Lappeenranta-Lahti University of Technology

School of Business and Management

Strategic Finance and Business Analytics

Master's thesis

2020

Residential real estate valuation: review and practical comparison of valuation methods

Elias Sairanen

1st Supervisor: Professor Mikael Collan

2nd Supervisor: Post-Doctoral Researcher, D.Sc. Jyrki Savolainen

ABSTRACT

Author:	Elias Sairanen
Title:	Residential real estate valuation: review and practical comparison of valuation methods
Faculty:	LUT School of Business and Management
Major:	Strategic Finance and Business Analytics
Year:	2020
Master's thesis:	56 pages, 7 figures, 21 tables
Examiners:	Professor, D.Sc. (Econ. & BA) Mikael Collan, Post-Doc. Researcher, D.Sc. (Econ. & BA) Jyrki Savolainen
Keywords:	residential real estate valuation, artificial neural network, multiple regression analysis, hedonic pricing, ANN, MRA, valuation methods

The purpose of this master's thesis is to find out what kind of housing valuation models exist in the previous literature. In addition, the purpose of this study is to compare more closely the two different methods, artificial neural network (ANN) and multiple regression analysis (MRA), and find out which of the methods is more accurate in determining the prices of apartment buildings in the Helsinki area.

The data used in the research has been collected from the Confederation of Finnish Real Estate's database and from the website of the Finnish Tax Administration. The combined data contains information about apartments in Helsinki. In addition to the housing information, the data contains regional income data. The research material of this study was analyzed using an artificial neural network as well as multiple regression analysis.

This research finds that the most often used residential real estate valuation method in the previous literature is the multiple regression method. The second most common is the hedonic pricing and 3rd common is the artificial neural network model. In the analysis section of this study, the artificial neural network model proved to be the most accurate way to estimate the prices of apartment buildings in Helsinki. According to this study, the most important variables influencing the price of an apartment were: number of square meters of the apartment, income level of the area, form of ownership of the plot and the condition of the apartment. In addition, the study found that in some cases the location of the apartment has a significant impact on the price.

TIIVISTELMÄ

Tekijä:	Elias Sairanen
Tutkielman nimi:	Asuinkiinteistöjen arvostaminen: arvostusmenetelmien läpikäynti ja käytännön vertailu
Tiedekunta:	Kauppatieteellinen tiedekunta
Pääaine:	Strateginen Rahoitus ja Bisnes Analytiikka
Vuosi:	2020
Pro Gradu -tutkielma:	56 sivua, 7 kuvaa, 21 taulukkoa
Tarkastajat:	Professor, D.Sc. (Econ. & BA) Mikael Collan, Post-Doc. Researcher, D.Sc. (Econ. & BA) Jyrki Savolainen
Avainsanat:	asuinkiinteistöjen arvottaminen, keinotekoinen neuroverkko, usean muuttujan regressioanalyysi, hedoninen hinnoittelu, ANN, MRA, arvottamismenetelmät

Tämän pro gradu -tutkielman tarkoituksena on selvittää, minkälaisia asuntojen arvonmääritysmalleja aiempi kirjallisuus tuntee. Lisäksi tavoitteena on vertailla tarkemmin kahta erilaista menetelmää, keinotekoista neuroverkkoa (ANN) sekä usean muuttujan regressioanalyysiä (MRA), ja ottaa kantaa siihen, kumpi menetelmistä on tarkempi määriteltäessä Helsingin seudun kerrostaloasuntojen hintoja.

Tutkielmassa käytetty data on kerätty Suomen Kiinteistövälitysalan Keskusliiton (KVKL) sähköisestä tietopalvelusta sekä Suomen Verohallinnon verkkosivuilta. Yhdistetty data sisältää tietoa Helsingin seudun asunnoista ja niiden omaisuuksista sekä alueellisista tulotiedoista. Aineisto analysoitiin hyödyntämällä keinotekoista neuroverkkoa sekä usean muuttujan regressioanalyysiä.

Tutkielman tulokset osoittavat, että aiemmassa kirjallisuudessa käytetyin asunnon arvonmääritysmalli on usean muuttujan regressio. Toiseksi yleisin on hedoninen hinnoittelumalli ja kolmanneksi keinotekoinen neuroverkko -malli. Tämän tutkimuksen aineiston analysoinnissa keinotekoinen neuroverkko -malli osoittautui tarkimmaksi malliksi arvioidessa Helsingin seudun asuntojen hintoja. Tutkielman mukaan Helsingin asuntojen kaikista tärkeimpiä hintaan vaikuttavia muuttujia ovat asuinneliöiden määrä, alueen palkkataso, tontin omistusmuoto sekä asunnon kunto. Lisäksi tutkimuksen mukaan asunnon sijainnilla on joissain tapauksissa merkittävä vaikutus asunnon arvoon.

ACKNOWLEDGEMENTS

Completing this master’s thesis was an intense yet very educative experience. I would like to thank my instructors Mikael Collain and Jyrki Savolainen very much for their vital advices during the research process. In addition to this, I want to thank my family and especially my girlfriend for providing me great support during the university studies. Although I still have some studies ahead, I would also like to thank LUT University in advance for the great opportunity to study one of Finland's most modern master's degrees in Strategic Finance and Analytics.

In Helsinki 16.09.2020

Elias Sairanen

SISÄLLYS

List of figures

List of tables

1 INTRODUCTION

1.1 RESEARCH BACKGROUND

Today, humans have more data at their disposal than ever before in human history. This is due to the large amount of data collected by various devices and increase in the storage capabilities of computers. Thanks to this large amount of data, home valuation models previously based on human judgment have been able to be replaced by fully automatic artificial intelligence models by several institutions.

Several methods of artificial intelligence have been developed to process, classify, and analyze housing data. One of these methods is the Artificial Neural Network (ANN). ANN seeks to evaluate the functional relationship between the input values in the model and the output values obtained from the model. In summary, the ANN model is constructed of neurons and the weighs between the neurons. Multilayer perceptron (MLP) is one of the most common ANN architectures (William, Peadar, Martin, Michael, & David, 2012). This study uses an MLP model with one input layer, one hidden layer, and one output layer. MLP provides nonlinear mapping between input and output vector (GUPTA & SINHA, 2000). This study estimates apartment prices in Helsinki area using an MLP model and MRA and compares their performance with a different error measures such as root mean squared error (RMSE), mean absolute error (MAE) and various accuracy thresholds (%). Literature review of this research reviews several advanced and traditional valuation models of residential properties familiar from the previous literature. The literature review also sorts valuation models according to their actual use in the literature.

This master's thesis is not done for any corporation but is done on a topic of most interest to the author. The author is aware that buying an apartment is possibly one of the biggest financial decision one makes during their lifetime. Buying or selling real estate makes it possible to accumulate or lose large amounts of wealth only as a follow-up to few decisions. This why it is important to learn as much as possible about the principles of housing valuation, usage of data analytics to support housing valuation, forming a good overall picture of the variables affecting housing value and to get acquainted with the most useful real estate valuation models.

1.2 RESEARCH QUESTIONS, FOCUS AND LIMITATIONS

There are three research questions in this thesis.

1. What kind of residential real estate valuation models exist in the previous literature and what are the most frequently used valuation models in the research data collected for this study?
2. Which model is more accurate when predicting apartment values in Helsinki: Simple Artificial Neural Network (ANN) model or multiple regression analysis (MRA)?
3. Which variables are the most significant variables in valuing an apartment based on this study as well as previous literature?

The focus of the study can be divided into two sub-categories based on the three research questions. The first sub-category of this research is to examine the previous literature on different residential real estate valuation models and form an overview of all valuation modeling options and their literary background for the reader. The first research question is answered within this sub-category. The second sub-category examines in more detail and delves into the use of ANN as well as MRA in residential real estate valuation. The ANN methodology in this research focuses only in ANN structure called Multilayer perceptron. This is simple feedforward ANN structure with one input layer one hidden layer and one output layer and a resilient weight backtracking algorithm. The MRA in this research includes three different regression equations: ordinary least squares (OLS) known also as linear regression, double-log regression and semi-log regression. Helsinki housing data is analyzed using ANN and MRA methodologies. The second and third research question is answered during the Helsinki housing data analysis process. In the literature review emerald is scanned for information acquisition purposes. Emerald has been chosen as the only search-string data source for this research. There are a few different reasons for this. The amount of data becomes too large if a lot of different databases are taken to research the topic. Emerald is known to be a reliable source of information, and one of the most important elements of this study is to use reliable and critical sources to examine the research topic. During the backward tracking process, relevant sources have also been added that are not included in Emerald. Other databases used within the backtracking process were Jstor, Scopus, Sciencedirect and ResearchGate. The data has been collected in way that only apartments in Helsinki are included. Many forms of apartment buildings have been removed from the data. Removed data included example, terraced houses, detached houses, semi-detached houses and other forms of housing.

1.3 STRUCTURE OF THIS THESIS

This research is divided into five different sections. The first section is the introduction. The second section discusses the most important theoretical backgrounds for this work, research history, and what is limited outside this research topic and why. The third section deals with the concepts related to the operation of research methods, the process of comparing methods, as well as the advantages and disadvantages of the chosen valuation methods. The fourth section introduces the data to be examined, explains how the data has been pre-processed and aggregated. Explains dummy variables and selection of error meters and finally performs MRA as well as analysis using an MLP model. The fifth section includes the results from this research, implications for the industry, as well as suggestions for future research.

2 LITERATURE REVIEW

The literature review is divided as follows: first chapter explains which data sources are used in this study, what keywords were used to search research articles in the study and why, what were the criteria for selecting particular articles and how the information retrieval process progressed from start to finish. This chapter also explains how the number of articles was distributed to the different stages of the information retrieval process and which articles were ultimately selected for this literature review section. We also go through in tabular form which are the most Frequently used valuation models in the research data collected for this literature review.

The second chapter classifies valuation models based on past literature practices. This section summarizes information about different property valuation models and their literature. This section is divided in two subgroups categorized by Elli et al. (2003) and Abidoye et al. (2019) where the first subgroup focuses on traditional valuation methods and second subgroup focuses on advanced valuation methods. A similar way of distinguishing models can be seen in several articles on real estate valuation theory, so it is in some level established and therefore it can also be used for literature review section of this research.

The third section presents a previous study of the ANN in relation to apartment price estimation. The ANN related to housing valuation theory is discussed in more detail to make it easier to place the results of the ANN model of this study to its frame of reference.

The fourth section explains the concept of hedonic pricing in relation to the valuation of residential real estate, as well as the micro and macro level variables that affect the price of apartment price. A broader review of hedonic pricing will help the reader understand the selection of coefficients for the ANN model as well as the selection preferences for MRA independent variables.

The fifth part of the literature review focuses on research articles comparing MRA and ANN models performance in residential real estate price estimation and finally summarizes previous studies comparing MRA and ANN models. It is very important to study the articles comparing the ANN and MRA models and a summary of the results so that we can place the results of this study in a natural continuum based on the results of the previous literature.

2.1 RESEARCH ARTICLES SEARCHING PROCESS

The article selection process can be seen in figure 1:

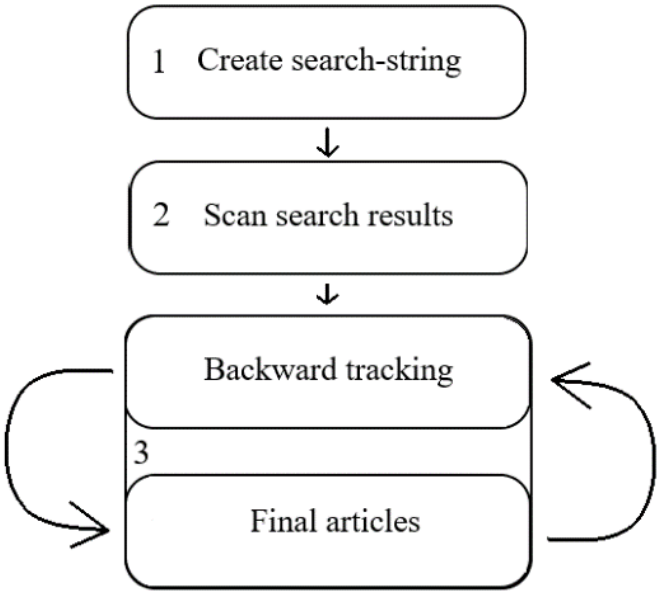

Figure 1. The article selection process in literature review by Webster & Watson (2002)

1. The first step was to create search string which included all the focus areas of the literature review. The search-string was built as follows: *residential real estate valuation and ANN or residential real estate valuation and regression or residential real estate valuation and comparison or residential real estate valuation and methods not Strategic Issues for Facilities Managers not property journals index*. We search data from Emerald database. The search was done only for articles that were accessible for the author and in written in English. In the first step there were 548 articles from Emerald.
2. The second step in the research process was to scan the titles, abstracts, and conclusions of the articles and dropping all the nonrelevant articles. The articles relevancy was decided by observing the title and abstract and looking how many times the keywords appeared in the research overall. When it was noticed that the title as well as the abstract were relevant as well as the keywords selected for the study were repeated often in the text, the text was examined further for double checking the relevancy by eye examining the full article through. All articles associated with topics other than the research questions or the research topic related research history were excluded. There were 14 relevant articles from Emerald selected for the literature review.

3. The third step was the continuum between the backward tracking of articles and final set of selected research papers. Sources that were relevant to address the research questions or background of the study were included in the study. The sources that were found in the backward tracking included sources from databases such as one article from Scopus, nine articles from Research gate, Three articles from Science direct and four articles from Jstor. The backward tracking was done until the data saturation was achieved. The data saturation was achieved once new articles with relevant information could no longer be found. There were total of 17 articles found in the backward tracking process. This means the total sum of articles used in the literature review process was 31 articles. The final set of articles chosen for the literature review is in the table 1 below. The table shows author names, year of publish and document titles. As it can be seen from the table 1 below, the majority of the studies examining residential real estate valuation models are from the 20th century and from the beginning of year 2000. The author believes that the small number of new surveys in the residential real estate area is due to factor that the data is not collected as much and accurately in the house valuation field as in many other more digitalized business areas. This makes it difficult to conduct new studies on housing price patterns.

