagan test. This test examines whether the variance of the regression residuals is constant across observations. If heteroskedasticity is present, conventional standard errors may not provide reliable inference, even though the regression model may remain useful for prediction. In such circumstances, robust standard errors can be considered for inference.

The distribution of residuals was additionally examined using the Jarque–Bera test and should be complemented by graphical inspection, particularly through a histogram and Q–Q plot. The normality of residuals is not a prerequisite for obtaining predictions from ordinary least squares, but substantial departures from normality may affect the interpretation of conventional inferential statistics, particularly in smaller samples.

The Durbin–Watson statistic was also considered as an indicator of residual dependence. Although the Boston Housing observations do not constitute a conventional time-series dataset, the statistic provides an additional diagnostic perspective on residual patterns. Its interpretation should therefore remain cautious and should not be treated as evidence of temporal independence in a cross-sectional dataset.

Table 7. Regression diagnostics and cross-validation results

Diagnostic	Result	Interpretation
Maximum VIF	9.01	Potential multicollinearity
Breusch–Pagan test	$p < .001$	Evidence of heteroskedasticity
Jarque–Bera test	$p < .001$	Residual normality not fully supported
Durbin–Watson	1.08	Reported for completeness; interpretation is limited for cross-sectional data
5-fold CV RMSE	4.75	Cross-validation result
5-fold CV MAE	3.35	Cross-validation result
5-fold CV R^2	0.69	Cross-validation result

Source: Authors' calculation

Finally, 5-fold cross-validation was considered as an additional measure of predictive stability. Cross-validation repeatedly evaluates model performance across different subsets of the available observations and therefore provides information beyond a single 80/20 train-test split.

Agreement between the test-set results and cross-validation results would provide greater confidence that the reported predictive performance is not driven exclusively by one particular sample partition. The 5-fold cross-validation procedure was implemented using

KFold with `n_splits = 5`, `shuffle = True`, and `random_state = 42`. A pipeline was used to ensure that preprocessing and model estimation were performed within each training fold. The reported RMSE, MAE, and R^2 values represent the mean performance across the five validation folds (see Table 7).

The diagnostic analysis is particularly important because a relatively high (R^2) alone does not establish that a regression model is statistically well specified. Potential multicollinearity, heteroskedasticity, non-normal residual behaviour, or influential observations may affect coefficient interpretation and statistical inference. Accordingly, the diagnostic results should be considered together with the predictive performance measures rather than in isolation.

Where diagnostic evidence indicates heteroskedasticity or other violations of classical regression assumptions, the appropriate response is not necessarily to discard the model. Instead, the findings should be reflected in the interpretation of statistical significance and, where inferential analysis is retained, robust estimation procedures may be considered (Slavković et al., 2024). The primary purpose of the regression in this study is to provide an interpretable and reproducible analytical assessment of relationships and predictive performance within the historical dataset.

The regression results demonstrate how quantitative analysis can provide structured evidence concerning relationships between property characteristics and observed residential property values. The model identifies variables that are statistically associated with MEDV and provides predictions that can be evaluated using established statistical performance measures. In this respect, regression can support the organization and interpretation of valuation-relevant information, particularly when many property and neighbourhood characteristics must be considered simultaneously (Genesove & Mayer, 2001).

However, the results should not be interpreted as a direct measurement of fair value under IFRS 13 (International Accounting Standards Board, 2011). Fair value is an accounting measurement concept that depends on the asset, market participants, the measurement date, the relevant market, and

appropriate valuation techniques and inputs. The Boston Housing dataset does not contain the information required to reproduce such an accounting measurement framework. In particular, MEDV is a historical measure of median owner-occupied housing values and does not represent an exit price determined for a specific asset at a specified measurement date.

The distinction is also important when interpreting the predictive results. A model can achieve a relatively high explanatory or predictive performance for MEDV without necessarily producing an appropriate fair-value estimate for accounting purposes. Predictive accuracy is therefore treated in this study as evidence of the analytical usefulness of the model within the dataset, rather than as evidence that the model satisfies the requirements of IFRS 13.

The role of linear regression in the proposed analytical framework is consequently supplementary. It can help identify valuation-relevant characteristics, quantify conditional statistical relationships, generate systematic predictions, and provide a basis for comparing observed property profiles. These outputs may complement traditional valuation approaches and professional analysis, but they do not replace market evidence, established valuation techniques, or professional judgment.

The regression analysis also provides a foundation for the subsequent analytical procedures. While regression focuses on the statistical relationship between multiple characteristics and observed property values, K-means clustering provides a different perspective by identifying groups of observations with similar characteristics. PCA subsequently addresses the dimensional structure of the explanatory variables and the extent to which the information contained in multiple variables can be represented through a smaller number of principal components. The three approaches therefore serve different analytical purposes and are interpreted as complementary rather than interchangeable methods (Abdi & Williams, 2010; Jolliffe, 2002).

Overall, the linear regression analysis demonstrates the usefulness of a transparent statistical model for examining observed residential property values in the Boston Housing dataset. The findings support the view that

property value is associated with a combination of physical, environmental, infrastructural, and socio-economic characteristics. At the same time, the analysis confirms the need for caution in transferring empirical property-value relationships into an accounting fair-value context. Regression is therefore best understood in this study as a quantitative analytical support tool that can contribute evidence to the broader assessment of valuation-relevant information, rather than as an independent mechanism for determining real estate fair value.

