VILNIUS TECH
Vilnius Gediminas Technical University

International Journal of Strategic Property Management
ISSN: 1648-715X / eISSN: 1648-9179
2021 Volume 25 Issue 2: 169–178
https://doi.org/10.3846/ijspm.2021.14293

PREDICTING LAND PRICES AND MEASURING UNCERTAINTY BY COMBINING SUPERVISED AND UNSUPERVISED LEARNING

Changro LEE *

Department of Real Estate, Kangwon National University, Chuncheon, Republic of Korea

Received 27 May 2020; accepted 21 November 2020

Abstract. Despite the popularity deep learning has been gaining, measuring the uncertainty within the result has not met expectations in many deep learning applications and this includes property valuation. In real-world tasks, however, rather than simply requiring predictions, assurance of the certainty of the predictions is also demanded. In this study, supervised learning is combined with unsupervised learning to bridge this gap. A method based on principal component analysis, a popular tool of unsupervised learning, was developed and used to represent the uncertainty in property valuation. Then, a neural network, a representative algorithm to implement supervised learning, was constructed, and trained to predict land prices. Finally, the uncertainty that was measured using principal component analysis was incorporated into the price predicted by the neural network. This hybrid approach is shown to be likely to improve the credibility of the valuation work. The findings of this study are expected to generate interest in the integration of the two learning approaches, thereby promoting the rapid adoption of deep learning tools in the property valuation industry.

Keywords: supervised learning, unsupervised learning, property valuation, land prices, uncertainty, principal component analysis, neural network.

Introduction

Deep learning, which has been used with great success for decades in engineering applications such as self-driving automobiles and chat bot services, is increasingly expanding its areas of application, with the real estate industry being no exception with respect to the uptake of this evolving technology. The applications observed in real estate vary greatly, and include the automated valuation of property, brokering and sales, building operations and property management, real estate law and contract services, and insurance.

Especially, property valuation is a central area in which deep learning algorithms such as neural networks have been emerging rapidly. Deep learning is being used across a wide range of areas of property valuation ranging from tax assessment to collateral evaluation and portfolio management. However, deep learning algorithms have not paid due attention to algorithm uncertainty, and thus, the results do not appear to be sufficient to convince property investors and lending institutions to trust the accuracy of the predicted price (Conway, 2018). Representing algorithm uncertainty is indispensable for real-world applications (Ghahramani, 2015). In terms of real-world tasks, rather than simply requiring predictions, the assurance of the nature of these predictions is also desired. For instance, when public-private partnerships (PPPs) are contracted, it would not only be desirable to calculate the monetary value of the infrastructure project but also to estimate the risk and to mitigate it at a controllable level, and the latter would be central to the success of the project. A crucial part of learning involves reserving one's judgment in the case of uncertainty. Surprisingly, measuring uncertainty has not been the aim of many deep learning applications including those used for property valuation.

The conventional valuation model under a neural network would predict a price even if it were given input that is completely unrelated to what it has learned. In the case of new input data in the form of a property the network has never processed before, the uncertainty inherent in the predicted price would be expected to be high, and the network would preferably have to express an appropriate warning for the price. The ability to quantify the uncertainty associated with the predicted price would significantly promote the adoption of deep learning tools in the

*Corresponding author. E-mail: *spatialstat@kangwon.ac.kr*

real estate industry. Supervised learning is well suited to predicting a target variable; thus, it is an excellent tool for predicting land prices. Unsupervised learning is an approach especially suitable for identifying the latent structure inherent in a dataset. The strength of this approach lies in its ability to capture the salient information in the dataset without the help of a target variable. The extraction of this salient information via unsupervised learning can be used to gauge the uncertainty in the predicted price.

In this study, the results obtained by using unsupervised learning are incorporated to the predictions of supervised learning. First, a method based on principal component analysis (hereafter, PCA) that is widely used for unsupervised learning, was developed and used to represent uncertainty in property valuation. Then, a neural network, a popular technique to implement supervised learning, was specified and trained to predict land prices. Last, by matching the uncertainty measured by PCA with the price predicted by the neural network, it was demonstrated that not only can the price of land be estimated, but the confidence that the price may not be equal to the true market value can also be expressed.

The use of artificial intelligence, such as neural networks, for property valuation has been studied by many researchers (Tadeusiewicz, 2011; Zimmermann & Eber, 2014; Mazur-Dudzińska, 2014; Jasiński & Bochenek, 2016). The contribution of our study is an attempt to measure the uncertainty in predicted prices, especially by combining supervised and unsupervised learning. The findings of this study show that a hybrid approach consisting of supervised and unsupervised learning can improve the credibility of the predicted price in property valuation.

This paper first presents background review on supervised and unsupervised learning. Section 2 discusses previous work in which these two learning approaches were combined, and considers the uncertainty associated with property valuation. Section 3 describes the study area and the data that were used, in addition to the specific algorithms that were chosen for measuring the uncertainty and predicting the land prices. Then, the results and implications are provided in Section 4, and finally, a summary of the study and conclusions are presented.

1. Supervised and unsupervised learning

Methods that involve learning from data can be divided into two major categories: supervised learning and unsupervised learning. In supervised learning, a model can be trained using data that are *labeled*, meaning that the data have target variables (land price in this study). The target variable plays a key role in enabling the supervised learning algorithm to predict the correct targets for unseen data. Typical algorithms for supervised learning include linear models (such as regression analysis), gradient boosting machines, support vector machines, and neural networks.