Table 1. Literature review articles

Authors	**Year**	**Document title**
Abidoye, R. B., Ma, J., Lam Terence Y M, Oyedokun, T. B., & Tipping, M. L.	2019	Property valuation methods in practice: Evidence from australia.
Elli, P., Vassilis, A., Thomas, H., & Nick, F.	2003	Real estate appraisal: A review of valuation methods.
Greenhalgh, P. M., & Soares, B. R.	2015	An investigation of development appraisal methods employed by valuers and appraisers in small and medium sized practices in brazil.
Bruce, T.	1994	The valuation of resort condominium projects and individual units.
Isakson, H.	2002	The linear algebra of the sales comparison approach.
Joshua, A. O.	2014	Critical factors determining rental value of residential property in ibadan metropolis, nigeria.
Wang, D., Li, V. J., & Yu, H.	2020	Mass appraisal modeling of real estate in urban centers by geographically and temporallyweighted regression: A case study of beijing's core area.
Makridakis, S., & Hibon, M.	1997	ARMA models and the Box–Jenkins methodology.
Hajnal, I.	2014	Continuous valuation model for work in process investments with fuzzy logic method
Borst, R. A.	1991	Artificial neural networks: The next modelling/calibration technology for the assessment community.
Lenk Margarita M, Worzala Elaine M, & Ana, S.	1997	High-tech valuation: Should artificial neural networks bypass the human valuer?
Rossini, P.	1997	Application of artificial neural networks to the valuation of residential property.
Stanley, M., Alastair, A., Dylan, M., & David, P.	1998	Neural networks: The prediction of residential values.
Chun, L. C., & Mohan Satish B.	2011	Effectiveness comparison of the residential property mass appraisal methodologies in the USA.

Limsombunchai, V., Gan, C., & Lee, M.	2004	House price prediction: Hedonic price model vs. artificial neural network.
Peterson, S., & Flanagan, A.	2009	Neural network hedonic pricing models in mass real estate appraisal.
Zurada, J., Levitan, A., & Guan, J.	2011	A comparison of regression and artificial intelligence methods in a mass appraisal context.
McCluskey, W. J., McCord, M., Davis, P., Haran, M., & McIlhatton, D.	2013	Prediction accuracy in mass appraisal: A comparison of modern approaches.
William, M., Peadar, D., Martin, H., Michael, M., & David, M.	2012	The potential of artificial neural networks in mass appraisal: The case revisited.
Cowling, K., & Cubbin, J.	1972	Hedonic price indexes for united kingdom cars.
Rosen, S.	1974	Hedonic prices and implicit markets: Product differentiation in pure competition.
Gaetano, L.	2019	Property valuation: The hedonic pricing model – location and housing submarkets.
Bartik, T. J.	1988	Measuring the benefits of amenity improvements in hedonic price models.
Hasanah, A. N., & Yudhistira, M. H.	2018	Landscape view, height preferences and apartment prices: Evidence from major urban areas in indonesia.
Karaganis, A.	2011	Seasonal and spatial hedonic price indices.
Ridker, R. G., & Henning, J. A.	1967	The determinants of residential property values with special reference to air pollution.
Sander, H. A., & Polasky, S.	2009	The value of views and open space: Estimates from a hedonic pricing model for ramsey county, minnesota, USA.
Thanasi (Boçe) Marsela.	2016	Hedonic appraisal of apartments in tirana.
Warren Clive M J, Peter, E., & Jason, S.	2017	The impacts of historic districts on residential property land values in australia.
Do, A. Q., & Grudnitski, G.	1992	A neural network approach to residential property appraisal.
Amri, S., & Tularam, G. A.	2012	Performance of multiple linear regression and nonlinear neural networks and fuzzy logic techniques in modelling house prices.

Table 2 shows how many times specific valuation method had been used in the research articles of this study. The data table is as follows:

Table 2.Frequency of Method usage in literature review's literature

Valuation method	**frequency of use in the literature**
Multiple regression	18
Hedonic pricing	12
ANN	10
Spatial analysis	9
Geographical information systems	6
Comparable/comparative	4
stepwise regression	2
Fuzzy logic	2
Income	1
ARIMA	1
Profits	0

As it can be seen in table 2. above, the most frequently used valuation method in the literature selected for this literature review is Multiple regression. Other most popular valuation methods in the selected

literature were Hedonic pricing, ANN, and Spatial analysis. The least used methods are Profits and the ARIMA.

2.2 PRESENTATION OF RESIDENTIAL REAL ESTATE VALUATION MODELS

Elli et al. (2003) and Abidoye et al. (2019) divided real estate valuation methods into both traditional and advanced property valuation methods. In this literature review, the valuation models were distributed in a similar way, utilizing the classification which they used in their researches. (Abidoye, Ma, Lam Terence Y M, Oyedokun, & Tipping, 2019; Elli, Vassilis, Thomas, & Nick, 2003)

2.2.1 Traditional valuation methods

Traditional methods are very often used for valuation real estate. In the previous literature traditional methods included Comparable method, Income method, Multiple regression method, Stepwise regression method, Profit method, Development method and Contractor's method (Abidoye et al., 2019; Elli et al., 2003)

Direct comparison also known as comparative method is often preferred among the operators at the residential industry to value residential real estate (Greenhalgh & Soares, 2015). The method seeks to use comparable properties recently sold to assign value for the target real estate, however, considering the differences between the target real estate and the comparable real estates (Bruce, 1994). The process functions in two steps: First the valuer leads intermediate prices from comparative real estates. After the intermediate prices have been led from the comparative real estates, all these intermediate values are converted to one final value of the target real estate (Isakson, 2002).

The income rate method includes few different methods. This method can construct the income flows to reflect the valuation of the home by using overall capitalization rate. This rate is constructed by multiplying the annual return of the property by a multiplier that expresses future returns. This method can also use the discount rate. This means discounting the property's future cash flows to the present. (Elli et al., 2003).

Development, contractor's method and profit method are used to value when it is desired to value uninhabited properties. (Elli et al., 2003). Therefore, we will not go into these methods in more detail during this research.

2.2.2 Advanced valuation methods

The advanced methods include methods such as: hedonic pricing, autoregressive integrated moving average (ARIMA), fuzzy logic, artificial neural network (ANN) and spatial methods. (Abidoye et al., 2019; Elli et al., 2003)

Spatial methods consider the effect of variability in spatial variables on the valuation. These variables can be, for example, accessibility, transportation, quality of living area or infrastructure. (Joshua, 2014)

The spatial analysis example includes methods such as geographically weighted regression (GWR) which considers socio-demographic and environmental factors in the assessment of house prices. The GWR model assumes that the independent variables affect the dependent variable differently at different points in the analyzed district unlike in linear OLS regression, which assumes that the formation of a dependent variable is always constant in every geographical district. These geographical differences that affect the formation of the dependent variable are considered by assigning weights to individual observations depending on the location of the observation. The mixed geographically weighted regression (MGWR) model differs from the GWR model in the way that some of the coefficients considered as static and others as non-static. Non-static coefficients are assigned with several different weights but static coefficients are not. GWR and MGRW models have the same limitations as OLS. This means the data will have to meet the same data quality requirements OLS does. This assumptions concern example multicollinearity and data linearity. Regressions models are seen also to be very sensitive to effects of outlier values and for poor quality data. (Wang, Li, & Yu, 2020)

Yuly (1926) first introduced the autoregressive models. Moving average were introduced by Slutsky in 1937. Wold (1938) combined autoregressive models with moving average concept. After that the ARIMA method which refers to box-jenkins method, was invented. The box-jenkins method integrated the autoregressive method and moving average concept together and formed an autoregressive integrated moving average method. This methodology became well known in the 1970s. The Box-Jenkins way has since gathered both opponents and supporters as some researchers believe that the method is not an accurate way to measure economic functions and some believe that it is very good form of modeling time series data. (Makridakis & Hibon, 1997)

The concepts of fuzzy logic were first introduced by Zadeh (1965) as he first introduced fuzzy set theory. In basic logic the certain elements are binary and either true as 1 or false as 0. The basic idea of fuzzy logic is that one certain element can belong to a set at different levels. Membership closer to

0 indicates that the element is not similar with the member group. Membership closer to 1 points out that the observed element is similar towards certain fuzzy set's members. Often the fuzzy set limit values are set to 1 and 0. The variable degrees will always be placed within the set constraint values. Due to nature of decision-making in house valuation it is not often possible to affirm with certainty that element is either true or false. Because of this, fuzzy logic is often a more realistic way to describe reality in many situations as fuzzy logic never has to make this assumption. (Hajnal, 2014) However, the use of fuzzy logic today is very limited in the field of residential real estate valuation as the phenomenon is still very new. However, there are few articles where theory has been used in housing valuation (Amri & Tularam, 2012).

2.3 ARTIFICIAL NEURAL NETWORK AND VALUATION OF REAL ESTATE

Borst (1991) was the first to study the use of an artificial neural network in real estate valuation. Lenk et al. (1997) modeled ANN using a sample of 288 observations in their study and found that the model results in significant estimation errors and in order to construct an optimal neural network model, the execution time of the model highly increases. At the time of the research of Rossini (1997), the ANN method was found to require too high computer power for forming an ANN model quickly enough to be used in valuation practices. He also found that the performance of the ANN model varies noticeably within different iterations. Rossini (1997) also points out that ANN modeling is still of a black-box type, as not near all values nor functions that end up on the final model can be explained perfectly.

Stanley et al. (1998) found in their research that the use of neural networking results in only 80% of predictions going less than 15% of the variance from the desired prediction result. However, they also found that predictive accuracy can be improved by using more homogeneous data and removing outliers from the data in the same way as when using, for example, regression in predicting apartment values.

Outliers refer to a single observation that differs considerably in characteristics from other observations (Lenk Margarita M et al., 1997). Stanley et al (1998) also reinforced the argument of ANN model being black-box -model and giving varying outcomes when applied.

Limsombunchai et al. (2004) modeled ANN using data from 200 different apartments in Christchurch New Zealand and found that ANN is good for finding recurring formulas hidden in data. They also found that the construction of the ANN model should be done with trial and error tactics in order to make the model optimal. Peterson and Flanagan (2009) investigated data sample of 46 000 properties with ANN model and found that ANN is capable of modeling complex nonlinearities within the

dataset. The ANN was seen to find new increasingly in-depth insights from the data as the amount of training data is increased from the previous. Zurada et al. (2011) investigated housing sample of 16 366 observations and found that neural network methods perform better compared to other methodologies when utilized in heterogeneous data. They also found that the connection weights between neurons are difficult to interpret. In the year of 2011, it also emerged that the use of ANN is a very cost-effective and reliable way to valuate large numbers of apartments as Chun and Satish (2011) investigated the housing sample consisting of 33 342 apartments.

McCluskey et al. (2012) found in their research that the ANN's black-box like characteristic prevents using the model to forecast residential real estate prices reliably. The author also stated that the ANN and MRA should be combined for hybrid model to get the best performance of both methodologies. MCcluskey et al. (2013) investigated dataset with 2694 observations and found that the non-linear regression model had higher predictive accuracy than ANN. They also found that it is difficult to make defensive arguments in favor of the results of the ANN because the model is black-box type, meaning it is not possible to explain exactly how the results are generated in the model. MCcluskey et al. (2013) also found that ANN's predictive capabilities are good despite its black-box type nature.

2.4 HEDONIC PRICING AND RESIDENTIAL REAL ESTATE VALUATION

First reference to the hedonic price analysis was created on Court (1939) Hedonic Price Indices - With Automotive Examples: The Dynamics of Automobile Demand. (Cowling & Cubbin, 1972)

Rosen (1974) created the theoretical background of hedonic modeling. Rosen's hedonic price theory work in such a way that the products are a package of different utilities bearing attributes and characteristics. These characteristics and utilities bearing attributes can be stored into different vectors. Implicit prices for these characteristics and attributes can be defined by step regression analysis by regressing a dependent variable with the vectors of attributes and characteristics. Once this is done, one can observe what is the individual characteristics price.

However, these characteristics cannot be differentiated from the model as individual, as the value of their impact towards the overall price can be valuated only indirectly as package. In hedonic pricing theory, different characteristics can be divided into macro and micro level characteristics. Macro-level variables include variables such as income, age structure, level of education, and economic variables such as unemployment, employment. Micro-level variables include variables such as area, distance to downtown, distance to park areas, distance to daily services, and building size, which is perceived in the literature as one of the most important characteristics In the field of residential real

estate valuation , hedonic pricing models aim to determine the impact of certain micro and macro-level factors on the price of a home. (Gaetano, 2019)

Ridker and Henning (1967) found that air pollution levels are relatively significant towards residential property values. Bartik (1988) discovered that the increase in services in the area increases property prices also. Sander & Polasky (2009) researched 4918 observations in Ramsey County and found that the house prices near streams, lakes, parks and trails are higher than elsewhere. Karaganis (2011) investigated 8685 apartments with Rosen's hedonic equations and discovered that property characteristics such as size, age, location and external characteristics such as economic situation and spatial differences affect housing prices. Thanasi (2015) found, that apartment characteristics as number of rooms, parking, furniture, view and surface of living affect the house pricing. Hasanah & Yudhistira (2018) found in their research that mountain, street and sport centers near the apartment are associated with higher valuations. They also state that apartment floor height is in significant correlation with value of the apartment. Gaetano (2019) confirmed in his research that location is one of the most important variables in hedonic pricing model. Warren et al. (2017) researched 4233 residential real estates in Brisbane, Australia and found that historic districts have a positive impact in the areas surrounding the area land price.

2.5 COMPARISON OF MRA AND ANN IN RESIDENTIAL REAL ESTATE VALUATION THEORY

Residential real estate valuation can be classified as part of pattern recognition (Borst, 1991). Lenk Margarita et al. (1997) stated that efforts have been made to develop models that can find increasingly complex connections in data that regression models may not be able to exploit. The author also stated that regression model obtains a rather high estimated error rate, making it important to obtain increasingly accurate models that can be used in home pricing. Rossini et al. (1997) stated that it has become very common to use the neural network to perform various complex statistical problems such as classification and pattern recognition.

Multiple regression analysis (MRA) and Artificial neural network (ANN) have been compared many times in the previous literature. Going from previous studies to newer ones, it is noticed that researchers often obtained divergent results from their researches, and the ultimate superiority between the methodologies has not been fully unequivocally established.

