ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

The gap between WOZ-waarde and market value on the Dutch housing market

Predicting the relative gap between the WOZ-waarde and the self-reported market value of a house on the Dutch housing market for 2008 till 2021

Name student: Rins Lukasse

Student number: 503338

Supervisor: dr. A Archimbaud

Second assessor: dr. SL Malek

Final version: 20-7-2022

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

The main goal of this research is to investigate the predictability of the relative gap between the self-reported market value and the WOZ-waarde in the Netherlands. The buyer-seller relationship in the Dutch housing market will improve when it becomes clearer which factors cause the gap. This paper focuses on the relative gap as the absolute gap is expected to differ across price categories on the market. A sub-question will zoom in on the irrational side of the relative gap and look into the effect of market sentiment by predicting house prices. The dataset used in this paper contains panel data for 2008 till 2021 with only regular Dutch houses. The variables include house characteristics, information on the inhabitants, time effects and macro-economic variables. Investigating the relative gap's predictability was done using a neural network and a decision tree operated as a surrogate model. The results showed that only 16.07% of the predictions fell within the 25% error bandwidth. *Household Income*, *WOZ-waarde* and the *Purchase price* are essential features to come to these predictions. House prices, and with that market sentiment, on the other hand, are easier to predict. Predicting the house prices happened primarily based on the variables *Inflation* and the *Supply of houses* in the Netherlands. Future research should implement the actual market value instead of the self-reported version used in this paper.

Keywords: Housing market, WOZ-waarde, self-reported market value, relative gap, market sentiment

Contents

1. Introduction

Understanding what a house is worth is quite challenging in the Netherlands nowadays. Both for buyers and sellers, it is essential to know and understand the price of a particular place. The Valuation of Immovable Property Act (In Dutch: 'WOZ-waarde') indicates what a house should be worth. But in reality, the actual price for which a home is sold on the market is often higher or even much higher than what should be expected looking at its WOZ-waarde. Is there a structural gap between the WOZ-waarde and the self-reported market value of a house, and if so, is it possible to predict this gap?

So, the main research question of this paper will be as follows:

> 'Is there a predictable gap between the WOZ-waarde and the self-reported market value of a house?'

Some literature by Lubberink, Post, and Veuger (2017) investigates if the WOZ-waarde can be used to indicate market value. They looked into the real estate owned by the Dutch government when evaluating the gap between the WOZ-waarde and the actual value. Their conclusion is that the WOZ-waarde is not a good indicator of market value. This paper will use ordinary houses owned by the Dutch people to predict the gap. It is unclear in the literature which factors might cause or at least play a role in this gap. From a governmental point of view, this is also exciting research. It is in their interest to establish the WOZ-waarde as good as possible. Better performance will result in clearer taxation of properties. Knowing that the WOZ-waarde of a property is a correct reflection of its market value and, therefore, its sale price could also create a housing market that is easier to understand for the Dutch population. It needs to be said that the WOZ-waarde should not be seen as an instrument that 'makes the market'. Waarderingskamer (2019) clearly states that the WOZ-waarde should follow the market. This statement suggests a lagging time effect between the WOZ-waarde and actual market value. Knowing that, in general, house prices increase over time leads to the first hypothesis that there is a positive gap between the self-reported market value and the WOZ-waarde of a dwelling. So, self-reported market value is expected to be higher than the WOZ-waarde of a house.

From a marketing perspective, the existence of this gap or the uncertainty about it might impact the buyer-seller relationship in the Dutch property market. Doney and Cannon (1997) explain that the trust-building process between buyers and sellers will depend on five different characteristics. For a housing market where a specific property, in the short term, is only sold once and where the buyer and the seller only meet once, two characteristics that Doney and Cannon (1997) use are most important, namely 'Calculative' and 'Transference'. In a housing market with uncertainty about the rightness of the WOZ-waarde and a gap between the actual market value and the WOZ-waarde 'Calculative' means that a buyer will spend more time overthinking the costs of a seller acting in an untrustworthy manner. Transference implies that in case of higher risks and unclarity, the buyer draws on proof sources to create some trust in the seller. So, clarifying the gap between the WOZ-waarde and the market value will lead to a housing market where buying and selling a property is more manageable. One should understand that changing the WOZ-waarde has some consequences as there is a connection between the WOZ-waarde and the taxes house owners need to pay. A higher WOZ-waarde leads to higher taxes which the government might like, but house owners, on the other hand, will not be happy if their WOZ-waarde increases.

The gap being predictable will create this extra clarity in the market. But answering this research's subquestions will also help remove some of the uncertainty. Answering the first subquestion will lead to a better understanding of why this gap is even existing. The second subquestion tries to determine what caused price changes in the housing market and how this influenced the gap between the WOZ-waarde and the self-reported market value.

The following sub-questions will be used to answer the main research question as good as possible:

- Which factors result in the fact that the gap exists?
- What is the impact of market sentiment on this gap?

The second sub-question mainly focuses on the irrational side of this problem. Market sentiment and human behaviour could also impact the size of the gap, but this might also be harder to capture in a particular variable. It will be interesting to see how the average price in the housing market reacted to the financial crisis and the Covid19 crisis. These events could lead to buyer or seller behaviour changes that eventually influence the gap between the WOZ-waarde and the self-reported market value. Understanding the factors that cause price changes will also help understand the overall market and provide possible explanations for why this gap exists. The second hypothesis is that more extreme market sentiment will lead to a gap between the WOZ-waarde and the market value that is harder to predict.

Market sentiment will be investigated using overall house prices in the Netherlands as the dependent variable. A random forest will be used to look into the different variables that cause price changes. A decision tree will, later on, be used as extra proof. Section 2 will explain the literature relevant to answering the research questions. Section 3 gives insides into the dataset that is used in this research. Section 4 shows the conceptual framework of this investigation, and section 5 explains the results. Lastly, section 6 will discuss the conclusions and limitations of this research.

2. Literature

WOZ-waarde

Lubberink, Post, and Veuger (2017) asked themselves if the WOZ-waarde is useful as an indicator of market value. They used government property transferred in the period 2010 up to and including 2015 to investigate this question. It is vital to notice that this government property consists of many different assets like offices, monumental buildings, houses and objects that might be hard to taxate correctly, like prisons. A certain margin between the WOZ-waarde and the market value was accepted before denying the WOZ-waarde as an indicator. For houses, for example, they allowed the WOZ-waarde to be 12.5% higher or lower before rejecting it. Their paper explicitly mentioned that any differences within this error bandwidth are due to 'normal taxation uncertainty'. The two main conclusions of Lubberink, Post, and Veuger (2017) are that the WOZ-waarde should not directly be used as an indicator of market value because the difference between these two is too big even when using the earlier mentioned margin of error. Their second conclusion is that this gap mainly exists because of differences in interpretation. For each area, a different municipality is responsible for real estate taxation. Although the big picture is the same, each municipality might give specific minor property characteristics a different weight in the taxation.

So, Lubberink, Post, and Veuger (2017) mention that these slight dissimilarities in interpretation might also cause this gap to exist. They also recommend that other researchers look deeper into the time effects because the WOZ-waarde might be a lagging indicator. There might be some time between movements in the market and the WOZ-waarde following these movements. Lubberink, Post, and Veuger (2017) express their concerns about the importance of market sentiment in this context. The WOZ-waarde might not correctly predict or consider the differences in market sentiment. This paper will further investigate these time effects and the impact that market sentiment has on the difference between the WOZ-waarde and the actual value of a house.

One of the main questions that need to be answered first is 'Which factors are represented in the WOZ-waarde?' Waarderingskamer (2019) provides information on what should be represented by the WOZ-waarde and what not. They divide the characteristics into two groups' primary characteristics' and 'secondary characteristics'. Primary characteristics are all the physical attributes that are objectively measurable, such as square feet, year of construction, location, and the presence of a garage or dormer window. The secondary characteristics contain all things that are more debatable, among them a comparison with other houses in the neighbourhood and information about the house's energy efficiency.

Both primary and secondary characteristics come with the following statement: “Which primary / secondary characteristics are registered, tracked and involved in the taxation is partly dependent on market analyses. After all, it is the market that decides which attributes are relevant for the value of a house” (Waarderingskamer 2019). The partly dependence means that each municipality is obligated to include the dwelling’s primary characteristics, which are also included in the ‘basic register of addresses and buildings’ (In Dutch: BAG). The municipalities do not have any obligations regarding the secondary characteristics. So, there is some room for the different local councils to do what they see fit based on market analyses. It is the job of municipalities to provide each house with its own WOZ-waarde. Although, this does not mean that each municipality works on its own to do this. Some municipalities formed a partnership to solve this task. This grouping results in the fact that the Netherlands have 345 municipalities but approximately 160 locations where the WOZ-waarde is calculated for a certain area with houses. It is important to note that the WOZ-waarde was never intended to perfectly reflect the actual market value of a house. The WOZ-waarde should follow the market and not make the market. This again proves that a time lag between the WOZ-waarde and the market value should be expected (Waarderingskamer 2019).

Housing market

Before diving into any research, it is essential to understand the context of the research question. This section will look into the literature on the housing market to see if there is anything special in this market that needs to be taken into account when doing the actual research. Ortalo-Magne and Rady (2006) built a model to predict housing prices. In this model, ‘the ability of young households to afford a down payment on a starter home’ is identified as a new driver of housing prices. To ensure that the different groups in the population are represented correctly, consumers are divided into four categories: ‘first-time buyers’, ‘credit-constrained repeat buyers’ and ‘unconstrained households that might move due to preference reasons’. The models show that there will be differences between these groups. The house prices could, for example, overshoot for a specific group but react more normally for another group. It will be essential later on to take into account these different groups that operate in the same market. The potential gap between the WOZ-waarde and the market value of a house is not necessarily equal for the various groups that Ortalo-Magne and Rady (2006) identified.

Which factors cause the price of a house to change? The most logical one is a change in the fundamentals, the demand for and supply of homes. House prices might not always change due to their underlying fundamentals. History provides us with examples of ‘housing bubbles’ because of which prices became unreasonably high. Stiglitz (1990) provided a general definition of asset bubbles in his journal: “If the reason that the price is high today is only because investors believe that the selling price is high tomorrow and fundamental factors do not seem to justify such a price then a bubble exists." These housing bubbles are a great example of extreme market sentiment that might make it harder to predict and understand the gap between the WOZ-waarde and the self-reported market value of a dwelling.

It is essential to realise that people who need a house to live in are not the only players in this market. Stiglitz (1990) uses the term 'investors' in his definition, which shows that there are also people on the housing market who are there just to earn money. Himmelberg, Mayer, and Sinai (2005) also looked into the correctness of house prices and possible bubbles. They mainly compared the price of a dwelling to the cost of ownership and its location. Himmelberg, Mayer, and Sinai (2005) conclude that house price dynamics are a local phenomenon and that there are crucial economic differences among cities. The second thing they stress is that the annual cost of ownership is a relevant comparison when one wants to compare one house price with another house price. When investigating the difference between market value and the WOZ-waarde, it might also be relevant to include data about the house's location and the ownership cost to create a complete picture of the situation.

Another interesting thought is changing house prices and their impact on consumption patterns for different age groups. Campbell and Cocco (2007) investigated this relationship between house prices and consumption patterns by using a dataset from the UK. They found that there is indeed an impact of house prices on overall consumption, but this effect is different between age groups. The most noticeable effect of house prices on consumption occurs when talking about older homeowners. On the other hand, the most negligible effect, insignificantly different from zero, occurred for the group of younger people. Ortalo-Magne and Rady (2006) already mentioned that there are different groups in the housing market, and age might be a good predictor to find out in which group or stage an individual is right now.

Veldhuizen, Vogt, and Voogt (2016) investigated the Dutch housing market and mainly looked at 2004 until 2015. They studied the relationship between Google search data of the word 'mortgage' and the actual transaction data. The results showed that last month's search data is positively and significantly related to the current month's transactional data. This conclusion proves that online search data might become a more and more relevant aspect of investigating real market behaviour. In this research, online search data is not included as a determinant to predict the gap. However, for further study, it would be great to take this approach and find out if online data will help in getting even more accurate predictions.

A second aspect which is underexposed in this research but presented in the literature is the variation among regions, cities and neighbourhoods in the Netherlands. Hochstenbach and Arundel (2019) were primarily interested in the trends of spatial housing-market polarisation, which they defined as increasing disparities between more and less expensive neighbourhoods in terms of house values. They concluded that there was increasing polarisation during the 2006-2018 period, meaning that there were uneven housing-market developments across time and space in the Netherlands. The locational data of the houses is unfortunately not present in the dataset of this research. However, it is crucial to keep in mind that differences in the location might also result in a different gap between the WOZ-waarde and the self-reported market value for two houses with similar characteristics and inhabitants.

House price dynamics & Macroeconomic determinants

Englund and Ioannides (1997) investigated the drivers behind house prices and compared 15 OECD countries, including the Netherlands. They found, first of all, a remarkable degree of homogeneity in the drivers of house prices for these fifteen countries. This finding implies that the big picture of a housing market is the same for each country. Therefore, conclusions of research that happened abroad might also be relevant to the housing market in the Netherlands. Englund and Ioannides (1997) first find that the GDP growth rate is enormously significant as a driver of the house price. A second macroeconomic determinant that helped the researchers predict house prices was the inflation rates. After seeing many similarities among the housing markets of these 15 OECD countries, Englund and Ioannides (1997) tried to answer the question if there is one international housing market with its own macroeconomic cycle. In the end, they conclude that their results do not show any significant evidence for this. As a reason, they suggest that there still might be policy differences among the countries by which there are still minor dissimilarities between countries.

