

## A HOUSE PRICE VALUATION BASED ON THE RANDOM FOREST APPROACH: THE MASS APPRAISAL OF RESIDENTIAL PROPERTY IN SOUTH KOREA

Jengei HONG <sup>1</sup>, Heeyoul CHOI <sup>2</sup>, Woo-sung KIM <sup>1,\*</sup>

<sup>1</sup> School of Management and Economics, Handong Global University, Pohang, Republic of Korea

<sup>2</sup> School of Computer Science and Electrical Engineering, Handong Global University, Pohang, Republic of Korea

Received 04 June 2019; accepted 14 October 2019

**Abstract.** Mass appraisal is the standardized procedure of valuing a large number of properties at the same time and is commonly used to compute real estate tax. While a hedonic pricing model based on the ordinary least squares (OLS) linear regression has been employed as the traditional method in this process, the stability and accuracy of the model remain questionable. This paper investigates the features of a house price predictor based on the Random Forest (RF) method by comparing it with that of a conventional hedonic pricing model. We used apartment transaction data from the period of 2006 to 2017 in the district of Gangnam, one of the most developed areas in South Korea. Using a data set covering 40% of all transactions in the sample area, we demonstrate that the accuracy of a machine learning-based predictor can be surprisingly high. The average of percentage deviations between the predicted and the actual market price was found to be only around 5.5% in the RF predictor, whereas it was almost 20% in the OLS-based predictor. With the RF predictor, the probability of the predicted price being within 5% of its actual market price was 72%, while only about 17.5% of the regression-based predictions fell within the same range. These results show that, in the practice of mass appraisal, the RF method may be a useful complement to the hedonic models, as it more adequately captures the complexity or non-linearity of actual housing markets.

**Keywords:** housing price forecasting, hedonic pricing model, random forest approach, mass appraisal, apartment, machine learning technique.

### Introduction

Mass appraisal, also called automatic valuation of real estate assets, is the introduction of mathematical statistics, computer technology, and geographic information technology to establish a mathematical model that serves as a systematic appraisal of a group of real estate properties and reveals its market value (Zhou, Ji, Chen, & Zhang, 2018). The Basel II Accord, issued by the Basel Committee on Banking Supervision (BCBS) in 2008, states that “the bank is expected to monitor the value of the collateral on a frequent basis and at a minimum once every year. More frequent monitoring is suggested where the market is subject to significant changes in conditions. Statistical methods of evaluation (e.g. reference to house price indices, sampling) may be used to update estimates or to identify collateral that may have declined in value and that may need re-appraisal. A qualified professional must evaluate the property when information indicates that the

value of the collateral may have declined materially relative to general market prices or when a credit event, such as default, occurs.” As a result of this accord, the value of the property is appraised more frequently than before, which leading to an increase in costs (time and money) for appraisals. Thus, a stable, accurate, and fast tool for appraisal is needed, and the mass appraisal model may be a viable solution. The mass appraisal model is also a widely accepted tool for the valuation of property for the purposes of taxation or mortgage for a loan. In the US, the Computer Assisted Mass Appraisal (CAMA), the computer system and software used for mass appraisal, is employed nearly universally by assessors nationwide for tax assessment. Owing to its importance, there has been a rich and diverse body of work addressing the appraisal techniques and performance.

Traditionally, the hedonic pricing model, originating from Lancaster’s consumer theory, has been one of

\*Corresponding author. E-mail: [wskim@handong.edu](mailto:wskim@handong.edu)

the most extensively employed models to estimate house prices and property values (Lancaster, 1966). The theoretical framework and foundation for hedonic pricing models were developed in a study by Rosen (1974). In hedonic price theory, it is assumed that a good can be regarded as a bundle of individual components or characteristics that provide utilities. Rosen (1974) defines the theory as “a model of product differentiation based on the hedonic hypothesis that goods are valued for their utility-bearing attributes or characteristics.” Rosen defines a set of “hedonic” prices as the amount of characteristics associated with the goods. Thus, a consumer who purchases a good acquires a collection of the characteristics embodied in it, and these attributes can be converted into utility. From this perspective, a house is a heterogeneous good embodying a package of inherent characteristics relevant to location, property attributes, and environmental amenities. The advantage of the hedonic pricing models is that the marginal implicit values of the characteristics can be obtained by differentiating the price function with respect to each attribute (McMillan, Reid, & Gillen, 1980).

Because house prices are influenced by a number of attributes, many studies employ the hedonic model to investigate relationship between house prices and their characteristics (Chau & Chin, 2003). The most common variables for the model involve the structural attributes, such as type, age of property, number of bedrooms and other rooms, and other amenities available within the property. Numerous studies have found a house’s number of bedrooms and bathrooms and its floor area to be positively related to its price (Fletcher, Gallimore, & Mangan, 2000; Li & Brown, 1980; Garrod & Willis, 1992; Rodriguez & Sirmans, 1994). Kain and Quigley (1970) revealed that the age of the property can impact house prices negatively. In addition, some researchers have analyzed the impact of locational features, such as racial composition, pollution level, and proximity to a central business district (CBD), transportation facilities, or retail stores on house prices (Palmquist, 1992; McMillan, Jarmin, & Thorsnes, 1992; Ridker & Henning, 1967). Dubin and Sung (1990) conducted a non-nested test to determine which set of neighborhood variables most accurately explained the variation in housing prices. To reveal the relationship between accessibility to a CBD and house prices, various measures have been proposed (Adair, McGreal, Smyth, Cooper, & Ryley, 2000; Hanson, 2004; Song & Sohn, 2007; Chen, Ong, Zheng, & Hsu, 2017). In So, Tse, and Ganesan (1997) and Debrezion, Pels, and Rietveld (2007), the effect of the proximity of public transportation infrastructure on house prices was studied. Location on a site with a desirable view, such as a lake or golf course, has been found to have a positive effect on the price in Benson, Hansen, Schwartz, and Smersh (1998), Gillard (1981), and Darling (1973). The neighborhood attributes can be implicitly valued through the hedonic model by comparing properties with differing neighborhood qualities (Goodman, 1989). Chau and Chin (2003) reviewed past studies and classified the attributes into three categories: socioeconomic variables (Garrod &

Willis, 1992; Richardson, Vipond, & Furbey, 1974), local government or municipal services (Clauret & Neill, 2000; Hayes & Taylor, 1996; Jud & Watts, 1981; Downes & Zabel, 2002; Huh & Kwak, 1997), and externalities such as crime rates (Thaler, 1978), noise (Wilhelmsson, 2000; Williams, 1991; Espey & Lopez, 2000), and air pollution (Harrison & Rubinfeld, 1978). Previous studies have suggested some key housing attributes included in most hedonic price models.

While the major advantage of the hedonic models is their simplicity in estimating and interpreting the regression coefficients, the pre-specified form of the models has been criticized for imposing strong assumptions, such as those regarding linearity parameters. The functional form of the conventional hedonic pricing model is based on the simplification of household’s preferences and strict assumptions about the housing. The model depends on the assumption that the effects from each attribute are separable and constant, which implies a separable preference, perfect competition, market equilibrium, and an integrated market (see Chau & Chin, 2003; Malpezzi, 2002; Sheppard, 1999). Thus, in practice, the accuracy of the OLS (ordinary least squares)-based model would be eroded insofar as the model simplifies the complexity or non-linearity of the real world. For instance, if the housing market is organized into a series of sub-markets by housing size or income group or if there is a non-linearity in household’s preference on an attribute, the predictor obtained from a single regression would fail to capture the complexities. This problem arises because we cannot directly observe the structure of preference and capture all the market characteristics causing the complexity in a market. In the real world, many market characteristics may intermingle, but there is no flexibility in the conventional hedonic pricing model to explore such complexity. These disadvantages are mentioned in Zurada, Levitan, and Guan (2011) as “failures [that] would result in untenable or imprecise coefficients caused by functional form misspecification, interaction among variables, multicollinearity, and non-linearity problems.”