Do & Grudnitski (1992) found in their research that the neural network can estimate the price of housing much more accurately than the multiple regression model. Lenk et al. (1997) found in their

research with 288 observations that hedonic regression model outperformed ANN by mean absolute error in both normal, and outlier sample dataset. The mean absolute error of the hedonic MRA model is smaller than the mean absolute error of the ANN model by 0,6 % units and also the maximum absolute error of MRA is 4,5% units smaller than ANN (Lenk Margarita M et al., 1997).

Rossini (1997) concluded in his study with 223 observations that multiple regression analysis is a better way to value properties than a neural network as the study found that the MRA model had mean absolute error 12.74% against ANN's mean absolute error of 19.97%. The correlation between actual and predicted prices was 0.86 model with MRA and 0.69 when modeling with ANN (Rossini, 1997).

Limsombunchai et al. (2004) researched predicting power of ANN and hedonic regression models. The study resulted that ANN is superior way of predicting housing values as the RMSE is lower than in hedonic MRA model and the predicted values are closer to actual values than in hedonic regression model. The best ANN model had R^2 of 0,9 and RMSE of 449,111 as the best hedonic model has the r^2 of 0.7499 and RMSE of 642,580 (Limsombunchai et al., 2004).

Peterson & Flanagan (2009) researched 46 467 apartment values using ANN model and regression model. They found, that the ANN model had lower difference between the predicted and actual outcome than the linear regression model.

Zurada et al (2011) found in their research of 16 366 observations that the neural network method had higher RMSE, MAE and lower R^2 value than MRA model. They stated that the MRA is more accurate way to predict house values than the artificial neural network. (Zurada et al., 2011)

Chun and Satish (2011) investigated the housing sample of 33 342 observations and found that the ANN model outperformed multiple regression model. In the training set, the mean absolute error of the ANN model was 21% lower than that of the multiple regression model. In the test set, the mean Absolute error of ANN was 18% lower than that of the multiple regression model. The RMSE of the ANN training set was 23% lower than the RMSE of the multiple regression model. The RMSE of the ANN test set was 11% lower than the RMSE of the multiple regression model (Chun & Mohan Satish B, 2011).

McCluskey et al. (2012) found in their research that semi-log and double-log regression model outperformed ANN model by the means of comparing the predicted values with actual values. Also the mean absolute percentage error was higher in ANN than with semi-log and double-log regression models (William et al., 2012).

(Amri & Tularam, 2012) found that the artificial neural network performed slightly better than linear regression in large parts of the data set but the multiple regression model had not been finalized in such a way that the model could not be refined. Their research showed that the neural network model has a higher R^2 value than the multiple regression model. For the ANN model, 31% of the valuation estimates were less than 10% spread away from actual prices. For the MR model the corresponding figure was 28%. For the ANN model, 56% of the valuation estimates were less than 20% spread away from actual prices. For the MR model the corresponding figure was 51% (Amri & Tularam, 2012).

McCluskey et al. (2013) researched ANN modeling with 2 694 observations and found that the ANN model was more accurate than traditional MRA model. (McCluskey et al., 2013)

The table 3 shows how artificial neural network and regression models have performed against each other. As it can be seen from table 3, it is not entirely clear which of the models works best for estimating house prices.

Table 3. Previous studies comparison of ANN and Regression models

Authors	**Result**	**Observations**
Do, A. Q., & Grudnitski, G. (1992).	Neural network outperforms regression	288
Lenk Margarita M, Worzala Elaine M, & Ana, S. (1997)	hedonic regression outperformed ANN	288
Rossini, P. (1997)	Regression outperformed ANN	223
Limsombunchai, V., Gan, C., & Lee, M. (2004)	ANN Outperformed regression analysis	200
Peterson, S., & Flanagan, A. (2009).	ANN outperforms OLS regression	46 467
Zurada, J., Levitan, A., & Guan, J. (2011).	Regression models outperform ANN	16 366
Chun, L. C., & Mohan Satish B. (2011).	ANN outperformed Multiregression and nonparametric regression	33 342
William, M., Peadar, D., Martin, H., Michael, M., & David, M. (2012)	semi-log regression outperforms ANN	-
Amri, S., & Tularam, G. A. (2012).	ANN outperformed multiregression	7 849
McCluskey, W. J., McCord, M., Davis, P., Haran, M., & McIlhatton, D. (2013).	ANN outperforms traditional MRA	2 694

3 METHODOLOGY

The first methodology chapter discusses ANN concept on a general level. The second chapter goes through model comparison process between the MRA methods and the ANN method.

The third chapter divides into two sub-categories. First sub-category goes through one of the most important forms of ANN, the Multilayer perceptron (William et al., 2012). First category also reviews what the architecture of the MLP is, what sources and statistical programs have been used to build the model and what kind of equations are inside the MLP model. The first sub-category also goes through how the MLP model process proceeds from start to finish step by step and what are the advantages and disadvantages of ANN.

The other sub-category of the third chapter explains the different formulas of the MRA for the reader. These formulas include methods such as OLS, double-log, and semi-log -regression. This category also introduces the advantages and disadvantages of using MRA method in residential real estate valuation.

Selection of models used to valuate Helsinki housing data is based on what have been the most frequently used methods in previous residential real estate valuation literature. Hedonic multiple regression as well as hedonic ANN structure MLP have been chosen for the focus methodologies in this research as the models have been compared many times in the residential real estate valuation earlier literature. This facilitates the placement of the results of this research as a continuation of previous literature. Other qualifying characteristics of the chosen models are that they are not too complex to use for commercial purposes. They do not take up a significant amount of processing time. They can be used with the data available.

In summary, it could be said that the models are selected so that the results of this study is easy to compare with the results of previous studies, the methodologies are easy to use and the processing time of the models is kept to a minimum. A great limiting factor of selected methodologies is the quality of the data used in the study. The data is not time-series nor accurately spatial in its nature.

3.1 ARTIFICIAL NEURAL NETWORK IN GENERAL

Artificial Intelligence has evolved tremendously due to increased research into the human brain in the last century. One of the important artificial intelligence applications is called Artificial Neural Network (ANN). Neuropsychologist Warren McCulloch and mathematician Walter Pitts first modeled a simple artificial neural network using electronic circuits in 1943. (Jimmy Pang,)

In recent years, Artificial Intelligence has become an increasing phenomenon across various different commercial industries because of the increase in the computing speeds. These areas include investing, marketing and, for example, healthcare. The ANN method is often used in tasks related to estimation, classification, prediction and approximation (Eija, 2004).

ANN seeks to simulate a highly simplified human neural system. The human brain system has dendrites, which are imitated by input neurons in ANN model. In human brain, the data is transferred from dendrites to a soma where the data is formed by a specific function. The function in human brain's soma is mimicked by sum and transformation function in hidden neurons in the ANN model. After that the data is transferred to axon in human brain, which can be described as the output neuron in the ANN model. Currently, one of the largests publicly known deep learning neural network models has 11.2 billion parameters in it. (Kalogirou, 2014; Le, 2013; Mora-Esperanza, 2004)

Neural network models can be classified into two different sections based on their learning manners: Supervised and unsupervised learning. In addition to these, there is reinforcement learning, which, however, belong to the category of supervised learning. (Diwan, 2019)

In the Supervised learning the process is as follows: the model is provided with certain datasets with both model input and output values. The ANN model then tends to process the input variables itself and compare them to the output variables. When the ANN model obtains output values that differ from the target values, it tends to change the weight values gradually so that the value of its output approaches the given target values. The second method is Unsupervised learning in which, no one looks at the learning process. The model is given certain output values from which the ANN model itself searches for certain iterative formulas. After the ANN model has processed the data itself, input vectors are classified based on their similarity. Eventually the input vectors activate the same output clusters, after which the user of the model must interpret what certain clusters mean. (Eija, 2004)

3.2 METHODOLOGIES COMPARISON PROCESS

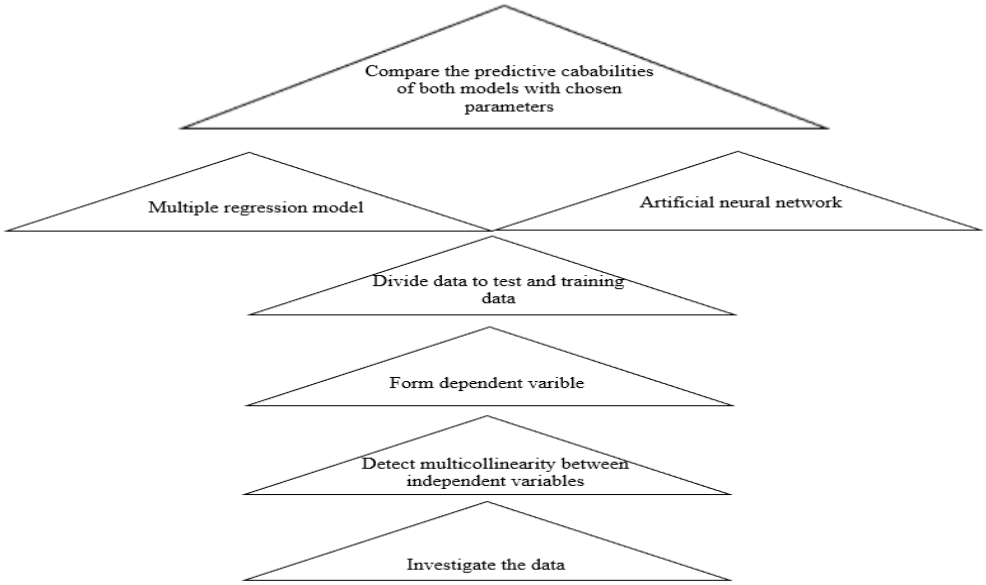

Figure 2. Model comparison process influenced by Chun & Mohan (2011)

The model comparison process is shown in figure 2 above.

1. The process starts with checking for the null values and values with clearly incorrect data values. All incorrect data is removed.
2. Second step is to detect multicollinearity between independent variables by making a variance inflation factor table. Variables with VIF value over 10 are removed.
3. Third step is to form dependent variable. The OLS uses the normal house price. Double-log and semi-log -regressions use the log prices of the house. Double-log regression also uses the log values of variables living space, income level and year of construction.
4. Fourth step is to divide the data into two parts. The first part contains training data which the models are trained with. The second piece of data contains test data that tests the ability of models to predict outcomes from untouched data. The data will be divided multiple times using for loop function in Rstudio -program. The final predictive results will be computed as the average results of multiple cross validation processes.

5. Comparing methodologies using RMSE, MAE and different accuracy thresholds (%).

3.3 METHODOLOGIES IN COMPARISON

3.3.1 Multilayer perceptron

The multilayer perceptron is one of the most important and most widely used forms of the ANN (William et al., 2012). The MLP model for this research is done using Rstudio software. The software has a few different packages with which an MLP model can be made. The packages are called "nnet" (Ripley, Venables, & Ripley, 2016) and "neuralnet" package (Fritsch, Guenther, & Guenther, 2016). The "neuralnet" package has been used in this study as it can be used to build MLP artificial neural network for regression analyzes, which we are constructing in the research. (Günther & Fritsch, 2010) The following figure 3 presents the architecture of simple MLP (Fritsch et al., 2016):

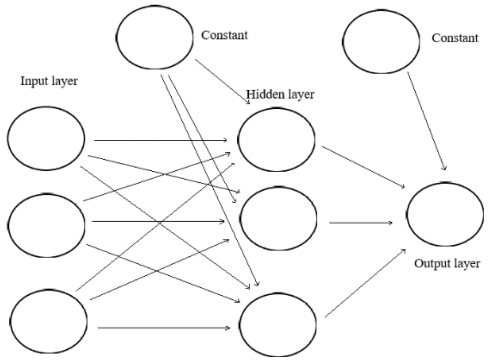

Figure 3. Architecture of simple Multilayer perceptron

Figure 3 is an example of MLP model presented by Günther & Fritsch (2010) with 3 input variables, 3 hidden neurons and 1 output neuron and one constant neuron connected to hidden layer and one constant connected to output layer. The constants are not directly affected by the independent variables (Günther & Fritsch, 2010). The MLP model consists of neurons arranged in layers. The arrows between the layers mean the weights between the neurons. These weights transfer the data between all the neurons (William et al., 2012). In the "neuralnet" package model, the weights can only be connected to subsequent layers (Günther & Fritsch, 2010). Input layer neurons consists of the coefficient variables and the output layer consists of the response variables. First Input layer values are transferred to the hidden layer according to the weights between input layer and hidden layer. Weighted Summation and the transformation function take place in the hidden layer. The summation function is combining the incoming signals and activation function determines connection between the input and output of the function. The transformation function can be, for example, a

linear function, a step linear function, a sigmoid function or a Gaussian function (Pagourtzi, Metaxiotis, Nikolopoulos, Giannelos, & Assimakopoulos, 2007).

The neural network model with one input layer, one hidden layer and one output layer is calculated as following function by Günther & Fritsch, (2010):

$$o(x) = f\left(w_0 + \sum_{j=1}^{J} w_j \cdot f\left(w_{0j} + \sum_{i=1}^{n} w_{ij}\, x_i\right)\right)$$

$$= f\left(w_0 + \sum_{j=1}^{J} w_j \cdot f\left(w_{0j} + w_j^T x\right)\right),$$

Equation 1. Simple MLP function (Günther & Fritsch, 2010)

Günther & Fritsch (2010) described that w_0 determines the intercept of the output neuron, w_{0j} is the intercept of the j:th hidden neuron. The w_j determines neuron weight corresponding to the weight starting at the j:th hidden neuron leading to output neuron $w_j = (w_{1j}, \ldots w_{nj})$ vector of all the weights leading to the j:th hidden neuron and $x = (x_1, \ldots, x_n)$ is the vector of covariates variables in the model (Günther & Fritsch, 2010). The formula calculates output o(x) in for given input variables x and selected weights. Learning algorithms seek to minimize the error function of the model as efficiently as possible with the given threshold limitation (Fritsch et al., 2016). Our model uses resilient backpropagation with weight backtracking algorithm "rprop+" (Riedmiller, 1994) .