To investigate these policy differences and the government's impact on the market, it becomes necessary to zoom in and look specifically at the situation in the Netherlands. Boelhouwer and Hoekstra (2009) write that the Dutch housing policy is quite developed in terms of money and instruments. However, there are also some problems resulting in the housing market not being as effective as possible. As examples of these problems, they mention 'housing shortages in areas of economic growth', 'high property prices' and segments of the Dutch population that experience' accessibility and affordability issues'. The government's actions impact both the demand side and the supply side. On the demand side, there is strong support from the government via mortgage interest reliefs for owner-occupiers and rent allowances for tenants. On the other hand, the supply side is not stimulated but counteracted by regulations and restrictions that are hampering the production of dwellings.

Boelhouwer and Hoekstra (2009) do not think that this government interference in the market is wrong per se. In such a small, densely populated country, it makes sense in their opinion that a government does some spatial planning and influences the market by doing so. Demand and supply being out of tune on the Dutch housing market both has a qualitative side: 'The characteristics of the available dwellings do not match the characteristics of the dwellings that home seekers desire' and a quantitative side: 'There are simply fewer homes than households'. Boelhouwer and Hoekstra (2009) show multiple solutions that could better the situation. Which one is best is mainly a political discussion which is not that interesting for now. The most relevant conclusion of this research is that the circumstances in the Dutch housing market are far from ideal. The problems that occur in the market need to be kept in mind when interpreting any results later on.

Genesove and Mayer (2001) also looked into the price dynamics in the housing market. They mainly focused on loss aversion and the effects of loss aversion on selling behaviour. The paper shows crucial insights into how we should interpret the market. Using the Boston housing market as an example, the researchers conclude that a real estate market is not a perfect market because they found during their research that a price decrease also resulted in a reduction of volume in the market. Economic principles of demand and supply suggest that more people should be willing to buy a dwelling if the price becomes lower, but this was not the case in Boston's real estate market.

Two effects cause this simultaneous occurrence of decreasing prices and decreasing volume, according to Genesove and Mayer (2001), namely 'loss aversion' and 'equity constraints'. Loss aversion results in people not wanting to sell their property at a price lower than their purchase price or at a price that they feel is too low. Loss aversion is one mechanism that results in decreasing volume when prices drop. In case of lower demand, it becomes harder to find someone who likes the characteristics of the property that is on the market. So, equity constraints further enhance the spiral of decreasing prices and decreasing volume. Genesove and Mayer (2001) end their paper with the remark that they suspect that there might be a third mechanism at play here. They do not show any proof, but they suspect that there could be a lagging adjustment of seller behaviour to new market conditions.

The marketing of a house

Buying or selling a house is a much more complicated and important procedure than buying or selling most other goods. One reason for this to be true is that one often needs to borrow money to buy a house. Which relationship is there between the access to and the cost of credit on one side and the overall house prices on the other? There has been much debate on this topic, but the general conclusion is that credit being more accessible and lower cost of credit will both increase house prices (Adelino, Schoar, and Severino 2012). But as always, such a rule also has its exceptions. When the supply of houses is elastic for a certain area, the supply of houses will start increasing when getting money from the bank becomes easier for buyers. In most instances, however, the supply of dwellings on the market will not change very quickly due to this exogenous factors like credit supply, and a price change will happen instead (Favara and Imbs 2015).

As mentioned in the introduction, there might be some unclarity in the housing market stemming from how a house's purchasing process is structured in the Netherlands. It is not mandatory, but there is, for example, the option to hire a real estate agent. A real estate agent is often the first one that knows which houses will go on sale, so using the help of a real estate agent will most likely result in a quicker purchasing process. After finding a place that matches the buyer's wishes, it is time to put in an offer. This is the moment it gets a bit complicated and unclear because the list price of a dwelling is just a starting point. So, the indicated price could both be too high or too low compared to the actual value. Overbidding and underbidding the listing price are also both allowed. For this reason, there is always some unclarity in the market about the fair price of a dwelling. After placing an offer, the seller can reject, accept or place a counteroffer. So, here is some negotiation going on. When an offer is accepted, the only steps that are left are applying for a Dutch mortgage, getting the necessary house insurances and signing the deeds at the notary (Hanno 2022).

Game theory and asymmetric information become applicable to the housing market when bidding on houses becomes normal. Asymmetric information in this context means that the buyer of the dwelling does not have the same kind or amount of information as the seller. Game theory shows different scenarios and negotiation styles in which both the buyer and the seller try the reach the outcome that is most optimal for them. Olaussen, Oust, and Sønstebø (2018) investigated different kinds of bidding behaviour under different market regimes in the Norwegian housing market. They compared different Norwegian cities. In some of these cities, house prices were increasing pretty quickly, while in others, prices were decreasing. They concluded that respondents have similar perceptions regarding strategic bidding measures, such as bid size and time factors, regardless of the market's situation. A weakness of this research is that there is no split between possible over- and underbidding. So, the Dutch market might cause different behaviour than the Norwegian market.

Aspects that play an essential role in the marketing of a house on the seller's side are setting a starting price. According to Anglin, Rutherford, and Springer (2003), there will always be a trade-off between a dwelling's selling price and its time on the market. The trade-off means that the seller can either choose to sell at a higher price or sell in less time. So what are the overall effects of price setting, and how does it influence the price for which the house is eventually sold? Using a high degree of overpricing when putting a dwelling on the market signals that one is patient with selling the property, so it will give a stronger position when there are negotiations about the price later on. A drawback is that potential buyers might lose interest when seeing the higher price and not even try to negotiate. Anglin, Rutherford, and Springer (2003) state that the most significant danger of setting a high price is that it will remain on the market for an overly long time and eventually become stigmatised. In practice, a house that becomes stigmatised means that current potential buyers assume that previous potential buyers found something wrong with the dwelling. Therefore buyers become distrusting or reluctant toward the property and its seller.

Intuitively, it seems most wise to start with a high price and slowly lower it when no one shows interest. By using this strategy, the seller should be able to get as much as possible for its property and still sells it before it becomes a stigmatised house. Knight (2002) used a different model to study the relationship between the listing price, time on the market and the selling price. Compared to the study of Anglin, Rutherford, and Springer (2003), the effect of price changes is also included in the model this time. Knight (2002) eventually concludes that mispricing the dwelling in the initial listing price is costly to the seller in time and money. Houses with substantial percentage changes in price take longer to sell and eventually sell at lower prices.

Predicting property prices

This research will try to predict the gap between the WOZ-waarde and the actual market value of a particular property. But is this something that has been done before? There are no examples that tried to predict this gap, but there are some papers in which the researchers attempted to predict property prices. It will be interesting to see which methods they used to do this and which variables they included to come to accurate predictions. Dubin (1998) mentions in his paper that the simplest way of predicting the price of a house is using Ordinary Least Squares as the statistical technique. All the property's characteristics should be used as independent variables in this regression. Dubin (1998) go on with saying that this approach misses a valuable source of information which are the neighbourhood effects. Each broker knows that the price of a dwelling hugely depends not only on its characteristics but also on the quality of the neighbourhood and on the prices of properties with which it has much in common. Dublin (1998) tries to incorporate this information into the equation by including a matrix K built of all correlations between the observations. Afterwards, he concludes that this approach leads to improved coefficient estimates.

Čeh et al. (2018) also tried to find a method that resulted in accurate predictions of apartment prices. They used price data of apartments in the Slovenian capital Ljubljana from 2008 to 2013. Their main goal was to compare the widely used multiple regression with a Random Forest model. It is interesting to note that they did not simply include all 36 variables in the regression but firstly built different topics by using Principal Component Analysis. So, the regression makes predictions by using those topics as independent variables. The main advantage of this approach is preventing multicollinearity between variables. Evaluating both the Random Forest and the regression based on R^2 values, the Mean Absolute Percentage Error (MAPE) and the coefficient of dispersion (COD) showed that the Random Forest made better predictions than the regression. Čeh et al. (2018) explain that this difference in the accuracy of the models might be caused by the non-linear nature of the prediction task and substantial price changes during specific years in the dataset. Both models performed best when predicting the price of an average apartment. The models' predictions overestimated the lower prices of flats and underestimated the higher prices of apartments. Ho, Tang, and Wong (2020) again try to predict the prices of houses, now in Hong Kong. This time they use a Random Forest (RF) again and compare this with even more complicated machine learning methods like a Support Vector Machine (SVM) and a gradient boosting machine (GBM). When discussing the entire dataset, the methods RF and GBM resulted in better predictions. It is crucial to note that this does not always mean that these methods outperform SVM. SVM has its own advantages. SVM performed, for example, better than the other methods when the researchers zoomed in on predictions within a tight time constraint. It remains challenging to give a clear answer to the question of which method will result in the best predictions. All these previous examples used multiple methods that they considered promising and compared the results afterwards. This approach might be the best way to come to excellent conclusions and insights.

3. Data

The first dataset used for this research is freely available via the LISS Panel (Scherpenzeel and Das 2010). This data contains a total of 14 waves, of which the first one was gathered in 2008 and the last one in 2021. The sum of observations ranges from 2,500 up to 5,000 per wave. Some people responded each year, and others only one or two times. The dataset contains self-reported data about the market value for which it is or could be sold at that point in time. It also provides the WOZ-waarde and characteristics like the number of floors and rooms, location and type of dwelling. This dataset makes it possible to further look into any factors that are perhaps not included in the WOZ-waarde. For example, any environmental problems, satisfaction levels about the house and the neighbourhood, or problems with the house itself. This dataset corresponds with the variables 1 up to 29 of the variable list in Appendix A.

A second dataset that will be used contains all kinds of background information on the people who filled out the survey. This second dataset contains, for example, information on gender, position in the household, financial situation, education and country of origin. This dataset is also freely available via the LISS Panel (Scherpenzeel and Das 2010). This second data source resulted in variables 30 up to 41 of the variable list in Appendix A.

The third dataset comes from CBS (CBS 2022). This government institute collects all kinds of data about the Netherlands. Any information on macro-economic market movements or the Dutch population can be gathered from this source. This information will mainly be helpful when looking into the impact of market sentiment on the WOZ-waarde and the gap between WOZ-waarde and the market value. When answering this sub-question, numbers on inflation, consumer confidence, GDP, and information on price changes in the Dutch housing market will all be relevant. All data necessary to account for specific time effects were also firstly collected by researchers of the CBS (CBS 2022). So, numbers that capture the impact of Covid19 and the impact of the financial crisis are also from this source. The total number of variables obtained from the CBS can be found in the variable list in Appendix A under the numbers 42 till 55. All these variables that are indexed take 2015 as their base year.

It is crucial to note that the context of this dataset and this research is the Netherlands and the period 2008 till 2021. So, investigating a different country or another time might lead to non-identical conclusions. The only properties in the dataset that is used are regular houses. So, there is, for example, no government-owned real estate in this dataset as Lubberink, Post, and Veuger (2017) used.

Data Cleaning

Before building any model or inspecting any relationship, it is crucial to investigate how the data looks. First, all values that represented answers as *'I do not know'* or *'I do not want to say'* were changed to NA-values. The second step was filtering for any unrealistic values that were still there. It is pretty arbitrary which value is still realistic and where it becomes unrealistic. In the end, the following decisions were made: Market value and WOZ-waarde only capture the value's between the boundaries of €30,000 and €10,000,000. The purchase price of a dwelling is considered valid when it falls in the range of €20,000 to €10,000,000. Here the lower bound is a bit more to the downside because a house might have been cheaper some years ago compared to the current value. The variable GAP is considered unrealistic and converted to NA-values, when smaller than -€3,000,000 or larger than €3,000,000. The year when a particular individual bought the house is only considered realistic when it falls from 1900 till 2022. Lastly, a place with more than 30 rooms is also viewed as unrealistic. After cleaning the data, a total of 50,630 observations are left in the dataset.

Data Visualisation

A first inspection of the variable GAP_WOZ resulted in the boxplots shown in Figure 1. The results show that the relative gap was primarily positive from 2008 to 2021. During 2012 and 2013, the gap divided by the WOZ-waarde lay close to zero, meaning that the WOZ-waarde and the self-reported market value were almost the same. Decreasing house prices due to the financial crisis might have caused the gap to nearly disappear, but there is, for now, no further proof that underwrites this statement. Figure 1 simply shows the numbers in this dataset so no real conclusions can be taken from this figure. During the entire period, the average relative gap was 0.066 or 6.6%, and the median equals 0.07 or 7%.

An inspection of the dataset tells us that the sample mainly consists of relatively older people than the Dutch population. Figure 2 reveals that most people are between 50 and 70 years old. The average age in this dataset is 58.5 years old. Figure 2 suggests that variance in age increased over time, but the average age stayed approximately the same during the studied period. Figure 3 shows the net monthly household income for the respondents. It is remarkable to see that the trend in Figure 3 seems to be a bit downwards. Investigating the codebook provided by the LISS panel, there is no reason to believe that these numbers are already corrected for inflation or otherwise manipulated. However, the dataset does show that over the years, more and more people said that their net monthly income as a household equaled zero.