In this case, the proposed data-driven modelling based on machine learning techniques could be a complement to the conventional regression methods. The main advantage of the proposed method is that it constructs the model, while exploring the complexity, without the modeler explicitly describing it. In recent years, the applicability of these methods has been expanding quickly, owing to the developments in data collection. In academic research on real estate, the application of machine learning techniques has grown (Fan et al., 2006; Selim, 2009; Antipov & Pokryshevskaya, 2012; Čeh, Kilibarda, Lisec, & Bajat, 2018). As discussed in Fan, Ong, and Koh (2006), the approach can be applied to investigate the linear or non-linear relationships between the dependent and independent variables and hierarchical structure of the determinants of house prices. In McCluskey and Anand (1999) and Verikas, Lipnickas, and Malmqvist (2002), artificial neural network models were employed to value properties. Limsom-

bunchai (2004) and Selim (2009) compared the predictive power of the hedonic model based on multiple regression with that of an artificial neural network model. Both studies demonstrate that an artificial neural network can be a more effective alternative to the hedonic model for appraising house prices. In Gu, Zhu, and Jiang (2011) and Mu, Wu, and Zhang (2014), supporting vector machine techniques were used to value house prices. Park and Bae (2015) developed a housing price appraisal model based on machine learning algorithms, such as C4.5, RIPPER, Naïve Bayesian, and AdaBoost, and analyzed the housing data for Fairfax County, Virginia, USA.

In spite of the wide application of machine learning techniques in house price valuation, there have been few studies using Random Forest (RF) techniques for appraisal. The RF method is a special type of the simple regression trees ensemble, which gives a prediction based on majority voting or by averaging predictions made by each of its trees (Antipov & Pokryshevskaya, 2012). The benefit of RF is that there are few hyperparameters with the potential to strongly influence its performance. It is defined only by the number of trees and the depth of each tree. Antipov and Pokryshevskaya (2012) “believe random forest may become one of the most appropriate techniques for mass appraisal ... it is expected to avoid fallacies of many other methods, commonly used for mass appraisal.” They also presented several advantages of RF. First, in many comparative studies, RF performed more strongly than other algorithms. Second, it can successfully manage categorical variables with many levels. In the case of multiple regression or neural networks, a large number of qualitative variables lead to a larger number of estimated parameters, which usually results in overfitting. Third, the method works adequately when there is missing data. Because the method is based on regression trees, the prediction is made from the part of the tree that has already been built, even when some data is missing. Fourth, it allows for nonlinear links and unsteadiness of variable influence across different segments. Fifth, its method does not require a detailed model specification. Thus, the RF method may become one of the most appropriate methods for mass appraisal, and it is for this reason that it was chosen for this paper.

In this paper, we investigate the features of a house price predictor based on the RF method by comparing it with those of a conventional, regression-based hedonic pricing model. We collected a data set covering 40% (16,601 samples) of all apartment transactions (39,564) during 2006–2017 in the district of Gangnam, one of the most developed areas in South Korea. The samples were randomly divided into a training set consisting of 90% of all transactions and a test set consisting of the remaining 10% of transactions. We compare several performance measurements for the predictions of the house prices in the test set. The results show that the machine learning approach can significantly enhance predictive performance. The average percentage deviation between the predicted

and the actual market price was only around 5.48% in the machine learning predictor. Further, the probabilities that the RF predictions fell within 3%, 5%, 10% of the actual market price were 53.5%, 71.9%, and 90.3%, respectively, while those of the OLS-based predictions were 10.4%, 17.4%, and 34.6%, respectively. Furthermore, we found that the RF predictor makes fewer outlier predictions than the conventional hedonic pricing model. The probability of the RF predictions deviating more than 50% from the actual price was only 0.5%, while that for OLS-based predictions was almost 3.8%.

The following can be derived from our results. From a theoretical perspective, the result may serve as evidence of high complexity in the price determination process of the housing market. The superiority of RF in appraisal accuracy indicates that the RF predictor can more successfully track the actual price determination process in housing market than the OLS predictor. In other words, there are some factors of the value determination process that cannot be fully explained in the simplified assumptions of the conventional hedonic pricing model (e.g., separability and constancy of an attribute’s effect on housing value).

From a practical perspective, our results show that the quality of mass appraisals or house price indices can be significantly improved by using the RF method. Relative to the predictive models in previous studies in Limsombunchai (2004), Selim (2009), Antipov and Pokryshevskaya (2012), and Čeh et al. (2018), the performance measures of RF— $R^2$  values (97.6%), mean absolute percentage error (MAPE, 5.482%), coefficient of dispersion (COD, 5.484%), and hitting rate—achieved significantly stronger results. Although it is difficult to compare experiments conducted in different samples, the results in this paper may also indicate that the accuracy of systemic appraisal can be surprisingly high (Note that the MAPE of human appraisals is 12% in Cannon and Cole, 2011).

We infer that the high predictive power of the RF-based model derives from a combination of the features of the RF method and the features of the data set we applied. In the RF method, a model is constructed by exploring the hierarchical structure of characteristics and the effect of each attribute on price is allowed to vary according to circumstances. The important advantage of this method is that it does not require assumptions about market complexity. The RF algorithm constructs the data-driven hierarchical structure of the model without the modeler explicitly describing it. Therefore, if the data set sufficiently covers the characteristics of the property, the RF model is expected to more sensitively replicate the complex structure of the house price determination process.

In addition, we presume that the features of our data also contributed to the high accuracy for two reasons. One is the geographic density of the samples. A large portion of a property’s value comes from its location. If the samples are sparsely located in a large area, it is difficult to accurately measure the effects related to location. We collected a relatively large sample (16,061 samples trained) in a

small area (39.55 km<sup>2</sup>) and expect that this high density of samples may have contributed to the high accuracy of our prediction. The other reason is the type of property that our data covers. We collected all of our apartment data in the same residential area (the district of Gangnam in Seoul), and the structural characteristics of the apartments can be sufficiently represented by a number of common and measurable features. A data set can contain only consolidated features of housing, such as number of rooms and floor level. Housing in different residential areas or in detached dwellings are usually more various in their amenities, interior decorations, and features and consequently are difficult to codify or consolidate in a data set, which eventually undermines the accuracy of predictors. In this context, we expect that our data on apartments in the same residential area (with a similar income group) would contribute to the accuracy of prediction.

The remainder of this paper is organized as follows. In Section 1, the data set and some basic statistics are described. In Section 2, we introduce the RF method and describe how it predicts house prices. Section 3 provides the quantitative results and interpretation. Concluding remarks are provided in final section.

