Journal of Innovation and Development
ISSN: 2958-4752 | Vol. 10, No. 3, 2025

The Mass Appraisal Study of Taxable Price of Commodity Housing Based on Automated Valuation Models

Kaiqing Wang[1], Yichao Wang[1], Zhuang Yang[1,*]

[1]School of Public Administration, Sichuan University, Chengdu, 610065, China

[*]Corresponding author Email: yangzhuang357@163.com

Abstract: The purpose of this study is to reveal the construction characteristics and advantages of the random forest automatic assessment model based on the transaction data of new commercial housing in X district of A city, and to introduce a rolling time window to enhance the model performance of random forest and guide its practice in the mass appraisal of taxable price of commodity housing. The methods of this study are GIS spatial information processing, random forest algorithm and rolling time window time series analysis. The research results show that: 1) the number of features in the feature subset at regression tree splitting and the node division rules at regression tree splitting have a large impact on the prediction accuracy of the random forest automatic assessment model. 2) The evaluation accuracy and generalization ability of the random forest automatic evaluation model are significantly better than those of the Hedonic model, with a goodness-of-fit of 98% and an average absolute error percentage of 2.7% in the training and test sets. 3) The introduction of a suitable rolling time window can improve the prediction accuracy and economic explanation ability of the random forest automatic valuation model. In conclusion, the random forest automatic valuation model based on rolling time windows can be widely used for automatic batch valuation of taxable prices of real estate in the study area.

Keywords: Taxable price of commodity housing; automatic valuation models; random forest; rolling time window.

1. Introduction

Adhering to the principle that "housing is for living, not for speculation," promoting the stable and healthy development of the real estate market is a key deployment and objective of the 14th Five-Year Plan. The introduction of a property tax, with a focus on the ownership phase, can curb speculative behavior in the real estate market while mitigating wealth inequality. To ensure the healthy growth of the real estate sector and uphold social equity, the implementation of a property tax is imperative. Using assessed property values as the tax base during taxation can enhance fairness. However, China's real estate market is characterized by rapid growth, frequent transactions, and a massive scale. Traditional property valuation methods, which primarily rely on appraisers and focus on individual properties, struggle to complete large-scale assessments efficiently. Additionally, due to limited data and incomplete standardization, traditional valuation results often carry subjectivity. In contrast, mass appraisal—capable of evaluating large numbers of properties under unified standards—has become the primary method for tax base assessment in countries levying property taxes[1]. Therefore, achieving high-precision mass appraisal has emerged as a critical research priority.

Entering the era of big data, the increased availability of real estate characteristics and transaction data has made the development of automated valuation models (AVMs) using computers a focal point in fields such as taxation and credit. In 2003, the International Association of Assessing Officers (IAAO) released the Standard on Automated Valuation Models, which standardized the construction processes, modeling techniques, and application scopes of AVMs. In contrast, China's research on mass appraisal and AVMs started later, initially focusing on integrating GIS with valuation methods to build appraisal systems [2, 3].

Real estate AVMs rely on computers and mathematical models, with their core objective being to uncover the quantitative relationship between property prices and their influencing factors, and to apply this relationship to predict prices for properties under appraisal. The hedonic price model is one of the most widely used mathematical frameworks for constructing AVMs[4, 5]. While hedonic models quantify the linear impact of individual factors on property prices, real estate prices are often shaped by complex, nonlinear interactions among multiple variables. To address this, scholars globally have begun integrating machine learning algorithms into AVM development to capture nonlinear and interactive effects of property characteristics on prices. Numerous studies have explored neural network-based AVMs, delving into their stability, accuracy, convergence speed, and generalization capabilities[6–9]. Additionally, research has investigated support vector regression (SVR)-based AVMs [10]. Compared to neural networks and SVR, regression tree algorithms—with their simplicity, computational efficiency, and ability to handle both continuous and categorical variables—have been widely adopted in AVM construction. Current studies indicate that ensemble methods based on regression trees, such as random forest models, demonstrate superior accuracy in AVMs[11–14]. However, existing research still faces limitations:

(1) Small sample sizes;

(2) Short temporal spans in datasets;

(3) Overreliance on listing prices rather than actual transaction data, failing to reflect real market conditions;

(4) Relatively low data quality, including inconsistent or incomplete quantification of property attributes and omitted variable biases.

To further elucidate the strengths and unique characteristics

of random forest AVMs, this study investigates how parameter settings affect their performance and compares them with the commonly used hedonic model in terms of accuracy, generalization, and economic interpretability. Innovatively, a rolling time window approach is introduced to enhance the model's performance in long-term price prediction, capturing evolving impacts of property characteristics on market prices across different periods. Finally, the application advantages of random forest AVMs are summarized, guiding their practical implementation in the mass appraisal of tax-assessable prices for commercial housing.

2. Model Construction Principle and Accuracy Evaluation Method

2.1. Characteristic price model

The hedonic price model, also known as the hedonic pricing model or Hedonic model, has been widely applied in real estate valuation in recent years. The core idea of the hedonic model is that a property possesses diverse characteristics, and its price is collectively determined by these attributes. Each factor influencing real estate prices corresponds to an implicit price. The hedonic model does not adhere to a fixed or uniform functional form; instead, its functional specification is determined based on empirical data. In valuation practice, common functional forms of the hedonic model include the linear function, linear-log function, log-linear function, and logarithmic function, with their respective formulas as follows:

$$P = \alpha_0 + \sum \alpha_i X_i + \varepsilon \quad (1)$$

$$P = \alpha_0 + \sum \alpha_i \ln X_i + \varepsilon \quad (2)$$

$$\ln P = \alpha_0 + \sum \alpha_i X_i + \varepsilon \quad (3)$$

$$\ln P = \alpha_0 + \sum \alpha_i \ln X_i + \varepsilon \quad (4)$$

Among them,P is the characteristic price of real estate; α_0 is a constant term, which refers to other constants other than the characteristic factors affecting the evaluation price;$\alpha_i$is the variable coefficient; X_i is the variable, specifically refers to the characteristic factors; ε is the error term, generally not of economic significance.