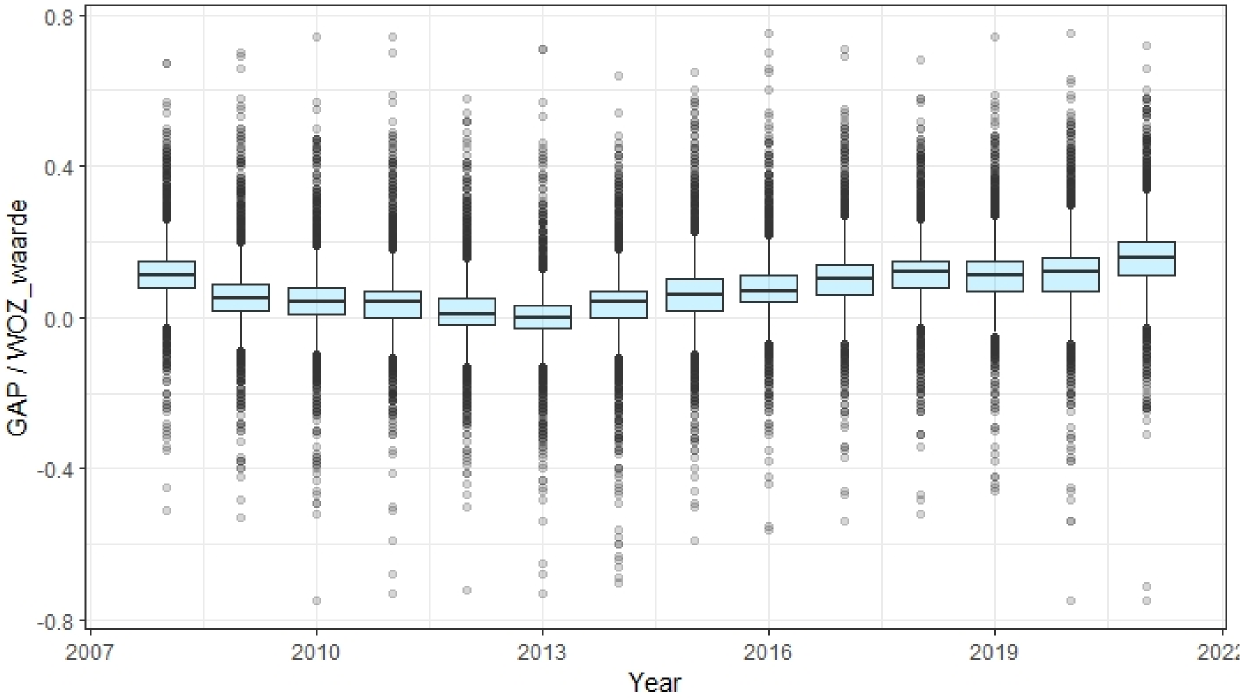

Figure 1: GAP divided by WOZ-waarde for each year

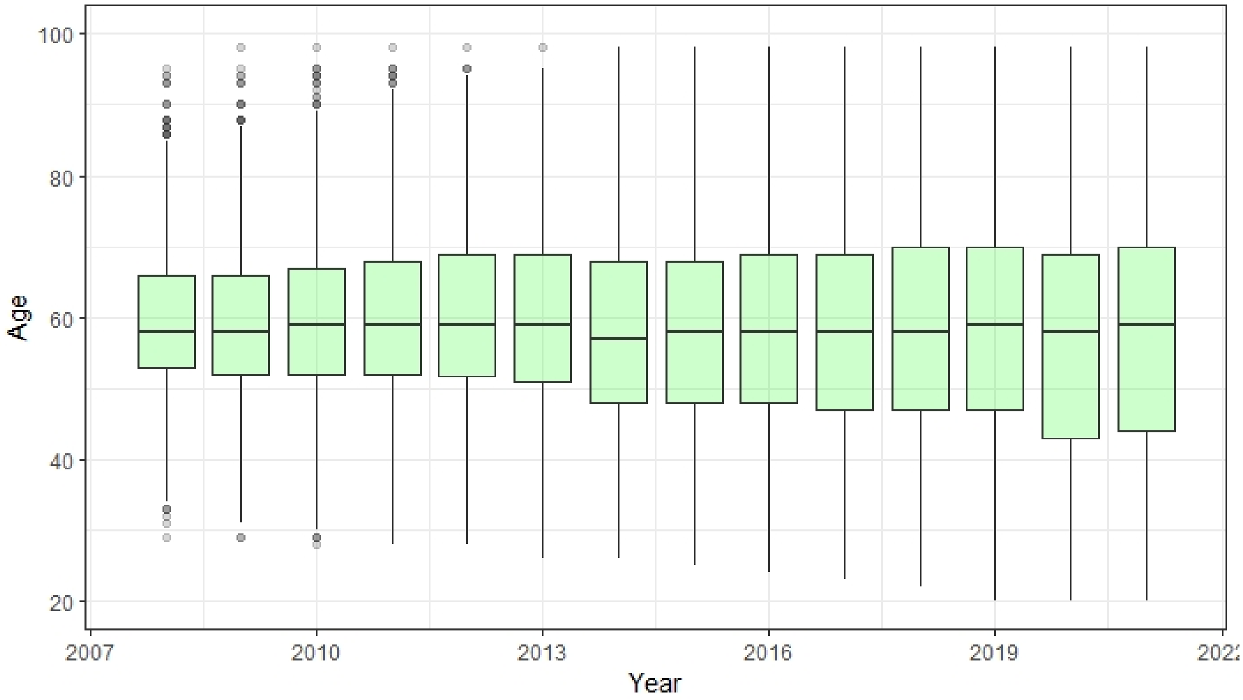

Figure 2: Age for each year present in the dataset

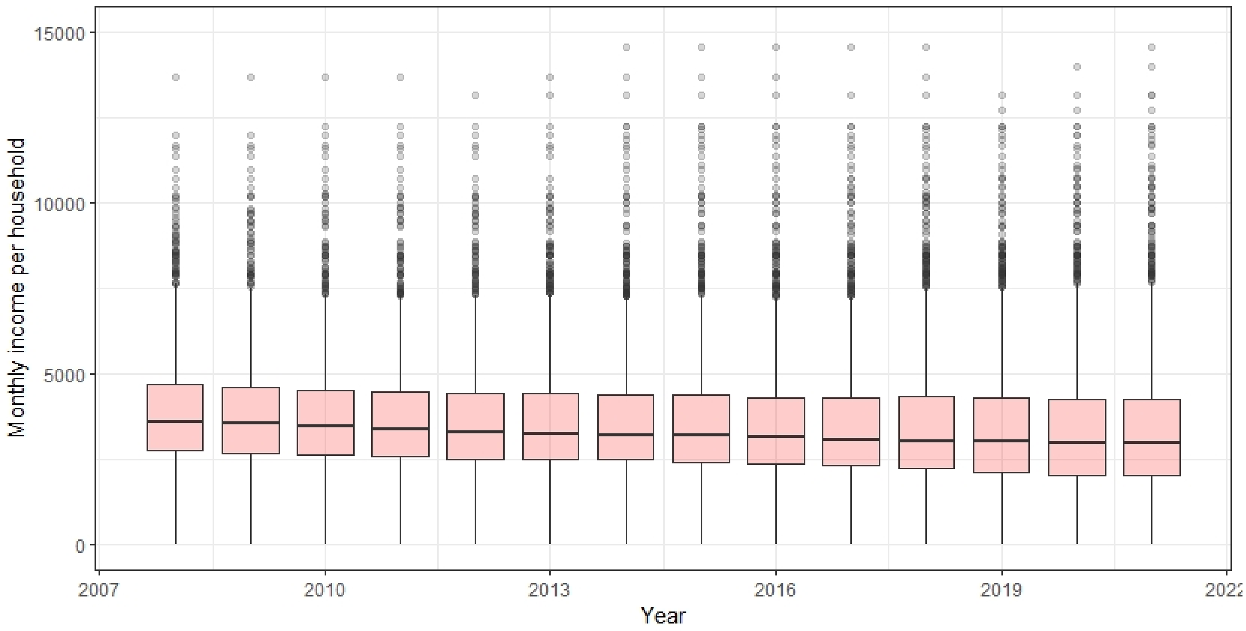

Figure 3: Monthly household income for each year present in the dataset

Missing values

Already at the beginning, there were some missing values in the dataset, and after cleaning the dataset, the number of missing values only increased. Before doing any analyses or building any model, it is essential first to solve the problem of missing values. According to Little & Rubin (2019), there are three different ways to deal with these missing values. The first option is to do nothing and start analysing the incomplete raw data. The problem with this approach is that certain methods, like Linear Regression, start ignoring the incomplete cases and only focus on the observations without any missing values. A second possibility is to give the complete cases certain weights. This second approach tries to compensate for the incomplete and omitted observations. It is still a bit unclear and uncertain with which weights one should compensate for the incomplete cases. A third approach is mainly based on the assumption that missingness hides meaningful value. (Little & Rubin, 2019) This third approach is called *Imputation*, which is predicting and replacing the missing values. Imputation is used to deal with the missing values in this dataset. Imputation is still not the ideal solution. It would be best if there were no missing values because each method used to predict missing values is based on assumptions that can not be checked.

Before picking any method to do imputations, it is essential to check where these missing values are located in the dataset. Are these values missing at random, or is something else going on? Little and Rubin (2019) state that there are three major possibilities here. Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). MCAR means that the missingness does not depend on both the independent and dependent variables. In the case of MAR, the missingness probability only depends on the observed data but not the unobserved data. For MNAR, the missingness probability can depend on observed and unobserved data. For this dataset, it is difficult to argue which option is most valid. Certain questions might have been skipped and left unanswered by accident. These missing values can be considered as being missing completely at random. But there are also instances that specific follow-up questions of the survey were automatically filled as missing when the first question was not answered. That is not entirely random. A second reason randomness can not be guaranteed is the answers 'I do not know' or 'I do not want to say' that were converted to missing values. For a question about net income, the respondent might be more likely to use these phrases in comparison with a question about their age. The missing values are spread out through the entire dataset but can not be considered as being entirely random.

In this paper, the R package *missForest* will be used. So, a random forest is built to predict the missing values. A significant advantage of this approach is that it also uses multiple iterations to capture the uncertainty that comes with using predictions in the dataset. Here, the number of iterations was set to 10, and a total of 100 trees were used to build the random forest. Theoretically, it would be best to set the number of iterations equal to infinity to take away all the simulation error (Buuren 2021). The problem is that more simulations also take more computational time and memory. Buuren (2021) also shows that most of the simulation error is already gone when only two or three iterations are used. Computational time and memory were not much of an issue in this paper, so the number of iterations was set equal to 10.

In the end, 84.23% of the dataset consists of actual data, and the other 15.77% were first missing values and are now replaced with values predicted by using a random forest. After checking the new imputations, the conclusion is that the random forest did not come to any unrealistic predictions by extrapolating outside the arbitrary boundaries that were set during the data cleaning process. Most of the imputations ended up close to the mean. So due to these imputations, the effect of outliers might become a little bit weaker. The results are more likely to underestimate than to overestimate the true effect because of this. It is crucial to keep this in mind when interpreting the results later. A second issue with these imputations is that the random forest made predictions with some decimals that are not always realistic. Suppose that the number of children was first unknown for an observation, then the algorithm predicted a number of children that equals 1.2. This number was later rounded to 1 because 1.2 is not a possible value in this scenario. Rounding imputations was done as little as possible, but it was necessary for some integer variables.

4. Conceptual framework

Predicting the gap between the WOZ-waarde and the market value of a dwelling and understanding which factors play an important role will be done with the following steps: 1. Linear regressions with Formula 1 to get first insights. 2. A neural network to make accurate predictions that do not require a lot of computational time. 3. Lastly, a decision tree will be used as a surrogate model to further understand which variables play a crucial role in predicting the gap.

The main function of this research contains the following rubrics with variables:

$$\text{GAP/WOZ}_{ij} = f(\text{house characteristics}_{ij}, \text{background characteristics}_{ij}, \text{time effects}_{j}, \text{macroeconomic variables}_{j}) \tag{1}$$

For each rubric, many different variables will be included, which can be found in the variable list in Appendix A. Formula 1 shows *GAP/WOZ* as the dependent variable. This dependent variable is the relative gap, so the difference between the self-reported market value and the WOZ-waarde divided by the WOZ-waarde. In Formula 1, the subscript i represents each individual in the dataset, and j represents each year that this data concerns ranging from 2008 up to and including 2021. So, house characteristics and background characteristics can change per individual and year, but time effects and macroeconomic variables, on the other hand, only differ among years and not among individuals.

First insights

Before making any predictions, it is essential to understand how the different variables are related and which factors might play an important role in the model later on. The data visualisation section already gives some first insights into the relationship between the most critical variables in this research. To better understand the relationships in the entire dataset, linear regressions will be used. These multiple linear regressions will help with getting a first indication of which variables are essential in predicting the gap between the WOZ-waarde and the self-reported market value. A regression is a straightforward method which is easy to implement as there are no problematic options or parameters that should be optimised.