## 1. Data set and basic statistics

Gangnam is one of the 25 local government districts of Seoul, the capital city of South Korea. With a population of 561,052 and an area of 39.5 km<sup>2</sup>, it is Seoul's third-largest district. The district is composed of 22 administrative divisions called "dongs" (Figure 1). While Seoul is known for its high housing prices (an average apartment cost approximately 5,500 USD per m<sup>2</sup> in 2011), the average housing price in Gangnam—approximately 10,000 USD per m<sup>2</sup>—is almost twice as high, and 3.5 times the national average. The district is also the place where the largest number of apartment transactions have occurred in the past decade. We collected 16,601 samples for 2006–2017 from the transaction records for apartments in Gangnam,

provided by South Korea's Ministry of Land, Infrastructure, and Transport (MOLIT). The data set covers about 40% of all apartment transactions in Gangnam during the sample period.

Because both models involve the regression of observed apartment prices against apartment attributes and economic variables hypothesized to be determinants of price, the factors assumed to contribute to the price are given in Table 1.

The structural attributes are related to inherent characteristics of the property. In this study, they include elapsed year (transaction year–construction year), area, floor level, and heating system. Regarding the heating system, the value of the dummy variable is set to 0 if an apartment has a central heating system. Otherwise, the value is set to 1.

For neighborhood attributes, we consider apartment brand, available units in the building, number of buildings in the apartment complex, parking lot, floor area ratio, building coverage ratio, and the top/lowest floor of the building. A dummy variable is employed for the ranking of apartment brands. The ranking is based on a report by the Korea Institute of Corporate Reputation, and the variable has a value of 1 if an apartment is not built by one of the ten highest-ranked apartment brands. The variable "parking lot" represents the average number of parking spaces available per apartment household. Floor area ratio (FAR) and building coverage ratio (BCR) are the ratio of total floor area (gross floor area) to land area and the ratio of the building area divided by the land (site) area, respectively.

The locational attributes of property, which also affect the price of the property, are considered in this study. To take the value of the geographical position into account, we consider latitude, longitude, and accessibility to nearby facilities. The facilities considered are national park, high school, redevelopment area, university, general hospital, museum, and subway station. While the information on the administrative division of the apartment was found in the data provided by MOLIT, other information (latitude,

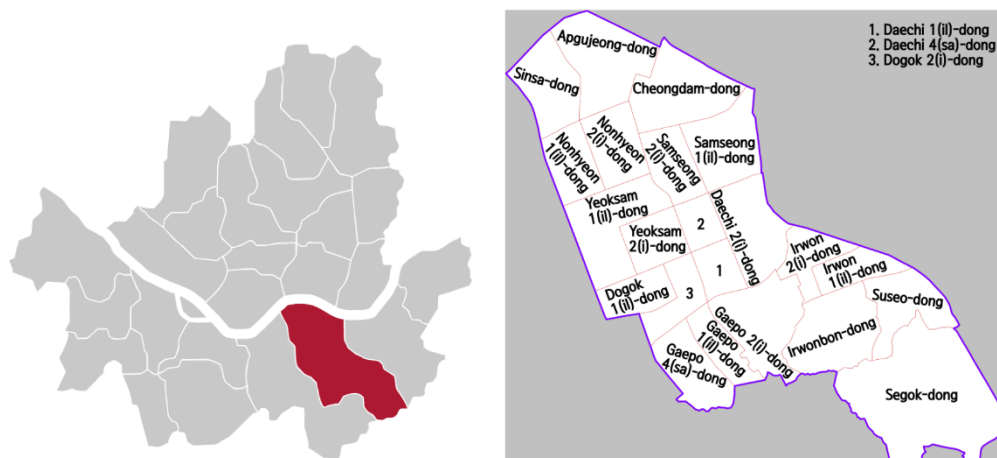


Figure 1. Location of Gangnam and its administrative divisions (Wikipedia)



Table 1. Variables used for forecasting house price

Category	Variables	Unit
Structural attributes	Elapsed year (transaction year-construction year) Area Floor level of a property Heating system	year m <sup>2</sup> floor level 0 if an apartment has a central heating system. Otherwise, the value is set to 1
Neighborhood attributes	Apartment brand  Number of units in the apartment complex Number of buildings in the apartment complex Parking lot Floor area ratio Building coverage ratio The top floor of an apartment The lowest floor of an apartment	0 if an apartment is built by one of the ten highest-ranked apartment brands Otherwise, the value is set to 1 number of units number of buildings number of parking spot /number of units ratio ratio floor level floor level
Locational attributes	Latitude Longitude Distance to national park Distance to high school Distance to redevelopment area Distance to university Distance to general hospital Distance to museum Distance to subway station	latitude of a property longitude of a property meter meter meter meter meter meter meter
Macro variable	Transaction period Gross domestic product (GDP) Annual growth rate in real GDP Land price fluctuation rate in Seoul Mortgage interest rate	year billion won % % %

Table 2. Descriptive statistics

Variables	Mean	Median	Standard deviation	Min	Max
Construction year	1992.946	1993	10.270	1978	2014
Area	71.567	59.96	35.0495	16.78	273.83
Floor level	7.464	5	5.633	-1	45
Units available in the building	1724.093	900	1855.321	7	5040
Number of buildings in an apartment complex	30.404	8	42.861	1	124
Parking lot	1.006	1	0.591	0.27	4.53
Floor area ratio	252.022	224	209.46	72	2435
Building coverage ratio	24.687	19	15.046	12	204
Latitude	37.494	37.493	0.0120	37.460	37.533
Longitude	127.060	127.058	0.0170	127.0181	127.104
Distance to national park	1065.147	1053.208	396.087	86.108	2142.469
Distance to high school	536.979	522.609	236.370	31.883	1531.516
Distance to redevelopment area	634.044	571.583	416.609	0	3,758.560
Distance to university	3,382.976	3,551.466	1272.367	24.587	7,136.498
Distance to general hospital	1,062.366	975.124	524.585	41.633	3,470.830
Distance to museum	986.106	1,032.572	373.803	87.490	3,323.865
Distance to subway station	678.640	579.455	394.342	47.487	2,559.068
GDP (billion won)	333,427	337,411	64,322	225,613	446,835
Growth rate in real GDP	3.641	3.4	1.791	-1.9	7.4
Land price fluctuation rate	0.0638	0.0165	0.319	-2.643	0.625
Mortgage interest rate	6.050	5.883	0.602	5.263	7.415

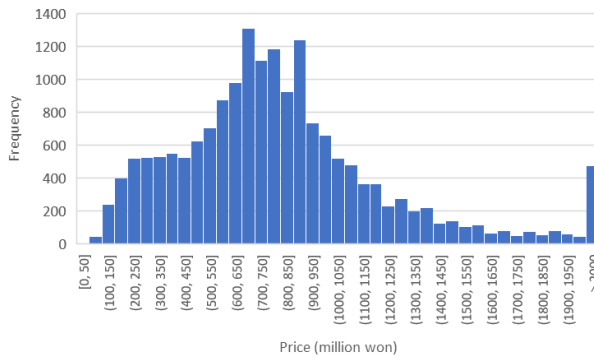


Figure 2. Histogram of traded apartment prices

longitude, and distance to nearest facilities) was calculated by using values obtained through the MAP open Application Programming Interface (API).

It has been observed in previous studies that macroeconomic variables may also affect the housing market (Case, Quigley, & Shiller, 2005; Miller, Peng, & Sklarz, 2011). As relevant macroeconomic factors, we consider the transaction period, the size of the economy (gross domestic product), business cycles (% growth rate in the real gross domestic product), land price fluctuation rate in Seoul, and mortgage interest rate. The values of the variables are measured for each year. The descriptive statistics of the data are given in Table 2. A summary of the traded apartment prices is shown as a histogram in Figure 2.

