The effect of location and environmental factors on the value of dwellings

Van Mulken, M.L.P.G.

Award date:
2019

THE EFFECT OF LOCATION AND ENVIRONMENTAL FACTORS ON THE VALUE OF DWELLINGS.

GRADUATION PROJECT M.L.P.G. VAN MULKEN

STUDENT NUMBER: 0750766

Technische Universiteit
Eindhoven
University of Technology

MASTER: ARCHITECTURE, BUILDING AND PLANNING
MASTER TRACK: URBAN SYSTEMS AND REAL ESTATE

1ST SUPERVISOR: PROF. DR. T.A. ARENTZE
2ND SUPERVISOR: DRS. IR. M.I.K. LEUSSINK
3RD SUPERVISOR: ASSOC. PROF. Q. HAN

DATE: JANUARY 29, 2019

PREFACE

This document is the result of my Graduation Project at the Eindhoven University of Technology. This graduation project turned out to be a long journey where collecting all the data was a pittance compared to merging and analysing the huge amount of data. Not only the amount of cases was a lot with over 26,000 cases, but the amount of variables was also very extensive. After the first commonly known analysis a hard process followed, for a layman, of applying datamining techniques. The results, with refreshing, new insights made everything more than worth it.

I would like to thank my supervisors for their professional counselling during my process. I would like to thank Drs. ir. M.I.K. Leussink for her pleasant counselling. She has a lot of knowledge about the process of a research and counselled me in a kindly way. I also would like to thank Assoc. Prof. Q. Han. Although she entered in a late phase of the process, I would like to thank her for her feedback. A special thanks is for my first supervisor Prof. Dr. T.A. Arentze. He always stayed calm and patient when I experienced problems with the datamining techniques. I could not find my way around with them in the beginning. My frustration had however no change against his calm personality.

To all readers, enjoy reading my thesis.

SUMMARY

INTRODUCTION

Correct valuation of the value of a dwelling is an important aspect for many municipalities, institutions, businesses and private persons. It can help with, for example, determining the value of property taxes. For determining the value of a dwelling it is important to know which factors determine the value of a dwelling. Ask any real estate agent what these factors are and the answer will probably be “location, location, location”. That means that besides the characteristics of the dwelling itself, also the location of a dwelling has a big influence on the value of a dwelling. But when the value of a dwelling is determined in practice, specific spatial factors are often not taken into account, only the location in general. Not knowing which specific spatial factors have an influence on the value of a dwelling can partly be explained by the absence of private and public data. Because of this absence it is hard to do research about spatial factors influencing the value of a dwelling. There are however done some scientific studies about value determination of dwellings and there are several studies that found spatial factors that have an influence on the value of a dwelling. A literature study about this topic has been done to look into the information that is already known. The previously called term spatial factors will be divided in location and environmental factors. Location factors are factors that consider the location of a dwelling in relation to specific facilities.

This factor is often expressed in distance. Examples are distance to highways and distance to schools. Environmental factors are quality characteristics that are in the near environment of a dwelling. Examples are the amount of green in the area and the density of the population. An overview is created about what is already known about the influence of location and environmental factors on the value of a dwelling but also about the characteristics of the dwelling itself. This overview is shown in Table 1.

Table 1 An overview of influencing factors found in the literature study.

Characteristics of the Dwelling Itself		**Location Factors**	**Environmental Factors**
Type of dwelling	Condition of dwelling	Good accessibility	Good traffic conditions
Age of the dwelling	Quality (luxurious)	Distance to CBD	Good parking facilities
Floor level of apartment	Presence of an elevator	Distance to school zone	Presence of green areas
Lot size	Presence of central heating	Distance to shopping centre	Presence of Recreational green
Volume	Presence of air conditioning	Distance to high school	Presence of shops
Floor area	Presence of fireplace	Distance to high way	Presence of schools
Presence and size of garage	Presence of pool	Distance to public transport	Quality of primary schools
Number of bathrooms	Presence of insulation	Distance to city centre	Type of neighbourhood
Number of rooms	Presence of plumbing	Distance to hospital	Degree of urbanization
Presence of garden	View of green spaces and water elements	Distance to hotel and catering industry	Density of the neighbourhood
Presence and size of the balcony	Windows with a south orientation	Distance to sport facilities	Compilation of population
	Sea view	Distance to coast	Average income in neighbourhood
		Climate	Quality of real estate in neighbourhood
		Size of island (when applicable)	Percentage rental and private owned
		Earthquake risk	Criminality

The literature study showed that the hedonic price analysis is the most used valuation method. This method is based on the idea that an item's total price can be considered as the sum of prices of attributes. The hedonic price analysis uses a regression analysis to determine the coefficient of every attribute.

Nowadays there are much more data available and also data mining techniques that can be used to analyse these big data sets. Data mining techniques are techniques that search for (statistical) relationships in data sets with the aim of pattern recognition. The use of data mining techniques when determining the value of a dwelling is also taken into account in the literature review and it shows that the use of data mining techniques is quite unknown and not much used for valuation of real estate yet. There are however some studies about this topic, but in all these studies the spatial factors are not specified or there is not much data used. The biggest study that was found contains 3,000 cases. The availability of new data and data mining techniques can help to determine which spatial factors have a positive or negative influence on the value of a property. Besides detecting which factors have an influence on the value of a dwelling, it is also interesting to investigate what the quantitative influence of these factors is and what the willingness to pay for these factors is. This leads to the research objective of this research. *Collect data about the value of a dwelling and spatial factors to analyse which influence specific spatial factors have on the value of a dwelling and what the willingness to pay is. In that way, a more complete housing valuation model can be created.*

Figure 1 shows the conceptual model of the research. The research uses several clusters of variables. Besides location and environmental factors, the characteristics of the dwelling itself also have to be taken into account. About the influence of the characteristics of the dwelling itself is much more information available and there will be little news to discover. But for the completeness of the research it is important to include these characteristics. Scarcity is another factor that will be used in the research. The value of a dwelling that is determined in this research is the market value. This is the selling price of a dwelling under normal selling conditions. The selling price is the maximum amount that people who buy a dwelling are willing to pay for the dwelling. The aim of this research is to examine the influence of characteristics of the dwelling itself, location factors and environmental factors on the value of a dwelling. Besides that, the influence of scarcity on the value of a dwelling

will also be examined. Not only commonly known research techniques will be used in this research but the possibility of data mining techniques that can be used for analysing the contribution of location and environmental factors to the value of a dwelling will also be investigated. This leads to the main research question; *What is the influence of location and environmental factors on the value of a dwelling?*

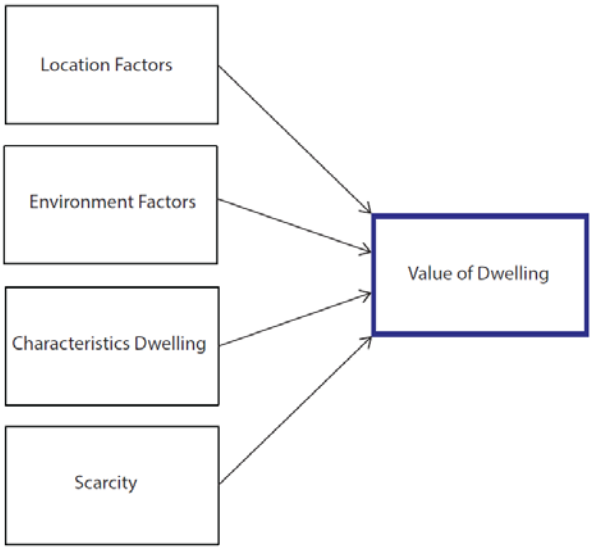

Figure 1 The conceptual model

METHOD

As stated before, this research investigates the influence of location and environmental characteristics on the value of a dwelling and whether scarcity has an influence on the value of a dwelling. The characteristics of the dwelling itself are also taken into account. To get insight in the influence of location and environmental characteristics on the value of a dwelling, a data set has to be created that combines characteristics of a dwelling itself together with location and environmental characteristics, scarcity and the transaction data of a dwelling. To realize this, nine municipalities in the west of the province of Noord-Brabant in The Netherlands have provided a data

set with information about dwellings that have been sold in the years 2011 up until 2017 in these municipalities. This data set contains information about characteristics of the dwellings itself and the selling price of the dwellings. After receiving the data set, the data set is extended with location and environmental factors and the scarcity in the region where the dwelling is located. After all the different types of data are combined, the data set is cleaned. A part of the data is transformed so the information can optimally be used; some variables are recoded into variables with larger categories, some variables are recoded into dummy variables and some variables are transformed into more meaningful variables. The variable transaction price of a dwelling is also transformed. A logarithm transformation is executed on the variable. Then the coefficients of the independent variables show the percentage influence of a variable on the value of a dwelling. After that, the cases that can be used for the analysis and which ones not are selected. Some variables have outliers that will be filtered out. Some cases will be removed from the data set because they are not included in the population of the research. Examples are, dwellings sold before 2016 or dwellings sold to family members. After selecting all the cases, 9,616 cases are left for the research.

At last the variables are checked to see if there are variables that represent the same factors and if there are variables that correlate with each other. The correlation between the variables is checked and variables that have a high correlation are transformed into components with the use of a principal component analysis. With this technique, the amount of variables in an analysis can be reduced to main components, the so called, principal components. The components that are the result of a principal component analysis do not correlate with each other. Eight principal component analysis are executed and in total, 80 variables are reduced to 15 components. The final data set contains 70 independent variables. 32 variables contain information about characteristics of the dwelling itself, 11 about environmental factors, 23 about location factors, 3 variables about the circumstances of the transaction and 1 about scarcity.

Three analytical techniques are selected to analyse which location and environmental factors have an influence on the value of a dwelling. These techniques are the hedonic price analysis (based on a regression analysis), a regression tree and a random forest. The hedonic price analysis is based on a regression analysis. A regression tree induction method is a machine learning algorithm that divides the data into subsets. These subsets are created by splitting the data set several times. Each time the data set is split, the residual sum of squared error (RSS) of the two remaining data sets is minimized. The methodology that is used is CART. Graphical models are created that predict outcomes based on the classification of cases using a tree structure. Various branches of variable length are formed and the most important factors are located at the top of the tree. When a model is derived with the use of a random forest, 500 regression trees are created with randomly chosen variables and cases. Combining all these regression trees gives an insight in which variables have the biggest influence on the value of a dwelling. The advantage of this method is that it overcomes problems that occur with single regression trees. These problems are multicollinearity and overfitting.

RESULTS

The results of the regression analysis show that the characteristics of the dwelling itself have the biggest influence on the value of a dwelling, especially the volume of a dwelling. But this information is not new. The most important environmental factors that have an influence on the value of a dwelling are; the amount of green in a neighbourhood, the population compilation and built up environment in a neighbourhood and the distance to daily facilities. The value of a dwelling increases when the amount of green increases and when the distance to daily facilities decreases. The neighbourhood where a dwelling is located has a positive influence on the value of a dwelling when the neighbourhood is characterised with a high percentage privately owned, single family houses and families with young children. The location factors that have the highest influence on the value of a dwelling are; distance to theme park and sauna, amount of daily facilities within a 3-5 km

radius and the amount of non-daily facilities within a 20-50 km radius. The variables distance to theme park and sauna probably represent another variable. They represent the distance to the nearest big city. When a dwelling is located closer to a big city, the value of a dwelling increases. It has a negative effect on the value of a dwelling when the amount of daily facilities within a 3-5 km radius increases, but a positive effect when the amount of non-daily facilities within a 20-50 km radius increases. The variable shortage also has an influence on the value of a dwelling, when there are less dwellings available for people when they are buying a house, the value of a dwelling increases. A result that is not mentioned in the literature yet is that a difference can be made according to the availability of facilities. This research shows that the distance to daily facilities is important and the amount of non-daily facilities in a specific radius are important. The regression model explains 84.1% of the variance, which is a good fit.

The results of the regression tree show that the most important variables in this regression tree are again characteristics of the dwelling itself. The location and environmental factors that have an influence on the value of a dwelling are; the population compilation and built environment of a neighbourhood, distance to leisure and high schools, the amount of facilities within a 10 km radius, the percentage green and the distance to daily facilities. The regression tree does not only show which variables have an influence on the value of a dwelling but also what this influence is. The value of a dwelling decreases when the population compilation and built up environment is characterised by high percentages vacant dwellings, inhabitants in the age category 15-24 years old and immigrants with a Western ethnicity and a low percentage of households with children. The value of a dwelling also decreases when the distance to leisure and high schools increases and when the distance to daily facilities decreases. The value of a dwelling increases when the amount of facilities within a 10 km radius increases and when the amount of green in a neighbourhood increases. The tree structure also shows which variables are important for specific sub groups. For smaller

dwellings, characteristics of a dwelling itself have a big influence on the value of a dwelling and location and environmental factors almost not. For bigger dwellings, location and environmental factors have together with characteristics of the dwelling itself an influence on the value of a dwelling. 75.3% of the variance is explained with this technique.

The results from the random forest model show again, the characteristics of the dwelling itself have the biggest influence on the value of a dwelling. The location factors that have the most influence on the value of a dwelling are; distance to an ice skating rink, distance to leisure and high schools, amount of facilities within a 10 km radius, distance to a railway station and cinema, distance to a sauna, the amount of facilities within a 20 to 50 km radius, distance to a high way and distance to a theme park. The environmental factors that have the most influence on the value of a dwelling are; the population compilation and built environment of a neighbourhood, distance to daily facilities, distance to a school, a day care and an after school care and the percentage green in a neighbourhood. Also the variable shortage has an influence on the value of a dwelling. Again, a difference can be made according to the availability of daily and non-daily facilities. The distance to daily facilities is important and the amount of non-daily facilities. 87.0% of the variance is explained with the use of a random forest. That makes it the technique with the best fitting model, but the random forest does not provide an insight in which way characteristics of the dwelling itself, location factors, environmental factors and scarcity have an influence on the value of a dwelling. This is done best by using a regression model. The disadvantage of the regression model is that it does not make a distinction between different sub groups. For retrieving an insight in which variables have an influence on the value of a dwelling in different sub groups, the regression tree can be used best.

CONCLUSIONS

From this research can be concluded that the characteristics of a dwelling have the biggest influence on the value of a dwelling. Location and environmental factors also have an influence on the value of a dwelling. The factors that have an influence on the value of a dwelling and what that influence is has been found in this research. The best of the three methods can be combined to retrieve an optimal result. The random forest model has the highest fit and the importance of variables is shown. With the use of the regression model can be seen if factors have a positive or negative influence on the value of a dwelling and how big their influence is. The regression tree model can be used to see what the influence on the value of a dwelling is in different sub groups. With the use of the three analysing techniques, a valuation tool can be created to determine the qualities of locations and neighbourhoods. The variables that have an influence on the value of a dwelling are indicators for the quality of a location.

With regards to future research, a recommendation is that a bigger research area is useful for a higher variance in specific factors. It is also interesting to take target groups into consideration and do the research again in a specific amount of time, to see if preferences of people have changed. There is also still some room for adding new variables, since the 'distance to the nearest big city' is now represented by the variables; distance to theme park, distance to swimming pool and distance to sauna. At last should be mentioned that this research is representative for The Netherlands, it is unknown how representative this is for other countries.

TABLE OF CONTENT

1
Introduction and Problem Definition

Determination of the value of housing is an important aspect for many municipalities, institutions, businesses and private persons. Municipalities in the Netherlands collect local taxes. If people own a house, they have to pay additional taxes for that. This is called property tax. Municipalities use a fixed percentage of the value of houses to determine the height of the property tax. Municipalities use these taxes to maintain local services (Rotterdam Centre, 2018).

Also social housing institutions are interested in the value of housing. Several of these institutions have stated that they want to map the current needs of their tenants and the current qualities of their dwellings. In this way the relationship between supply and demand can be improved and the dwellings meet the needs of the tenants (Cobalt Recruitment, 2018).

Another type of institute that is interested in the value of housing are investment businesses. These businesses want to know which type of housing can offer an optimal profit for their organization. When an investment business has an overview of the value of specific types of dwellings, they can calculate in which type of housing they should invest to retrieve an optimal profit.

Finally, private individuals are interested in the value of housing. Not only private individuals who buy a dwelling for the purpose of rental in the private sector or selling it with profit are interested in the value of dwellings. The valuation of housing is also important for people who buy a house for private use. People who buy or sell a house want to know that the price they will pay or receive for the house is accurate.

1.1 Problem Definition

As stated before, several municipalities, companies, institutes and private persons are interested in correct valuation of housing. But how is this valuation of housing in practice determined? Ask any real estate agent what determines the value of a dwelling and the answer will probably be “location, location, location”. This is of course just a statement and it means that the location of a house has a big influence on the value of the house. The value of identical properties can increase or decrease as a result of their location (The Balance, 2017). The location of a dwelling plays therefore an important role in decision making when people buy a dwelling, but also the characteristics of a dwelling itself are important (Vereniging Eigen Huis, 2018). When people buy a dwelling they do not only buy the dwelling itself but also the location. And where there is mostly room for adjustments on the dwelling itself, the location of a dwelling is unchangeable (Ruimtelijk Planbureau, 2006).

So both, characteristics of the dwelling itself and the spatial factors of a dwelling are important in determining the value of a dwelling. Some characteristics have a positive influence on the value, others a negative influence. All these characteristics together determine the total value of a dwelling. To get a clear picture of the value of a dwelling it is important to map which characteristics of a dwelling have which specific effect on the value of a dwelling. This type of valuation can be helpful

for several cases. Some examples are; objective valuation of a dwelling, investment analyses and determination of property tax (Van Sprundel, 2014).

The characteristics that are taken into account when a real estate agent determines the value of a dwelling are mainly the characteristics of the dwelling itself (Vereniging Eigen Huis, 2018). And also determination of property tax is mainly based on characteristics of the dwelling itself (Kadaster, 2018). Examples are plot size, the volume of a dwelling and the presence of a garden. Specific spatial factors of a house are not often included in practice in determining the value of housing. The location of a house is taken into account in determining the value of a dwelling, but the value is based on values of other dwellings in proximity of that location. Vereniging Eigen Huis (2018) states that if people want to know the value of a dwelling they are interested in, they should take the average square meter price of other dwellings in that street and multiply that with the amount of square meters of the house they are interested in. In that way, the location of a dwelling is taken into account, but it is unknown which specific spatial factors influences the value of a dwelling.

Not knowing which specific spatial factors influence the value of a dwelling can partly be explained by the absence of private and public data. Because of this absence it is hard to do research about spatial factors influencing the value of a dwelling. There are however some scientific studies about value determination of dwellings and there are several studies that found spatial factors that have an influence on the value of a dwelling. The presence of green, the type of neighbourhood, a good accessibility and the distance to a central business district are the most important spatial factors found in scientific studies. But also in the studies that already have been done, the availability of private and public data is a problem. This made it hard to investigate the specific influence of spatial factors on the value of a dwelling. This will all be further discussed in chapter 2. Nowadays there is much more data available and also much more data mining techniques that can be used to analyse

this data. Data mining techniques are techniques that search for (statistical) relationships in data sets with the aim of creating profiles. The availability of new data and data mining techniques can help to determine which characteristics of housing have a positive or negative influence on the value of a property. But also these data mining techniques are not much used yet in scientific studies.

1.2 Research Objective

A lot of organizations could benefit from good valuation of housing. To achieve this it is required to determine which characteristics of a dwelling have an influence on the value of a dwelling. In practice, the current housing valuation models only focus on the characteristics of the dwelling itself (De Hypotheker, 2018; Ruimtelijk Planbureau, 2006; Vereniging Eigen Huis, 2018; and Kadaster, 2018). Through previous studies it can be concluded that the value of a dwelling is not only determined by characteristics of the dwelling itself but also by its spatial factors (Hill and Melser, 2007; d'Amato, 2010; Ranzato, 2013; DeSimone, 2013). Currently spatial factors are however in practice almost always only taken into account as one factor, namely 'location'. It is interesting to investigate which specific spatial factors have an influence on the value of a dwelling. Besides that it is also interesting to investigate what the quantitative influence of these factors is. The quantitative influence of a specific factor on the value of a dwelling can be expressed in terms of 'willingness to pay'. Willingness to pay is an indicator that shows how much people are willing to spend on something. For example, most people would appreciate living in a neighbourhood with an extensive amount of green (Zoppi et. al., 2014). But when they buy a dwelling, how much money are they willing to spend to live in a neighbourhood with a lot of green areas? In other words, what is their willingness to pay for living in a green neighbourhood? Further research is needed to get better insight in the influence of spatial factors and the willingness to pay for these factors. Next, the research objective of this research is formulated.

Collect data about the value of a dwelling and spatial factors to analyse which influence specific spatial factors have on the value of a dwelling and what the willingness to pay is, so a more complete housing valuation model can be created.

1.2.1 VARIABLES

To do further research some terms have to be specified or distinguished. First the previously named term spatial factors will be divided in location and environmental factors. Location factors are factors that consider the location of a dwelling in relation to specific facilities. This factor is often expressed in distance. Examples are distance to highways, schools and central business districts. Environmental factors are quality characteristics that are in the near environment of a dwelling. Examples are the amount of green in the area and the density of the population. But besides location and environmental factors, the characteristics of the dwelling itself also have to be taken into account. About the influence of the characteristics of the dwelling itself is much more information available and there will be little news to discover. But for the completeness of the research it is important to include these characteristics.

Another factor that will be used in this research is scarcity. The expectation is that scarcity has an influence on the value of a dwelling as well. When an apartment is located in an area with only a small amount of apartments, the value of that apartment can increase because a lot of people want an apartment but there are not many available. When the demand is high and the supply low, economic models learn that the price of an object will increase (Kramer,2018). This is also the expectation in valuation of housing. At last the term 'value of a dwelling' will be defined. The value of

a dwelling that is determined in this research is the market value. This is the selling price of a dwelling under normal selling conditions. Examples of selling condition that are not normal is when a dwelling is sold to family or when the sale is the result of a foreclosure. The selling price is the maximum amount that people who buy a dwelling are willing to pay for the dwelling (De Hypotheker, 2017).

1.2.2 Research Question

The aim of this research is to examine the influence of characteristics of the dwelling itself, location factors and environmental factors on the value of a dwelling. Besides that the influence of scarcity on the value of a dwelling will also be examined. The influence of each factor on the value of a dwelling will be expressed in willingness to pay. Not only commonly known research techniques will be used in this research but the possibility of data mining techniques that can be used for analysing the contribution of location and environmental factors to the value of a dwelling will also be investigated. This leads to the main research question as described below.

What is the Influence of Location and Environmental Factors on the Value of a Dwelling?

To answer the main research question sub questions, as described next, will be answered step-by-step;

-How can data sets be created with information about the value of a dwelling, characteristics of the dwellings itself, location factors, environmental factors and scarcity?

-How can information about the influence of location and environmental factors be extracted from these data sets? Which data mining methods can be used?

-Which location and environmental factors have an influence on the value of a dwelling?

-What quantitative effect do these location and environmental factors have on the value of a dwelling?

-What is the influence of scarcity on the value of a dwelling?

1.2.3 Conceptual Model

Figure 1.1 shows the conceptual model of the research. As stated before, this research investigates the influence of location and environmental characteristics on the value of a dwelling, whether scarcity has an influence on the value of a dwelling and what the willingness to pay for these characteristics is. The characteristics of the dwelling itself also have to be taken into account for the completeness of the research. The value of a dwelling is in this research defined as the market value. This is the selling price of a dwelling under normal selling conditions. The aim of this research is to examine the influence of location and environmental factors on the value of a dwelling. The value of a dwelling is therefore the dependent variable. Characteristics of the dwelling itself, location factors, environmental factors and scarcity are the independent variables. The willingness to pay of the factors is in the conceptual model represented by the arrows from the independent variables to the dependent variable.

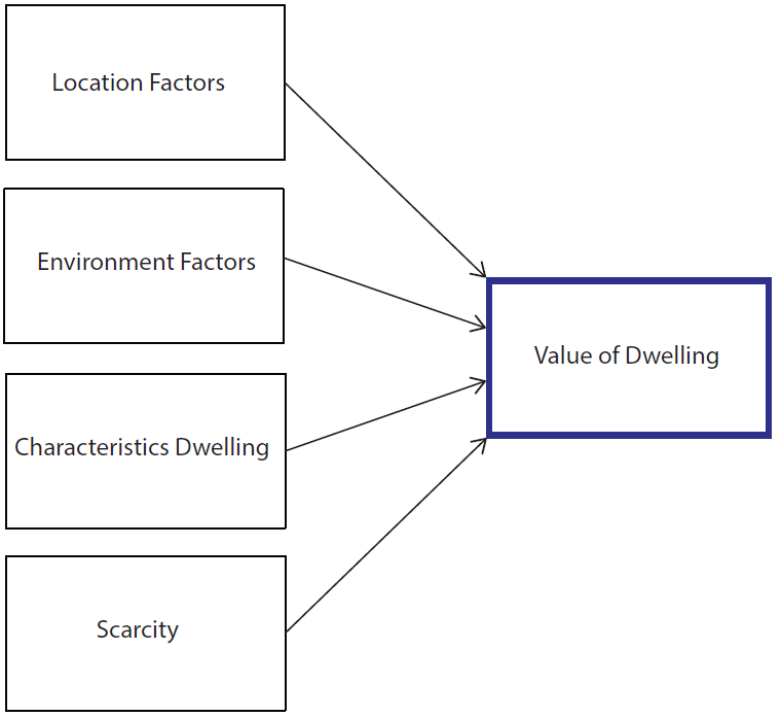

Figure 1.1 The conceptual model

1.3 Relevance

A distinction can be made between the theoretical relevance and the practical relevance of this research.

1.3.1 Theoretical Relevance

The proposed research is theoretically relevant because previous research has proven that there is a relation between the value of a dwelling and its location and environmental characteristics. The availability of data strongly increased the last years so there is more data available to do research about this topic. There is a gap in the literature when it comes to a proper valuation tool that includes all these location and environmental characteristics and there is room for data mining

techniques that can be explored. There is also a gap in the literature when it comes to information about the 'willingness to pay' for these characteristics. As discussed earlier, studies have been done already about this topic but only smaller parts of the research question have been investigated. It is interesting to create a better insight in which specific location and environmental factors have which specific influence on the value of housing. Finally the use of data mining techniques is very promising but quite unknown and not much used yet. There is literature available that points out the opportunities that data mining techniques offer. Real estate is put forward as one of the subjects that can be used in combination with data mining techniques (Yanchang and Yonghua, 2013; and Watada, 2012). There are already a few studies that used data mining techniques in combination with real estate, but also in these studies the spatial factors are not specified and there is a lack of available data. This will be discussed further in chapter 2.5.

1.3.2 Practical Relevance

The research is also practically relevant. Municipalities have to determine the value of housing every year because of the property tax that homeowners have to pay. In practice, only characteristics of the dwelling itself are often taken into account and not the location and environmental characteristics. For example, the municipality of Tilburg has mentioned that they want a property valuation model that takes into account the characteristics of the dwelling itself as well as the location and environmental characteristics of a dwelling. A good property valuation model is also interesting for future housing development projects. When a new project is set up, it is interesting to understand which plan has the highest profit for the organization. This could mean to build as many houses as possible. However for example, when the willingness to pay for living in a green area is high, it can be interesting to create some green open spaces in the area. In the end this can result in a higher profit and a better fit of supply and demand. So when the research shows what the influence of specific location and environmental characteristics on a dwelling is together with the

willingness to pay for these characteristics, this can be taken into account in new housing development projects. Also housing associations have stated that they are interested in the influence of characteristics on the value of a dwelling. They are interested in this so they can map qualities of locations. When this is known they can better anticipate on the needs of tenants. At last is the research practically relevant because it shows the possibilities of data mining techniques in relation to real estate and especially in creating a proper housing valuation model.