Predicting the gap by using a Neural Network

The information that will be included to predict the gap between the WOZ-waarde and the self-reported market value is all visible in the variable list in Appendix A. It is important to note that this formula will not include the self-reported market value. This information on the self-reported market value is available for almost all observations, so it will be possible to compare the predictions with the actual gaps to check the accuracy of the model. A neural network will be used to build the model with which the predictions will be made. Compared to other models, an advantage of a neural network is its high accuracy. A neural network is essentially a network that consists of one or multiple hidden layers, and each hidden layer is made of hidden nodes. When the number of hidden layers or the number of hidden nodes increases, the model becomes more detailed, and results will most likely become more accurate as well. Maskara, Kubica, and Jochym-O'Connor (2019) further explain that the number of hidden layers, the number of nodes per hidden layer, the size of each batch and the total number of training steps can all be optimised by using a grid search. A second possibility would be to manually change the hyperparameters and see how they affect outcomes.

A potential drawback of using a neural network is its high computational time when the model consists of many hidden layers and hidden nodes. This research will make use of the H2O R package to build the neural network to avoid a high computational time as much as possible. Candel and LeDell (2017) explain that using this package requires the user to install Java because it uses this platform to hugely decrease the computational time for a specific model. It is also important to keep in mind that this H2O package overrides certain functions in R, so one should only load this package when it is actually used. This function comes with many different arguments that all impact the performance of the model. (H2O-Innovation-Inc 2016)

The predictions made by this neural network will be evaluated by using a bandwidth of error. Lubberink, Post, and Veuger (2017) shows in their paper that municipalities accept a bandwidth of error when calculating the WOZ-waarde. Everything within the bandwidth of 25% (from 0.875 to 1.125; +/- 12.5%) is seen as correct taxation when discussing houses. For other kinds of real estate, even a more considerable bandwidth of error is accepted. This research looks into the WOZ-waarde of dwellings, so the bandwidth of 25% will be the main rule by which it will be determined if a prediction is considered accurate or not. So, within the bandwidth is a correct prediction and outside of the bandwidth means that the prediction is too much different from the actual value.

Some essential arguments of this H2O neural network will be discussed. Before building the model, the dataset was randomly split into training and testing datasets. 70% of the data is used for training, and 30% is used for testing the model. The first argument of the neural network is *Nfolds*, which will be set equal to 5. This activates 5-fold cross-validation when building the model. 5-fold cross-validation means that the train data is split into five parts. For each iteration, all parts of the training data are used to train the model except one, which is used for validation. Cross-validation stops when all parts of the training data have been the validation data during an iteration. Each iteration will get an accuracy score. The average of all five accuracy scores is the overall performance of the model on training data. The parameter *Standardise* defaults to true in this formula, which means that all the numeric variables are standardised automatically. *Rectifier* will in this formula be used as the activation function. An activation function is used to transform input into output for each hidden layer in the neural network. A neural network would simply be a linear function without many learning capabilities if an activation function is not used. So, an activation function is needed to extract complicated non-linear information from the data that is used as input for the neural network (Sharma, Sharma, and Athaiya 2020).

It is also crucial to set a seed in the formula itself because it will do some of the computation on Java, and this part must also be reproducible. The number of hidden layers and hidden nodes can be changed with the argument *hidden*. As said earlier, Maskara, Kubica, and Jochym-O'Connor (2019) stated that more layers and nodes will result in a higher accuracy of the model. But a neural network which is too large will lead to a very high computational time and a network that is overfitting to the training data. So, the goal here is to optimise the model's size to a point where the predictions are as accurate as possible, but the model is still not overfitting to the training data. The researcher will manually test which kind of cross-validation, number of hidden layers and number of neurons results in the most accurate model.

	Model	Hidden layers	Neurons per hidden layer	Cross-validation
1	h2o_nn2x_200	2	200	5-fold
2	h2o_nn2x_200_10	2	200	10-fold
3	h2o_nn3x_200	3	200	5-fold
4	h2o_nn3x_200_10	3	200	10-fold
5	h2o_nn4x_200	4	200	5-fold
6	h2o_nn5x_200	5	200	5-fold

Table 1: Settings for different neural networks

Besides the R^2, the following metrics will be used to evaluate which settings are the best choice:

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2} \tag{2}$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_1 - \hat{y}_i| \tag{3}$$

The RMSE, as shown in Formula 2, calculates the square root of the mean of all squared errors. In Formula 2, O_i is the actual observation and S_i is the predicted value for a certain variable. The subscript i ranges from 1 to n, which is the total number of observations. The MAE, as shown in Formula 3, gives the average of the absolute errors. The real value is shown by y_i and $\hat{y}_i$ is the predicted value. The observation is again shown by the subscript i ranging from 1 till n, which is the total number of observations.

Surrogate model

To increase the interpretability of the results, the third step after getting the first insights and evaluating the predictions will be creating a surrogate model. A surrogate model is a model that uses the same data as the black box but which is, contrary to the black box, interpretable and simpler to understand. A crucial aspect of a black box is that it is almost impossible to entirely understand what is happening inside the model. It is also unclear which variables play an essential role in making the predictions. A surrogate model functions as an extra simpler model to gain more insights into what happens in the black box. Blanco-Justicia and Domingo-Ferrer (2019) show in their research that each surrogate model is based on a trade-off between comprehensibility and representativeness. This trade-off essentially means that when a surrogate model becomes more detailed, it might explain more about the black box, but at the same time, the surrogate model also becomes harder to understand. Blanco-Justicia and Domingo-Ferrer (2019) further explain this trade-off using decision trees as a surrogate model. Here, a small tree with only one root node and one decision node is very comprehensible but not very representative for the entire black box model it tries to explain. Increasing the number of decision and leaf nodes will increase representativeness but decrease comprehensibility.

In this research, a decision tree will also be used as a surrogate model. When building this model, it will be essential to keep this trade-off between comprehensibility and representativeness in mind. The surrogate model should give as much information about the black box as possible while remaining understandable. This will mostly be important when optimising the number of nodes and the number of leaves the tree should be built with.

Formula pruning:

$$SSE + \alpha|T| \tag{4}$$

The R package *rpart* will be used to build the decision tree shown in Figure 6. The decision tree will be trained by optimising the complexity parameter (cp) and the number of observations per leaf (minbucket). The complexity parameter is shown as α in Formula 4. α penalizes the number of terminal nodes (T). The abbreviation *SSE* in Formula 4 stands for Sum of Squares Error. The SSE is the sum of the squared differences between each observation and its group's mean. Optimizing the complexity parameter is called *pruning*. When one prunes a decision tree, one actual starts with a large tree and makes the tree smaller by optimizing the function shown in Formula 4. After a grid search on 10-fold cross-validated training data *cp* being equal to 0.001 was the best option. This decision was made based on the RMSE as shown in Formula 2. The RMSE stopped decreasing heavily around the area of 0.001. The decision tree kept becoming more accurate for lower values of cp but this also results in far larger and more detailed trees and that is not desirable here. When optimizing the parameter *minbucket*, the grid search again showed the same pattern: A lower number of observations per leaf comes with a larger and more accurate tree. In the end, the parameter minbucket was set to 500. To be fair, this was mainly decided on the interpretability and size of the eventual tree and not on the performed grid search as the decision tree is used as a surrogate model and not as method to make the most accurate predictions.

Market sentiment analyses

$$\text{House Pricce Netherlands } = \alpha + \beta_1 \times \text{ Inflation}_j + \beta_2 \times GDP_j + \beta_3 \times \text{ Change in house supply}_j + \beta_4 \times \text{ Covid19}_j + \beta_5 \times \text{ Financial crisis}_j + \epsilon \tag{5}$$

The sub-question “What is the impact of market sentiment on this gap?” will be answered as well as possible using Formula 5. Overall sentiment on the market is captured by the average house price in the Dutch housing market during a specific year. The year is shown in the formula with subscript j and ranges in the data from 2008 to 2021. Variables that might have an impact on the circumstances of the Dutch housing market are, firstly, macro-economic factors like Inflation and GDP. Time effects might also play a role, such as the impact of Covid19 or the financial crisis. The last thing that is also included is changes in the supply of houses in the Netherlands. The change in house supply is an indexed variable that takes 2015 as its base year. This variable equalled 100 in 2015 and 100.59 in 2016. This difference means that the total number of houses in the Netherlands had risen by 0.59% in 2016 compared to 2015.

In this research, market sentiment is meant as Jin, Soydemir, and Tidwell (2014) defined it in the context of the housing market: *‘Market sentiment is conceived of as expectations or judgments that are not fully justified by available information on market fundamentals, and thus beliefs on future market conditions can be misguided. (...) Theoretically, this phenomenon can be explained by a behavioral concept referred to as overreaction. In this study, the concept of overreaction posits that homebuyers respond disproportionately to new information. This causes housing prices to fluctuate more than fundamentals might suggest as homebuyers overreact to fundamental market information.’* Hui and Wang (2014) investigated the impact of market sentiment in the private housing sector in Hong Kong. They developed their sentiment index to capture investors’ behaviour and the corresponding market scenario. The sentiment index is the positive sentiment divided by the total sentiment (positive + negative sentiment). Sentiment itself is measured by dividing the expected number of sentiment-based trades and the expected total number of trades in the market. Furthermore, the sentiment index was found to be an efficient predictor of the price level, return rate of price and trading volume of the housing market in Hong Kong. In this paper, there is no such thing as a sentiment index, but as this sentiment index is shown to be a good predictor of the price level, the price of a house will here be used to say something about the overall sentiment on the market.

An important question that remains is: ‘What can be said about overall market sentiment when certain price movements occur?’ The assumption is that both strong positive and strong negative moves result from an unstable, unpredictable housing market in which market sentiment plays an important role. The methods that will be used to find out which factors play an essential role in explaining price movements and market sentiments are a random forest and a decision tree. The random forest will function as a method to come to accurate predictions. The decision tree will be used to control or further underwrite the random forest's results. Kingsford and Salzberg (2008) warn in their paper that decision trees are pretty sensitive to the data it is trained on. A decision tree can become very large or detailed and, because of that, overfit the training data. It is necessary to examine the gain in extra predictive power for each new split to ensure that each split is worth the additional risk of overfitting. Gini impurity should be used to decide how and if a new split is created. For each feature in the dataset, it is calculated how well it divides the data into the correct classes. The variable that makes the best division will have the lowest Gini Impurity Score and will be picked for that particular split.

Formula 6 shows how the Gini Impurity Score is calculated. The training data is represented with the letter D. The probability of datapoints belonging to class i is denoted by p_i. The total number of classes is represented by n. A lower Gini Impurity Score means that a relatively larger part of the observations belongs to one class and that an extra split will become more valuable.

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{6}$$

A decision tree is often called 'greedy' as it tries to use the best possible variables to explain as much variance as possible at each split. Because of this, variables that are higher in the tree play a more significant role in the outcomes that are located at the bottom. The root node, the variable at the top, could be seen as having the highest predictive power.

The random forest is a combination of a certain number of decision trees. Each decision tree gives its own prediction of the price level on the market. Stacking all those different predictions eventually leads to one more reliable prediction and, most likely, more accurate than just the prediction made by one tree. But if we keep feeding the algorithm the same data, why does it not lead to a random forest with all similar trees? Bootstrap Aggregation (Bagging) solves this problem. Bagging and eventually building the random forest happens through three steps. First, it randomly divides the training data into different sub-samples. Secondly, it trains a tree based on each data sub-sample. The last step is averaging the predictions of all trees into one final prediction.

A couple of different parameters should be optimised to make the model's results as accurate as possible. But before talking about the parameters of a random forest, it needs to be clear what the out-of-bag error (OOB) means. The out-of-bag error is calculated using the following steps: For each sample, x_i find the prediction $\hat{y}_i$ for all bootstrap samples which do not contain the sample x_i. Average these predictions to obtain $\hat{y}_i^{\text{oob}}$. Obtain the error by squaring the difference between $\hat{y}_i$ and $\hat{y}_i^{\text{oob}}$. In the end, the OOB is equal to the average of the errors for all observations n. Each observation is notated as i ranging from 1 till n.

One should decide how many different trees are used to build the random forest. Adding more trees leads to more robust and more accurate predictions but also to a higher computational time. It takes already quite some time to build the random forest when one tunes the parameter *ntree*. The goal here is to use enough trees so that an extra tree does not lead to a considerable decrease in the out-of-bag error. But, on the other hand, not too many trees might lead to a random forest that is overfitting to the training data. When creating the random data sub-samples, some observations were left out. These data points left out during the process are called the out-of-bag points. Later on, these out-of-bag points are used to calculate the out-of-bag error. So, the out-of-bag error shows how well the model performs on unseen data. If this out-of-bag error does not decrease much when extra trees are included, it becomes unnecessary to increase the number of trees. There are different ways to stop a random forest from overfitting. One could, for example, set a maximum number of nodes or include a minimum number of observations for each node. But the most prominent way with which overfitting can be prevented is by optimising the parameter *mtry*. This parameter's value decides how many variables will be randomly selected from the dataset for each split. Suppose it equals three, then three variables are randomly selected, and the algorithm makes the best possible split with these three variables. Using a higher value for mtry will result in a 'greedier algorithm' and more overfitting to the training data. The mtry will be optimised by using a grid search in this research. The number of trees will be decided based on the out-of-bag error movements.

Relationship between the WOZ-waarde and the self-reported market value

The biggest challenge when doing this research is that the WOZ-waarde should reflect the market value as good as possible. But it is not unthinkable that it works the other way around. The WOZ-waarde might have some impact on the market value as well. Although, the Waarderingskamer tries to prevent the WOZ-waarde from having a real impact on the market. Waarderingskamer (2019) clearly mentions that the WOZ-waarde is meant to follow the market and not the other way around. Multicollinearity mainly occurs as a problem when both these variables are used as independent variables in the same model. In this scenario, the possible multicollinearity of these two variables will lead to wider confidence intervals and results that are not reliable. It is not sure yet how WOZ-waarde and the self-reported market value are related, but it is essential to remember that multicollinearity might hurt the results.

