Determinants of house prices in Amsterdam: Hedonic regression versus Neural Network performance

Student name: Mahdiazhari Austian (Amsterdam University College)

Student email: mahdi.mahdiazhariaustian@student.auc.nl

Student number: 11206004

Major: Social Sciences (Economics and Information Track)

Supervisor: prof. dr. Jos van Ommeren (Vrije Universiteit Amsterdam)

Supervisor email: jos.van.ommeren@vu.nl

Reader: dr. Michael P. McAssey (Amsterdam University College)

Reader email: M.P.McAssey@auc.nl

Tutor: dr. D.L. (Diederik) van Werven

Capstone Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Arts.

Date: 19-12-2018

Word Count: 7805 words

Abstract

Ongoing studies of machine learning within the context of the Dutch housing market are scarce. To add to this, and to explain the factors affecting housing estimation, this research has two objectives: 1. To investigate determinants of house prices in Amsterdam. 2. To improve price prediction performance. In this thesis, hedonic regression is firstly implemented. After that, neural networks were tested to see if further improvements of the prediction performance could be found. The study was conducted on a dataset of house prices in Amsterdam from the Dutch real estate agent association (NVM). The models were tested on a dataset of around 52,000 samples. It is found that some house features in Amsterdam are negatively related with the predicted house price, there is a u-shaped relation between construction year and expected house price, and the model expects that apartments cost more than the other house types except canal houses. With regards to prediction performance, the neural network model performed 10% better compared to the baseline hedonic regression method.

Keywords: hedonic regression, price prediction, Amsterdam, real-estate, neural network

Table of Contents

List of Tables

List of Figures

Introduction

Real estate agencies are usually involved in activities such as being the intermediary between buyers and sellers, advising customers on home buying and selling, and residential property valuation. The valuation of property results in reports that are used for mortgage requests. This valuation report is created by real estate agents and validated by an external institution which determines if the estimated price is a good indication for the residence. This manual appraisal of houses costs time and money. It takes around two weeks for this valuation process to be done and would cost the applicant around 300 euros. Moreover, the result of this valuation is mostly subjective and not accurate (Schekkerman, 2004). Schekkerman also found that the average accuracy for two-thirds of valuations has an error of less than 20%.

There are cases where it is more feasible to have a housing price indication rather than having to evaluate these houses manually. These estimations are done by the Automated Valuation Model (AVM), services that provide house valuations using mathematical modeling on the current database of existing transactions. Contrary to manual valuations, AVM provides accurate and objective property valuations. This price estimation for housing has been done most commonly with hedonic regression methods. In the Netherlands, the leading AVM provider is Calcasa. They provide a reliable, quick, and accurate estimate of the current market value of a given residential place based on hedonic regression techniques (Calcasa, 2018).

The research on modeling housing prices and property values has been done since as early as the 1960s, with the advent of the hedonic pricing model (Kain & Quigley, 1970). It has been the most widespread method as this approach allows total housing expenditure to be constructed from the values of its individual components. In the pioneering work done by Rosen (1974), goods are defined as having characteristics that possess utility. He defined hedonic prices as “the implicit prices of attributes and are revealed to economic agents from observed prices of differentiated products and the specific amounts of characteristics associated with them.” With the availability of housing transactions data, hedonic regression can be applied to solve this problem by assuming property to be a heterogeneous good that can generally be divided into three main kinds of characteristics: structural, locational, and environmental.

In the topic of the housing and real estate market, the hedonic model is not only used for predicting prices but also useful for its model structure. For example, finding out which characteristics have the highest factor on the price or what the influence of each characteristic is on the price. Malpezzi Ozanne, and Thibodeau (1987) paralleled housing to different grocery bundles bought by customers of a supermarket. Each customer buys a unique bundle of groceries in which the different items inside each bundle determine their overall price. Each house characteristic has a contributing effect on the overall price which can be estimated by regression and the hedonic method.

Malpezzi et al. also suggested that this approach provides a direct way to adjust for quality differences between old and new housing, thus providing a way to make estimates of economic depreciation. Sirmans, Macpherson, and Zietz (2005) studied various price prediction hedonic models over the previous decade. Their study identified the effect on the house price of the most common housing characteristics that were used in several well-known studies. They found that the most common and important attributes are the characteristic variables such as lot size, square footage of the house (area), and number of rooms. This is supported by Marjan, Kilibarda, Lisec, & Bajat (2018), who concluded that structural characteristics were the most contributing characteristics to these house prices.

Visser, Van Dam, and Hooimeijer (2008) demonstrated that house characteristics do contribute to house price variations in the Netherlands. Additionally, Debrezion, Pels, and Rietveld (2011) tested a hedonic house price model in Amsterdam, Rotterdam, and Enschede with the addition of railway access as explanatory variables. These studies found that most common structural attributes have a positive relationship with the house price in Amsterdam. The common structural attributes that they have are: size, total area, garage, number of rooms, and garden. Contrary to Sirmans, Macpherson, and Zietz (2005) and Debrezion et al., Visser et al. did not find a negative relation between house price and age of the house. In fact, Visser found that Dutch houses are more expensive the older it is compared to 1970. However, house prices start to rise if its construction year is after the 1990s. In Debrezion et al. (2011), there is a negative relationship between construction year and house price.

Machine learning has also been employed to the problem of house price prediction. Nguyen and Cripps (2001) compared multiple regression analysis with artificial neural networks (ANN) on a dataset of single-family houses with three different training set sizes. They used house characteristics such as size in square feet area, number of rooms, number of bathrooms, year built, and garage availability. Nguyen and Cripps found that the performance of ANN starts to outperform multiple regression as the dataset size increases. They also concluded that multiple regression performs better if the comparison was done on smaller datasets.

Limsombunc, Gan, and Lee (2004), found that hedonic price models based on regression do not perform better than neural networks. Selim (2009) obtained improved price prediction performance in Turkey after utilizing a feedforward ANN. He found that the ANN predictions produced significantly less errors. The neural network model does not always perform better than the standard regression, this is not the case in Kontrimas & Verikas (2011), Mach (2017) and

Worzala, Lenk, & Silva (1995), who concluded that neural networks performed worse in comparison to standard regression. However, Mach (2017) concluded that the differences between the performances of both models were statistically insignificant so neither model performed better than the other.

Feng and Jones (2015) found that most studies concerning ANN in the context of housing utilized relatively small sample sizes of less than 6,000 observations. They indicated the need for more relatively large-scale neural network implementation results within the realm of the housing market. This is in line with Nguyen and Cripps' findings, where a larger dataset should benefit the neural network more.

Not only within Amsterdam, but also in the Netherlands, literature concerning application of machine learning in the context of residential market seem to be scarce. In Kagie and Wezel (2007), a hedonic model based on boosting is applied to data from the Dutch housing market. Their boosting model is combined with machine learning's decision trees to create predictions. They found a 40% improvement over a baseline linear regression approach. In van der Burgt (2017), the regression approach for house price prediction was improved by utilizing ensemble models. They did not find a major accuracy improvement after utilizing ensemble, however, the average error and standard deviation was better compared to the standard linear regression. These findings suggest that the prospect of applying machine learning to the context of the Dutch housing market could prove beneficial. Besides ensemble and decision trees, no published literature was found regarding the use of other machine learning techniques in predicting Dutch house prices.