Günther & Fritsch (2010) presented the that the error term in the model is sum of squared errors which measures difference between predicted and observed output values. The l = 1,…,L indexes observations, h = 1,…,H is the output nodes. All the weights adapt according to the rule of learning algorithm and the model reconstructs weights between the neurons until the absolute partial derivatives of error function respect to weights $\partial E / \partial w$ are smaller than the threshold value which is 0.01. The function of the SSE is as follows (Günther & Fritsch, 2010):

$$E = \frac{1}{2}\sum_{l=1}^{L}\sum_{h=1}^{H}(0lh - ylh)^2$$

Equation 2. Sum of squared errors (SSE) (Günther & Fritsch, 2010)

Several different learning algorithms have been made for the feedforward neural network model. Most of these algorithms are based on gradient descent algorithms.

Rumelhart et al. (1986) introduced a backpropagation scheme and understood that one might only look for a local minimum value and not necessarily a global minimum value. Johansson et al. (1991) presented a backpropagation algorithm using a faster conjugate gradient method than previous backpropagation algorithms for the MLP model. Riedmiller (1993) developed a resilient backpropagation algorithm with weight backtracking (rprop+). All learning algorithms seek to minimize the error function by increasing learning rate to the connecting weights that go in different directions from the gradient. Weight backtracking in the resilient algorithm refers to undoing the last iteration and adding a slighty lower value to the weight in the next group of iterations. (Johansson, Dowla, & Goodman, 1991; Riedmiller & Braun, 1993; Rumelhart, Hinton, & Williams, 1986)

Ciaburro & Venkateswaran (2017) and Esperanza (2004) outlined the neural network in step by step processing as follows:

1. select random values for both weights and Biases, also select the transformation function into the hidden layer
2. Divide the data into two different parts. The first part is the training data set. The second part is the test data set.
3. enter training data set for the model on the input nodes
4. calculate the output values for each neuron from the input layer through the hidden layer to the output layer
5. calculate output error (original values - predicted values)
6. Use the output error to calculate the error signals to previous layers. Partial derivation of the activation function is used to compute error signals
7. use error signals to change the weight of the inputs
8. change the weights

repeat points from 4 to 8 in a looping process until the error is within the allowable limits when training the model, the updated values and biases are set at the beginning of the next cycle after each training round. (Ciaburro & Venkateswaran, 2017; Mora-Esperanza, 2004)

Artificial neural network manages to learn how to solve problems and find different repetitive patterns in data without certain algorithms set by coding. ANN is a fast learning and adapting structure. (Tay Danny P H & Ho David K H, 1992)

ANN is well suited for the estimation of housing data. The data contains lots of dummy variables. Peterson and Flanagan (2009) noted that ANN is not dependent on rank of regressor matrices. The

ANN has also higher accuracy than linear models when it is estimating values of properties that are outliers in the data (Mora-Esperanza, 2004).

The ANN has disadvantages as well. Setting the error threshold in ANN too small can cause over training of the ANN model. The over trained model will not be able to generalize the recurring formulas found in the existing data to the new stream of data generated. This results in low prediction capability of the model. Similarly, if the error threshold is set too wide, the predictive power of the model is formed low due to the under training of the model. In this research, the error threshold is set to a default so that the model does not become over or underfitting. Artificial neural network's black box-like and complex behavior makes it difficult to conduct straightforward research on model development as well as to strive to develop consistent studies on method performance. These problems are related, for example, how many hidden layers there should be in the model. How many neurons should be placed in the hidden layer and what should be the relationship between input layers and hidden layers etc. (Lenk Margarita M et al., 1997; Tay Danny P H & Ho David K H, 1992)

3.3.2 Multiple regression analysis

The housing price is chosen to be dependent variable in the model. The housing characteristics and attributes are the independent variables in the multiple regression model. The model uses characteristics like apartment size, number of rooms or balcony. The model also explains how variables like income or location impact to real estate pricing. The independent variables are chosen to the final regression model by observing their p-value in stepwise regression. OLS data qualification requirements affect also the variables chosen in the valuation models. (William et al., 2012)

McCluskey et al. (2012) presented three different regression models where Y is dependent variable, B0 is constant variable, B1…Bk are coefficients (beta estimates), X1…Xk are the independent variables and ε is the error term. The three different regression models are as follows (William et al., 2012).

The OLS regression:

$$Y = \beta 0 + \beta 1 * X1 \dots \beta k * Xk + \varepsilon$$

Equation 3. OLS function (McCluskey et al.,2012)

Semi-log regression:

$$Ln(Y) = \beta 0 + \beta 1 * X1 \ldots \beta k * Xk + \varepsilon$$

Equation 4. Semi-log regression function (McCluskey et al.,2012)

Double-log regression:

$$Ln(Y) = \beta 0 + \beta 1 LnX1 \ldots \beta k LnXk + \varepsilon$$

Equation 5. Double-log regression function (McCluskey et al.,2012)

OLS regression can be used to measure how well a linear modeling approach works for housing data. Semi-log regression and double-log regression can be used to more accurate estimate nonlinear data using the model. (William et al., 2012)

MRA is a very transparent and defensive valuation method. The method is very familiar in the literature in many statistical fields. Often in apartment valuation modeling, it is important to get the repetitive and stable results which multiple regression analysis usually produces for its author. The hedonic price method using multiple regression is considered to be the dominant method in calculating the effects of various internal and external independent variables on residential real estate values. (William et al., 2012)

However, also multiple regression models have some disadvantages. If the training data of the regression model is incomplete or the data has many outlier values, the predictive results may be poor. However, the predictive result can be improved by removing outlier values from the data and preprocessing the data before use. The model also assumes that training data is normally distributed even if it is not. This leaves it up to the user of the model to determine if the regression technique can be used for a specific data. Often there are correlations between different variables in the dataset. The multicollinearity between independent variables can destroy the predictive performance of the regression model. The author must make his own choices about which variables are retained and which are deleted, and also what degree of correlation is accepted between the independent variables. The data used in regression analysis must be specified and measured in quantitative form which causes data availability to be one of the disadvantages of MRA. For example, numerating location information into a model can be challenging. The regression analysis can be also very easily over or underfitted to the training data. (Paul, Michael, & Matthew, 1996)

It can be summarized that the major problem of multiple regression analysis is that several decisions affecting the outcome remain the concern of the model user, which can lead to human error.

4 CASE: VALUATION OF FINNISH HOUSING DATA

This chapter consists of the following topics: first, the data is introduced on a general level, second it is described how and where the data was collected, third it is explained how the data is preprocessed before modeling and consolidation, fourth it is explained in what way the data is consolidated into a single file and how the dummy variables are formed for all the variables, last it will be discussed what kind of research methods are used in this research. In the last part also a MRA and ANN are performed to estimate housing prices within the Helsinki housing data.

4.1 DATA PROCESSING DIAGRAM

Figure 4 is a diagram of the data processing. The data processing starts to advance from the bottom and ends at the top of the pattern as follows:

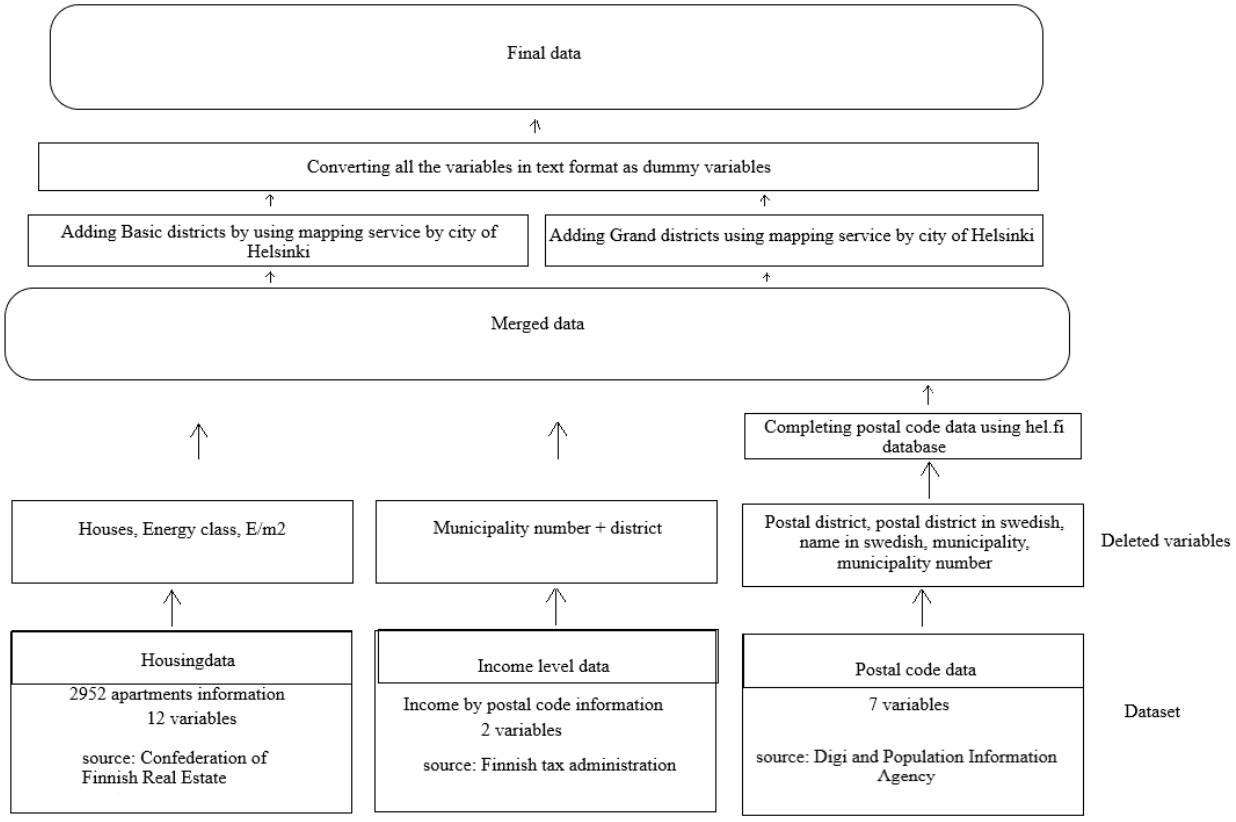

Figure 4. Data processing diagram

The figure 4 above is a simplified view of data processing in this research. As the figure shows, data processing starts with the processing of separate datasets. Once these data sets have been processed, the combined data is modified. the combined data is processed by adding grand and basic district

variables to the data. last text files will be converted to a computable format using dummy variables and the final dataset is formed.

4.2 DATA DESCRIPTION

The housing dataset has been collected from the Confederation of Finnish Real Estate (KVKL) website. The service is used to retrieve actual home transaction data for the last 12 months in Helsinki area. It should be noted that there are no dates in the Helsinki housing data. The housing data contains data on 2952 dwellings classified into the following categories: District, Apartment, Houses, m2, Vh €, €/m2, Rv, Krs, Elevator, Condition, Plot, Energy class. There are 12 variables in this data as well as 14741 observations overall. The variables are in following table 4:

Table 4. Variable description KVKL (2020)

Variable	**describtion**
District	Districts displayed from intermediaries and may differ from the official districts used by the cities. However, in terms of the smooth running of the research as well data handling perspective, It is assumed that neighborhoods are the official neighborhoods used by the city.
Apartment	Consists of several parts: studio, one bedroom, triangle, 4 or larger. This category includes a description of the apartment's rooms, utility rooms and equipment.
Houses	Includes only apartment buildings class.
M2	M2 includes the surface area of the apartment in square meters.
Vh €	refers to the debt-free price of the apartment. It includes the sale price of the dwelling and the total amount of corporate debt that may be incurred by the dwelling at the time of the transaction.
€/m2	means the debt-free price per square meter of the apartment.
Rv	variable refers to the year of construction, which can also be the year of commissioning in completely renovated houses.
Krs	refers both to the floor number of the apartment and the number of floors of the house.
Elevator	indicates that the condominium either has or does not have an elevator.
Condition	can be either bad / satisfactory / good.
Plot	can be either own or rent.
Energy class	refers to the apartment’s energy solution. It can either be E / F / C or D + year.

The second data has been collected from the tax administration's statistical database for personal income tax section. The website is vero2.stat.fi. The data has two different variables: postal number and average income level (including both men and women) with 166 observations. This income data is located to Helsinki area. The income data used in the research is from 2018. It should therefore be noted that it is not fully comparable with the 2019 housing price data. We did not use 2019 statistics as they were not available. However, we also wanted to use external factor variables in the research.

One should notice if were to use the model to predict in the future, to pay attention to the separate timing of income data relative to home sales. (Finnish tax administration, 2020)

The third data included all Postal Codes in the Helsinki area and the corresponding names of districts both in Finnish and Swedish. This postal code data is located to Helsinki. The postal code data used is from 2020. The postal code data was collected from Digital and Population Data Services Agency website. (Digital and Population Data Services Agency, 2020)

4.3 DATA PREPROCESSING AND CONSOLIDATION

4.3.1 Data preprocessing and consolidation in general

The preprocessing of this research as well as the consolidation of different data packages is done using Excel's built in data consolidation and preprocessing software called Power Query.

The original housing data has 12 variables. District, Apartment, Houses, m^2, VH €, €/m^2 RV, Floor, Elevator Condition, Plot and energy class. The data contains a total of 2951 different apartments and their information.

Data processing is started by deleting nonrelevant variables from the data. The Energy Class variable is removed as the data for 471 apartments do not have the information of energy class. Second removed variable is the "houses" variable. This variable indicates whether the apartment is an apartment building or not. This variable contains the observation "KT" for all data rows, which makes it nonrelevant.

The square price per m^2 (E /m2) will be removed as the debt-free selling price variable VH / E will be set as independent variable in the models. To make the estimation of the models meaningful, the model cannot use the dept free selling price as dependent and price per square as independent variable.