2.2. Regression tree model

Given a real estate transaction data set with a capacity of n, it is randomly divided into a training set with a capacity of nt and a test set with a capacity of nr (n=nt+nr), and the characteristic factors affecting the real estate price are labeled as$x_j(j = 1,2, \dots,)$. In this paper, the training set is used to train the automatic evaluation model, and the test set is used to analyze the accuracy.

Regression tree is a kind of regression tree that is applied to the dependent variable as a continuous variable [15], and it is a machine learning that automatically evaluates the model of random forest. The regression tree model uses the tree structure to divide the real estate transaction data set into several subsets, and makes use of the sample prices in the subsets for price evaluation, that is $\widehat{P_{C_r}} = \text{average}(P_i | i \in C_r)$. A typical regression tree consists of root nodes, inner nodes and leaf nodes. The root node is the node that contains all the data in the data set; The root node contains the judgment condition and takes it as the starting point to grow the regression tree. The splitting process will form multiple internal nodes. Each internal node represents the data set that meets all the judgment conditions from the root node to the node in the data set. The resulting subsets of data with similar characteristics are called leaf nodes. That is, the data set used to make real estate price forecasts.

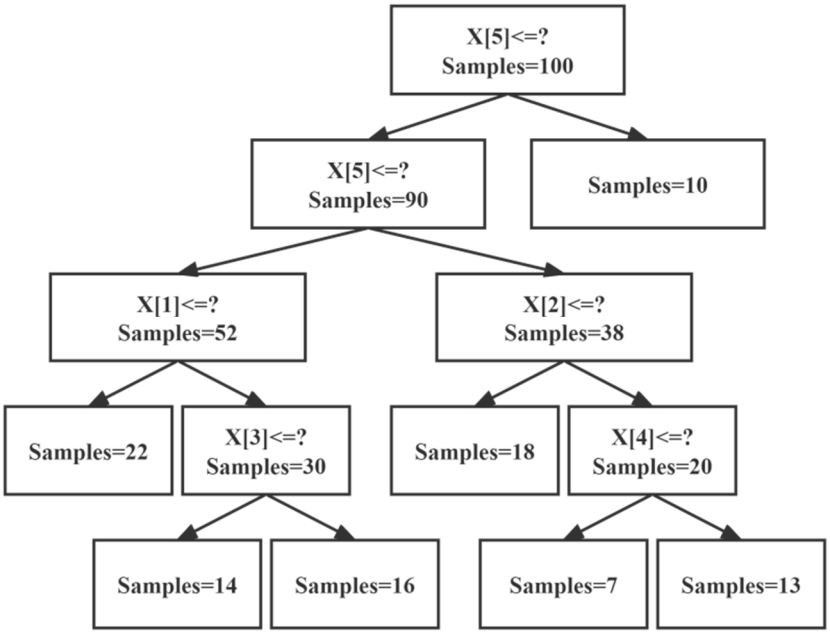


Figure 1. Regression tree schematic

When splitting root or internal nodes to grow a regression tree-based automated valuation model, specific node-splitting rules are applied as splitting criteria. Using corresponding purity metrics and splitting rules, a heuristic approach determines the optimal splitting feature variable (jj) and split point. Additionally, constructing the regression tree model requires defining the tree's height or depth. Pre-pruning and post-pruning are two common methods: the former trims the tree during construction (e.g. setting a maximum depth in advance), but this approach is less effective; the latter first builds a large tree and then decides whether to prune subtrees based on changes in prediction errors before and after merging leaf nodes [16].The most widely used regression tree growth algorithm is the Classification and Regression Tree (CART) algorithm, which employs the Gini Index as the node-splitting rule. In this study, the extended CART algorithm implemented in the R-language ranger package is utilized for regression tree construction. This algorithm also supports multiple node-splitting rules. Common splitting rules for regression problems include:

(1) Variance: Uses variance to measure node purity. In regression, variance quantifies the dispersion of the target variable, with node purity evaluated via mean squared error (MSE). The feature minimizing variance is selected for splitting.

(2) Extratrees: Applies the Extremely Randomized Trees (ExtraTrees) algorithm for node splitting. During tree construction, this algorithm randomly selects a subset of features as candidate splits and randomly chooses a split threshold. Among multiple randomized feature-threshold combinations, the split maximizing purity gain is selected.

2.3. Random forest automatic assessment model

When the node-splitting depth of a regression tree-based automated valuation model is excessive, overfitting becomes likely [16]. Random Forest Regression (RFR), an ensemble algorithm based on regression trees enhanced by the Bagging algorithm, effectively mitigates overfitting in regression tree models. The construction process of the random forest automated valuation model using the R-language ranger package is as follows:

(1) Bootstrap Sampling: Randomly sample k training subsets (each of size n) with replacement from the original training set. Each bootstrap sample typically contains approximately two-thirds of the original training data, with the remaining one-third forming the out-of-bag (OOB) data.

(2) Feature Selection and Splitting: For each training subset, at every splitting node, randomly select h features from m housing characteristics and employ the extended CART algorithm to split nodes and grow the regression tree（m ≤ h）.

(3) The generated k regression trees are formed into a random forest, and the mean value of the predicted values of k trees determines the final prediction result. The random forest model can evaluate the importance of feature variables using out-of-bag (OOB) data. When applying each regression tree to its corresponding OOB data, the root mean squared error (RMSE) of the OOB predictions is obtained. Subsequently, random noise is introduced to a specific feature variable in the OOB samples (by randomly perturbing its values), and the RMSE is recalculated. The importance of x_j is determined by the magnitude of the increase in RMSE before and after the perturbation. A larger indicates that variable x_j plays a more critical role in the regression model construction [17].