This study examines determinants of Amsterdam house prices using housing transactions data from the Dutch Brokers Association (NVM) in order to paint a clearer picture of the factors affecting house prices. To estimate the hedonic model, multiple linear regression analysis will be utilized. Considering the role of ANN in possibly improving the accuracy of price estimation, this study also implements an ANN model from the data and compares its performance with the baseline regression model. It could be expected that ANN will outperform the baseline model, based on previous literature.

Real estate agencies can benefit from the analysis by knowing what factors buyers look when considering a house purchase. Results from the performance comparison could help companies such as Calcasa in assessing whether to incorporate newer methods such as ANN for their AVM. Moreover, with the application of newer machine learning concepts, relevant theory will be more complete and solid as previous literature suggests application of ANNs in the context of housing markets with larger datasets seem lacking. The remainder of this paper proceeds with a brief introduction on the models: hedonic and ANN, after which is the data and methods section. Next is the results section, and finally, the paper ends with the conclusion and discussion section.

Theoretical Background

Hedonic Price Model

Rosen (1974) was the first to present the hedonic pricing theory; where an item is valued by its characteristics. The total price of this item can be considered a total of the price of each of its attributes, therefore there is a unique implicit price per attribute assuming an equilibrium market. The implication from Rosen is that the price of an item can be regressed on its characteristics to investigate how each attribute uniquely contributes to the overall price of the unit. Since this formulation, this theory has been extended to the residential market, becoming a tool for urban analysis and property markets. The regression of a house price on various property describing factors gives the marginal contribution of the house characteristic on its price (Rosen, 1974). This theory guides both consumers and producers to make decisions regarding important aspects of residential housing. The standard linear regression is the default method used to build these hedonic models.

The basic model is most commonly represented as:

$$Y = f\ (S\beta, L\gamma, E\delta) + \mathcal{E} \qquad (1)$$

where Y is the observed house price, S the matrix of structural characteristics of the housing, L the matrix of locational characteristics, E the matrix of environmental characteristics. Next, β, γ, and δ parameters for S, L, and E respectively. Finally, there is Ɛ as the vector of random error terms. Thus, this is a multiple linear regression problem that can be represented as:

$$Y = \beta_0 + \beta_1 X_1 + \dots. + \beta_n X_n\ + \mathcal{E}_1 + \dots. + \mathcal{E}_n \qquad (2)$$

with β_0 as the intercept, and $\beta_1,\dots.,\beta_n$ are the regression coefficients. The error is the difference between the predicted and actual value $\mathcal{E} = \hat{Y} - Y$. If X is a vector containing explanatory variables, and β the vector of unknown regression coefficients, then the coefficients β can be estimated by:

$$\beta^{\wedge} = (X^T X)^{-1} X^T Y \qquad (3)$$

Artificial Neural Networks

ANN was originally designed to mimic the learning process of a human brain. It consists of neurons and weighted connections between them. The human brain is a vastly interconnected network of neurons, where the output of a certain neuron becomes the input for another. The process of learning happens when certain neural connections are activated often, causing a reinforcement process (Thomas, 2017). Mimicking this structure, ANN consists of an artificial neuron (node) with weighted input(s) and an output. This structure is illustrated in Figure 1:

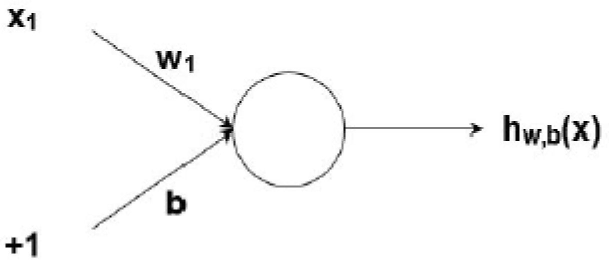


Figure 1. Illustration of a Neuron. Retrieved from: (Thomas, 2017)

where w_1 refers to the weight (a real-valued number) of the connection between an input x1 and the node. A weight is multiplied by the inputs and then summed up inside the node. Sometimes, nodes might include a bias. The node in Figure 1 would then have a weighted input of: x_1w_1 + b. This input is calculated inside the activation function contained per node. The output of the activation function is h. Sometimes, this structure is also called a perceptron in some literature.

In a fully-fledged ANN, several or large amounts of these neurons are connected with each other, forming a large network of nodes. This network would then consist of an input layer, hidden layer(s), and an output layer. The more hidden layers a network has, the more complex its structure becomes. The input layer is where external input data enters the network. The hidden layer just refers to the layer that is not the input nor output. Lastly, the output layer describes the output of the network. Figure 2 shows an example of an ANN with an input, output, and two hidden layers.

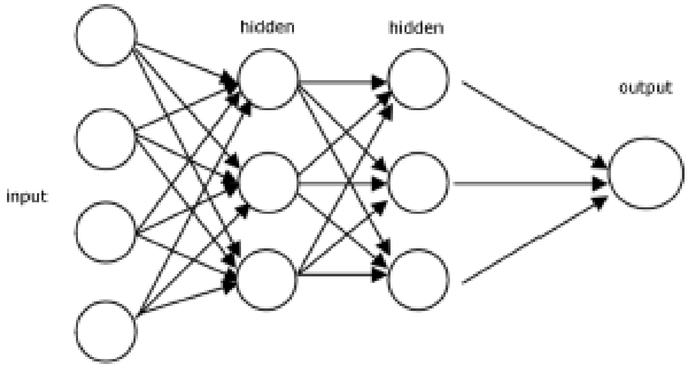


Figure 2. Illustration of a feedforward neural network with two hidden layers. Retrieved from: (Goodfellow, Bengio, & Courville, 2016)

One kind of ANN is the feedforward network, which is a neural network model with no feedback or cycle from the output to the input (Goodfellow et al., 2016). Networks with such feedback cycles are classified as recurrent neural networks (RNN). Considering the prediction nature of the problem in this study, only a feedforward network will be considered as it is best for modeling relationships between a set of inputs and output(s). In a feedforward network, a mathematical function maps combination of input to a certain output. This function is achieved by a mixture of several simpler functions (Goodfellow et al., 2016).

Suppose $f^*(x) = y$ is a function mapping x to y, a feedforward network will define a mapping $y = (x; \Theta)$ and attempt to learn the value of Θ in order to have its output be close to y. If a network is made of several such functions, such as $f^{(1)}$, $f^{(2)}$,$f^{(3)}$ then the feedforward network's function is: $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. In Figure 2, $f^{(1)}$ represents the first hidden layer, $f^{(2)}$ the second hidden layer, and $f^{(3)}$ the output third layer.

A Back-propagation training algorithm is commonly used to train these feedforward networks (Goodfellow et al., 2016). This algorithm calculates the errors between predicted and target output. The errors will then be used to adjust the weights between the neurons, by means of backpropagation to find the most optimal connection between the nodes. During the training/learning phase, these feedforward models often need to learn the dataset more than once. This is quantified as an “epoch”, an epoch is measured after learning the full dataset once (Thomas, 2017). To find the optimal number of epochs, the training should be done for n epochs, and stop when there is no further reduction in errors or no further improvement in the optimization function in a procedure called early stopping.

