한국경제지리학회지 제25권 제1호 2022(183~201)

How the Pattern Recognition Ability of Deep Learning Enhances Housing Price Estimation

Jinseok Kim* · Kyung-Min Kim**

딥러닝의 패턴 인식능력을 활용한 주택가격 추정

김진석* · 김경민**

Abstract: Estimating the implicit value of housing assets is a very important task for participants in the housing market. Until now, such estimations were usually carried out using multiple regression analysis based on the inherent characteristics of the estate. However, in this paper, we examine the estimation capabilities of the Artificial Neural Network(ANN) and its 'Deep Learning' faculty. To make use of the strength of the neural network model, which allows the recognition of patterns in data by modeling non-linear and complex relationships between variables, this study utilizes geographic coordinates (i.e. longitudinal/latitudinal points) as the locational factor of housing prices. Specifically, we built a dataset including structural and spatiotemporal factors based on the hedonic price model and compared the estimation performance of the models with and without geographic coordinate variables. The results show that high estimation performance can be achieved in ANN by explaining the spatial effect on housing prices through the geographic location.

Key Words : housing market, housing price estimation, hedonic price model, deep learning, spatial autocorrelation

요약: 주택가격을 정확히 추정하기 위한 많은 연구가 진행되어 왔다. 선행연구들은 주택의 고유 특성과 인근 지역 특성을 통제하는 계량경제모형을 활용한 분석이 많았다. 본 연구에서는 인공신경망 모형(ANN)을 활용하여 주택가격을 추정하였다. 딥러닝 기술의 장점은 변수 간의 복잡하고 비선형적인 특성을 모델링하고 데이터의 패턴을 인식할 수 있다는 것이다. 본 연구에서는 부동산 시장에서 공간적 분포도 패턴으로 인식할 수 있다는 가정하에 지리좌표를 설명변수로 ANN에 투입하였다. 선형회귀분석과 ANN 모형 간 비교 결과, 선형 모형 대비 ANN 모형의 설명력이 높았으며, 특히 ANN 모형은 지리좌표를 투입하였을 때 더 높은 정확도를 보여주었다. 또한 ANN 모형의 경우 지리좌표를 통해 모형 잔차의 공간적 자기 상관성이 크게 감소하였다는 점을 확인하였다. 이를 통해 ANN 모형의 패턴인식 능력을 활용하면 공간적 패턴을 학습시킴으로써 주택가격을 정확히 추정할 수 있음을 밝혔다.

주요어: 주택시장, 주택가격추정, 헤도닉 가격 모형, 딥 러닝, 공간적 자기상관

* Ph.D. Student, Department of Environmental Planning, Graduate School of Environmental Studies, Seoul National University (서울대학교 환경대학원 박사과정, jinski71@snu.ac.kr)

** Associate Professor, Department of Environmental Planning, Graduate School of Environmental Studies, Seoul National University (서울대학교 환경대학원 부교수, kkim2@snu.ac.kr)

https://doi.org/10.23841/egsk.2022.25.1.183

1. Introduction

Real estate is characterized by heterogeneous attributes and rarely-repeated transactions of individual assets. Thus, it is necessary to accurately predict the implicit price of a hypothetical transaction between buyer and seller. In the housing market, estimating the appropriate value of assets is very important to market participants. The hedonic price model is widely used for these tasks. This model is based on the hedonic hypothesis that the attributes or components inherent in a product can be used to estimate unobserved prices. (Rosen, 1974; Goodman, 1978) On this premise, there have been many attempts to estimate housing prices through the characteristics of houses with multiple linear regression analysis. (Zietz *et al.*, 2008)

However, conventional regression methods require many premises, such as linear relationships between the variables. To solve this problem, some studies have tried to utilize the Artificial Neural Network (which will heretofore be simply termed 'ANN') to estimate housing prices. An important feature of the ANN is that it can recognize patterns in data and can learn about the linear and non-linear relationships between variables. (Tay and Ho, 1992) Because of this, ANN-based analysis has been acknowledged as a substitute for classical regression analysis. Since the 1990s, studies have been conducted using ANN to evaluate housing prices and comparing them with the results of multiple regression analysis. It is reported that the performance of ANN can be superior to multiple regression (Tay and Ho, 1992), while some said that no significant difference was found. (Worzala *et al.*, 1995) However, around 2010, the introduction of the Deep Neural Network (DNN), also known as 'Deep Learning,' which uses two or more hidden layers, has shown superb performance in many realms of analysis. (Chollet, 2017)

Beyond the prior hedonic price studies that replace the conventional method with ANN, this study aims to use variables that can harness pattern recognition, which is the main strength of ANN. Prior studies have pointed out the impact of locational factors on housing prices due to differences in neighborhood environments. (Dubin, 1992; Anselin, 1998) Thus, the methodologies to explain the spatial heterogeneity of variables so far, such as the Spatial Autoregressive Model and Geographically Weighted Regression, have attempted to explain spatial aspects. (Can, 1992; Fotheringham *et al.*, 1998; Osland, 2010; Huang *et al.*, 2010; Lee *et al.*, 2011; Lu *et al.*, 2014) However, the extended models based on linear regression have limitations in modeling complex and non-linear spatial distributions of housing prices. In this study, we try to explain the spatial factors that determine housing prices by injecting the geographic location patterns of housing prices, i.e., the three-dimensional geographic variables, directly into the ANN model. Just as a person recognizes locational patterns, training a model to interpret these patterns through the non-linear pattern recognition ability of ANN may allow it to effectively capture the human perception and

valuation of the location conditions, which is a crucial factor in estimating the price that a house would sell for.

Prior studies that directly used geographic coordinate variables reported better predictability when using ANN than conventional regression methods (Lee and Kim, 2018) or spatial autoregression analysis (Garc a *et al.*, 2008; Mimis *et al.*, 2013) under the same conditions. However, these studies only compared the performance of different analysis methodologies and did not examine how the coordinate variables explain spatial heterogeneity. In this study, we assume that geographic coordinates explain the spatial aspects of housing prices, and validate it by comparing the spatial autocorrelation of the models with and without the geographic coordinates. Through analyzing the prediction error, we see that the deep learning model can use geographic coordinates to capture the spatial distribution of housing prices and thereby provide improvements to housing price estimation.