2.4. Accuracy evaluation method

After constructing the random forest automated valuation model using the training set, the model's performance can be validated and evaluated with test samples. This evaluation is based on the prediction errors between the model's predicted values and the actual values of the test samples. Drawing on prior research [18], this study employs the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) as metrics to assess prediction accuracy. The formulas for RMSE, MAE, and MAPE are as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{nr}\left(P_i - \widehat{P}_i\right)^2}{nr}} \tag{5}$$

$$MAE = \frac{1}{nr}\sum_{i=1}^{nr}\left|P_i - \widehat{P}_i\right| = \frac{1}{nr}\sum_{i=1}^{nr}|e_i| \tag{6}$$

$$MAPE = \frac{\sum_{i=1}^{nr}\frac{\left|P_i - \widehat{P}_i\right|}{P_i} \times 100}{nr} \tag{7}$$

Where, is the real price value of the test sample P_i, and $\widehat{P}_i$ is the predicted price value of the test sample. R^2 is calculated as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{8}$$

Where, Total sum of squares(SST) , $SS_{tot} = \sum_i\left(P_i - \widehat{P}_i\right)^2$,sum of squares of residuals(SSE), $SS_{res}=\sum_t\left(P_i - \widehat{P}_i\right)^2 = \sum_i e_i^2$.

2.5. Rolling time window model

The Time Window Model is a classical time series analysis framework that divides events within a specific period into discrete time windows, using historical data within each window to forecast future trends. Its strength lies in segmenting time series data into discrete intervals, making it adaptable to both short-term and long-term forecasting. In the Rolling Time Window Model, the time series is partitioned into consecutive windows, each containing data points from a defined time span. By sliding these windows across the time series, data within each window can be extracted and applied to subsequent analysis and prediction. This approach is widely used in tasks such as time series forecasting, anomaly detection, and feature extraction. For instance, in time series forecasting, the rolling time window model can capture patterns and trends from past intervals, enabling predictions of future values based on these insights.

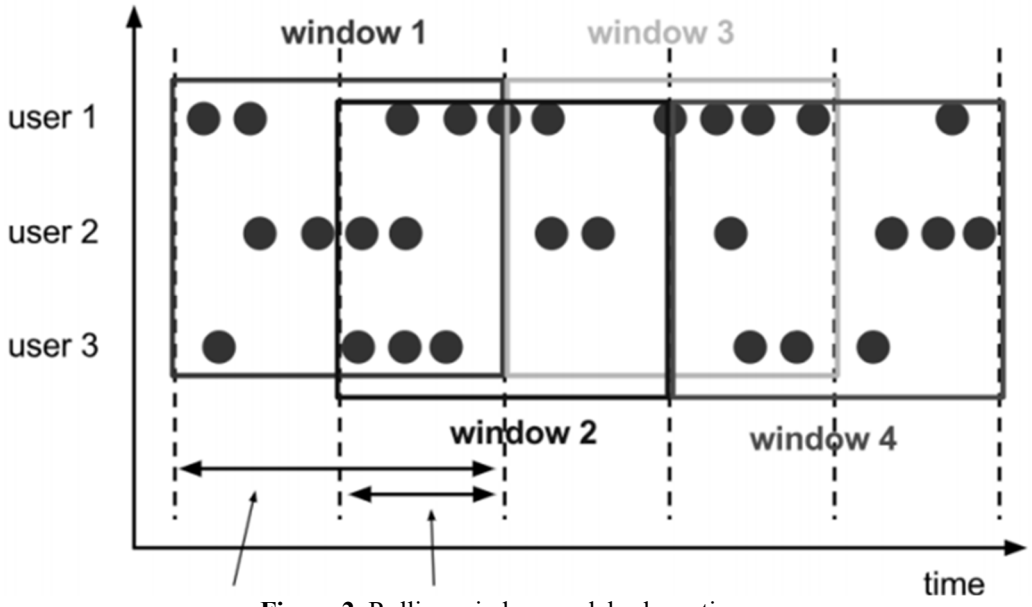


Figure 2. Rolling window model schematic

The window size and sliding step length are two critical parameters in the rolling time window model. The window size determines the temporal span of each window, while the sliding step length defines the interval between consecutive windows. A larger window size allows the use of more data to estimate the hedonic price model, thereby improving estimation validity, but reduces the model's responsiveness to market fluctuations. Conversely, a smaller window size enhances the alignment of shadow prices with market conditions but sacrifices estimation reliability. The optimal choice of these parameters depends on the specific task requirements and data characteristics.In this study, constrained by the original transaction data's temporal granularity (accurate only to the month), the sliding step length is set to 1 month. To enable the random forest model to more precisely capture the dynamic influence of variables over specific periods, and informed by prior research [19], the window size is optimized to 12 months.

3. Data and Variables

3.1. Study area and data

This study focuses on newly constructed commercial residential properties in X district of A city. The data base is based on the new commercial residential transaction cases from January 2016 to December 2022 in District X of City A. There are 102,210 records covering 136 neighborhoods. The original dataset contains 102,210 records covering 136 residential complexes. After removing invalid entries (e.g., affordable housing units, villa-type properties) and outliers in unit price per construction area, the final sample comprises 72,767 records across 114 residential complexes. In addition to transaction prices, the dataset includes six key attributes: transaction date, geographic location, building number, floor level, total floors, and construction area.

To enrich the property characteristics, this study integrates multi-source big data. For physical attributes, transaction data are combined with property listings from platforms such as Beike (www.ke.com), Anjuke (www.anjuke.com), and Fangtianxia (www.fang.com). For locational attributes, point of interest (POI) data for X district of A city are obtained from the Amap Open Platform. These are further integrated with third national land survey data, remote sensing data, and publicly available government documents. Using GIS, annual POI layers are constructed, including:Transportation hubs (e.g., intercity transit stations, metro stations, bus stops)、Commercial facilities (shopping centers)、Educational and cultural services (schools, cultural institutions)、Recreational amenities (sports and leisure venues)、Healthcare facilities、School districts (primary and middle schools)、Tourist attractions (AAA-rated or above)、Green spaces (parks)、Blue spaces (water bodies).