In this study, a neural network was chosen from the toolbox of available supervised learning approaches. The neural network can be viewed as a generalization of linear models that perform multiple stages of processing to predict a target. A neural network is a multi-layered network of neurons (also known as nodes), an example of which is shown in Figure 1, which presents a diagram of a simple neural network with one input layer, one hidden layer with six neurons, and one output layer.

Attempts to estimate property values via neural networks can be found in numerous studies: most of them fall under the category of comparative studies between neural networks and other baseline valuation models. Amri and Tularam (2012) compared the performance of a model based on a neural network with a fuzzy logic system, McCluskey et al. (2013) studied the prediction accuracy of a neural network and geographically weighted regression (GWR), and many other studies compared the performance of neural networks with that of regression models (Morano & Tajani, 2013; Sampathkumar et al., 2015). More recently, algorithms based on neural networks are being utilized for broader areas other than property valuation. For example, Hsu and Juan (2016) designed a decision model for building reuse, and the model was specified on the basis of neural networks.

The advantage of using a neural network over conventional approaches such as linear regression analysis lies in its capability to capture the complex nonlinear relationships inherent in data (Peterson & Flanagan, 2009). Additionally, a neural network can process non-traditional data such as images and texts efficiently, and studies using such non-traditional data for property valuation have started appearing in the literature (Poursaeed et al., 2018; Guo et al., 2020).

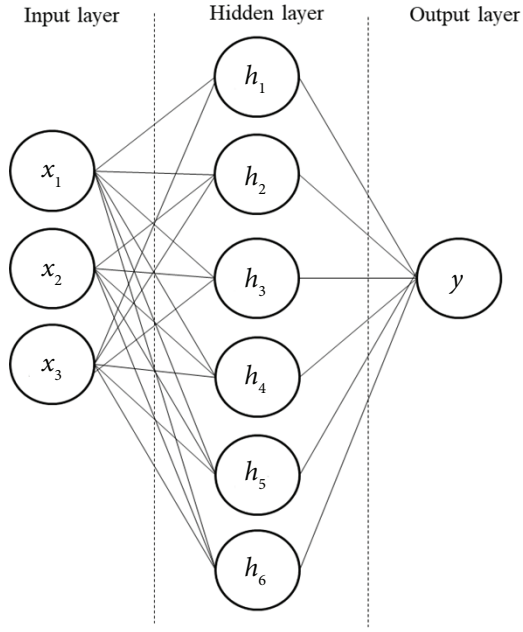


Figure 1. Example of a neural network architecture

In unsupervised learning, the target variable is unavailable and the task of the algorithm is not clearly defined. Contrary to supervised learning, unsupervised learning allows the algorithm to run on its own to discover the salient information and latent structure inherent in the dataset. Well-known algorithms include dimensionality reduction and clustering analysis.

In this study, PCA was chosen from available unsupervised learning approaches. PCA, the most common dimensionality reduction technique, finds a low-dimensional representation of the data while retaining as much of the variation as possible. PCA accomplishes this by considering the correlation among variables. In the case of highly correlated variables, PCA attempts to combine these variables to represent these data with a smaller number of uncorrelated variables. The algorithm performs this correlation reduction iteratively to identify the directions of maximum variance in the original high-dimensional data and projecting them onto a smaller dimensional space (Patel, 2019). The newly derived components are known as principal components.

These components make it possible to reconstruct the original dataset, and PCA strives to minimize the reconstruction error during the search for the optimal number of components. Figure 2 illustrates this process by way of an example dataset.

PCA applications in real estate are observed across various areas: Bourassa et al. (1999) used PCA to extract a set of factors from the original variables and defined the boundaries of housing submarkets in Sydney and Melbourne by using these factors. Wang and Zhang (2013) identified the factors influencing the development of China's real estate market using PCA, and Jiang and Shen (2013) measured urban competitiveness using PCA. More recently, Wilkinson (2014) derived primary attributes using PCA and used them to examine building adaptation events. Chiang and Perng (2018) utilized PCA to eliminate non-discriminatory variables in a study on quality in property management, and Budie et al. (2019) also employed the PCA technique to reduce the number of mediating variables in the path analysis when studying the impact of the office environment on employee satisfaction. Finally, Mrówczyńska et al. (2020) used PCA to compress the input data for application to the vertical displacement measurements of a building.

In summary, the advantage of using PCA lies in the capability of reducing unimportant or redundant variables such that only primary variables should be used for the subsequent analysis. This is especially applicable to analysis involving supervised learning. Table 1 summarizes the differences between supervised and unsupervised learning, as explained in this section.

2. Combination of two learning approaches in property valuation

In this section, the advantages of combining supervised and unsupervised learning are explained. Then, the uncertainty in property valuation is described, and the need to measure the uncertainty in the predicted price by using both supervised and unsupervised algorithms in combination is suggested.