5. Results

Linear Regression Main Function

The first step in analysing this data and answering the research questions was building a Multiple Linear Regression. The formula used for this method is shown as Formula 1 under Contextual Framework. The problem with using this method is the unknown values. A Linear Regression ignores a specific observation if it encounters an unknown value in the dataset. Initially, 15.77% of the data that is used here is an unknown value. These unknown values caused the linear regression to only use about 30% of the dataset to come to its results. Even certain variables were excluded initially because only one level was present in the data without missing values. The missing values were imputed using a random forest to avoid this issue.

Forward selection was implemented to prevent using all 56 variables present in the dataset. Using all variables in one regression would result in much unnecessary noise. Forward selection begins with a model that contains no independent variables at all. Then it starts adding one variable after another and decides for each variable whether it is a valuable contribution. This process stops when all variables are considered once. The decision to include a variable can be made on different metrics. Here, these decisions were simply made based on the p-value. In the end, forward selection results in a list of variables that should be added to the model. In the scenario, 29 of the 56 variables are considered crucial contributions to the model. These 29 variables are all used for the regression shown in Table 2.

Multiple variables are significant and might play a role in predicting the relative gap between the WOZ-waarde and the self-reported market value of a house. Significant house characteristics are *WOZ-waarde*, *Purchase_Price*, *Rooms*, *Type_Of_House*, *Owner*, *Municipality* and *Mortgage*. But also personal characteristics like *Household_Income*, *Education*, *No_Problems*, *Too_Small*, *Satisfied_House*, *Second_Dwelling* and *Noise_Neighbors*. Time effects and macroeconomic variables like *FinancialCrisis*, *Covid19*, *Year_WOZ* and *Year_Of_Residence* are also shown to significantly impact the relative gap between the WOZ-waarde and the self-reported market value of a dwelling. Looking at these results, one can conclude that a bigger house or personal characteristics that suggest a higher standard of living both lead to a more significant gap between the WOZ-waarde and self-reported market value of a dwelling. Being satisfied with the house or the neighbourhood and having no considerable issues with the property or the immediate environment are also characteristics that will most probably result in a larger gap. It needs to be said that a simple Linear Regression might be too little evidence on its own for drawing conclusions.

There are a couple of assumptions regarding the residuals of a linear regression. The residuals should, for example, follow a normal distribution and have a mean value around zero. Figure 4 shows a Q-Q plot of the residuals for the regression in Table 2. The values in this Q-Q plot should follow the red line when the residuals have a normal distribution. In this scenario, the red line is not followed, resulting in the conclusion that the residuals are not normally distributed. This violation of the assumptions is probably caused by a fixed or random effect that is not captured by this basic linear regression.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.2400	2.5034	5.29	0.0000***
WOZ_waarde	-0.0000	0.0000	-49.46	0.0000***
HousePriceNetherlandsWest	0.0004	0.0014	0.30	0.7605
Purchase_Price	0.0000	0.0000	8.54	0.0000***
Household_Income	0.0000	0.0000	4.71	0.0000***
Owner	-0.0078	0.0015	-5.23	0.0000***
Partner	0.0057	0.0041	1.39	0.1650
FinancialCrisis	-0.0338	0.0057	-5.98	0.0000***
Year_WOZ	-0.0041	0.0017	-2.33	0.0198**
Education	0.0042	0.0010	4.07	0.0000***
Mortgage	-0.0110	0.0034	-3.21	0.0013***
Noise_Neighbors	-0.0073	0.0032	-2.32	0.0203**
Type_Of_House	0.0028	0.0007	3.87	0.0001***
Satisfied_House	0.0023	0.0010	2.41	0.0162**
Second_Dwelling	0.0210	0.0069	3.03	0.0024***
Net_Monthly_Income	-0.0000	0.0000	-1.93	0.0539*
HousePriceNetherlandsNorth	0.0008	0.0014	0.56	0.5736
Year_Of_Residence	0.0004	0.0001	3.43	0.0006***
Municipality	-0.0002	0.0001	-1.97	0.0487**
Covid19	-0.0175	0.0084	-2.08	0.0372**
Rooms	0.0019	0.0010	1.97	0.0492**
Too_Small	0.0130	0.0054	2.38	0.0172**
No_Problems	0.0078	0.0036	2.19	0.0283**
Primary_Occupation	0.0005	0.0004	1.24	0.2160
GDP	0.0000	0.0000	1.69	0.0904*
Reside	-0.0110	0.0082	-1.34	0.1791
Civil_Status	-0.0015	0.0011	-1.32	0.1860
Year	-0.0029	0.0024	-1.20	0.2293
Special_Adjustment	0.0047	0.0043	1.09	0.2758
Position_Household	0.0019	0.0018	1.05	0.2928

P-value: * <0.10, ** <0.05 and *** <0.01

Table 2: Linear Regression with Forward Selection; Formula 1

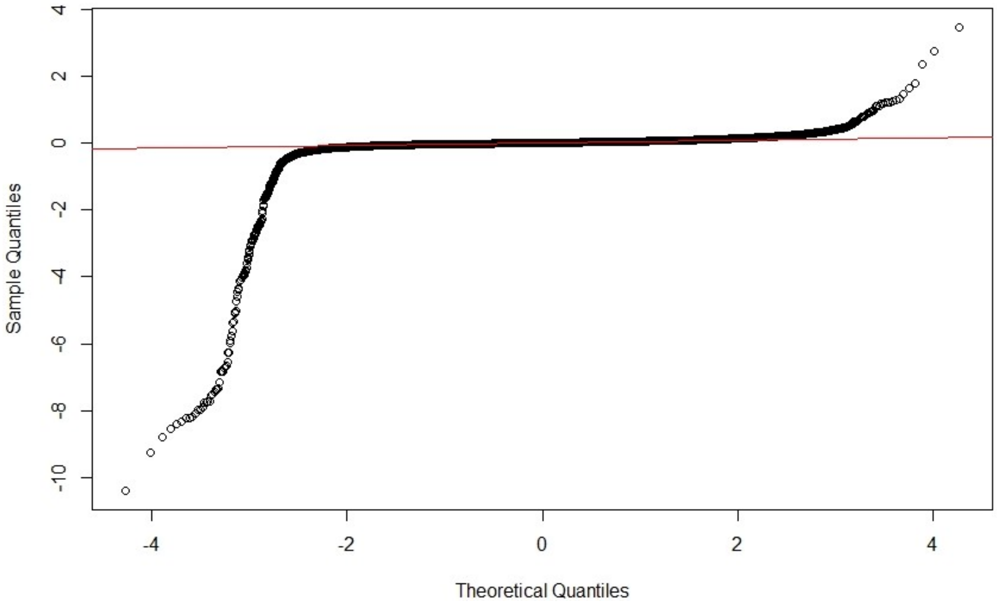

Figure 4: Q-Q Plot of the residuals

Panel Regressions

The following regressions, of which the results are visible in Appendix B, are primarily meant to use on Panel data. Panel data is a dataset that contains the same variables and information for a certain number of years. In this research, the Panel data is collected for the years 2008 to 2021. The main advantages of Panel data are that one can establish trends and compare individuals with themselves during the years. Tables 7 to 10 in Appendix B show an overview of the results of three separate regressions on Panel data. Column 1 shows a Pooled OLS, which could be seen as a basic Linear Regression. Column 2 presents the results of a Fixed-effects Regression. Lastly, column 3 gives the output of a regression with Random-effects. A Fixed-effects Regression creates a constant for each individual in the dataset to capture all time-invariant effects, so it essentially compares an individual with itself over time. A Random-effects Regression, on the other hand, allows for time-invariant variables to play a role as independent variables. Random effects should be preferred if there is some reason to assume that differences across respondents affect the dependent variable.

But how can one decide which of the three options in Appendix B gives the best way to explain the dependent variable *GAP_WOZ*? Pooled OLS and a Fixed-effects Regression were compared by using an F-test. The results of this F-test are shown in Table 3 in row 1. The p-value is smaller than 0.05, meaning the null hypothesis is rejected, and the alternative hypothesis is true. This result indicates that the Fixed-effects Regression performs better than the Pooled OLS regression. Afterwards, the Fixed-effects and the Random-effects model were compared using a Hausman test. The p-value for this Hausman test is in Table 3 on row 2. The resulting p-value showed that the Fixed-effects Regression is also preferred over the Random-effects Regression. But a Random-effects Regression is still preferred over a Pooled OLS regression according to the Lagrange Multiplier Test (Breusch Pagan) on row 5 of Table 3.

There are still other options that might improve the Fixed-effects Regression. Should, for example, only individual Fixed-effects be included or also Time-fixed effects? Or should the model also contain certain variables to capture a time lag? An F-test and a Lagrange Multiplier Test (Breusch Pagan) were performed to determine if it is necessary to include Time-fixed effects in the model. The results of these two tests are shown in rows 3 and 4 of Table 3. Both tests showed strong support for having the time effects. As a result, the variable *Year* was incorporated as a factor in the Fixed-effects Regression shown in column 2 of Appendix B. Lastly, a Dickey-Fuller test was used to check for stochastic trends. The results of this Dickey-Fuller test can be found in row 6 of Table 3. The p-value of the Dickey-Fuller test resulted in the conclusion that including variables to account for a time lag is unnecessary in this scenario.

	Test	Description	P-value
1	F-test	Comparing OLS with Fixed-effects model	0.0000***
2	Hausman Test	Comparing Fixed-effects with Random-effects	0.0000***
3	F-test	Comparing Time-fixed effects with no Time-fixed effects	0.0131**
4	Lagrange Multiplier Test (Breusch-Pagan)	Comparing Time-fixed effects with no Time-fixed effects	0.0084***
5	Lagrange Multiplier Test (Breusch-Pagan)	Comparing Random-effects with OLS	0.0000***
6	Dickey-Fuller Test	Comparing a model with and without time lag included	<0.01***

P-value: * <0.10, ** <0.05 and *** <0.01

Table 3: Tests with Panel data models

As the tests showed that the Fixed-effects Regression is the best option, these coefficients in column 2 of Appendix B are most relevant. The time effects are all significant on a 5% level except for the years 2019 and 2020. These time effects suggest that it might be hard to predict the impact of Covid19 on an individual level. The results of this Fixed-effects regression show that the following house characteristics have a significant effect on the relative gap on a 5% significance level: *WOZ-waarde*, *Owner*, *Purchase price*, *Rooms*, *Special Adjustment*, *Type of House* and environmental problems due to *Traffic or Industry.*

There are also a couple of background characteristics of the individuals that are proven to be significant on a 5% significance level: *Position (in) Household*, *Partner*, *Civil status*, *Primary Occupation*, *Net Monthly Income* and *Household Income*. The Fixed-effects Regression only looks into a linear relationship while the actual relationship might be a lot more complicated; because of that, it seems unrealistic to already give a reasonable interpretation of the coefficients.

Linear Regression Market Sentiment

The second Linear Regression that is built uses Formula 5. This regression is meant to give a first indication of which variables might cause price changes. The variables *FinancialCrisis* and *Covid19* are both dummy variables. The regression results are shown in Table 4. Comparing the coefficients to the intercept reveals that both Covid19 and the financial crisis have a significant but surprisingly low impact on the dependent variable *House prices in the Netherlands*. Although the coefficient of *Covid19* and *FinancialCrisis* are pretty high, the total effect is relatively low as *Covid19* and *FinancialCrisis* are both dummy variables that only take the values one and zero. Compared with, for example, *SupplyHouses*, which has a value of around a hundred, the total effect of the financial crisis or Covid19 becomes relatively low.

The results further show that a higher GDP and a higher supply of houses will result in a price increase. It seems a little unrealistic that a higher supply, so more dwellings that become available on the housing market, results in increasing prices. results in increasing prices. In reality, it might be the case that an increasing supply is often the result of rising prices, and because of that, rising supply becomes a good predictor of increasing prices. Lastly, inflation has a decreasing effect on the average price change of houses in the Netherlands. The decreasing effect of inflation is pretty strange as higher inflation typically increases prices. Inflation could also be an indicator of economic circumstances that are relatively bad. An economic downturn might lead to less overbidding and a smaller gap between the market value and WOZ-waarde of a dwelling. It remains unclear through which mechanisms inflation leads to a smaller relative gap. But in the end, these linear regressions can not prove any conclusion on their own. As the name already says, this method only investigates a simple linear relationship in this data, while the actual relationships in the data might be more complex. So, these results should only be used as a first step to gain some insights and not as evidence for certain conclusions.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-663.7263	8.3711	-79.29	0.0000***
Inflation	-6.6574	0.0675	-98.68	0.0000***
GDP	0.0001	0.0000	33.97	0.0000***
SupplyHouses	13.7618	0.1579	87.16	0.0000***
Covid19	28.4869	0.1742	163.51	0.0000***
FinancialCrisis	14.7057	0.1641	89.62	0.0000***

P-value: * <0.10, ** <0.05 and *** <0.01

Table 4: Linear Regression Market Sentiment; Formula 5

Neural Networks

A black box model is one of the best ways to come to the most accurate predictions. A black box and especially a neural network is a self-learning method that can detect and capture complex structures that might be present in the data. In this paper, five different neural networks are built to determine which one has the best settings and will eventually result in the best predictions. An overview of the various neural networks can be found in Table 1. The main differences among these models are the number of hidden layers and how cross-validation was performed when building the model. The models 1, 3 and 5 all use 5-fold cross-validation, but the models 2 and 4 are based on 10-fold cross-validation. Models 1 and 2 use two hidden layers with each 200 neurons. Models 3 and 4 consist of three hidden layers with 200 neurons, and model 5 uses four hidden layers with again 200 neurons each.