3.2. Variable Setting

The factors that affect the housing price include general factors, location factors and physical factors. Since the real estate transaction data collected in this paper are all located in X district of A city, with similar social, economic and administrative conditions, and general factors are homogeneous in the study area, the influence of general factors on housing prices is not considered, and only the selection of location factors and individual factors are considered. In addition, the transaction data collected in this paper has a large time span, so the influence of time factor should be considered. Finally, based on data requirements and studies of other scholars [10,20,21], this paper sets 1 time factor, 12 physical factors and 28 location factors as characteristic attributes of real estate. Different attributes were quantified by setting continuous variables and virtual variables. Finally, 141 individual factor characteristic variables were obtained. The definitions of each variable are shown in Table 1.

Table 1. Variable definition table

Target layer	Criterion layer	Executive layer	Variable name
Dependent variable	Transaction price	Total real transaction price	price
Time	Trading hours	Real estate actual transaction hours	Y16M1~Y22M12
Entity	Building characteristics	Floor area	area
		Number of chambers	room1~room5_more
		Number of halls	hall1~ hall3_more
		Number of toilets	toilet1~toilet3_more
		Building age	Housing under construction/prospective_house
			1_5years
			6_10years
		General floor	low_rise
			multi_storey
			small_high_rise
			high-rise
			super_tall_buildings
		floor	lower_floor
			middle_floor
			higher_floors
		Whether it is at the bottom	first-layer
		Top-level or not	top-floor
	Cell characteristics	Cell size	apartment_complex_size
		Plot ratio	volume_ratio
		Greening rate	greenery_rate
location	Distance from city center	Distance from city center	szxjl
	Regional location	Latitude and longitude coordinates	xcoordinate、ycoordinate
	Traffic facilities condition	Whether there are external traffic nodes within 1000/1500 meters	X1000jtjd/X1500jtjd
		Is there a subway station within 500/1000 meters	X500dtzd/X1000dtzd
		Number of bus routes at bus stops within 500/1000 meters	X500gjlx/X1000gjlx
	Educational facilities Facility condition	Distance from the nearest primary school	xxjl
		Distance from the nearest secondary school	zxjl
		Is located in a secondary school district	zxxq
	Living service configuration facility status	Number of shopping places within 500/1000 m	X500gwcs/X1000gwcs
		Distance from the nearest shopping place	gwcsjl
		Number of scientific, educational and cultural service facilities within 500/1000 meters	X500kjwh/X1000kjwh
		Distance from the nearest scientific, educational and cultural service facilities	kjwhjl
		Number of sports and leisure facilities within 500/1000 meters	X500tyxx/X1000tyxx
		Distance from the nearest sports and leisure facilities	tyxxjl
		Number of medical facilities within 500/1000 m	X500ylbj/X1000ylbj
		Distance to the nearest Tier 3 hospital	sjyyjl（Continuous variable）
		Is there a 3A or above scenic spot within 1000/1500 meters	X1000fjms/X1500fjms
		Blue space coverage within 1000 meters	X1000lanskj
		Green space coverage within 1000 meters	X1000lvskj
		Distance from the nearest park green space	gyldjl

4. Model Construction and Comparative Analysis

This paper divides the training set and the test set by stratified random sampling based on the cell names of the transaction data. The training set contains 80% of the total sample (58,171 records in total), and the test set contains 20% of the total sample (14,596 records in total). In order to ensure the same distribution of the training data set and the test data set, so that the prediction accuracy and generalization ability of the model can be accurately evaluated, the two-sample Kolmogorov-Smirnov test (KS test) is used to test the stability of the two samples of the test set and the training set. The test shows the maximum vertical difference between the cumulative distribution curves of the two samples, that is, D=0.0045549 (p-value=0.9688) of KS test, and the two samples show the same distribution. Therefore, this training set is used to construct automatic evaluation model of feature

price, automatic evaluation model of random forest and automatic evaluation model of random forest based on rolling time window, and this test set is used for model comparison and analysis.

4.1. Construction of characteristic price model

Multiple linear regression equations were used to establish the characteristic price model for this evaluation object, and 1 dependent variable and 134 independent variables were selected to enter the model. In order to select the best model suitable for sample data, experiments were conducted on Hedonic models with different functional forms, and lm() function in R language was used. Four kinds of characteristic price models in the form of linear model, linear logarithm model, log-linear model and log-linear model are initially constructed, and regression analysis is carried out to further study the model with good fitting effect. The formula of the four characteristic price models is as follows:

$$P = \alpha + YM \cdot \beta + FC \cdot \gamma + FD \cdot \delta + \varepsilon \quad (9)$$

$$P = \alpha + YM \cdot \beta + log(FC) \cdot \gamma + FD \cdot \delta + \varepsilon \quad (10)$$

$$log(P) = \alpha + YM \cdot \beta + FC \cdot \gamma + FD \cdot \delta + \varepsilon \quad (11)$$

$$log(P) = \alpha + YM \cdot \beta + log(FC) \cdot \gamma + FD \cdot \delta + \varepsilon \quad (12)$$

Where, P is the transaction price column vector, YM is the time dummy variable matrix, FC is the continuous attribute feature variable matrix, and FD is the attribute feature dummy variable matrix. The trial fitting results of each form of model are summarized as follows:

Table 2. Hedonic model trial fitting results summary table

Model type	Dependent variable	Significant variable（p-value≤0.05）	Significance（p-value）	AdjustR^2
Linear model	price	129 variables passed the test	< 2.2e-16	0.8924
Linear logarithmic model	price	128 variables passed the test	< 2.2e-16	0.8913
Linear logarithmic model	Ln(price)	130 variables passed the test	< 2.2e-16	0.9204
Logarithmic model	Ln(price)	126 variables passed the test	< 2.2e-16	0.9247

According to the trial calculation results, the above four forms of models have all passed the model test, and the F-test results are significant. From the adjustment of R2 value, the logarithmic form model has a relatively good fitting effect on the price of new housing in District X of City A. Therefore, this paper takes logarithmic model as the representative of characteristic price model to conduct comparative study.