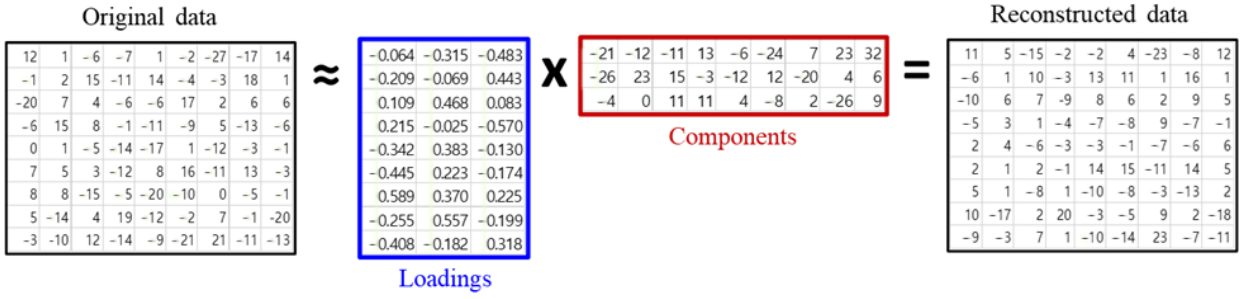


Figure 2. Original and reconstructed dataset in PCA

Table 1. Comparison between supervised and unsupervised learning

	Supervised learning	Unsupervised learning
Data	Input and target data given	Only input data given
Training process	Algorithms are trained using labeled data	Algorithms are trained using unlabeled data
Example	Neural network, support vector machine, gradient boosting machine, random forest	PCA, clustering, association
Accuracy of results	Highly accurate	Less accurate
Goal	To predict the target when new input data are given	To identify the hidden patterns or underlying structure in the input data
Applications	Image recognition, speech recognition, predicting	Pre-processing the data, data dimensionality reduction, outlier detection

2.1. Combining supervised and unsupervised learning

The supervised learning approach has spread throughout the property valuation industry along with the boom in artificial intelligence and the big data hype, and several novel attempts have been made to apply supervised learning to the valuation process. The most representative method to implement supervised learning is the neural network, which has been established as the de facto standard model in a wide range of property valuation projects (Abidoye & Chan, 2017; Sandbhor & Chaphalkar, 2019; Talaga et al., 2019). Especially, a popular application of supervised learning is found in the valuation of houses (Morano et al., 2015; Bazan-Krzywoszanska & Bereta, 2018; Poursaeed et al., 2018). However, most of these studies focused on the accuracy of the estimated house price itself. For example, Morano et al. (2015) estimated apartment unit prices in Bari, Italy by using neural networks, but primarily paid attention to the capability of the model to capture accurate market values of residential properties.

Applications of unsupervised learning are also observed frequently in property valuation, but they are found mainly in the area of data reduction techniques. All the studies concerned with unsupervised learning presented in the previous section are examples of data dimensionality reduction. That is, they used unsupervised learning such as PCA to eliminate redundant variables for the subsequent analysis. To summarize, valuation studies in supervised learning did not pay due attention to the uncertainty in the estimated prices, and most research that focused on unsupervised learning did not expand the application area beyond the reduction of the size of the dataset. Therefore, it is desirable for supervised learning to take into account uncertainty in its prediction, hopefully facilitated by unsupervised learning.

Attempts to combine the benefits of both supervised and unsupervised learning are frequently found in fields other than real estate. For example, Bao et al. (2019) proposed a strategy for integrating unsupervised learning with supervised learning for credit risk assessment. In their study, various supervised algorithms (random forest, gradient boosting, support vector machine, etc.) were combined with unsupervised algorithms such as k-means to improve the performance of credit scoring models. Similar studies in which the benefits of these two approaches are combined have rarely been reported in the property valuation literature.

2.2. Uncertainty in property valuation

Despite the widespread success of neural networks, they have not considered uncertainty with sufficient attention. This shortcoming becomes especially critical for industry applications where the representation of uncertainty is indispensable (Ghahramani, 2015). In this regard, medical diagnosis would be a good example in which confidence is a key factor in decision-making (Johnson et al., 2016). Property valuation is another application field in which an indication of the degree of uncertainty plays a critical role in the decision-making process, for example the approval of real estate loans. Kucharska-Stasiak (2013) stated that the uncertainty in property valuation is exceptionally high owing to both the characteristics of real estate (e.g., fixed location, long useful life, variations in physical features) and the characteristics of the real estate market (e.g., low efficiency, low elasticity of supply and demand). For example, a concert hall would be a good example of a case with rare characteristics, and it would certainly be difficult to predict the price of this property with sufficient confidence.

In valuations prepared for the purposes of loans or insurance, predicted prices are usually provided as single figures, and these figures are often accepted as confident and certain prices by stakeholders. However, from both the point of view of valuation theory and practice, a single figure is a myth (Kucharska-Stasiak, 2013; Mooya, 2016). Although the valuer needs to provide a single figure, a description should be developed to explain the uncertainty in the final figure (Mallinson & French, 2000).

A description of the uncertainty would greatly assist many decision-makers and improve the credibility of the valuation work (Mallinson & French, 2000). In this study, supervised and unsupervised learning algorithms are used in combination to gauge the uncertainty in predicted prices.

3. Application of PCA and a neural network

In this section, the study area and the dataset that was used are briefly explained. PCA is reviewed as a tool to capture the uncertainty in property valuation, and the concept of a reconstruction error is introduced to calculate the uncertainty in the predicted price. The successful application of PCA largely depends on the number of principal components it chooses to use, and the optimal number of components is identified by using trial and error. Finally, a neural network architecture is specified, and an error function is defined to train the neural network.

3.1. Study area and dataset

Seocho-gu, one of the 25 districts in Seoul, South Korea, was chosen for the analysis. In this district, properties are traded more often than in other districts, which implies that it is relatively easy to obtain a larger trade dataset for this district. More importantly, Seocho-gu is well known for its high standard of life and extremely high property prices and is often spotlighted by the real estate media.

The dataset we used consisted of property acquisitions for three consecutive years (2016~2018), which includes the attributes of 3,980 sample sites. These attributes include the size of the site, assessed price, and year reported. The dataset al.o includes the price of the land, which is the price filed by a taxpayer for the purpose of acquisition tax. Figure 3 shows the locations of the sample sites, most of which are densely distributed across the northwestern part of the district. The southeastern area of the district comprises small mountains such as Mt. Guryong, accounting for the sparseness of sample sites there.