## 2. Model description

### 2.1. Conventional hedonic pricing model

We considered a conventional hedonic pricing model estimated by OLS regression. The hedonic pricing model is theoretically based on Lancaster (1966) and Rosen (1974). In Lancaster's characteristics demand theory, consumers are described as deriving utility not from goods themselves but from their characteristics. Thus, the consumption of a good can be considered the consumption of the composite attributes of the good. Rosen (1974) extended the characteristics demand theory to the hedonic pricing model. He suggested that the value of a good can be divided into the values of its attributes. Under the assumption that each attribute has a unique implicit price in an equilibrium market, the price of good can be interpreted as the sum of the attribute prices, implying that the price of a good can be regressed on the characteristics.

However, these theories provide little specification for the functional form and list of variables considered. In this paper, we start with the conventional assumption for the hedonic pricing model, which can be expressed as:

$$p = X\beta + \varepsilon, \quad (1)$$

where:  $p$  represents a  $n \times 1$  vector of the natural logarithm of apartment prices.  $X$  is the matrix containing explana-

tory variables.  $\beta$  is the coefficient vector corresponding to  $X$ , and  $\varepsilon$  is the vector of the white-noise error.

Since the main purpose of this paper is to compare the performances of OLS and RF predictors, we set the explanatory variables considered in this hedonic pricing model to be the same as the variables considered in the RF model. However, when the hedonic pricing model includes time dummy variables, it is meaningless to include each macroeconomic variable, such as annual GDP growth rate, because the effects from those variables are already embedded in the coefficients of the time dummy. Therefore, in this paper, we set the hedonic pricing model to include the time dummy variable and no other macroeconomic variables, even if the RF model explicitly uses several macroeconomic variables.

The predictor of this OLS-based approach,  $\hat{p}$ , can be conveniently obtained from the expression as:

$$\hat{p} = X(X'X)^{-1}X'p. \quad (2)$$

### 2.2. Random forest

Decision trees (DTs) are decision support tools based on tree-like graph models, in which each branch represents a decision result on a feature and its threshold. For example, suppose that a node has a branch based on a feature  $A$  with threshold  $T$ . If a new sample's feature  $A$  is lower than  $T$ , then it takes the left branch; otherwise, it takes the right one. DTs build classification or regression models. For classification models, each leaf node of the tree represents a class, and classification is based on following the branches from the root node to a leaf node. Regression is based on local linear regression in the divided subspaces defined by leaf nodes after following the branches, as in classification.

To train trees, one should select one feature and one threshold at a time to make a branch at a node such that each branch has similar samples after the split. There are a couple of metrics, including standard deviation reduction, for choosing a feature for a new branch. The tree grows in depth by adding one new node at a time.

RF is an ensemble of DTs, which gives a prediction based on averaging (the case of regression) predictions made by each tree in the ensemble using some input data. Figure 3 depicts an example of RF. When given training data, RF runs the trees first. All of the trees in the ensemble are built independently according to the algorithm as follows. Let  $P$  denote the set containing all predictors. A subset of  $P$ , randomly chosen predictors, is used to grow each tree on a bootstrap sample of the training data. For each of the bootstrap samples, an unpruned regression tree is grown. After a large number of trees are generated, predictions are averaged over the different trees.

Since RF is a decision tree-based technique, it has some advantages in our mass appraisal problem. In RF, a categorical variable with  $n$  classes is recoded into  $n-1$  dichotomous ones, only a fraction of which is used in building a tree (Antipov & Pokryshevskaya, 2012). This

helps to avoid overfitting problems caused by the large number of classes. In the case of multiple regression or neural networks, such categorical variables lead to an increased number of estimated parameters, which results in overfitting. Since there are qualitative variables, such as apartment brand and heating system, in our problem, RF techniques can be advantageous for predicting the price of a property. RF can also deal with nonlinear links and the unsteadiness of variable influence across different segments, since it is based on regression trees. In many previous studies on mass appraisal, the predictive power of models based on nonparametric methods, such as neural network or support vector machine, is greater than that of OLS-based models. It seems that there are significant market complexities that cannot be fully explained using the conventional hedonic pricing model. RF is more appropriate for dealing with this complexity.

Another benefit of DTs and RFs is the interpretability of the trained model: humans can understand how the trained model works. In addition, trees are trained easily and make faster inferences than other machine learning algorithms. To train RF models, there are only two hyperparameters: number of DTs and depth of each tree. With more DTs, the result would be more stable in their computation cost, and deeper trees find more accurate results by dividing the sample space into smaller parts, which may lead to overfitting. In our experiment, after trying many different combinations, the RF model consisted of 50 trees with a depth of 17, although there is no significant difference in performance with slightly different combinations. In this study, we used the sklearn toolkit from scikit-learn.org.

### 3. Results

#### 3.1. Feature selection

We investigated the 26 variables in Table 3 to determine which of them have a dominant or significant impact on the price. We fixed a random forest architecture from 50 decision trees with depth 17, after many trials with different configurations on training and validation samples. Once training the RF model with the training samples, the

model has importance values which indicate the predictive power of the variables—that is, how much the variable decreases variance (or error) in the split space. In decision trees, every node is a condition of how to split values in a single feature, so that similar values of the dependent variable (price) belong to the same set after the split. The condition is based on impurity, which is Gini impurity in case of classification problems, while mean squared error (MSE) and its variance are used for regression trees. So when a tree is trained, the importance is how much each feature contributes to decreasing the weighted impurity. In the case of Random Forest, we use the average of the decrease in impurity over trees by a feature as the importance of the feature.

Figure 4 shows the importance of the variables in the trained RF model. Note that “area” is the most important factor for price, followed by “number of buildings in the apartment complex”. “Transaction date” and “construction year” are also significant. Interestingly, distances to places of interest, such as a subway station, seem to have no effect on price.

We selected features based on performance while training the RF model after removing the least important variables one at a time. To measure its performance, we used mean absolute percentage error (MAPE), a straightforward measurement that captures the average percentage deviation of predictions from the actual transaction prices. The formula is expressed as:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{\hat{p}_i - p_i}{p_i} \right|, \tag{3}$$

where:  $p_i$  and  $\hat{p}_i$  are the actual price and predicted price of apartment  $i$ , respectively.

Figure 5 shows the MAPE curve with different numbers of variables, while removing the least important variables one at a time. The horizontal axis indicates the number of variables used in prediction. Notice that the error is minimal when 16 features are used (or 10 features are removed) to train the RF model. By this feature selection, we can avoid potential overfitting problems. The 16 features are listed in Table 4. We use these 16 features for the subsequent experiments.