5-fold and 10-fold cross-validation are two common ways cross-validation is often applied. Both 5-fold and 10-fold cross-validation were used for the same model size to determine which performs best. The results in Table 5 show that model 3 performs better than model 2 and that there is not much difference between models 4 and 5. So, it is assumed that 5-fold cross-validation does the best job here. There is also experimentation with the model size; the model size varies from 2 to 5 hidden layers. Looking at Table 5, it becomes clear that the R^2 does not improve much when the model size becomes larger than three hidden layers, as the R^2 does not vary much for models 3, 4 and 5. The *RMSE* is the lowest for model 5, and the *MAE* is the lowest for model 3, but both metrics only show minor differences among the models. In the end, it is decided to use model 3 to make the predictions on the test data.

	Model	R^2 Train data	R^2 Test data	RMSE	MAE
1	h2o_nn2x_200	0.52	0.21	0.255	0.070
2	h2o_nn2x_200_10	0.52	0.23	0.246	0.069
3	h2o_nn3x_200	0.64	0.33	0.239	0.066
4	h2o_nn3x_200_10	0.60	0.30	0.244	0.067
5	h2o_nn4x_200	0.66	0.26	0.237	0.067
6	h2o_nn5x_200	0.64	0.29	0.241	0.069

Table 5: Performance of different Neural Networks on train and test data

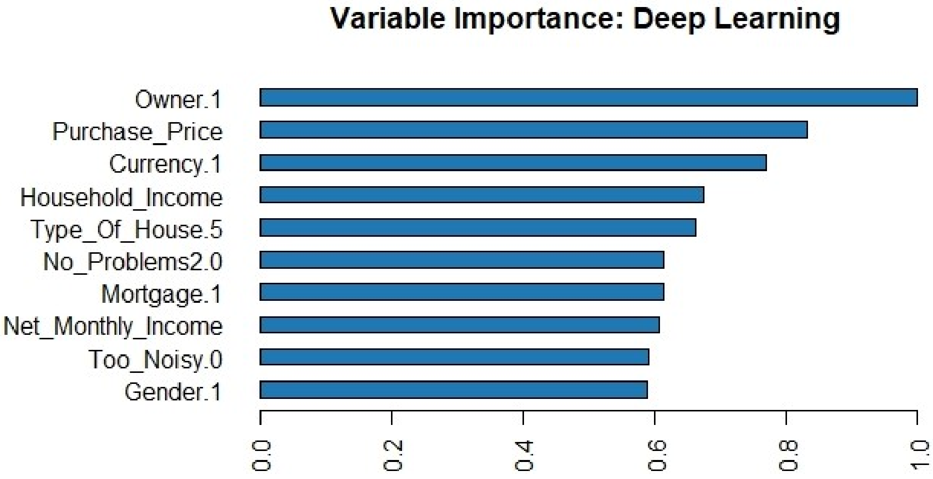

Figure 5: Variable Importance Neural Network model 3

Unfortunately, a neural network does not come with a variable list with coefficients and p-values. Because of that, it is not easy to see which variables have a significant impact and which do not. One way of getting insights into variables' relative importance is by creating a Variable Importance Plot. Such a plot shows the contribution of the most critical variables to the predictions and the model's accuracy. A Variable Importance Plot for the neural network can be found in Figure 5. This figure shows the top 10 in relative importance for the independent variables. A higher importance level also means a higher contribution to the predictive power of the neural network. Figure 5 indicates which variables play a crucial role in predicting the relative gap. *Owner.1* and *Purchase_Price* are the most important variables for the neural network according to the Variable Importance Plot. The variable *Owner.1* means that the respondent rents the dwelling, and *Owner.4* stands for 'other', for instance, cost-free accommodation or anti-squatting accommodation. The type of house variable should be interpreted as follows: *Type_Of_House.5* stands for *flat, apartment, floor or maisonette*. *Currency.1* means that the property was purchased with euros, which refers to the number of years someone already lived there.

The first insights into the neural network's predictive power can be obtained by looking at the differences between the performance on training data and the performance on testing data. To say something about this, one needs to go back to Table 5 and compare the R^2 of the same model for training and testing data. No gap or a small gap between these two values would suggest that a model's predictions are as good on known data as they are for new, unknown data. A relatively small decline in the R^2 from training to testing is pretty standard as a neural network is always overfitting a little bit to the known training data. A model that performs better on training than on testing data might also be a bit off. The decline in R^2 between training and testing is pretty significant for all models in Table 5. This gap between training and testing is the first indication that the models might not lead to perfect predictions and high accuracy.

It is vital to go back to some literature before jumping to conclusions about the neural networks' predictive power. Lubberink, Post, and Veuger (2017) show in their paper that municipalities accept a bandwidth of error when calculating the WOZ-waarde. Everything within the bandwidth of 25% (from 0.875 to 1.125; +/- 12.5%) is considered a correct taxation when discussing houses. For other kinds of real estate, even a more considerable bandwidth of error is accepted. Lubberink, Post, and Veuger (2017) investigated if the WOZ-waarde could be used as a market value indicator. They found that only 33% of the taxations fell within the bandwidth of error equal to 25% and that all other properties were taxed wrongly.

	Bandwidth of error	Correct predictions
1	10%	6.37%
2	25%	16.07%
3	40%	24.31%
4	100%	46.22%

Table 6: Percentages of correct predictions using neural network model 3

When evaluating the predictive power of the neural network again, a bandwidth of error will be used to capture an acceptable error. Table 6 shows the predictive power of model 3. A bandwidth of error equal to 10% means that a prediction is correct when it is not more than 5% higher or 5% lower than the actual value. Using a bandwidth of error equal to 25% means that the relative gap is predicted correctly in 16.07% of the occasions. The main conclusion when also looking at Lubberink, Post, and Veuger (2017) results is that predicting the relative gap is even more inaccurate than using the WOZ-waarde as a market value indicator without predicting any gap.

As the results of Lubberink, Post, and Veuger (2017) showed that the market value could only be used in 33% of the observations, is it still interesting to investigate why the gap exists and which variables play a crucial role here. The variable importance plot of Figure 5 already gave first insights, but a decision tree was built to come to even more robust conclusions. This decision tree can be seen in Figure 6. This decision tree shows that *WOZ-waarde*, *HousePriceNetherlands*, *Year* and *Household_Income* are the most useful variables present in the dataset to predict the relative gap. The decision tree in Figure 6 should be interpreted as follows: One starts at the top with the relative gap (the dependent variable) being equal to 0.065. Suppose that the WOZ-waarde is higher than 775,000 for a particular individual, then one should pick 'yes' at the top and move on to the most left leaf. For this individual, the relative gap is predicted to be -0.44. The same leaf also shows 1%, which means that 1% of the dataset ended up in this leaf.

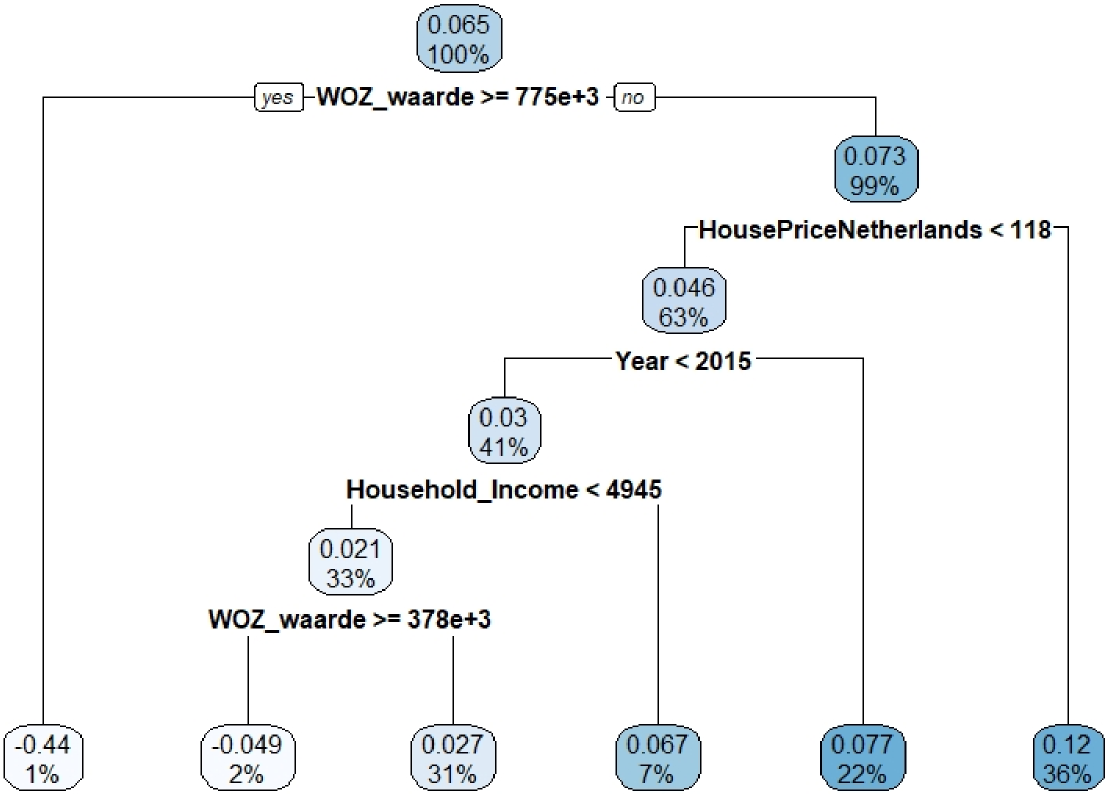

Figure 6: Decision tree as surrogate model for neural network model 3

Random Forest

The second research question of this paper is as follows: 'What is the impact of market sentiment on the gap?' Olaussen, Oust, and Sønstebø (2018) showed in their paper that human bidding behaviour and bid size do not change when market conditions change. On the other hand, it might still be possible that the gap will be different when real estate prices are heavily increasing or decreasing. Changes in house prices will be used to say something about overall market sentiment. This part of the paper will investigate if it is possible to predict house prices and show which factors play an essential role in causing the price changes. A random forest will be built to predict house prices in the Netherlands. The dependent variable is an indexed number that takes 2015 as its base year. The independent variables consist of macroeconomic variables like *Inflation* and *GDP*. Some time effects are included like *Covid19* and *Financial Crisis*. Lastly, the *supply of houses* is also present in the formula to capture all market conditions that might result in a price change.

Different parameters were optimised to develop a model that performs as good as possible. The first step was finding the best value for the parameter *mtry* by using a grid search. This resulted in a list of models that were evaluated based on the R^2. Mtry equal to three was found to be the best option. A higher value of mtry resulted in overfitting to the training data. Mtry gives the model some restrictions on the number of variables it can randomly pick for each split. Mtry being equal to three means that the model can select three different variables and use these variables to come to a new split in a tree that explain as much of the variance as possible. The second step was tuning the parameter *maxnodes*. An interval that ranges from 5 to 25 was used to find the optimal value. The model's performance stopped increasing when the parameter maxnodes equaled 22. A higher value for this parameter results in more terminal nodes and a more detailed model. A higher value for the parameter maxnodes also increases the likelihood of overfitting to the training data.

A third parameter that also influences the number of nodes for each tree is *nodesize*. Maxnodes results in a maximum for the number of nodes, but nodesize, on the other hand, sets a minimum for the number of nodes. After using a grid search, the nodesize was eventually set to 14.

The last parameter that was optimised was *ntree*. The parameter ntree determines how many decision trees will be used to build the random forest. The default value of this parameter is 500, but here different options were tested to come to a model that might perform even better. Models with the following values for the ntree were tested: 250, 300, 350, 400, 450, 500, 550, 600, 800, 1000 and 2000. After comparing the results of these options, it became clear that a model with 800 trees was the best option. Using more trees did not result in the explanation of more variance. This conclusion was made after comparing the R^2 of the different options.

The random forest can give very accurate predictions for each year's house prices. The correlation between the actual values in the test set and the predicted values is 98.97%. Comparing the predicted values with the actual values also showed that all the predicted values were within a bandwidth of error equal to 5%. So, compared to the actual value, the predicted values were no more than 2.5% higher or 2.5% lower than the actual value.

The Variable Importance Plot (VIP) in Figure 7 gives two different graphs with which the importance of all independent variables is evaluated, namely *IncMSE* and *IncNodePurity*. IncMSE stands for the increase in the Mean Squared Error of predictions due to a certain variable being permuted. The higher the number, the more influential the variable is for the model. IncNodePurity relates to the loss function with which the best splits are chosen. More useful variables achieve a better node purity. Using the results of both metrics shows that the variables *Supply of houses* and *Inflation* are the two factors that have the most significant impact on the house prices in the Netherlands. It is remarkable to see that the financial crisis is both in place one and four in these importance graphs. The variable *GDP* also affects real estate prices, but the impact of this variable is smaller than the impact of inflation or the supply of houses. There are only five independent variables present in the formula. This explains the fact that variable *Covid19* is visible as well, although it has very low importance levels for both metrics.