4.2. Construction of random forest automatic assessment model

This study employs the R-language ranger package to train the random forest automated valuation model, utilizing the caret package for hyperparameter optimization. One dependent variable and all quantified independent variables (141 in total) are included in the model. When performing hyperparameter tuning with the train() function from the caret package, the following four hyperparameters are critical for constructing a well-fitted random forest model:

(1) Mtry: The number of features randomly selected from the total feature set at each node split of the regression trees.

(2) Splitrule: The rule governing node splitting (i.e., selection of optimal features for splits).

(3) Min.node.size: The minimum number of samples required in a terminal node.

(4) Num.trees: The number of regression trees in the random forest.

As previously discussed, during node splitting, the random forest randomly selects mtry features from the total feature set as candidate splitting features. The value of mtry influences model complexity and performance: smaller values may lead to underfitting, while larger values risk overfitting. By default, the caret package sets mtry to the square root of the total number of features. However, optimal mtry varies across datasets. Given the total feature count (141), the optimization range for mtry is set to [1, 141], with all possible values evaluated.

Splitrule defines the criterion for node splitting in regression trees. For this regression task, two rules are compared: Variance and Extratrees, as detailed earlier.

Min.node.size specifies the minimum number of samples in a terminal node. If a node's sample size falls below this threshold, splitting stops, and the node becomes a leaf. This parameter controls tree depth: smaller values allow deeper trees (higher complexity), while larger values produce shallower trees (lower complexity). To avoid artificial constraints, min.node.size is set to 1, allowing the algorithm to autonomously determine tree growth.

Num.trees determines the number of regression trees in the forest. Increasing num.trees enhances model stability and robustness but escalates computational costs. Here, num.trees is tested at {500, 1,000, 1,500}. By exhaustively searching across these hyperparameter ranges, 846 distinct random forest models are constructed. Their goodness-of-fit and root mean squared error (RMSE) on the training set are calculated, with the corresponding relationships illustrated in Figures 3 and 4.

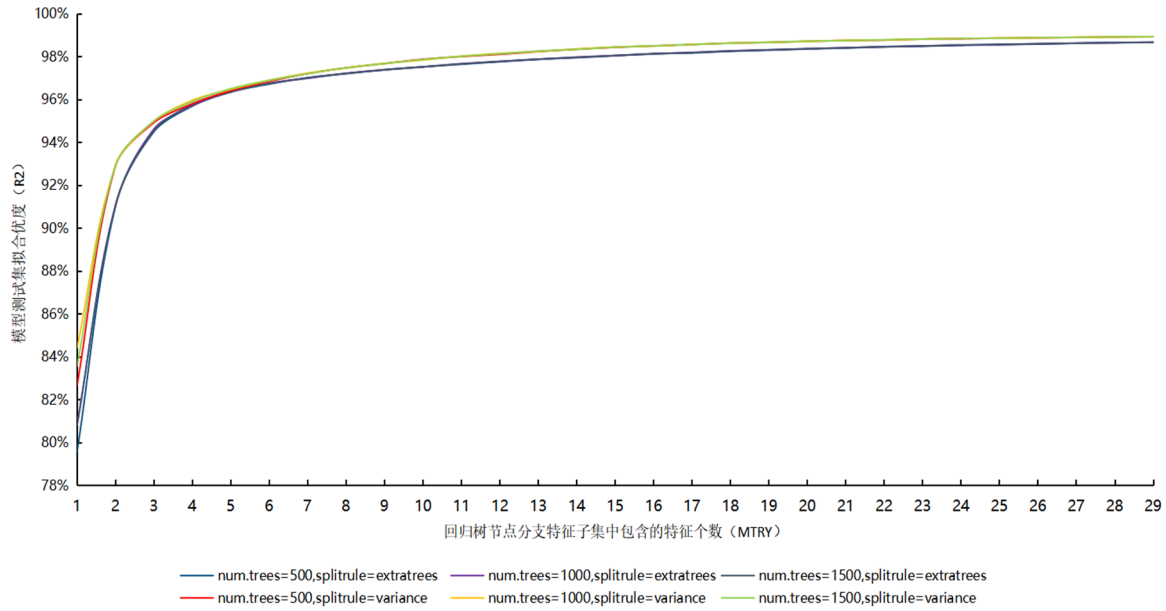


Figure 3. Model training set goodness of fit (R2)

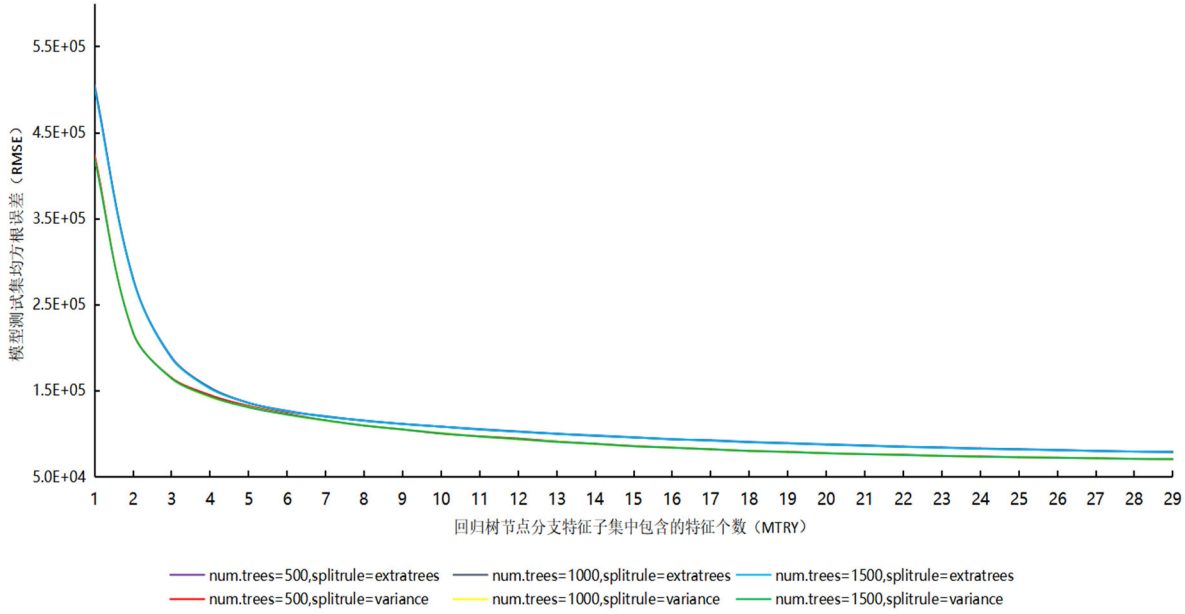


Figure 4. Model training set root mean square error (RMSE)