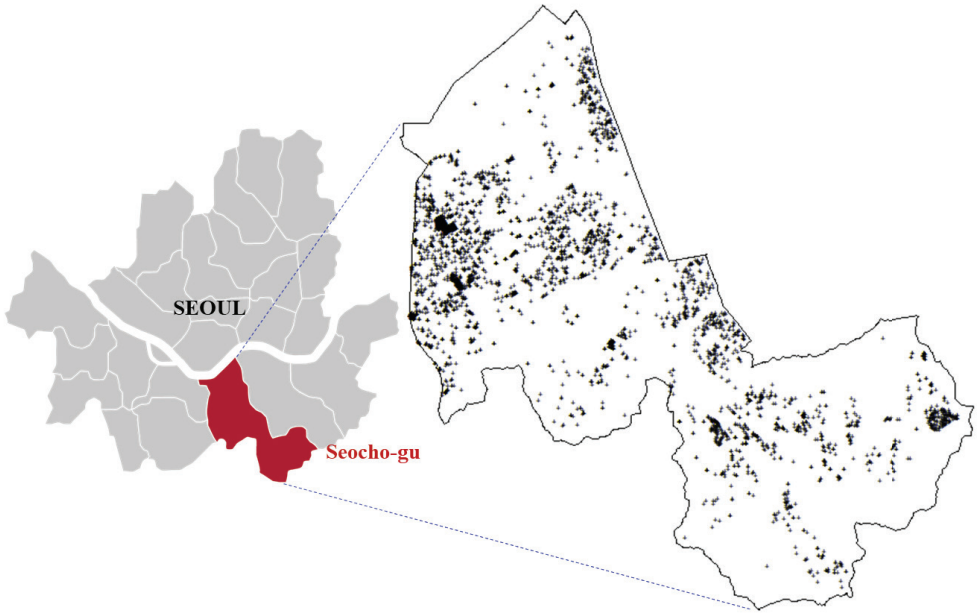


Figure 3. Locations of the 3,980 sites

Table 2. Descriptive statistics of the 3,980 sites

	Min.	Mean	Median	Max.
Acquisition price (USD per m^2)	1	6,767	6,726	41,312
Assessed price (USD per m^2)	17	4,226	3,931	24,812
Site area (m^2)	0.1	1,077	273	928,453
Site zoning	Green belt: 602(15.1%), Preserved: 260(6.5%), Residential: 2,877(72.3%), Quasi-residential: 13(0.3%), Commercial: 228(5.7%)			

Table 2 lists descriptive statistics for the 3,980 sites. The typical acquisition price for a site is approximately USD 6,700 per square meter, and the corresponding assessed price is approximately USD 4,000 per square meter. The median area of a purchased site is 273 m^2, and most sites in the dataset are zoned for residential use (72.3%). The 3,980 sites in Table 2 were randomly divided into training data (80%, 3,182 sites) and test data (20%, 798 sites) for subsequent analysis[1].

3.2. PCA and the number of principal components

PCA aims to capture the most salient information of the original observations in such a way that the original values can be reconstructed from the reduced dataset as accurately as possible. However, PCA cannot capture all the information of the original dataset as it reduces the dimensionality of data; therefore, a certain error is introduced as PCA reconstructs the reduced dataset back to the original number of components.

[1] A 5-fold cross validation strategy was applied during neural network training. 636 sites (20% of 3,182 sites) were used as the validation dataset, and the remaining 2,546 sites (80% of 3,182 sites) were used as the training dataset.

In the context of the property acquisition dataset, PCA is expected to obtain the largest reconstruction error on those sites that are traded the least often and have rare characteristics. The rarer the site is, the more likely it is to be a specialized site. Because a specialized site is presumably different than common sites, the reconstruction error for a specialized site would be the largest. Thus, the reconstruction error is a good measure to calculate the rarity of each site. This error is defined as the sum of the squared difference between the original data matrix and the reconstructed matrix, as shown in Formula 1.

$$\textit{reconstruction error} = \Sigma\left(\mathbf{O}-\mathbf{R}\right)^2, \quad (1)$$

where **O** is the original matrix, and **R** is the reconstructed matrix. The sum of the squared difference is scaled by the min-max range scaler, such that all the reconstruction errors lie between zero and one. More specifically, a specialized site would have a reconstruction error close to one, whereas that of a common site would be close to zero.

The reconstruction error for rare properties–those of which the prices are likely to be the most difficult to predict–should preferably be as large as possible and that of the remaining properties as small as possible. When using PCA, the reconstruction error largely depends on the number of principal components the algorithm maintains

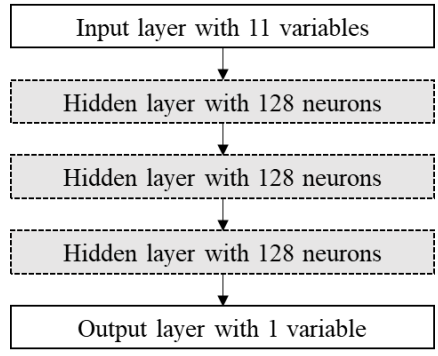
and uses to reconstruct the original observations. If the algorithm was to maintain too many principal components, PCA may too easily reconstruct the original observations, with the result that the reconstruction error is minimal for all of the observations. Conversely, if the algorithm was to maintain too few principal components, PCA may not be able to reconstruct any of the original observations, not even the common sites.