1.4 Research Procedure Thesis Outline

To execute this research, various steps are taken. The steps taken in this research, that will be described in this section are in the same order as the structure of the paper. The research starts with a literature study to define the problem. Not only should the problem be defined, but also must be investigated how the problem has arisen and how it can be resolved. In case of this research, a part of the problem can be resolved because of new data mining techniques and the amount of data that is available nowadays. Then the literature study will be resumed to obtain more information about the influence of location and environmental factors that influence the value of a dwelling. The literature study focusses on several areas. Examples are; from which variables do we know they influence the value of a dwelling?; what is known about scarcity?; which research techniques can be used in this research?; what are data mining techniques and how can these be used in this research? The context that will be created in this literature study will be described in chapter 2.

After creating the context of the research, data about the value of dwellings, the characteristics of the dwellings itself, location factors, environmental factors and information about scarcity should be collected. This will be described in chapter 3. The nature of the variables has to match with the chosen techniques. When all the data are collected, one big data set has to be created. This data set

has to contain all the obtained information per dwelling. When the total amount of information has been collected in one data set, the data set has to be cleaned up.

Chapter 4 will describe the hedonic price analysis. First the method will be described, then the analysis procedure and finally the results of the hedonic price analysis. In chapter 5 these steps will be described for data mining techniques. The results of the data mining techniques are considered to see whether the accuracy of the prediction of the hedonic price analysis can be increased. The paper will be sealed with a conclusion and discussion in chapter 6.

2
LITERATURE STUDY OF INFLUENCING SPATIAL FACTORS

The context of location and environmental factors influencing the value of a dwelling should be clear before the analyses are made. It is therefore necessary to look into the information that is already known in various study areas. A definition has to be made about the term 'value of a dwelling' and an overview should be created about what is already known about the influence of location and environmental factors on the value of a dwelling. At first, some qualitative results will be discussed and secondly some quantitative results. An analysis can then be made about the currently known information and the gaps in the information that have been found. In the last section a literature study about data mining techniques is discussed. A closer look will be taken to what data mining techniques exactly are and how they are used up till now to predict the value of real estate.

As stated before is it important to not only take into account the location and environmental factors in this research but also the characteristics of a dwelling itself. This is for the completeness of the research and to exclude correlations that are due to separate effects. In this way the effects of the characteristics of the dwelling itself on the value of a dwelling are controlled. Values of dwellings in green neighbourhoods might for example have a higher value, but this can also be due to the fact

that bigger houses are located in green neighbourhoods. Therefore information about which characteristics of the dwelling itself have an influence on the value of a dwelling is taken into account in this research.

2.1 Value of Housing

Before the literature study is discussed the definition of the term 'value of housing' needs to be defined. The first person who mentioned the term value was the philosopher Plato. Plato discussed the value of knowledge and the value of life but also the value of things (Robbesom, 2016). Since this research investigates the value of dwellings, the value of things is what needs to be determined. Plato states that the value of something is assigned by every person differently, for every possible thing that you can name (Chambers, 1936). Buczynski (2017) states that an object is worth whatever someone will pay for it. According to the Cambridge Dictionary is value 'The amount of money that can be received for something'. Van Sprundel (2014) states that all types of disciplines (economics, psychology, humanities etc.) interpret value within their own discipline. In case of valuation of real estate properties, this would mean the economic value. This is the value that is used most frequently in real estate. The definition of the economic value is the highest amount a consumer is willing to pay for a product or service in a free market economy (MBN, 2018). Within the economic value there are still several values that can be used. These are; the market value, the liquidation value and the asset or property tax value (Van Sprundel, 2014). The market value is the selling price of a product under normal selling conditions (De Hypotheker, 2017). The liquidation value is a value with the restriction that an object must be sold within a limited time frame. The asset value is a value that is determined by authorities that collect property tax. This value is an estimated value based on the value of other dwellings in proximity of a dwelling. Dwellings that are recently sold are used for this estimation (Scharm, 2006). The value of a dwelling that will be used in this

research is the market value. For this research, the value of a dwelling is the selling price of a dwelling under normal selling conditions.

2.2 Qualitative Results

Various studies have shown that the location of a real estate property influences its value. A part of these studies are studies that use surveys to see what preferences of people are. These studies are most of the time based on a qualitative method. For example, Hill and Melser (2007) declare that besides characteristics of the dwelling itself, also 'neighbourhood and accessibility attributes' influence the value of a dwelling. And also d'Amato (2010), Ranzato (2013) and DeSimone (2013) conclude that when people buy a dwelling, the location of that dwelling has an influence on the price they are willing to pay. There are also several studies that specify these spatial factors. In this way Stanghellini et. al. (2015) concluded from their survey study in the South of Italy that besides building characteristics, public services, the amount of green areas and accessibility are important factors that influence the choice people make when they buy a house. Other spatial factors that influence this choice and are mentioned in qualitative studies are, recreational green, parking space, the density of the neighbourhood, compilation of population, percentage dwellings in the rental sector and private owned property, quality of real estate in neighbourhood, criminality, and distance to hospital, high school, hotel and catering industry, coast, high way, public transport, sport facilities, shopping centre and city centre (Hill and Melser, 2007; Davidoff and Leigh, 2008; Kruk, 2012; Abelson et. al., 2013; Pandya and Patel, 2017; Rymarzak and Sieminska, 2012; and Haynes et. al., 2012). The results of these survey studies are all qualitative.

2.3 HEDONIC PRICE ANALYSIS

To retrieve a more quantitative and extensive overview about the influence of spatial factors on the value of a dwelling, hedonic price analysis can be used. The hedonic price analysis is a tool that can be used to determine the value of an object, in this case, a dwelling. The theory behind the hedonic price analysis will be explained followed by the results of hedonic price analysis in determining the value of housing.

2.3.1 THEORY

To retrieve a quantitative and extensive overview about the influence of spatial factors on the value of a dwelling, hedonic price analysis can be used. The hedonic price analysis is a valuation method based on a regression analysis. Rosen (1974) was the first that presented the hedonic price analysis. Rosen stated that an item's total price can be considered as the sum of prices of attributes. Each attribute has a unique price which is formed by an equilibrium market. Because the variance in dwellings is so high, it is difficult to determine their value. Rosen stated that the hedonic price analysis is therefore very suitable for valuation of housing. A dwelling can be decomposed into several characteristics where each characteristic has its own value. The sum of the value of each characteristic of a dwelling determines the total value of the dwelling. Examples of characteristics are number of bedrooms, size of lot, or distance to the city center.

So the hedonic price analysis is based on the assumption that the value of an item is determined by the sum of the value of several characteristics of the item. But not all characteristics have an equal contribution to the total value of an item. Some characteristics have a bigger influence on the value of an item than others. In case of a dwelling can the presence of a garden for example be more appreciated than the presence of a basement. In a hedonic price analysis the weight of each characteristic is detected by means of a regression analysis (Keskin, 2008). The weight of each

characteristic that is found in a regression analysis is called a coefficient. To determine the influence of specific characteristics on the value of a dwelling an equation can be formulated. The symbols 'X' in the equation below represent the characteristics of a dwelling. The symbols 'β' represent the coefficients of the characteristics. The equation is here presented.

$$P_H = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \ldots\ldots\ldots + \beta_n X_n + \varepsilon$$

A hedonic price analysis is used when the price of a product is known and the influence of each characteristic is unknown. The estimates of the coefficients of each characteristic are than the results of the hedonic price analysis. In case of determining the value of a dwelling, the coefficients in the equation represent the influence of a characteristic on the value of a dwelling. Examples are the influence of the age of the dwelling, the number of bedrooms, the floor area and the presence of a garage. The coefficients can represent the marginal value of the factors (Xiao, 2016). In this way the presence of a garage has mostly a higher marginal value than the unit floor area. This is because the garage is either present or not and floor area is a continuous variable. The variable floor area is measured on a bigger scale. A regression analysis can also use a log transformation for the dependent variable. The coefficients than represent the influence of a variable in terms of a percentage. The results of a hedonic price analysis by using a regression analysis gives an extensive insight of the influence of specific factors on the price of a dwelling.

2.3.2 Results

This section will discuss the results of studies that use the hedonic price analysis to determine which factors influence the value of a dwelling. The results of hedonic price analyses are not unequivocal and can be presented in different ways.. First studies will be discussed that use a log

transformation for the dependent variable. In that way, the influence a factor has on the value of a dwelling is expressed in a percentage. After that studies that use marginal values will be described. The results of characteristics of the dwelling itself will be discussed, then location and environmental factors. Next, some examples of studies that are not applicable to every region will be discussed. Finally two models that are used to value dwellings in The Netherlands will be discussed. All the studies that are described in this chapter have data sets with a maximum of 3,000 cases.

When a study uses the hedonic price analysis to determine the value of a dwelling, the results are not always presented in the same way. Sometimes the influence of a specific category is expressed in percentages, because a log transformation for the dependent variable is used. Nowak and Smith (2017) conclude that spatial factors determine 25% of the total price of a dwelling. This study is done with information retrieved from real estate agents. Wen et. al. (2005) did a study in China and stated that architecture characteristics, by which characteristics of the dwelling itself are meant, determine 60.0% of the price of a dwelling, location characteristics 19.8% and neighbourhood characteristics 16.5%. Also the percentage of specific factors that influence the value of a dwelling can be given. Jim and Chen (2006) also did a study in China and conclude that 9.2% of the selling price of an apartment is influenced by the floor level of the apartment in an apartment block. View of green spaces and water elements can increase the selling price of a dwelling with 7.1% to 13.2% and windows with a southern orientation with 1.0%. Partly retrieved from developers of the dwellings and partly collected in the field.

The studies that are described below are all studies that used the hedonic price analyses and present their results as a marginal value. These studies present more specifically what the influence of certain factors is when determining the value of a dwelling. For example, not only the total influence of characteristics of the dwelling itself is given but the influence of specific characteristics such as

floor area, age of the dwelling or the condition of the dwelling. These marginal values differ for every study, because not all studies include the same type of factors in their studies. The characteristics of the dwelling itself that are almost always mentioned as having a significant effect on the value of a dwelling are floor area and age of the building (Kuminoff et. al., 2010; Mok et. al., 2005; Bonnetain, 2003; An et. al., 2010; Wen et. al., 2005; Morancho, 2003; Boardman et. al., 2006; Keskin, 2008; Jim and Chen, 2006; and Pandya and Patel, 2017). When the floor area of a dwelling increases, the value of a dwelling also increases. This is inverse for the age of the building because this marginal value is negative. This means on average that older dwellings have a lower value. Because not all studies have the same type of information available, the results of hedonic price analyses have some differences. In that way there are some characteristics that have an influence on the value of a dwelling when they are taken into account, but information about these characteristics is not available for every study. The number of bathrooms in a dwelling is such an example. Almost all the studies that include the number of bathrooms in the hedonic price analysis, conclude that the number of bathrooms has a significant influence on the value of a dwelling (Kuminoff et. al., 2010; An et. al., 2010; Morancho, 2003; and Wen et. al., 2005). This type of information about characteristics of the dwellings itself is hard to retrieve and is often not included in studies. There are more characteristics of a dwelling itself that have been found significant several times but not always. Examples are; the condition of the dwelling, the presence of insulation, the type of dwelling, lot size, number of rooms, story level of apartment, the presence of a garden, the presence and size of a balcony, the presence of a garage, kitchen and bathroom equipment, the presence of a pool, the presence of an elevator, the presence of air conditioning, and the presence of a fireplace (Kuminoff et. al., 2010; Hill and Melser, 2007; Visser and van Dam, 2006; Stanghellini et. al., 2015; Jim and Chen, 2006; Morancho, 2003; Wen et. al., 2005; Kenkin, 2008; An et. al., 2010; and Zoppi et. al., 2014). Not only is this kind of information not always available, there are also clusters of dwellings that do not contain these factors. For example when a study focusses on stand-alone houses, the

presence of an elevator is rather unlikely. It is therefore important to take into account all types of dwellings when doing research and not focus on a specific type.

Hedonic price analysis has also been used to investigate the influence of location and environmental factors on the value of a dwelling. A location factor that is mostly found to have a significant influence on the value of a dwelling is the distance to the nearest Central Business District (CBD) (Pandya and Patel, 2017; Wen et. al., 2005; Du and Mulley, 2011; Jim and Chen, 2006; Abelson et. al., 2013; and Visser and van Dam, 2006). This is a location factor that is not always taken into account in studies but when it is taken into account it has a significant influence in the value of a dwelling in most of the studies. Some studies take the distance to the closest CBD into account, but when the study focusses on dwellings that are all located in the same area there is only little variance in that location factor. When the variance is low it is hard to identify a significant effect. This occurs not only with the factor distance to CBD but with multiple location and environmental factors. Other examples of location and environmental factors that have most of the time a significant influence on the value of a dwelling when they are taken into account are; presence of green areas, traffic conditions, well developed neighbourhoods, average income in neighbourhood, sea view, distance to school zone and distance to sport facilities (Mok et. al., 2005; An et. al., 2010; Pandya and Patel, 2017; Wen et. al., 2005; and Jim and Chen, 2006). Because of the use of relative small data sets where the variance is low, this is a difference that sometimes cannot be made.

At last there are some studies that give some significant results that are not applicable to every study, city or even country. Sea view (Zoppi et. al., 2014; and Abelson et. al., 2013) is for example a factor that is not applicable for the majority of the dwellings, but is interesting when a dwelling is located near a lake or sea. Also the presence of a pool (Hill and Melder, 2007) is not usual in every country. Bonnetain (2003) did a hedonic price analysis about dwellings that are located at islands and

found that the size of the island and the average temperature on the island have a significant effect on the value of a dwelling. And Keskin (2008) investigated the effect on the value of a dwelling of living in an earthquake risky location. At last there are some studies that have significant results about factors that have an influence on the value of a dwelling that are not quite up to date to modern standards. In that way, Hill and Melser (2007) concluded that the presence of central heating has a significant effect on the value of a dwelling and White (2015) found that result for the presence of plumbing. In this last study, dwellings in Alaska were included. A place where the presence of plumbing is not standard yet. As stated before these results are not applicable to every region.

The studies described above are executed all over the world. For this research it is interesting to look at the results of studies in The Netherlands, especially studies that investigate housing valuation. First the OrtaX model of Francke will be described. OrtaX is a model that is used to determine the value of real estate (Francke and Broekmeulen, 2016; and Francke and van de Minne 2017). It is used since 2006 and valuates every year 2.3 million real estate properties. It is developed by the company Ortec Finance and it is based on a data set with 2,658 cases of real estate sold between 2001 and 2004. The model is a Hierarchical Trend Model that determines the value of a dwelling based on characteristics of the dwelling itself, spatial factors and time. A Hierarchical Trend Model is a parametric statistical model which brings structure to the data and is an example of a State-Space model. A State-Space model can be seen as a regression model with parameters changing in time. The selling price of a dwelling at a specific time is the dependent variable in the OrtaX model. Next, segments are made to bring structure to the data. Examples of these segments are neighbourhoods or clusters of types of dwellings. Houses that are placed in the same segments have the same price development. The characteristics that have been found in the model to have a significant influence on the value of a dwelling are; volume, age of the dwelling, type of dwelling, lot size, level of

maintenance, amount of dormers, size of the garage, size of the carport, size of the living room, size of the basement and location.

The 'MarktPositie' model is another model that is used in The Netherlands. It is a web application that is used by the NVM, the Dutch Organization of real estate agents (Op 't Veld et. al., 2006). This model uses the hedonic price analysis to determine the value of a dwelling. Characteristics of a dwelling are used in the analysis. The model uses transaction data that is known by the NVM and the data set contains information about characteristics of the dwelling itself, where the dwelling is located and the sales status of the dwelling. When a real estate agent has a dwelling that needs to be valued, the characteristics of that dwelling can be uploaded in the application. Next, the application compares these characteristics with characteristics of sold dwellings that are already uploaded in the data set. Characteristics that through previous comparisons have been found to be the most important have priority. These characteristics are; lot size, type of dwelling, size of dwelling and location. The values of the 150 dwellings that have the best match with the new dwelling will be used to estimate the value of the new dwelling. The differences between the 150 dwellings are subjected to a regression analysis. In that way, the best match with the new dwelling can be found. The values of the sold dwellings will be converted to the present time for this comparison. In this way, thousands of dwellings can be compared and MarktPositie is always up to date with the actual state of the housing market. When a new dwelling is sold, the selling price will be added to the information about the dwelling and this dwelling can again be compared to new dwellings.

Both the OrtaX model and the MarktPositie model show how dwellings in The Netherlands are valued. However, both models do not take specific location or environmental factors into account for this. It is on the other hand useful to already see which factors have an influence on the value of a dwelling in The Netherlands.

2.4 Conclusions

As mentioned before, the spatial factors that are taken into account in this research are split up into location factors and environmental factors. For the completeness of the research the characteristics of the dwelling itself also have to be taken into account. There is already much known about the influence of these characters, but when these characteristics are not taken into account, the effects of the characteristics of the dwelling itself on the value of a dwelling are not controlled for. An overview can be created with the factors that have been found in the literature study to have an influence on the value of a dwelling. The factors in table 2.1 are all factors that have been described in previous sections. The factors are divided into characteristics of the dwelling itself, location factors and environmental factors.

Table 2.1 An overview of influencing factors found in the literature study.

Characteristics of the Dwelling Itself		**Location Factors**	**Environmental Factors**
Type of dwelling	Condition of dwelling	Good accessibility	Good traffic conditions
Age of the dwelling	Quality (luxurious)	Distance to CBD	Good parking facilities
Floor level of apartment	Presence of an elevator	Distance to school zone	Presence of green areas
Lot size	Presence of central heating	Distance to shopping centre	Presence of Recreational green
Volume	Presence of air conditioning	Distance to high school	Presence of shops
Floor area	Presence of fireplace	Distance to high way	Presence of schools
Presence and size of garage	Presence of pool	Distance to public transport	Quality of primary schools
Number of bathrooms	Presence of insulation	Distance to city centre	Type of neighbourhood
Number of rooms	Presence of plumbing	Distance to hospital	Degree of urbanization
Presence of garden	View of green spaces and water elements	Distance to hotel and catering industry	Density of the neighbourhood
Presence and size of the balcony	Windows with a south orientation	Distance to sport facilities	Compilation of population
	Sea view	Distance to coast	Average income in neighbourhood
		Climate	Quality of real estate in neighbourhood
		Size of island (when applicable)	Percentage rental and private owned
		Earthquake risk	Criminality

Not all factors that have been found in the literature study are relevant for every study. Some factors are very specific and are only applicable in exceptional circumstances. Depending on the location of the study should be considered which factors should be taken into account in a study.

The studies mentioned in this chapter show that there is a relation between location and environmental factors and the value of a dwelling. But the results of these studies are not satisfying enough to determine which location and environmental factors have which influence on the value of a dwelling. There are some gaps in the information of these studies. The studies that use surveys are based on qualitative methods. These studies describe the preferences people have, but the effect on the value of a dwelling is missing. The studies that use quantitative methods only investigate the influence of a small amount of factors with a maximum of 25. There is no study where a wide range of factors is analysed. Despite the fact that a list of spatial factors that have an influence on the value of a dwelling can be created, the effect that these factors have on the price is missing. Also the effects of these spatial factors in combination with characteristics of the dwelling itself are missing. This is needed to control the effect of the characteristics of the dwelling itself on the value of a dwelling. At last, the willingness to pay for all these factors is missing. Besides the lack of a study that investigated a wide range of factors, the data sets that are used up until now are not very big. The biggest data set in the studies described before completed 3,000 dwellings.

Besides that, a lot of studies are done in a small area. The dwellings are located very close to each other and the variance is low. Therefore it is hard to distinguish the effect of a specific location factor on the value of a dwelling. When all the dwellings have about the same distance to, for example, the nearest CBD, it is hard to identify what the effect of distance to a CBD is. For that, dwellings close to the CBD and not close to the CBD are needed. Another disadvantage of the previous mentioned

studies is that the majority of them took place a longer time ago. Ten years ago there were not much transaction data available as is now and the data sets that are used for the studies are with a maximum of 3,000 dwellings small in contrast with the amount of data that is nowadays available. Also the type of information that was available differs from the type of information that is available now. For example, in the example of the study of Mok et. al. (2005) is stated that taking into account the number of bedrooms in the analysis would provide a much better insight in the factors that determine the price. But this kind of information was not available.

A lot of the lacks that are just described are due to the fact that it is hard to retrieve information about dwellings. What is needed is a lot of information. Information about the characteristics of a dwelling itself, the location of a dwelling and the value of a dwelling. All this information is needed on a big scale and with a lot of variance.

2.5 Data Mining Techniques

The use of data mining techniques is very promising but it is quite unknown and not much used for valuation of real estate yet. There are however already studies that have been done that used data mining techniques in combination with real estate. Because the availability of data has strongly increased the last years, these kind of studies can now be executed on a bigger scale. Data mining techniques are namely very suitable for studies with big data sets (Philip, 2017). In this chapter a closer look will be taken to what data mining techniques exactly are and how they are so far used to predict the value of real estate.

2.5.1 Theory Data Mining Techniques

Data mining techniques are techniques that wade through (mostly) huge amounts of data to find useful information (Philip, 2017). Data mining techniques use specialized and highly advanced analysis tools to extract previously unknown patterns and relationships in data sets. These tools include most of the times statistical and mathematical algorithms (Gaur, 2012). Data mining techniques search for statistical relations in data sets with the goal to identify patterns. With these patterns a model can be created that can predict the behaviour of people or systems (Adomavicus and Tuzhilin, 2001). Much used examples of data mining techniques are association rules, neural networks, regression trees, classification trees and random forests (Han and Kamber, 2006 and Gaur, 2012). Association rule learning is an algorithm technique that finds frequent item sets in data sets. Item sets are items that often occur together. For example it has appeared that milk and bread are often bought together in a store. In other words, when a person buys milk, there is an increased chance this person will also buy bread. This can for example be interesting for marketing teams of stores. Neural networks are networks that are inspired by the way brains work. Neural networks recognize patterns in data sets and learn from them. This results in a network that recognizes, classifies and predicts new data. The network also puts a weight to each connection that represents the importance of a connection. Regression trees are graphical models that predict outcomes based on classification of cases using a tree structure. The tree structures are derived from the data by tree induction algorithms. These models are represented as flowcharts with branches, each branch represents a specific profile. Each branch ends with a leave that shows an expected utility or value. Regression trees have a continuous target variable and classification trees have a categorical target variable. A random forest is a follow up of regression trees. A random forest is a collection of many regression trees to achieve an even better prediction. Because a random forest exists of many regression trees, it analyses which factors have the biggest influence on the end of a leave where a utility or value is predicted. In this way more specific predictions can be made. Data mining

techniques can be used in the discipline of real estate. It can for example help finding patterns in valuating real estate.

2.5.2 Data Mining Techniques and Real Estate

The previous described data mining techniques can be of great interest in several business industries and real estate is one of them. For example, real estate agents and financial institutions are interested in behaviour of residents in advance of their own business. When behaviour of residents can be predicted they can respond to that. The price of recent sold dwellings can, with the use of data mining techniques help these business. Now they can for example discover which markets perform well and invest in these markets (JWB Real Estate Capital, 2018). But it can also help with the valuation of real estate. This is, as described earlier in chapter 1 interesting for municipalities, institutions, businesses and private persons. Watada (2012) explains that data mining techniques are very innovative and that they can be used in all kinds of disciplines. Real estate is mentioned as one of them. Also Yanchang and Yonghue (2013) describe that data mining techniques can be used in several industries and mention as example the prediction of real estate value.

There are some studies that used data mining techniques in combination with real estate valuation. Hromada (2015) analysed with the use of data mining techniques the evaluation of price development in the real estate market. A method is created that can analyse real estate investments in the long-term and can predict if investments are profitable. The method is based on Hadoop, an open-source software framework for distributed storage and processing big amounts of data. The study took place in Czech Republic and 650,000 cases from all over the country were used. These cases consist of dwellings, business properties and building lots. The characteristics that are taken into account are floor area, presence of a balcony or garden, presence of air conditioning and location. The values of the real estate are asking prices which are found online with real estate

agents. The study concludes that in general real estate prices will continue to decrease, but there is a difference between ‘strong locations’ and ‘weak locations’. The study takes into account the factor location, but specific spatial factors are not included in the study. Hromada (2016) continued with a study about predicting the value of real estate based on historical selling prices. An innovative method called Historical Market Price is created to valuate real estate. Normally comparable constructions, costs and profit analysis are used to predict the value of real estate. This new method is based on mathematical and statistical algorithms. Next, the software application EVAL is created. This specialized software is used to systematically collect, analyse and valuate data. Every six months, 650,000 cases are added. These cases relate to the sale and rental of dwellings, business properties and building lots. The method estimates the current value of real estate, based on characteristics of the real estate and previous selling prices. Acciani et. al. (2011) did a study to value a specific type of real estate. In their study farm houses are chosen since this type of real estate has an increasing market value. The study took place in the South of Italy and information about 169 farms is used. These are the farms that were sold from 2008 up to and including 2010. The data mining technique that is used is Model Tree. A classification model that combines decision tree learning with regression trees. For this study, thirteen variables are used. The variables that have the most influence on the value of a farm are the size of the farm, whether the farm is renovated or not, the distance to the nearest town and the presence and size of a trullo. A trullo is a typical South Italian, little building, that people can for example rent for a holiday. In the study of Bárcena et. al. (2012) the values of dwellings in Spain are investigated. For the study 12,000 dwellings in Bilbao are used. To investigate what determines the value of a dwelling several characteristics of a dwelling are taken into account as well as the location of a dwelling. To analyze which characteristics have an influence on the value, a regression analysis is executed, to select the variables that are used in a regression tree. The characteristic that has the biggest influence on the value of a dwelling is the volume of a

dwelling. The fit of the model is reasonably good with an R-squared of 0.55. There are however locations that have a better fit that other locations. For this result is no explanation found.

The studies that are presented here all are examples of studies that use data mining techniques to predict the value of real estate. But in all these studies the spatial factors are not specified or there is not much data used. The study of Acciani et. al. (2011) is the only study that investigated a specific spatial factor in their study, namely distance to nearest town. The other studies concluded that location has an influence on the value of real estate, but did not investigate which specific factors influenced this. The data set of Acciani et. al. is however very small with only 169 cases. The data mining techniques that will be used for the research of this paper are regression trees and random forest. This will be discussed further in chapter 5. The Model Tree that is used in the study of Acciani et. al. (2011) is quite similar to the regression trees technique that will be used for this research. There is however, to the best of our knowledge, no study known that uses random forest to predict real estate values.