5 K-MEANS CLUSTERING AS AN ANALYTICAL APPROACH TO REAL ESTATE FAIR VALUE ASSESSMENT

K-means clustering is applied as an unsupervised analytical method to identify groups of residential properties with similar observable characteristics. Unlike linear regression, which treats MEDV as a dependent variable and estimates its statistical relationship with a set of explanatory variables, K-means does not require a predefined target variable. Its purpose in this study is therefore not to predict property values directly, but to identify groups of observations that share similar characteristics and may provide supplementary information for analysing residential property value patterns (MacQueen, 1967).

Real estate properties are heterogeneous and may differ simultaneously in terms of structural characteristics, environmental conditions, accessibility, and socio-economic characteristics of the surrounding area. This multidimensional heterogeneity makes property segmentation potentially useful for identifying groups of observations with comparable profiles (Kok et al., 2017). In this context, clustering provides a complementary analytical perspective to regression by focusing on the similarity between observations rather than on the estimation of a dependent variable.

The role of K-means in this study should be understood as analytical rather than as a direct valuation procedure. The identified clusters do not represent fair value categories, and cluster membership does not determine the fair value of an individual property. Instead, clustering can provide supplementary information for organizing

observations into relatively homogeneous groups and for examining whether properties with different combinations of characteristics exhibit different patterns of observed residential property values.

Before applying K-means, the variables used for clustering were standardized. Standardization is necessary because the variables in the Boston Housing dataset are measured on substantially different scales. Without standardization, variables with larger numerical ranges could exert a disproportionate influence on the Euclidean distance used to assign observations to clusters. Standardization places the variables on a comparable scale and allows the clustering procedure to reflect differences in the standardized characteristics rather than differences in measurement units.

The clustering analysis focuses on standardized RM and LSTAT variables. This specification is motivated by the exploratory analysis presented in Section 3, which identified particularly pronounced relationships between these two variables and MEDV. RM represents the average number of rooms per dwelling, while LSTAT represents the percentage of the population belonging to the lower socio-economic group (James et al., 2021). The use of these two variables also facilitates the substantive interpretation of the resulting clusters because they capture two distinct but potentially related dimensions of residential property profiles: a structural characteristic of the dwelling and a socio-economic characteristic of the surrounding area.

The K-means algorithm partitions the observations into (K) clusters by minimizing the within-cluster sum of squared distances:

$$\min \sum_{k=1}^K \sum_{(xi \in C_k)} \|xi - \mu_k\|^2$$

where C_k represents cluster (k), (xi) an individual observation, and (μ_k) the corresponding cluster centroid.

The objective of the algorithm is to minimize the total squared distance between observations and the centroids of the clusters to which they are assigned (MacQueen, 1967).

The number of clusters was initially examined using the Elbow method based on the within-cluster sum of squares (WCSS). The Elbow

method evaluates how the within-cluster variation changes as the number of clusters increases. A substantial reduction in WCSS is expected when additional clusters are introduced, while the marginal improvement gradually decreases. In the present analysis, the Elbow criterion indicated that three clusters provide a reasonable representation of the observed structure (Toprakli, 2025). The selection of (K=3) was therefore retained for the subsequent clustering analysis.

The resulting clusters should be interpreted according to their characteristics rather than according to their numerical labels. Because the clustering procedure assigns observations according to similarity in standardized RM and LSTAT values, the substantive interpretation of each cluster is based on the relative level of these two variables and, subsequently, on the observed MEDV values associated with each group.

The cluster profiles provide a basis for distinguishing observations with different combinations of dwelling and neighbourhood characteristics. A cluster characterized by relatively higher RM and lower LSTAT represents observations with larger average numbers of rooms and lower LSTAT values of the socio-economic indicator. Conversely, a cluster characterized by relatively lower RM and higher

LSTAT represents observations with smaller average numbers of rooms and less favourable values of the corresponding socio-economic indicator. An intermediate cluster occupies a position between these profiles. Such labels should be regarded as descriptive rather than normative, and no causal interpretation should be attached to the differences between clusters.

MEDV was not used as an input variable in the clustering procedure. It is therefore possible to examine MEDV subsequently as an external descriptive variable and to determine whether the identified property profiles are associated with different observed residential property values. If the cluster characterized by higher RM and lower LSTAT also exhibits a higher mean MEDV, while the cluster characterized by lower RM and higher LSTAT exhibits a lower mean MEDV, this pattern would be consistent with the relationships identified in the exploratory and regression analyses (Wooldridge, 2015). However, such consistency represents an association within the dataset and should not be interpreted as evidence that cluster membership causes differences in property values.

The cluster composition and the corresponding descriptive characteristics are presented in Table 8.