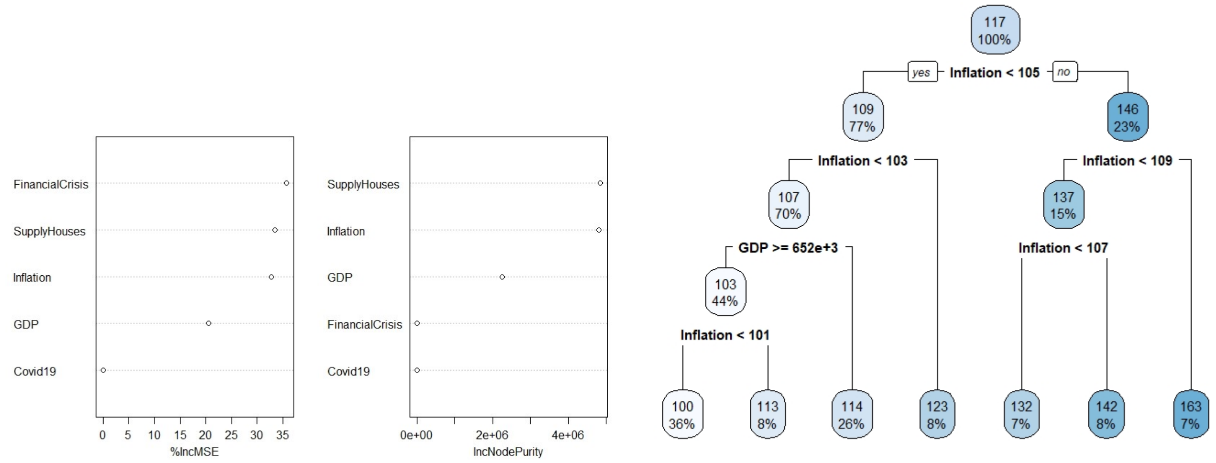

Figure 7: VIP of the random forest and a decision tree as surrogate model

Decision tree

A decision tree was used as extra proof to get an even better understanding of the variables that play a role in predicting the price levels of houses in the Netherlands. This decision tree was built using the package *rpart* and uses Formula 5 as well. It is interesting to see that the predictions made by one tree are only a little less accurate than the predictions made by the random forest. When time is limited, only creating this tree would already give reasonable predictions. In this scenario, the predictions are not that interesting because they were already made by the random forest. The decision tree is only created to find out which variables it will use. The decision tree on the right part of Figure 7 shows that it only used two variables, namely *Inflation* and *GDP*. Inflation is located at the root node and is also used for four splits, so this variable could be considered as being the most prominent variable in this tree. In contrast to the variable Inflation, GDP is present in the tree but only used for one split. Looking both at the VIP and the decision tree, it makes sense to conclude that both macroeconomic factors and the supply of houses play a crucial role in causing changes in the prices of real estate. Time effects like the Covid19 or financial crisis are less relevant than the other independent variables when predicting price levels. It might be the case that house prices are not very elastic and can not quickly increase or decrease in a one-year time span due to a crisis.

6. Conclusion and Discussion

This research tries to answer a research question and two sub-questions. The main research question is as follows: ‘Is there a predictable gap between the WOZ-waarde and the self-reported market value of a house?’ Although it is phrased as a closed question, it is more complicated than that. The relative gap is predictable but far from perfect. When using a bandwidth of error equal to 25%, only 16.07% of the predictions fell within this bandwidth. A neural network is often a method that comes with high accuracy, and a large part of the predictions that end up being correct. In this light, 16.07% within the bandwidth is pretty disappointing. So, the main conclusion regarding this research question is that this neural network cannot give close to perfect predictions for all observations.

There are several reasons why the model might not be able to come up with excellent predictions. Lubberink, Post, and Veuger (2017) tried to predict the WOZ-waarde of the government’s real estate, and they also found that many of the predictions laid outside the accepted bandwidth of error. Lubberink, Post, and Veuger (2017) mentioned that the model’s lack of specific property features might have caused the difference between the predictions and the actual values. An essential element that is, for example, not included here is information on the property’s location.

A second issue is the subjectivity of certain parts of the dataset. The survey that was used for this dataset, for example, asked the respondents to indicate on a scale from 1 to 10 how satisfied he or she is with their neighbourhood. The respondents were also asked to reveal any maintenance problems with their property. Thirdly, they were also asked to give a well-educated guess of the market value of their own dwelling. All these questions could be answered with information that does not entirely conform to reality. The respondents are biased because they are the people that live there. People from somewhere else might have totally different ideas about that house and neighbourhood and might have given completely different answers that are more objective and less biased. Another reason that may well cause an unpredictable gap is unpredictable human behaviour. According to the model, a buyer may place a bid on a dwelling that is too high or unrealistic, but the buyer is so enthusiastic about the property that they are still willing to pay that price.

The neural network’s predictive power may be a bit disappointing, but that does not mean that no lessons can be learned from the results. The Linear Regressions and the neural network’s results revealed that certain factors help predict the gap between the WOZ-waarde and the self-reported market value. A concrete answer to the first sub-question: ‘Which factors result in the fact that the gap exists?’ would be as follows. The following variables are the most important variables with which one can predict the gap: *being the owner or not, purchase price, household income* and the *type of the house*. These aspects all played a more prominent role in the model than the WOZ-waarde. So, the WOZ-waarde on its own might not be that useful when predicting the gap between the WOZ-waarde. Lubberink, Post, and Veuger (2017) already concluded in their paper that the WOZ-waarde on itself should not be used as a market value indicator. The results of this research further underwrite this conclusion.

The second sub-question is as follows: 'What is the impact of market sentiment on this gap?' Investigating market sentiment using the house prices as a dependent variable revealed that a random forest does a pretty good job predicting these house prices. The random forest primarily relies on the variables *inflation* and *supply of houses* to come to its predictions. As the house prices are pretty predictable, market sentiment does not directly lead to sudden, unexpected changes in the price of houses in the Dutch housing market. Market sentiment might still impact the gap because it might influence human behaviour in the market. Increasing prices could lead to people who are more eager to buy, or decreasing prices might make buyers more hesitant. Olaussen, Oust, and Sønstebø (2018) concluded in their paper that bidding behaviour did not change among Norwegian buyers when the market conditions did change. This might be some proof for the statement that changes in human behaviour are only a marginal issue, but the effect of human behaviour is still a bit unclear.

Besides information on properties and their inhabitants, macroeconomic variables and time effects were also included in the models. Waarderingskamer (2019) mentioned that there might be a lagging effect between the WOZ-waarde and the market value, but the time effects did not play a crucial role in this research. Both for the random forest and the neural network, the variables *year*, *financial crisis* or *Covid19* did not significantly impact the predictions. All these time effects had low variable importance. On the other hand, the macroeconomic determinants played a more prominent role. *Inflation* was found to be an excellent indicator with which price changes in the housing market can be predicted.

Weaknesses and future research

There are a couple of areas in which improvements are possible to eventually come to even more detailed conclusions about the gap between the WOZ-waarde and the self-reported market value. The first aspect is the self-reported market value being self-reported in this investigation. Using historical real estate transactions could be an excellent source of information that might reveal more insights into the gap.

Investigating market sentiment by using the prices of houses is also a concept that could be replaced by a different way of reasoning. For example, Hui and Wang (2014) created a sentiment index that measures traders' confidence or optimism in the housing market. Such a sentiment index could also be used in the context of predicting the gap between the (self-reported) market value and the WOZ-waarde.

A third aspect is the unpredictable or irrational part of the gap because the market value is partly based on human behaviour. It is not always clear and obvious why people decide to start over- or underbidding in the Dutch housing market. This gap is also partly a result of behaviour that was not captured or predicted by the models in this paper.

Appendix A: Variable list

	Variable	Description
	Dependent variable	
1	GAP/WOZ	The dependent variable is the relative gap. This is the gap divided by the WOZ-waarde.
	House characteristics	
2	WOZ-waarde	A numeric variable that gives the WOZ-waarde of a respondent's dwelling
3	Satisfied house	A numeric variable that shows on a scale of 0-10 how satisfied one is with his or her dwelling.
4	Satisfied vicinity	A numeric variable that shows on a scale of 0-10 how satisfied one is with his or her neighbourhood.
5	Owner	Variable that groups respondent into either being the owner of the house, tenant or subtenant for example.
6	Year House	A numeric variable that shows in which year the house was bought.
7	Mortgage	A variable that indicates if there is a loan or mortgage on the house.
8	Purchase price	A numeric variable that gives the purchase price of the property.
9	Currency	Indicates the currency of the house's purchase price.
10	Rooms	A numeric variable that gives the number of rooms in the dwelling.
11	Special adjustment	A variable that shows if there are any special adjustments that help people with a disability.
12	Type of house	Variable that groups all houses into different categories based on type.
13	Floors	A numeric variable that indicates the number of floors that are present in the house.
14	Steps	A numeric variable that shows the number of steps one needs to climb up or down to reach street level.
15	Too small	Dummy variable that shows if a house is experienced as being too small.
16	Too dark	Dummy variable that shows if a house is experienced as being too dark.
17	Inadequate heating	Dummy variable that shows if a house is experienced as having inadequate heating.
18	Leaking roof	Dummy variable that shows if a house is experienced as having a leaking roof.
19	Damp walls or floors	Dummy variable that shows if a house is experienced as having damp walls or floors.
20	Rotten frames or floors	Dummy variable that shows if a house is experienced as having rotten frames or floors
21	Too noisy	Dummy variable that shows if a house is experienced as being too noisy.
22	No problems	Dummy that indicates if any of the before mentioned problems are present or not.
23	Noise neighbours	Dummy variable that shows if a house is experienced as having noisy neighbours.
24	Noise environment	Dummy variable that shows if a house's environment is experienced as being noisy.
25	Traffic or industry	Dummy variable that shows if the respondent experiences nuisance caused by traffic or industry.
26	Vandalism or crime	Dummy variable that shows if the respondent experiences problems caused by vandalism or crime.
27	No problem 2	Dummy that indicates if any of the before mentioned problems are present or not.
28	Reside	Variable showing if the respondent regularly resides elsewhere or not.
29	Second dwelling	Dummy indicates if the respondent has a second dwelling or not.
	Background characteristics	
30	Gender	Dummy that explains if the inhabitant is either male or female.
31	Position within household	Variable that groups all individuals according to the different positions within a household.
32	Age	A numeric variable that contains one's age.
33	Number of household members	A numeric variable that shows how many people live in the dwelling.
34	Number of children	A numeric variable that contains the number of children that someone has.
35	Partner	A dummy variable that indicates if someone has a partner or not.
36	Civil status	Variable that groups all individuals according to the different forms of civil status. For example, wedded or unwedded.
37	Primary occupation	A variable that explains the mainstream of income. For example, a full-time or part-time job.
38	Net monthly income	A numeric variable that shows the net monthly income.
39	Household income	A numeric variable that gives the total net income of the entire household.
40	Education	Variable that groups the individuals based on the highest finished education.
41	Origin	Variable that groups the respondents based on the country of origin.
	Time effects	
42	Year	A numeric variable that indicates in which year the survey was filled out by the respondent.
43	Year WOZ	A numeric variable that gives the year for which the WOZ-waarde is valid.
44	Year of residence	A numeric variable that shows in which year one started living in the house
45	Covid19	Dummy variable that indicates that the covid19 crisis hit from 2019 till 2021.
46	Financial crisis	Dummy variable that indicates that the financial crisis hit from 2008 till 2013.
47	Municipality	A numeric variable that indicates since what year one lived in this municipality.
	Macroeconomic variables	
48	House price Netherlands	A numeric variable that gives the indexed differences in house price.*
49	House price Netherlands North	A numeric variable that gives the indexed differences in house prices for houses located in the north.*
50	House price Netherlands East	A numeric variable that gives the indexed differences in house prices for houses located in the east.*
51	House price Netherlands South	A numeric variable that gives the indexed differences in house prices for houses located in the south.*
52	House price Netherlands West	A numeric variable that gives the indexed differences in house prices for houses located in the west.*
53	GDP	A numeric variable that gives the indexed differences in GDP between years.*
54	Inflation	A numeric variable that gives the indexed differences in inflation between years.*
55	Supply houses	A numeric variable that gives the indexed differences in the supply of houses between years.*

* = These indexed variables all take 2015 as their base year.