2. Literature Review

1) Hedonic Price Model

Unlike other markets, the real estate market is composed of heterogeneous goods, and transactions are occurring very rarely. Even when the area and types of housing are the same, the value varies depending on the location, due to differences in accessibility to the surrounding environment and other factors. Likewise, the value of each house also changes over time, because of amortization or economic fluctuation in the housing market. As transactions of individual houses occur at intervals of at least several months and usually several years, it is necessary to estimate the implicit price at a time when market prices are not observed. Thus, Rosen (1974) suggested a hedonic hypothesis that the potential value can be calculated by summing up the utilities or attributes inherent in the product. But housing prices are not only influenced by the inherent features of the house; their valuation is highly affected by the location, mainly due to differences in neighborhood quality. (Dubin, 1992; Anselin, 1998) Therefore, prior studies of the hedonic price model largely classified the explanatory variables into structural and locational factors (Goodman and Thibodeau, 1998), and in some cases, locational factors were further delineated into locational and neighborhood environmental factors. (Dodgson and Topham, 1990; Stamou *et al.*, 2017) Many prior studies fixed the time and did not consider economic fluctuations over time, (Yang, 2017) but some argued that housing demand is affected by the overall trend of the economy over time, (Case *et al.*, 2004; Chun, 2012; Ghorbani and Afgheh, 2017) thus the time dimension should be added as a determinant of housing prices to capture price fluctuations due to economic trends. (Huang *et al.*, 2010; Herath and Maier, 2011) In sum, prior studies largely considered three types of factors to estimate housing prices: structural, locational, and temporal economic factors. Therefore, the studies dealing with the hedonic price model have

employed methodologies that can effectively explore spatiotemporal factors contributing to housing prices.

In its early stages, the analytical methodology of the hedonic price model was 'multiple regression' using a linear formula and its extensions. (Edmonds, 1984) However, the linear regression model presumes linear relationships between variables, so non-linear relationships are hard to capture and the researcher has to identify the relationships in advance and model them. (Tay and Ho, 1992) This problem is well established as it is known that many variables that determine housing prices have non-linear relationships. (Do and Grudnitski, 1993; Garc a *et al.*, 2008) Also, to explain spatial aspects of housing price distribution, some studies utilized spatial econometric models such as the Spatial Autoregressive model and the Geographically Weighted Regression model. (Can, 1992; Fotheringham *et al.*, 1998; Osland, 2010; Huang *et al.*, 2010; Lu *et al.*, 2014) These models explain the spatial factors of housing prices by taking the prices of nearby houses into account, or by changing the coefficients of the models according to geographic location. But these methodologies do not aim to train the model to capture the spatial pattern itself, but rather indirectly weigh the model based on the First Law of Geography. (Tobler, 1970) Furthermore, even if the spatial context is considered, it still has limitations as an extension of the linear methodology in modeling complex relationships between the variables. Since the 1990s, however, the ANN methodology has been drawing attention as a potential solution to the limitations of existing methodologies through its capacities of "pattern recognition" and "self-learning."

2) Neural Network Methods in Real Estate Studies

ANN is a branch of machine learning that learns expression from data, developed by motifs from the signal transmission structure of brain neurons. The structure of the network varies depending on the field of study, but the model used in this study is the basic Multi-Layer Perceptron (MLP) consisting of fully-connected layers. The model consists of the input layer, hidden layers, and the output layer, in order, and each layer consists of several nodes. The number of nodes in the input and output layers is equal to the number of input and output variables, respectively, and the number of hidden layers and nodes can be controlled. In the hidden and output layers, the input value of a node is a linear combination of output values from the previous layers and weights connecting each layer. It becomes the output for the next layer after conversion using a non-linear activation function. If a loss is calculated based on the difference between the final output and the label, a back-propagation is made with this loss value to adjust the weight value. In the pre-learning initialization phase, weights are randomly assigned based on the probability distribution, and they are adjusted to account for the relationships between variables as learning progresses. Figure 1 illustrates the basic structure of the Deep Neural Network, ANN with

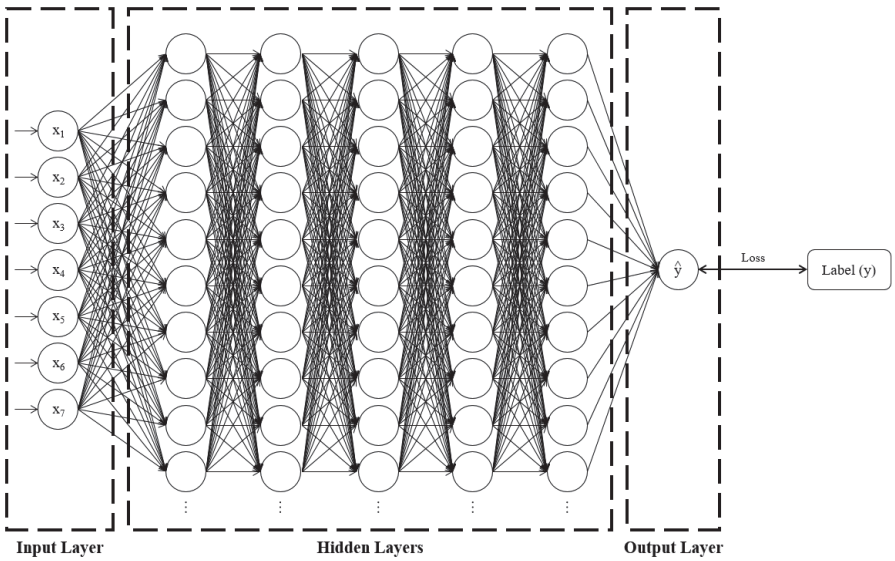

Figure 1. Basic Structure of Artificial Neural Network (Muilti-Layer Perceptron)

two or more hidden layers.

Input values are converted through several hidden layers with the non-linear transformation. In the process, the inherent features in noisy raw data are refined. (Haykin, 2003; Chollet, 2017) Through this process, we can also effectively capture non-linear relationships between variables, and consequently, enable the model to recognize patterns inherent in the data. Moreover, the ANN model can create rules between variables on its own without previous model settings.