3
DATA

To get insight in the influence of location and environmental characteristics on the value of a dwelling, a data set has to be created that combines characteristics of a dwelling itself together with location and environmental characteristics, scarcity and the value of a dwelling. To realize this, nine municipalities in the west of the province of Noord-Brabant in The Netherlands have provided a data set with information about dwellings that have been sold in the years 2011 up until 2017 in these municipalities. This data set contains information about characteristics of the dwellings itself and information about the selling price of the dwellings. Not only the selling price itself is given, but also information about the circumstances under which the dwelling is sold such as the estimated value of a dwelling and the year in which the dwelling is sold. This retrieved data set has to be extended with information about the spatial factors of the dwellings and the scarcity in the region the dwelling is located. The factors that have an influence on the value of a dwelling according to the literature study that is done, will be taken into account in this data set as much as possible. When the data set is extended with the spatial factors and scarcity, cleaning the data set can start. For example, some factors can be left out of the data set, some have to be recoded and others have to be converted into dummy variables. When the variables are transformed, the correlations between the variables will be checked for multicollinearity. This correlation cannot be too high. The definition of the factors,

retrieving the information, cleaning the data set and creating the final data sets will be described in this chapter.

3.1 VARIABLES DEFINITION

Before explaining how the data is retrieved and merged, the definitions of the different variables will be discussed. There are five different types of variables: the value of a dwelling, the characteristics of the dwelling itself, location factors, environmental factors and scarcity. In this research, the value of a dwelling is the market value of a dwelling. Since the market value is the selling price of a dwelling, the selling prices of dwellings will be used to represent the dependent variable 'value of a dwelling'. The characteristics of a dwelling itself are characteristics that belong explicitly to the dwelling itself, independently of the location of the dwelling. Location factors are factors that consider the location of a dwelling in relation to specific facilities. This factor is often expressed in distance but it can also show the amount of specific facilities in a particular radius. Environmental factors are quality characteristics that are in the near environment of a dwelling. These factors are often represented per neighbourhood.

The last variable is scarcity. Scarcity is the result of supply and demand. When the supply is low and the demand is high, the price of a product can increase. In this case the value of a dwelling could increase. The scarcity of a dwelling is dependent of the location of a dwelling. But unlike the other variables it cannot be expressed as a physical character of its location or the environment of the dwelling. Scarcity of dwellings is related to the supply and demand of the housing market in a specific region, often even of a specific type of housing. The variable scarcity will therefore be treated separately. Scarcity will in this research be represented by the so called 'shortage indicator'. This is an indicator that shows how many dwellings were available for people when they were buying a

house. It is a ratio of supply and transactions and it is used by the NVM, the Dutch Organization of real estate agents.

3.2 Retrieving Data

As stated in chapter 1, not knowing which specific location factors influence the value of a dwelling can partly be explained by the absence of private and public data. The availability of data has increased a lot the last years and information about sold dwellings is retrieved from municipalities. Information about the value of a dwelling and characteristics of the dwelling itself are retrieved from nine municipalities in the west of the province of Noord-Brabant. The municipalities are; Bergen op Zoom, Dongen, Etten-Leur, Oosterhout, Woensdrecht, Zundert, Halderberge, Roosendaal and Moerdijk. Figure 3.1 shows where the municipalities are situated in The Netherlands.

Figure 3.1 Municipalities participating in the research

The biggest city in the west of the province of Noord-Brabant is Breda. This city has 183,448 inhabitants and is the ninth biggest city in The Netherlands (CBS, 2018). The municipality of Breda has not provided information about sold dwellings so the municipality is not included in this research. To show where Breda is located relatively to the other municipalities, figure 3.2 shows in red the municipality of Breda.

Figure 3.2 The location of Breda relatively to the participating municipalities

The participating municipalities provided information about dwellings that are sold in these municipalities from 2011 up and including 2017. The information that is retrieved is information that already was available with the municipalities, and sometimes supplemented with information found on Funda. Funda is an online sales platform created by the NVM. At this website, real estate agents can offer their portfolio. For every dwelling, 40 variables are provided by the municipalities. These variables give information about where the dwelling is located, the selling price of the dwelling, the circumstances under which the dwelling is sold and the characteristics of the dwelling itself. Some examples are; the postal code, the selling price, the value estimated by assessors, the reason of difference between the selling price and the estimated value, the date of the transaction, the volume of a dwelling, the building year of a dwelling, the amount of bedrooms, the type of dwelling and whether the dwelling has specific characteristics such as a basement, dormer and garage. There are also some variables that represent qualitative aspects of the dwelling such as the quality of the

dwelling, the maintenance condition and the appearance of the dwelling. These variables are estimated by an assessor. The selling price of the dwellings will be used as dependent variable in this research because this is the market value of a dwelling, the maximum amount that consumers were willing to pay for the dwelling when the dwelling was for sale. Overviews of the variables that are retrieved from the municipalities are shown in the appendix in Tables 3.1 and 3.2. Table 3.1 shows an overview of the variables about the sale of the dwelling. Table 3.2 shows an overview of variables about characteristics of the dwelling itself. The total amount of cases retrieved from the municipalities is 26,301 cases.

The information retrieved from the municipalities has to be enriched with information about the location and environment of the dwellings. It would be ideal to retrieve this information per dwelling. This information is unfortunately not available. This information is however available at municipal, district and neighbourhood level and for all these groups an average is available. For example the average distance in a neighbourhood to a primary school. Most information about location and environmental factors is available on StatLine. StatLine is a service of the CBS, the Central Bureau for Statistics. This organization publishes reliable and coherent statistical information. Besides the Dutch official, national statistics, they also provide the European (communication) statistics (CBS, 2018). The information is available on municipal, district and neighbourhood level. To obtain an impression as specific as possible, the information for this research is obtained at neighbourhood level. Information is obtained about location and environmental factors that have an influence on the value of a dwelling, according to the literature study. Besides this information, a lot of information about other location and environmental factors is available. This information will also be included in the research. These factors might have an influence on the value of a dwelling but are not known yet. The information is retrieved from the files: Land Use, Employment, and Proximity of Services. The neighbourhoods about which this information is given, are provided with a

neighbourhood code. The CBS, has provided every municipality, district and neighbourhood in The Netherlands with a code that is used in general.

The literate study has shown that the amount of criminality in a neighbourhood has an influence on the value of a dwelling. But this information is not available on StatLine. Therefore information of the Ministry of Home Affairs and Kingdom Relations is obtained. They provide indicators of the liveability in districts and neighbourhoods in The Netherlands. This liveability shows the situation in a district or neighbourhood and criminality is part of that. To indicate liveability they use 100 indicators. These indicators are selected after statistical research has shown that these indicators show the liveability best (Ministerie van Binnenlandse Zaken en Koninkrijkrelaties, 2017). The liveability is obtained at neighbourhood level. These neighbourhoods are also provided with the neighbourhood code and can be linked to the information retrieved from StatLine.

The literature study has also shown that the population compilation and built up environment of a neighbourhood have an influence on the value of a dwelling. This information is available on StatLine in the file regional key figures of The Netherlands. The file contains information about the dwellings, inhabitants and income of the inhabitants of a neighbourhood. This information is also provided with the neighbourhood codes and can be linked to the other information about location and environment of neighbourhoods. At last information about the variable scarcity must be obtained. This information is provided by the NVM. The NVM works with a so called shortage indicator. A shortage indicator is an indicator that shows how many dwellings were for sale when people were buying a house. It is a ratio of supply and transactions and this ratio will be used for the variable scarcity in this research. Scarcity is often applied on a bigger scale than a neighbourhood. This information is therefore not available on neighbourhood level but it is at district level. Through the codes used by the CBS it is known which neighbourhoods are part of which districts. Therefore the

information of scarcity can be added to the other retrieved location and environment information. The total amount of location and environmental factors is 145 variables. An overview of the information that is retrieved from StatLine, the Ministry of Home Affairs and Kingdom Relations, and the NVM is shown in the appendix in Tables 3.3 and 3.4. Table 3.3 shows an overview of the environmental variables. Table 3.4 shows an overview of the location variables. The variable shortage is added to the environmental variables.

The information about the value and characteristics of the dwelling itself, that is provided by the nine municipalities, contains also the location of each dwelling. Because the location of each dwelling is known, the official neighbourhood codes which are provided by the CBS, can be linked to these cases. The retrieved information about the location and environment of a neighbourhood can now be combined with the information about the value and characteristics of the dwelling itself. The neighbourhood code of each dwelling is taken and the retrieved information about the location and environment of this neighbourhood is added to the case.

3.3 Cleaning Data

All the data that is needed for the analysis is retrieved. A part of the data must be transformed so the information can optimally be used. It is also necessary to select which cases can be used for the analysis and which ones not. At last it is necessary to check if there are variables that represent the same factors and if there are variables that correlate with each other.

3.3.1 Recoding Variables

There are several ways to transform the variables. Some variables will be recoded into variables with larger categories. For example, there are 65 different object codes for different types of dwellings. This will be reduced to six types of dwellings so they can be compared better to each

other. The new categories are: standalone, semidetached, row house, corner house, apartment and business property. Also the building year of a dwelling will be classified in categories. Each category covers 10, 20 or 30 years.

Next, the variables that are transformed into larger category variables will be transformed into dummy variables. But there are also other categorical variables that will be transformed into dummy variables. A dummy variable is a binary variable to represent subgroups in the data. An example is the municipality of the dwellings. For each of the nine municipalities a dummy variable will be created and every dwelling is either located in a specific municipality (value 1) or not (value 0). In a regression analysis, all the dummy variables minus one can be included. One dummy variable cannot be included and must function as a reference. The influence of the other variables is relative to this reference variable. A result may be that a house located in a specific municipally has a positive effect on the value of a dwelling. The variables that will be transformed into dummy variables are; dwelling type, municipality, building year, level of liveability and year of transaction.

There are also some variables that will be transformed into more meaningful variables. There are, for example, several variables that represent land use in a neighbourhood. These variables have absolute values. To receive a better insight of the land use in a neighbourhood, the surface of a specific land use in a neighbourhood will be divided by the total area of that neighbourhood. In that way the percentages of land use in a neighbourhood will be shown. This will be done for the following variables: park and green area surface, day recreation surface, residence recreation surface, inland water surface, sport terrain surface and forest, open and natural terrain surface. After that, the variables percentage day recreation surface and percentage residence recreation surface will be merged into the variable percentage recreation green. The variables percentage park and green area surface and percentage forest, open and natural terrain surface will be merged into the

variable percentage green. Also some information about the inhabitants of a neighbourhood will be transformed from absolute value into percentages. The variables about age, ethnicity, type of household and the amount of income receivers in a neighbourhood will be transformed in that way. The data base contains also a group of variables that represent attributes of a dwelling, such as a dormer or basement. All these attributes have multiple variables and these will be reduces to one variable for each attribute. The new variable is binary and represents if the attribute is present or not. This will be done for the attributes; bay window, dormer, balcony, basement, garage, barn, garden shed and garden apartment building. The data set misses information about the size of a garden. The total lot size and the amount of surface the dwelling covers is however present. A new variable will be created. The amount of surface covered by the dwelling is abstracted from the total lot size. The surface that remains represents the surface of the garden.

At last the transaction price of a dwelling will be transformed. Figure 3.3 shows the distribution of the transaction price. The normal distribution is also drawn in the figure. As can be seen in the figure the distribution of the transaction price differs from the normal distribution. This is mostly due to the long tail to the right. To create a clear picture of the distribution, only the cases with a transaction price between €0 and €1,250,000 are shown.

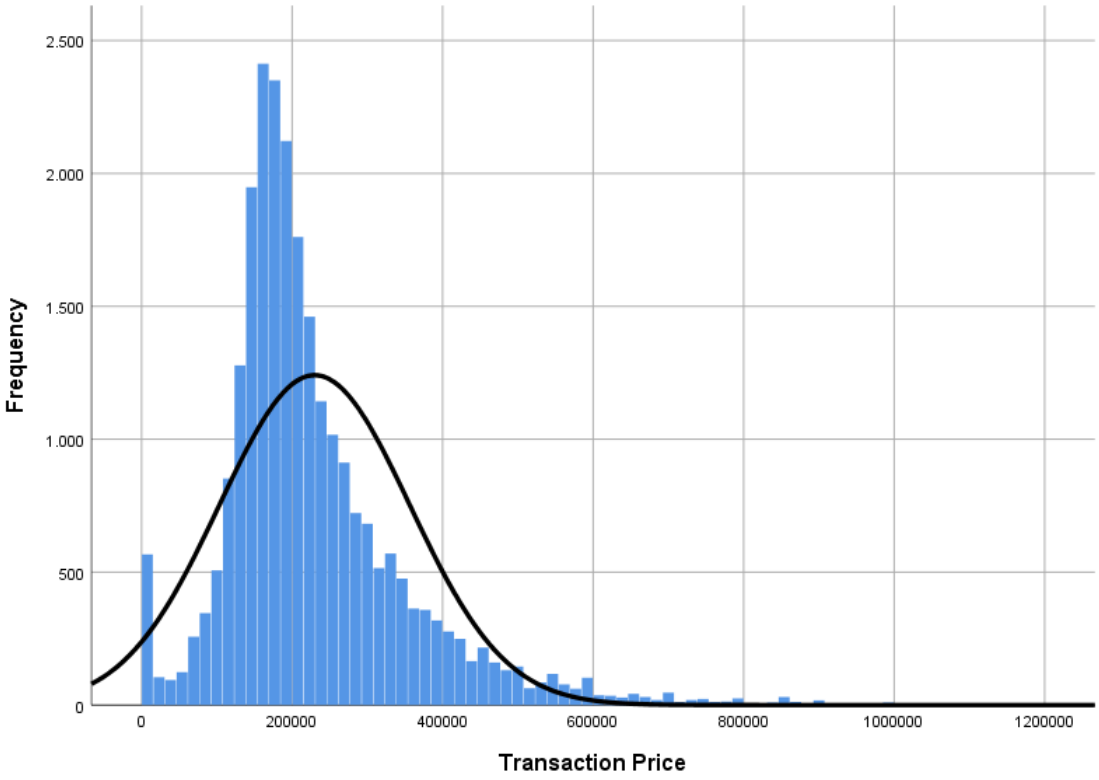

Figure 3.3 Histogram of the transaction price with normal distribution

Sometimes, researches use percentages to show the influence of a specific variable. This is possible when a logarithm transformation will be executed on the dependent variable. For this research a logarithm transformation will be executed on the variable transaction price. When a variable is transformed into a logarithm, the coefficients of the independent variables show the percentage influence of a variable on the value of a dwelling. The model is then multiplicative instead of additive. Figure 3.4 shows the distribution of the log-transaction price. The normal distribution is also drawn in the figure. As can be seen this distribution is much closer to a normal distribution than the distribution of the transaction price before the transformation. Therefore the log-transaction price will be used for the analysis.

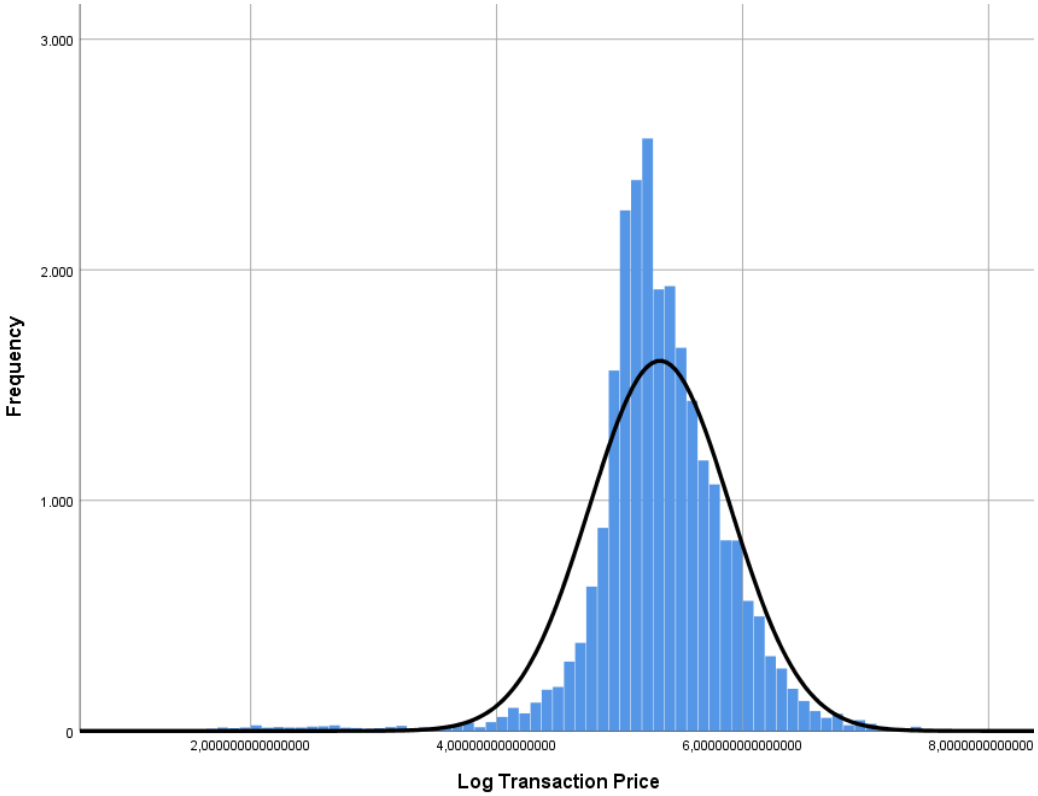

Figure 3.4 Histogram of the log-transaction price with normal distribution

3.3.2 Filtering Cases

When the variables are transformed into better fitting variables, the usable cases will be selected. All the cases that cannot be used have to be filtered out. Some variables have outliers that will be filtered out. These outliers are values that deviate too much from the mean. A standard statistical rule will be used to determine whether a value is an outlier. According to this rule, all values have to be in a specific range. This range is from the mean value minus three times the standard deviation to the mean value plus three times the standard deviation. If a value differs more than three times from the mean, the case will be considered an outlier and filtered out. Examples of variables where this rule will be applied to are the transaction price and the volume of a dwelling. This process of filtering out outliers will be done to prevent that the results of the analysis are misleading and incorrect because of extreme cases.

Besides filtering out the outliers, some cases will be removed from the data set because they are not included in the population of the research. In that way, business properties and dwellings with an office will be filtered out. These properties do often also have a dwelling included, but the research focuses on dwellings specifically. There are also properties that have a very big lot size. If the lot size is several acres big, it might function for another function such as agriculture. Therefore the biggest 5% lots will be filtered out. In practice that means the lot size has a maximum of 1,524 m^2. There is also a variable that shows the amount of dwellings that are covered by a case. For this research, only the cases with one dwelling will be selected. There are several variables that have missing values. This is not always a problem for the analysis, but the cases that have a missing value for the variable transaction price will be filtered out. The dwellings in the data set are sold between 2011 and 2017. For the research it is best to use dwellings that are sold recent and in a short time. Then, factors such as changes in the welfare of a country have a minimum effect. Therefore the dwellings that are sold before 2015 will be filtered out. At last dwellings with an abnormal transaction price will be filtered out. This will proceed in two different ways. The first is to compare the estimated value of the dwelling with the actual selling price. When the difference is more than 25% the case will be filtered out. There is also a variable that shows specific information about the sale of a dwelling. For example whether the dwelling is sold to family or is the result of a foreclosure. These are factors that can have an influence on the selling price of a dwelling and not represent the actual value of a dwelling. Therefore all the dwellings sold to family or sold in a foreclosure will be filtered out. Filtering out all these cases means that the analysis will be more valid.

After all the previous described cases are filtered out, there are 9,616 cases left in the data set. An overview of the distribution of the transaction price, dwelling types, municipalities and transaction year is shown in the figures 3.5 to 3.8. Figure 3.5 shows the distribution of the transaction prices. The mean transaction price is €220,451. The range of the transaction prices is from €21,000 to €605,000.

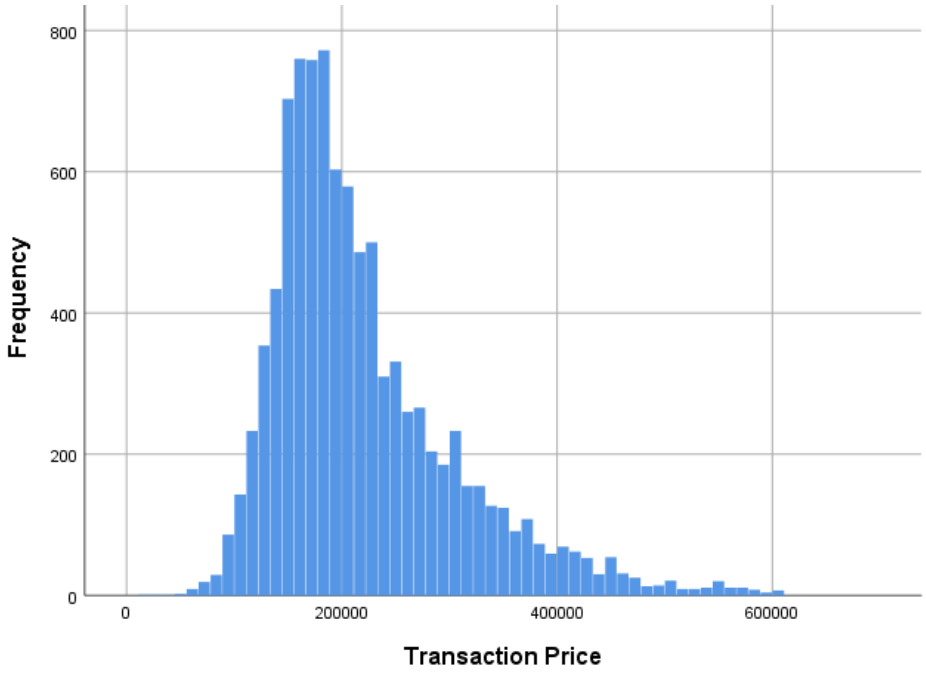

Figure 3.5 Histogram of transaction price distribution

Figure 3.6 shows the percentage distribution of dwellings across the different dwelling types. There are five dwelling types left since the business properties are removed from the data set. The most frequent dwelling type is the row house with 32.6%. After that the semi-detached houses are the most frequent with 22.2%. The least frequent dwelling types are standalone houses, corner houses and apartments with 16.0%, 15.5% and 13.6%.

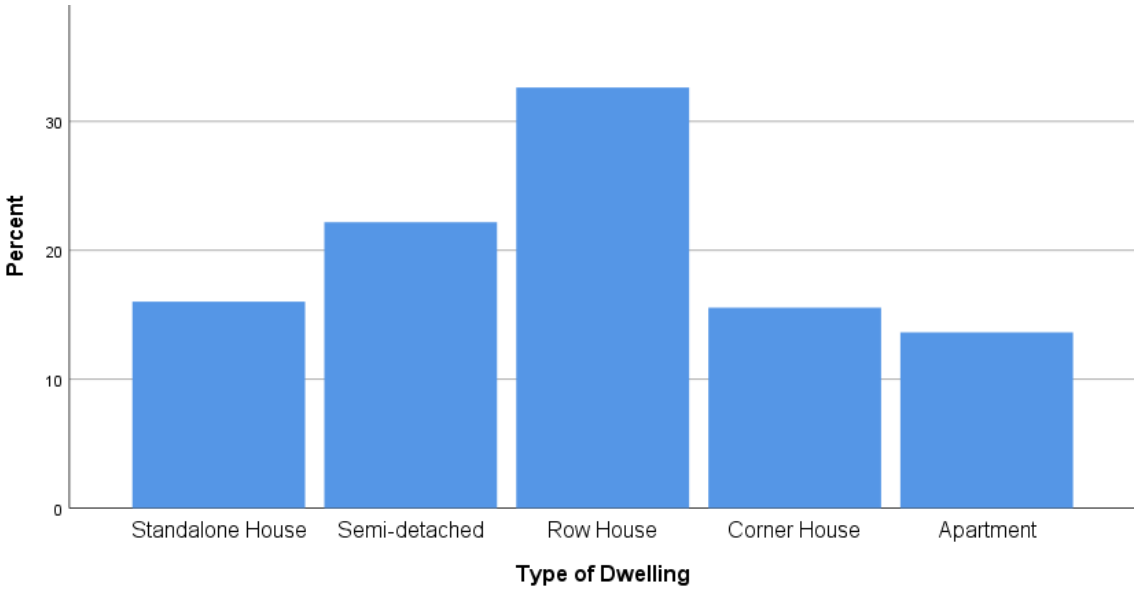

Figure 3.6 Bar chart of dwelling type distribution

Figure 3.7 shows where the dwellings that will be used for the analysis are located. The distribution is shown in percentages. Most of the dwellings are located in the municipalities of Roosendaal, Bergen op Zoom and Oosterhout with 22.2%, 17.8% and 16.1%. The least common municipality is Zundert with 4.5% of the dwellings.

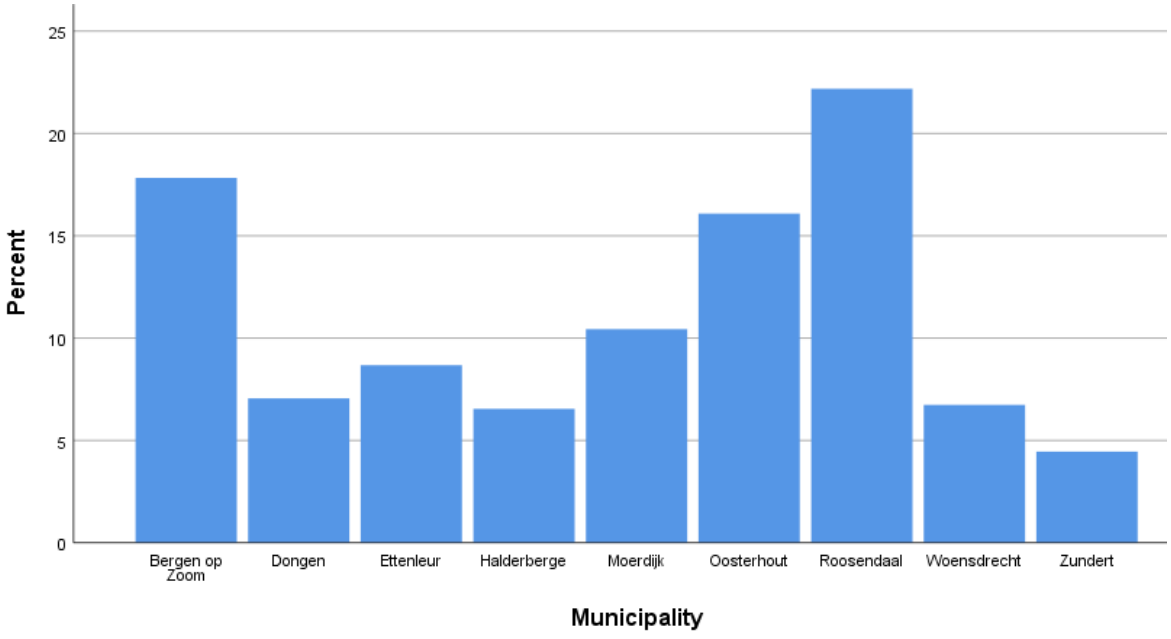

Figure 3.7 Bar chart of dwelling distribution in municipalities

Figure 3.8 shows the distribution of the transaction year of the dwellings. Most dwellings are sold in the years 2015 and 2016. This is because the transaction date of the sold dwellings is not later than July 2017. The dwellings that are sold in the second half of 2017 are not included in the data set.