ANNs have been frequently cited as black box methods. Even if they can approximate any function, attempts to learn from its structure would prove to be futile. There are no connections between the weights of the node and the function that is being approximated. For a given dataset and neural network nodes and layers for example, it is possible to obtain two separate networks with different weights but the same outputs (Sendhil Mullainathan & Jann Spiess, 2017). Machine learning tools are experts on prediction of ŷ. In contrast, economic tools centers around estimation of parameters β*. It should not be assumed that an algorithm built for ŷ, could have properties explaining β* (Sendhil Mullainathan & Jann Spiess, 2017). Therefore, within economics, machine learning is employed in places where improved prediction has large applied value.

Data and Methodology

The dataset used for this research consists of actual property transactions from the NVM collected in the year 2009. It includes 54,000 observations of house transactions in Amsterdam. The houses in the dataset belong within the municipality of Amsterdam. This includes all the houses located in the eight boroughs: Amstedam-Centrum, Amsterdam Noord, Amsterdam Oost, Amsterdam Zuid, Amsterdam West, Amsterdam Nieuw-West, Amsterdam Zuidoost, and Westport. Details of the transactions include house prices and house characteristics, such as area, construction year, number of rooms, and address. The address of the houses is identified with a six-digit post code, house number and street name.

Data Description

The data from the NVM included mostly structural characteristics. Table 1 shows and explains the full list of the explanatory variables. Variables such as size, type, rooms, year of construction, and the house features (garden, carport, parking, etc) are considered structural characteristics.

Variable	*Definition*	*Data Type*
price (in euros)	Transaction Price of the house	Numerical
size (in m^2)	Living area size	Numerical
type-apartment	Indicator if house is apartment	Dummy
type-terraced	Indicator if house is terraced	Dummy
type-semidetached	Indicator if house is semidetached	Dummy
type-detached	Indicator if house is detached	Dummy
canalhouse	Indicator if house is canalhouse	Dummy
rooms	Number of rooms in the house	Numerical
balcony-roofterrace	Indicator if house has a balcony or a roof terrace	Dummy
noparking	Indicator if house does not have parking spots nearby	Dummy

parking	Indicator if house has parking spots nearby	Dummy
carport	Indicator if house has carport	Dummy
garage	Indicator if house has a garage	Dummy
garden	Indicator if house has a garden	Dummy
maintenancegood	Indicator of house maintenance quality, 1 for good, 0 for bad	Dummy
centralheating	Indicator for availability of central heating in the house	Dummy
insulation	Indicator for availability of insulation in the house	Dummy
listed	Indicator if house is listed[1]	Dummy
constryearunknown	Indicator if construction year of the house is unknown	Dummy
constryear_1500-1905	Indicator if the house was constructed between 1500-1905	Dummy
constryear_1906-1930	Indicator if the house was constructed between 1906-1930	Dummy
constryear_1931-1944	Indicator if the house was constructed between 1931-1944	Dummy

[1] Listed houses are houses that are officially recognized as having historical/architectural interest, but is also limited by regulations in the degree of allowable renovation (Cultural Heritage Agency, 2018)

constryear_1945-1959	Indicator if the house was constructed between 1945-1959	Dummy
constryear_1960-1970	Indicator if the house was constructed between 1960-1970	Dummy
constryear_1971-1980	Indicator if the house was constructed between 1971-1980	Dummy
constryear_1981-1990	Indicator if the house was constructed between 1981-1990	Dummy
constryear_1991-2000	Indicator if the house was constructed between 1991-2000	Dummy
constryear_above2001	Indicator if the house was constructed after 2001	Dummy
street	Address of the house	
number	House number	
postcode	The six-digit postcode of the house	

Table 1. The list of explanatory variables in the dataset

Preprocessing

To be useful for predictive modeling, the data must first be cleaned. This is to improve the quality by detecting and removing measurement errors or inconsistencies from the data. Observations not meeting the following inclusion criteria were removed from the data set:

1. Property with living area size of 0 m^2 or less

2. Property with less than 50.000-euro price or above 5,000,000-euro price
3. Property with zero rooms
4. Property with conflicting characteristics: such as a house that is both an apartment and a canal house (variable type-apartment of 1 and canalhouse of 1)

The motivation for such restrictions for house prices and rooms were obtained from observing the lower bound and upper bounds of the house prices and number of rooms. These bounds were observed from funda.nl, an online housing marketplace founded by the NVM (Funda, 2001). Around 3000 observations were filtered out after the filtering process, resulting in a total of 51,728 observations.

This study will focus on house price attributes that explain its characteristics. Thus, the explanatory variables connected with its locational characteristics are omitted from the explanatory variables that are included in the analysis. Moreover, no information about the address could be incorporated into the analysis. These variables include: postcode, number, and street. Upon examination of the variables, both parking and noparking represent the same concept. They indicate whether a house has parking spots, moreover, noparking is the opposite of parking. Whenever parking has a value of 1, noparking has a value of 0. Therefore, to eliminate redundancy, only parking will be used. It is also more intuitive to use parking; the presence of parking can be indicated by 1 and its absence by 0.

Data Descriptives

First of all, Table 2 describes the number of houses belonging to each dummy characteristic. The majority of houses in the municipality of Amsterdam are apartments, with 38,000 houses. They make up around 73.6 % of the total number of houses in the city. While canal houses are very rare, making up only 0.4% of Amsterdam’s houses. Related to the year of construction, the majority of houses in this dataset were constructed between 1906-1930 (around 20% of the total houses). Around 60% of the houses have either balconies or roof-terraces. Finally, almost 90% of the houses have some form of central heating equipped.

House Type	*Number of Houses*
Apartment	38087
Terraced	8061
Semi-detached	4391
Detached	1191
Canal houses	232

Table 2. The Number of houses corresponding to each house type

Table 3 reports the summary statistics of house prices, the living area and number of rooms. The typical house is smaller in Amsterdam (85 m^2) compared to the Dutch average of 117 m^2. An average citizen would be expected to live in a house with 90 m^2 in living area with 3 rooms. The average house price for all types is 280,032 euros in the city. The data shows that the mean price of canal houses are the most expensive, followed by detached, semi-detached, terraced, and finally apartments. Figure 3 reports the boxplot of all the house prices, this shows that the data is riddled with outliers.