As illustrated in Figures 3 and 4, the goodness-of-fit and prediction accuracy of the random forest automated valuation model improve with increasing values of mtry (number of features sampled per split), following a nonlinear growth pattern. When mtry rises from 1 to 5, both metrics exhibit a sharp increase, after which the improvement slows and stabilizes. Beyond mtry = 25, the goodness-of-fit and prediction accuracy reach their upper limits. Simultaneously, comparing models with varying num.trees (number of trees) reveals that values exceeding 500 have no significant impact on model performance. Furthermore, the choice of splitrule (node-splitting criterion) critically influences the model's effectiveness. Two key observations emerge: For mtry < 5, models using the Variance splitrule consistently achieve higher goodness-of-fit and accuracy compared to Extratrees under identical parameter settings. As mtry increases toward its limit, models employing Variance still demonstrate superior upper limits in both metrics relative to Extratrees. This underscores the need to carefully select node-splitting rules based on data characteristics and prediction objectives.

Finally, the optimal model is selected from the 846 candidate models by minimizing the out-of-bag (OOB) prediction error. The chosen model parameters are:mtry = 72,num.trees = 1,000,splitrule = "variance",min.node.size = 1.

4.3. Construction of random forest automatic assessment model based on rolling time window

Given that the real estate transaction data collected in this study spans from 2016 to 2022, and recognizing that shifts in market dynamics—such as changes in developer supply and buyer preferences—lead to heterogeneous impacts of property characteristics on market prices across different periods, this study adopts a 12-month rolling time window with a 1-month sliding step to segment the transaction data. This approach aims to enhance prediction accuracy by capturing the varying impacts of real estate characteristics on prices over time. For instance, to train the random forest automated valuation model for June 2018,

transaction data falling between July 1, 2017, and June 30, 2018 is used, and the trained model is exclusively applied to predict prices for June 2018.

After segmenting the data via the rolling time window, the hyperparameter optimization process described earlier is applied to each window's dataset. A total of 61,758 submodels are constructed. From these, the submodel with the smallest out-of-bag (OOB) prediction error within each window is selected for inclusion in the final ensemble. Ultimately, the ensemble comprises 73 optimized submodels.

4.4. Comparison of model evaluation accuracy and generalization ability

To evaluate the prediction accuracy and generalization performance of each automated valuation model, the constructed models were applied to the test set for validation, with results summarized in Table 3. Additionally, the predicted values from the Hedonic automated valuation model and the random forest automated valuation model were visualized against observed values in the test set, as illustrated in Figures 5 and 6.

Table 3. Model Evaluation Validity Summary Table

Model name	R^2	MAE	RMSE	MAPE
Training set				
The characteristic price is in logarithmic form	0.91	118473	203261	9.59
Random forest	0.99	29443	36815	2.68
Random forest based on rolling window	0.98	27566	32255	2.62
Test set				
The characteristic price is in logarithmic form	0.89	120465	235458	10.63
Random forest	0.98	29084	36485	2.64
Random forest based on rolling window	0.99	27425	31216	2.59

As shown in Table 3, across both the test and training sets, the four accuracy metrics consistently indicate that the Hedonic automated valuation model underperforms the random forest and rolling time window-based random forest models in prediction accuracy. A visual comparison of predicted versus observed values in Figures 5 and 6 further demonstrates the superior performance of the random forest models, particularly in handling extreme values. This discrepancy arises because the Hedonic model only captures linear relationships (or linearizable relationships) between property prices and features, neglecting nonlinear and interactive effects of property characteristics on prices [22]. In contrast, the random forest and rolling window-based random forest models account for interdependencies among features, enabling deep learning of complex relationships between features and prices, thereby yielding more accurate predictions. In the test set, both random forest models exhibited excellent performance: The mean absolute percentage error (MAPE) was 2.64% for the standard random forest model and 2.59% for the rolling window-enhanced version. For a property valued at CNY 1 million, the prediction intervals are CNY 0.9736–1.0264 million (random forest) and CNY 0.9741–1.0259 million (rolling window random forest).The improved accuracy of the rolling window model stems from its ability to dynamically capture the varying impacts of property characteristics across different market periods, thereby enhancing predictive power in long-term temporal data.