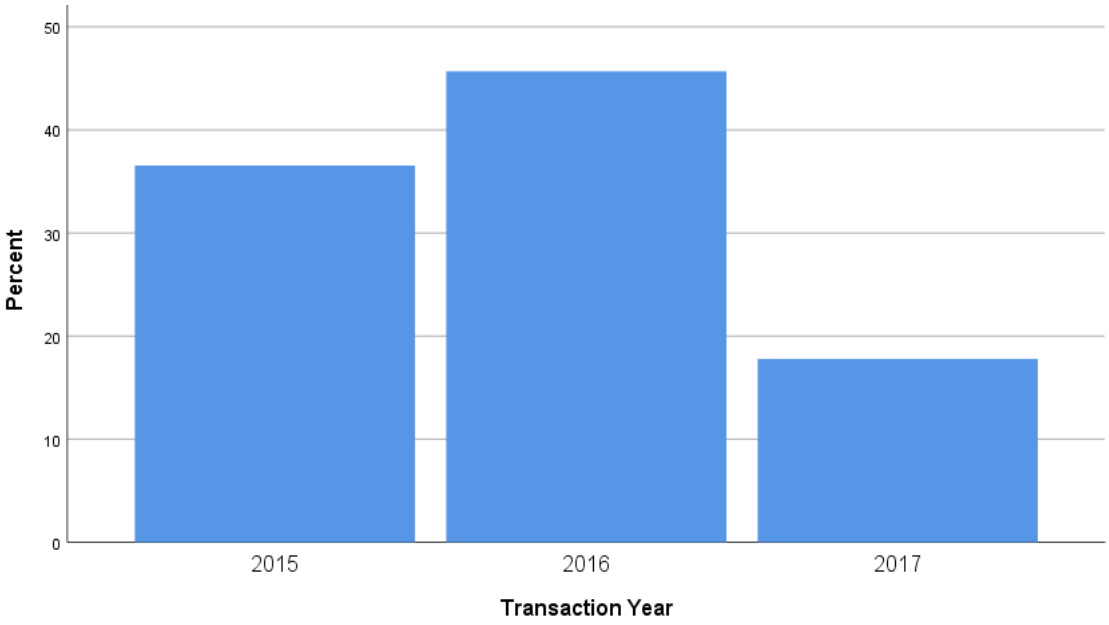

Figure 3.8 Bar chart of transaction year distribution

Table 3.1 shows the mean transaction price per year in euro´s. The biggest difference in transaction price is between the years 2015 and 2016. In 2016 the average selling price has increased with 5.8%. In 2017 it has decreased with 1.6%. These changes are probably the result of a growing economy in The Netherlands. These differences are negligible, but the variable transaction year will be included in the research as a dummy variable to control for the influence of the year the dwelling was sold.

Table 3.1 Average Selling Price Per Year

Transaction Price			
Transaction year	Mean	N	Std. Deviation
2015	212870	3513	81455
2016	225151	4392	89239
2017	223951	1711	85541
Total	220451	9616	85994

3.3.3 Combining Variables

The variables and cases that will be included in the research have now been selected. But variables in a research often have the tendency to represent the same and to correlate with each other. This can cause difficulties in estimating the effects in an analysis. A method to cope with this phenomenon is a principal component analysis. With this technique, the amount of variables in an analysis can be reduced to main components, the so called, principal components. The components that are the result of a principal component analysis do not correlate with each other. Each principal component analysis will be done with two or more variables.

At first, principal component analysis will be executed with variables that are selected together because of their content. Examples are a group of variables that represent the population of a neighbourhood or several variables that have information about daily shopping facilities in a neighbourhood. It is expected that when a principal component analysis is executed on these groups of variables, the explained variance is high.

When these first principal component analysis are done, a correlation analysis will be executed to see which remaining variables correlate strongly with each other. Variables that correlate strongly with each other tend to cause problems in a model, because their effect cannot be estimated properly. Therefore, additional principal component analysis will be executed with variables that correlate strongly with each other. The components that are created in this analysis do not correlate with each other and can be used in the model.

The first principal component analysis that will be executed is about neighbourhood types. The variables that will be used represent something of the population compilation or built up environment of a neighbourhood. Table 3.2 and 3.3 show the results of that first principal

component analysis. The amount of variance explained by the component is shown in Table 3.2.The explained variance for each of the created components is shown in Table 3.2. As can be seen in Table 3.2, in total 85.7% of the variance can be explained by six components.

Table 3.2 Variance of the Component Neighbourhood Type

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.156	37.071	37.071	8.156	37.071	37.071
2	4.143	18.830	55.901	4.143	18.830	55.901
3	2.712	12.325	68.226	2.712	12.325	68.226
4	1.711	7.778	76.004	1.711	7.778	76.004
5	1.123	5.103	81.107	1.123	5.103	81.107
6	1.018	4.626	85.733	1.018	4.626	85.733
7	.732	3.329	89.063			
8	.620	2.820	91.883			
9	.496	2.254	94.137			
10	.422	1.920	96.057			
11	.334	1.517	97.574			
12	.269	1.221	98.795			
13	.167	.757	99.552			
14	.060	.274	99.826			
15	.019	.085	99.911			
16	.010	.044	99.955			
17	.006	.030	99.984			
18	.003	.014	99.998			
19	.000	.002	100.000			
20	3.592E-5	.000	100.000			
21	8.811E-6	4.005E-5	100.000			
22	1.295E-6	5.887E-6	100.000			

Table 3.3 shows the loads of the variables on the components. Only the components with an eigenvalue bigger than one are taken into account. This shows which component is associated with which variables. The load of each variable is between -1 and 1. The loads can be seen as weights. A high positive value of a variable on a component means that a high value on this variable leads to a

high value of the component. That is the same with negative values. So high positive or high negative values have a strong connection to a component. When a case scores high on a component then this case also scores high on the variables that are included in the component. Based on the pattern of connections with variables a label can be given to each component.

Table 3.3 Loads of the Variables on Components, Neighbourhood Types

	Component Neighbourhood Type					
	1	2	3	4	5	6
Percentage single family house	.856	-.071	.167	.002	.294	-.172
Percentage apartment building	-.856	.072	-.166	-.002	-.294	.172
Percentage inhabited	.601	.353	.223	-.492	-.359	-.211
Percentage vacant	-.600	-.354	-.223	.492	.359	.211
Percentage privately owned	.779	-.460	-.259	.085	-.154	-.102
Percentage rental	-.765	.473	.270	-.096	.151	.089
Percentage housing association	-.469	.655	.379	-.279	.192	.120
Percentage remaining rental	-.568	-.388	-.224	.370	-.094	-.061
Percentage built before 2000	-.073	-.345	.883	.031	-.001	-.165
Percentage built since 2000	.073	.345	-.883	-.031	.001	.165
Average household size	.930	.173	-.077	.109	.117	.149
Percentage inhabitants 0-14	.627	.684	-.179	.097	.080	-.007
Percentage inhabitants 15-24	.444	.059	.517	.470	-.078	.376
Percentage inhabitants 25-44	-.176	.695	-.180	.279	-.251	-.385
Percentage inhabitants 45-64	.514	-.559	.335	.149	-.163	.178
Percentage inhabitants 65+	-.654	-.449	-.127	-.455	.234	.026
Percentage Western	-.237	-.083	.068	.491	.179	-.604
Percentage non Western	-.333	.644	.397	.129	.228	.104
Percentage household 1p	-.951	.013	.064	.041	-.171	-.067
Percentage household without child.	.319	-.648	-.111	-.394	.347	-.093
Percentage household with child.	.901	.284	-.022	.175	.004	.158
Percentage income receivers	-.244	-.457	.139	-.010	-.443	.146

The components in Table 3.3 all represent dimensions of a type of neighbourhood. The first dimension of neighbourhood type, which is based on the first component, characterizes neighbourhoods with a high percentage single family houses, low percentage apartment buildings, low vacant percentage, high percentage privately owned houses, low percentage rental houses, high

average household size, high percentage inhabitants in the age categories 0-14 years and 45-64 years, low percentage inhabitants in the age category 65 years and older, low percentage single person households and high percentage households with children. Summarizing this, the first dimension of a neighbourhood type can be defined as the extent to which a neighbourhood includes privately owned, single family houses and families with young children. The second dimension of a neighbourhood type characterizes neighbourhoods with a high percentage of social housing, high percentage of inhabitants in the age categories 0-14 years and 25-44 years old, low percentage of inhabitants in the age category 45-64 years old, high percentage immigrants with a non-Western ethnicity and a low percentage of households without children. The third dimension of a neighbourhood type characterizes neighbourhoods with dwellings built in 2000 or later and with a high percentage inhabitants in the age category 15-24 years old. The fourth dimension of a neighbourhood type characterizes neighbourhoods with a relatively high percentage vacant dwellings and immigrants with a Western ethnicity in the age category 15-24 years old. The fifth dimension of a neighbourhood type characterizes neighbourhoods with a relatively high percentage vacant dwellings, high percentage of households without children and low percentage of income receivers. The sixth and last dimension of a neighbourhood type characterizes neighbourhoods with a high percentage immigrants with a non-Western ethnicity, high percentage inhabitants in the age category 15-24 years old and low percentage inhabitants in the age category 25-44 years old.

The variables that will be used for the second principal component analysis are variables regarding to facilities in a neighbourhood. The variables that are included in the principal component analysis are: distance to doctor, distance to pharmacy, distance to supermarket, distance to other daily shops and distance to cafeteria. In Table 3.4 the explained variance of the components are represented. As can be seen in the table, 62.1% of the variance can be explained by only one component and this is the only component with an eigenvalue bigger than one.

Table 3.4 Variance of the Components Distance to Daily Facilities

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.106	62.110	62.110	3.106	62.110	62.110
2	.711	14.217	76.328			
3	.559	11.175	87.502			
4	.355	7.105	94.607			
5	.270	5.393	100.000			

Table 3.5 Loads of the Variables on Components, Distance to Daily Facilities

Component Distance to Daily Facilities	
	1
Distance Doctor	.863
Distance Pharmacy	.720
Distance Supermarket	.804
Distance Other Daily Shops	.725
Distance Cafeteria	.818

Table 3.5 shows the loads of the variables on the components that have been created with the principal component analysis. Only the components with an eigenvalue bigger than one are selected. All the variables that are included in the component in Table 3.5 have a high, positive score on the component. This means that if the distance to these facilities increases, the value of the component will also increase. So if a case scores high on this component, the dwelling has a big distance to the daily facilities that are included in this component.

The variables that will be used for the next principal component analysis are variables about day care and school facilities for children until they go to high school. The variables that are included in the principal component analysis on this level are: distance to day care, distance to elementary school and distance to after school care. The group of components that will be created will be called ‘school facilities children 0-11’. Table 3.6 shows the explained variance of the components. As can be seen in

the table, 77.8% of the variance can be explained by only one component and only one component has an eigenvalue bigger than one.

Table 3.6 Variance of the Components Distance to School Facilities Children 0-11

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.336	77.872	77.872	2.336	77.872	77.872
2	.442	14.731	92.603			
3	.222	7.397	100.000			

Table 3.7 Loads of the Variables on Components, School Facilities Children 0-11

Component Distance to School Facilities Children 0-11	1
Distance to Day Care	.902
Distance to Elementary School	.913
Distance to After School Care	.830

Table 3.7 shows the loads of the variables on the components of 'school facilities children 0-11', created by the principal component analysis. Only the component with an eigenvalue bigger than one is taken into account. All the variables that are included in the component have a high, positive value. So if a case has a high score on this component, it means that the dwelling has overall a big distance to the three facilities that are included in the component.

The three principal component analysis that are executed up and until now all are created based on variables that are selected together because of their content. Therefore it was already expected that the explained variance of the components is high. There are also variables that correlate highly with each other but the combination of those variables is not that obvious. Therefore the correlation between variables will be checked. With the use of the Pearson correlation the correlation between variables at interval or ratio level will be checked. When the variables that have an ordinal level are checked for correlating with other variables, the Spearman correlation will be used. When variables

have a correlation of 0.7 or higher, this is called a strong correlation. This is a standard statistical rule that will be used. Variables that correlate strongly with each other tend to cause problems in a model, because their effects cannot be estimated properly.

Checking the correlations between variables is a iterative process. First, the variables that correlate strongly with each other will be combined and a principal component analysis will be executed to see if the variables create components with a high explained variance. Next, the correlation will be checked between the new components and the remaining variables. It is possible that after creating the components, some variables correlate strongly with the created components. In that case, the variables still can be added to the components. This process will be repeatedly run through until there are no strong correlations left between variables. Several new components are created because of this correlation analysis.

Correlation analysis turns out that there is a high correlation between the following variables: Amount of department stores, hotels, cinemas, theme parks and employment opportunities in a radius of 10 km. The results of this correlation analysis is shown in Table 3.8. As can be seen in the table, the variables cinema 10 km and hotel 10 km do not have a very strong correlation with each other, but they correlate both strongly with the other variables.

Table 3.8 Correlations Variables 'Amount of Facilities within 10km'.

Correlations						
		Departm. store 10km	Hotel 10 km	Cinema 10 km	Theme park 10 km	Employ. opp. 10 km
Departm. store 10km	Pearson corr.	1	.836**	.739**	.805**	.844**
	Sig. (2-tailed)		.000	.000	.000	.000
Hotel 10 km	Pearson corr.	.836**	1	.534**	.680**	.769**
	Sig. (2-tailed)	.000		.000	.000	.000
Cinema 10 km	Pearson corr.	.739**	.534**	1	.806**	.902**
	Sig. (2-tailed)	.000	.000		.000	.000
Theme park 10 km	Pearson corr.	.805**	.680**	.806**	1	.849**
	Sig. (2-tailed)	.000	.000	.000		.000
Employ. opp. 10 km	Pearson corr.	.844**	.769**	.902**	.849**	1
	Sig. (2-tailed)	.000	.000	.000	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Since the variables in Table 3.8 correlate strongly with each other a principal component analysis will be executed. Table 3.9 shows the explained variance of the components. Table 3.10 shows the loads of the variables on the components. Since the variables that are included in the principal component analysis all show how many facilities are located in a radius of 10 km, the group of components will be named 'amount of facilities within 10 km'.

Table 3.9 Variance of the Components Amount of Facilities Within 10 km.

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.115	82.293	82.293	4.115	82.293	82.293
2	.520	10.396	92.690			
3	.192	3.843	96.532			
4	.133	2.660	99.192			
5	.040	.808	100.000			

Table 3.10 Loads of the Variables on Components, Amount of Facilities Within 10 km.

Component Amount of Facilities Within 10 km	1
Departm. store 10 km	.932
Hotel 10 km	.839
Cinema 10 km	.881
Theme park 10 km	.915
Employ. opp. 10 km	.964

Table 3.10 shows that all the variables that are included in the component have a high, positive value. So if a case has a high score on this component, it means that there is overall a high amount of department stores, hotels, cinemas, theme parks and employment opportunities located within a radius of 10 km of the dwelling.

The correlations analysis also shows that there is a high correlation between the next variables: The amount of department stores, hotels, cinemas, theme parks and employment opportunities within a 20 km radius, the amount of theme parks and employment opportunities within a 50 km radius and the distance to the nearest artificial ice skating rink. This correlation analysis is shown in Table 3.11. As can be seen in the table, all the correlations are very strong. Only the variable 'amount of theme parks within a radius of 50 km' has correlations that are just under 0.7. The correlations with all the other variables are high, so if the variable 'amount of theme parks within a radius of 50 km' would be left out of the principal component analysis, it would correlate strongly with the created components and this would give problems when estimating the effect. The table shows also that all the correlations are positive, except the correlations of the variable 'distance to the nearest artificial ice skating rink'. This means that if the distance to the nearest ice skating rink decreases, the amount of all the other facilities increases.

Table 3.11 Correlations Variables 'Amount of Facilities Within 20-50km'..

Correlations									
		Depart store 20 km	Hotel 20 km	Dist. ice skating	Cinema 20 km	Theme park 20 km	Theme park 50 km	Employ opp. 20 km	Employ opp. 50 km
Depart. store 20km	Pearson corr.	1	.969**	-.837**	.961**	.943**	.664**	.982**	.745**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
Hotel 20 km	Pearson corr.	.969**	1	-.883**	.951**	.937**	.731**	.980**	.802**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000
Dist. ice skating	Pearson corr.	-.837**	-.883**	1	-.884**	-.846**	-.823**	-.873**	-.879**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
Cinema 20 km	Pearson corr.	.961**	.951**	-.884**	1	.940**	.681**	.979**	.761**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
Theme park 20 km	Pearson corr.	.943**	.937**	-.846**	.940**	1	.650**	.939**	.706**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000
Theme park 50 km	Pearson corr.	.664**	.731**	-.823**	.681**	.650**	1	.702**	.973**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
Employ. opp. 20 km	Pearson corr.	.982**	.980**	-.873**	.979**	.939**	.702**	1	.777**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000
Employ. opp. 50 km	Pearson corr.	.745**	.802**	-.879**	.761**	.706**	.973**	.777**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Since the variables in Table 3.11 correlate strongly with each other a principal component analysis will be executed. Table 3.12 shows the explained variance of the components. Table 3.13 shows the loads of the variables on the components Since the variables that are included in the principal component analysis almost all show how many facilities are located in a radius of 20 or 50 km, the group of components will be named 'amount of facilities within 20-50 km'.

Table 3.12 Variance of the Components Amount of Facilities Within 20-50 km.

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.967	87.083	87.083	6.967	87.083	87.083
2	.741	9.263	96.346			
3	.125	1.563	97.909			
4	.078	.981	98.890			
5	.044	.544	99.434			
6	.022	.279	99.713			
7	.015	.188	99.901			
8	.008	.099	100.000			

Table 3.13 Loads of the Variables on Components, Amount of Facilities Within 20-50 km.

Component Amount of Facilities Within 20-50 km	
	1
Department store 20 km	.956
Hotel 20 km	.975
Distance to artificial ice skating rink	-.941
Cinema 20 km	.963
Theme park 20 km	.937
Theme park 50 km	.826
Employ. opp. 20 km	.973
Employ. opp. 50 km	.884

Table 3.13 shows that all the variables, except distance to artificial ice skating rink, have a high, positive value. The variable distance to artificial ice skating rink has a high, negative value. This means that if a case scores high on the created component, there is overall a high amount of department stores, hotels, cinemas, theme parks and employment opportunities located within a radius of 20 or 50 km of the dwelling. It also means that the location to the nearest artificial ice skating rink is low.

The next principal component analysis that is executed contains the following variables: Address density, amount of high schools within a 3 and 5 km radius, amount of pre-vocational education

within a 3 and 5 km radius, amount of pre-university education within a 3 and 5 km radius, amount of doctors within a 3 and 5 km radius, amount of pubs within a 3 and 5 km radius, amount of cafeterias within a 3 and 5 km radius, amount of restaurants within a 3 and 5 km radius, amount of day cares within a 3 and 5 km radius, amount of after school cares within a 3 and 5 km radius, amount of elementary schools within a 3 and 5 km radius, amount of supermarkets within a 3 and 5 km radius, amount of other daily shops within a 3 and 5 km radius, the distance to the nearest hospital inclusive external departments and the distance to the nearest hospital exclusive external departments. The list of variables that correlate with each other is relatively big. Therefore the correlation matrix of these variables can be seen in Appendix Table 3.6. Appendix Table 3.5 functions as an index table for Table 3.6. The values that are shown in Table 3.6 are the Pearson correlations.

Appendix Table 3.6 shows that almost all the correlations are strong. Only the variables distance to nearest hospital inclusive and exclusive external departments have lower correlations with some variables. But overall the correlations with the other variables are high so the variables will be included in the principal component analysis. Appendix Table 3.6 shows also that all the correlations are positive, except the correlations of the variables distance to nearest hospital inclusive and exclusive external departments. This means that if the distance to the nearest hospital decreases, the amount of all the other facilities increases.

Since the variables in Appendix Table 3.6 correlate strongly with each other a principal component analysis will be executed. Table 3.14 shows the explained variance of the components. Table 3.15 shows the loads of the variables on the components. Since the variables that are included in the principal component analysis mostly show how many daily facilities are located in a radius of 3 or 5 km, the group of components will be named ‘amount of daily facilities within 3-5 km’.

Table 3.14 Variance of the Components Amount of Daily Facilities Within 3-5 km.

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	21.708	80.401	80.401	21.708	80.401	80.401
2	1.872	6.934	87.335	1.872	6.934	87.335
3	1.195	4.425	91.760	1.195	4.425	91.760
4	.540	2.001	93.762			
5	.364	1.346	95.108			
6	.286	1.060	96.168			
7	.266	.985	97.154			
8	.214	.791	97.944			
9	.114	.421	98.365			
10	.089	.329	98.694			
11	.076	.280	98.974			
12	.054	.202	99.176			
13	.046	.169	99.344			
14	.036	.134	99.478			
15	.032	.119	99.597			
16	.030	.111	99.708			
17	.019	.069	99.777			
18	.012	.045	99.822			
19	.010	.038	99.860			
20	.009	.033	99.893			
21	.007	.027	99.920			
22	.006	.021	99.941			
23	.005	.020	99.962			
24	.003	.013	99.974			
25	.003	.012	99.986			
26	.002	.008	99.994			
27	.002	.006	100.000			

Table 3.15 Loads of the Variables on Components, Amount of Daily Facilities Within 3-5 km.

Component Amount of Daily Facilities Within 3-5 km			
	1	2	3
Address density	.841	.334	-.191
The amount of high schools within a 3 km radius	.909	.322	.153
The amount of high schools within a 5 km radius	.935	-.240	.225
The amount of pre-vocational education within a 3 km radius	.880	.333	.184
The amount of pre-vocational education within a 5 km radius	.898	-.252	.297
The amount of pre-university education within a 3 km radius	.820	.410	.191
The amount of pre-university education within a 5 km radius	.900	-.173	.251
The amount of doctors within a 3 km radius	.944	.207	-.042
The amount of doctors within a 5 km radius	.925	-.254	.008
The amount of pubs within a 3 km radius	.885	.265	.164
The amount of pubs within a 5 km radius	.887	-.246	.325
The amount of cafeterias within a 3 km radius	.925	.262	-.076
The amount of cafeterias within a 5 km radius	.933	-.256	.040
The amount of restaurants within a 3 km radius	.909	.271	-.143
The amount of restaurants within a 5 km radius	.937	-.261	-.003
The amount of day cares within a 3 km radius	.909	.146	.191
The amount of day cares within a 3 km radius	.896	-.229	.263
The amount of after school cares within a 3 km radius	.894	.136	-.191
The amount of after school cares within a 5 km radius	.910	-.256	-.222
The amount of elementary schools within a 3 km radius	.898	.227	-.201
The amount of elementary schools within a 5 km radius	.919	-.238	-.169
The amount of supermarkets within a 3 km radius	.899	.297	-.214
The amount of supermarkets within a 5 km radius	.928	-.217	-.144
The amount of other daily shops within a 3 km radius	.947	.261	-.059
The amount of other daily shops within a 5 km radius	.952	-.276	.065
Distance to nearest hospital inclusive external departments	-.699	.249	.502
Distance to nearest hospital exclusive external departments	-.787	.325	.339

Table 3.15 shows that there are three components created in the principal component analysis with an eigenvalue bigger than one. In general, these three components show the dimensions of how daily facilities are available. The first dimension of daily facilities within 3-5 km characterizes that there are overall a lot of daily facilities in the area of the dwelling. More specific, in a 3 or 5 km radius of the dwelling. The distance to the nearest hospital inclusive or exclusive external departments is overall low. The second dimension of daily facilities within 3-5 km characterizes that the amount of

high schools, pre-vocational education and pre-university education within a 3 km radius is high, there is a high address density and the distance to the nearest hospital exclusive external departments is high. Scoring high on the third component means that the amount of pubs within a 5 km radius is high and the distance to hospitals inclusive or exclusive external departments is overall high.

Correlation analysis show that there are also strong correlations between the variables: Distance to nearest department store, hotel, solarium, high school, pre-vocational education and pre-university education, and the amount of department stores within a 5 km radius. Table 3.16 shows the correlation analysis of the variables. As can be seen in the table, most correlations are very strong. There are some variables that have one or two correlations that are less strong but they correlate overall strongly so these variables will be included in the principal component analysis. Table 3.16 shows also that all the correlations are positive, except the correlations of the variable 'amount of department stores within a 5 km radius'. This means that if the amount of department stores within a 5 km radius decreases, the distance to all the other facilities increases.

Table 3.16 Correlations Variables 'Distance to Leisure and High School'.

Correlations								
		Dist. depart. store	Depart. store 5 km	Dist. hotel	Dist. solarium	Dist. high school	Dist. pre-voc	Dist. pre-uni
Dist. depart. store	Pearson corr.	1	-.764**	.576**	.806**	.759**	.757**	.670**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
Depart. store 5 km	Pearson corr.	-.764**	1	-.508**	-.669**	-.636**	-.636**	-.516**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000
Dist. hotel	Pearson corr.	.576**	-.508**	1	.428**	.612**	.607**	.523**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000
Dist. solarium	Pearson corr.	.806**	-.669**	.428**	1	.610**	.596**	.572**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
Dist. high school	Pearson corr.	.759**	-.636**	.612**	.610**	1	.994**	.868**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000
Dist. pre-voc	Pearson corr.	.757**	-.636**	.607**	.596**	.994**	1	.857**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000
Dist. pre-uni	Pearson corr.	.670**	-.516**	.523**	.572**	.868**	.857**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Since the variables in Table 3.16 correlate strongly with each other a principal component analysis will be executed. Table 3.17 shows the explained variance of the components. Table 3.18 shows the loads of the variables on the components . Since the variables that are included in the principal component analysis almost all show the distance to leisure facilities and high schools, the group of components will be named 'distance to leisure and high schools'.

Table 3.17 Variance of the Components Distance to Leisure and High Schools.

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.030	71.852	71.852	5.030	71.852	71.852
2	.754	10.765	82.616			
3	.569	8.123	90.740			
4	.334	4.766	95.506			
5	.171	2.440	97.946			
6	.139	1.979	99.925			
7	.005	.075	100.000			

Table 3.18 Loads of the Variables on Components, Distance to Leisure and High Schools.

Component Distance to Leisure and High Schools	1
Distance department store	.902
Department store 5 km	-.794
Distance hotel	.787
Distance solarium	.705
Distance high school	.935
Distance pre-vocational education	.930
Distance pre-university education	.854

Table 3.18 shows that all the variables, except amount of department stores within a 5 km radius, have a high, positive value. The variable about the amount of department stores is high and negative. This means that when a case scores high on the component, the distance to the leisure facilities and high schools is relatively high and the amount of department stores within a 5 km radius is relatively low.

The last set of variables that have a strong correlation are the following variables: Distance to nearest transfer railway station, distance to nearest cinema and the amount of cinemas within a 5 km radius. Table 3.19 shows the correlations of the variables. As can be seen in the table, the correlations are strong. The variables distance to nearest cinema and transfer railway station correlate positively with each other. They both correlate negatively with the variable amount of cinemas within a 5 km radius. This means that if the distance to a cinema or transfer railway station increases, the amount of cinemas in a 5 km radius decreases.

Table 3.19 Correlations Variables 'Cinemas and Transfer Railway Stations'.