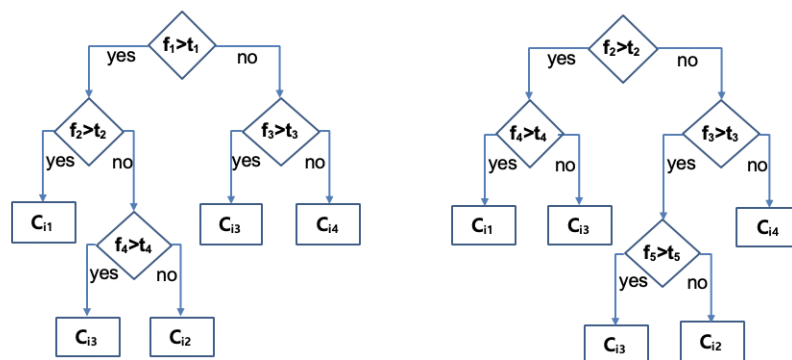


Figure 3. Example of RF: an ensemble of 2 decision trees (DTs) with a depth of 4 on 5 features

Table 3. List of attributes

Index	Description
1	Transaction date
2	Elapsed year
3	Area
4	Floor level of a property
5	Latitude
6	Longitude
7	Gross domestic product (GDP) in the quarter
8	Economic growth rate during the quarter
9	Land price fluctuation rate in Seoul
10	Mortgage interest rate
11	Distance to national park
12	Distance to high school
13	Distance to redevelopment area
14	Distance to university
15	Distance to general hospital
16	Distance to museum
17	Distance to subway station
18	Apartment brand
19	Units available in the apartment building
20	Number of buildings in the apartment complex
21	Parking lot
22	Heating system
23	Floor area ratio
24	Building coverage ratio
25	The top floor of an apartment
26	The lowest floor of an apartment

Table 4. The 16 features after feature selection

Selected feature description
Area
Number of buildings in the apartment complex
Transaction date
Elapsed year
The lowest floor of an apartment
Units available in an apartment
Parking lot
The top floor of an apartment
Gross domestic product (GDP) in the quarter
Floor area ratio
Building coverage ratio
Land price fluctuation rate in Seoul
Mortgage interest rate
Floor level of a property
Longitude
Apartment brand

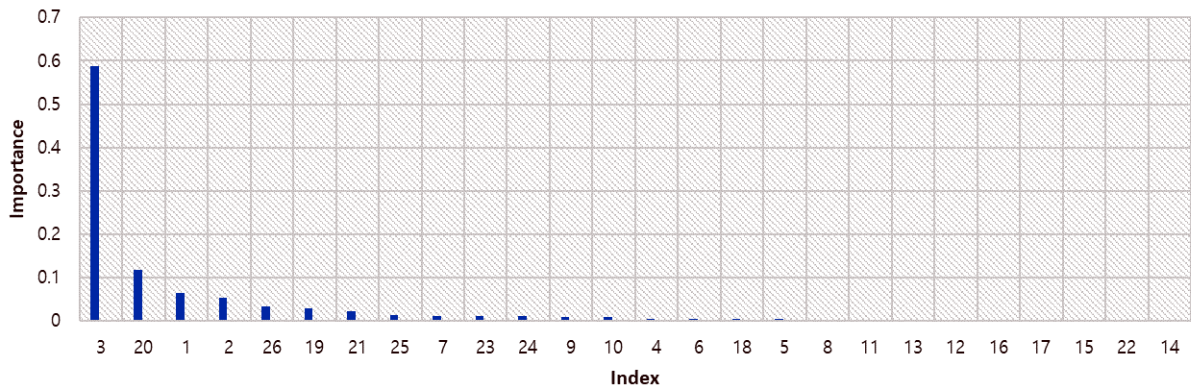


Figure 4. Importance of features in the trained RF model. “Importance” means the contribution of each variable to the model

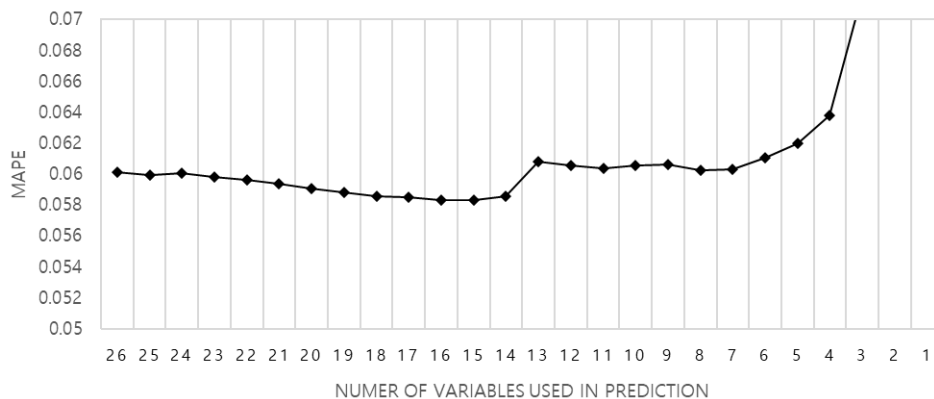


Figure 5. MAPE when removing the least important variables one at a time



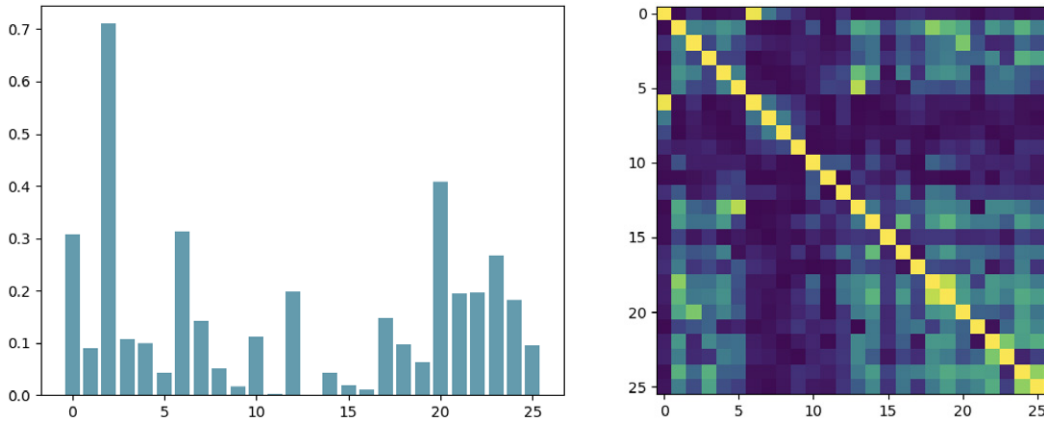


Figure 6. Absolute values of correlation between variables and prices (left) and between variables (right). The brighter cells have higher correlations

On the other hand, we also need to look at the correlations between the variables and prices (target) and those between variables. Even when the importance is low, some variables can have strong predictive power on the price if they are not correlated to other variables. In Figure 6 (left), “parking lot” (index 21) has a strong correlation with price, while it is not important in Figure 4. This phenomenon can be explained by Figure 6 (right), where “parking lot” has a strong correlation to “area” (index 3), probably because “area” includes much of the same information contained in “parking lot.” Thus, “parking lot” is not important when “area” is one of factors.

### 3.2. Comparison between RF and OLS predictor

The predictive performances of OLS and RF regression can be compared using measurements that capture the distance between predicted and observed transaction price. We considered three measurements: MAPE, coefficient of dispersion (COD), and R-squared.

MAPE measures the average percentage error of predictions from the actual transaction prices. Percentage errors from each sample are averaged after taking absolute value to ignore the sign of the errors. MAPE is frequently used because it is convenient and can be understood intuitively. Its formula is shown in the equation (3).