Studies using ANN have reported better prediction performance than classical multiple regression (Peterson and Flanagan, 2009; Loo, 2019), but there have also been studies that report no significant difference in predictability. (Worzala *et al.*, 1995) In the past, a lack of reliable training methods for the deep model meant that the shallow model with a single hidden layer was generally used. Those studies that used two or more hidden layers could not achieve better performance than those using a single layer. (Khalafallah, 2008) However, with significant improvements in computing performance around 2010, the usage of activation functions that are well-suited for multiple hidden layers, increased efficiency of optimizers over the conventional Stochastic Gradient Descent (SGD), and advanced methods of weight initialization, "Deep Learning" was eventually able to significantly outperform the shallow model. (Chollet, 2017)

As the performance of the neural network methodology is drawing attention, there have been many studies using deep learning in real estate market research. (Tay and Ho, 1992; Khalafallah, 2008; Peterson and Flanagan, 2009; Garc a *et al.*, 2008; Mimis *et al.*, 2013; Lee and Kim, 2018; Loo, 2019; Zhou *et al.*, 2019; Law *et al.*, 2020) However, what matters is not just to improve performance by replacing the methodology

with the latest one but to model the relationship between variables that traditional methodologies could not properly capture. Recent studies of house price estimation have adopted the data that traditional methods cannot utilize such as image data of house and neighborhood streets (Poursaeed *et al.*, 2018; Law *et al.*, 2019), textual information (Zhou *et al.*, 2019), or geographic coordinates of individual homes. (Garc a *et al.*, 2018; Lee and Kim, 2018; Pai and Wang, 2020)

However, many prior studies have only compared prediction performance with or without the data, and have rarely analyzed "how that variable accounts for housing prices". This is because, unlike traditional methods, the neural network model has "black-box" properties that are difficult to directly confirm the coefficient or significance of each variable. (Loo, 2019) However, revealing the structure of how the variables of the model describe the dependent variables is a very important problem. To overcome this, some efforts have been made such as simulating changes of one explanatory variable with keeping others fixed (Peterson and Flanagan, 2009; Abidoye and Chan, 2018) or quantifying image data of houses via CNN models then reveal the relationship using linear regression. (Law *et al.*, 2019)

In this study, we demonstrate more directly that neural network models can better model complex pattern relationships than traditional methodologies. This study focuses on the geographic coordinates of houses. As we have seen earlier, the location of houses has also been used in previous studies, as housing prices are affected by locational factors, such as neighborhood conditions and traffic accessibility. (Basu and Thibodeau, 1998) However, putting the geographic coordinates into the linear model presumes that the price varies linearly when the location of a house moves a certain distance along the direction, which does not make sense. However, if the deep learning methodology that can model the non-linear relationships manage to utilize the variables, the performance of the model can be improved by describing the spatial effects of housing prices.

We present the following hypotheses. Given that the prediction error, defined as the difference between the model's predictions and the actual price, refers to the unexplained part of the model, the spatial effect of the prediction error would be reduced if the model captures the locational factor of the house. To prove this, in this study, we compare Global Moran's I, a spatial autocorrelation indicator, of prediction error for a deep learning model with and without three-dimensional location coordinates. Then we compare linear regression in the same way. If the model using the geographic coordinates shows a reduction in the tendency towards spatial autocorrelation of the residuals, then it can be inferred that the geographic coordinate variables account for the housing price as the information containing spatial distribution patterns. Through this, we demonstrate that the spatial effect of housing prices can be explained and model prediction performance can be enhanced with non-linear locational information through deep learning methodology rather than traditional linear regression.

3. Research Methods

1) The Data

Based on the literature review, this study estimates housing prices in Seoul. As stated earlier, the explanatory variables from the data can be broken down into three groups: structural, locational, and temporal variables. To establish the dataset, we utilize the housing transaction data, the spatial data of land parcels, and the house price index.

The variables that describe the structural factors are the area, age, and the number of floors of the house, which can be obtained from the transaction data. The housing transaction data came from the Korean Government's Ministry of Land, Infrastructure and Transport (MOLIT). It is a database of all housing transactions that have occurred throughout the country, including information on transaction dates, address, area, year of construction, floors, and price. The main feature of this data is that it is open to the public and gathered on legally mandated transaction reporting.

The variables that describe the locational factors are three-dimensional geographic coordinates and appraisal value of the land where the house is located, which supports the coordinates in capturing the location patterns. We obtain these variables from the spatial data of land parcels provided by MOLIT. This dataset contains polygon data along with the annual appraisal land price of all land parcels in Korea. The land data is created in the planar coordinate system with the Transverse Mercator projection method (EPSG:5174), and the spatial unit is the 'meter'. The longitude and latitude are obtained by calculating the centroids of all land parcel polygons, and the altitude above sea level is obtained by merging raster Digital Elevation Model (DEM) data provided by MOLIT. We merge annual appraisal value and coordinates of land parcels into housing transaction data based on their address.

The variable that represents the economic factor is the monthly house price index that shows relative price levels at a particular time. Since the transaction data used in this study contains spatiotemporal information over several years and the housing market is heavily influenced by government policies and seasonality, it is necessary to capture the changes throughout the entire housing market over time. The monthly house price index is provided by the Korea Appraisal Board. It is calculated using the Repeat Sales Method, which combines houses with certain similarities into the same product and calculates the rate of change in prices when repeated transactions involving the same house occur over time.

This study targets apartment buildings: multi-family complexes of more than five stories. This is because, in the case of Korea, apartment buildings are relatively standardized as housing assets and easily traded. (Jeong, 2014; Kim *et al.*, 2015) The spatial scope is limited to Seoul, which is considered the bellwether for trends in the housing market in Korea. The time range is 14 years, from 2006 to 2019. The final number of data records,

excluding observations with null values, is 1,128,758. Table 1 shows the details of the final dataset and Table 2 shows descriptive statistics of the variables.

To understand the spatial context of apartment distribution in Seoul, we interpolate the distribution of housing prices with the Inverse Distance Weighting tool of ArcGIS (version 10.3). Figure 2 shows the results. Overall, the apartments in Gangnam-gu, the southeastern district of Seoul, and near the Han River are among the highest-priced.

Table 1. Dataset

Classification	Variable	Details	Unit	Sample Data
Metadata (Not included in the model)	Location	The administrative district where the house is located	-	Gaepo-dong, Gangnam-gu, Seoul
	Address	Detailed lot number	-	658-1
	APTName	The name of the apartment complex	-	Gaepo 6th Woosung APT, Bldg Vol.1-8
	Contract Month	The year and month when the transaction was made	YYYYMM	200601
Input Variables (Model 1, 3)	Area	The area of the house	square meter	67.28
	Floor	The floor of the house is located on	-	3
	Age	The number of years since completion	years	19
	PriceIndex	The house price index published by the Korean Appraisal Board	2006.01=100	100
	Appraisal Value	The official price of the land where the house is located	KRW per square meter	5,700,000
Input Variables: Geographic Coordinates (All Model)	Longitude	Geographic coordinates of the house (EPSG:5174)	meter	204958.0160
	Latitude		meter	441568.0148
	Altitude	The elevation of the land where the house is located	meter	66.5234
Target Variable	Price	The trade price of the house	10,000 KRW	54,500