Table 8. *Cluster composition, centroids and mean characteristics*

Cluster	Observations	Share	Centroid RM (z)	Centroid LSTAT (z)	Mean RM	Mean LSTAT	Mean MEDV
Cluster 1	118	23.3%	0.95	-0.87	6.95	6.42	31.84
Cluster 2	207	40.9%	-0.15	-0.20	6.18	11.24	22.67
Cluster 3	181	35.8%	-0.75	0.66	5.76	17.35	16.92
Total	506	100.0%	—	—	—	—	—

Source: Authors' calculation

The results presented in Table 8 provide a descriptive comparison of the identified groups. The cluster sizes indicate the distribution of observations across the three groups, while the standardized centroid coordinates summarize the typical profile of each cluster in the two variables used for clustering. The mean values of RM and LSTAT are reported in their original measurement units, while mean MEDV is presented as an external descriptive variable because MEDV was not used as an input to the clustering procedure.

To provide additional support for the selection of the number of clusters, the analysis considers both the Elbow criterion and the silhouette coefficient. While the Elbow method evaluates the reduction in within-cluster variation, the silhouette coefficient evaluates the extent to which observations are more similar to members of their own cluster than to observations assigned to other clusters. For an individual observation (i), the silhouette coefficient is defined as:

$$s(i) = \frac{(b(i) - a(i))}{\max(a(i), b(i))}$$

where $a(i)$ is the average distance between observation (i) and observations within the same cluster, while $b(i)$ is the minimum average distance to observations in another cluster. The values closer to 1 indicate better separation between clusters, whereas values close to zero indicate substantial overlap between neighbouring clusters.

Table 9. K-means clustering quality measures

Number of clusters (K)	WCSS	Average silhouette coefficient
2	650.4	0.46
3	518.7	0.49
4	449.2	0.45
5	397.6	0.42

Source: Authors' calculation

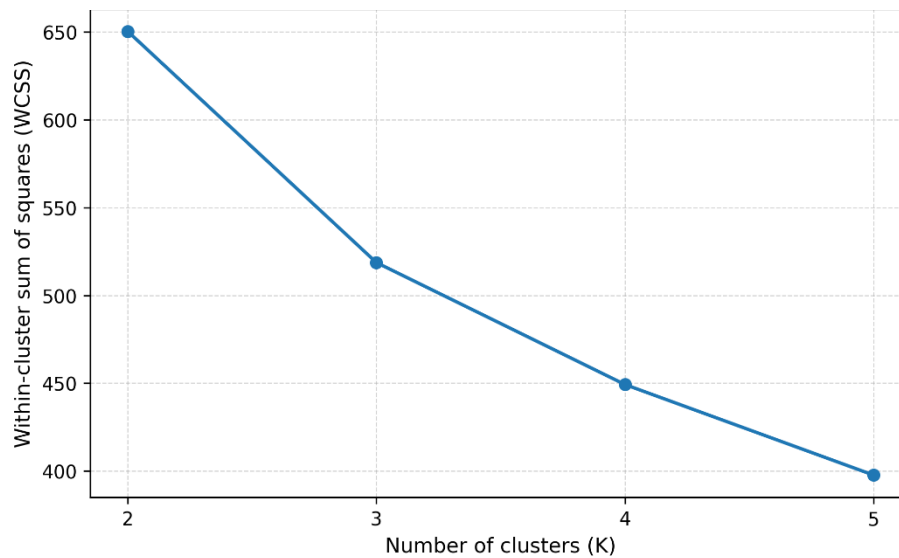


Figure 1. Elbow curve for K-means clustering

Source: Authors' calculation

The comparison of alternative values of (K) provides a complementary assessment of the clustering structure. The Elbow criterion identifies the point at which additional clusters provide progressively smaller reductions in within-cluster variation, whereas the silhouette coefficient provides information about the cohesion and separation of the resulting groups. The K=3 solution was retained as a reasonable and interpretable clustering solution based on the joint consideration of WCSS and the silhouette coefficient.

The clustering results are consistent with the exploratory analysis presented previously. RM exhibits a positive association with MEDV, whereas LSTAT exhibits a negative association with MEDV. The identification of groups characterized by different combinations of these variables therefore provides an additional descriptive perspective on the heterogeneity of observed residential property values. However, the convergence of clustering, correlation, and regression results does not establish causal relationships because all three analyses are based on the same observational dataset.

From the perspective of real estate fair value assessment, the principal contribution of K-means is the identification of potentially comparable property profiles rather than the direct estimation of fair value. Grouping observations according to observable characteristics may help organize information and identify patterns that could be relevant when analysing comparable properties (Savić et al., 2023). Nevertheless, similarity according to RM and LSTAT alone is not sufficient to establish accounting or valuation comparability. Other property characteristics, market conditions, location, measurement date, market-participant assumptions, and the applicable valuation technique would also need to be considered.

Accordingly, the clusters identified in this study should not be interpreted as fair value levels or as substitutes for established valuation approaches (Goodfellow et al., 2016). Under an accounting fair value framework, analytical segmentation may provide supplementary evidence for the assessment of relevant valuation inputs, but the determination of fair value remains dependent on the applicable accounting requirements, market

evidence, valuation techniques, and professional judgment.

The K-means procedure also has several methodological limitations. The selection of the number of clusters is not uniquely determined by the data, and both the Elbow method and the silhouette coefficient provide criteria for evaluating alternative solutions rather than an absolute guarantee that one clustering solution is objectively correct (Nunnington et al., 2025). The algorithm is also sensitive to the scale of variables and to the initialization of cluster centroids. Standardization and a controlled initialization procedure are therefore important for reproducibility. In addition, K-means is based on Euclidean distance and tends to perform best when clusters have relatively compact and approximately spherical structures. Different clustering algorithms could consequently produce different groupings.