The dataset used in this study has 11 variables (11 dimensions), and PCA attempts to project the dataset onto a smaller subspace of p dimensions. PCA projects the data to a lower dimensional space using singular value decomposition (SVD). The PCA used in this study utilizes a randomized truncated SVD by employing the method of Halko et al. (2011). It generates new variables known as principal components, which is represented by the eigenvectors of the covariance matrix (the correlation matrix in this study). A trial-and-error approach was used to select four principal components ($p = 4$), to build an efficient reconstruction model. Interpreting the extracted principal components is a task that involves subjective judgment, and we did not attempt to interpret the meanings of the four principal components because our primary goal is to reconstruct the original variables efficiently[2].

3.3. Neural network architecture

Eleven variables are utilized to estimate the price of land in this study: year of acquisition; site area; size zoning; site use; slope of the site; the width of the road (m) onto which the site abuts; assessed price; the ratio of the assessed price over the overall market value[3]; latitude and longitude; and the median assessed price for each ZIP code area[4]. The target variable is the acquisition price (sales price) reported on the tax return.

The architecture of a neural network is usually created in a layer-by-layer manner. That is, the input layer is first initiated, after which more than one hidden layer is created and added to the input layer, and then the output layer is added at the end of the architecture. The power of a neural network comes from the ability to learn the information in the data via the hidden layers, and the number of hidden layers depends on the complexity of the data structure and the judgment of the researcher. The final architecture, shown in Figure 4, was chosen based on a grid search that uses cross-validation to evaluate the possible combinations of hyperparameter values (the number of hidden layers and neurons in this case). As shown in the figure, a neural network with three hidden layers and 128 neurons on each layer was employed for the analysis.



Notes: dropout with 20% ratio is used for each hidden layer to prevent over-fitting during training. See Chollet (2018) for details related to layers and dropouts.

Figure 4. Network architecture chosen for the analysis. Total parameters to be trained is 34,689

When training a neural network, an error function needs to be determined to evaluate convergence. In this study, we used the mean squared error (MSE) and mean absolute error (MAE) functions:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y} - y\right)^2; \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y} - y\right|, \tag{3}$$

where $\hat{y}$ is the price predicted by the network, and y is the observed price. These two indexes are used to determine the network convergence. A neural network is said to converge when the network yields an error close to the minimum, that is, when the error no longer significantly decreases during training.

4. Results and discussion

The distribution of the reconstruction error obtained with PCA is presented, and its implications are discussed. The neural network specified in the preceding section is trained, and the goodness-of-fit of the network is examined. Then, the relationship between the reconstruction errors from PCA and the residuals from the network is investigated, and the application of the findings of this study to real-world valuations is discussed.

4.1. Results

Figure 5 shows the distribution of the reconstruction error for the 3,980 sites. The errors are min-max scaled; therefore, the minimum and maximum values are always zero and one, respectively.

As shown in Figure 5, the reconstruction errors are distributed in a right-skewed manner (positive skewness). Most of the errors are less than 0.2, with the histogram in the inset showing the errors below 0.2 in a more detailed manner. Figure 5 suggests that PCA has little difficulty in recovering most cases, as evidenced by the fact that most errors have scores below 0.2. However, PCA experienced difficulty in reconstructing a few rare cases for which error scores above 0.2 were obtained.

[2] Interpretation of the principal components is usually based on finding which original variables are the most strongly correlated with each component, and this is a subjective decision.

[3] This ratio is a useful indicator of the assessed price level.

[4] The area corresponding to each ZIP code is a good alternative to a neighborhood, and its median price indicates an average level of land price for the neighborhood.

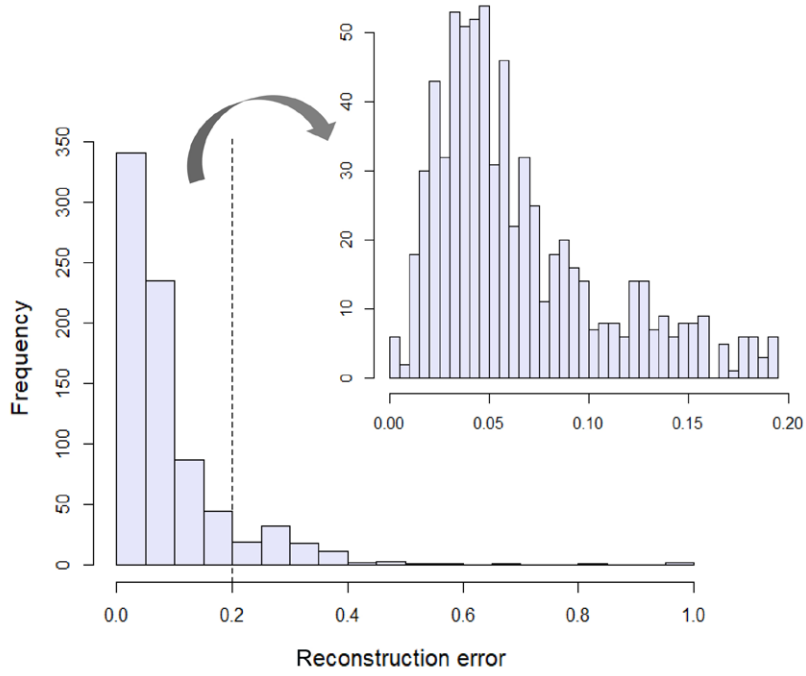


Figure 5. Distribution of the reconstruction error

Table 3. Sites with the two highest and two lowest reconstruction errors

Reconstruction error	Site zoning	Use	Site area (m^2)	Acquisition price (USD per m^2)
1.0000	Residential	Forest	2,045.9	984
0.9991	Preserved	Electric transformer facility	3,504.9	390
0.0004	Residential	Single family house	187.7	6,938
0.0000	Residential	Single family house	254.5	5,264

Table 3 lists the sites with the two highest and two lowest reconstruction errors. A forest in an area zoned for residential use and a site for an electric transformer facility are those with the highest reconstruction errors. Needless to mention, these two sites have rare characteristics and would be difficult to find in the real estate market. In contrast, PCA produced considerably low reconstruction errors for sites occupied by single-family houses, because they are relatively standardized properties with commonplace characteristics, and are traded frequently in the real estate market.