Table 2. Descriptive Statistics

Variables	Mean	Median	Standard Deviation	Number of Data
Price	49613.7	39700	37780.6	1,128,758
Area	77	80.4	28.8	
Floor	8.8	8	5.9	
Age	15.2	14	9.1	
PriceIndex	142.6	136	25	
AppraisalValue	4349560	3370000	2911223	
Longitude	199825.6	201823.9	7975.9	
Latitude	450356.1	449336.9	6384.9	
Altitude	32.9	24.7	28.0	

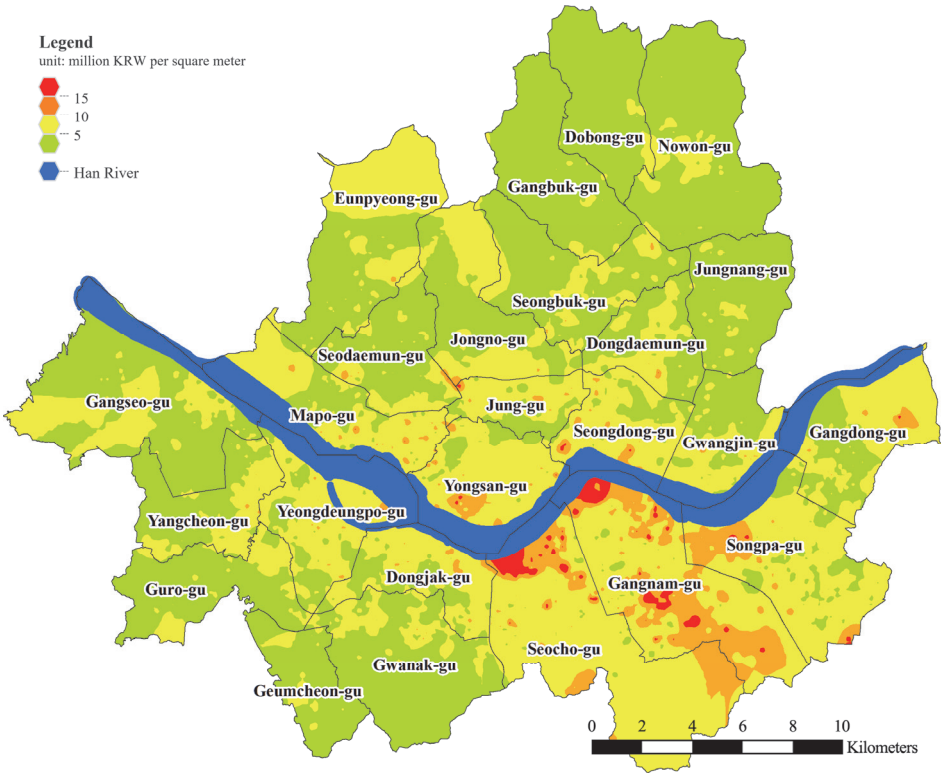

Figure 2. Interpolated Spatial Distribution of the Housing Prices in Seoul

2) Model Specification

As discussed, the models are processed both with and without geographic coordinates, and each conducted an optimal model search separately. Model 1 is the base ANN model where the coordinates are not included, and Model 2 is the ANN model including the coordinates. For comparison, we also perform multiple regression analysis with the same combination of variables. Model 3 is a multiple regression model without geographic coordinates and Model 4 includes the coordinates.

The analytical tool of Models 1 and 2 is Keras (version 2.3.1), the neural network analysis library of Python. The explanatory variables are standardized using the mean and standard deviation calculated for each variable before being put into the model. This is because the training can be difficult if the values with different scales are injected into the ANN model. (Chollet, 2017)

To establish the best-fit model for ANN analysis, it is necessary to find the optimal combination of hyperparameters, which are the variables related to the structure and analytical methods of the model. However, there is no globally established rule for determining hyperparameters, so they need to be found through repeated trial-and-error. (Lenk *et al.*, 1997) We change only the number of hidden layers and nodes, and other parameters are fixed

at the optimal values suggested by the literature. This is because the number of hidden layers and nodes, which are equivalent to the structure of the model, are directly related to the characteristics of the dataset and have a significant impact on the results of the study compared to other hyperparameters. In this study, the number of hidden layers goes from three to nine, and the number of nodes multiplies from 256 to 2048.

First, for the fixed hyperparameters we use ReLU as the activation function for the hidden layers. ReLU is the most commonly used activation function in deep learning. (Chollet, 2017) It returns zero if the input is negative, and returns the input if positive. ReLU is suitable for the deep network as it solves vanishing gradient problems due to the nature of returning 'one' when input values are positive. (Raschka and Mirjalili, 2019)

Second, the train loss function of the model is Mean Squared Error (MSE), and the validation loss function is Mean Absolute Percentage Error (MAPE). In general, when using ANN for regression problems, MSE is often used as a loss function. In the case of validation loss, which is used to determine the best-fit model, Mean Absolute Error (MAE) or MAPE is often used. In this case, we use MAPE because the price range is so wide that the ratio of error is more important than the absolute value of error. When calculating MAPE, we use the valid set that is separated from the training set, instead of the test set. This is to prevent hyperparameters from being adjusted to suit the test set. (Koki, 2016) Thus, the test set is only used to identify the performance of the final model after hyperparameters are adjusted. In this study, the proportions of the training, valid, and test sets are 80%, 10%, and 10% respectively.

Third, the weights of the model are initialized with the He Initialization method. The Xavier Initialization, which is commonly used, uses the values that form a normal distribution with a standard deviation of where the number of nodes in the previous layer is n. According to He (2015), when using the ReLU activation function, the standard deviation of the distribution should be increased to, which is times more than that of the Xavier Initialization, to enable efficient model convergence.

Fourth, we use the Adam optimizer. Adam is known for having a low requirement to tune hyperparameters, ensuring better performance and faster convergence than other methods with default settings. (G ron, 2017) The learning rate, which is the rate at which weights are updated in one epoch, is the default value of 0.001 for the Adam optimizer in Keras.

Finally, the number of train epochs is set to 200. The MAPE trend for each epoch is recorded to confirm that no overfitting and underfitting have occurred in the learning process. Overfitting means that the model is excessively optimized for the training set, leading to generalization performance that is rather poor. In contrast, underfitting means that not enough training has occurred, leaving room for the model's prediction accuracy to rise through further training. If the MAPE trend in each epoch indicates that the errors in the valid set tend to increase from a threshold, we can conclude that

Table 3. Summary of Hyperparameters

Parameters	Values	Notes
Number of Hidden Layers	[3, 4, 5, 6, 7, 8, 9]	-
Number of Nodes	[256, 512, 1024, 2048]	Not varying with hidden layers
Activation Function	ReLU	Only for hidden layers
Train Loss Function	MSE (Mean Squared Error)	Fixed
Validation Loss Function	MAPE (Mean Abs. Percentage Error)	Fixed
Optimizer	Adam	Fixed
Epochs	200	Fixed

overfitting has occurred. The details of the parameters are summarized in Table 3.