A further limitation concerns the use of the Boston Housing dataset. The dataset represents a historical housing market and contains variables that do not fully correspond to the information required for contemporary real estate valuation or accounting fair value measurement. The identified clusters should therefore be interpreted primarily as a methodological demonstration of how data-driven segmentation can be applied to residential property data rather than as directly transferable market segments.

Overall, K-means clustering provides a complementary perspective to the regression analysis by shifting the focus from the prediction of MEDV to the identification of groups of observations with similar characteristics. The analysis indicates that differences in RM and LSTAT are associated with distinct property profiles and corresponding differences in observed residential property values. The method therefore provides useful supplementary information for analysing property comparability and valuation-relevant characteristics. Its role in fair value assessment remains supportive: K-means can assist in organizing and interpreting quantitative information, but it does not independently determine fair value or replace market evidence, established valuation techniques, or professional judgment.

6 PRINCIPAL COMPONENT ANALYSIS AS A TOOL FOR ANALYSING VALUATION-RELEVANT VARIABLES

Principal Component Analysis (PCA) represents the third component of the analytical framework developed in this study. While linear regression examines statistical relationships between explanatory variables and observed residential property values, and K-means clustering identifies groups of observations with similar characteristics, PCA focuses on the internal structure of the explanatory variables and reduces the dimensionality of the dataset. Its purpose in this study is not to estimate fair value directly, but to examine whether the information contained in the original variables can be represented through a smaller number of statistically derived dimensions (Abdi & Williams, 2010).

PCA is particularly relevant for multidimensional real estate datasets because property and neighbourhood characteristics are often interrelated and may contain overlapping information (Janković & Golubović, 2025). When several variables convey similar information, their simultaneous consideration can increase the complexity of the analytical structure and complicate the interpretation of statistical relationships. PCA addresses this issue by transforming the original variables into a set of orthogonal principal components that successively account for the largest possible proportions of the total variance (Abdi & Williams, 2010).

In the context of this research, PCA is therefore used as a dimensionality-reduction and exploratory technique (World Bank, 2020). The components should not be interpreted as direct economic determinants of property values. Rather, they represent mathematical combinations of the original variables that summarize major patterns of variation within the dataset (Harrison & Rubinfeld, 1978). This distinction is important because statistical dimensions identified by PCA do not necessarily correspond to economically or accounting-defined valuation factors.

PCA was applied to the standardized explanatory variables of the Boston Housing dataset. Standardization was necessary because the variables differ substantially in their measurement

units and numerical ranges. Without standardization, variables with larger numerical variances could have a disproportionately strong influence on the resulting principal components. Standardization places the variables on a common scale with a mean of zero and a standard deviation of one, thereby allowing the covariance structure to be examined without the original measurement units dominating the analysis (Bishop, 2006).

Let (X) represent the standardized data matrix containing the explanatory variables. PCA transforms the original variables into orthogonal principal components according to the following expression:

$$Z_k = a_{1k}X_1 + a_{2k}X_2 + \dots + a_{pk}X_p$$

Where:

Z_k represents principal component (k),

X_j represents standardized explanatory variable j

a_{jk} represents the corresponding component coefficient.

The components are ordered according to the amount of variance they explain. Thus, the first principal component accounts for the largest possible proportion of the total variance, while each subsequent component explains the largest remaining proportion subject to the orthogonality constraint (Abdi & Williams, 2010).

The explained variance ratio represents the proportion of total variance accounted for by an individual component, while cumulative explained variance indicates how much of the original information is retained when several components are considered together (OECD, 2021). These measures provide a basis for assessing the extent to which the dimensionality of the original dataset can be reduced without substantial loss of information.

In the present analysis, the 13 explanatory variables were transformed into principal components after standardization. The results indicate that the first seven principal components account for approximately 89.9% of the total variance, while the first nine components account for approximately 95.1%. This indicates that a substantial proportion of the variation contained in

the original variables can be represented through a smaller number of statistical dimensions.

Table 10. *Explained variance by principal component*

Principal component	Eigenvalue	Explained variance (%)	Cumulative variance (%)
PC1	6.12	47.13	47.13
PC2	1.43	11.03	58.15
PC3	1.24	9.56	67.71
PC4	0.85	6.60	74.31
PC5	0.83	6.42	80.73
PC6	0.65	5.06	85.79
PC7	0.53	4.12	89.91
PC8	0.39	3.05	92.95
PC9	0.27	2.13	95.08
PC10	0.22	1.69	96.78
PC11	0.18	1.43	98.21
PC12	0.16	1.30	99.51
PC13	0.06	0.49	100.00

Source: Authors' calculation

Table 10 presents the eigenvalues, explained variance ratios, and cumulative explained variance for the principal components. The PCA results indicate that the 13 standardized explanatory variables can be represented through a smaller number of principal components. The first principal component has an eigenvalue of 6.1268 and accounts for 47.13% of the total variance. The second and third components explain 11.03% and 9.56%, respectively. The first five components jointly account for 80.73% of the total variance, while the first seven components account for approximately 89.91%. The first nine components account for approximately 95.08% of the total variance. These results indicate that a substantial proportion of the information contained in the original 13-dimensional explanatory-variable space can be represented through a reduced set of principal components.