	Min	*Max*	*Mean*	*Median*	*S. D*
Price (€)	50,000	5,000,000	280,032	226,500	202,513
Living Area (m^2)	26	500	93.58	85	41.990

Number of rooms	1	20	3.46	3	1.371
Apartment Price (€)	50,000	5,000,000	247,264	203,500	155,465
Terraced (€)	61,500	4,715,000	340,491	277,000	249,639
Semi-detached (€)	75,000	4,000,000	367,339	307,500	249,693
Detached (€)	50,000	4,500,000	596,836	500,000	411,607
Canal House (€)	169,500	2,500,000	822,812	717,500	447,356

Table 3. Summary statistics for house price, living area (in m^2), number of rooms, and price per house type

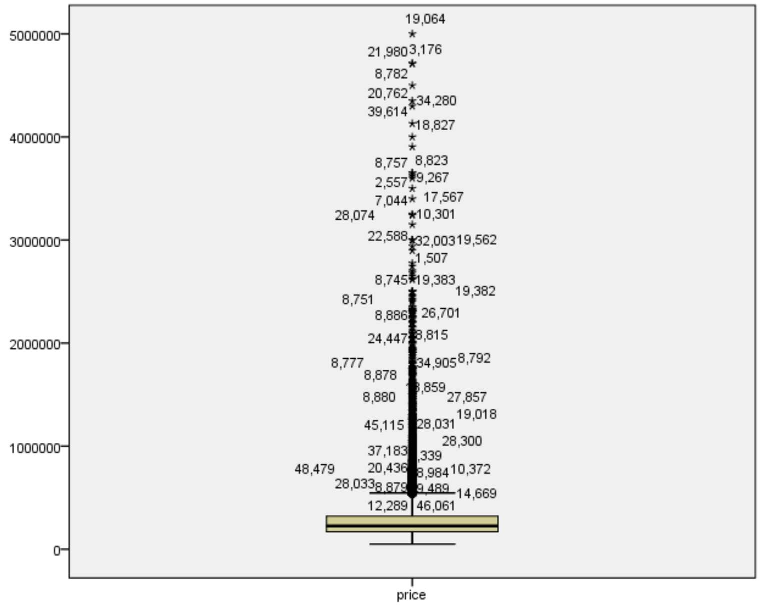


Figure 3. Boxplot of the house prices.

Data Transformation

The data was also investigated to check whether the assumptions needed for linear regression were fulfilled. This is done in order to validate inference made for the house price model. Some of these assumptions will be tested with visualizations. On basis of the VIF (Variance

Inflation Factor), no indication of multicollinearity was found. Therefore, no other independent variable (IV) was required to be removed from the analysis. Further transformations of the variables were also conducted in order to fulfill some of these assumptions.

Linear regression assumptions.

Independent observations.

This assumption is satisfied because the occurrence of one observation does not explain the occurrence of another. Each house in the data is different from one another.

Normality of the residuals.

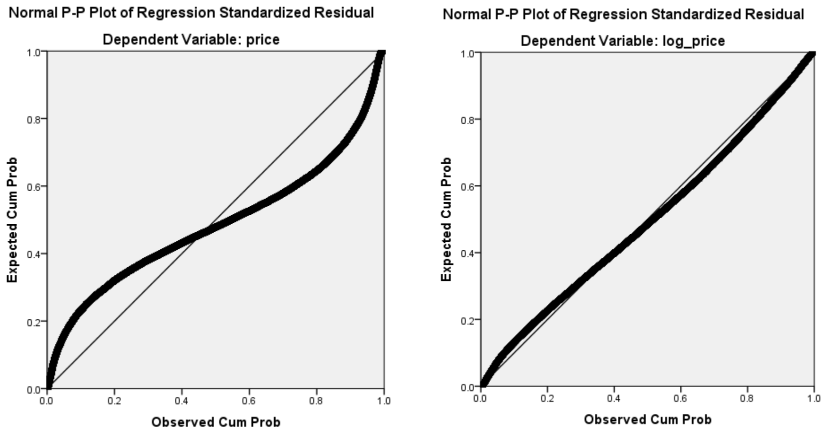


Figure 4. The normal plot before (left) and after transformation (right)

The residuals of the house prices have to be normally distributed. This assumption will be checked by examining the P-P plot. Initially, the assumption was not satisfied. but, after a log transformation of the house price, the points follow the line more closely (see Figure 4). Therefore, this assumption seems to have been verified.

Homoscedasticity (constant variance of the residuals).

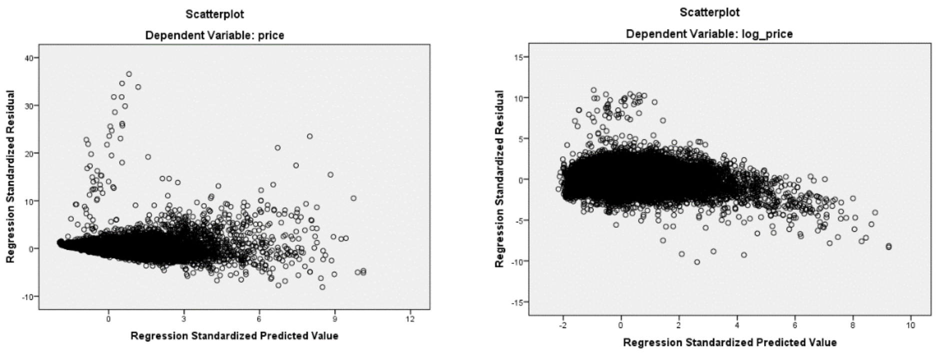


Figure 5. The residual plot before and after data transformation

Similar to the second assumption, the assumption is somewhat met after the data is transformed. Figure 5 describes the process of checking for heteroscedasticity before and after transformation of the variable. The second residual plot has a random scatter over the predicted values, and they fit more within a strip with constant width. In contrast, the first graph is clearly heteroscedastic, with values scattered largely on the left and right part of the strip.

Linearity between the dependent variable (DV) with each of the IVs.

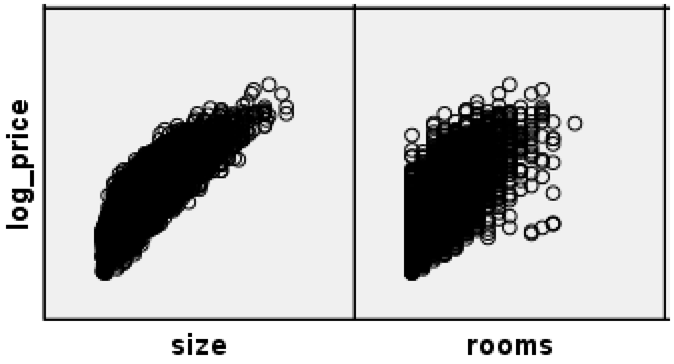


Figure 6. The relationship between log price, size and rooms

Figure 6 visualizes the relationship between the log-price, size of the house, and number of rooms. From this we can infer that this assumption is fulfilled. The rest of the IVs are dummies, which have fulfilled the linearity assumption, by definition.

No outliers or influential points.

From the residual graph of the log-transformed price, we can see that outliers might still be present. An observation will be considered an outlier if it meets one of these criteria:

1) Cook's distance larger than or equal to 1.

2) The absolute value of its standardized residuals is bigger than 3.

It is found that the maximum Cook's distance in the data is 0.032, therefore no cases were excluded based on the first criteria. However, around 429 points were considered outliers based on the standardized residuals, the absolute value of these residuals ranged from 3 to 11. Therefore, these observations were removed considering that they comprise a small portion of only 0.8% of the data. After the removal of these outliers, the residual graph seems to fit more within a band of certain width. On top of that, the P-P plot follows closer to the line as shown in Figure 7.