4. Empirical Outcomes

1) Model Performance Comparison

This section presents the results of the ANN model training and compares them with OLS. Tables 4 and 5 show the trend of MAPE values calculated with the valid set according to the ANN model structure. The optimal model configurations in this study are 9 hidden layers and 1024 nodes in Model 1, and 6 hidden layers and 2048 nodes in Model 2. Meanwhile, overall MAPE values are lower in Model 2 with geographic coordinates. With optimal model configuration, Model 1 shows 11.62% of MAPE, while Model 2 shows 5.11%. Figure 3 shows the training progress by epochs in the optimal model configuration. It is confirmed that the MAPE converges at a specific value, meaning that the training progresses well without overfitting.

We conduct prediction for the test set using ANN and multiple regression analysis and calculate MAPE in the same way as with the valid

Table 4. MAPE of Model 1 by the Hidden Layer Structure (unit: %)

Model 1	3 Layers	4 Layers	5 Layers	6 Layers	7 Layers	8 Layers	9 Layers
256 Nodes	15.52	13.91	13.05	13.67	12.31	12.81	12.67
512 Nodes	14.61	12.95	12.20	12.62	11.78	11.97	11.92
1024 Nodes	13.96	12.48	11.78	12.02	11.64	11.70	11.62
2048 Nodes	13.09	12.14	11.77	11.70	11.67	11.85	11.88

Table 5. MAPE of Model 2 by the Hidden Layer Structure (unit: %)

Model 2	3 Layers	4 Layers	5 Layers	6 Layers	7 Layers	8 Layers	9 Layers
256 Nodes	8.86	7.01	6.23	6.23	5.68	5.74	5.58
512 Nodes	6.98	5.74	5.45	5.28	5.41	5.49	5.65
1024 Nodes	6.19	5.42	5.27	5.47	5.15	5.25	5.15
2048 Nodes	5.78	5.40	5.24	5.11	5.17	5.47	5.44

set. Figure 4 shows the result. The test set MAPEs of ANN are 12.58% for Model 1 and 5.63% for Model 2, showing no significant difference from the results for the valid set. On the other hand, both multiple regression models show MAPE values of more than 28%, indicating that they are significantly less reliable than ANN. It is noteworthy that for multiple regression, the error percentage in Model 4, with the addition of the geographic coordinates, is greater than in Model 3. The regression results presented in Table 6 show that the R-squared value calculated in the process of fitting the model rise slightly from 0.678 in Model 3 to 0.687 in Model 4, but the generalized prediction performance for the test set is rather poor with more explanatory variables. Therefore, it can be determined that the accuracy of housing price predictions using the non-linear pattern data of ANN is greater than when using the linear data constructs of multiple regression.

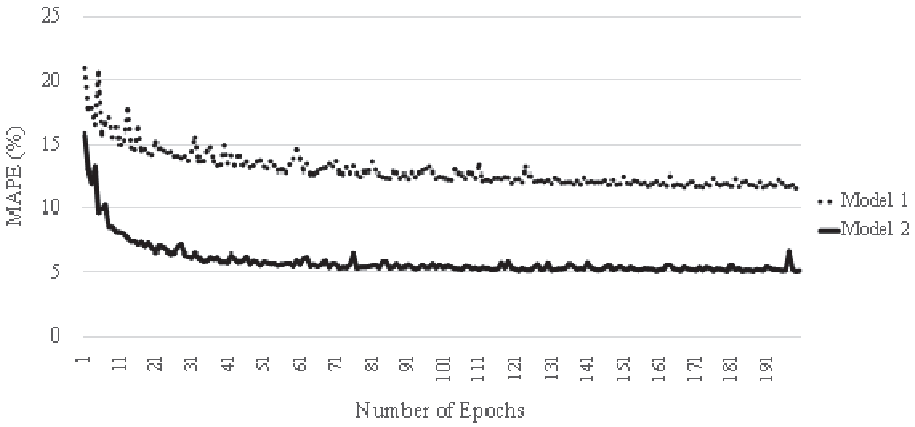

Figure 3. Training Progress by ANN Models (Optimal Configuration)

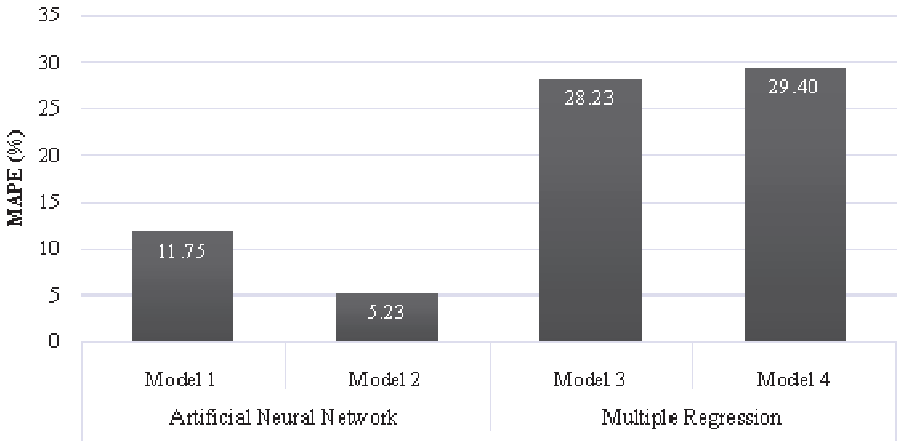

Figure 4. Prediction Performance based on Test Set

Table 6. OLS Multiple Regression Results

Independent Variables	Model 3	Model 4
Intercept	4.963e+04*** (2328.968)	4.963e+04*** (2362.169)
Area	2e+04*** (909.508)	1.995e+04*** (916.616)
Floor	492.2233*** (22.303)	653.7899*** (29.963)
Age	1825.8418*** (81.054)	1580.9397*** (70.083)
PriceIndex	6465.9683*** (288.555)	7093.2793*** (316.425)
AppraisalValue	1.731e+04*** (758.125)	1.49e+04*** (553.134)
Longitude		2372.7439*** (99.510)
Latitude		-3600.2269*** (-140.269)
Altitude		-2109.0055*** (-97.688)
R-Squared	0.678	0.687
Adj. R-Squared	0.678	0.687
No. Observations	1,015,882	1,015,882
Df Residuals	1,015,876	1,015,873
F-statistic	4.270e+05***	2.782e+05***

2) Prediction Error Analysis

Now, in the optimal configuration of the ANN model, we analyze the prediction errors. Figure 5 shows the prediction error based on selling price in all models with the price on the x-axis and the prediction error on the y-axis. At this point, error values greater than zero mean that the estimated price is higher than the actual price, whereas error values less than zero mean the opposite. In the case of Model 1, the scatter plot has a downward bias, but Model 2 shows a significant decrease in this tendency. The linear regression line in red indicates a significant decrease in the slope in Model 2 compared to Model 1. Also, Meanwhile, both models using OLS show strong downward biases compared to ANN. In summary, the model with geographic coordinates using ANN has less tendency to underestimate high-priced houses and overestimate low-priced houses.