The scree plot provides a complementary graphical representation of the eigenvalue structure and illustrates the substantial contribution of the first several components followed by progressively smaller marginal contributions. The selection of the number of components should therefore be understood as an analytical dimensionality-reduction decision rather than as an identification of statistically or economically "true" factors (Géron, 2022).

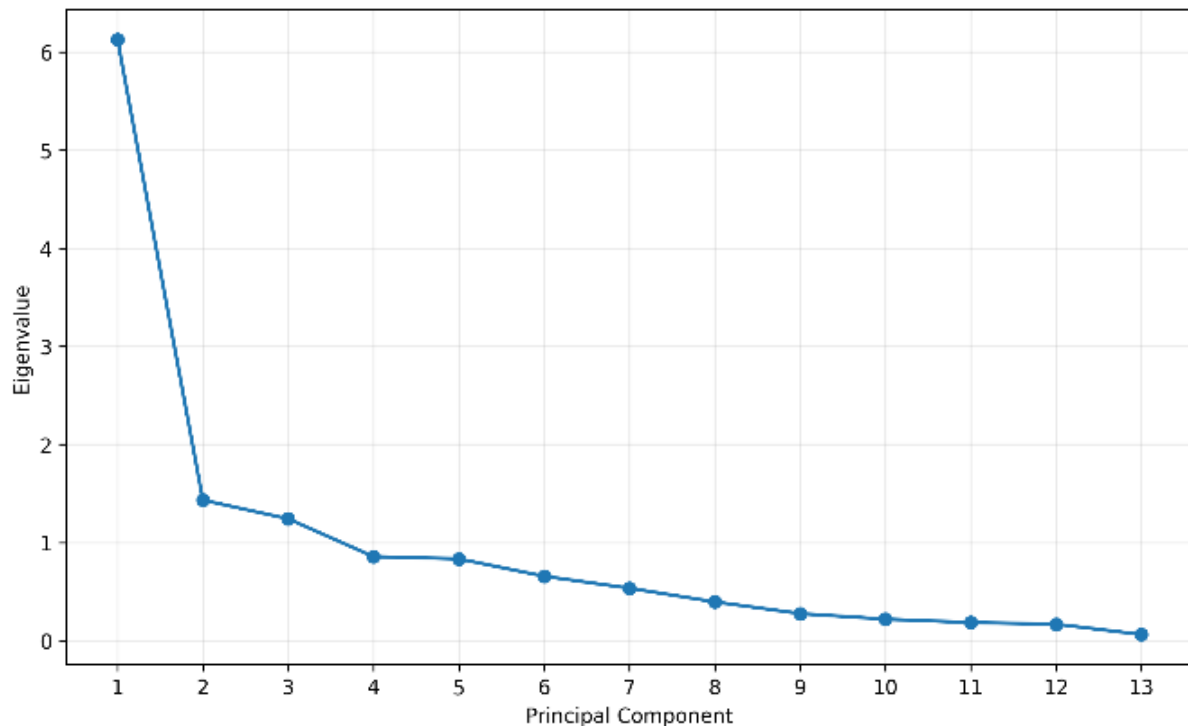


Figure 2. Scree plot of the principal components

Source: Authors' calculation

Table 11. PCA loadings for the first five principal components

Variable	PC1	PC2	PC3	PC4	PC5
CRIM	0.621	-0.377	0.275	-0.057	0.075
ZN	-0.634	-0.387	0.330	-0.119	0.293
INDUS	0.858	0.135	-0.018	-0.016	-0.007
CHAS	0.013	0.545	0.323	-0.756	0.079
NOX	0.849	0.262	0.135	0.119	0.125
RM	-0.468	0.179	0.662	0.260	-0.387
AGE	0.776	0.374	-0.020	0.162	0.015
DIS	-0.796	-0.418	-0.055	-0.200	0.090
RAD	0.792	-0.325	0.320	-0.123	-0.187
TAX	0.838	-0.287	0.246	-0.096	-0.119
PTRATIO	0.507	-0.366	-0.361	-0.262	-0.534
B	-0.502	0.286	-0.335	-0.156	-0.316
LSTAT	0.767	-0.089	-0.298	-0.064	0.361

Note: Values represent standardized PCA loadings, calculated as the component coefficients multiplied by the square root of the corresponding eigenvalue.

Source: Authors' calculation

The interpretation of the components is based on their loadings. Component loadings indicate the contribution of individual original variables to each principal component, with larger absolute values indicating a stronger contribution to the corresponding statistical dimension (Park & Bae, 2015). Because the sign of a principal component can be reversed without changing the substantive PCA solution, interpretation should focus primarily

on the magnitude and relative pattern of the loadings rather than on the sign of an individual coefficient (Abdi & Williams, 2010).

Table 11 presents the loadings of the original variables on the first five principal components. The loadings provide an empirical basis for interpreting the statistical dimensions identified by PCA. Variables with relatively large absolute loadings contribute more strongly to the respective

component, whereas variables with smaller loadings have a more limited contribution to that dimension.