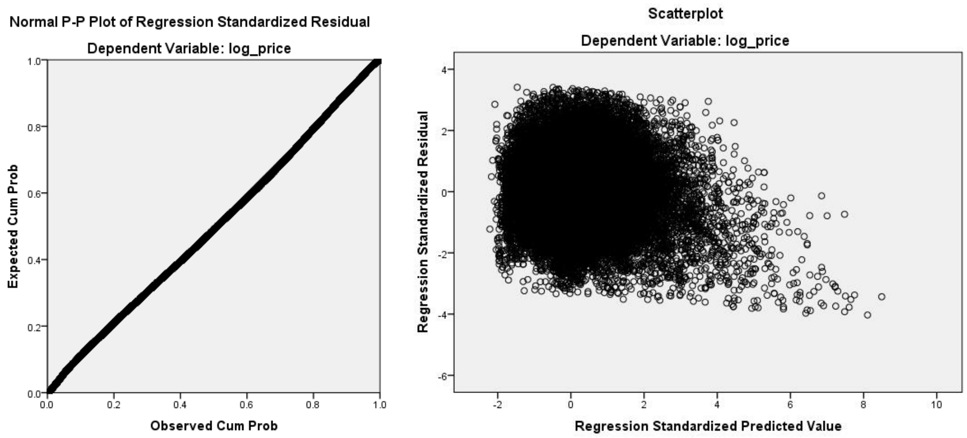


Figure 7. Further improvement in residual and P-P plot after removal of outliers

For this thesis, SPSS (version 24) is used extensively for data processing and visualization. SPSS is selected because of the user-friendly interface. STATA (version 15.1) is used to implement the multiple linear regression model, STATA is selected as it eases the interpretation of results with easily reproducible commands. Python could be utilized for the regression; however, it is less specialized for statistical analysis, and its results are not easily interpretable. Implementation of a neural network requires a programming language, which is why Python was used The ANN is implemented using Python version 3.7 with TensorFlow as its backend. Within Python, the first library used is Keras, an open source neural network library. Keras is selected for its user-friendliness and extensive documentation of its functions. Besides Keras, sci-kit learn is another library utilized to aid implementation of the neural network and compute performance metrics. Finally, the last package imported was matplotlib, which was mainly used to plot the results.

Experiment Environment

Experiment was conducted on a PC with the specifications listed in Table 4.

Operating System	Windows 10 Home ver. 1803
CPU	Intel® Core™ i7-7700HQ
GPU	NVidia GTX 1050 – 2GB GDDR5
RAM	8 GB

Table 4. Testing Hardware Specifications

Models, Procedures and Hypothesis

Hedonic Price Model.

The hedonic model includes regressing the house price against the given attributes which are expected to be the determinants of the price. In the literature, finding an appropriate functional form to analyze housing prices has been a frequent concern. This concern arises due to inadequate guidance from economic theory about the proper functional relationship between housing prices and its attributes (Selim, 2009). Selim found that the most common recommended form is the semi-logarithmic form, which is preferred by researchers because the coefficient estimates from the model are highly interpretable as the proportion of a good's price that is attributable directly to the characteristics of this good. Therefore, the semi-logarithmic form is selected, also it fits more with the data (see Data transformation section above). In practice, this hedonic function is estimated by a Linear Regression model: Log (Price) = f* (X, Ɛ).

Where X is the vector of characteristics defining the house, Ɛ describes the error term, and Price is the price of the house. Because this study examines the effect of house characteristics on the price, the vector X is constructed from the following:

-House size (in m^2)

-Number of rooms

-House type (apartment, detached, semi-detached, terraced, canal house)

-Construction year (in the form of its dummies)

-Other features (Balcony-roofterrace, Parking, Carport, Garage, Maintenance, Garden, Insulation, Central-heating, Listed).

Due to the nature of house type being a dummy, apartment house types will be selected as the reference (type-apartment = 1) in accordance with the best practice of selecting the most common type (dummy with most number of observations) as the dummy. In addition, the reference for year of construction will be construction year between 1906-1930. This is because for this dataset, most houses were constructed between that time.

Based on previous literature, the hypothesis is that variables such as the house size, number of rooms, and other features will have a positive relationship with the price in the model. (Debrezion et al., 2011; Visser et al., 2008). Thus, houses that have more facilities are expected to cost more. With regards to year of construction, the older the construction year, the larger the expected coefficient will be (Visser et al., 2008). In other words, prices are higher the older the house.

Visser et al. (2008) also found that prices are also higher compared to the base year of 1970, when the house is constructed after 1990. Debrezion et al. (2011) did not find this however, they reported a linear negative relationship between construction year and price in the form of negative relationship between house age and price. Considering this study's dataset was closer to the year of Visser et al.'s study, their results will be used as the expectations or the hypothesis.

Finally, canal houses are expected to have the highest house prices in the model compared to the rest. This is observed in the mean of the dataset shown in Table 3. Canal house's coefficient is expected to be negative and its absolute the largest compared to the reference category of apartment. Detached and semi-detached houses are expected to be more expensive than terraced houses (Visser et al., 2008). Table 3 shows that apartment houses might be estimated to cost the least in the model based on its mean house prices.

To achieve the research objective of explaining the determinants of house prices in Amsterdam, the first set of hypotheses for the hedonic model is formulated as follows:

H1: House size, number of rooms, and other features show a positive relationship with the house price.

H2: Prices are higher the older the construction year, but is also higher if constructed after 1990.

H3: Canal houses will be the most expensive, followed by detached, semi-detached, terraced and apartment.

Neural Networks.

The neural network is firstly trained from a set of data. An output will be estimated from a set of particular inputs: in this case the vector X of house characteristics. The feedforward network is applied in this study. As explained in the theory, this type of ANN has three separate layers: the input layer, the hidden layer(s), and the output layer. The value to be estimated in the output layer will be the log of house prices. Performance of this model will be influenced by the number of hidden layers and the number of nodes each hidden layer has. Moreover, the value of weights and number of epochs are also factors to be optimized in the training phase.

There has been little theory to assist in obtaining the optimal number of nodes and hidden layers (Limsombunc et al., 2004). Most practitioners utilize a trial and error procedure to find the optimal model by comparing the performance of different models, each with different hidden layers and nodes (Selim, 2009). Based on Selim, this study will firstly employ 20 nodes and two hidden layers, and subsequently added more nodes/layers until no performance increase is found. Regarding the number of epochs, training will be done for initially 100 epochs, and stop at the epoch with no further reduction in the RMSE. For each epoch, the last saved best weights are used and updated accordingly by backpropagation algorithm. After this training, the best weights for the model will be selected. After obtaining the neural network, its performance will be measured and compared to the baseline model with metrics explained in the following sub-section.

Performance metrics.