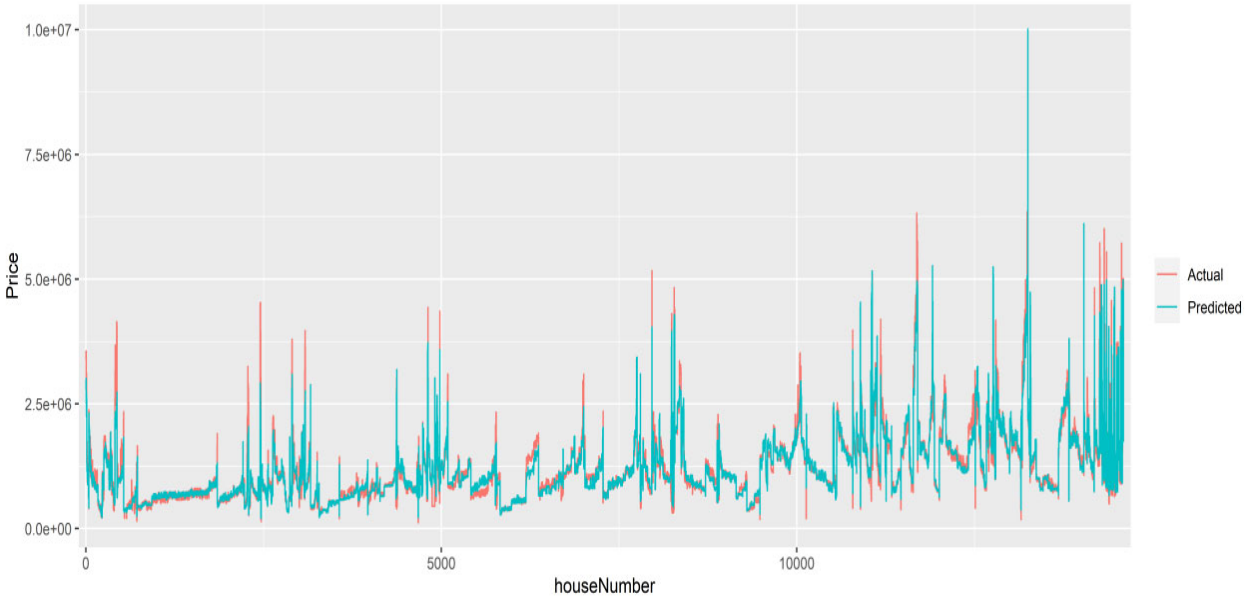


Figure 5. Hedonic model predicted values versus observed values

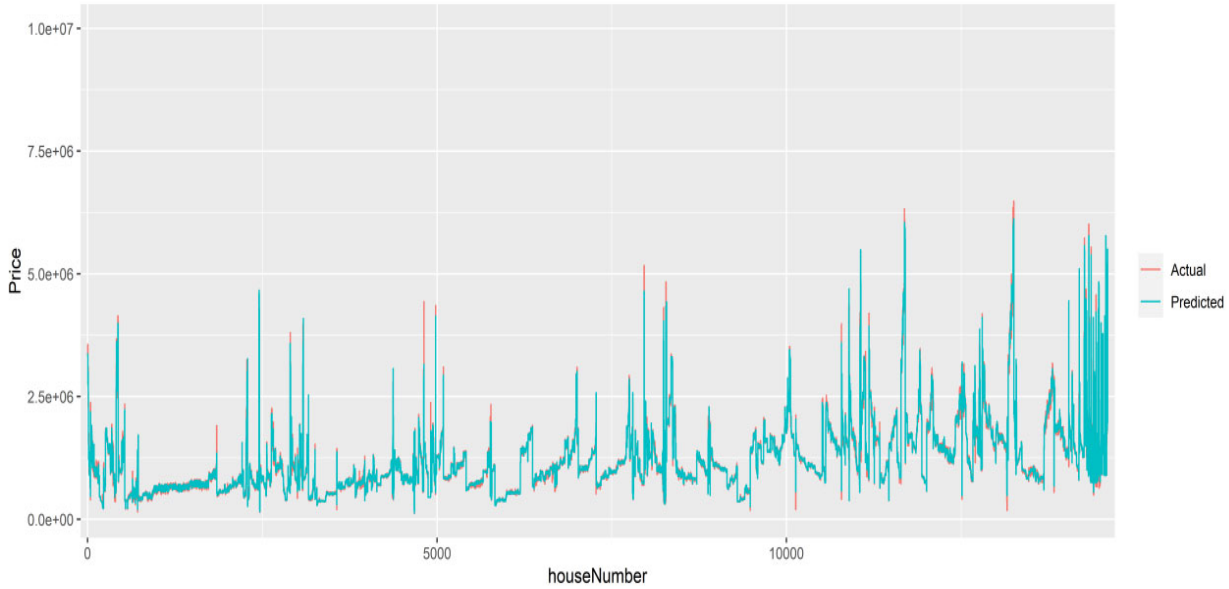


Figure 6. Random forest model predicted values versus observed values

Regarding model generalization capabilities, as evidenced by the performance metrics in Table 3 across training and test sets, the Hedonic model exhibits significant overfitting. This is attributed to the high dimensionality of variables in this study, which increases model complexity. Excessive variables introduce noise or redundancy, making the model overly flexible and prone to overfitting random fluctuations in the training data. Furthermore, the inclusion of numerous dummy variables for scientific quantification introduces multicollinearity (highly correlated variables), further exacerbating overfitting risks. In contrast, the random forest model demonstrates robustness to noise and outliers while capturing nonlinear relationships between features and prices. By aggregating predictions across multiple regression trees, it mitigates errors inherent in individual models. The results show that, except for the R^2 metric, the random forest model outperforms its training set performance in the test set for all other three metrics. Notably, the rolling time window-enhanced random forest model surpasses its training set performance across all four metrics in the test set.

In summary, the random forest automated valuation model significantly outperforms the Hedonic model in both prediction accuracy and generalization capability. The integration of a rolling time window further enhances the random forest model's performance, making it a robust tool for automated tax-assessable price valuation of real estate in the study region.

4.5. Model economic interpretation

Unlike the "black-box" nature of many other machine learning models, the random forest model can evaluate variable importance through metrics such as node impurity or mean squared error (MSE), thereby identifying features with significant impacts on real estate prices and illuminating the underlying mechanisms of price formation. Among these, floor area emerges as the most critical determinant of property prices, followed closely by the number of sports and leisure facilities within a 1,000-meter radius. Other influential variables include residential complex scale, floor area ratio (FAR), and locational factors—particularly those related to environmental quality, urban vibrancy, and proximity to public amenities such as schools and hospitals.

Overall, locational factors exert a stronger influence on property prices than individual property attributes, a finding that underscores the critical role of location in the real estate market.

A variable importance analysis was conducted on submodels within the rolling time window-based random forest automated valuation model, each responsible for predicting prices in different months. The results reveal significant shifts in variable importance rankings between the December 2017 and December 2021 prediction models. While floor area remained the dominant factor in both periods, the relative importance of other variables diverged markedly. Notably, variables related to bathroom count exhibited higher contributions to property prices in the December 2017 model, reflecting temporal shifts in market preferences for housing characteristics.