The reduction in the error of the housing prices when using the geographic coordinates suggests that there is a spatial effect on the distribution of error. Therefore, we identify the distribution of errors by the geographic location. Figure 6 shows the percentage error rate according to the location in all models. As the color approaches red, it means that the house values are underestimated, and as the color gets closer to green, it means the opposite. Model 1 contains many red dots, especially around southeast Seoul and near the Han River, while other regions contain many green dots. Model 2 shows a significant decrease in the number of over- and underestimated house values in most regions, with a large number of yellow dots representing an accurate estimation compared to Model 1. Also, the prediction errors in Models 3 and 4 with OLS show apparent spatial cluster tendency. As previously confirmed in Figure 3, Seoul has a large number of high-priced apartments in Gangnam-gu, the coveted district of southeast Seoul, and near the Han River. The results suggest that the ANN model with geographic coordinates as input variables effectively captures the spatial distribution of the housing prices.

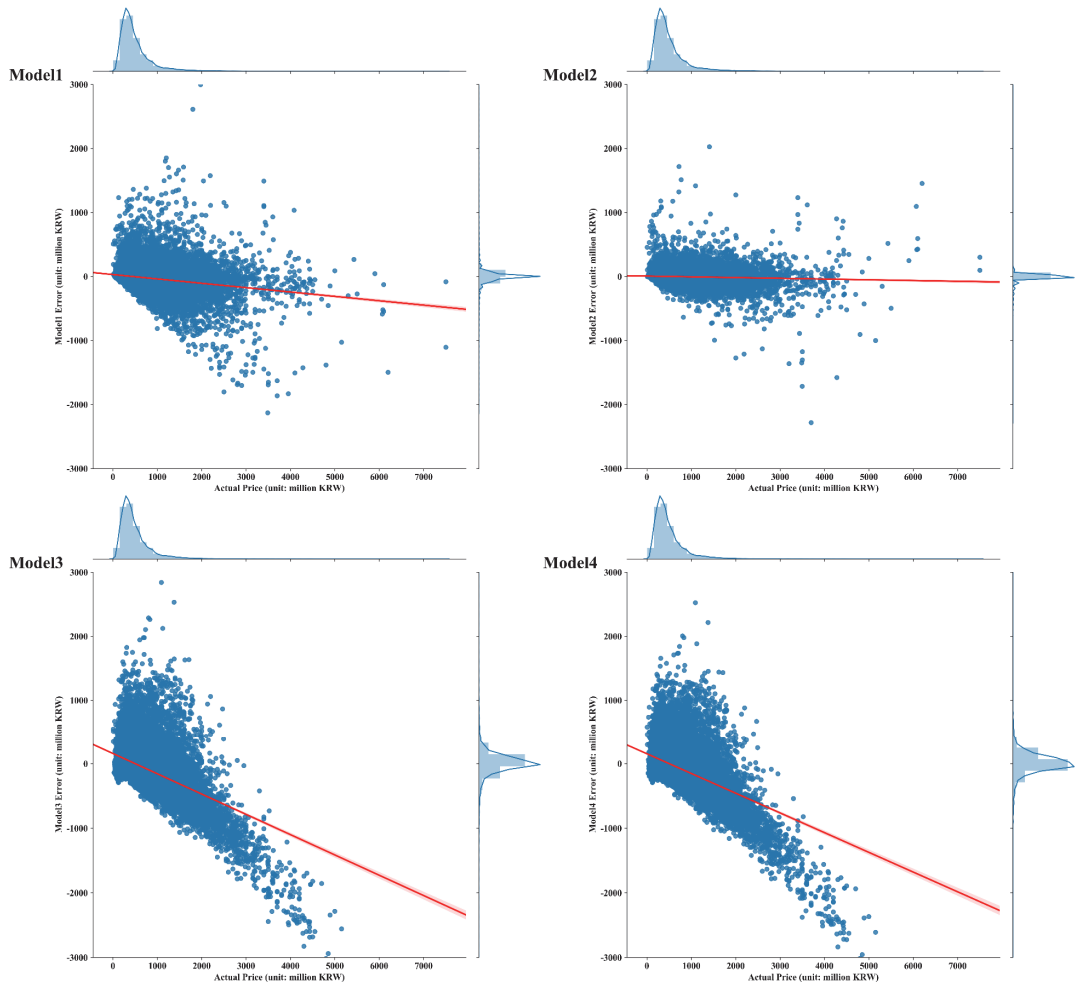

Figure 5. Scatter Plots of House Prices vs. Prediction Errors

Comparing the prediction error distributions of the models confirms that using the geographic coordinates reduces the overall error rate and resolves the spatial cluster tendency of the error. In Table 7, we calculate the percentage error values of all models using Global Moran's I with ArcGIS by generating a spatial weights matrix that considers a house located within a radius of 1,000 meters as a nearby sample.

As the distribution can be significantly clustered when the z-score of Moran's I is higher than 2.58, we can determine that the prediction error of all Models has significant spatial autocorrelation. However, the ANN models have significantly lower Moran's I values than the OLS models and among the ANN models, the error of Model 2 with geographic coordinates is even less than Model 1. Overall results indicate that, unlike OLS, ANN can model overall spatial distribution only with geographic coordinates, and training geographic coordinates in the model accounts for a significant portion of the spatial