Appendix B: Results panel regressions

	Dependent variable:		
	GAP_WOZ		
	Pooled OLS	Fixed-effects OLS	Random-effects OLS
	(1)	(2)	(3)
2009		-0.041^{***} (0.009)	
2010		-0.062^{***} (0.009)	
2011		-0.072^{***} (0.009)	
2012		-0.104^{***} (0.009)	
2013		-0.120^{***} (0.009)	
2014		-0.092^{***} (0.009)	
2015		-0.076^{***} (0.009)	
2016		-0.065^{***} (0.009)	
2017		-0.044^{***} (0.009)	
2018		-0.022^{**} (0.009)	
2019		-0.003 (0.010)	
2020		0.019^{*} (0.010)	
2021		0.084^{***} (0.010)	
WOZ_waarde	-0.00000^{***} (0.000)	-0.00000^{***} (0.000)	-0.00000^{***} (0.000)

Table 7: Panel regressions, part 1

	Dependent variable:		
		GAP_WOZ	
	Pooled OLS	Fixed-effects OLS	Random-effects OLS
	(1)	(2)	(3)
SupplyHouses	-0.039^{***}		0.054^{***}
	(0.009)		(0.004)
Covid19	-0.024^{*}		0.085^{***}
	(0.015)		(0.005)
Satisfied_House	0.006^{***}	0.002	0.006^{***}
	(0.001)	(0.002)	(0.001)
Satisfied_Vicinity	0.002^{*}	-0.0001	0.002^{*}
	(0.001)	(0.002)	(0.001)
Owner	-0.009^{***}	-0.017^{***}	-0.010^{***}
	(0.002)	(0.005)	(0.002)
House_Bought_Year	0.00001	0.001	
	(0.0003)	(0.001)	
Mortgage	-0.006	-0.0002	-0.007^{*}
	(0.004)	(0.007)	(0.004)
Purchase_Price	0.00000^{***}	0.00000^{***}	0.00000^{***}
	(0.000)	(0.000)	(0.000)
Currency	-0.010^{**}	-0.001	-0.011^{***}
	(0.004)	(0.009)	(0.004)
Rooms	0.015^{***}	0.014^{***}	0.015^{***}
	(0.001)	(0.003)	(0.001)
Special_Adjustment	0.003	-0.024^{**}	0.003
	(0.005)	(0.011)	(0.005)
Year_Of_Residence	0.0003	-0.00005	
	(0.0003)	(0.001)	
Municipality	-0.0002^{**}	-0.0003	
	(0.0001)	(0.0003)	
Type_Of_House	0.007^{***}	0.009^{***}	0.007^{***}
	(0.001)	(0.002)	(0.001)
Inflation	0.014^{**}		-0.027^{***}
	(0.006)		(0.002)
Floors	0.001	0.003	0.002
	(0.001)	(0.002)	(0.001)

Table 8: Panel regressions, part 2

	Dependent variable:		
	GAP_WOZ		
	Pooled OLS	Fixed-effects OLS	Random-effects OLS
	(1)	(2)	(3)
Steps	0.0002 (0.0001)	0.00002 (0.0002)	0.0001 (0.0001)
Too_Small	0.033*** (0.007)	0.006 (0.009)	0.032*** (0.007)
Too_Dark	0.014 (0.010)	0.006 (0.014)	0.013 (0.010)
Inadequate_Heating	0.014 (0.008)	0.009 (0.011)	0.013 (0.009)
Leaking_Roof	0.005 (0.011)	0.007 (0.014)	0.005 (0.012)
Damp_Walls_Or_Floors	0.013* (0.008)	0.009 (0.011)	0.013* (0.008)
Rotten_Frames_Or_Floors	−0.006 (0.009)	−0.001 (0.011)	−0.007 (0.009)
Too_Noisy	0.001 (0.006)	−0.007 (0.009)	−0.001 (0.006)
No_Problems	0.022*** (0.006)	0.005 (0.008)	0.020*** (0.006)
Noise_Neighbors	−0.006 (0.006)	−0.007 (0.008)	−0.006 (0.006)
Noise_Environment	0.006 (0.006)	−0.0003 (0.007)	0.007 (0.006)
Traffic_Or_Industry	0.013* (0.007)	0.019** (0.009)	0.013* (0.007)
Vandalism_Or_Crime	0.007 (0.006)	0.001 (0.008)	0.006 (0.006)
No_Problems2	0.004 (0.006)	0.003 (0.008)	0.004 (0.006)
Reside	−0.050*** (0.010)	−0.029* (0.015)	−0.047*** (0.010)
Second_Dwelling	−0.022*** (0.008)	−0.027* (0.016)	−0.022** (0.008)

Table 9: Panel regressions, part 3

	Dependent variable:		
	GAP_WOZ		
	Pooled OLS	Fixed-effects OLS	Random-effects OLS
	(1)	(2)	(3)
Gender	−0.005	0.014	-0.009^{***}
	(0.003)	(0.012)	(0.003)
Position_Household	0.002	0.062^{***}	0.002
	(0.002)	(0.015)	(0.002)
Age	0.0001	0.002	0.00000
	(0.0002)	(0.001)	(0.0002)
Nr_Household_Memebers	−0.004	0.004	−0.006
	(0.006)	(0.016)	(0.006)
Nr_of_Childeren	0.001	−0.029	0.002
	(0.006)	(0.019)	(0.006)
Partner	0.007	-0.061^{***}	0.008
	(0.007)	(0.023)	(0.007)
Civil_Status	−0.0001	-0.029^{***}	−0.002
	(0.001)	(0.008)	(0.001)
Primary_Occupation	0.003^{***}	0.013^{***}	0.003^{***}
	(0.001)	(0.004)	(0.001)
Net_Monthly_Income	-0.00001^{***}	0.0001^{***}	-0.00001^{***}
	(0.00000)	(0.00001)	(0.00000)
Household_Income	0.00002^{***}	0.0001^{***}	0.00002^{***}
	(0.00000)	(0.00001)	(0.00000)
Education	0.009^{***}	0.013	0.009^{***}
	(0.001)	(0.009)	(0.001)
Origin	0.00005	0.0004	0.00005
	(0.00003)	(0.0003)	(0.00003)
FinancialCrisis	-0.071^{***}		
	(0.011)		
Constant	−0.030		-2.605^{***}
	(0.429)		(0.198)
Observations	50,630	50,630	50,630
R^2	0.123	0.166	0.118
Adjusted R^2	0.122	−0.040	0.117
F Statistic	128.806^{***}	147.107^{***}	$6{,}721.592^{***}$
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 10: Panel regressions, part 4

Results are based on calculations by Rins Lukasse (Student at Erasmus University Rotterdam) using public (micro)data from Statistics Netherlands

References

Adelino, Manuel, Antoinette Schoar, and Felipe Severino. 2012. "Credit Supply and House Prices: Evidence from Mortgage Market Segmentation." *Nber.org*. https://doi.org/10.3386/w17832.

Anglin, Paul M, Ronald Rutherford, and Thomas M Springer. 2003. "The Trade-Off Between the Selling Price of Residential Properties and Time-on-the-Market: The Impact of Price Setting." *The Journal of Real Estate Finance and Economics* 26 (1): 95–111.

Blanco-Justicia, Alberto, and Josep Domingo-Ferrer. 2019. "Machine Learning Explainability Through Comprehensible Decision Trees." *Lecture Notes in Computer Science*, 15–26. https://doi.org/10.1007/978-3-030-29726-8_2.

Boelhouwer, Peter, and Joris Hoekstra. 2009. "Towards a Better Balance on the Dutch Housing Market? Analysis and Policy Propositions." *European Journal of Housing Policy* 9 (4): 457–75. https://doi.org/10.1080/14616710903357235.

Buuren, Stef. 2021. *Flexible Imputation of Missing Data*. Chapman &amp; Hall/CRC.

Campbell, John Y., and João F. Cocco. 2007. "How Do House Prices Affect Consumption? Evidence from Micro Data." *Journal of Monetary Economics* 54 (3): 591–621. https://doi.org/10.1016/j.jmoneco.2005.10.016.

Candel, Arno, and Erin LeDell. 2017. "Deep Learning with H2o." Edited by AngelaEditor Bartz. *H2o.ai, Inc.* Sixth Edition (November).

CBS. 2022. *CBS Statline*. https://opendata.cbs.nl/CBS/nl/navigatieScherm/thema?themaNr=51440.

Čeh, Marjan, Milan Kilibarda, Anka Lisec, and Branislav Bajat. 2018. "Estimating the Performance of Random Forest Versus Multiple Regression for Predicting Prices of the Apartments." *ISPRS International Journal of Geo-Information* 7 (5): 168. https://doi.org/10.3390/ijgi7050168.

Doney, Patricia M., and Joseph P. Cannon. 1997. "An Examination of the Nature of Trust in Buyer-Seller Relationships." *Journal of Marketing* 61 (2): 35. https://doi.org/10.2307/1251829.

Dubin, Robin A. 1998. "Predicting House Prices Using Multiple Listings Data." *The Journal of Real Estate Finance and Economics* 17 (1): 35–59. https://doi.org/10.1023/a:1007751112669.

Englund, Peter, and Yannis M. Ioannides. 1997. "House Price Dynamics: An International Empirical Perspective." *Journal of Housing Economics* 6 (2): 119–36. https://doi.org/10.1006/jhec.1997.0210.

Favara, Giovanni, and Jean Imbs. 2015. "Credit Supply and the Price of Housing." *American Economic*

Review 105 (3): 958–92. https://doi.org/10.1257/aer.20121416.

Genesove, David, and Christopher Mayer. 2001. "Loss Aversion and Seller Behavior: Evidence from the Housing Market." https://doi.org/10.3386/w8143.

H2O-Innovation-Inc. 2016. "H2o Innovation Inc, Canada." *Filtration Industry Analyst* 2016 (10): 11. https://doi.org/10.1016/s1365-6937(16)30247-7.

Hanno. 2022. "Buying a House in the Netherlands in 10 Steps:" *Hanno*. https://www.hanno.nl/expat-mortgages/buying-a-house-in-the-netherlands/.

Himmelberg, Charles, Christopher Mayer, and Todd Sinai. 2005. "Assessing High House Prices: Bubbles, Fundamentals and Misperceptions." *Journal of Economic Perspectives* 19 (4): 67–92. https://doi.org/10.1257/089533005775196769.

Ho, Winky K. O., Bo-Sin Tang, and Siu Wai Wong. 2020. "Predicting Property Prices with Machine Learning Algorithms." *Journal of Property Research* 38 (1): 48–70. https://doi.org/10.1080/09599916.2020.1832558.

Hochstenbach, Cody, and Rowan Arundel. 2019. "Spatial Housing Market Polarisation: National and Urban Dynamics of Diverging House Values." *Transactions of the Institute of British Geographers* 45 (2): 464–82. https://doi.org/10.1111/tran.12346.

Hui, Eddie Chi-man, and Ziyou Wang. 2014. "Market Sentiment in Private Housing Market." *Habitat International* 44: 375–85. https://doi.org/10.1016/j.habitatint.2014.08.001.

Jin, Changha, Gökçe Soydemir, and Alan Tidwell. 2014. "The u.s. Housing Market and the Pricing of Risk: Fundamental Analysis and Market Sentiment." *Journal of Real Estate Research* 36 (2): 187–220. https://doi.org/10.1080/10835547.2014.12091390.

Kingsford, Carl, and Steven L Salzberg. 2008. "What Are Decision Trees?" *Nature Biotechnology* 26 (9): 1011–13. https://doi.org/10.1038/nbt0908-1011.

Knight, John R. 2002. "Listing Price, Time on Market, and Ultimate Selling Price: Causes and Effects of Listing Price Changes." *Real Estate Economics* 30 (2): 213–37. https://doi.org/10.1111/1540-6229.00038.

Little, Roderick, and Donald Rubin. 2019. "Statistical Analysis with Missing Data, Third Edition." *Wiley Series in Probability and Statistics*. https://doi.org/10.1002/9781119482260.

Lubberink, A. S., J. van der Post, and W. Veuger. 2017. "De WOZ-Waarde Als Marktwaarde-Indicator. Real Estate Research Quarterly." *Real Estate Research Quarterly* 16 (4): 28–37. https://doi.org/10.1257/089533005775196769.

Maskara, Nishad, Aleksander Kubica, and Tomas Jochym-O'Connor. 2019. "Advantages of Versatile Neural-Network Decoding for Topological Codes." *Physical Review A* 99 (5). https://doi.org/10.1103/physreva.

99.052351.

Olaussen, Jon, Are Oust, and Ole Sønstebø. 2018. “Bidding Behavior in the Housing Market Under Different Market Regimes.” *Journal of Risk and Financial Management* 11 (3): 41. https://doi.org/10.3390/jrfm11030041.

Ortalo-Magne, Francios, and Sven Rady. 2006. “Housing Market Dynamics: On the Contribution of Income Shocks and Credit Constraints*.” *Review of Economic Studies* 73 (2): 459–85. https://doi.org/10.1111/j.1467-937x.2006.383_1.x.

Scherpenzeel, and Das. 2010. “““True’ Longitudinal and Probability-Based Internet Panels: Evidence from the Netherlands.” *Social and Behavioral Research and the Internet*, 77–104. https://doi.org/10.4324/9780203844922-4.

Sharma, Siddharth, Simone Sharma, and Anidhya Athaiya. 2020. “Activation Functions in Neural Networks.” *International Journal of Engineering Applied Sciences and Technology* 04 (12): 310–16. https://doi.org/10.33564/ijeast.2020.v04i12.054.

Stiglitz, Joseph E. 1990. “Symposium on Bubbles.” *Journal of Economic Perspectives* 4 (2): 13–18. https://doi.org/10.1257/jep.4.2.13.

Veldhuizen, Sander van, Benedikt Vogt, and Bart Voogt. 2016. “Internet Searches and Transactions on the Dutch Housing Market.” *Applied Economics Letters* 23 (18): 1321–24. https://doi.org/10.1080/13504851.2016.1153785.

Waarderingskamer. 2019. “Waarderingsinstructie.” *Waarderingskamer* 4 (2): 110–11. https://doi.org/10.1257/jep.4.2.13.