COD measures the dispersion of sales ratio, the quotient obtained by dividing the predicted price with actual transaction price, around the median sales ratio. It is used to measure appraisal uniformity. It is obtained by the average percentage deviation of sales ratio from the median value; thus, lower COD implies a more uniform prediction. This measurement can be expressed as:

$$\text{COD} = \frac{100}{SR_m} \left( \frac{\sum_{i=1}^n |SR_i - SR_m|}{n} \right), \quad (4)$$

where:  $SR_i$  is the ratio between the predicted price and actual sale price for the apartment  $i$ ;  $SR_m$  is the median of the quotient, and  $n$  is the sample size for the prediction.

R-squared shows the predictable portion of the observed transaction price. It is measured by the proportion

of the variance in the target variable (actual transaction price) that is accounted for by the models. R-squared is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (p_i - \hat{p}_i)^2}{\sum_{i=1}^n (p_i - \bar{p})^2}, \quad (5)$$

where:  $\bar{p}$  is the sample mean of the actual transaction price for apartment  $i$ .

We took the average of 10 experiments for each measurement to determine whether the results were obtained by chance. In each experiment, the measurements were obtained both inside and outside of the sample prediction context. The 16,061 observations were randomly divided into training sets (90% of all transactions) and test sets (10% of all transactions).

Table 5 presents a comparison of the measurements obtained from both predictors. The values of MAPE and COD for RF are only 5.482 and 5.484, respectively, while those for the OLS predictor are 19.605 and 19.571, respectively. The MAPEs indicate that the percent deviation of the RF prediction from the actual contract price is only about 5% on average, while that of the OLS predictor is about 20%. The R-squared of the RF is also noticeably higher than the R-squared of the OLS. The R-squared of the RF model is 0.9761, which implies that 97% of the variability of the dependent variable has been accounted for while the remaining 3% of the variability has not.

The predictive performance can also be considered in terms of the hitting rate. If we define a successful prediction as an event in which the predicted price is within a certain range of the actual price, hitting rate indicates the

Table 5. Measurements for accuracy (average of 10 trials)

	OLS	Random Forest
MAPE	19.60567	5.482407
COD	19.57161	5.484705
R-squared	0.726056	0.976198

proportion of successful prediction. Table 6 compares the hitting rates obtained from both methods when we define the range of successful prediction as 1%, 3%, 5%, 10%, and 15%, respectively. In the RF predictor, when the difference between the market price and the forecast price is less than 15%, the hitting rate is about 95%. This means that the RF predictor allows us to make more sophisticated predictions.

Table 6. Hitting rates (average of 10 trials)

	OLS	Random Forest
Within 1%	3.4%	21.0%
Within 3%	10.4%	53.5%
Within 5%	17.4%	72.0%
Within 10%	34.6%	90.3%
Within 15%	50.9%	95.6%

The results can be interpreted as follows.

First, in the comparison between the accuracies of both methods, we can conclude that the RF predictor is significantly more accurate than the OLS predictor in all measurements (MAPE, COD, R-squared, and hitting rates). This finding is notable because the quality and quantity of information used in both methods were the same. The difference lied only in the form of the models. The functional form of the conventional hedonic pricing model represents a form of our intuition about housing value with the assumption that each attribute is separable and its influences constant. This means that, in the OLS-based model, the effect of each attribute is extremely simplified, with a single coefficient. In the RF model, since the predictor explores the hierarchical structure of features, it can more sensitively track the possibility that the effect of each attribute on price varies by context. The result implies that there are substantial losses resulting from the simplified nature of the OLS-based model and that at least some of these losses can be recovered using the RF predictor.

Second, the results show that the accuracy of the RF predictor can be surprisingly high. The average MAPE means that the percent deviation of a prediction from the actual contract price is only about 5% on average. We hypothesize that this accuracy is not due to the superiority of RF modeling alone and that the features of our data set also contribute to the high accuracy in the absolute perspective. One reason is the geometric density of our samples. A large portion of a property value comes from its location. If the samples are sparsely located in a large area, it is difficult to accurately measure the value from its location. We collected a relatively large sample (16,061 samples trained) from a small area (39.55 km<sup>2</sup>) and expect that this high density of samples may contribute to the high accuracy of prediction. The other reason is the type of property that our data covers. The coverage of observable characteristics can be an important factor affecting the accuracy of the estimation, as a data set contains only consolidated or measurable features of housing, such

as the number of rooms and floor level. We collected all apartment data from the same residential area (Gangnam in Seoul) because the structural characteristics of the apartments can be well-represented by a number of common characteristics. However, other types of dwelling (for example, detached houses) are usually more heterogeneous in their amenities, interior decorations, and other features that are difficult to codify or consolidate in a data set. If a large portion of attributes are unmeasured or unobservable, the predictive power of the model will be undermined by the lack of information rather than any modelling issue.

In addition, we will discuss the frequency of outliers, which is potentially related to the complexity of the prediction structure in the data-driven model constructed by the machine learning approach. For an OLS-based predictor, the prediction is made by the linear projection of observed attributes; thus, a large deviation from the actual value occurs only when the values of the attributes for which the coefficients are overestimated or underestimated are extremely large. For the RF predictor, it is difficult to formalize when outliers occur. However, it is important that, in the RF model, the order of variables is constructed by a data-driven process and the effect of an attribute on housing value can vary according to ordering structure. Therefore, if the ordering structure greatly distorts the actual value determination process in the housing market, the non-linearity can make largely deviated predictions. In the opposite case, if the complicated structure of the actual housing market is captured by the data-driven ordering structure, the occurrence of outliers will be significantly reduced. Rather, the rigidity of the model in the OLS-based technique might lead to more frequent outliers.

The frequency of outliers is displayed in Table 7, in which we define an “outlier” as a case in which the prediction deviates from the actual price by more than a certain percent range (50%, 100%, and 200%). Under these criteria, the occurrence of outliers is markedly reduced with the RF predictor. If we define the outliers as deviations greater than 50% from the actual value, then about 3.8% of OLS-based prediction are revealed to be outliers, compared to about 0.5% of the RF predictions. This result implies that the hierarchical structure of features constructed by the RF technique is not distortive and that the predictor is not easily over-fitted to a training set.

Table 7. Proportion of outliers (average of 10 trials)

Criteria	OLS	Random Forest
$\left  \frac{\hat{p}_i - p_i}{p_i} \right  > 0.5$	3.82%	0.52%
$\left  \frac{\hat{p}_i - p_i}{p_i} \right  > 1$	0.39%	0.26%
$\left  \frac{\hat{p}_i - p_i}{p_i} \right  > 2$	0.18%	0.19%

### 3.3. Comparison by time period

In the previous section, we shuffled the whole sample and randomly selected the training sets and test sets. In doing so, we ignored the time order of samples. For instance, in that case, samples from 2016 could be trained to be used to appraise a property in 2010. However, in the actual practice of mass appraisal, the information we can access is usually constrained to the present and past. Therefore, if we use more recent information to make a less recent appraisal, we may overstate the model's predictive power in reality. To address this problem, this section presents the performance of the predictors within limited time segments.

As in the previous experiments, we divided the sample in each time period into 90% training sets and 10% test sets and compared the average MAPE from 10 experiments for OLS and RF predictions. Table 8 presents the results for each time segment, as divided into years.

We noted several features of the results. At first, the overall level of performance measured by MAPE is lower than that in the previous section. This is natural, since the samples for each time segment are smaller than the total sample. In this case, the performance of predictors inevitably decreased.