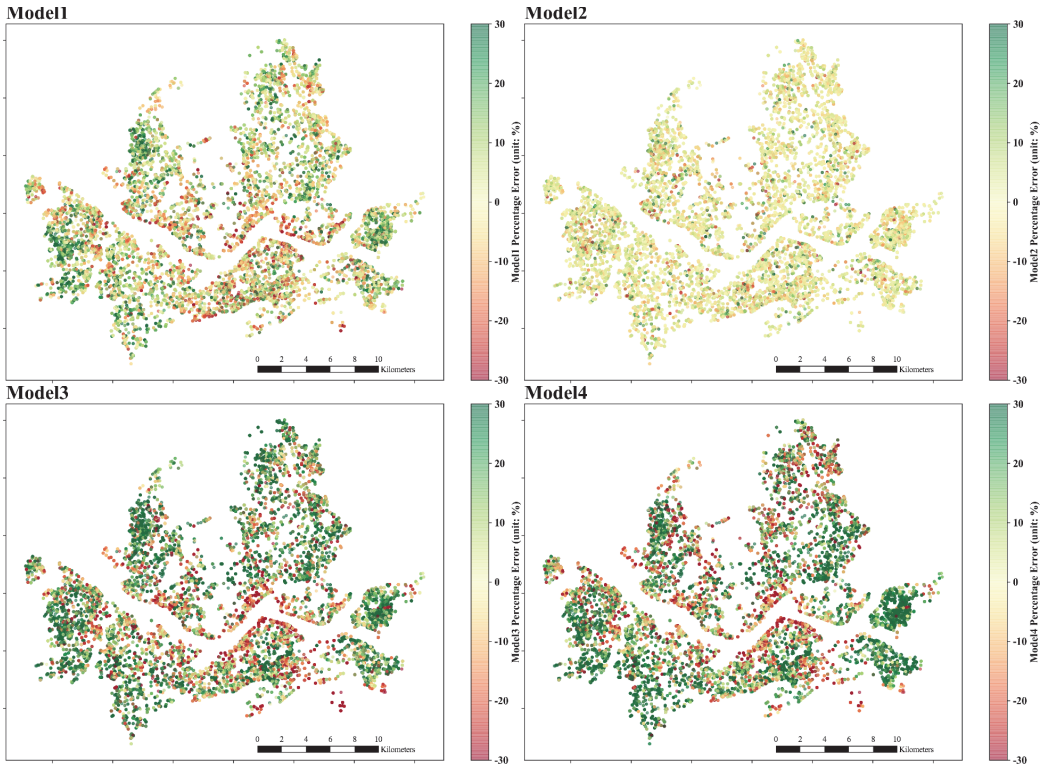

Figure 6. Percentage Error by Location

Table 7. Moran's I Summary

Classification	Model 1	Model 2	Model 3	Model 4
Moran's Index	0.033289	0.001177	0.204559	0.218293
Expected Index	-0.000009	-0.000009	-0.000009	-0.000009
Variance	0.000000	0.000000	0.000000	0.000000
z-score	96.989840	4.448409	483.436181	519.392212
p-value	0.000000	0.000009	0.000000	0.000000

distribution patterns of dependent variables that other variables did not explain.

5. Conclusion

In the real estate market, property values within a small neighborhood can vary widely due to several locational traits. In many cases, these differences are dramatic; a property adjacent to a major subway station may double the price of one next to it. Therefore, a model for evaluating real estate prices should be able to capture small differences that occur with the slight change of location. It means that spatial effect should be considered in estimating housing prices. This has led existing spatial econometrics models to

consider spatial autocorrelation or vary the coefficient according to the location to explain the spatial effect.

In this study, we further introduce deep learning methodology, which has recently received wide attention by replacing several conventional methodologies. We confirm that this technique can also explain the impact of spatial characteristics on housing prices. An important feature of ANN is its pattern recognition ability, which self-models and abstracts the relationships inherent in the data with complex non-linear patterns. (LeCun *et al.*, 2015) Recognition of spatial distribution patterns is a useful feature in the task of estimating the price of real estate, as people also intuitively recognize that certain neighborhoods are more or less affluent depending on their location. In other words, people's recognition of real estate markets can be understood based on how people recognize spatial patterns of a city's real estate markets. Therefore, to understand how ANN can capture spatial distribution patterns of housing prices, we conducted the following experiments: using the housing transaction data of Seoul, we built a hedonic price model based on ANN and OLS methodologies and compared the prediction performance and the spatial autocorrelation of prediction errors between models with and without geographic coordinates.

The overall results reveal that when the ANN model includes geographic coordinates, price estimation errors are significantly lower than the others. In the same condition, the modeling performance of OLS was poor when compared to the ANN methodology. In the ANN models, high-end apartments mainly located in luxury residential areas in Seoul tend to be undervalued and low-end apartments in other areas tend to be overvalued without the geographic coordinates. But this tendency was significantly reduced when the coordinates were utilized in the model. We compared the Global Moran's I for the models' percentage error, i.e., the ratio of error to the actual price, and confirmed that the Moran's I value in the ANN model with coordinates was significantly lower. Based on this result, we concluded that utilizing the pattern recognition ability of ANN makes it possible to train spatial distributions directly into the model, and allows the model to capture spatial characteristics of the housing price that the other features cannot. In conclusion, this study reveals that it is possible to capture the spatial effects of the housing price through locational information by utilizing the pattern recognition ability of machine learning techniques. Further from this study, we believe that the theoretical framework of mass appraisal tasks in the real estate market can be strengthened through additional model optimization and testing various variables in existing hedonic model studies.

Acknowledgement

This work was supported by SNU Environmental Planning Institute.

References

Abidoye, R. B., and Chan, A. P., 2018, "Improving property valuation accuracy: A comparison of hedonic pricing

model and artificial neural network," *Pacific Rim Property Research Journal* 24(1), pp.71-83.

Anselin, L., 1998, "GIS research infrastructure for spatial analysis of real estate markets," *Journal of Housing Research* 9(1), pp.113-133.

Basu, S., and Thibodeau, T. G., 1998, "Analysis of spatial autocorrelation in house prices," *The Journal of Real Estate Finance and Economics* 17(1), pp.61-85.

Can, A., 1992, "Specification and estimation of hedonic housing price models," *Regional science and urban economics* 22(3), pp.453-474.

Case, B., Clapp, J., Dubin, R., and Rodriguez, M., 2004, "Modeling spatial and temporal house price patterns: A comparison of four models," *The Journal of Real Estate Finance and Economics* 29(2), pp.167-191.

Chollet, F., 2017, *Deep learning with Python*, Manning Publications Company.

Chun, H. J., 2012, "Liquidity-related Variables Impact on Housing Prices and Policy Implications," *Journal of the Economic Geographical Society of Korea* 15(4), pp.585-600. (Korean)

Do, Q., and Grudnitski, G., 1993, "A neural network analysis of the effect of age on housing values," *Journal of Real Estate Research* 8(2), pp.253-264.

Dodgson, J. S., and Topham, N., 1990, "Valuing residential properties with the hedonic method: A comparison with the results of professional valuations," *Housing Studies* 5(3), pp.209-213.