To gauge how well these models will perform for predicting house prices in Amsterdam, their performance will be gauged by utilizing several metrics. For model comparison, there are several possible metrics:

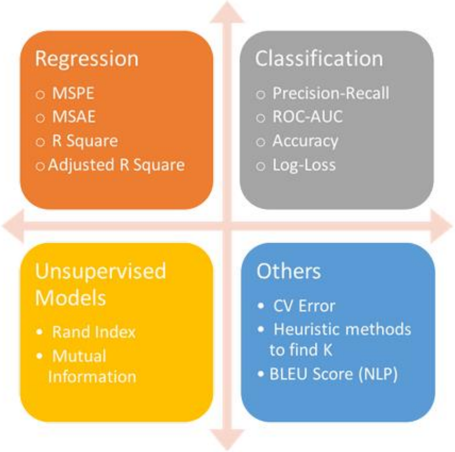


Figure 8. The different metrics for model comparison. Retrieved from: (Swalin, 2018)

Considering the scope of this study is for regressing prices, RMSE (Root Mean Square Error), R-squared, and MAE (Mean Absolute Error) are considered. These metrics were selected as they were the most common and reliable metric utilized to cross-examine the performance of different models (Limsombunc et al., 2004; Selim, 2009). R^2 will be used as a goodness-of-fit measure and RMSE & MAE will be utilized as an accuracy measure.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{SSE}{SST} \qquad (4)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Yi - \hat{Y}_i)^2} \qquad (5)$$

$$\text{MAE} = \frac{1}{N} \sum_{j=1}^{n} |yj - \hat{y}j| \qquad (6)$$

Equation 4 describes the measurement for R-squared. Y_i is the actual house price, n the total number of cases, $\hat{Y}_i$ the estimated price, and $\bar{Y}$ is the mean house price. The SST (Sum of Squares Total) quantifies how much Yi varies around their mean and the SSE (Sum of Squares Error) is a measure of how much Yi varies around the estimated regression line. The R-square measures the proportion of the total variability in Y which is explained by the IVs. It is measured in values between 0 and 1, and the closer it is to 1, the better the IVs are at explaining the dependent variable. Equation 5 describes the RMSE (Root Mean Square Error) formula. This calculates the sample standard deviation of the residuals (differences between observed and predicted values). A model with a lower RMSE score means it is a relatively better model. The MAE (Mean Absolute Error) describes the average of the absolute difference between the predicted and observed values. The MAE has all the individual differences weighted equally in the average.

As stated in the literature review above, predictive performance might vary. Drawing from Nguyen & Cripps (2001), and the fact that the dataset is relatively large with 51,728 observations; Neural Networks are expected to perform comparatively better than the hedonic model in this case. Therefore, it is expected that there will be a lower MAE and RMSE for the ANN, and better model fit, i.e. larger R-squared estimate. The hypothesis is formulated as follows:

H4: The ANN will have lower MAE and RMSE, and larger R-squared estimate value.

Results

This section outlines the results obtained by both regression analysis and neural networks. Firstly, Table 5 describes the goodness of fit tests for the hedonic model. Table 6 describes the hedonic regression model coefficient results. As explained in the Data & Methodology section, there were no indications of multicollinearity based on the VIF and the assumptions to validate inference were satisfied. Table 7 reports the performance metric comparison results. The predicted prices of both hedonic and neural network model are illustrated in Figure 9. Implications of the following results are explained in the Discussion section.

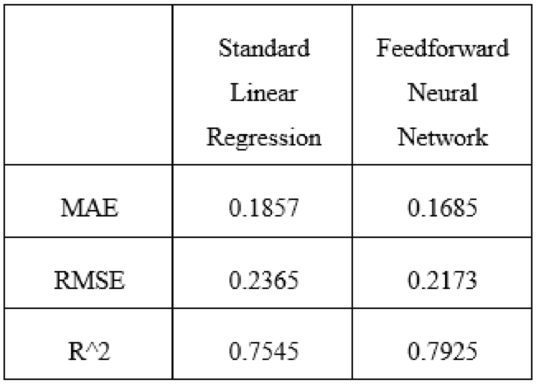
Tests	Value
F-Test (Prob >F)	0.0000
R-squared	0.7545
Adjusted-Rsquared	0.7544

Table 5. The F-test, R-squared, and adjusted R-squared of the hedonic model

log_price	**Coef.**	**Std. Err.**	**t**	**P>t**
size	0.009119	0.0000477	191.20	0.000
typeterraced	-0.17753	0.0041544	-42.73	0.000
typesemidetached	-0.1731	0.0049174	-35.20	0.000
typedetached	-0.05365	0.0080703	-6.65	0.000
canalhouse	0.100132	0.0170208	5.88	0.000
rooms	0.01326	0.0013739	9.65	0.000
balconyroofterrace	-0.00015	0.0024164	-0.06	0.950
parking	0.03022	0.0053158	5.68	0.000
carport	0.048898	0.0071242	6.86	0.000
garage	0.08961	0.0064637	13.86	0.000
garden	0.068034	0.0031523	21.58	0.000

maintenancegood	0.103317	0.0027921	37.00	0.000
centralheating	0.084294	0.0035635	23.66	0.000
insulation	-0.01211	0.0024736	-4.90	0.000
listed	0.115398	0.007817	14.76	0.000
constryearunknown	-0.02975	0.0287218	-1.04	0.300
constryear_15001905	0.114844	0.0039703	28.93	0.000
constryear_19311944	-0.07327	0.0043362	-16.90	0.000
constryear_19451959	-0.18921	0.0048873	-38.71	0.000
constryear_19601970	-0.3376	0.0035884	-94.08	0.000
constryear_19711980	-0.34244	0.0049651	-68.97	0.000
constryear_19811990	-0.25617	0.0038244	-66.98	0.000
constryear_19912000	-0.09462	0.003904	-24.24	0.000
constryear_above2001	-0.08553	0.0054221	-15.77	0.000
_cons	11.5432	0.0047723	2418.77	0.000

Table 6. Hedonic model results, with typeapartment and constryear_19061930 as the reference



	Standard Linear Regression	Feedforward Neural Network
MAE	0.1857	0.1685
RMSE	0.2365	0.2173
R^2	0.7545	0.7925

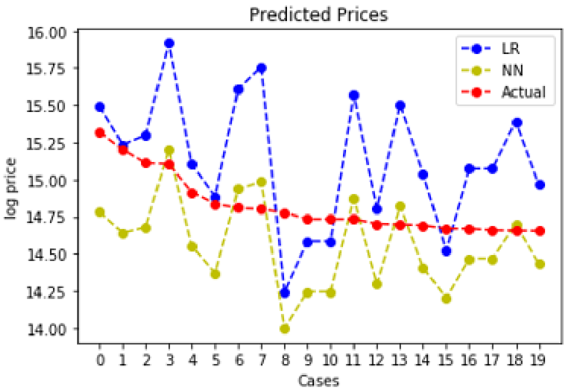


*Table 7.*Metric Comparison

*Figure 9.*Predicted Prices obtained

Discussion

F-test and Goodness of Fit of the Regression

Table 5 reports the R-squared, F-test, and adjusted R-squared of the multivariate linear regression model. This F-test tests the null hypothesis that all coefficients of the IVs are equal to zero. The result indicate that this null hypothesis is rejected with very high confidence. An R^2 of 0.75 shows that this model explains around 75% of the variance, which indicates that this is a very good fit. Since this model is a log-level multivariate regression, when interpreting a certain variable's coefficient, the other IVs are assumed constant.