Second, the performances of both predictors are unstable over the period. We hypothesize that this instability resulted from the smaller sample size and some economic events causing higher volatility in certain periods. From 2006 to 2011, the housing market in Korea was impacted by the global housing boom-bust cycle and subsequent financial crisis, and the annual rate of change in apartment prices was relatively more volatile than in the other periods. However, our data provides only the year of the contract, not its exact date. Hence, if the annual change in housing price is more severe, the importance of the un-

observable information (the exact dates of the contracts) becomes more important, and the predictive power of the models would decrease. Thus, we can expect that the average performance of both predictors is poorer from 2006 to 2011 (see that they simultaneously reach a peak in 2011) than from 2012 to 2017.

Finally, we noted that the RF predictor is still more accurate than the OLS predictor in all individual time segments. Although the performance of the RF predictor is volatile, the results show that it is always stronger than that of OLS predictor. Roughly, it seems that the percent deviation rates of RF are lower than half of those of the OLS predictor and that the gap between the OLS and RF predictors is similar in the main result. Conclusively, the advantage of the RF predictor seems to remain even with a smaller data set and on different timelines.<sup>1</sup>

### Conclusions

In this paper, we discussed the features of the RF predictor in comparison to the conventional OLS-based predictor. This paper shows that the predictive performance of a machine learning-based predictor can be superior to that of the OLS-based approach. We used apartment transaction data from 2006–2017 in Gangnam, one of the most developed areas in Korea. We collected a data set covering 40% of all transactions in the selected area, and the samples were randomly divided into training sets consisting of 90% of all transactions and test sets consisting of the remaining 10% of transactions. We used the averages of 10 experiments to compare the performance measurements in order to eliminate the possibility that the results occurred by chance.

The average percentage deviation between the predicted and actual market price was only around 5.5% for the machine learning predictor and almost 20% for the OLS-based predictor. Moreover, the probabilities that the RF prediction was within 3%, 5%, and 10% of the actual market price were 53.5%, 72%, and 90.3%, respectively, whereas those of the OLS-based prediction are 10.4%, 17.4%, and 34.6%, respectively. Furthermore, we found that the RF predictor made fewer outlier predictions than the conventional hedonic pricing model. The probability of the RF predictions deviating more than 50% from the actual price was found to be only 0.5%, while that of OLS-based predictions doing so was almost 3.8%.

The contribution of this paper can be discussed in two ways. From a theoretical perspective, this paper shows that there are significant market complexities making the value determination process unable to be fully accounted for in the simplified assumptions of the conventional hedonic pricing model (separability and constancy of an attribute's effect on housing value). From a practical perspective, the

Table 8. MAPE for each time segment (average of 10 trials)

	OLS	Random Forest
2006	23.11313	4.574283
2007	19.07825	8.020659
2008	18.65792	8.058142
2009	23.96337	9.916612
2010	23.46693	10.28114
2011	25.52543	11.8575
2012	15.28604	6.439383
2013	14.02813	4.597863
2014	14.28471	5.189567
2015	12.89255	4.125248
2016	12.96526	5.764724
2017	15.17713	4.313319

<sup>1</sup> It also implies that the superiority of the RF predictor in the previous section did not result from the fact that we were allowed to use information from the “future”—i.e., more recent information for less recent appraisals.

results are a demonstration that the accuracy of a machine learning-based mass appraisal can be surprisingly high in some cases (Note that the MAPE of human appraisals is 12% in Cannon and Cole, 2011). We infer that the high predictive power derives from a combination of the features of the RF model and the data set we applied.

It is important to obtain an accurate estimation of the value of a house whose market price is not observed in order to construct a reliable house price index or to conduct a successful mass appraisal. Traditionally, the hedonic pricing model has been adopted as the appraisal machine, but, for several reasons, the accuracy of the OLS-based predictor can be undermined. This paper suggests that the RF predictor could be a complement to this linear regression method. Its results show that there is a significant loss in accuracy resulting from the simplification of reality in the OLS-based model and that some of that loss can be recovered by the RF predictor. This implies that the RF method can more successfully track the complexity of the value determination process that the OLS-based models cannot fully capture.

### Author contributions

W. Kim and J. Hong conceived the study and were responsible for the design and development of the data analysis. W. Kim were responsible for data collection and J. Hong and H. Choi were responsible for data analysis and interpretation.

### Disclosure statement

There are no conflicts of interest.

### References

- Adair, A., McGreal, S., Smyth, A., Cooper, J., & Ryley, T. (2000). House prices and accessibility: the testing of relationships within the Belfast urban area. *Housing Studies*, 15(5), 699-716. <https://doi.org/10.1080/02673030050134565>
- Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: an application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2), 1772-1778. <https://doi.org/10.1016/j.eswa.2011.08.077>
- Benson, E. D., Hansen, J. L., Schwartz, A. L., & Smersh, G. T. (1998). Pricing residential amenities: the value of a view. *Journal of Real Estate Finance and Economics*, 16(1), 55-73. <https://doi.org/10.1023/A:1007785315925>
- Cannon, S. E., & Cole, R. A. (2011). How accurate are commercial real estate appraisals? Evidence from 25 years of NCREIF sales data. *Journal of Portfolio Management*, 37(5), 68-88. <https://doi.org/10.3905/jpm.2011.37.5.068>
- Case, K. E., Quigley, J. M., & Shiller, R. J. (2005). Comparing wealth effects: the stock market versus the housing market. *Advances in Macroeconomics*, 5(1). <https://doi.org/10.2202/1534-6013.1235>
- Čeh, M., Kilibarda, M., Liseč, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information*, 7(5), 168. <https://doi.org/10.3390/ijgi7050168>
- Chau, K. W., & Chin, T. L. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and its Applications*, 27(2), 145-165.
- Chen, J. H., Ong, C. F., Zheng, L., & Hsu, S. C. (2017). Forecasting spatial dynamics of the housing market using Support Vector Machine. *International Journal of Strategic Property Management*, 21(3), 273-283. <https://doi.org/10.3846/1648715X.2016.1259190>
- Clautrie, T. M., & Neill, H. R. (2000). Year-round school schedules and residential property values. *Journal of Real Estate Finance and Economics*, 20(3), 311-322. <https://doi.org/10.1023/A:1007841326833>
- Darling, A. H. (1973). Measuring benefits generated by urban water parks. *Land Economics*, 49(1), 22-34. <https://doi.org/10.2307/3145326>
- Debrezion, G., Pels, E., & Rietveld, P. (2007). The impact of railway stations on residential and commercial property value: a meta-analysis. *Journal of Real Estate Finance and Economics*, 35(2), 161-180. <https://doi.org/10.1007/s11146-007-9032-z>
- Downes, T. A., & Zabel, J. E. (2002). The impact of school characteristics on house prices: Chicago 1987–1991. *Journal of Urban Economics*, 52(1), 1-25. [https://doi.org/10.1016/S0094-1190\(02\)00010-4](https://doi.org/10.1016/S0094-1190(02)00010-4)
- Dubin, R. A., & Sung, C. H. (1990). Specification of hedonic regressions: non-nested tests on measures of neighborhood quality. *Journal of Urban Economics*, 27(1), 97-110. [https://doi.org/10.1016/0094-1190\(90\)90027-K](https://doi.org/10.1016/0094-1190(90)90027-K)
- Espey, M., & Lopez, H. (2000). The impact of airport noise and proximity on residential property values. *Growth and Change*, 31(3), 408-419. <https://doi.org/10.1111/0017-4815.00135>
- Fan, G. Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: a decision tree approach. *Urban Studies*, 43(12), 2301-2315. <https://doi.org/10.1080/00420980600990928>
- Fletcher, M., Gallimore, P., & Mangan, J. (2000). Heteroscedasticity in hedonic house price models. *Journal of Property Research*, 17(2), 93-108. <https://doi.org/10.1080/095999100367930>
- Garrod, G. D., & Willis, K. G. (1992). Valuing goods' characteristics: an application of the hedonic price method to environmental attributes. *Journal of Environmental Management*, 34(1), 59-76. [https://doi.org/10.1016/S0301-4797\(05\)80110-0](https://doi.org/10.1016/S0301-4797(05)80110-0)
- Gillard, Q. (1981). The effect of environmental amenities on house values: the example of a view lot. *The Professional Geographer*, 33(2), 216-220. <https://doi.org/10.1111/j.0033-0124.1981.00216.x>
- Goodman, A. C. (1989). *Topics in empirical urban housing research*. In R. Muth, & A. Goodman (Eds.), *The economics of housing markets* (pp. 49-146). Chur, Switzerland: Harwood Academic.
- Gu, J., Zhu, M., & Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications*, 38(4), 3383-3386. <https://doi.org/10.1016/j.eswa.2010.08.123>
- Hanson, S. (2004). The context of urban travel: concepts and recent trends. In S. Hanson, & G. Giuliano (Eds.), *The geography of urban transportation* (pp. 3-29). New York: The Guilford Press.
- Harrison Jr, 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)