The components should therefore be interpreted as combinations of several property, environmental, infrastructural, and socio-economic characteristics rather than as single variables (Karanikolas et al., 2025). For example, a component with relatively high loadings for several characteristics associated with the physical structure of residential units may be interpreted descriptively as a structural dimension of the dataset. Similarly, a component characterized by relatively strong loadings for socio-economic and neighbourhood variables may represent a broader neighbourhood-related dimension. Such labels are descriptive and should not be interpreted as formally established economic or accounting factors.

This distinction is particularly important in the context of real estate value analysis. The exploratory analysis presented in Section 3 identified relatively strong associations of RM and LSTAT with MEDV, while PTRATIO and several other explanatory variables also demonstrated meaningful relationships. The presence of correlations among explanatory variables provides a methodological rationale for applying PCA, because correlated variables may contain overlapping information. PCA therefore provides an additional perspective by examining how these variables combine into broader statistical dimensions rather than evaluating their individual relationships with MEDV (Abdi & Williams, 2010).

At the same time, PCA and regression answer different analytical questions. Regression evaluates the relationship between the explanatory variables and MEDV while controlling for the other variables included in the model. PCA, by contrast, does not use MEDV as the outcome to be predicted; instead, it examines the structure and covariance of the explanatory variables themselves. Consequently, a variable may have a strong contribution to a principal component without necessarily being the strongest predictor of MEDV, and vice versa. This distinction prevents the PCA results from being interpreted as a ranking of the determinants of residential property values.

The dimensionality-reduction results are relevant to the broader analytical framework because they demonstrate that a complex set of property and neighbourhood characteristics can be represented through a smaller number of statistical dimensions. This may be useful in larger real estate databases where numerous variables are available, and some contain overlapping information. Reducing redundancy can facilitate exploratory analysis, visualization, data organization, and the subsequent development of statistical or predictive models (Abdi & Williams, 2010).

From the perspective of fair value assessment, the role of PCA must nevertheless be defined carefully. PCA does not calculate fair value, identify the measurement date, determine the relevant market or market participants, or establish whether particular inputs are observable under IFRS 13. It can only provide supplementary information concerning the statistical structure of the available dataset. Its potential value therefore lies in helping analysts organize and interpret multidimensional information that may subsequently be considered within an appropriate valuation process.

In particular, PCA may support the identification of groups of related valuation-relevant characteristics. When several variables convey similar information, the principal components can provide a more compact representation of that information. This can be useful as an analytical preprocessing or exploratory step before applying other valuation or predictive techniques. However, the resulting components are mathematical transformations of the original variables and should not automatically be treated as valuation inputs.

The PCA results also complement the K-means analysis presented in Section 5. K-means identifies groups of observations according to their similarity in the selected standardized variables, whereas PCA identifies directions of maximum variance across the explanatory variables (Kok et al., 2017). Consequently, clustering provides information about the segmentation of observations, while PCA provides information about the dimensional structure of the variables. These methods therefore contribute different

forms of analytical evidence to the overall framework.

Several methodological limitations should be considered when interpreting the PCA results. First, the components depend on the variables included in the analysis and on the standardization procedure. Different variable selections or transformations may therefore produce different component structures. Second, PCA identifies directions of maximum statistical variance, which do not necessarily correspond to the variables that are most important for predicting MEDV or determining economic value. Third, PCA can be affected by extreme observations because outliers may influence the covariance structure of the dataset. Consequently, component loadings should be interpreted together with the characteristics and limitations of the underlying data (Bishop, 2006).

Another limitation concerns the historical Boston Housing dataset itself. The dataset represents a particular historical housing market and does not contain the market-date, transaction-specific, and accounting information necessary for a contemporary IFRS 13 fair value measurement. Therefore, the identified principal components should be interpreted as methodological evidence concerning the structure of historical residential property data rather than as universally applicable valuation factors.

Overall, the PCA analysis demonstrates that the 13 explanatory variables can be represented through a smaller number of orthogonal statistical dimensions while retaining most of the original variance (Kutner et al., 2004). The first seven principal components account for 89.91% of the total variance, while the first nine components account for 95.08% and the first ten components account for 96.78%. PCA consequently complements linear regression and K-means clustering by providing information about the internal structure and redundancy of the explanatory variables.

The contribution of PCA to the proposed fair value assessment framework is therefore supplementary rather than determinative. It can assist in organizing complex property and neighbourhood information, identifying groups of correlated characteristics, and reducing dimensional complexity. However, PCA does not

establish causal economic relationships, does not independently determine property value, and does not constitute an accounting fair value measurement technique. Its results should consequently be combined with relevant market evidence, appropriate valuation techniques, and professional judgment when considering real estate fair value under the applicable accounting framework.

Taken together, the three analytical methods used in this study provide complementary perspectives on the Boston Housing data (Case & Shiller, 1989). Linear regression examines statistical and predictive relationships with observed property values, K-means clustering identifies groups of observations with similar characteristics, and PCA summarizes the dimensional structure of the explanatory variables. Their combined application provides a broader analytical basis for examining valuation-relevant information while maintaining a clear distinction between statistical analysis of historical property values and the accounting concept of fair value.