Effect of Size, Number of Rooms, and Other Features on Price

The results reported in Table 6 show that many house features have positive coefficients. This is in line with the expectations from (Visser et al., 2008). Contrary to this however, two house attributes, insulation and balcony/roof terrace show negative coefficients. With that said, the coefficient size of balcony/roof terrace seem to be miniscule. Moreover, with a p-value of 0.950 there is evidence to conclude that balcony/roof terrace is not statistically significant at the alpha level of 0.05. The size of the living area has a positive relation with the log price of the house.

Results in Table 6 indicate that per 1 m^2 increase in living area, there is an expected increase in house prices by 0.9%. If there are two exact houses, the addition of one more room in one of the houses would lead to a house price increase of 1.3%. Having parking facilities nearby could add the house price by 3% but having a personal garage/carport would increase it even further. For example, having a carport would on average increase the house price by 4.89% while a garage increases it by 8.97%. Houses with gardens are indeed more attractive, on average they cost 6.8% more than houses that lack it.

Both good maintenance and being listed are indicators of a more attractive house. Having good maintenance increases the house price by 10.3%. Listed houses are granted certain privileges but are also restricted in the amount of renovation, but this does not make them less attractive. Instead, they are expected to cost 11.54% more than the average house. Having central heating would increase the house price by 8.43%, but a house with insulation is 1.2% less expensive on average. Since most of these structural attributes show a positive relation with the house price, this finding corroborates with Visser et al. (2008) and Debrezion, Pels, and Rietveld (2011).

The negative coefficient from insulation is also found in Debrezion et al., who included more structural characteristics in their hedonic model compared to Visser et al. However, they did not explain the reason for this finding. Thus, houses in Amsterdam which has insulation are expected to cost less.

H1: House size, number of rooms, and other features show a positive relationship with the house price.

With regards to the first hypothesis, two factors show a negative relationship with the house price: balcony/roof terrace and insulation. Balcony/roof terrace was found to be statistically insignificant at the 0.05 level. However, insulation is still a significant predictor. Based on these results, the first hypothesis cannot be confirmed.

Effect of House Type on Price

Compared to apartments, terraced houses in Amsterdam cost 17.753 % less on average. This is followed by semi-detached houses which cost 17.3% less. Detached houses costs 5.35% less and canal houses are 10% more expensive than apartments. This is in line with the situation in Amsterdam, as canal houses are often considered to be the most luxurious. In Visser et al. (2008), the type of houses also show this trend, with detached houses having the highest prices on average. This trend is continued by semi-detached, and lastly, terraced houses are the least expensive. However, Visser et al. did not include canal houses and apartments in their model.

H3: Canal houses will be the most expensive, followed by detached, semi-detached, terraced and apartments.

Contrary to expectations, the estimated hedonic model coefficients are all negative for the other house types besides canal house. However, some order is still present as detached houses are more expensive than semi-detached, and terraced houses less expensive compared to the other two. In conclusion, this hypothesis cannot be concluded.

Effect of Construction Year on Price

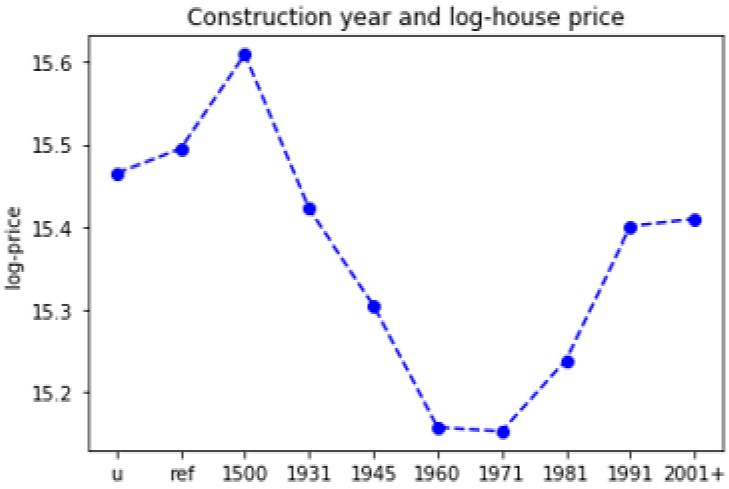


Figure 10. The expected house price assuming all other variables constant, with u = unknown construction year and ref = constructed between 1906-1930.

In relation to construction year, houses with unknown date of establishment costs 2.9% less on average in comparison to the reference of between 1906 and 1930. However, this characteristic is not significant at the 0.05 level. Houses built before the 20th century until 1905 were more valuable, with a higher price of 11.5% on average. After this sharp increase, the relation between house price and construction year starts to decline, with those built between 1931-1944 costing 7.3% less in contrast to the reference. In the period of 1945-1959, houses cost almost 19% less on average compared to the reference. This declining trend continues with houses between 1960-1970 and between 1971-1980 costing around 34% less on average than the reference construction year. The expected house price is the lowest when it is constructed between these two periods.

In the period between 1981 and 1990, prices started to get more expensive, with only a 25.6% decrease in price compared to the reference. Houses built near the beginning of the 21st century are considerably more expensive, with only 9% less price on average compared to the reference. This relationship between construction year and price is illustrated in Figure 10, where the house prices closely follow a U-curve, with older houses being more expensive until the beginning of the 21st century.

Such behavior is also observed in Visser et al. (2008), who used 1970 as the reference category and found that prices were higher the older the house but rose sharply when constructed in the 1990s and above. This suggests that Amsterdam residents find older and classical houses or newer houses built after the 1981 or the 1990s to have more value compared to houses in between the years 1960-1970s. However, this thesis' findings contradict those of Debrezion, Pels, & Rietveld (2011). They found a direct and negative relationship between age of the house and price.

H2: Prices are higher the older the construction year, but is also higher if constructed after 1990

Based on Figure 10, the older the house construction year, the higher the price. Additionally, the data also shows that the price hiked rose if constructed after 1981, but a sharp increase was observed if the construction was after 1991.

Neural Network: Prediction Performance Improvement

The second research objective was to improve price prediction, to achieve this, a neural network model was implemented and fit to the data. Training started with 20 nodes per layer with two hidden layers. The number of nodes and layers were varied until three hidden layers with 40, 70, and 100 nodes are found to be the most optimal fit for the model. Training was intended to run until 100 epochs, but no further improvements were found after 11 epochs based on early stopping.

H4: The ANN will have lower MAE and RMSE, and higher R-squared estimate value.

As reported in Table 6, the feedforward network performed better with a lower MAE score of 0.1685 versus the 0.1857 score of the baseline regression. In terms of RMSE, the ANN obtained 0.2173 which is lower than the 0.2365 score of the regression model. Overall, the ANN obtained 10% improved RMSE and MAE scores in comparison with the hedonic model. In terms of model fit, the neural network model could explain an extra 3.8% variance in the data, with an R^2 of 0.7925 compared to the 0.7545 of the standard regression. In conclusion, hypothesis 4 can be concluded.

Figure 9 describes the line fit of the ANN and regression model. As seen from the graph, the deviation from the original price in the case of ANN seems smaller. On the other hand, the deviation in the regression model is considerably larger. Considering these improvements, the results corroborate with Limsombunc, Gan, & Lee (2004), Nguyen & Cripps (2001), and Selim (2009). They found improvements from the NN model in comparison to standard regression models. This paper's results contradict those of Kontrimas & Verikas (2011), Mach (2017) and Worzala, Lenk, & Silva (1995).