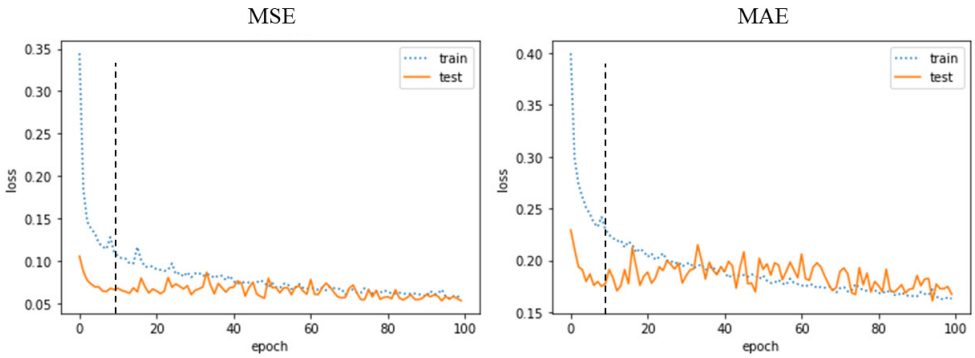
As for the neural network, Figure 6 shows the decrease in the MSE and MAE during the training process after 100 epochs, indicating that the network seems to converge sufficiently after 10 epochs. Thus, the network used in this study was trained for 10 epochs to avoid over-fitting[5].

The goodness-of-fit of the neural network for the test data (20% of the entire dataset) is shown in Figure 7. Both the predicted and observed prices were normalized, and the predicted prices appear to follow the observed ones closely[6]. This led us to conclude that the use of a network trained in this way to predict the prices of untraded sites in subsequent analysis is not problematic.

A large residual from the neural network signifies a large prediction error, and can be interpreted to indicate that the neural network has difficulty in predicting its price accurately. In the same context, a large reconstruction error indicates that PCA has difficulty in recovering the original input values, implying that the site is unordinary. Figure 8 shows the relationship between the reconstruction errors and the residuals from the neural network for the test data. Most data points lie in the bottom-left

[5] The implementation details are as follows: Adaptive moment estimation (Adam) optimizer and Glorot initialization with a uniform distribution were used. A constant learning rate of 0.001 was used. Different learning rates or different learning schedules (exponential scheduling, power scheduling, etc.) made little impact on the result. A rectified linear unit (ReLU) activation function was used for the three hidden layers, and the linear activation function was used for the output layer. A dropout rate of 0.2 was applied to each hidden layer to prevent over-fitting, and the batch size was 128.

[6] The average percent error [(observed price – predicted price) / observed price × 100] for the test data is 12%, and this degree of error is acceptable compared to established standards such as those provided by the International Association of Assessing Officers (IAAO, 2013): IAAO and many states in the US set an average percent error of 25% as the acceptance threshold for tax assessment.



Notes: X-axis: epoch, Y-axis: MSE (left) and MAE (right), Dotted line for training data, Solid line for test data. Vertical dotted line indicates the 10th epoch.

Figure 6. Convergence of the neural network

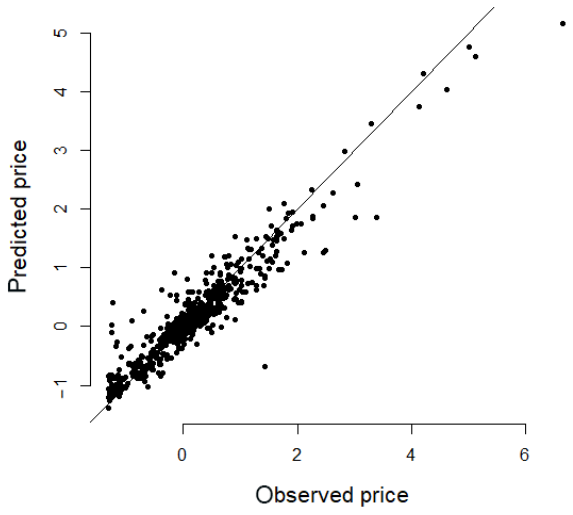


Figure 7. Goodness-of-fit in the neural network for the test data

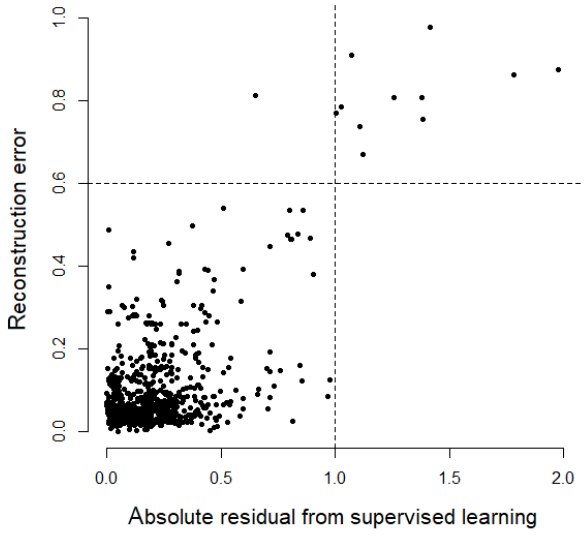


Figure 8. Reconstruction errors vs. residuals

region, meaning that most sites are not problematic in terms of predicting their prices and reconstructing their original input values. However, a few data points are located far from the origin, strongly implying that they represent rather unique sites of which the prices cannot be predicted accurately. Figure 8 indicates that most cases with relatively large residuals (larger than 1.0) could be identified by setting the threshold of the reconstruction error at 0.6.