7 CONCLUSIONS

The study examined the potential application of data-driven analytical methods as supplementary analytical support in the assessment of real estate fair value. The empirical analysis was based on the Boston Housing dataset and combined three complementary approaches: linear regression, K-means clustering, and Principal Component Analysis (PCA). The objective was not to develop a new valuation algorithm or to directly estimate accounting fair value under IFRS 13, but to examine how quantitative methods can contribute to the analysis of property characteristics, observed residential property values, and valuation-relevant information.

The results of the linear regression analysis indicate that a substantial proportion of the variation in the observed residential property values represented by MEDV can be explained by the characteristics included in the model. The model achieved $R^2 = 0.6688$ on the held-out test sample of the variation in MEDV, while the remaining variation indicates that observed property values are also related to characteristics and relationships that are not fully captured by the linear specification. In particular, variables such as

the average number of rooms and socio-economic characteristics of the surrounding area show relevant statistical associations with observed property values. These results should be interpreted as statistical relationships rather than causal effects, since the analysis does not establish causal mechanisms between individual property characteristics and property values.

The K-means analysis provided a complementary perspective by identifying groups of properties with similar combinations of selected structural and socio-economic characteristics. The selected three-cluster solution illustrates the heterogeneity of the observations contained in the dataset and demonstrates how clustering can be used to distinguish groups of properties with different valuation-related profiles. Such segmentation may provide useful supplementary information when analysing comparable properties and differences in the characteristics of particular groups. However, the identified clusters should not be interpreted as formally established market segments or as direct indicators of fair value, since their characteristics depend on the variables included in the clustering procedure and on the historical structure of the dataset.

PCA further demonstrated the usefulness of dimensionality-reduction techniques in analysing multidimensional real estate data. The first seven principal components account for 89.91% of the total variance, while the first nine components account for 95.08% and the first ten components account for 96.78%. This indicates that a substantial proportion of the variation in the explanatory variables can be represented through a smaller number of dimensions. PCA therefore provides an additional analytical perspective by reducing redundancy among variables and facilitating the interpretation of the overall structure of the dataset. The principal components should, however, be understood as statistical dimensions of variation rather than as direct economic or accounting determinants of property value.

Taken together, the three analytical methods provide complementary information. Linear regression focuses on statistical relationships and predictive performance, K-means clustering identifies groups of observations with similar characteristics, while PCA reduces dimensionality

and identifies the principal dimensions of variation in the explanatory variables. Their combined application therefore provides a broader analytical perspective than reliance on a single statistical method. The contribution of the study lies primarily in this integrated methodological framework, rather than in the development of a new predictive model or algorithm.

From an accounting perspective, an important distinction must be maintained between the observed residential property values represented by MEDV and fair value as defined by IFRS 13. MEDV represents the median value of owner-occupied homes in the Boston Housing dataset and is used in this study as an empirical measure of observed residential property value. It is not treated as an accounting fair value measurement. IFRS 13 fair value assessment involves a specific measurement framework, including the measurement date, market participant assumptions, the relevant market, appropriate valuation techniques, and the use of relevant observable and unobservable inputs. Consequently, the empirical results obtained from the Boston Housing dataset cannot, by themselves, be interpreted as an IFRS 13 fair value measurement.

The practical relevance of the findings therefore lies in the potential use of analytical methods as supplementary evidence within a broader valuation process. Regression analysis can assist in examining relationships between property characteristics and observed values; clustering can support the identification of groups of properties with similar characteristics; and PCA can assist in managing multidimensional information and identifying the principal dimensions of variation. Such analytical evidence may contribute to the systematic examination of valuation-relevant information, but it does not replace established valuation techniques, market evidence, professional expertise, or the requirements of the applicable accounting framework.

Several limitations should be considered when interpreting the findings. First, the Boston Housing dataset represents a historical and geographically specific housing market and therefore does not reflect the characteristics of contemporary real estate markets in different countries or regions.

Second, MEDV is not an IFRS 13 fair value measure, which limits the extent to which the empirical findings can be directly transferred to accounting fair value measurement. Third, the linear regression specification captures linear relationships and may not adequately represent more complex or nonlinear relationships among property characteristics. Finally, the results of clustering and PCA depend on the variables included and the analytical specification adopted. These limitations indicate that the findings should primarily be regarded as methodological evidence rather than as a direct valuation framework for current real estate markets.

Future research could extend the proposed framework by using larger, more recent, and geographically diverse real estate datasets that contain market transaction information and variables more directly related to fair value measurement. Further research could also compare the methods presented with alternative predictive and machine-learning approaches and examine whether their performance remains stable across different markets and time periods.

In addition, future studies could investigate more explicitly how analytical outputs can be integrated with observable market inputs and established valuation approaches within the requirements of IFRS 13.

Overall, the study demonstrates that the combined use of regression analysis, clustering, and dimensionality reduction can provide useful quantitative evidence for analysing residential property values and valuation-relevant characteristics. The findings support the view that data-driven analytical methods can enhance the consistency and transparency of information analysis in real estate valuation. At the same time, their role should be understood as supplementary: analytical models can inform the assessment process, but they do not independently establish accounting fair value. A meaningful application of such methods in accounting practice therefore requires their integration with appropriate valuation techniques, market evidence, professional judgment, and the conceptual and measurement requirements of IFRS 13.