Dubin, R. A., 1992, "Spatial autocorrelation and neighborhood quality," *Regional science and urban economics* 22(3), pp.433-452.

Edmonds Jr, R. G., 1984, "A theoretical basis for hedonic regression: A research primer," *Real Estate Economics* 12(1), pp.72-85.

Fotheringham, A. S., Charlton, M. E., and Brunsdon, C., 1998, "Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis," *Environment and planning A* 30(11), pp.1905-1927.

García, N., Gámez, M., and Alfaro, E., 2008, "ANN+GIS: An automated system for property valuation," *Neurocomputing* 71(4-6), pp.733-742.

Géron, A., 2017, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, O'Reilly Media.

Ghorbani, S., and Afgheh, S. M., 2017, "Forecasting the house price for ahvaz city: the comparison of the hedonic and artificial neural network models," *Journal of Urban Economics and Management* 5(19), pp.29-44.

Goodman, A. C., 1978, "Hedonic prices, price indices and housing markets," *Journal of urban economics* 5(4), pp.471-484.

Goodman, A. C., and Thibodeau T. G., 1998, "Housing market segmentation," *Journal of housing economics* 7(2), pp.121-143.

Haykin, S., 2003, *Neural Networks: A Comprehensive Foundation 2nd Edition*, Prentice Hall PTR.

Herath, S. K., and Maier, G., 2011, "Hedonic house prices in the presence of spatial and temporal dynamics," *Territorio Italia land Administration, Cadastre, Real Estate* 1(1), pp.39-49.

Huang, B., Wu, B., and Barry, M., 2010, "Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices," *International Journal of Geographical Information Science* 24(3), pp.383-401.

Jeong, J. H., 2014, "An Analysis of Network Structure in Housing Markets: the Case of Apartment Sales Markets in the Capital Region," *Journal of the Economic Geographical Society of Korea* 17(2), pp.280-295. (Korean)

Khalafallah, A., 2008, "Neural network based model for predicting housing market performance," *Tsinghua Science and Technology* 13(S1), pp.325-328.

Kim, S. J., Jeong, J. H., and Seo, K. C., 2015, "The Conversion Trend of Jeonsei to Monthly Rent Contracts and Its Major Characteristics: The Case of Three Gangnam Districts' APT Rental Market in Seoul," *Journal of the Economic Geographical Society of Korea* 18(3), pp.348-365. (Korean)

Koki, S., 2016, *Deep Learning from Scratch*, OREILLY JAPAN.

Law, S., Paige, B., and Russell, C., 2019, "Take a look around: using street view and satellite images to estimate house prices," *ACM Transactions on Intelligent Systems and Technology* 10(5), pp.1-19.

Law, S., Seresinhe, C. I., Shen, Y., and Gutierrez-Roig, M., 2020, "Street-Frontage-Net: urban image classification using deep convolutional neural networks," *International Journal of Geographical Information Science* 34(4), pp.681-707.

LeCun, Y., Bengio, Y., and Hinton, G., 2015, "Deep learning," *nature* 521(7553), pp.436-444.

Lee, C., and S. H. Kim, 2018, "The Deep Learning Approach to Property Valuation: An Application of a Multilayer Neural Net Model for Estimating House Prices," *Journal of The Korean Regional Development Association* 30(4), pp.179-201.

Lee, W. D., Won, J. S., and Joh, C. S., 2011, "A Study of Correlation between Air Environment Index and Urban Spatial Structure: Based On Land Use and Traffic Data In Seoul," *Journal of the Economic Geographical Society of Korea* 14(2), pp.143-156. (Korean)

Lenk, M. M., Worzala, E. M., and Silva, A., 1997, "High-tech valuation: should artificial neural networks bypass the human valuer?" *Journal of Property Valuation and Investment* 15(1), pp.8-26.

Loo, W. K., 2019, "Predictability of HK-REITs returns using artificial neural network," *Journal of Property Investment & Finance* 38(4), pp.291-307.

Lu, B., Charlton, M., Harris, P., and Fotheringham, A. S., 2014, "Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data," *International Journal of Geographical Information Science* 28(4), pp.660-681.

Mimis, A., Rovolis, A., and Stamou, M., 2013, "Property valuation with artificial neural network: the case of Athens," *Journal of Property Research* 30(2), pp.128-143.

Osland, L., 2010, "An application of spatial econometrics in relation to hedonic house price modeling," *Journal of Real Estate Research* 32(3), pp.289-320.

Pai, P. F., and Wang, W. C., 2020, "Using machine learning models and actual transaction data for predicting real estate prices," *Applied Sciences* 10(17), pp.5832.

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

Poursaeed, O., Matera, T., and Belongie, S., 2018, "Vision-based real estate price estimation," *Machine Vision and Applications* 29(4), pp.667-676.

Raschka, S., and Mirjalili, V., 2019, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, Packt Publishing Ltd.

Rosen, S., 1974, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of political economy* 82(1), pp.34-55.

Stamou, M., Mimis, A., and Rovolis, A., 2017, "House price determinants in Athens: a spatial econometric approach," *Journal of Property Research* 34(4), pp.269-284.

Tay, D. P., and Ho, D. K., 1992, "Artificial intelligence and the mass appraisal of residential apartments," *Journal of Property Valuation and Investment* 10(2), pp.525-540.

Tobler, W. R., 1970, "A computer movie simulating urban growth in the Detroit region," *Economic geography* 46, pp.234-240.

Worzala, E., Lenk, M., and Silva, A., 1995, "An exploration of neural networks and its application to real estate valuation," *Journal of Real Estate Research* 10(2), pp.185-201.

Yang, J. S., 2017, "The Spillover Effect of Public Hosing Policy on Rental Housing Market: The Case of Seoul, Korea," *Journal of the Economic Geographical Society of Korea* 20(3), pp.403-416. (Korean)

Zhou, X., Tong, W., and Li, D., 2019, "Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual Information and Deep Learning," *ISPRS International*

Journal of Geo-Information 8(8), pp.349.

Zietz, J., Zietz, E. N., and Sirmans, G. S., 2008, "Determinants of house prices: a quantile regression approach," *The Journal of Real Estate Finance and Economics* 37(4), pp.317-333.

Correspondence: Kyung-Min Kim, Graduate School of Environmental Studies, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Korea, E-mail: kkim2@snu.ac.kr

교신: 김경민, 08826, 서울특별시 관악구 관악로 1 서울대학교 환경대학원, 이메일: kkim2@snu.ac.kr

최초투고일 2022년 02월 24일
수 정 일 2022년 02월 25일
최종접수일 2022년 03월 15일