The size of the observations is a contributing factor for poorer relative performance of an ANN over a regression model. As stated in Nguyen & Cripps (2001), neural network starts to outperform regressions when the data size increases. The small dataset is an issue for example, in Worzala, Lenk, and Silva (1995) and Kontrimas & Verikas (2011). Worzala et al. compared predictive performance of ANN and hedonic regression on only 288 observations. While Kontrimas and Verikas did so for 200 samples of houses.

As expected in the hypothesis, the performance of ANN is better compared to the regression model. Feng and Jones (2015) highlighted the need for more larger data experimentation of neural networks in the housing market. Within this context, this paper provides more data to be compared regarding empirical analysis of an ANN on relatively large sample size of houses. Results indicate that AVM providers or real estate agents could benefit from predicting house prices more accurately by utilizing neural networks in their AVM models.

Limitations

One limitation is the date of the dataset. Due to reorganizations from the data provider, requests to obtain more recent data cannot be processed. The available housing data is the one provided by the NVM containing transactions in 2009. Therefore, the results might not reflect recent conditions from the Amsterdam housing market. Results should be understood in the context of the 2009 Amsterdam housing market. Another limitation is the lack of neighborhood, location or environmental characteristics in the analysis. There are locational/neighborhood characteristics, in the form of addresses, but they cannot be incorporated into the analysis.

Therefore, this paper's findings should be interpreted with the assumption that the whole Amsterdam is one neighborhood, with the main difference in the houses being their structural characteristics.

Conclusion

This study has firstly utilized hedonic regression for prediction and interpretation in the Amsterdam housing market context. The hedonic model using multivariate regression was initially constructed on the 2009 Amsterdam housing market. Results from the hedonic analysis are reported and analyzed in relation with the Dutch hedonic price literature. This paper found that many but not all house features have a positive relationship with the expected price. This is confirmed by previous work. Another finding is that there is a U-shaped relationship between year of construction and expected house price, which is corroborated in a past study. Lastly, the model suggests that expected prices of canal houses would be the highest, followed by apartments, detached, semi-detached, and terraced houses.

To see whether prediction performance can be enhanced, a neural network model was trained from the data. The performance of the ANN and hedonic model was compared using several metrics selected based on existing literature. ANN performed better in comparison to the baseline regression model. Considering the lack of published literature regarding utilization of neural networks for Dutch housing, this paper could be considered the very few to attempt such a method. Moreover, the results add to the relatively scarce comparison of large-scale housing data using neural network. Real estate agents or consumers can also benefit from having a more accurate prediction model to assess house prices.

Future research should consider utilizing more recent data. This is in order to make sure that the findings can correspond with the latest condition of the housing market. Additionally, other explanatory variables could be implemented in the analysis, such as environmental, neighborhood or location. This study’s scope was situated within the context of Amsterdam’s housing market. Considering the viability of using neural networks to large data, future works could also include other regions in the Netherlands. Finally, as a steadily growing field, other machine learning techniques could be tested and compared as well.

References

CALCASA. (2018). Calcasa - Technology. Retrieved December 3, 2018, from http://www.calcasa.co.uk/technology

Chiarazzo, V., Caggiani, L., Marinelli, M., & Ottomanelli, M. (2014). A neural network based model for real estate price estimation considering environmental quality of property location. *Transportation Research Procedia*, *3*(July), 810–817. https://doi.org/10.1016/j.trpro.2014.10.067

Cultural Heritage Agency. (2018). Listed buildings. Retrieved December 5, 2018, from https://erfgoedmonitor.nl/en/indicators/listed-buildings-numbers

Debrezion, G., Pels, E., & Rietveld, P. (2011). The impact of rail transport on real estate prices: An empirical analysis of the Dutch housing market. *Urban Studies*, *48*(5), 997–1015. https://doi.org/10.1177/0042098010371395

Feng, Y., & Jones, K. (2015). Comparing multilevel modelling and artificial neural networks in house price prediction. *ICSDM 2015 - Proceedings 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, 108–114. https://doi.org/10.1109/ICSDM.2015.7298035

Funda. (2001). About funda. Retrieved November 5, 2018, from https://content.funda.nl/over/funda/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. Retrieved from http://www.deeplearningbook.org/

Kagie, M., & Wezel, M. Van. (2007). Hedonic price models and indices based on boosting applied to the Dutch housing market. *Intelligent Systems in Accounting, Finance & Management*, *15*(3–4), 85–106. https://doi.org/10.1002/isaf.287

Kain, J. F., & Quigley, J. M. (1970). Measuring the Value of Housing Quality. *Journal of the American Statistical Association*, *65*(330), 532–548. Retrieved from https://www.jstor.org/stable/2284565

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing Journal*, *11*(1), 443–448. https://doi.org/10.1016/j.asoc.2009.12.003

Limsombunc, V., Gan, C., & Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *American Journal of Applied Sciences*, *1*(3), 193–201. https://doi.org/10.3844/ajassp.2004.193.201

Mach, Ł. (2017). The Application of Classical and Neural Regression Models for the Valuation of Residential Real Estate. *Folia Oeconomica Stetinensia*, *17*(1). https://doi.org/10.1515/foli-2017-0004

Malpezzi, S., Ozanne, L., & Thibodeau, T. G. (1987). Microeconomic Estimates of Housing Depreciation. *Land Economics*, *63*(4), 372–385. Retrieved from https://www.jstor.org/stable/3146294

Marjan, C., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information*, *7*(5). https://doi.org/10.3390/ijgi705168

Nguyen, N., & Cripps, A. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research*, *22*(3), 313–336. https://doi.org/Journal of Real Estate Research. 22(3): 313–336

Prajapati, P., Patel, N., Macwan, R., Kachhiya, N., & Shah, P. (2014). Comparative Analysis of DES , AES , RSA Encryption Algorithms. *International Journal of Engineering and Management Research*, *4*(1), 132–134.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55. Retrieved from https://www.jstor.org/stable/1830899

Schekkerman, C. (2004). *Nauwkeurigheid in taxaties*. Amsterdam School of Real Estate.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2 PART 2), 2843–2852. https://doi.org/10.1016/j.eswa.2008.01.044

Sendhil Mullainathan, & Jann Spiess. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2—Spring), 87–106. https://doi.org/10.1257/jep.31.2.87

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*((1)), 1–44. https://doi.org/Article

Swalin, A. (2018). Choosing the Right Metric for Evaluating Machine Learning Models — Part 1 Image. Retrieved December 15, 2018, from https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

Thomas, A. (2017). *An introduction to neural networks for beginners*. *Adventures in Machine*

Learning. Retrieved from http://adventuresinmachinelearning.com/ebook_author/dr-andrew-thomas/

van der Burgt, E. J. T. G. (2017). *Data Engineering for house price prediction*. Eindhoven University of Technology.

Visser, P., Van Dam, F., & Hooimeijer, P. (2008). Residential environment and spatial variation in house prices in the Netherlands. *Tijdschrift Voor Economische En Sociale Geografie, 99*(3), 348–360. https://doi.org/10.1111/j.1467-9663.2008.00472.x

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