4.3.2 Data preprocessing, removal of data

Table 5. Rows deleted from the data

Deleted rows	description
25	The value of the plot variable is unknown
27	The condition of the apartment is not defined
11	Unknown floor variable are deleted
1	Postal number outside Helsinki will be deleted
9	Targeted under the unfocused district "Helsinki"
4	Have district heading "center"
12	Apartment column is left blank, marked as offices, commercial apartments or garage
121	Marked as outlier as the z-score of the sale price of the apartment is more than 2

210	**Overall**	

We use the z-score for extracting outlier properties from the dataset. Lenk et al (1997) calculated the z-score by subtract the property price from the average house selling price and dividing it by selling prices standard deviation when x is the average Selling price (VH) of the apartments u is the price of an individual apartment and σ is the standard deviation of apartment selling prices as follows:

$$Z = \frac{x-\mu}{\sigma}$$

Equation 6. Z-score (Lenk et al., 1997)

4.3.3 Preparation of postcode-specific income level data

Table 6. Example of postcode-specific income level data

00100 Helsinki (091 Helsinki)	61 905
00120 Helsinki (091 Helsinki)	61 859
00130 Helsinki (091 Helsinki)	89 530

Postcode-specific income level data consists of the average income of men and women in different postal code numbers in Helsinki area. Separations are made for the column with the combined postal code + Helsinki + municipality number. All rows that are not in the Helsinki municipality number "091 Helsinki" are deleted from the income data. After this, editing the postal code data begins with deleting all other information than the district and the postcodes from the data.

4.3.4 Preparation of postal code data

The postal number data consists of postal number, post office, office, and name of the district as follows:

Table 7. Postal code data before preprocessing

00100	HELSINKI	HELSINGFORS	Helsinki Keskusta - Etu-Töölö
00120	HELSINKI	HELSINGFORS	Punavuori
00130	HELSINKI	HELSINGFORS	Kaartinkaupunki

This data is processed as follows: all but the first and last columns are deleted as there are irrelevant information. Last column will be divided into many rows. For example, the first row becomes two parts: 00100 Keskusta and 00100 Etu-Töölö. This means that the rows with many locations are separated into two separate rows so that the merge function of the power query identifies both of the locations. Eventually all the text columns are reduced to lowercase letters to ease the "merge" functions task.

4.3.5 Combining the data

1. Combining postal code data with postcode-specific income level data to obtain district for all zip codes.
2. Combine this data with housing data using the power query's "merge" function. Equivalences 1746/2741 are obtained.
3. The postal number data will be complemented from the City of Helsinki website. Search is done one at a time by typing districts with missing postal codes example "Oulunkylä" in the search function and looking at which postal code area the neighborhood belongs to, and filling the missing pieces to the merged income + postal code data.
4. Combining the final postal number + income level information dataset with the housing data. Result obtains 2741/2741 association result.
5. All houses in the combined data are divided into grand as well as basic districts. This is done by setting the address of the apartment in the hel.fi map service and looking at which basic and large district the apartment belongs to and writing the district in the file. This is done for all the different locations.

The data used was not the highest quality for conducting a quantitative study. Most of the data was in text format and had to be converted to quantitative format. Housing characteristic column of the data was very poor as the column contained a lot of abbreviations, periods, commas, and other special characters. The way the house characteristics of the apartments were marked was not standard enough. Converting text format as dummy variables was very challenging. When data was converted to a quantitative format using dummy variables, some of the data may have lost in the process. The author also combined the data from many sources. Poor targeting, data processing, typos or other human errors in data processing stage could affect the quality of the Helsinki housing data and the research results.

4.4 DUMMY VARIABLES

Indicator variables include dummy, binary and dichotomous variables. These variables can have either a value of zero or one. They express if a feature exists or not. These features can be, for example, the existence of balcony or alcove in the apartment. On the other hand, they also tell you if a certain condition is true or false. In this study, such a condition can be for example, whether the dwelling is located in the area of a particular grand/basic district or not. (Hill, Griffiths, & Lim, 2018)

An advantage of using dummy variables in determining home locations is that it is very simply and easy way to address the location of the apartment quantitatively. Contrarily it is not as accurate as using spatial methods for valuating the location effects on home prices. (Gaetano, 2019)

Table 8 shows in the first column which variable is in question. Second column is the metric of the variable and third column includes the content of the variable presented. The table is as follows:

Table 8. variables description and measurement (Kayode and Modupe., 2012)

Variable	Variable description	Category
Sales price	Eur	debt-free price of the apartment. It includes the sale price of the dwelling and the total amount of corporate debt that may be incurred by the dwelling at the time of the transaction.
m^2	m^2	surface area of the apartment in square meters.
Income level	Eur	The average income level of men and women in this postcode area in euros
Year of construction	Number	Construction year variable refers to the year of construction, which can also be the year of commissioning in completely renovated houses.
Elevator	Dummy	The elevator variable indicates that if the condominium has elevator = 1 or does not have an elevator = 0.
Plot	Dummy	The plot variable has only two values. The plot can be own = 1 or rental = 0.
Grand district	Dummy	Helsinki is divided into eight grand districts. The apartment belongs to one of these districts. The districts are listed in the appendices due to their large number. The variable has value 1 under the grand district if located in in it. If it is not = 0
Standard district	Dummy	There are 34 basic districts in Helsinki. All apartments are located in some of these districts. They are marked as either belonging = 1 or not belonging = 0 to a particular district.
Apartment characteristic		
Balcony	Dummy	If the apartment characteristic row includes abbreviation for a balcony = 1 If the apartment characteristics row does not include abbreviation for a balcony = 0
Open plan kitchen	Dummy	If the apartment has Open plan kitchen = 1 If the apartment does not have Open plan kitchen = 0
Sauna	Dummy	If the apartment has a sauna = 1, If the apartment does not have a sauna = 0
Kitchenette	Dummy	If the apartment has a kitchenette = 1, If the apartment does not have a kitchenette = 0
Alcove	Dummy	If the apartment has a alcove = 1, If the apartment does not have a alcove = 0
1h	Dummy	If the apartment is classified as an studio apartment = 1, If it is not classified as an apartment studio = 0
2h	Dummy	If the apartment is classified as an two unit apartment = 1, If it is not classified as an two unit apartment = 0
3h	Dummy	If the apartment is classified as an triangle = 1, If it is not classified as triangle = 0
4h	Dummy	If the apartment is classified as an four room apartment = 1, If it is not classified as four room apartment = 0
5h+	Dummy	If the apartment is classified as an five room apartment or more = 1, If it is not classified as five room apartment or more = 0
Floor		
Ground floor	Dummy	If the apartment is on the lowest or second lowest floor = 1, if it is not in the lowest or second lowest floor = 0
Top floor	Dummy	If the apartment is located on the top floor = 1, if the apartment is not located in the top floor = 0
Middle floor	Dummy	If the apartment is on a floor other than the top floor or the lowest floor = 1 if it is in the Ground floor or Top floor group = 0
Apartment condition		
Good	Dummy	If the real estate agent's or homeowner's assessment of the condition of the home is good = 1, if not = 0
Bad	Dummy	If the real estate agent's or homeowner's assessment of the condition of the home is bad = 1, if not = 0
Satisfactory	Dummy	If the real estate agent's or homeowner's assessment of the condition of the home is satisfactory = 1, if not = 0

4.5 SELECTION OF ERROR METRICS

In this study, the following types of error metrics were selected to measure the performance of different models relative to each other: RMSE, MAE and prediction accuracy thresholds (%).

The accuracy threshold values are ±10%, ±20% and ±25%. The Prediction accuracy threshold (%) indicates how many percent of all estimated home prices fall within these already stated threshold values. (Aurélio Stumpf González Marco, Lucio, & Torres, 2005; Chun & Mohan Satish B, 2011; Stanley et al., 1998; William et al., 2012).

The term Mean Absolute error (MAE) and the term Root Mean Square Error (RMSE) are very often used as an error measure in statistical modeling and especially in measuring the accuracy of housing valuation models in the previous literature. These measures aim to compensate for the predictability of dependent variability (Chun & Mohan Satish B, 2011; Limsombunchai et al., 2004; Michael, Bourassa Steven C, Martin, & Donato, 2019; Rossini, 1997; Zurada et al., 2011).

RMSE is the square root of the mean of the squared errors between the predicted house prices and the actual test set house selling prices (Pierluigi, Francesco, & Marco, 2018). RMSE is considered by some authors to be sensitive to outlier values and should not be used in comparing models using large data sets (Hyndman & Koehler, 2006).

MAE is the average of the absolute values of the prediction errors. MAE processes errors very evenly according to their magnitude (Zurada et al., 2011). Zurada et al. (2011) present the formulas for both error parameters as follows, when y_t parameters are the actual apartment selling prices and $\hat{y}_t$ parameters are the predicted apartment selling prices.

The RMSE equation is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_t (y_t - \hat{y}_t)^2}$$

Equation 7. RMSE (Chun & Satish., 2011)

The MAE equation is as follows:

$$MAE = \frac{1}{n}\sum_t |y_t - \hat{y}_t|$$

Equation 8. MAE (Chun & Satish., 2011)

4.6 MULTIPLE REGRESSION ANALYSIS

Before starting a regression analysis or ANN model, a stepwise multiple regression is performed. The objective is to find and remove those variables p-value is greater than 0.05 in two-tailed statistical test. It follows that all the variables included in the final MRA, as well as in the ANN, are those that fell below this significance level.

4.6.1 Stepwise regression

Before analyzing the Helsinki housing data using OLS, semi-log regression or double- log regression, a stepwise regression is performed. The stepwise regression is performed to remove variables that have p-value over 0.05 threshold value. The insignificant variables are excluded by starting to delete the variables with the highest p-values one by one in order of magnitude. The final members of the model can be seen in table 11.

The formula of stepwise regression is as follows:

$$Y = \beta 0 + \beta 1 * X1 \ldots \beta 61 * X61 + \varepsilon$$

Equation 9 Stepwise regression

MCcluskey et al. (2012) form the formula as Y is the housing price, β0 is the (Intercept) β1… β61 are the regression coefficient estimates X1…X61 are the independent variables in the model and ε is the error term.

Following table shows the first stepwise regression model:

Table 9. Stepwise regression

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-868913	104414,6	-8,32175	1,36E-16
2	m.2	4453,866	119,2316	37,35475	2,59E-246
3	Income.level	2,030232	0,186634	10,87812	5,31E-27
4	Year.of.construction	377,0329	50,07383	7,529541	6,91E-14
5	Top.floor	7748,157	29349,11	0,264	0,7918
6	Ground.floor	-6805,88	29328,38	-0,23206	0,816511
7	Middle.floor	3884,123	29465,71	0,131818	0,895138
8	ElevatorÂ.	8143,625	2747,327	2,964199	0,003061
9	PlotÂ.Â.	39085,73	3049,539	12,81693	1,52E-36
10	GoodÂ.	31348,74	2469,04	12,69673	6,50E-36
11	BadÂ.	-23458,4	6513,035	-3,60177	0,000322

12	SatisfactoryÂ.	NA	NA	NA	NA
13	X1h	55314,55	11873,39	4,6587	3,34E-06
14	X2h	55276,7	9955,162	5,552566	3,09E-08
15	X3h	51543,25	8564,772	6,018053	2,01E-09
16	X4h	21817,92	8077,405	2,701105	0,006954
17	X5h.	NA	NA	NA	NA
18	Balcony	1114,535	2994,911	0,372143	0,709816
19	Open.plan.kitchenÂ.	20155,12	3332,199	6,048594	1,66E-09
20	Sauna	5814,062	3028,795	1,919596	0,055015
21	Kitchenette	16710,14	3363,927	4,967451	7,21E-07
22	Alcove	4391,48	5972,702	0,735258	0,462246
23	French.balcony	18893,93	9584,636	1,971273	0,048795
24	kaakkoinen	-128636	14444,38	-8,90559	9,57E-19
25	etelÃ.inen	64532,62	13165,97	4,901473	1,01E-06
26	keskinen	3740,576	13372,62	0,279719	0,779715
27	lÃ.ntinen	15345,42	9605,298	1,597599	0,11025
28	itÃ.inen	-113835	15249,26	-7,46493	1,12E-13
29	koillinen	-128556	21633,43	-5,94248	3,17E-09
30	pohjoinen	-150087	35658,59	-4,20899	2,65E-05
31	herttoniemi	90757,65	10174,36	8,920234	8,42E-19
32	kampinmalmi	-5239,06	7088,548	-0,73909	0,459918
33	ullanlinna	-1177,87	8530,047	-0,13809	0,890183
34	alppiharju	19843,68	9004,118	2,203845	0,02762
35	vironniemi	16704,99	8501,393	1,964971	0,049521
36	lauttasaari	-40601	7524,913	-5,39555	7,43E-08
37	taka.tÃ.Ã.lÃ.	NA	NA	NA	NA
38	haaga	-54547,7	8231,982	-6,62632	4,14E-11
39	kallio	14416,98	8580,713	1,680161	0,093042
40	vallila	32701,47	10319,66	3,168853	0,001548
41	vartiokylÃ.	9156,581	11466,8	0,79853	0,424634
42	mellunkylÃ.	-35476,1	10675,28	-3,3232	0,000902
43	munkkiniemi	-21883,3	9146,918	-2,39243	0,016806
44	pitÃ.jÃ.nmÃ.ki	-94393,6	9336,901	-10,1097	1,32E-23
45	pasila	-11141,9	11200,41	-0,99478	0,319933
46	kaarela	-113375	9440,163	-12,0099	2,10E-32
47	kulosaari	83931,76	15445,63	5,434014	6,00E-08
48	malmi	-9134,59	16108,98	-0,56705	0,570728
49	vanhakaupunki	NA	NA	NA	NA
50	vuosaari	-2360,32	10823,42	-0,21808	0,827387
51	latokartano	19151,04	13282,58	1,441816	0,149471
52	suutarila	-46638,5	18469,27	-2,5252	0,01162
53	reijola	NA	NA	NA	NA
54	puistola	-62442,3	16864,68	-3,70255	0,000218
55	pukinmÃ.ki	314,9435	17288,83	0,018217	0,985467
56	maunula	78186,74	34295,33	2,279807	0,022697
57	oulunkylÃ.	24601,12	37305,04	0,659458	0,509658

58	laajasalo	NA	NA	NA	NA
59	lÃ.nsi.pakila	54329,86	66895,9	0,812155	0,416774
60	myllypuro	NA	NA	NA	NA
61	itÃ..pakila	NA	NA	NA	NA

Several NA values can be considered from the table 9. The NA value indicate that data is missing. Several p-values exceed the upper limit of the acceptable p-value of 0.05. All NA values as well as high p-values are cleaned.