- Hayes, K. J., & Taylor, L. L. (1996). Neighborhood school characteristics: what signals quality to homebuyers? *Economic Review-Federal Reserve Bank of Dallas*, 2-9.
- Huh, S., & Kwak, S. J. (1997). The choice of functional form and variables in the hedonic price model in Seoul. *Urban Studies*, 34(7), 989-998. <https://doi.org/10.1080/0042098975691>
- Jud, G. D., & Watts, J. M. (1981). Schools and housing values. *Land Economics*, 57(3), 459-470. <https://doi.org/10.2307/3146025>
- Kain, J. F., & Quigley, J. M. (1970). Measuring the value of housing quality. *Journal of the American Statistical Association*, 65(330), 532-548. <https://doi.org/10.1080/01621459.1970.10481102>
- Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2), 132-157. <https://doi.org/10.1086/259131>
- Li, M. M., & Brown, H. J. (1980). Micro-neighborhood externalities and hedonic housing prices. *Land Economics*, 56(2), 125-141. <https://doi.org/10.2307/3145857>
- Limsombunchai, V. (2004, June). House price prediction: hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference* (pp. 25-26), New Zealand.
- Malpezzi, S. (2002). Hedonic pricing models: a selective and applied review. *Housing Economics and Public Policy*, 67-89. <https://doi.org/10.1002/9780470690680.ch5>
- McCluskey, W., & Anand, S. (1999). The application of intelligent hybrid techniques for the mass appraisal of residential properties. *Journal of Property Investment & Finance*, 17(3), 218-239. <https://doi.org/10.1108/14635789910270495>
- McMillan, D., Jarmin, R., & Thorsnes, P. (1992). Selection bias and land development in the monocentric model. *Journal of Urban Economics*, 31, 273-284. [https://doi.org/10.1016/0094-1190\(92\)90056-Q](https://doi.org/10.1016/0094-1190(92)90056-Q)
- McMillan, M. L., Reid, B. G., & Gillen, D. W. (1980). An extension of the hedonic approach for estimating the value of quiet. *Land Economics*, 56(3), 315-328. <https://doi.org/10.2307/3146034>
- Miller, N., Peng, L., & Sklarz, M. (2011). House prices and economic growth. *Journal of Real Estate Finance and Economics*, 42(4), 522-541. <https://doi.org/10.1007/s11146-009-9197-8>
- Mu, J., Wu, F., & Zhang, A. (2014). Housing value forecasting based on machine learning methods. *Abstract and Applied Analysis*, 2014, Article ID 648047. <https://doi.org/10.1155/2014/648047>
- Palmquist, R. B. (1992). Valuing localized externalities. *Journal of Urban Economics*, 31, 59-68. [https://doi.org/10.1016/0094-1190\(92\)90032-G](https://doi.org/10.1016/0094-1190(92)90032-G)
- Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: the case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6), 2928-2934. <https://doi.org/10.1016/j.eswa.2014.11.040>
- Richardson, H. W., Vipond, J., & Furbey, R. A. (1974). Determinants of urban house prices. *Urban Studies*, 11(2), 189-199. <https://doi.org/10.1080/00420987420080341>
- Ridker, R. G., & Henning, J. A. (1967). The determinants of residential property values with special reference to air pollution. *Review of Economics and Statistics*, 49(2), 246-257. <https://doi.org/10.2307/1928231>
- Rodriguez, M., & Sirmans, C. F. (1994). Quantifying the value of a view in single-family housing markets. *Appraisal Journal*, 62, 600-603.
- Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55. <https://doi.org/10.1086/260169>
- Selim, H. (2009). Determinants of house prices in Turkey: hedonic regression versus artificial neural network. *Expert Systems with Applications*, 36(2), 2843-2852. <https://doi.org/10.1016/j.eswa.2008.01.044>
- Sheppard, S. (1999). Hedonic analysis of housing markets. *Handbook of Regional and Urban Economics*, 3, 1595-1635. [https://doi.org/10.1016/S1574-0080\(99\)80010-8](https://doi.org/10.1016/S1574-0080(99)80010-8)
- So, H. M., Tse, R. Y., & Ganesan, S. (1997). Estimating the influence of transport on house prices: evidence from Hong Kong. *Journal of Property Valuation and Investment*, 15(1), 40-47. <https://doi.org/10.1108/14635789710163793>
- Song, Y., & Sohn, J. (2007). Valuing spatial accessibility to retailing: a case study of the single family housing market in Hillsboro, Oregon. *Journal of Retailing and Consumer Services*, 14(4), 279-288. <https://doi.org/10.1016/j.jretconser.2006.07.002>
- Thaler, R. (1978). A note on the value of crime control: evidence from the property market. *Journal of Urban Economics*, 5(1), 137-145. [https://doi.org/10.1016/0094-1190\(78\)90042-6](https://doi.org/10.1016/0094-1190(78)90042-6)
- Verikas, A., Lipnickas, A., & Malmqvist, K. (2002). Selecting neural networks for a committee decision. *International Journal of Neural Systems*, 12(05), 351-361. <https://doi.org/10.1142/S0129065702001229>
- Wilhelmsson, M. (2000). The impact of traffic noise on the values of single-family houses. *Journal of Environmental Planning and Management*, 43(6), 799-815. <https://doi.org/10.1080/09640560020001692>
- Williams, A. W. (1991). A guide to valuing transport externalities by hedonic means. *Transport Reviews*, 11(4), 311-324. <https://doi.org/10.1080/01441649108716793>
- Zhou, G., Ji, Y., Chen, X., & Zhang, F. (2018). Artificial neural networks and the mass appraisal of real estate. *International Journal of Online Engineering*, 14(3), 180-187. <https://doi.org/10.3991/ijoe.v14i03.8420>
- Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research*, 33(3), 349-387.