4.2. Quantifying uncertainty

The largest concern with the automated valuation model is that an appropriate tool to gauge the confidence for the predicted prices of untraded sites is not available. Because they have not been traded, their sales prices are unknown, thus their residuals (sales prices minus predicted prices), an indication of the degree of prediction errors, cannot be calculated. Figure 8 shows the potential to represent uncertainty in supervised learning with the help of unsupervised learning. As shown in the figure, the reconstruction error also tends to increase as the residual increases. Thus, the reconstruction error could be employed to capture the confidence levels of the predicted prices for untraded sites as the reconstruction error can also be calculated for these sites. For example, a site with a reconstruction error larger than 0.6 could be classified as one with a predicted price of grade B, whereas a site with an error below 0.6 could be classified as one with a grade A predicted price. It would additionally be possible to classify the uncertainty in prices in a more detailed manner, for example, grades A through D depending on the magnitude of the reconstruction error[7].

In real-world land valuations such as the administration of tax assessment, the approach proposed in this study could be used as follows: first, a supervised learning algorithm such as a neural network would be trained on the basis of a few sales comparables (labeled samples), that

[7] Grading A through D is conventionally used by financial institutions in South Korea for indicating the loan risk.

is, traded sites of which the sales prices are known. Second, this trained supervised learning algorithm would be used to assess all the sites in the jurisdiction area, including those that are untraded. Third, an index obtained from an unsupervised learning method such as PCA would be used to represent the uncertainty in the assessed value of each property including those that are both traded and untraded. Assessed sites of which the values are characterized by high uncertainty may be sent to a tax assessor for reappraisal or field inspection. By using this process, the assessed value pronounced by the government would win the taxpayer's confidence, reducing unnecessary administration costs such as tax appeals.

Conclusions

In this study, the benefit of unsupervised learning was combined with the predictive ability of supervised learning. It was proposed that the uncertainty in property valuation could be quantified by the reconstruction error produced by PCA. The reconstruction error was used as a proxy for gauging the confidence level of the prediction. Then, a neural network with three hidden layers was specified to predict land prices. Finally, incorporation of the uncertainty measured via PCA in the prices predicted by the neural network enabled red flags to be placed alongside predictions with large reconstruction errors. Based on the analysis results, the following conclusions can be drawn:

- PCA can be an effective approach to filter sites with rare characteristics (sites with "hard to predict prices"), by producing larger reconstruction errors for specialized properties.
- Neural networks can be employed with little difficulty as an effective tool to predict land prices as shown in Figure 7, where the predicted prices follow the observed prices closely.
- The residuals from the neural network and the reconstruction errors from PCA were found to have an approximately positive correlation, and this indicates that the reconstruction errors can be used to represent uncertainty in the predicted prices of untraded sites, of which the residuals cannot be estimated because of the absence of observed prices.

Incorporating the uncertainty could drastically improve the applicability of supervised learning approaches in tasks for which expressing confidence is crucial. The inclusion of an algorithm in our toolset to determine the uncertainty enables us to use an explicit approach for properties with high uncertainty.

This study attempted to embrace the strengths of both supervised and unsupervised learning in the field of property valuation, and the results in this study are expected to promote the integration of the two learning approaches in real-world projects, such as collateral valuation and property tax assessment.

Funding and disclosure statement

The author declares no funding and no conflict of interest.

References

Abidoye, R. B., & Chan, A. P. (2017). Artificial neural network in property valuation: application framework and research trend. *Property Management*, *35*(5), 554–571. https://doi.org/10.1108/PM-06-2016-0027

Amri, S., & Tularam, G. A. (2012). Performance of multiple linear regression and nonlinear neural networks and fuzzy logic techniques in modelling house prices. *Journal of Mathematics and Statistics*, *8*(4), 419–434. https://doi.org/10.3844/jmssp.2012.419.434

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*, 301–315. https://doi.org/10.1016/j.eswa.2019.02.033

Bazan-Krzywoszanska, A., & Bereta, M. (2018). The use of urban indicators in forecasting a real estate value with the use of deep neural network. *Reports on Geodesy and Geoinformatics*, *106*, 25–34. https://doi.org/10.2478/rgg-2018-0011

Bourassa, S. C., Hamelink, F., Hoesli, M., & MacGregor, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics*, *8*(2), 160–183. https://doi.org/10.1006/jhec.1999.0246

Budie, B., Appel-Meulenbroek, R., Kemperman, A., & Weijs-Perree, M. (2019). Employee satisfaction with the physical work environment: the importance of a need based approach. *International Journal of Strategic Property Management*, *23*(1), 36–49. https://doi.org/10.3846/ijspm.2019.6372

Chiang, T. Y., & Perng, Y. H. (2018). A new model to improve service quality in the property management industry. *International Journal of Strategic Property Management*, *22*(5), 436–446. https://doi.org/10.3846/ijspm.2018.5226

Chollet, F. (2018). *Deep Learning mit Python und Keras: das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG.