These findings demonstrate that the rolling time window framework enhances the model's ability to capture evolving market dynamics, thereby improving its economic interpretability by quantifying how feature preferences vary across different market phases.

5. Conclusions

Under the background of big data era, this paper takes the 72,767 sets of new commercial residential property transaction data from January 2016 to December 2022 in District X of City A as a sample, explores the construction characteristics of the Random Forest real estate price automatic assessment model, and carries out a comparative analysis with the Hedonic model from the three perspectives of assessment accuracy, generalization ability, and economic explanation, and introduces a rolling time window to improve the prediction accuracy and economic explanation ability. At the same time, a rolling time window is introduced to improve the prediction accuracy and economic interpretation ability of Random Forest, and the following conclusions are finally drawn:

(1) Model Construction Insights. The number of features (mtry) sampled during regression tree splitting significantly impacts prediction accuracy. As mtry increases, model performance improves rapidly before stabilizing. The choice of node-splitting rules (e.g., Variance vs. Extratrees) critically

influences prediction accuracy and the model's upper performance limits. Optimal rules should be selected based on data characteristics and prediction goals. The number of regression trees (num.trees) has negligible effects on prediction outcomes.

(2) Performance Superiority. The random forest automated valuation model achieves a 98% goodness-of-fit (R^2) and a 2.7% mean absolute percentage error (MAPE) across training and test sets, outperforming the Hedonic model in both accuracy and generalization.

Variable importance analysis reveals that locational factors (e.g., proximity to amenities, environmental quality) exert a stronger influence on property prices than individual property attributes (e.g., floor area, bathroom count), reaffirming the paramount importance of location in real estate markets.

(3) Enhanced Capabilities via Rolling Time Window.The integration of a rolling time window improves prediction accuracy and economic interpretability by capturing temporal shifts in market preferences for property characteristics. For instance, variables like bathroom count showed higher importance in 2017 models but diminished in 2021 models.Given its superior performance, the rolling time window-enhanced random forest model holds significant potential for mass appraisal of tax-assessable property prices in commercial housing markets.

References

[1] Geng Xing. Assessment Issues in the Levying of Property Tax [J]. Taxation Research, 2004(04): 53-55. (in Chinese)

[2] Wang Xiuli, Li Hengkai, Liu Xiaosheng. GIS-Based Comparative Approach Model for Real Estate Market Valuation [J]. China Land Science, 2011, 25(10): 70-76, 97. (in Chinese)

[3] Geng Jijin. Urban Real Estate Mass Appraisal [M]. Beijing: China Financial Publishing House, 2012. (in Chinese)

[4] Sun Yuhuan. Construction of Hedonic Price Models Based on Massive Transaction Data [J]. Statistics and Decision, 2011(02): 9-13. (in Chinese)

[5] Zhu Chenming. Application of Hedonic Price Model in Mass Appraisal of Property Tax Bases [D]. Baoding: Hebei University, 2014. (in Chinese)

[6] Wang Rui. Construction of BP Neural Network Model for Real Estate Valuation [J]. Journal of Shandong Agricultural University (Natural Science Edition), 2016, 47(6). (in Chinese)

[7] Xiao Chenglian. Research on Mass Appraisal of Residential Tax Bases Using Grey Correlation-Based GA-BP Neural Network [D]. Ganzhou: Jiangxi University of Science and Technology, 2016. (in Chinese)

[8] Wang Azhong, Li Qian. Research on Mass Appraisal of Property Tax Bases Using Rough Set Neural Networks [J]. Journal of Fuzhou University (Philosophy and Social Sciences Edition), 2019, 33(5). (in Chinese)

[9] Chiarazzo V, Caggiani L, Marinelli M, et al. A Neural Network Based Model for Real Estate Price Estimation Considering Environmental Quality of Property Location [J]. Transportation Research Procedia, 2014, 3: 1-10.

[10] Song Zujie. Application of Support Vector Regression-Based Mass Appraisal Model for Second-Hand Housing [D]. Chongqing: Chongqing University, 2016. (in Chinese)

[11] Antipov E A, Pokryshevskaya E B. Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics [J]. Expert Systems with Applications, 2012, 39(2): 1772-1778.

[12] Xu Ge, Zhang Ke. Real Estate Price Valuation Based on Random Forest Model [J]. Statistics and Decision, 2014, (17): 82-85. (in Chinese)

[13] Pang Feng. Application of Random Forest Model in Mass Appraisal of Second-Hand Housing [D]. Chongqing: Chongqing Jiaotong University, 2017. (in Chinese)

[14] Chen Shiqin, Wang Hongwei. Machine Learning-Based Mass Appraisal Model for Real Estate [J]. Statistics and Decision, 2020, 36(09): 26-30. (in Chinese)

[15] Breiman L, Friedman J H, Olshen R A, et al. Classification and Regression Trees [M]. Belmont: Wadsworth, 1984.

[16] Li Hang. Statistical Learning Methods [M]. Beijing: Tsinghua University Press, 2012. (in Chinese)

[17] Liu Yanli. A Review of Random Forest [D]. Tianjin: Nankai University, 2008. (in Chinese)

[18] Kok N, Koponen E, Martinez-Barbosa C A. Big Data in Real Estate? From Manual Appraisal to Automated Valuation [J]. The Journal of Portfolio Management, 2017, 43(6): 202-211.

[19] Calainho F D, Francke M K. A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate [J]. The Journal of Real Estate Finance and Economics, 2022, 64: 1-28.

[20] Han Hui. Application of Hedonic Price Model in Mass Appraisal of Property Tax Bases [D]. Shijiazhuang: Hebei University of Economics and Business, 2014. (in Chinese)

[21] Bai Xue. Research on Mass Price Appraisal of Existing Housing in Nantong Based on Hedonic Price Model [D]. Hefei: Hefei University of Technology, 2015. (in Chinese)

[22] Del Giudice V, De Paola P, Forte F, et al. Real Estate Appraisals with Bayesian Approach and Markov Chain Hybrid Monte Carlo Method: An Application to a Central Urban Area of Naples [J]. Sustainability, 2017, 9(12): 1-15.