4.6.2 ORDINARY LEAST SQUARES (OLS)

Ordinary least squares (OLS) also called linear regression begins by testing whether the basic assumptions of the data used in linear regression are met. Assumptions can be defined as follows (Doszyń Mariusz, 2020; Singla & Priyanka, 2019):

1. The data is independent and indentically distributed
2. Conditional means of error terms is zero and errors are uncorrelated with regressors
3. No multicollinearity between regressors
4. No heteroskedasticity

The fulfillment of basic assumptions of linear regression are examined visually from the housing data. Quantile-Quantile plot figure 5 and Residuals vs Fitted values plot figure 6 are created as follows:

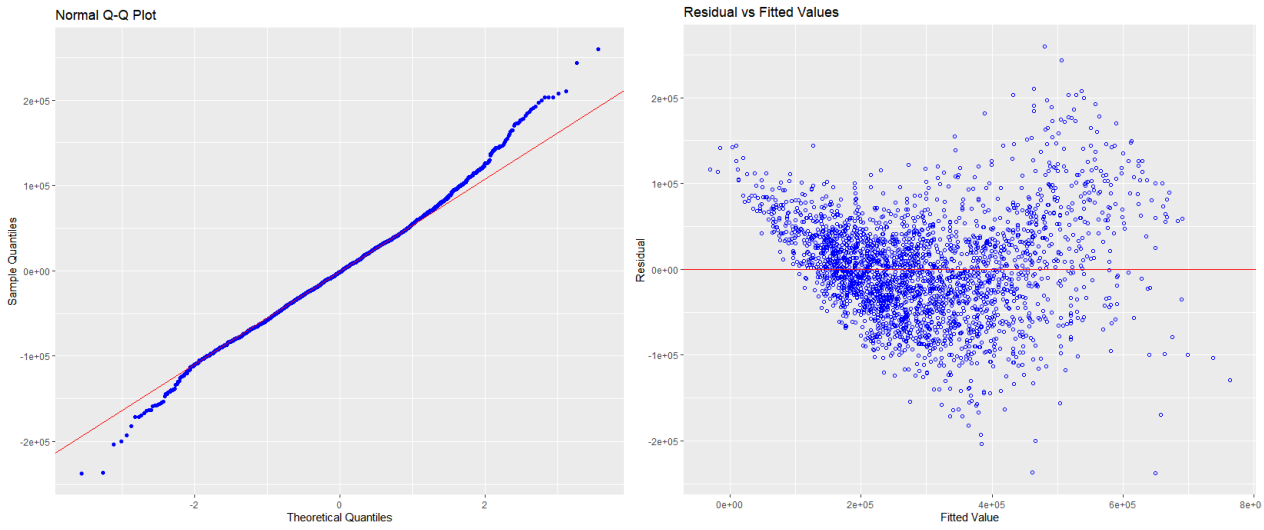

Figure 5. Left side Residual vs fitted values plot

Figure 6. Right side Normal Quantile-quantile plot

It can be seen from Figure 5 that the observations at the lowest quantiles as well as the highest quantiles do not fit the line drawn. This indicates that the first assumption is not accepted as data is

not normally distributed. It can be seen from Figure 6 that the residuals do not appear to propagate completely randomly around the 0 line. From the graph 6, it is seen that the residuals have nonequality in the variance of error terms, as the values do not spread evenly around 0, but form as if U decomposing towards the end. Residuals horn-like propagation toward the right edge of the image may indicate heteroscedasticity in the data.

Next, it is examined whether the data satisfies the multicollinearity assumption by making the variables a variance inflation factor table. Any variables that exceed the VIF value of 10 are seen to be due to multicollinearity within the data. All these variables causing multicollinearity are dropped from the model. (Montgomery, Peck, & Vining, 2012; Singla & Priyanka, 2019)

As we can see the variance inflation factors in table 10, there are no VIF values over 10 left in the final model.

Table 10.Final variables variance inflation factors (Montgomery et al., 2012)

	Variable	VIF
1	m.2	1,6
2	Income.level	3,3
3	Year.of.construction	1,7
4	Ground.floor	1
5	ElevatorÂ.	1,4
6	PlotÂ.Â.	1,5
7	GoodÂ.	1,2
8	BadÂ.	1,1
9	X2h	1,4
10	X3h	1,5
11	Sauna	1,2
12	Kitchenette	1,3
13	French.balcony	1
14	kaakkoinen	4,1
15	etelÃ.inen	5,5
16	keskinen	3,7
17	lÃ.ntinen	3,8
18	herttoniemi	3
19	alppiharju	1,5
20	lauttasaari	1,4
21	haaga	2
22	vallila	1,2
23	vartiokylÃ.	1,4
24	mellunkylÃ.	1,8
25	pitÃ.jÃ.nmÃ.ki	1,5
26	kaarela	1,6
27	kulosaari	1,5

28	malmi	1,5
29	vuosaari	1,8
30	latokartano	1,3
31	suutarila	1,2
32	puistola	1,2
33	pukinmÃ.ki	1,2
34	oulunkylÃ.	1

McCluskey et al. (2014) address that in the OLS regression model Y is housing price, β0 is the Intercept, β1… β34 are the regression coefficient estimates, X1…X34 are the independent variables in the model and ε as the error term. The final OLS model is as follows in the equation 12:

$$Y = \beta 0 + \beta 1 * X1 \dots \beta 34 * X34 + \varepsilon$$

Equation 10. OLS model (McCluskey et al., 2014

The Final form of OLS regression can be seen in the table 11:

Table 11. OLS regression

	Variable name	Regression coefficient Estimate	Standard error	T value	P value
0	(Intercept)	-838800,4334	98163,737	-8,544911402	2,12E-17
1	m.2	3892,316776	63,997741	60,81959598	0
2	Income.level	2,181808458	0,1534654	14,21694464	2,73E-44
3	Year.of.construction	358,0580277	49,612032	7,217161126	6,87E-13
4	Ground.floor	-12913,28635	2916,4879	-4,427683833	9,90E-06
5	ElevatorÂ.	10204,23546	2700,4874	3,778664387	0,0001611
6	PlotÂ.Â.	44451,4843	2919,1313	15,22764146	2,67E-50
7	GoodÂ.	34172,5362	2485,7011	13,74764506	1,28E-41
8	BadÂ.	-25159,54834	6624,1024	-3,798182301	0,000149
9	X2h	13516,9532	2782,7548	4,857400062	1,26E-06
10	X3h	19126,48051	3054,4821	6,261775355	4,41E-10
11	Sauna	6523,350154	3079,2401	2,118493484	0,0342243
12	Kitchenette	9726,007766	3245,134	2,997105108	0,0027504
13	French.balcony	20617,7983	9708,9152	2,123594443	0,0337947
14	kaakkoinen	-54904,99214	9454,2914	-5,807414823	7,09E-09
15	etelÃ.inen	138401,1752	6045,4815	22,89332547	3,62E-106
16	keskinen	87216,27631	5573,4791	15,64844405	6,66E-53
17	lÃ.ntinen	67863,45835	5213,5134	13,01683776	1,29E-37
18	herttoniemi	94536,74646	10351,313	9,132826319	1,27E-19
19	alppiharju	16889,87533	6177,3217	2,734174477	0,0062945
20	lauttasaari	-38017,31991	5891,982	-6,452382204	1,30E-10
21	haaga	-31550,63078	5995,1989	-5,262649513	1,53E-07
22	vallila	28572,11376	8445,3101	3,383192961	0,0007267
23	vartiokylÃ.	-25490,50651	8062,0343	-3,161795821	0,0015854

24	mellunkylÃ.	-70753,75551	6732,0404	-10,5100016	2,38E-25
25	pitÃ.jÃ.nmÃ.ki	-70521,89624	7522,555	-9,374726626	1,42E-20
26	kaarela	-86789,89366	7621,2451	-11,3878891	2,22E-29
27	kulosaari	75603,29427	15548,986	4,862265295	1,23E-06
28	malmi	-60891,94504	8231,685	-7,39726371	1,85E-13
29	vuosaari	-41606,33906	6745,2637	-6,168230146	7,94E-10
30	latokartano	-28961,07052	6459,5834	-4,483426989	7,65E-06
31	suutarila	-96618,63962	12138,327	-7,959799037	2,51E-15
32	puistola	-110702,7178	10215,873	-10,83634458	8,17E-27
33	pukinmÃ.ki	-50100,75609	10992,923	-4,557546233	5,41E-06
34	oulunkylÃ.	-41629,49762	16813,614	-2,4759399	0,0133494

Table 12. Summary statistics OLS regression

Residual standard error: 59090 on 2703 degrees of freedom
Multiple R-squared: 0.8293
Adjusted R-squared: 0.8272
F-statistic: 386.4 on 34 and 2703 DF
p-value: < 2.2e-16
Root Mean Square Error 59266
Mean Absolute Error 46383

Table 11 shows the regression coefficient estimates, Standard errors, t-values, and p-values of all the model variables. The summary of the model statistics is in the table 12.

The housing valuation using OLS regression would appear to be challenging using the Helsinki housing data. It seems that the first assumption of normal data distribution does not hold. On the other hand, figure 5 shows evidence that the data would have nonlinearities. The data also included noncollinearity before removing variables with high VIF values. The figure 6 probably indicates that there might be heteroscedasticity in the data. Because of these violations of the OLS assumptions, it can be concluded that OLS regression is probably not the ideal method to estimate this housing data. However, this model achieved an adjusted r-squared 0.8272. The F statistic is 386.4 on 34 and 2703 DF. The model is likely to be significant, as the p-value of the model is less than 2.2e-16. The root mean square error (RMSE) of the model is 59266 and the Mean Absolute error is 46383.

4.6.3 SEMI-LOG REGRESSION

McCluskey et al. (2014) presented the function in semi-log regression is as follows:

$$Ln(Y) = \beta 0 + \beta 1 * X1 \ldots \beta 34 * X34 + \varepsilon$$

Equation 11. Semi-log regression

when Y is the Apartment price, β0 is the Intercept, β1… β34 are the regression coefficient estimates, X1…X34 are the independent variables in the model and ε is the error term. The semi-log regression can be seen in the table 13. This table also shows the regression coefficient estimates, Standard errors, t-values, and p-values of all the model variables. Also the summarized information of the model is in the table 14.

Table 13. Semi-log regression

Variable	term	coefficient Estimate	standard error	t-value	p-value
0	(Intercept)	11,37199	0,028934	393,0356	0
1	m.2	1,926252	0,037125	51,88587	0
2	Income.level	0,507587	0,041443	12,24784	2,16E-33
3	Year.of.construction	0,248465	0,02804	8,860993	1,62E-18
4	Ground.floor	-0,03807	0,009823	-3,87553	0,00011
5	ElevatorÂ.	0,037687	0,009052	4,163288	3,26E-05
6	PlotÂ.Â.	0,143344	0,009682	14,80503	2,64E-47
7	GoodÂ.	0,119878	0,008285	14,4696	2,35E-45
8	BadÂ.	-0,13477	0,021678	-6,21713	6,06E-10
9	X2h	0,070041	0,009252	7,570144	5,49E-14
10	X3h	0,090966	0,010151	8,961554	6,77E-19
11	Sauna	0,034488	0,010296	3,349694	0,000823
12	Kitchenette	-0,01468	0,010871	-1,35009	0,17713
13	French.balcony	0,064288	0,032745	1,963308	0,049739
14	kaakkoinen	-0,17826	0,031969	-5,57587	2,77E-08
15	etelÃ.inen	0,407672	0,019794	20,59521	3,53E-86
16	keskinen	0,29587	0,018466	16,02286	1,12E-54
17	lÃ.ntinen	0,220883	0,017485	12,63294	2,41E-35
18	herttoniemi	0,310973	0,034826	8,929447	8,95E-19
19	alppiharju	0,026048	0,02082	1,251097	0,211035
20	lauttasaari	-0,0928	0,019543	-4,74837	2,19E-06
21	haaga	-0,08761	0,020025	-4,3749	1,27E-05
22	vallila	0,073922	0,028275	2,614367	0,009002
23	vartiokylÃ.	-0,1414	0,027246	-5,18959	2,31E-07
24	mellunkylÃ.	-0,39941	0,022647	-17,6366	3,90E-65
25	pitÃ.jÃ.nmÃ.ki	-0,28427	0,024776	-11,4737	1,27E-29
26	kaarela	-0,33255	0,025406	-13,0893	1,00E-37
27	kulosaari	0,262968	0,050817	5,174817	2,49E-07
28	malmi	-0,2826	0,028108	-10,0542	2,84E-23
29	vuosaari	-0,14898	0,022592	-6,59466	5,35E-11
30	latokartano	-0,13473	0,021202	-6,35451	2,54E-10
31	suutarila	-0,47536	0,041454	-11,4671	1,37E-29

32	puistola	-0,52005	0,03303	-15,7445	5,96E-53
33	pukinmÃ.ki	-0,24735	0,034918	-7,08375	1,89E-12
34	oulunkylÃ.	-0,15539	0,063291	-2,45519	0,01416

Table 14. Summary statistics Semi-log regression

Semi-log regression
Residual standard error: 0.1746 on 2155 degrees of freedom
Multiple R-squared: 0.8524
Adjusted R-squared: 0.85
F-statistic: 366 on 34 and 2155 DF
p-value: < 2.2e-16
Root Mean Square Error 53 991
Mean Absolute Error 39 203

Semi-log regression adjusted r-squared value is 0.8447. The F-statistic value is 366 on 34 and 2155 DF. The model is significant because the p-value is less than 2.2e-16. Root mean square error is 53 991 and mean absolute error is 39 203. RMSE and MAE has been calculated for the semi-log model by cross-folding the training and test dataset 100 times and calculating mean of the error measures from all the iterations.

4.6.4 DOUBLE-LOG REGRESSION

McCluskey et al. (2014) presented the function in double-log regression is as follows:

$$Ln(Y) = \beta 0 + \beta 1 LnX1 + \ldots + \beta 3 LnX3 + \beta 4 X4 + \cdots + \beta 34 X34 + \varepsilon$$

Equation 12. Double-log regression

when Y is the Apartment price, β0 is the Intercept, β1… β34 are the regression coefficient estimates, X1…X34 are the independent variables in the model and ε is the error term. The double-log regression can be seen in the equation 14. The values of m^2, income level and year of construction are in log form. All others are in normal form as they are dummy variables so it would not be meaningful to present them in log formation. The following double-log table and summary statistics are below in table 15. The log terms are not seen in the table before the terms, as the log data is converted in log form before modeling.