Correlations				
		Distance to cinema	Cinemas 5km	Distance to transfer railway station
Distance to cinema	Pearson correlation	1	-.846**	.793**
	Sig. (2-tailed)		.000	.000
Cinemas 5km	Pearson correlation	-.846**	1	-.693**
	Sig. (2-tailed)	.000		.000
Distance to transfer railway station	Pearson correlation	.793**	-.693**	1
	Sig. (2-tailed)	.000	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Since the variables from Table 3.19 correlate strongly with each other a principal component analysis will be executed. Table 3.20 shows the explained variance of the components. Table 3.21 shows the loads of the variables on the components. Since the variables that are included in the component only say something about cinemas and transfer railway stations, the group of components will be called 'cinemas and transfer railway stations'.

Table 3.20 Variance of the Components Cinemas and Transfer Railway Station.

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.557	85.240	85.240	2.557	85.240	85.240
2	.312	10.383	95.623			
3	.131	4.377	100.000			

Table 3.21 Loads of the Variables on Component, Cinemas and Transfer Railway Station.

Component Cinemas and Transfer Railway Station	
	1
Distance cinema	.955
Cinemas 5 km	-.918
Distance to transfer railway station	.895

Table 3.18 shows that the variables distance to cinema and transfer railway station have a high, positive value and that the variable amount of cinemas within a 5 km radius has a high, negative value. This means that when a case has a high score in this component, the distance to a cinema and transfer railway station is relatively high and the amount of cinemas within a 5 km radius is relatively low.

3.4 Conclusions

The data set that will be used for the analysis is ready now. First all the retrieved data about the transaction price of the dwelling, the characteristics of the dwelling itself, the scarcity, location factors and environmental factors was combined into one single data set. After that, the data set was cleaned. Variables that were redundant were filtered out and some other variables were recoded. Also the amount of cases was reduced. Some cases were filtered out because they had values that were outliers. These outliers are filtered out to prevent that the results of the analysis are misleading and incorrect. Other cases were filtered out because of other reasons such as, sold before 2016, sold to family members or the dwelling was sold as foreclosure. By filtering out the cases, the amount of cases is reduced from 26,301 to 9,616 cases. At last the correlations between the variables were checked for multicollinearity. Variables that correlate strongly with each other tend to cause problems in a model, because their effect cannot be estimated properly. With the use of a principal component analysis, several components were created. The variables that are included in the components will be removed from the data set and the created components will be added. Now, there is no substantial correlation between the variables (and components) present anymore. The final data set contains 71 variables and 9,616 cases.

4

HEDONIC PRICE ANALYSIS METHOD AND RESULTS

A data set with all the requested information about dwellings and their spatial information is now created and the data can be analysed. The data will be analysed using the hedonic price analysis and data mining techniques. Analysing the data set with data mining techniques will be described in chapter 5. This chapter focusses on analysing the data set using the hedonic price analysis. First, the method will be described. Then, the different steps that will be taken in this process will be described. At last the results of the analysis will be shown and discussed. These results will be split up into characteristics of the dwelling itself, environmental factors, location factors, circumstances of the transaction and shortage.

4.1 METHODOLOGY

Because the data set is now complete, analysing the data set to determine what the influence of different spatial factors on the value of a dwelling is can start. The analysis that will be used and discussed in this chapter is the hedonic price analysis. As described in chapter 2, the hedonic price analysis is a valuation method based on a regression analysis. The hedonic price analysis is based on the assumption that the value of an item is determined by the sum of the value

of several characteristics of the item. Every characteristic has a unique contribution to the total value of an item. A dwelling can be decomposed into several characteristics where each characteristic has its own value. The sum of the value of each characteristic of a dwelling determines the total value of the dwelling. Since the transaction price of the dwelling is known, the influence of each characteristic can be determined by using the hedonic price analysis. In the hedonic price analysis the weight of each characteristic is detected by means of a regression analysis. To determine the influence of specific characteristics on the value of a dwelling, the next equation will be used.

$$P_H = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \ldots\ldots\ldots + \beta_n X_n + \varepsilon$$

P_H represents the value of the dwelling. The symbols 'X' represent the characteristics of a dwelling. The symbols 'β' represent the coefficients of the characteristics. The coefficients in the equation represent the influence of a characteristic on the value of a dwelling. The estimates of the coefficients of each characteristic are the results of the hedonic price analysis because they represent the influence each characteristic has on the value of a dwelling. This influence is the willingness to pay of a characteristic. In other words, what are people who buy a dwelling willing to pay for a characteristic. The regression analysis will be executed in the statistical program SPSS.

4.2 Analysis Process

With the use of the regression analysis, the influence of specific characteristic of a dwelling on the value of a dwelling can be determined. So not only which characteristics have an influence on the value of a dwelling but also which specific influence they have, in other words, the willingness to pay. Therefore the value of a dwelling will be the dependent variable in the regression analysis and characteristics of the dwelling will be the independent variables. As stated in chapter 3, it is possible to show the percentage influence of specific variables on the value of a dwelling. This is possible when a logarithm transformation will be executed on the dependent variable. For this research a

logarithm transformation will be executed on the variable transaction price. The coefficients of the independent variables will then show the percentage influence of a variable on the value of a dwelling. The model will be multiplicative instead of additive. Also, the distribution of the log-transaction price is much closer to a normal distribution than the distribution of the transaction price before the transformation. Therefore the log-transaction price will be used for the analysis. All the variables that are remaining after adjusting the data set in chapter 3 will be used as independent variables in the regression analysis. These variables contain information about the dwelling itself, environmental factors, location factors, the circumstances of the transaction and scarcity.

In total, 70 independent variables are remaining for the regression analysis. An overview of all the variables is shown in Appendix Table 4.1. As can be seen in the table, 32 variables contain information about characteristics of the dwelling itself, 11 about environmental factors, 23 about location factors, 3 variables about the circumstances of the transaction and 1 about scarcity. A regression analysis will now be executed with all these variables as independent variables and the log-transaction price as the dependent variable. For each of the dummy variables, one variable will be left out. The results of the other variables are then in relation to the variable that is left out. There are three dummy variables, so the total amount of independent variables is 67. The results of the regression analysis are shown in Table 4.1, 4.2 and 4.3.

Table 4.1 Model 1 Summery Regression Analysis

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.921	0.848	0.844	0.122

Table 4.1 shows how much of the total variance of the dependent variable, the log-transaction price, can be explained by the independent variables. The R square shows this value. The adjusted R square is a modified version of R squared that has been adjusted for the number of predictors in the model.

In this research the adjusted R square will be used. As can be seen in the table, 84.4% of the total variance has been explained.

Table 4.2 Anova Regression Analysis 1

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	251.101	67	3.748	251.711	.000
	Residual	45.114	3030	.015		
	Total	296.215	3097			

Table 4.2 shows the statistical significance of the regression analysis that is executed. It indicates how significant the regression model predicts the dependent variable. When the level of significance is 0.05 or smaller, the prediction is significant. This is a standard statistical rule that will be used in this research. As can be seen in the table, the level of significance is smaller than 0.0005. This means the regression model is significant.

Table 4.3 Coefficients of Regression Analysis 1

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.	Beta		
1	(Constant)	3.879	0.070		55.163	0.000
	Quality luxury	0.091	0.006	0.135	16.183	0.000
	Maintenance condition	0.073	0.004	0.158	18.379	0.000
	Appearance	0.066	0.008	0.064	8.036	0.000
	Volume	0.001	0.000	0.466	41.253	0.000
	Residential layers	0.004	0.002	0.019	2.327	0.020
	Amount rooms	0.002	0.004	0.007	0.491	0.624
	Amount bathrooms	-0.003	0.006	-0.004	-0.490	0.624
	Amount bedrooms	0.017	0.004	0.053	3.747	0.000
	Standalone house	0.105	0.013	0.086	7.974	0.000
	Semi-detached house	0.075	0.009	0.078	8.182	0.000
	Row house	-0.027	0.007	-0.044	-4.219	0.000
	Apartment	0.016	0.011	0.021	1.475	0.140
	Size garden	0.000	0.000	0.210	16.492	0.000
	Bay window	0.100	0.007	0.121	15.300	0.000
	Dormer	0.022	0.006	0.031	3.805	0.000
	Balcony	0.036	0.009	0.036	4.049	0.000
	Basement	0.044	0.023	0.014	1.894	0.058

Garage	0.097	0.007	0.119	14.514	0.000
Barn	-0.045	0.009	-0.061	-4.876	0.000
Garden shed	0.024	0.010	0.019	2.496	0.013
Garden apartment building	0.007	0.033	0.002	0.218	0.828
Built between 1900-1929	-0.011	0.025	-0.009	-0.457	0.648
Built between 1930-1949	0.028	0.025	0.020	1.100	0.271
Built between 1950-1959	-0.035	0.025	-0.028	-1.368	0.171
Built between 1960-1969	-0.052	0.025	-0.061	-2.100	0.036
Built between 1970-1979	-0.013	0.024	-0.018	-0.540	0.589
Built between 1980-1989	0.028	0.025	0.032	1.144	0.253
Built between 1990-1999	0.164	0.025	0.192	6.596	0.000
Built between 2000-2009	0.240	0.025	0.239	9.444	0.000
Built after 2009	0.177	0.028	0.091	6.339	0.000
Percentage sport terrain	0.001	0.000	0.013	1.484	0.138
Percentage green	0.001	0.000	0.056	4.664	0.000
Percentage recreation terrain	0.008	0.003	0.022	2.629	0.009
Neighbourhood type 1	0.011	0.004	0.037	2.564	0.010
Neighbourhood type 2	-0.012	0.003	-0.037	-3.947	0.000
Neighbourhood type 3	-0.009	0.003	-0.033	-3.490	0.000
Neighbourhood type 4	-0.015	0.003	-0.042	-4.609	0.000
Neighbourhood type 5	-0.002	0.003	-0.006	-0.599	0.549
Neighbourhood type 6	0.013	0.004	0.037	3.661	0.000
Distance to daily facilities	-0.016	0.008	-0.036	-2.041	0.041
Distance to school facilities children 0-11	-0.004	0.005	-0.008	-0.782	0.434
Hospital inclusive 5 km	0.014	0.018	0.030	0.819	0.413
Hospital inclusive 10 km	-0.006	0.010	-0.017	-0.581	0.561
Hospital inclusive 20 km	-0.010	0.008	-0.049	-1.226	0.220
Hospital exclusive 5 km	0.061	0.026	0.096	2.383	0.017
Hospital exclusive 10 km	-0.021	0.016	-0.040	-1.293	0.196
Hospital exclusive 20 km	-0.019	0.010	-0.097	-2.008	0.045
Distance pub	0.027	0.005	0.080	5.387	0.000
Distance restaurant	-0.021	0.008	-0.038	-2.686	0.007
Hotel 5 km	0.015	0.005	0.118	2.933	0.003
Distance library	-0.018	0.005	-0.054	-3.643	0.000
Distance swimming pool	0.014	0.004	0.061	3.661	0.000
Distance sauna	0.008	0.002	0.170	3.058	0.002
Distance theme park	-0.018	0.002	-0.234	-7.922	0.000
Distance firehouse	0.005	0.006	0.014	0.774	0.439
Distance highway	0.010	0.005	0.025	1.815	0.070
Distance railway station	0.004	0.002	0.044	2.209	0.027
Amount of facilities within 10 km	-0.021	0.011	-0.071	-1.891	0.059
Amount of facilities within 20-50 km	0.050	0.014	0.166	3.620	0.000
Amount of daily facilities within 3-5 km 1	-0.056	0.031	-0.170	-1.791	0.073
Amount of daily facilities within 3-5 km 2	0.020	0.009	0.062	2.078	0.038
Amount of daily facilities within 3-5 km 3	-0.045	0.010	-0.155	-4.567	0.000
Distance to leisure and high schools	-0.032	0.008	-0.084	-4.095	0.000
Cinemas and transfer railway stations	0.002	0.019	0.007	0.122	0.903
Transaction year 2016	0.035	0.005	0.057	6.471	0.000
Transaction year 2017	0.064	0.007	0.085	9.380	0.000
Shortage indicator	-0.004	0.001	-0.051	-3.814	0.000

Table 4.3 shows the coefficients of all the independent variables that are included in the regression analysis. The coefficients are the estimated effects of the variables. Table 4.3 also shows the standardized coefficients and the significance of all the variables. With the standardized coefficients, the strengths of the effects of the independent variables can be compared to each other. The significance of all the variables shows if the influence of the variable, the coefficient, is significant or not. The standard statistical rule of a significance level of 0.05 will be used.

Table 4.3 shows the effects that the different variables have on the log-transaction price of a dwelling. There are several variables that have an effect that is not significant. All the variables with a signification level bigger than 0.2 are removed from the model and the regression analysis will be executed again. The reason to remove the variables with a significance level bigger than 0.2 is consciously. This is chosen because it is possible that if one variable is removed from the model, the effect and significance of another variable will change. This is possible due to correlations between variables. The strong correlations between variables are resolved but mild correlations also can have an effect on the model. Therefore the variables with a significance level bigger than 0.2 will be removed first. In the next phases of the analysis will be worked towards a significance level of 0.05.

The variables that will be removed for the next regression analysis are variables that were also not found in the literature study. It is therefore not strange that they do not have a significant effect on the value of a dwelling. Only the variable ‘distance to school facilities children 0-11’ was expected to have an effect. Therefore a closer look will be taken to the variable. The variable is a component created with the variables, distance to elementary school, distance to day care and distance to after school care. Especially the variable distance to elementary school was expected to have an effect. When looking at the distribution of the variable distance to elementary schools, can be seen that 95% of the dwellings has an elementary school within a radius of 1 km. To investigate the variable

further, the cases with an elementary school within a radius of 1 km are temporarily removed from the data set. The results show then that the distance to elementary schools has a negative effect with a 0.073 significance level. In other words, the distance to the nearest elementary school does not have an effect on the value of a dwelling until the nearest elementary school is more than 1 km away. From then on, if the distance to an elementary school increases, the value of a dwelling decreases. The distance to the nearest day care and after school care do not have an effect on the value of a dwelling. Also when the nearest day care or after school care is located more than 1 km away, it does not have an effect. It is possible that this effect only occurs with distance to elementary school because children are considered to go to elementary school by themselves when they are old enough. And children have to be picked up at day care and after school care anyway. Because this effect only occurs within 5% of the cases, this will no longer be taken into account. Maybe the effect is stronger in districts or countries where elementary schools are not always so close to dwellings.

The 14 variables with a significance level of 0.2 and higher are removed from the model and a new regression analysis is executed. This second model has two variables that have a significance level of 0.2 or higher, while in the previous model their significance level was 0.2 or lower. The first variable is apartment. In the first model, the variable apartment had a significance level of 0.140. This changed to 0.227. This is probably due to the fact that there is a correlation between the variables apartment and amount of rooms. This correlation is 0.520. Also the significance level of the variable 'hospital exclusive 10 km' changed from 0.196 to 0.201. This is probably due to removing the variables 'hospital exclusive 5km, 10 km and 20 km'. The correlation with the three removed variables is small and respectively 0.216, 0.367 and 0.263. The variables apartment and hospital exclusive 10 km will be removed when the next regression analysis is executed. The third variable that will be removed is percentage sport terrain. This variable has a significance level of 0.182. All the

other variables have a significance level of 0.05 or lower. When the three variables that are described are removed, a new regression analysis will be executed. The results are described in Table 4.4, 4.5 and 4.6

Table 4.4 Model 3 Summery Regression Analysis

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.917	.841	.839	.124

Table 4.5 Anova Regression Analysis 3

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	303.569	50	6.071	394.429	.000
	Residual	57.492	3735	.015		
	Total	361.061	3785			

Table 4.6 Coefficients of Regression Analysis 3

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients		
Model	B	Std. Error	Beta	t	Sig.
1 (Constant)	3.771	0.040		94.182	0.000
Quality luxury	0.093	0.005	0.140	18.436	0.000
Maintenance condition	0.074	0.004	0.160	20.039	0.000
Appearance	0.067	0.007	0.067	9.181	0.000
Volume	0.002	0.000	0.480	51.512	0.000
Residential layers	0.006	0.002	0.024	3.440	0.001
Amount bedrooms	0.009	0.002	0.034	4.198	0.000
Standalone house	0.103	0.012	0.085	8.892	0.000
Semi-detached house	0.075	0.008	0.079	9.501	0.000
Row house	-0.033	0.005	-0.053	-6.233	0.000
Size garden	0.000	0.000	0.191	17.180	0.000
Bay window	0.097	0.006	0.119	16.642	0.000
Dormer	0.027	0.005	0.037	5.158	0.000
Balcony	0.025	0.008	0.025	3.210	0.001
Basement	0.059	0.021	0.019	2.758	0.006
Garage	0.097	0.006	0.120	16.010	0.000
Barn	-0.036	0.008	-0.049	-4.625	0.000
Garden shed	0.027	0.009	0.020	2.952	0.003
Built between 1950-1959	-0.031	0.009	-0.025	-3.564	0.000
Built between 1960-1969	-0.043	0.006	-0.051	-6.667	0.000

Built between 1990-1999	0.165	0.007	0.193	23.706	0.000
Built between 2000-2009	0.236	0.009	0.234	26.735	0.000
Built after 2009	0.171	0.015	0.084	11.800	0.000
Percentage green	0.001	0.000	0.049	4.789	0.000
Percentage recreation terrain	0.006	0.003	0.016	2.163	0.031
Neighbourhood type 1	0.012	0.004	0.041	3.311	0.001
Neighbourhood type 2	-0.011	0.003	-0.035	-4.109	0.000
Neighbourhood type 3	-0.009	0.002	-0.031	-3.662	0.000
Neighbourhood type 4	-0.014	0.003	-0.040	-5.042	0.000
Neighbourhood type 6	0.013	0.003	0.036	4.319	0.000
Distance to daily facilities	-0.017	0.006	-0.036	-2.686	0.007
Hospital exclusive 5 km	0.069	0.017	0.108	4.039	0.000
Hospital exclusive 20 km	-0.018	0.008	-0.089	-2.342	0.019
Distance pub	0.025	0.004	0.073	5.907	0.000
Distance restaurant	-0.015	0.006	-0.027	-2.262	0.024
Hotel 5 km	0.021	0.002	0.166	9.069	0.000
Distance library	-0.019	0.004	-0.055	-4.452	0.000
Distance swimming pool	0.012	0.003	0.054	4.028	0.000
Distance sauna	0.008	0.001	0.185	7.879	0.000
Distance theme park	-0.017	0.002	-0.217	-9.851	0.000
Distance highway	0.017	0.004	0.044	4.247	0.000
Distance railway station	0.004	0.001	0.043	2.479	0.013
Amount of facilities within 10 km	-0.036	0.006	-0.124	-5.878	0.000
Amount of facilities within 20-50	0.046	0.012	0.156	3.944	0.000
Amount of daily facilities within 3-	-0.056	0.013	-0.169	-4.166	0.000
Amount of daily facilities within 3-	0.020	0.005	0.064	4.308	0.000
Amount of daily facilities within 3-	-0.048	0.005	-0.167	-9.086	0.000
Distance to leisure and high	-0.031	0.007	-0.077	-4.753	0.000
Transaction year 2016	0.029	0.005	0.047	6.252	0.000
Transaction year 2017	0.059	0.006	0.074	9.632	0.000
Shortage indicator	-0.004	0.001	-0.054	-4.964	0.000

Table 4.4 shows that 84.1% of the total variance of the log-transaction price can be explained by the remaining independent variables. Table 4.5 shows the significance level of the model. This is again smaller than 0.0005 so the regression analysis is again significant. Table 4.6 shows the coefficients of all the independent variables that are included in the regression analysis. All variables have a significance level of 0.05 or smaller. This means, that the effects are significant.

That all the variables have a significance level of 0.05 or smaller means that the variables in the regression analysis all have an influence on the value of a dwelling. The individual effects of all the variables can be different. When analyzing the effect that a variable has on the value of a dwelling, a

closer look can be taken to the size of the coefficient and whether it is positive or negative. Besides that, also the standardized coefficients can be taken into account to compare the variables with each other. Analyzing the results of this regression analysis will be done in the next section. When a variable has an unexpected effect, for example a positive coefficient instead of a negative coefficient, this will be discussed.

4.3 Results

Table 4.6 shows which variables have an influence on the value of a dwelling and also which influence each variable has. The coefficients of the variables represent the marginal value of the variables. In this way the presence of a balcony has a higher marginal value than the volume of a dwelling. This is because the balcony is either present or not and the volume of a dwelling is a continuous variable. The variable volume is measured on a bigger scale. To be able to compare the effects of different variables on the value of a dwelling, the standardized coefficients can be taken into account. With the standardized coefficients, the scale of the variables is also taken into account. Therefore variables can be compared to each other. Because the dependent variable of the regression analysis has used a log-transformation, the coefficients represent the influence of a variable in terms of a percentage.

4.3.1 Characteristics of the Dwelling Itself

First, a closer look will be taken to the influence of characteristics of the dwelling itself on the value of a dwelling. The variables quality and luxury, maintenance condition, appearance, volume of dwelling, amount of residential layers, amount of bedrooms and the size of the garden all have a positive coefficient. That means that if the quality, maintenance, appearance, volume, amount of residential layers, amount of bedrooms and size of the garden of a dwelling increases, the value of a dwelling also increases. The variables bay window, dormer, balcony, basement, garage, barn and

garden shed show if these attributes are present or not. The size of these attributes is not taken into account. The presence of a bay window, dormer, balcony, basement, garage or garden shed have a positive influence on the value of a dwelling. The presence of a barn has a negative effect on the value of a dwelling. This may be caused by the fact that people do not use barns anymore that often. They might prefer that space as garden.

The type of dwelling and building year of a dwelling are added to the regression analysis as dummy variables. For the variable type of dwelling, a corner house is used as base level. This means that the effects of the other type of dwellings are compared to a corner house. The results show that when a dwelling is a standalone or a semi-detached house, this has a positive influence on the value of a dwelling compared to a corner house. When a dwelling is a row house, this has a negative influence on the value of a dwelling compared to a corner house. When a dwelling is an apartment, this has no effect on the value of a dwelling compared to a corner house. The base level for building year is the group of dwellings built before 1900. This means that the effects of the other groups of building years are compared to the dwellings built before 1900. The results show that when a dwelling is built between 1950 and 1969, this has a negative influence on the value of a dwelling compared to dwellings built before 1900. When a dwelling is built after 1990, this has a positive influence on the value of a dwelling compared to dwellings built before 1900. When a dwelling is built between 1900 and 1949 or 1970 and 1989, this has no effect on the value of a dwelling compared to dwellings built before 1900.

Table 4.7 shows all the characteristics of a dwelling itself that have an influence on the value of a dwelling. The variables are sorted by the size of the standardized coefficients. The volume of a dwelling has the biggest influence on the value of the dwelling, followed by dwellings built between 2000-2009 and dwellings built between 1990-1999.

Table 4.7 Characteristics of the Dwelling Itself

	Coefficients				
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Volume	0.002	0.000	0.480	51.512	0.000
Built between 2000-2009	0.236	0.009	0.234	26.735	0.000
Built between 1990-1999	0.165	0.007	0.193	23.706	0.000
Size garden	0.000	0.000	0.191	17.180	0.000
Maintenance condition	0.074	0.004	0.160	20.039	0.000
Quality luxury	0.093	0.005	0.140	18.436	0.000
Garage	0.097	0.006	0.120	16.010	0.000
Bay window	0.097	0.006	0.119	16.642	0.000
Standalone house	0.103	0.012	0.085	8.892	0.000
Built after 2009	0.171	0.015	0.084	11.800	0.000
Semi-detached house	0.075	0.008	0.079	9.501	0.000
Appearance	0.067	0.007	0.067	9.181	0.000
Row house	-0.033	0.005	-0.053	-6.233	0.000
Built between 1960-1969	-0.043	0.006	-0.051	-6.667	0.000
Barn	-0.036	0.008	-0.049	-4.625	0.000
Dormer	0.027	0.005	0.037	5.158	0.000
Amount bedrooms	0.009	0.002	0.034	4.198	0.000
Built between 1950-1959	-0.031	0.009	-0.025	-3.564	0.000
Balcony	0.025	0.008	0.025	3.210	0.001
Residential layers	0.006	0.002	0.024	3.440	0.001
Garden shed	0.027	0.009	0.020	2.952	0.003
Basement	0.059	0.021	0.019	2.758	0.006

4.3.2 Environmental Factors

Secondly a closer look will be taken to the influence of environmental factors on the value of a dwelling. The percentages green and recreation terrain in a neighbourhood have a positive coefficient. The variable green is a compilation between the variables park and green area, forest, open and natural terrain and inland water. The variable recreation terrain is a compilation between day recreation and residence recreation. That the variables have a positive coefficient means that if the percentages green and recreation terrain in a neighbourhood increases, the value of a dwelling also increases. The variables neighbourhood type 1, 2, 3, 4, 5 and 6 are components that are created with a principal component analysis. The variables in the components tell something about the

population compilation and built up environment of a neighbourhood. In Table 3.3 in chapter 3, the loads of the variables on the components can be seen. The created components all represent dimensions of a type of neighbourhood. These are also discussed in chapter 3. Table 4.8 shows an overview of these dimensions of neighbourhoods.

Table 4.8 Dimensions of Neighbourhoods

Dimensions of Neighbourhoods	Description
1	Privately owned, single family houses, families with young children.
2	Social housing, non-Western inhabitants, households with children.
3	Dwellings built in 2000 or later, relatively much 15-24 years old.
4	Vacant dwellings, Western inhabitants, relatively much 15-24 years old.
5	Vacant dwellings, households without children, relatively less income receivers.
6	Western inhabitants low, relatively much 15-24 years old and less 25-44 years old.

The dimensions of neighbourhood types 1 and 6 have a positive influence on the value of a dwelling. That means, that regardless all the other variables, if a dwelling is located in a neighbourhood that scores high on the dimensions of a neighbourhood as described, the value of the dwelling increases. The dimensions of neighbourhood types 2, 3 and 4 have a negative influence on the value of a dwelling. The dimensions of neighbourhood type 5 does not have an influence on the value of a dwelling. The variable, distance to daily facilities is also a created component. The variables that are included in the component are distance to doctor, pharmacy, supermarket, other daily shops and cafeteria. The coefficient of the variable is negative. This means that when the distance to these facilities increases, the value of the dwelling decreases.

Table 4.9 shows all the environmental factors that have an influence on the value of a dwelling. The variables are sorted by the size of the standardized coefficients. The percentage green in a neighbourhood has the biggest influence on the value of the dwelling, followed by having a dwelling located in a neighbourhood with a high percentage privately owned, single family houses and

families with young children. The third influential factor is having a dwelling located in a neighbourhood with a high percentage vacant dwellings and high percentages immigrants with a Western ethnicity, and 15-24 years old.