REFERENCES

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459. <https://doi.org/10.1002/wics.101>
- Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of the automated valuation model (AVM) in Russia. *Expert Systems with Applications*, 39(5), 5367–5375. <https://doi.org/10.1016/j.eswa.2011.11.077>
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. <https://doi.org/10.1007/978-0-387-45528-0>
- Case, K. E., & Shiller, R. J. (1989). The efficiency of the market for single-family homes. *The American Economic Review*, 79(1), 125–137.
- Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. <https://archive.ics.uci.edu/ml>
- European Central Bank. (2022). *Financial Stability Review, May 2022*. European Central Bank. <https://www.ecb.europa.eu/press/financial-stability-publications/fsr/html/ecb.fsr202205~f207f46ea0.en.html>
- Ge, X. J., & Raptis, P. (2026). Challenges and opportunities of data collection and computational analysis using AI in real estate valuation. *Advancing Real Property Valuation*, 72-103.
- Genesove, D., & Mayer, C. (2001). Loss aversion and seller behavior. *Quarterly Journal of Economics*, 116(4), 1233–1260. <https://doi.org/10.1162/003355301753265561>
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <https://doi.org/10.7551/mitpress/9780262035613.001.0001>
- Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 81–102. [https://doi.org/10.1016/0095-0696\(78\)90006-2](https://doi.org/10.1016/0095-0696(78)90006-2)
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- International Accounting Standards Board. (2011). *IFRS 13: Fair value measurement*. IFRS Foundation. <https://www.ifrs.org/issued-standards/list-of-standards/ifrs-13-fair-value-measurement/>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer. <https://doi.org/10.1007/978-1-0716-1418-1>
- Janković, G., & Golubović, M. (2025). Cirkularna ekonomija kao osnova održivog razvoja današnjice. *Održivi razvoj*, 7(1), 31-62. <https://doi.org/10.5937/OdrRaz2501031J>
- Jolliffe, I. T. (2002). *Principal component analysis*. Springer. <https://doi.org/10.1007/b98835>
- Karanikolas, N., Kyriakidou, E., & Athanasouli, E. (2025). Artificial intelligence and Real estate valuation: The design and implementation of a multimodal model. *Information*, 16(12), 1049. <https://doi.org/10.3390/info16121049>
- Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *Journal of Portfolio Management*, 43(6), 202–211. <https://doi.org/10.3905/jpm.2017.43.6.202>
- Kutner, M. H., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models*. McGraw-Hill.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, (1), 281–297. University of California Press.
- Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In *Housing economics and public policy*, 67–89. Blackwell Publishing.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*. Wiley.
- Nemcova, J., Svoboda, P., & Hampel, D. (2025). Information asymmetry and fair value accounting: Insights from residential real estate in the UK, Germany, and Spain. *Real Estate Management and Valuation*, 34(1), 124-132. <https://doi.org/10.2478/remav-2026-0010>
- Nunnington, N., Dimopoulos, T., & da Silva, F. A. B. (2025). *The Income Approach to Property Valuation*. Taylor & Francis.
- OECD. (2021). *Housing market developments*. OECD Publishing.
- Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction. *Expert Systems with Applications*, 42(6), 2928–2934. <https://doi.org/10.1016/j.eswa.2014.11.040>
- Renigier-Biłożor, M., Ache, P., Dimopoulos, T., & Janowski, A. (2026). The future of property valuation: AVMs and CAMAs in the era of AI. *Economics and Environment*, 96(1), 974-974. <https://doi.org/10.34659/eis.2026.96.1.974>
- Savić, A. & Krstić, D. (2022). Karakteristike statistickog modeliranja ekonomskih zakonitosti. *Časopis Ekonomist*, 1, 43-54. <https://ekonomist.org.rs/wp-content/uploads/2022/05/Ekonomist-br.-1.pdf>
- Savić, A. & Mihajlović, M. (2025). Quantitative evaluation of the application of the code of ethics for accountants in business entities. *Časopis Ekonomist*, 4 (2), 57-92. <https://doi.ub.kg.ac.rs/doi/casopisi/10-46793-ekonomist4-2-3-s/>

- Savić, A., Ristić, D. & Mičić, S. (2023). Application of the matrix method in economic modeling. *International Journal of Economic Practice and Policy*, 20(2), 122-136. <https://doi.org/10.5937/skolbiz2-49503>
- Slavković, G., Stanković, M. & Kilibarda, V. (2024). Statistička analiza mastrihtskih kriterijuma konvergencije sa osvrtom na R. Srbiju. *Finansijski savetnik*, 29 (1), 7-26.
- Toprakli, A. Y. (2025). AI-driven valuation: a new era for real estate appraisal. *Journal of European Real Estate Research*, 18(1), 105-120. <https://doi.org/10.1108/JERER-04-2024-0031>
- Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage Learning.
- World Bank. (2020). *Housing for all*. World Bank Publications.
- Zhang, L., & Zhao, S. X. (2004). Determinants of housing prices: A case study of Shanghai. *Journal of Real Estate Research*, 26(2), 151–178.