Table 15. Double-log regression

Variable	term	coefficient Estimate	standard error	t-value	p-value
0	(Intercept)	-13,7892	2,430671	-5,67299	1,59E-08
1	m.2	0,644938	0,01204	53,56764	0
2	Income.level	0,363991	0,02595	14,02658	7,73E-43
3	Year.of.construction	2,578195	0,31484	8,188907	4,46E-16
4	Ground.floor	-0,03432	0,00944	-3,63538	0,000284
5	ElevatorÂ.	0,033938	0,008745	3,880901	0,000107
6	PlotÂ.Â.	0,142986	0,009518	15,02235	1,38E-48
7	GoodÂ.	0,124066	0,008022	15,46503	3,05E-51
8	BadÂ.	-0,10694	0,021342	-5,01088	5,86E-07
9	X2h	-0,0153	0,00894	-1,71162	0,08711
10	X3h	0,013245	0,01034	1,280954	0,200348
11	Sauna	0,037536	0,009936	3,777619	0,000163
12	Kitchenette	0,006805	0,010707	0,635581	0,525117
13	French.balcony	0,068652	0,030534	2,248413	0,024651
14	kaakkoinen	-0,21649	0,030484	-7,10171	1,67E-12
15	etelÃ.inen	0,388408	0,020052	19,37038	3,44E-77
16	keskinen	0,287215	0,017868	16,07395	5,38E-55
17	lÃ.ntinen	0,215626	0,016716	12,89947	9,99E-37
18	herttoniemi	0,351712	0,034101	10,31375	2,23E-24
19	alppiharju	0,066895	0,020238	3,305436	0,000964
20	lauttasaari	-0,10087	0,018997	-5,30953	1,21E-07
21	haaga	-0,0904	0,019577	-4,61791	4,10E-06
22	vallila	0,104624	0,027178	3,849622	0,000122
23	vartiokylÃ.	-0,11876	0,02689	-4,41645	1,05E-05
24	mellunkylÃ.	-0,34607	0,022851	-15,1448	2,58E-49
25	pitÃ.jÃ.nmÃ.ki	-0,25818	0,024192	-10,672	6,04E-26
26	kaarela	-0,30616	0,02498	-12,256	1,97E-33
27	kulosaari	0,264981	0,051763	5,11909	3,34E-07
28	malmi	-0,27494	0,025348	-10,8465	1,00E-26
29	vuosaari	-0,14113	0,021753	-6,48773	1,08E-10
30	latokartano	-0,09933	0,021547	-4,60982	4,27E-06
31	suutarila	-0,43802	0,037929	-11,5484	5,62E-30
32	puistola	-0,53181	0,031652	-16,8016	1,25E-59
33	pukinmÃ.ki	-0,25091	0,035511	-7,06566	2,15E-12
34	oulunkylÃ.	-0,15972	0,055233	-2,89174	0,003869

Table 16. Summary statistics double-log regression

Double-log regression
Residual standard error: 0.1706 on 2155 degrees of freedom.
Multiple R-squared: 0.854
Adjusted R-squared: 0.8517
F-statistic: 370.8 on 34 and 2155 DF

p-value: < 2.2e-16
Root Mean Square Error 51779
Mean Absolute Error 38316

The double-log regression model has a multiple r-squared of 0.854 and adjusted R-squared of 0.8517. The model is highly significant as the p-value is under 2.2e-16. The root mean square error is 51779 and the mean absolute error is 38316.

4.6.5 Multiple Regression Analysis summary statistics

In the table 17, in the vertical row there are residual standard error, multiple r-squared, adjusted r-squared f-statistic, p-value, root mean square error, and mean Absolute error. On the horizontal axis, double-log, semi-log, and OLS regressions.

Table 17. Summary statistics of Multiple Regression Analysis

	Double-log regression	**Semi-log regression**	**Ordinary least squares**
Residual standard error	0.1706 on 2155 degrees of freedom.	0.1746 on 2155 degrees of freedom.	59090 on 2703 degrees of freedom.
Multiple R-squared	0.854	0.8524	0.8293
Adjusted R-squared	0.8517	0.85	0.8272
F-statistic	370.8 on 34 and 2155 DF	366 on 34 and 2155 DF	386.4 on 34 and 2703 DF
P-value	< 2.2e-16	< 2.2e-16	< 2.2e-16
Root Mean Square Error	51779	53 991	59266
Mean Absolute Error	38316	39 203	46383

The table 17 shows a few points. Double-log regression has the highest multiple r-squared, OLS has the smallest r-squared. Double-log regression has highest adjusted r-squared and OLS has the smallest r-squared. All the models are significant with p-values less than 2.2e-16. Double-log regression has smallest RMSE value of 51 779. OLS has largest RMSE value of 59 266. Double-log regression has smallest MAE value of 38 316. OLS has largest MAE value of 46 383. Semi-log regression and double-log regression performed very evenly. OLS performed much worse than the nonlinear models.

4.7 ARTIFICIAL NEURAL NETWORK

4.7.1 General formation of the model

The artificial neural network -model uses the same variables that were below the stepwise regression p-value of 0.05. Also, all NA values were removed during stepwise regression. The ANN model is a multilayer perceptron model that is constructed by one input layer, one hidden layer and one output layer. The input layer includes all 34 input variables. based on the previous literature, 50-200% of the number of neurons out of the number of neurons contained in the input layer should be selected for the hidden layer (Mora-Esperanza, 2004). In reference to previous literature about 12-70 neurons should be selected for the hidden layer. However, no more than 13 neurons are selected in this research for the valuation model for a few different reasons. The results of the neural network model vary widely. In order to obtain even reasonable prediction accuracy, the model must be run at least ten times through with each number of hidden nodes (Chun & Mohan Satish B, 2011). However, this is impossible with larger numbers of neurons as the processing time lengthens when more neurons are added to the model. The ANNs performance will be tested with yet other amounts of hidden neurons using the so-called bracketing technique. To save time, all neuron quantities are first tested once for model functionality. Once the best possible number of neurons has been found, 5-fold testing is performed on this particular model to ensure that the performance of the model remains at least relatively the same as the first time. The testing quantities used were 10, 7, 4 and 2. The final model is the one with the lowest error rates measured with metrics as RMSE and MAE).

4.7.2 Multilayer perceptron function parts

Starting weights refer to weights in connections between different neurons. These starting weights are set to random in this model. The output of this MLP model is set to be linear and not categorical, as we want to produce estimated housing prices as output (Günther & Fritsch, 2010).

the MLP model consists of the following elements: Algorithm that is used to train the model. The model of this research is trained using resilient backpropagation algorithm “rprop+” (Fritsch et al., 2016; Riedmiller & Braun, 1993). The next factor in the code is the “act.fct” which indicates the activation function of the model. In this research it is the sigmoid function also called as logistic function. The logistic function is presented as follows by Günther & Fritsch (2010):

$$(f(u) = \frac{1}{1 + e^{-u}}$$

Equation 13. Logistic function by Günther & Fritsch (2010)

The next piece of the code is the data, which implies the data used in the model. Our training data consists of 2190 apartment's information. The test data consists of 548 apartment's information. Next member of the code is called hidden. This member sets the number of neurons in the hidden layers as well as the number of hidden layers in the model. In this model, the number of hidden layers is determined by the bracketing method. The number of hidden neurons in the final model is either 10,7,4 or 2. In this research simple MLP model there is only one hidden layer used. The next part of the code is the "stepmax" part. This "stepmax" member calculates the maximum steps in the ANN model training. When the model exceeds that number of steps, the model stops training. We use a "stepmax" quantity in the model which is 1e6.

Next term is the error factor. The error factor indicates the metric that the model in question seeks to minimize. In this model, we use the sum of squared errors also named as Residual sum of squares. The formula is presented by Agostino (2020) in following way: $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.

Threshold member sets the threshold value. This value sets at which point the model stops training. Threshold is the partial derivatives of the error function as stopping criteria (Günther & Fritsch, 2010). In this model the threshold is set as 0.01. The last part is to set the output type of the model. In this model, the linear output is set as TRUE (Günther & Fritsch, 2010). This indicates that the wanted output type is linear and not categorical.

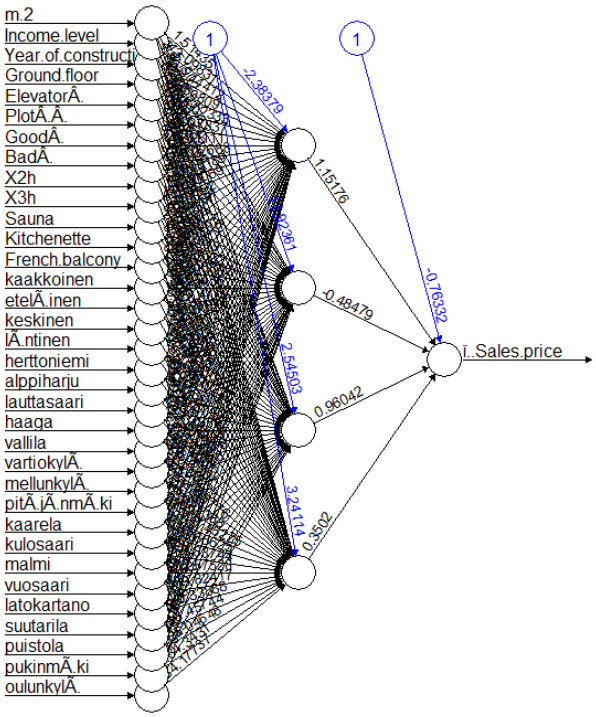

Figure 7. MLP architecture

Table 18. Summary statistics MLP

MLP	
Root Mean Square Error	38902
Mean Absolute Error	29010
Accuracy of the model (%)	
±10%	53
±20%	84
±25%	89

The final architecture of the MLP model can be seen in figure 7. The summary statistics of the model can be seen in the table 18. The final MLP model has 34 input neurons, 4 hidden neurons and 1 output neuron. The model also has 2 constants. The first constant is connected to each of the hidden neurons and the second constant is connected to the output neuron. MLP model does not produce the multiple r-squared or adjusted r-squared values.

Estimating the price of an apartment is measured using several different accuracy level %. The accuracy (%) of the ANN model is calculated as the percentage of the real estates which price estimation fell within the specified accuracy thresholds. This metric is used as this method of measurement has been used in several works in the residential real estate valuation literature (Aurélio Stumpf González Marco, Lucio, & Torres, 2005; Chun & Mohan Satish B, 2011; Stanley et al., 1998;

William et al., 2012). As noted in Table 18. The ANN model is able to estimate 53% of the apartment values ranging ±10% from the true value of the apartment. 84% of the estimated prices are ranging ± 20% from the actual values of the apartment. 89% of the estimated values are ranging ±25% from the value of the apartment. Another error metrics of the ANN model is MAE and RMSE. MAE of the Final MLP model is 29 010 and the RMSE is 38 902. Cross-folding has been performed for the MLP model three times as the computing time got too high at higher volumes of cross-folding.

5 CONCLUSION

The Conclusion part of the study consists of three separate sections. The answers to all the research questions are discussed first. The second section reviews the implications of research for the industry. The third section discusses research limitations and suggestions for future research.

5.1 RESEARCH RESULTS

Answering to research questions:

1. *What kind of residential real estate valuation models exist in the previous literature and what are the most frequently used valuation models in the research data collected for this study?*

The first research question was answered in the literature review section. Abidoye et al. (2019) and Elli et al. (2003) classified the valuation methodologies into two different categories. Advanced valuation methods, as well as traditional valuation methods. In this study, a similar division was made between different methods. The traditional real estate valuation methods found in the previous literature were comparable or comparative method, income method, multiple regression method, stepwise regression method and profits method. The advanced methods included spatial analysis, fuzzy Logic, artificial neural network and geographical information systems, ARIMA and hedonic pricing.

The most frequently used valuation method in the literature used in this study was multiple regression. The 2nd used method was hedonic pricing, the 3rd used was ANN and 4th frequently used was spatial analysis. Based on the frequency of the use of these models, it was possible to justify why this study has selected to model Helsinki housing data using hedonic multiple regression and the ANN model.

2. Which model is more accurate when predicting apartment values in Helsinki: Simple Artificial Neural Network (ANN) model or multiple regression analysis (MRA)?

In this research there was Three MRA methods and One ANN method used to estimate apartment prices in Helsinki. All models used both housing characteristic variables and external variables in the modeling process. The MRA models were: OLS, double-log and semi-log -regression. In addition to this, one artificial intelligence method: feedforward artificial neural network was used for the modelling process.

All of these valuation models were compared using three different error measures: square root of the average of the squared values of the prediction errors (RMSE), mean average of the absolute values of the predicted errors (MAE) and accuracy frequencies (%) of ± 10%, ± 20% and ± 25%.

The predictive capabilities of the models were calculated so that the OLS, double-log and semi-log models produced an multiple r-squared and adjusted r-squared values. The ANN model produced only the accuracy levels (%). The comparison of the models can be seen in the table 19 as follows:

Table 19. Method comparison table

Metric	OLS	Semi-log	Double-log	ANN
RMSE	59 266	53 991	51 779	42 968
MAE	46 383	39 203	38 316	31 617
Multiple R-squared	0.8293	0.8524	0.858	-
Adjusted R-squared	0.8272	0.85	0.8558	-
Accuracy of the model (%)	-	-	-	-
±10%	41	45	45	53
±20%	69	75	77	84
±25%	77	83	85	89

As it can be seen from the table 19, the ANN model has the lowest RMSE and MAE of all the models. The highest r-squared and adjusted r-squared values were obtained by the double-log regression model. The most accurate model in 10%, 20% and 25% accuracy levels % was ANN. This means that ANN predicted real estate prices fell more often inside the presented accuracy levels than other models. In all the factors OLS performed the worst. The performance comparison of the models can be seen in the table 20:

Table 20. Order of performance

					Accuracy of the model (%)		
	RMSE	**MAE**	**Multiple R-squared**	**Adjusted R-squared**	**±10%**	**±20%**	**±25%**
1st	ANN	ANN	Double-log	Double-log	ANN	ANN	ANN
2nd	Double-log	Double-log	Semi-log	Semi-log	Double-log, Semi-log	Double-log	Double-log
3rd	Semi-log	Semi-log	OLS	OLS	OLS	Semi-log	Semi-log
4th	OLS	OLS	-	-	-	OLS	OLS

The comparison of the models indicate that the ANN model is the most accurate way to predict apartment prices in the Helsinki housing data in various meters. The second most accurate model is the double-log regression model, the third most accurate is the semi-log regression model. OLS regression was last. However, based on the previous literature, it can be concluded that there are limitations to the use of the ANN model. These limitations are based on the poor transparency of the model as well as the difficulty in explaining the values of the model. The model was previously seen in the literature as a so-called black-box model (Lenk Margarita M et al., 1997; Limsombunchai et al., 2004; Stanley et al., 1998; William et al., 2012).