Conway, J. (2018). *Artificial intelligence and machine learning: current applications in real estate* [Master's thesis]. Massachusetts Institute of Technology, Cambridge, Massachusetts.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459. https://doi.org/10.1038/nature14541

Guo, J. Q., Chiang, S. H., Liu, M., Yang, C. C., & Guo, K. Y. (2020). Can machine learning algorithms associated with text mining from internet data improve housing price prediction performance? *International Journal of Strategic Property Management*, *24*(5), 300–312. https://doi.org/10.3846/ijspm.2020.12742

Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, *53*(2), 217–288. https://doi.org/10.1137/090771806

Hsu, Y. H., & Juan, Y. K. (2016). ANN-based decision model for the reuse of vacant buildings in urban areas. *International Journal of Strategic Property Management*, *20*(1), 31–43. https://doi.org/10.3846/1648715X.2015.1101626

International Association of Assessing Officers. (2013). *Standard on ratio studies*. IAAO.

Jasiński, T., & Bochenek, A. (2016). Prognozowanie cen nieruchomości lokalowych za pomocą sztucznych sieci neuronowych [Apartment prices forecasting by the artificial

neural networks]. *Studia i Prace WNEiZ US*, *45*, 317–328. https://doi.org/10.18276/sip.2016.45/1-25

Jiang, Y., & Shen, J. (2013). Weighting for what? A comparison of two weighting methods for measuring urban competitiveness. *Habitat International*, *38*, 167–174. https://doi.org/10.1016/j.habitatint.2012.06.003

Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., & Clifford, G. D. (2016). Machine learning and decision support in critical care. *Proceedings of the IEEE*, *104*(2), 444–466. https://doi.org/10.1109/JPROC.2015.2501978

Kucharska-Stasiak, E. (2013). Uncertainty of property valuation as a subject of academic research. *Real Estate Management and Valuation*, *21*(4), 17–25. https://doi.org/10.2478/remav-2013-0033

Mallinson, M., & French, N. (2000). Uncertainty in property valuation – the nature and relevance of uncertainty and how it might be measured and reported. *Journal of Property Investment & Finance*, *18*(1), 13–32. https://doi.org/10.1108/14635780010316636

Mazur-Dudzińska, A. (2014, 12–16 maja). Sztuczne sieci neuronowe w modelowaniu zjawisk zachodzących na rynku nieruchomości [Application of the artificial neural networks to the real estate market analysis]. In *XVIII Międzynarodowa Szkoła Komputerowego Wspomagania Projektowania, Wytwarzania i Eksploatacji* (pp. 381–388), Szczyrk, Polska. Stowarzyszenie Inżynierów i Techników Mechaników Polskich.

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, *30*(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

Mooya, M. M. (2016). *Real estate valuation theory*. Springer Books. https://doi.org/10.1007/978-3-662-49164-5

Morano, P., & Tajani, F. (2013). Bare ownership evaluation. Hedonic price model vs. artificial neural network. *International Journal of Business Intelligence and Data Mining*, *8*(4), 340–362. https://doi.org/10.1504/IJBIDM.2013.059263

Morano, P., Tajani, F., & Torre, C. M. (2015). Artificial intelligence in property valuations: an application of artificial neural networks to housing appraisal. *Advances in Environmental Science and Energy Planning*, 23–29.

Mrówczyńska, M., Sztubecki, J., & Greinert, A. (2020). Compression of results of geodetic displacement measurements using the PCA method and neural networks. *Measurement*, *158*, 107693. https://doi.org/10.1016/j.measurement.2020.107693

Patel, A. A. (2019). *Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data*. O'Reilly Media, Inc.

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

Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. *Machine Vision and Applications*, *29*(4), 667–676. https://doi.org/10.1007/s00138-018-0922-2

Sampathkumar, V., Santhi, M. H., & Vanjinathan, J. (2015). Evaluation of the trend of land price using regression and neural network models. *Asian Journal of Scientific Research*, *8*(2), 182–194. https://doi.org/10.3923/ajsr.2015.182.194

Sandbhor, S., & Chaphalkar, N. B. (2019). Impact of outlier detection on neural networks based property value prediction. In *Information systems design and intelligent applications* (pp. 481–495). Springer. https://doi.org/10.1007/978-981-13-3329-3_45

Tadeusiewicz, R. (2011). Artificial intelligence applied to the intelligent buildings. In *6th International Congress on Intelligent Building Systems* (pp. 1–11). https://www.academia.edu/39590553/Artificial_Intelligence_Applied_to_the_Intelligent_Buildings

Talaga, M., Piwowarczyk, M., Kutrzyński, M., Lasota, T., Telec, Z., & Trawiński, B. (2019, September). Apartment valuation models for a big city using selected spatial attributes. In *International Conference on Computational Collective Intelligence* (pp. 363–376). Springer. https://doi.org/10.1007/978-3-030-28377-3_30

Wang, X., & Zhang, J. (2013). Principal component analysis of influencing factors of the development of China's real estate market. In *ICCREM 2013: Construction and Operation in the Context of Sustainability* (pp. 1027–1035). https://doi.org/10.1061/9780784413135.098

Wilkinson, S. (2014). The preliminary assessment of adaptation potential in existing office buildings. *International Journal of Strategic Property Management*, *18*(1), 77–87. https://doi.org/10.3846/1648715X.2013.853705

Zimmermann, J., & Eber, W. (2014). Consideration of risk in PPP-projects. *Business, Management and Education*, *12*(1), 30–46. https://doi.org/10.3846/bme.2014.03