Table 4.9 Environmental Factors

Coefficients					
	Unstandardized		Standardized	t	Sig.
	B	Std. Error	Beta		
Percentage green	0.001	0.000	0.049	4.789	0.000
Neighbourhood type 1	0.012	0.004	0.041	3.311	0.001
Neighbourhood type 4	-0.014	0.003	-0.040	-5.042	0.000
Neighbourhood type 6	0.013	0.003	0.036	4.319	0.000
Distance to daily facilities	-0.017	0.006	-0.036	-2.686	0.007
Neighbourhood type 2	-0.011	0.003	-0.035	-4.109	0.000
Neighbourhood type 3	-0.009	0.002	-0.031	-3.662	0.000
Percentage recreation terrain	0.006	0.003	0.016	2.163	0.031

4.3.3 Location Factors

Next, a closer look will be taken to the influence of location factors on the value of a dwelling. These location are factors that consider the location of a dwelling in relation to specific facilities. These factors are often expressed in distance to nearest facility or the amount of facilities within a specific radius. The variables distance to the nearest restaurant, library, theme park and leisure and high schools all have a negative coefficient. This means that if the distance to these nearest facilities decreases, the value of a dwelling increases. The amount of theme parks in the research region is not very big. The theme parks that are located in the research region are all located close to big cities. The effect of distance to nearest theme park therefore probably represents the distance to nearest big city. The variable distance to leisure and high schools is a component that is created with a principal component analysis. The variables that are included in the component are; Distance to nearest department store, hotel, solarium, high school, pre-vocational education and pre-university education, and the amount of department stores within a 5 km radius. The loads of the variables on the component are all positive, except for ‘amount of department

stores within a 5 km radius', This means that when the distance to these facilities decreases and the amount of department stores within a 5 km radius increases, the value of a dwelling increases. The variables distance to the nearest pub, swimming pool, sauna, highway and railway station all have a positive coefficient. This means that if the distance to these nearest facilities increases, the value of a dwelling also increases. When looking at the variable distance to nearest pub, this effect can be explained by the fact that people experience noise disturbance when living close to a pub. This explanation of noise disturbance is probably not applicable to saunas and swimming pools. Therefore a closer look will be taken to these facilities. When the locations of all the saunas and swimming pools in the region of the research are located can be seen that they are not highly represented in the biggest city, Breda. They are however highly represented in the surrounding municipalities. It is therefore possible that the variables sauna and swimming pool represent another character. A character that is not included in the research. Since the facilities saunas and swimming pools are not highly represented in the city Breda, the effects of the variables might represent that when the distance to the sauna and swimming pool increased and therefore the distance to the city Breda decreases, the value of a dwelling increases. In other words, when a dwelling is located close to the city Breda, the value of a dwelling increases. Also the locations of high ways and railway stations in the region of the research will be located. When looking at these facilities can be seen that there are a lot of high ways and railway stations in the research region. So even the dwellings that are relatively located further away from these facilities, are in fact still very close to high ways and railway stations. It is therefore possible that people living in dwellings relatively further away from high ways and railway stations still think they live close enough to these facilities, while people that live very close to these facilities experience noise disturbance. The combination of these factors might explain why the value of a dwelling increases when the distance to high ways and railway stations increases.

The variables amount of hospitals within 5 km, amount of hotels within 5 km and amount of facilities within 20-50 km all have a positive coefficient. This means that if the amount of these facilities within a specific radius increases, the value of a dwelling increases. The variable amount of facilities within 20-50 km is a component that is created with a principal component analysis. The variables that are included in the component are; The amount of department stores, hotels, cinemas, theme parks and employment opportunities within a 20 km radius, the amount of theme parks and employment opportunities within a 50 km radius and the distance to the nearest artificial ice skating rink. The loads of the variables on the component are all positive, except for 'distance to ice skating rink'. This means that when the amount of these facilities within a specific radius increases and the distance to the nearest ice skating rink decreases, the value of a dwelling increases. The influence of the variable distance to ice skating rink might be strange. It is therefore possible that this variable represents another character which is not included in the research. When the locations of ice skating rinks in the region of the research are located can be seen that there is only one ice skating rink outside of Breda. All the other ice skating rinks are located in Breda. Therefore the variable ice skating rink might have the same effect as the variables sauna and swimming pool. It is probably not the ice skating rink that causes the value of a dwelling to increase, but the fact that these dwellings are located close to a bigger city.

The variables amount of hospitals within 20 km and amount of facilities within 10 km have negative coefficients. This means that if the amount of these facilities within a specific range increases, the value of a dwelling decreases. In the previous section was noted that the amount of hospitals within 5 km has a positive influence on the value of a dwelling. This might mean that people appreciate a higher amount of hospitals within a smaller radius but not in a bigger radius. The variable amount of facilities within 10 km is a component. The variables that are included in the component are the same variables that are included in the variable amount of facilities within 20-50 km, only the radius

is smaller and the variable ice skating rink is not included. This might mean that people appreciate a high amount of facilities within a 20-50 km radius but not having all these facilities in a smaller radius.

The variables amount of daily facilities within 3-5 km are components. The variables in the components tell something about the address density, distance to nearest hospital and amount of daily facilities in a neighbourhood. In Table 3.15 in chapter 3, the loads of the variables on the components can be seen. Table 4.10 shows an overview of these components.

Table 4.10 Components Amount of Daily Facilities

Component	Description
1	High address density, small distance to nearest hospital and high amount of high schools, doctors, pubs, cafeterias, restaurants, day cares, after school cares, elementary schools, supermarkets and other daily shops.
2	High address density, small distance to nearest hospital and high amount of high schools, doctors, pubs, cafeterias, restaurants, day cares, after school cares, elementary schools, supermarkets and other daily shops within 3 km.
3	High distance to nearest hospital, high amount of high schools, and pubs.

The components 1 and 3 have a negative coefficient in the regression analysis. That means that a high address density, a small distance to a hospital and a high amount of daily facilities in a 3-5 km radius have a negative effect on the value of a dwelling. This might mean that people do not like these facilities too close to their dwelling because they might suffer from disturbance. Component 2 has a positive coefficient in the regression analysis. That means that a high address density, a small distance to a hospital and a high amount of daily facilities in a 3 km radius have a positive effect on the value of a dwelling. The loads of the variables on component 2 are however very weak.

Table 4.11 shows all the location factors that have an influence on the value of a dwelling. The variables are sorted by the size of the standardized coefficients. The distance to the nearest theme

park has the biggest influence on the value of the dwelling, followed by the distance to the nearest sauna. These effects probably represents the distance to a bigger city. The third influential factor is the amount of daily facilities within a 3-5 km radius.

Table 4.11 Location Factors

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Distance theme park	-0.017	0.002	-0.217	-9.851	0.00
Distance sauna	0.008	0.001	0.185	7.879	0.00
Amount of daily facilities within 3-5 km 1	-0.056	0.013	-0.169	-4.166	0.00
Amount of daily facilities within 3-5 km 3	-0.048	0.005	-0.167	-9.086	0.00
Hotel 5 km	0.021	0.002	0.166	9.069	0.00
Amount of facilities within 20-50 km	0.046	0.012	0.156	3.944	0.00
Amount of facilities within 10 km	-0.036	0.006	-0.124	-5.878	0.00
Hospital exclusive 5 km	0.069	0.017	0.108	4.039	0.00
Hospital exclusive 20 km	-0.018	0.008	-0.089	-2.342	0.01
Distance to leisure and high schools	-0.031	0.007	-0.077	-4.753	0.00
Distance pub	0.025	0.004	0.073	5.907	0.00
Amount of daily facilities within 3-5 km 2	0.020	0.005	0.064	4.308	0.00
Distance library	-0.019	0.004	-0.055	-4.452	0.00
Distance swimming pool	0.012	0.003	0.054	4.028	0.00
Distance highway	0.017	0.004	0.044	4.247	0.00
Distance railway station	0.004	0.001	0.043	2.479	0.01
Distance restaurant	-0.015	0.006	-0.027	-2.262	0.02

4.3.4 Circumstances of Transaction

Next, a closer look will be taken to the influence of the circumstances of transactions in the value of a dwelling. By that is meant the year of the transaction. This variable is included in the regression analysis as a dummy variable to control for the influence of the year the dwelling was sold. For the variable transaction year, the year 2015 is used as base level.. This means that the effects of the other transaction years are compared to the year 2015. Selling a dwelling in the year 2016 has a positive influence on the value of a dwelling compared to a transaction in the year 2015. Selling a dwelling in the year 2017 has an even bigger positive influence on the value of dwelling.

These effects are probably the result of a growing economy in The Netherlands. Table 4.12 shows the standardized coefficients of these variables,

Table 4.12 Transaction Circumstances

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Transaction year 2016	0.029	0.005	0.047	6.252	0.000
Transaction year 2017	0.059	0.006	0.074	9.632	0.000

4.3.5 Shortage

Next, a closer look will be taken to the influence of shortage on the value of a dwelling. In the regression analysis the variable shortage indicator is included. This variable is an indicator that shows how many dwellings were available for people when they were buying a house. It is a ratio of supply and transactions. Since the name of the variable is 'shortage indicator', it is expected that the bigger the variable is, the more shortage there is, but it is the other way around. That means that the bigger the indicator is, the more dwellings were available for people when they were buying a house and the smaller the shortage. The variable of shortage has a negative coefficient in the regression analysis. That means that when the shortage indicator decreases, the value of a dwelling increases. In other words, when there are less dwellings available for people when they are buying a house, the value of a dwelling increases. This effect was expected. Table 4.13 shows the standardized coefficient of this variable.

Table 4.13 Shortage

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Shortage indicator	-0.004	0.001	-0.054	-4.964	0.000

4.3.6 INTERPRETATIONS

The results show that with the regression model 84.1% of the total variance of the value of a dwelling can be explained. When a model can explain 70% or more of the total variance can be concluded that the model is very strong (Reasearch Gate, 2018). Since the model of this research can explain 84.1% of the total variance can be concluded that the regression model of this research is very strong.

The effects of the variables on the value of a dwelling are already discussed and when the variables have an unexpected effect, possible explanations are given. All the standardized coefficients of variables in a specific group are summed up so they can be compared to each other. The characteristics of the dwelling itself have the biggest influence on the value of a dwelling. The sum of the standardized coefficient of these variables is 2.288. The volume of a dwelling has the biggest influence with a standardized coefficient of 0.480. Also the environmental factors of a dwelling have an influence on the value of a dwelling. The sum of these standardized coefficients is 0.284. The most important environmental factor is the percentage green in a neighbourhood with a standardized coefficient of 0.049. The last group of variables that has an influence on the value of a dwelling are the location factors. The sum of these standardized coefficients is 1.744. The influence of location factors on the value of a dwelling is bigger that the influence of environmental factors but not as big as the characteristics of the dwelling itself.

As a result from the literature study in chapter 2 can be concluded that the distance to specific facilities has an influence on the value of a dwelling. From the regression model of this research can be concluded that a difference can be made between the effects of different types of facilities. For this research, several types of variables are used for the facilities. For every facility, there is not only a

variable that shows the distance to this nearest facility, but there are also variables that show the amount of these facilities within in specific radius. The regression model shows, that the distance to daily facilities such as a doctor, supermarket, pharmacy, cafeteria and other daily shops has an influence on the value of a dwelling. The amount of daily facilities within a specific radius does not have an effect on the value of a dwelling. With the non-daily facilities, the effects are the other way around. Non-daily facilities are department stores, hotels, cinemas and theme parks. The distance to these nearest facilities do not have an effect on the value of a dwelling, except for the distance to a theme park. The amount of these facilities that are available in a 20-50 km range do have an effect on the value of a dwelling. From these results can be concluded that people want the distance to daily facilities to be small, but how many of these daily facilities are located close to their dwelling is not important. People want the amount of non-daily facilities within a 20-50 km radius to be high. The distance to the nearest non-daily facility is not important. The most important location or environmental factor is the distance to a theme park. The closer a dwelling is located to a theme park, the higher the value of a dwelling. This effect is probably caused because another factor is represented. In this case, the distance to the nearest big city.

The year of the transaction does also have an effect on the value of a dwelling. This effect is probably the result of a growing economy in The Netherlands. The variable shortage was not found in the literature study but does have an effect on the value of a dwelling. When there is shortage in a specific region the value of a dwelling increases. Although this effect was not found in the literature study, it is an expected result.

4.4 CONCLUSIONS

The regression model that is created in this chapter is very strong and it is a good model to predict the value of a dwelling. One of the most important findings of the regression analysis is that a difference can be made between daily and non-daily facilities. There are also characteristics that are still not included in the model. This can for example be seen from the effects that distance to theme parks, saunas and swimming pools have on the value of a dwelling. These effects represent probably the distance to the nearest big city.

In this chapter, the value of a dwelling is analysed by means of a hedonic price analysis. The values of specific factors of the dwelling are determined and combining all the values together determines the value of the dwelling. But there are also other techniques that can help determining the value of a dwelling. In the next chapter, two data mining techniques will be used to determine the value of a dwelling. The model fit of these techniques can be compared to the fit of the regression model of this chapter.

5 DATA MINING TECHNIQUES METHOD AND RESULTS

In the previous chapter, the influence of characteristics of a dwelling itself, scarcity and location and environmental factors on the value of a dwelling is analysed using a regression analysis. In this chapter the influence of these characteristics on the value of a dwelling will be analysed using data mining techniques. The goal is to investigate how data mining techniques can be used to extract information about the influence of spatial factors on the value of a dwelling. Data mining techniques are techniques that search for (statistical) relationships in data sets with the aim of creating profiles. The data mining techniques that have been chosen are creating a regression tree and a random forest. First, a model will be derived from the data with the use of a regression tree. The method will be described, the different steps that will be taken in this process will be described and at last the results will be shown and discussed. Secondly, a model will be derived with the use of a random forest. Again, the method will be described, the different steps in this process will be described and the results will be shown and discussed. The results of the data mining techniques will be considered to see whether the accuracy of the prediction of the hedonic price analysis can be increased.

5.1 Regression Tree

First a regression tree will be created to determine which factors have an influence on the value of a dwelling. A regression tree induction method is a machine learning algorithm that divides the data into subsets. Graphical models are created that predict outcomes based on the classification of cases using a tree structure. Various branches of variable length are formed and the most important factors are located at the top of the tree. Regression trees have the aim to create easy readable overviews of categories that are created to predict outcomes.

5.1.1 Methodology

Regression trees are a subsection of decision trees. Decision trees are derived by repeatedly dividing data into multiple subsets. A subset is as homogeneous as possible. A regression tree is created when the target variable, or dependent variable, is continuous. When the target variable is categorical, a classification tree is created. The difference between these two trees is that the subsets that are created with classification trees are fixed and the subsets that are created with regression trees are variable. The subsets that will be created in a classification or regression tree are called leaves. These leaves are located at the end of a branch and an estimated value or classification is assigned to this leaf. The value or category that is assigned to a leaf, is the predicted value or category of an object that is classified to this leaf.

There are many methodologies for creating regression trees but the most commonly used is CART (Classification And Regression Tree). This method has been developed by Breiman et. al. (1984) and will be used for creating a regression tree. There are several functions that can be used with CART. For creating a regression tree, a logical way to choose the split criteria is to reduce the sum of squares. Because the F-statistic is used, this method is called Anova. When this method is used,

regression trees divide a data set into smaller groups and then fit a constant value to each sub group. This constant is the mean of each sub group.

The most important aspect of the method is how the subsets are created. The regression tree is created top-down and when using CART, the splitting points are binary. The model begins with the entire data set and every value of every variable will be analyzed to find the best fitting variable and split value that divides the data set into two regions. The first split, resulting in the first node, will be done with the variable that has the highest association with the dependent variable. It is not important that the split results in two remaining data sets with an equal amount of cases. It is important that the residual sum of squared error (RSS) of the two remaining data sets is minimized. Now the first two branches of the regression tree are created. The splitting will be repeated in each of the two branches and will happen in the same way as the first split. For every split the RSS has to be minimized. This process is continued until no further splits can be made. No further splits can be made when one of the stopping criteria is reached. There are three stopping criteria. The first is when all leaves are pure with a single class or value. This will probably not occur in this research because the variable value of a dwelling has a big range. The second stopping criteria is when a pre-specified minimum number of cases cannot be assigned to a leaf when the splitting continuous. The third and last stopping criteria is when dividing the data set further is not significant. Whether splitting is significant is determined by a value called the complexity parameter. With Anova this means that the R-squared must increase with a minimum of this value at each split. The standard value of the complexity parameter is 0.01.

5.1.2 Results

The data set that will be used for creating a regression tree is the data set from the nine municipalities in the west of the province of Noord-Brabant in The Netherlands. A tree structure with

variables that determine in which category a dwelling is placed will be created. All the categories that will be created have a value. This value is the mean value of the dwellings that are placed in that category. So when the value of a new dwelling needs to be estimated, the estimated value is the value of the category where this dwellings will be located in. First must be checked if the data set that is used for the regression analysis in the previous chapter needs to be adjusted. This is because some of the original variables are changed into components or dummy variables. When two variables have a high correlation and they both have an influence on the value of a dwelling, the best variable will be chosen for a split. The other variable will probably not be used further in the tree anymore because it makes no significant distinction between two branches. When the variable is not visible in the tree, it is unknown that it has a significant influence on the value of a dwelling. Therefore the components that are created for the regression model will also be used for the regression tree. In that way, correlating variables are taken into account together. There are also some variables that are taken into account in the regression model as dummy variables. This is done to compare the effects of different categories with each other. The variable type of dwelling is one of these variables but in the regression tree the variable will not be included as a dummy variable. Only one variable that indicates the type of dwelling will be included. This will be a categorical variable. The variable building year was in the regression model also included as a dummy variable. From the results in the previous chapter can be seen that the building year of a dwelling does not have a linear effect on the value of a dwelling. Therefore the variable building year will also in the regression tree be included as a dummy variable. The dependent variable of the regression tree is the transaction price and not, as in the regression model, the log-transaction price. When the data set is changed a regression tree can be created. The results are shown in figure 5.1.

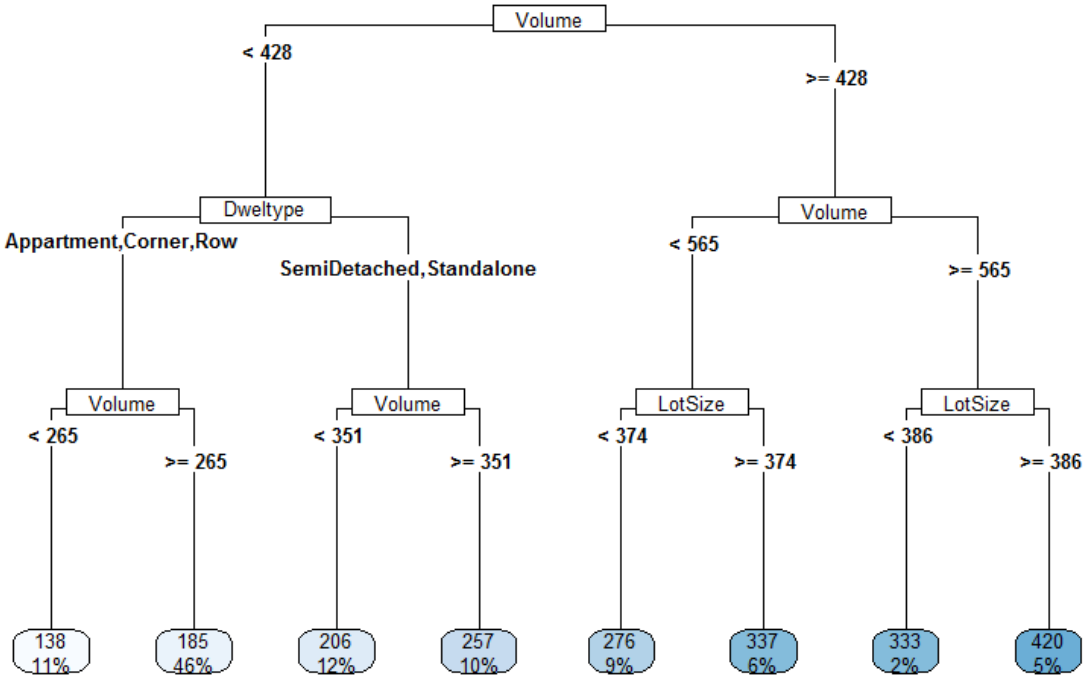

Figure 5.1 Regression Tree

The regression tree in figure 5.1 shows the variables that have a key function when estimating the value of a dwelling. The root node is placed at the top of the tree and the first split of the regression tree is done with the variable that has the highest association with the dependent variable. This is in this case the volume of a dwelling. Dwellings with a volume smaller than 428 cubic meter are placed in the left branch of the tree and dwellings with a volume equal to or bigger than 428 cubic meter are placed in the right branch. Then both branches are split again. The variable that has the highest association with the dependent variable in the left branch is the type of dwelling. Apartments, corner houses and row houses are placed in the left branch. Semi-detached and standalone houses are placed in the right branch. The variable that has the highest association with the dependent variable in the right branch is again the volume of the dwelling. All dwellings in this branch have a volume equal to or bigger than 428 cubic meter, but the dwellings with a volume smaller than 565 cubic meter are divided to the left and dwellings with a volume equal to or bigger than 565 cubic meter are divided to the right. In the four branches, the last separation is done using the variables volume and lot size. All branches lead to a leaf and in the leaf two numbers are placed. The top number in a leaf

is the estimated value of the dwelling. This value has to be multiplied with 1.000. The lower number in a leaf is the percentage of dwellings that is placed in that category. This is how the regression tree should be interpreted. In other words, when a dwelling has a bigger volume than 428 cubic meter, but smaller than 565 cubic meter and the lot size of the dwelling is smaller than 374 square meter, the estimated value of the dwelling is €276.000. Nine percent of the dwellings is located in this category.

The leaves that represent the created categories are coloured, the darker the leaf, the higher the estimated value of a dwelling. When there is a splitting point, the values of the splitting variable that result in a lower estimated value of a dwelling go to the left branch of the regression tree and the values of the splitting variable that result in a higher estimated value of a dwelling go to the right branch. The result is that the estimated values of the dwellings are sorted. In general, the estimated values are increasing from left to right.

With the first regression tree, eight categories are created. The regression tree did not split any further because one of the stopping criteria is reached. The stopping criteria that is reached is the third stopping criteria; dividing the data set further is not significant. The complexity parameter is 0.01 and this is its standard value. The complexity parameter can be lowered so more splitting points are created, more categories are created and more different variables are used for these splitting points. When more variables are used in the regression tree, more information about the influence of these variables can be retrieved. Therefore the results of several complexity parameters are analyzed to see which regression tree contains the most information while it is still readable. This is a regression tree with a complexity parameter of 0.0017. Figure 5.2 shows how the relative error changes while splitting points are added. In the figure can be seen that adding more splitting points will not decrease the relative error much more. The R square of the model is 1 minus the relative

error. That means that the R square and therefore the fit of this model is 0.753. Figure 5.3 shows the overview of the regression tree.

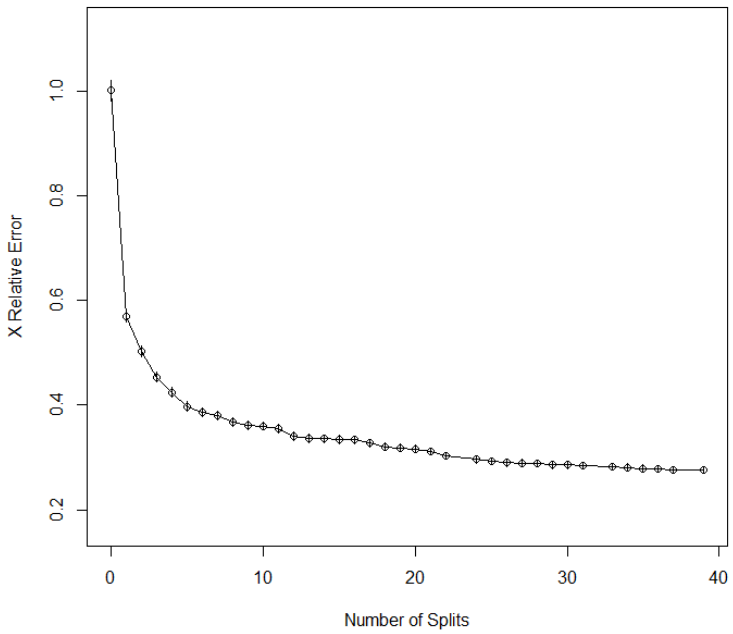

Figure 5.2 Relative error of regression tree.

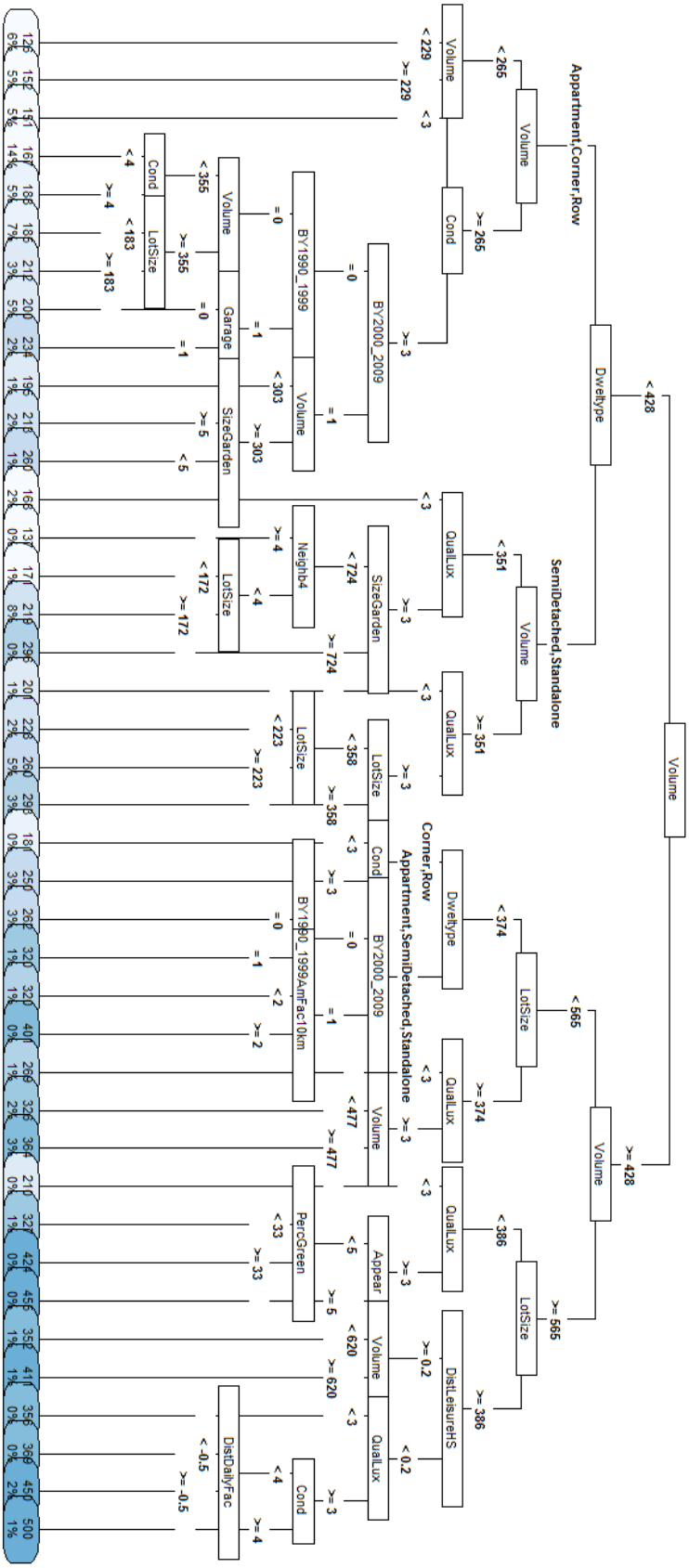

Figure 5.3 Regression Tree

Since the size of the regression tree is still very extensive, the regression tree will be split up virtually after the first node. The left side of the tree is shown in figure 5.4.