In this study the ANN was composed by doing fewer cross-folding loops than one might have wished. It would have been important to be able to do the analysis with more computing speed as the limitation in computing speed did affect the ANN prediction accuracy measuring. The cross-folding should have been done several dozen times with the same number of neurons as ANN model produces slightly different results every time it is used with same number of neurons. It would have also important to perform the ANN process multiple times with many different network sizes. Both the neurons in each layer, and the number of layers in the model should have been modified. The same limitations in computing power were not encountered when using regression models.

3. Which variables are the most significant variables in valuing an apartment based on this study as well as previous literature?

Based on previous literature, it was found that important variables in determining the value of apartments were house characteristics, such as the size, age, location, parking space, number of rooms, view and height of the apartment. External variables such as air quality, state of the economy, increase in services in the region. Spatial variables such as the apartment's distance to streams, rivers, parkland, running trails, sports halls, shopping malls, mountains. The most important variable was location based on previous literature. (Bartik, 1988; Gaetano, 2019; Hasanah & Yudhistira, 2018; Karaganis, 2011; Ridker & Henning, 1967; Sander & Polasky, 2009; Thanasi (Boçe) Marsela, 2016; Warren Clive M J et al., 2017)

In this research the important variables in the apartment valuation process were found in a stepwise regression process for both the ANN model and the multiple regression model. P-value under 2e-16, was given to the following variables: number of squares, income level, land ownership (dummy), good condition of the dwelling (dummy).

The grand district dummy variables with P-value under 2e-16 were Southeast, South, Central and West. The Basic District dummy variables with P-value under 2e-16 were: Mellunkylä, Pitäjänmäki, Kaarela, Puistola. Thus, it can be generalized that significant apartment characteristic variables in this data were living space, land ownership, condition. The significant external characteristics were income level of the living area and parts of grand and basic districts.

5.2 IMPLICATIONS FOR THE INDUSTRY

This section explores what kind of concepts and thoughts real estate companies and other agents may implement to their processes from this research. Previous literature indicates that the ANN is able to

find more and more repeated hidden formulas as well as insights in the data when the amount of data is increased. However, regression models, for example, do not improve significantly with increasing data (Peterson & Flanagan, 2009). Several real estate operators could take advantage of artificial neural network models good performance in apartment valuation. However, in order for these companies to be able to take advantage of these insights from the model, the data collecting should be both standardized and applied with better methods.

The data should be collected in highly standardized quantitative format. For example, written notation such as 1h + k + kp + s should be replaced with forms that would be clicked on and not written, as in the following table 21:

Table 21. Standardized data collection form

Characteristic	0	1	2	3	4	5
Number of toilets			x			
Number of rooms						x
Number of windows					x	
parking space		x	-	-	-	-
Own plot	x		-	-	-	-

Apartment plus code	**Enter the address**
5W9H+5F Helsinki	Runeberginkatu 2, 00100 Helsinki

There would be several benefits to such a labeling approach. First there would be fewer typing errors in the labeling process. The labeling would not be time consuming as clicking the right sections is quick. In this way, increasingly accurate property variables example estimation of the condition of the floor, the number of windows, the height of the room could also be measured and analyzed.

Labeling the full location information with example coordinates, could be used to focus more closely on the most important variables determining the price of an apartment, such as the analysis of spatial data. The table 21 is just a rough beta type of data collection form but in this the table there is a space for the coordinates for the apartment. The effect of variables such as apartment's distance to Streams, Rivers, parkland, running trails, sports halls, shopping malls, mountains on the value of the apartment could be measured and analyzed in more detail if more accurate coordinate data would be collected.

When the list would be filled, it could be converted directly to numeric format as well as text file. In this case, the information would not be lost in the conversion process as it was found during the process of this research that could easily happen.

Standardized data collected this way could benefit many operators. The collector of standardized housing data could rent the data for use of external parties, such as banks, real estate investor or real estate agencies. Banks for example would be able to make very insightful collateral calculations for their customers loans, and real estate companies to be able to value homes more accurately.

5.3 LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

Spatial analysis was seen in previous literature as one of the most important aspects in valuing residential real estate. However, due to data limitations, the location analysis was done using only dummy variables, although it would have been much more accurate to do this analysis utilizing more accurate spatial analysis methods. The data used in the modeling should be of both spatial and time-series qualities. New variables such as economic indices as well as, for example, crime statistics could also be added to improve data predictability.

References

Abidoye, R. B., Ma, J., Lam Terence Y M, Oyedokun, T. B., & Tipping, M. L. (2019). Property valuation methods in practice: Evidence from australia. *Property Management, 37*(5), 701-718.

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.

Aurélio Stumpf González Marco, Lucio, S., & Torres, F. C. (2005). A new approach to spatial analysis in CAMA. *Property Management, 23*(5), 312-327.

Bartik, T. J. (1988). Measuring the benefits of amenity improvements in hedonic price models. *Land Economics, 64*(2), 172-183.

Borst, R. A. (1991). Artificial neural networks: The next modelling/calibration technology for the assessment community. *Property Tax Journal, 10*(1), 69-94.

Bruce, T. (1994). The valuation of resort condominium projects and individual units. *Journal of Property Valuation and Investment, 12*(4), 9-36.

Chun, L. C., & Mohan Satish B. (2011). Effectiveness comparison of the residential property mass appraisal methodologies in the USA. *International Journal of Housing Markets and Analysis, 4*(3), 224-243.

Ciaburro, G., & Venkateswaran, B. (2017). *Neural networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles* Packt Publishing Ltd.

Cowling, K., & Cubbin, J. (1972). Hedonic price indexes for united kingdom cars. *The Economic Journal, 82*(327), 963-978.

Digital and Population Data Services Agency. (2020). *Helsinki metropolitan postal code areas*.https://www.opendata.fi/data/en_GB/dataset/paakaupunkiseudun-postinumeroalueet/resource/b989b251-6190-44f7-87cf-2be0413cb5c1

Diwan, S. A. (2019). Proposed study on evaluating and forecasting the resident property value based on specific determinants by case base reasoning and artificial neural network approach. *Indonesian Journal of Electrical Engineering and Computer Science, 17*(3), 1467-1473.

Do, A. Q., & Grudnitski, G. (1992). A neural network approach to residential property appraisal. *The Real Estate Appraiser, 58*(3), 38-45.

Doszyń Mariusz. (2020). Algorithm of real estate mass appraisal with inequality restricted least squares (IRLS) estimation. *Journal of European Real Estate Research, 13*(2), 161-179.

Eija, K. (2004). Artificial neural networks in analytical review procedures. *Managerial Auditing Journal, 19*(2), 191-223.

Elli, P., Vassilis, A., Thomas, H., & Nick, F. (2003). Real estate appraisal: A review of valuation methods. *Journal of Property Investment & Finance, 21*(4), 383-401.

Finnish tax administration. (2020). *Verohallinnon tilastotietokanta*.http://vero2.stat.fi/PXWeb/pxweb/fi/Vero/Vero__Henkiloasiakkaiden_tulover ot__lopulliset__postinum/postinum_104.px/

Fritsch, S., Guenther, F., & Guenther, M. F. (2016). Package 'neuralnet'. *The Comprehensive R Archive Network,*

Gaetano, L. (2019). Property valuation: The hedonic pricing model – location and housing submarkets. *Journal of Property Investment & Finance, 37*(6), 589-596.

Greenhalgh, P. M., & Soares, B. R. (2015). An investigation of development appraisal methods employed by valuers and appraisers in small and medium sized practices in brazil. *Journal of Property Investment & Finance, 33*(6), 530-547.

Günther, F., & Fritsch, S. (2010). Neuralnet: Training of neural networks. *The R Journal, 2*(1), 30-38.

GUPTA, P., & SINHA, N. K. (2000). CHAPTER 14 - neural networks for identification of nonlinear systems: An overview. In N. K. SINHA, & M. M. GUPTA (Eds.), *Soft computing and intelligent systems* (pp. 337-356). San Diego: Academic Press.

Hajnal, I. (2014). Continuous valuation model for work-in-progress investments with fuzzy logic method. *Procedia Engineering, 85*, 206-213. doi:https://doi.org/10.1016/j.proeng.2014.10.545

Hasanah, A. N., & Yudhistira, M. H. (2018). Landscape view, height preferences and apartment prices: Evidence from major urban areas in indonesia. *International Journal of Housing Markets and Analysis, 11*(4), 701-715.

Hill, R. C., Griffiths, W. E., & Lim, G. C. (2018). *Principles of econometrics* John Wiley & Sons.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679-688.

Isakson, H. (2002). The linear algebra of the sales comparison approach. *Journal of Real Estate Research, 24*(2), 117-128.

Jimmy Pang. ***History: The 1940's to the 1970's***.https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history1.html

Johansson, E. M., Dowla, F. U., & Goodman, D. M. (1991). Backpropagation learning for multilayer feed-forward neural networks using the conjugate gradient method. *International Journal of Neural Systems, 2*(04), 291-301.

Joshua, A. O. (2014). Critical factors determining rental value of residential property in ibadan metropolis, nigeria. *Property Management, 32*(3), 224-240.

Kalogirou, S. (2014). Designing and modeling solar energy systems. Retrieved from https://www.sciencedirect.com/book/9780123972705/solar-energy-engineering#book-description

Karaganis, A. (2011). Seasonal and spatial hedonic price indices. *Journal of Property Investment & Finance, 29*, 297-311.

KVKL. (2020). *Asuntojen hintatieto palvelu*.https://asuntojen.hintatiedot.fi/haku/

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8595-8598.

Lenk Margarita M, Worzala Elaine M, & Ana, S. (1997). High-tech valuation: Should artificial neural networks bypass the human valuer? *Journal of Property Valuation and Investment, 15*(1), 8-26.

Limsombunchai, V., Gan, C., & Lee, M. (2004). House price prediction: Hedonic price model vs. artificial neural network. *American Journal of Applied Sciences, 1*

Makridakis, S., & Hibon, M. (1997). ARMA models and the Box–Jenkins methodology. *Journal of Forecasting, 16*(3), 147-163.

McCluskey, W. J., McCord, M., Davis, P., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: A comparison of modern approaches. *Journal of Property Research, 30*(4), 239-265.

Michael, M., Bourassa Steven C, Martin, H., & Donato, S. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research, 12*(1), 134-150.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* John Wiley & Sons.

Mora-Esperanza, J. G. (2004). Artificial intelligence applied to real estate valuation: An example for the appraisal of madrid. *Catastro.April,*

Pagourtzi, E., Metaxiotis, K., Nikolopoulos, K., Giannelos, K., & Assimakopoulos, V. (2007). Real estate valuation with artificial intelligence approaches. *International Journal of Intelligent Systems Technologies and Applications, 2*(1), 50-57.

Paul, G., Michael, F., & Matthew, C. (1996). Modelling the influence of location on value. *Journal of Property Valuation and Investment, 14*(1), 6-19.

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

Pierluigi, M., Francesco, T., & Marco, L. (2018). Multicriteria analysis and genetic algorithms for mass appraisals in the italian property market. *International Journal of Housing Markets and Analysis, 11*(2), 229-262.

Pinder, J. P. (2016). *Introduction to business analytics using simulation* Academic Press.

Ridker, R. G., & Henning, J. A. (1967). The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics, 49*(2), 246-257.

Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. *Computer Standards & Interfaces, 16*(3), 265-278.

Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *IEEE International Conference on Neural Networks,* pp. 586-591.

Ripley, B., Venables, W., & Ripley, M. B. (2016). Package 'nnet'. *R Package Version, 7*, 3-12.

Rocco, C., Elena, F., & Patrizia, S. (2015). Listing behaviour in the italian real estate market. *International Journal of Housing Markets and Analysis, 8*(1), 97-117.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34-55.

Rossini, P. (1997). Application of artificial neural networks to the valuation of residential property. *Third Annual Pacific-Rim Real Estate Society Conference. Palmerston North, New Zealand,*

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536.

Sander, H. A., & Polasky, S. (2009). The value of views and open space: Estimates from a hedonic pricing model for ramsey county, minnesota, USA. *Land use Policy, 26*(3), 837-845. doi:https://doi.org/10.1016/j.landusepol.2008.10.009

Singla, H. K., & Priyanka, B. (2019). Factors affecting rentals of residential apartments in pune, india: An empirical investigation. *International Journal of Housing Markets and Analysis, 12*(6), 1028-1054.

Stanley, M., Alastair, A., Dylan, M., & David, P. (1998). Neural networks: The prediction of residential values. *Journal of Property Valuation and Investment, 16*(1), 57-70.

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

Thanasi (Boçe) Marsela. (2016). Hedonic appraisal of apartments in tirana. *International Journal of Housing Markets and Analysis, 9*(2), 239-255.

Wang, D., Li, V. J., & Yu, H. (2020). Mass appraisal modeling of real estate in urban centers by geographically and temporallyweighted regression: A case study of beijing's core area. *Land, 9*(5)

Warren Clive M J, Peter, E., & Jason, S. (2017). The impacts of historic districts on residential property land values in australia. *International Journal of Housing Markets and Analysis, 10*(1), 66-80.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, , xiii-xxiii.

William, M., Peadar, D., Martin, H., Michael, M., & David, M. (2012). The potential of artificial neural networks in mass appraisal: The case revisited. *Journal of Financial Management of Property and Construction, 17*(3), 274-292.

Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research, 33*