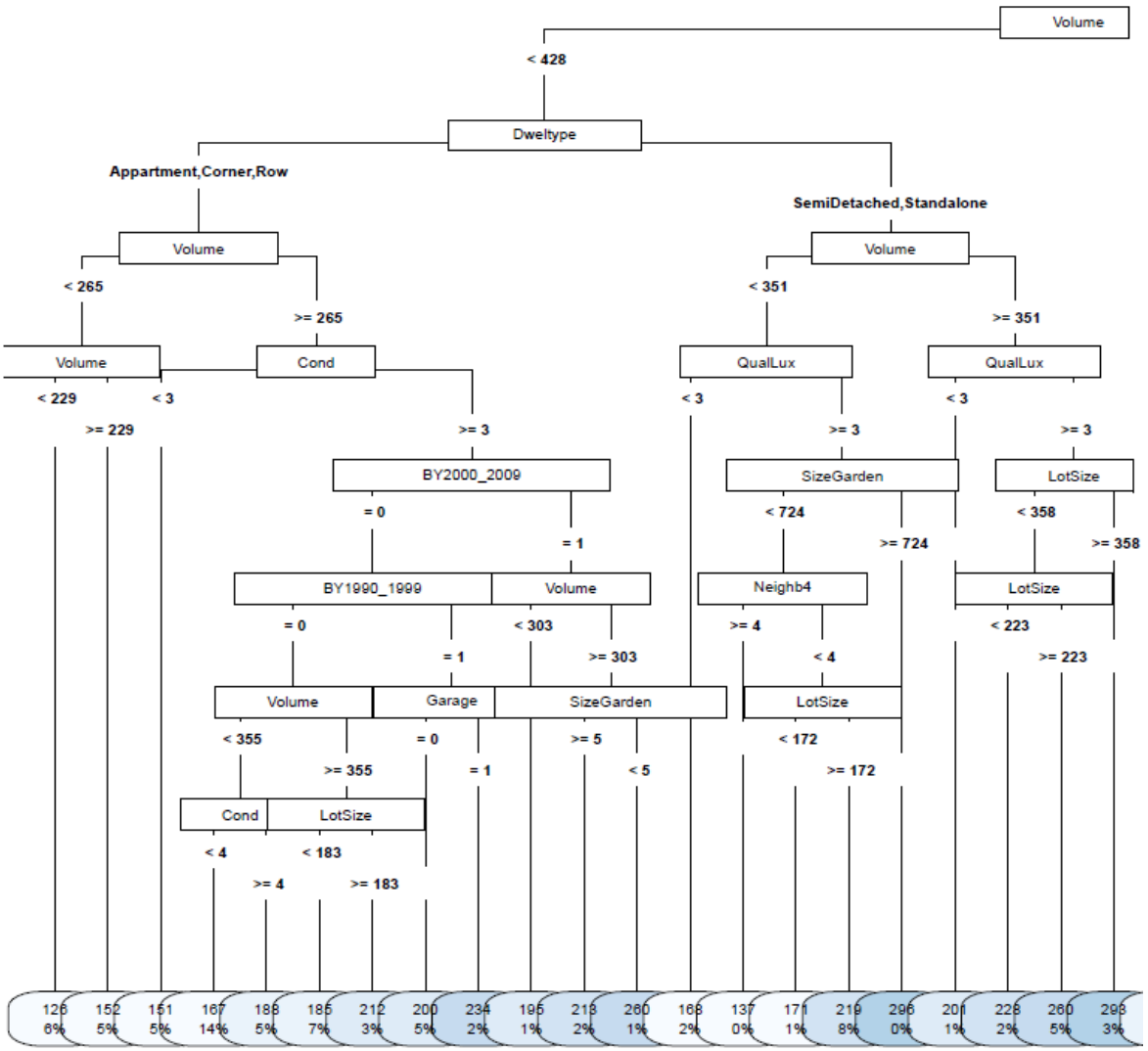

Figure 5.4 Left Part of Regression Tree

Figure 5.4 shows the part of the tree after the first splitting is completed. The dwellings that are placed in this part of the tree are the dwellings with a volume smaller than 428 cubic meter. That the characteristics of a dwelling itself have the most influence on the value of a dwelling is already known from the regression analysis in chapter 4. This part of the regression tree shows again that these characteristics have a bigger influence on the value of a dwelling than location and environmental factors. The only environmental factor that is mentioned is neighbourhood type 4.

Since the aim of this research is to investigate the influence of location and environmental factors on the value of a dwelling, this splitting point will be evaluated. The variable neighbourhood type 4 is a component created with variables that tell something about the population compilation and built up environment of a neighbourhood. The dimensions that are associated with this neighbourhood type are a high percentage vacant dwellings, inhabitants in the age 15-24 years old and immigrants with a Western ethnicity and a low percentage of households with children. When a dwelling scores higher on this component, to be precisely four or more, the estimated value of a dwelling is lower than when a dwelling scores lower than a four on this component. This distinction is only made in a specific part of the regression tree. Figure 5.5 shows the right part of the regression tree.

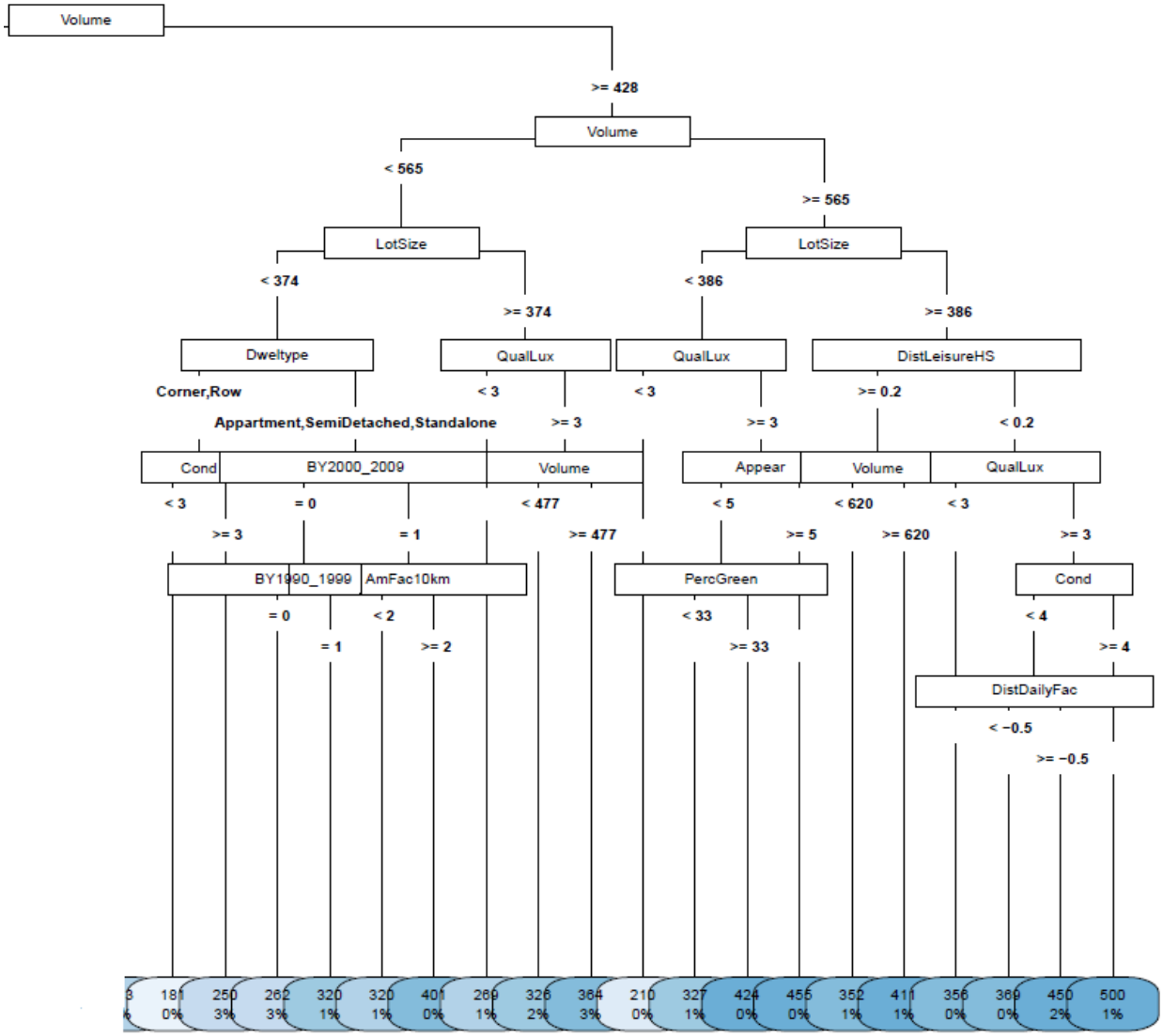

Figure 5.5 Right Part of Regression Tree

Figure 5.5 shows the categories that can be created with dwellings which have a volume of 428 cubic meters or bigger. Also on this side of the tree can be seen that the characteristics of the dwelling itself have a big influence on the value of a dwelling. These variables are often chosen as splitting variables, which means they have the highest association with the value of a dwelling. On this side of the tree there are however more location and environmental factors chosen as splitting point than at the left side of the tree. Again, since the aim of this research is to investigate location and environmental factors these splitting points will be evaluated.

The variables distance to leisure and high schools, the amount of facilities within a 10 km radius, the percentage green and the distance to daily facilities are all used as splitting points in this regression tree. The variable distance to leisure and high schools is located the closest to the top and has therefore the biggest influence. This variable is a component created with the variables distance to nearest department store, hotel, solarium, high school, pre-vocational education and pre-university education, and the amount of department stores within a 5 km radius. When the component increases, the distance to these facilities increases and the amount of department stores decreases. The regression tree shows that when a dwelling scores higher on this component, the estimated value of a dwelling is lower than when a dwelling scores lower on this component. This means that when the distance to these facilities increases and the amount of department stores decreases, the value of a dwelling decreases.

The variable amount of facilities within a 10 km radius is also a component created with the variables amount of department stores, hotels, cinemas, theme parks and employment opportunities in a radius of 10 km. When the component increases, the amount of these facilities increases. The regression tree shows that when a dwelling scores higher on this component, the estimated value of

a dwelling is higher than when a dwelling scores lower on this component. This means that when the amount of these facilities increases, the value of a dwelling also increases.

The percentage green in a neighbourhood is also a splitting point in the right side of the tree. As can be seen in the figure, it is the last splitting point before the branches continue into leaves. When a dwelling follows a certain path in the regression tree and the percentage green in a neighbourhood is 33% or more, the estimated value of the dwelling is €424.000. When a dwelling follows the same path, but the percentage green in that neighbourhood is lower than 33%, the estimated value of the dwelling is €327.000. That is a very big difference.

The last variable that will be discussed is the variable distance to daily facilities. This is a component created with the variables distance to doctor, pharmacy, supermarket, other daily shops and cafeteria. When the component increases, the distance to these facilities increases. The regression tree shows that when a dwelling from a specific part of the regression tree scores lower on this component, the estimated value of a dwelling is lower than when a dwelling scores higher on this component. This means, that when the distance to these facilities increases, the value of a dwelling increases. This effect is the other way around as with the regression analysis in the previous chapter. The dwellings where this effect occurs to in the regression tree are however very specific. The regression analysis shows an overall effect.

5.1.3 Interpretations

Using the regression tree gives an insight in which factors have an influence on the value of a dwelling. A distinction can be made for several clusters of dwellings. In general, the characteristics of the dwelling itself have the biggest influence. The variables volume of a dwelling, lot size, quality of a dwelling and condition of a dwelling are highly represented in the regression tree. That means that

these variables several times have the highest association with the dependent variable. But location and environmental factors are also represented in the regression tree. From the regression tree can be noted that on the left side of the tree, where the smaller dwellings are located, almost only the characteristics of the dwelling itself have an influence on the value of a dwelling. When a dwelling is bigger, the location and environmental factors have a more significant influence.

In this way can be identified which factors have an influence on the value of a dwelling for each cluster of dwellings. This can for example be useful for investment businesses. When an investment business is building smaller dwellings, creating good characteristics of the dwelling itself is important for increasing the value of a dwelling. When bigger dwellings are built, the location according to facilities is worth taken into account. The location and environmental factors; neighbourhood type 4, the distance to leisure and high schools, the amount of facilities within a 10 km radius, the percentage of green and the distance to daily facilities all have an influence on the value of a dwelling in a specific part of the regression tree.

The R square of the model is 0.753. The R square of the regression model from the previous chapter is 0.841. This means that the regression model has a better fit and therefore explains more of the variance. However, the regression tree can be an adjustment to the regression model because it gives an insight of the influence of location and environmental factors for different clusters of dwellings.

5.2 Random Forest

The second data mining technique that will be used to analyse which spatial factors have an influence on the value of a dwelling is a random forest. A random forest induction method is a machine learning algorithm. It builds multiple classification or regression trees and combines them to

retrieve a better prediction than the prediction of a single regression tree. So the aim is to reduce the variance error of a single regression tree. A random forest is designed in such a way that it provides insight about which variables have an influence on the dependent variable. In the results is no tree structure included as seen with a regression tree.

5.2.1 Methodology

One of the advantages of a regression tree is that immediately a clear overview is created about the influence of specific variables on the dependent variable. A disadvantage of a regression tree is that in every split, the variable that has the highest association with the dependent variable is chosen as splitting variable. With this methodology, it is possible that there is a variable that has an influence on the dependent variable but it will not be used in any splitting point because it never has the highest association with the dependent variable. That variable will then not be visible in the regression tree and it is unknown that this variable has an influence on the dependent variable. The high correlations are removed from the data set of the nine municipalities in the west of the province of Noord-Brabant, but there are still low correlations that contribute to this problem. Multicollinearity is however not a problem when deriving a model with the use of a random forest. Another problem that can occur with regression trees is overfitting. This is when a model does not generalize the used data set to a new unseen data set. In other words, the created regression tree can predict the value of the dwellings in the data set very well, but when new data is added the model predicts less good. Overfitting is also not a problem when deriving a model with the use of a random forest.

A random forest is created by combining multiple regression trees (Breiman, 2001). The algorithm for building a random forest differs in two ways from the algorithm for building a regression tree. First, known under the name ‘tree bagging’, not all the cases from a data base are used to create a

regression tree but a random sample is used. Secondly, known under the name 'feature bagging', not all variables are used to create a regression tree but the variables are chosen again randomly. In tree bagging, one third of the cases of a data set is left out and not used for creating the regression tree. These cases are used as a training set to see how well the tree fits. With this method, the problem of overfitting is handled. With the use of feature bagging, the output and fitting of a random forest model are the mean output of all the regression trees. With this method, the problem of multicollinearity is handled.

A disadvantage of deriving a model with the use of a random forest is that a large number of regression trees makes running the algorithm slow. This is especially a disadvantage when the algorithm is used to do real time predictions. The more trees are used for a random forest, the more precise the predictions are. This results however in a slower model. The amount of regression trees in a random forest can be set. It is possible to create a graph that shows how much the error of the model decreases when using a specific amount of regression trees. This will be used to determine how many regression trees can be used for a model.

5.2.2 Results

The data set that will be used for deriving a random forest is again the data set from the nine municipalities in the west of the province of Noord-Brabant. It is possible to remove the created components out of the data set and include the variables of these components again. This could be done because multicollinearity is not a problem when deriving a model with the use of random forest. To compare the results of the random forest with the results of the regression model and the regression tree, the data set with the created components will be used again. In that way, it is easier to compare the different effects of the same variables in different analysing techniques.

With the use of the program R, a model is derived with the use of a random forest. Figure 5.6 shows the out of bag error of the random forest. This is an error term of a random forests. The graph can help to determine how many regression trees should be used for a random forest. The more trees are used, the better the model will fit. There will be a point where each additional tree does not improve the overall model performance anymore and when more trees are added, more time is needed to create the model. As can be seen from figure 5.6, the first 100 regression trees have the biggest influence on decreasing the error of the model. From then on, the error only decreases minimally. For this research only one random forest will be created and it is not important that it takes 15 to 20 minutes to create this model. Therefore the standard amount of 500 regression trees will be maintained to create the random forest for this research.

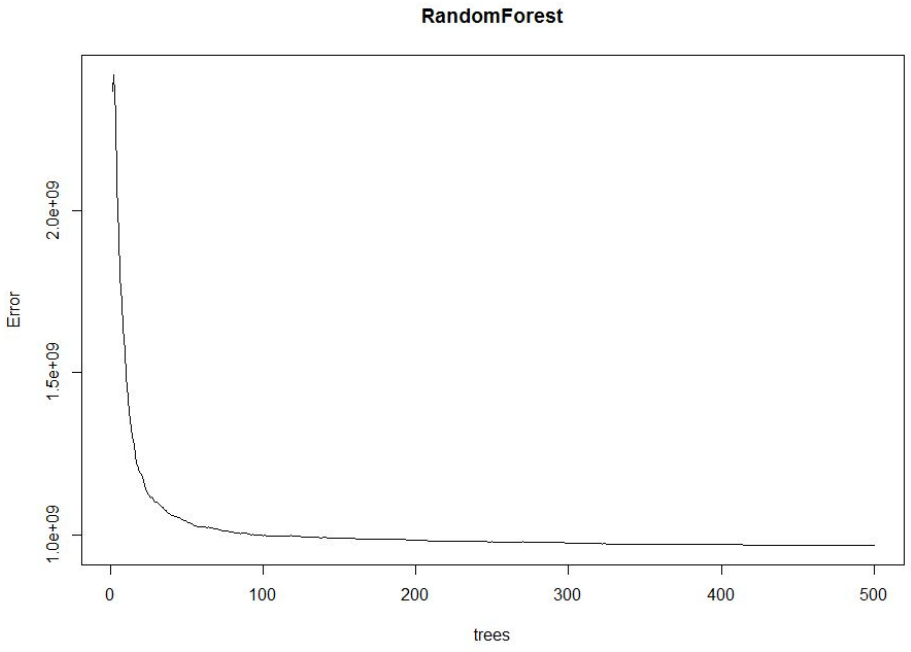

Figure 5.6 Out of Bag Error.

Figure 5.7 shows the variable importance plot of the random forest. This plot shows 30 variables that have the biggest influence on the value of a dwelling. The plot also shows how important each

variable is. This can be seen from the node purity on the x-axis. The node purity is a measurement of how each variable contributes to the purity in each node in a regression tree. The node purity represents the mean marginal influence of a variable. Figure 5.7 shows the variable importance plot of the random forest. This plot shows 30 variables that have the biggest influence on the value of a dwelling. The plot also shows how important each variable is. This can be seen from the node purity on the x-axis. The node purity is a measurement of how each variable contributes to the purity in each node in a regression tree. The node purity represents the mean marginal influence of a variable.

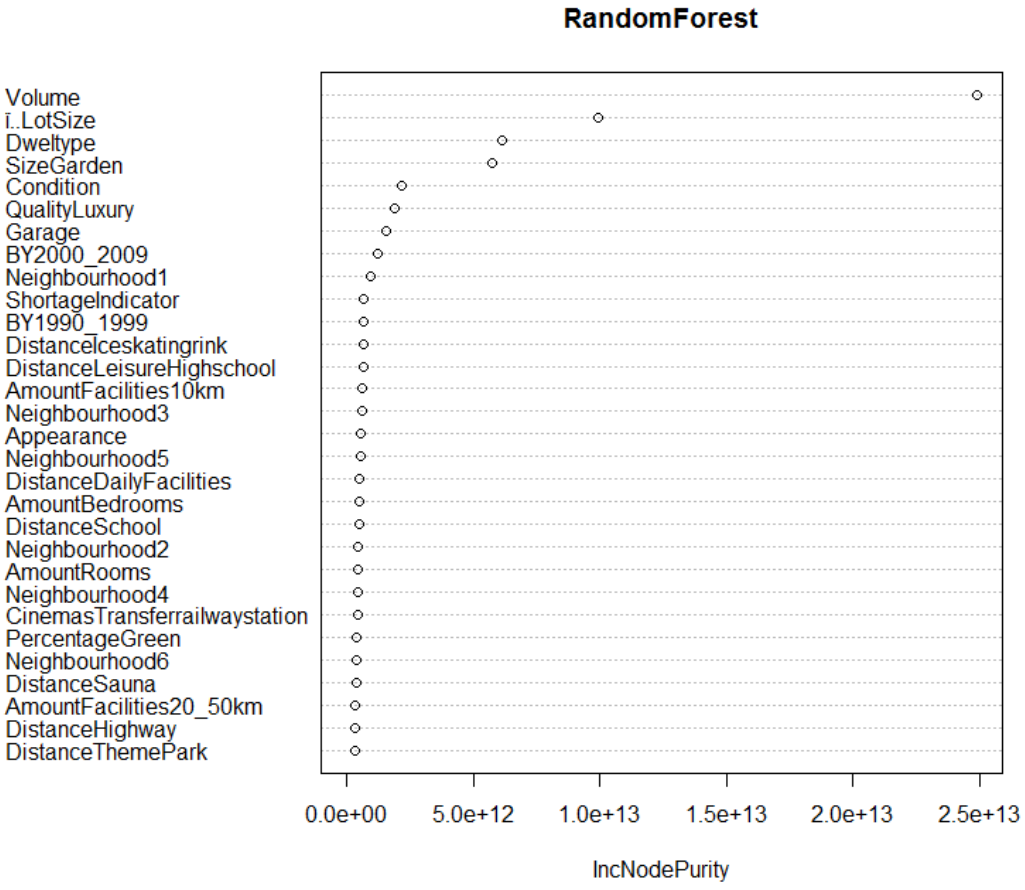

Figure 5.7 Variable Importance Plot.

The variable importance plot shows that the volume of a dwelling has by far the biggest influence on the value of a dwelling followed by the lot size, type of dwelling and the size of the garden. After

that, some variables that have a node purity close to each other follow. These variables are; the condition of a dwelling, the quality of a dwelling, the presence of a garage, the building year between 2000 and 2009 and dimensions of neighbourhood type 1. This represents neighbourhoods with a high percentage privately owned, single family houses and families with young children. This dimension of a neighbourhood is the first environmental or location factor in line and therefore it is the spatial variable with the highest influence on the value of a dwelling. The variables that follow have node purities that are close to each other. The variables that are characteristics of the dwelling itself are; building year between 1990 and 1999, the appearance of a dwelling, the amount of bedrooms and the amount of bedrooms. The variables that are location factors are; distance to an ice skating rink, distance to leisure and high schools, amount of facilities within a 10 km radius, distance to a railway station and cinema, distance to a sauna, the amount of facilities within a 20 to 50 km radius, distance to a high way and distance to a theme park. The variables that are environmental factors are; dimensions of neighbourhood types 2, 3, 4, 5 and 6, distance to daily facilities, distance to a school, a day care and an after school care and the percentage green in a neighbourhood. The last variable that has an influence on the value of a dwelling according to the random forest is shortage.

Since the aim of this research is to investigate the influence of location and environmental factors on the value of a dwelling, these variables will be discussed. The first group of environmental variables that are not used in the regression tree are the dimensions of the types of neighbourhoods. Only the dimension of neighbourhood type 4 was used in the regression tree. Apparently the built up environment and population compilation have an influence on the value of a dwelling. Especially, living in a neighbourhood with a high percentage privately owned, single family houses and families with young children. The variable distance to the nearest school was also not included in the regression tree. This can be caused by a correlation between the variables distance to school and

distance to daily facilities. The variable distance to school is a component created with the variables; distance to elementary school, distance to day care and distance to after school care. This variable also has an influence on the value of a dwelling. The environmental factors distance to daily shops and the percentage green have an influence on the value of dwelling in the regression tree as well as in the random forest.

The location factors distance to ice skating rink, distance to cinema and transfer railway station, distance to sauna, distance to high way and distance to theme parks that are not used for the regression tree but are important variables according to the random forest. As explained in chapter 4, most of these variables probably represent the distance to the nearest big city. In the regression model in chapter 4 was found that the variables distance to nearest transfer railway station and high way have a positive correlation with the value of a dwelling, That means that when the distance to these facilities increases, the value of a dwelling also increases. This was probably caused by noise disturbance when living close to these facilities.

The location factors distance to leisure and high schools and the amount of facilities within a 10 km radius have an influence on the value of dwelling in the regression tree as well as in the random forest. The variable amount of facilities within a 20 to 50 km radius is not mentioned in the regression tree. This is probably caused by a correlation with the variable amount of facilities within a 10 km radius. The variable shortage was also not mentioned in the regression tree but it is according to the random forest the tenth most important variable. The regression model in chapter 4 showed that when there are less dwellings available for people when they are buying a house, the value of a dwelling increases.

The variable importance plot only shows the importance of the variables, not in which way they have an influence on the value of a dwelling. For example, the plot shows that the size of a garden has an important influence on the value of a dwelling. It is unknown if the value of a dwelling increases or decreases when the size of the garden increases. The explained variance of the random forest is 0.870. This is a higher variance than the regression model from chapter 4 (0.841) and the regression tree (0.753).

5.2.3 Interpretation

The results of the random forest show that the model explains 87.0% of the total variance. That is a good fit. The regression model in the previous chapter explains 84.1% of the variance and the regression tree 75.3% so the random forest is the best fitting model. The random forest also gives a new insight in which variables have an influence on the value of a dwelling. Again the volume of a dwelling has the biggest influence on the value of a dwelling, followed by other characteristics of the dwelling itself. The group of location and environmental factors that have an influence on the value of a dwelling according to the random forest model is more extensive than the group of location and environmental factors in the regression tree. This is partly due because in a regression tree it is possible that a variable is used multiple times as a splitting variable. In that way, the variable volume was used nine times as a splitting variable and the variable lot size six times. Another reason why more location and environmental factors are included in the random forest than in the regression tree is because of multicollinearity. Sometimes variables correlate with each other and then only the variable that has the highest association with the dependent variable is used as splitting point.

5.3 Comparing Techniques

The regression tree and the random forest show, as well as the regression analysis in chapter 4, that characteristics of a dwelling have the biggest influence on the value of a dwelling but location and environmental factors and shortage have an influence as well. Most of the variables that have an influence on the value of a dwelling are mentioned in the regression model. From the characteristics of the dwelling itself, only the lot size and the amount of rooms were not mentioned in the regression model. The variable lot size was left out of the model because the correlation with the size of the garden was too high. The variable amount of rooms was included in the regression model but there was a small correlation with the volume of the dwelling. This could be the reason why the amount of rooms is not found as an important variable in the regression model and the regression tree but it was found to be an important variable in the random forest.

From the environmental factors, neighbourhood type 5 and the variable distance to school were not mentioned in the regression model. Why the variable distance to school has no significant influence on the value of a dwelling in the regression model is explained in chapter 4. It is unknown why neighbourhood type 5 has an influence on the value of a dwelling in the random forest but not in the regression model. All the location factors that have an influence in the random forest also have an influence in the regression model. The shortage indicator also has an influence in the regression model as well as in the random forest. There are no variables that have an influence in the regression tree and are do not have an influence in the random forest. Again, this is can be due to the fact that variables are sometimes used multiple times as splitting variable in the regression tree.

In the results of all of the three analysing techniques can be seen that the distance to daily facilities has an influence on the value of a dwelling and the amount of non–daily facilities in a specific radius has an influence on the value of a dwelling. Other location and environmental factors that are always

shown in the results are; the distance to leisure and high schools, the percentage of green and the compilation of the population and the built up environment . Especially the difference between the distance to daily facilities and the amount of non-daily facilities within a specific radius is a result that was unknown before this research.

The random forest has the best model fit of the three analysing techniques. The regression model explains 84.1% of the variance, the regression tree 75.3% and the random forest 87.0%. But using one of the techniques does not have to exclude the other ones. Combining the three techniques gives the best insight in analysing the influence of location and environmental factors on the value of a dwelling. Especially because they all have a different kind of results.

5.4 Conclusions

The regression tree provides an easy readable overview of sub groups of dwellings and predicts a value for these dwellings. It also provides an insight of the influence of location and environmental factors for different clusters of dwellings. Deriving a model using the random forest helps overcome problems in the regression tree with multicollinearity and overfitting. The random forest also provides a more extensive overview of which variables have an influence on the value of a dwelling. The specific influence these variables have on the value of a dwelling is however not given. Therefore the combination of the three analysing techniques gives the best results.

6
CONCLUSIONS AND DISCUSSION

The thesis ends with a conclusion and a discussion. In the section conclusion, the research question is answered and the conclusions are represented. The section discussion shows the problems that were faced during the research, where the results can be used for and some recommendations for further research are done.

6.1 CONCLUSION

Municipalities, companies, institutes and private persons are interested in correct valuation of housing. It is already known that besides the characteristics of the dwelling itself, also the location of a dwelling has an influence on the value of a dwelling. But specific spatial factors of a house are not often included in practice in determining the value of housing, only the location in general. A literature review has been done and several studies show specific spatial factors that have an influence on the value of a dwelling. But the amount of factors that is used in the researches is still not extensive and the data base of the researches is relatively small. Therefore the aim of the research was to investigate the influence of characteristics of the dwelling itself, location factors and environmental factors on the value of a dwelling. Besides that, also the influence of scarcity had to

be investigated. The goal was to not only use commonly known research techniques, but also investigate how data mining techniques can help with this research goal. To answer the main research question, sub questions were formulated.

The first sub question was; How can a data set be created with information about the value of a dwelling, characteristics of the dwellings itself, location factors and environmental factors? A data base with information about the selling price of dwellings and the characteristics of the dwelling itself was provided by nine municipalities in the west of the province of Noord-Brabant in The Netherlands. Information about location and environmental factors and scarcity of these dwellings was added to this data base. Most of the information about location and environmental factors is retrieved from the Central Bureau for Statistics. But also from the Ministry of Internal Affairs and the NVM, the Dutch Organization of real estate agents information is retrieved. In total, 145 variables about location and environmental factors are retrieved and added to the data base of the municipalities. When the data base was complete, the data base was cleaned up. A part of the variables was transformed, the cases that can be used for the analysis were selected and at last the variables were checked to see if there were variables that represent the same factors and if there are variables that correlate with each other. The correlation between the variables was checked and variables that have a high correlation were transformed into components with the use of a principal component analysis.

The second sub question was; How can information about the influence of location and environmental factors be extracted from the created data set and more specific which data mining methods could be used? For extracting the information from the data set, three analysing techniques were selected to analyse the influence of specific factors on the value of a dwelling. These analysing

techniques are a hedonic price analysis with the use of a regression model, a regression tree and a random forest.

The third, fourth and fifth sub question were; Which location and environmental factors have an influence on the value of a dwelling? What is the quantitative effects of these factors? And what is the influence of scarcity on the value of a dwelling? The answers to these sub questions is combined. All three techniques have provided different results. The regression model shows which variables have an influence on the value of a dwelling and what their percentage influence is. The coefficients represent the willingness to pay of a characteristic and can be formulated in terms of euro's. With the use of a standardized coefficient is shown what the relative importance of a variable is. The volume of a dwelling has the most influence on the value of a dwelling, followed by the dwelling being built between 2000 and 2009, the distance to a theme park, the building being built between 1990 and 1999 and the size of a garden. It has a positive effect on the value of a dwelling when the volume and size of the garden increase. It has also a positive effect on the value of a dwelling when a dwelling is built between 1990 and 2009. The value of a dwelling decreases when the distance to the nearest theme park increases. As it appears, the most important variables are characteristics of the dwelling itself. All together, these characteristics of the dwelling itself have the biggest influence on the value of a dwelling. The location factors that have the biggest influence on the value of a dwelling are distance to theme park, distance to sauna and the amount of daily facilities within a 3 to 5 km radius. The effect of the variable distance to theme park is already discussed. The value of a dwelling increases when the distance to the nearest sauna increases. And the value of a dwelling decreases when the amount of daily facilities within a 3 to 5 km radius increases. The environmental factors that have the biggest influence on the value of a dwelling are the percentage green in a neighbourhood, the population compilation and built environment of a neighbourhood and the distance to daily facilities. The value of a dwelling increases when the amount of green increases. The

value of a dwelling also increases when a neighbourhood has a lot of, single family houses and families with young children. The value of a dwelling decreases when the distance to daily facilities increases. Also scarcity has an influence on the value of a dwelling. When there are less dwellings available for people when they are buying a house, the value of a dwelling increases. A result that is not mentioned in the literature yet is that a difference can be made according to the availability of facilities. This research shows that the distance to daily facilities is important and the amount of non-daily facilities in a specific radius are important. The disadvantages of this technique are that the model can have problems with multicollinearity and the model provides just one coefficient for each variable. There is no distinction made between different
sub groups of dwellings.

When a model is derived with the use of a regression tree, a tree structure with sub groups is created. The most important variables in this regression tree are again characteristics of the dwelling itself. The location and environmental factors that have an influence on the value of a dwelling are; the population compilation and built environment of a neighbourhood, distance to leisure and high schools, the amount of facilities within a 10 km radius, the percentage green and the distance to daily facilities. The regression tree also shows what the influence on the value of a dwelling is regarding these variables. The value of a dwelling decreases when the population compilation and built up environment is characterised by high percentages vacant dwellings, inhabitants in the age category 15-24 years old and immigrants with a Western ethnicity and a low percentage of households with children. The value of a dwelling also decreases when the distance to leisure and high schools increases and when the distance to daily facilities decreases. The value of a dwelling increases when the amount of facilities within a 10 km radius increases and when the amount of green in a neighbourhood increases. The tree structure also shows which variables are important for specific sub groups. For smaller dwellings, characteristics of a dwelling itself have a big influence on

the value of a dwelling and location and environmental factors almost not. For bigger dwellings, location and environmental factors have together with characteristics of the dwelling itself an influence on the value of a dwelling. The disadvantages of a regression tree are problems with multicollinearity and overfitting. Multicollinearity is a problem when two variables have a high correlation and they both have an influence on the value of a dwelling. Then the best variable will be chosen for a split. The other variable will probably not be used further in the tree anymore because it makes no significant distinction between two branches. When the variable is not visible in the tree, it is unknown that it has a significant influence on the value of a dwelling.

When a model is derived with the use of a random forest, 500 regression trees are created with randomly chosen variables and cases. Combining all these regression trees gives an insight in which variables have the biggest influence on the value of a dwelling and multicollinearity and overfitting are not a problem anymore. Again, the characteristics of the dwelling itself have the biggest influence on the value of a dwelling. The location factors that have the most influence on the value of a dwelling are; distance to an ice skating rink, distance to leisure and high schools, amount of facilities within a 10 km radius, distance to a railway station and cinema, distance to a sauna, the amount of facilities within a 20 to 50 km radius, distance to a high way and distance to a theme park. The environmental factors that have the most influence on the value of a dwelling are; the population compilation and built environment of a neighbourhood, distance to daily facilities, distance to a school, a day care and an after school care and the percentage green in a neighbourhood. Also the variable shortage has an influence on the value of a dwelling. Again, a difference can be made according to the availability of daily and non-daily facilities. The distance to daily facilities is important and the amount of non-daily facilities. The disadvantage of a random forest is that deriving a model can be a slow process and the created model only shows which variables are important, not in which way they contribute to the value of a dwelling.

The regression model explains 84.1% of the variance, the regression tree 75.3% and the random forest 87.0%. That makes the random forest the model with the best fit. In other words, the random forest can predict the value of a dwelling the best. But the random forest does not provide an insight in which way characteristics of the dwelling itself, location factors, environmental factors and scarcity have an influence on the value of a dwelling. This is done best by using a regression model. The disadvantage of the regression model is that it does not make a distinction between different sub groups. For retrieving an insight in which variables have an influence on the value of a dwelling in different sub groups, the regression tree can be used best. From this research can be concluded that the characteristics of a dwelling have the biggest influence on the value of a dwelling and the location and environmental factors that have an influence are specified. The best of the three methods can be combined to retrieve an optimal result. The random forest model has the highest fit and the importance of variables is shown, with the use of the regression model can be seen if factors have a positive or negative influence in the value of a dwelling and how big their influence is, and the regression tree model can be used to see what the influence on the value of a dwelling is in different sub groups.

The results of this research can be used to determine the quality of a location for housing. For example, values of dwellings are higher when they are located close to leisure and high schools and there is a high amount of non-daily facilities in a specific radius. But the results of this research can also be used to determine which factors could be added to improve the quality of a neighbourhood. In that way, values of dwellings are higher when the amount of green in a neighbourhood is higher and daily facilities are located close to dwellings.

6.2 Discussion

With the use of the three analysing techniques, a valuation tool can be created to determine the qualities of locations and neighbourhoods. The variables that have an influence on the value of a dwelling are indicators for the quality of a location. The level of appreciation for the variables in the research is known now for the function of living. How attractive specific locations are for dwellings can be determined with the use of the analysing techniques. It can also be used to improve specific locations or neighbourhoods. Investment business can, for example, use it when they are looking for a new location for a project. Municipalities can, for example, use it to analysing the quality of neighbourhoods and improve this quality if necessary. Now that is known which location factors are appreciated, for example, shrinking areas can be detected in an early stage and improvements can be made to slow down the shrinkage.

During the research some small problems were faced. The literature study showed that the criminality in a neighbourhood has an influence on the value of a dwelling, but it does not have an influence on the value of a dwelling according to this research. When looking at the levels of criminality in the neighbourhoods of this research was shown that the variance was very low. The factor criminality was represented by the variable liveability and the liveability was in all the neighbourhoods good, very good or excellent. The level of criminality in a neighbourhood might have an influence on the value of a dwelling, but because of the low variance this was not proven in this research. To test this, the research area should be extended with neighbourhoods with a high level of criminality. There are more variables that have a low variance in this research and might have a different result when the variance is bigger, an example is the variable distance to nearest railway station. Another issue was the influence of the distance to the nearest theme park, swimming pool and sauna on the value of a dwelling. The influences of these variables are very big and seemed strange, but after analysing the locations of these facilities in the research area is concluded that

these variables probably represent the distance to the nearest bigger city. This problem was also caused by the use of the relatively small research area. This problem could be solved by extending the research area or by adding the variable ‘distance to a big city’.

Several recommendations can be done for this research. The first is to extend the research area. In that way, the variance increases and the quality of different locations and neighbourhoods can be determined. It is also interesting to execute this research for different target groups. The preferences of families with children might be different than the preferences of elderly people or singles. It was not possible to include this in this research because this data was not available. It is unknown which people bought which dwelling. It is also interesting to repeat this research in a couple of years to see if the preferences have changed over time. The last recommendation is to execute this research in other countries. Some of the results of this research are specific for The Netherlands such as, the distance to elementary school has almost no effect because 95% of the dwellings are located very close to an elementary school because the address density in The Netherlands is very high and the presence of an outdoor swimming pool was not included because of the Dutch climate.

7
REFERENCES

Abelson, P., Joyeux, R. and Mahuteau, S. (2013) Modeling House Prices across Sydney. In: *Australian Economic Review,* 46 (269-285).

Acciani, C., Fucilli, V. and Sardaro, R. (2011) Data Mining in Real Estate Appraisal: a Model Tree and Multivariate Adaptive Regression Spline Approach. In: *Aestimum,* 58 (27-45).

Adomavicus, G. and Tuzhilin, A. (2001) Using Data Mining Methods to Build Customer Profiles. In: *Journal Computer,* 34 (74-82).

Amato, M. d' (2010) A Location Value Response Surface Model for Mass Appraising: An "Iterative" Location Adjustment Factor in Bari, Italy. In: *International Journal of Strategic Property Management*, 14 (231-244).

An, Y., Qiu, G. and Liu, L. (2010) A Study of Real Estate Prices in Jilin Province. In: *The Chinese Economy,* 43 (53).

Bácena, M.J., Menéndez, P., Palacios, M.B. and Tusell, F. (2012) A Real-Time Property Value Index Based on Web Data. In: Watada, J. (2012) *Intelligent Decision Technologies: Proceedings of the 4th International Conference on Intelligent Decision Technologies.* Conferention took place by the 4th Symposium on Intelligent Decision Technologies, Gifu-shi.

Boardman, A., Greenberg, D., Vining, A. and Weimer, D. (2006) Cost–Benefit Analysis, Concepts and Practice, 3rd edn. Prentice Hall, Upper Saddle River, New Jersey.

Bonnetain, P. (2003) A Hedonic Price Model for Island. In: *Journal of Urban Economics*, 54 (368-377).

Breiman, L. (2001) Random Forests. Machine Learning. In: *Journal of Biomedical Science and Engineering,* 45 (5-32).

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984) *Classification and Regression Trees.* Taylor and Francis.

Buczynski, B. (2017) How to Determine Home Value and Why It Matters. Retrieved on August 17th 2018, from https://www.huffingtonpost.com/entry/how-to-determine-home-value-and-why-it- matters_us_5a1f1345e4b039242f8c8158?guccounter=1

CBS (2018) *Regionale Kerncijfers Nederland. Retrieved on October 5th 2018, from http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=70072ned&D1=0&D2=154&D3=21-23&HDR=T&STB=G1,G2&VW=T*

CBS (2018) *Organisatie*. Retrieved on September 10th 2018, from https://www.cbs.nl/nl-nl/over-ons/organisatie.

Chambers, L.P. (1936) Plato's Objective Standard of Value. In: *The Journal of Philosophy*, 33 (596-605).

Cobalt Recruitment (2018) *The Future of Social Housing and New Building Sector*. Retrieved on June 28th 2018, from https://www.cobaltrecruitment.com/news-blog/item/the-future-of-social-housing-and-new-building-sector

Davidoff, I. and Leigh, A. (2008) Howmuch do Public Schools Really Cost? Estimating the Relationship Between House Prices and School Quality. In: *Economic Record,* 84 (193-206).

De Hypotheker (2017) *Marktwaarde.* Retrieved on October 26th 2017, from https://www.hypotheker.nl/begrippenlijst/huis-kopen/marktwaarde/

De Hypotheker (2018) *Woningwaarde Berekenen.* Retrieved on August 6th 2018, from https://www.hypotheker.nl/zelf-berekenen/wat-is-de-woning-waard/

DeSimone, B. (2013) *Why Location Matters in Real Estate*. Retrieved on March 16th from www.foxbusiness.com

Donges, N. (2018) *The Random Forest Algorithm,* Retrieved on December 10th 2018, from https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Du, H. and Mulley, C. (2011) Understanding spatial variation in the impact of accessibility on land value using geographically weighted regression. Paper presented to World Symposium of Transport and Land Use Research, Whistler, British Columbia, Canada, 28–30 July.

Gaur, P. (2012) Neural Networks in Data Mining. In: *International Journal of Electronics and Computer Science Engineering,* 1 (1449-1453).

Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann, San FranciscoHaynes, B., Rymarzak, M. and Sieminska, E. (2012) Factors Affecting the Location of Real Estate. In: *Journal of Corporate Real Estate,* 14 (214-225).

Hill, R.J. and Melser, D. (2007) Comparing house prices across regions and time: an hedonic approach. In: *School of Economics Discussion Paper,* 33**.** University of New South Wales.

Hromada, E. (2015) Mapping of Real Estate Prices Using Data Mining Techniques. In: *Procedia Engineering,* 123 (233-240).

Hromada, E. (2016) Real Estate Valuation Using Data Mining Software. In: *Procedia Engineering,* 164 (284-291).

Francke, M.K. and Broekmeulen, B. (2016) OrtaX modelwaardesplitsing. Ortec Finance Research Center, Rotterdam.

Francke, M.K. and A.M. van de Minne (2017) Land, Structure and Depreciation. In: *Real Estate Economics,* 45 (415-451).

Jim, C. and Chen, W. (2006) Impacts of Urban Environmental Elements on Residential Housing Prices in Guangzhou (China). In: *Landscape and Urban Planning,* 78 (422-434).

JWB Real Estate Capital (2018) *What is Data Mining for Real Estate?* Retrieved on August 22th 2018, from https://www.jwbrealestatecapital.com/what-is-data-mining-for-real-estate/

Kadaster (2018) *WOZ-waarde.* Retrieved on August 11th 2018, from https://www.kadaster.nl/woz-waarde

Keskin, B. (2008) Hedonic Analysis of Price in the Istanbul Housing Market. In: *International Journal of Strategic Property Management,* 12 (125-138).

Kramer, L. (2018) *How Does the Law of Supply and Demand Affect Prices?* Retrieved on August 6th 2018, from https://www.economicmodels/how-does-law-supply-and-demand-affect-prices.asp

Kruk, J. van der and Pellenbarg, M. (2012) *Maatschappelijk Vastgoed.* Dataland and Kadaster, Maastricht

Kuminoff, N. V., Parmeter, C.F. and Pope, J.C. (2010) Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities? In: *Journal of Environmental Economics and Management,* 60 (145-160).

Market Business News (MBN) (2018) What is Economic Value. Retrieved on August 17th 2018, from https://marketbusinessnews.com/financial-glossary/economic-value/

Ministerie van Binnenlandse Zaken en Konkrijksrelaties (2017) *Wat is de Leefbaarometer?* Retrieved on December 17th 2017, from https://www.leefbaarometer.nl/page/leefbaarometer

Mok, H., Chan, P. and Cho, Y. (2005) A Hedonic Price Model For Private Properties in Hong Kong. In: *The Journal of real Estate Finance and Economics,* 10 (37-48).

Morancho, A. (2003) A Hedonic valuation of urban green areas. In: *Landscape and Urban Planning,* 66 (35-41).

Nowak, A. and Smith, P. (2017) Textual Analysis in Real Estate. In: *Journal of Apploed Econometrics,* 32 (896-918).

Pandya, J. and Patel, V. (2012) Loaction Factors Affecting Project Choice and Timing in Real Estate Project Decisions: A Case of Ahmedabad, India. In: *Real Estate Finance*, 34 (52-63).

Philip, S. (2017) 8 Concrete Data Mining Techniques That Will Deliver the Best Results. Retrieved on August 21th 2018, from https://www.datanami.com/2017/06/14/8-concrete-data-mining-techniques-will-deliver-best-results/

Ranzato, A. (2013) *Huizenprijs bepaald door veel factoren.* Retrieved on March 16th 2018, from www.mistermoney.nl/huizenprijs-bepaald-door-veel-factoren.asp

Research Gate (2018) *Research Interpretations.* Retrieved on December 10th 2018, from https://www.researchgate.net/

Robbesom, D. (2016) Plato (ca. 427-347 v. Chr) – Griekse Filosoof, Bedenker van Ideeënleer. Retrieved on August 17th 2018, from https://historiek/plato-griekse-filosoof/64694/

Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. In: *Journal of Political Economy,* 82 (34-55).

Rotterdam Centre (2018) *Local taxes in Rotterdam.* Retrived on June 28th 2018, from https://rotterdamexpatcentre.nl/expats/taxes/local-taxes-in-rotterdam/

Schram , J.F. (2006) *Real Estate Appraisal.* Rockwell Publishing, Bellevue.

Sprundel, W.P.A. van (2014) Importnace of attributes in the valuation of owner-occupied housing in the Netherlands capturing the determanants of market value. Master thesis, Eindhoven University of Technology, Eindhoven.

Stanghellini, S., Morano, P., Bottero, M. and Oppio, A. (2015) *Appraisal: From Theory to Practice.* Springer, Berlin.

The Balance (2017) *Location, location, location in Real Estate.* Retrieved on January 28th 2017, from www.thebalance.com

Vereniging Eigen Huis (2018) *Waarde Bepalen*. Retrieved on August 13th 2018, from https://www.eigenhuis.nl/huis-kopen/bestaande-bouw/onderhandelen/waarde-huis-bepalen

Visser, P. and Dam, F. van (2006) Ruimtelijk Planbureau; De Prijs van de Plek. Retrieved on February 12th 2017, from www.ruimtelijk planbureau.nl

Watada, J. (2012) *Intelligent Decision Technologies: Proceedings of the 4th International Conference on Intelligent Decision Technologies.* Conferention took place by the 4th Symposium on Intelligent Decision Technologies, Gifu-shi.

Wen, H., Sheng-Hua, J. and Xiao-Yu, G. (2005) Hedonic Price Analysis of Urban Housing: An Empirical research on Hangzhou, China. In: *Journal of Zhejiang University-Science A: Applied Physics & Engineering,* 6 (907-914).

White, R. (2015) Alaska Water and Sewer Challenge. In: *Alaska Business Monthly,* 31 (41-43).

Xiao, Y. (2016) *Urban Morphology and Housing Market.* Springer, Tungji University Press.

Yanchang, Z. and Yonghua, C. (2013) *Data Mining Applications with R.* Academic Press, Boston.

Zoppi, C., Argiolas, M. and Lai, S. (2014) Factors influencing the value of houses: Estimates for the city of Cagliari, Italy. In: *Land Use Policy,* 42 (367-380).

8
APPENDIX

Table 3.1 Variables about the Sale of the Dwelling

Variable Name	Variable Description
Estimated Value	The value of a dwelling estimated by a real estate agent before the sale
Transaction price	The selling price of a dwelling
Transaction date	The date of the transaction year-month-day
Transaction year	The year of the transaction
Kind of Market info	Whether the transaction price is based on the real transaction price or the asking price
Reason deviation	The reason of the difference between the estimated value and the transaction price

Table 3.2 Variables about Characteristics of the Dwelling Itself

Variable Name	Variable Description
Village	Village where dwelling is located
Postal code	Postal code
Lot size	The size of the lot in m^2
Use code	Code that defines where the dwelling is used for
Type of Object	Code that defines the type of dwelling
Building year	The year in which the dwelling is built
Volume	Volume of the dwelling in m^3
Surface dwelling	The surface of a lot that is covered by the dwelling
Usable surface	The total usable surface of the dwelling
Amount dwellings	The amount of dwellings
Residential layers	The amount of residential layers
Amount rooms	The amount of rooms
Amount bathrooms	The amount of bathrooms
Amount bedrooms	The amount of bedrooms
Surface bay window	The total surface of bay windows
Amount bay window	The amount of bay windows
Surface dormer	The total surface of dormers
Amount dormer	The amount of dormers
Surface balcony	The total surface of balconies
Amount balcony	The amount of balconies
Surface basement	The total surface of basements
Amount basement	The amount of basements
Surface garage	The total surface of garages
Amount garage	The amount of garages
Surface barn	The total surface of barns
Amount barn	The amount of barns
Surface garden shed	The total surface of garden sheds
Amount garden shed	The amount of garden sheds
Surface garden apartment building	The total surface of gardens of an apartment building
Amount garden apartment building	The amount of gardens of an apartment building
Quality luxury	Expresses the quality of the dwelling
Maintenance condition	Expresses the maintenance condition of the dwelling
Appearance	Expresses the appearance of the dwelling
Expediency	Does the usability match with the requirements

Table 3.3 Environmental Variables

Variable Name	Variable Description
Neighbourhood code	Official neighbourhood code
Amount inhabitants	The amount of inhabitants
Amount inhabitants 0-14	The amount of inhabitants between 0 and 14 years old
Amount inhabitants 15-24	The amount of inhabitants between 15 and 24 years old
Amount inhabitants 25-44	The amount of inhabitants between 25 and 44 years old
Amount inhabitants 45-64	The amount of inhabitants between 45 and 64 years old
Amount inhabitants 65+	The amount of inhabitants 65 years old and older
Amount Western	The amount of Western inhabitants
Amount non Western	The amount of non-Western inhabitants
Amount household	The amount of households
Amount household 1p	The amount of single person households
Amount household without children	The amount of households without children
Amount household with children	The amount of households with children
Average household size	The average household size
Amount dwellings	The amount of dwellings
Percentage single family house	The percentage single houses
Percentage apartment building	The percentage apartment building
Percentage inhabited	The percentage inhabited dwellings
Percentage vacant	The percentage empty dwellings
Percentage privately owned	The percentage dwellings privately owned
Percentage rental	The percentage rental dwellings
Percentage housing association	The percentage dwellings owned by housing associations
Percentage remaining rental	The percentage remaining rental dwellings
Percentage ownership unknown	The percentage dwellings with unknown ownership
Percentage built before 2000	The percentage dwellings built before 2000
Percentage built since 2000	The percentage dwellings built since 2000
Average house value	The average value of a house
Amount income receivers	The amount of inhabitants that receive income
Percentage income low 40%	The percentage inhabitants with an income belonging to the lowest 40% of The Netherlands
Percentage income high 20%	The percentage inhabitants with an income belonging to the highest 20% of The Netherlands
Percentage inhabitants income	The percentage inhabitants between 15 and 75 with income from labour
Percentage household income low	The percentage households with an income belonging to the lowest 40% of The Netherlands
Percentage household income high	The percentage households with an income belonging to the highest 20% of The Netherlands
Percentage households income low	The percentage households with a low income
Percentage households social minimum	The percentage households at or below the social minimum
Average income per income receiver	The average income per income receiver
Average income inhabitant	The average income per inhabitant
Population density	The density of the population
Address density	The density of addresses
Liveability	The liveability presented on a scale level
Total surface	The total surface
Park and green area surface	The surface of park and plantation
Day recreation surface	The surface of day recreation
Residence recreation surface	The surface of residence recreation
Forest, open and natural terrain surface	The surface of forest, open and natural terrain
Inland water surface	The surface of inland water
Sport terrain surface	The surface of sport terrain
Shortage indicator	The shortage indicator, a ratio of supply and transactions

Table 3.4 Location Variables

Variable Name	Variable Description
Distance doctor	The distance to the nearest doctor
Doctor 1 km	The amount of doctors within a radius of 1 km
Doctor 3 km	The amount of doctors within a radius of 3 km
Doctor 5 km	The amount of doctors within a radius of 5 km
Distance Pharmacy	The distance to the nearest pharmacy
Distance hospital inclusive	The distance to the nearest hospital inclusive external departments
Hospital inclusive 5 km	The amount of hospitals incl. external dep. within a radius of 5 km
Hospital inclusive 10 km	The amount of hospitals incl. external dep. within a radius of 10 km
Hospital inclusive 20 km	The amount of hospitals incl. external dep. within a radius of 20 km
Distance hospital exclusive	The distance to the nearest hospital exclusive external departments
Hospital exclusive 5 km	The amount of hospitals excl. external dep. within a radius of 5 km
Hospital exclusive 10 km	The amount of hospitals excl. external dep. within a radius of 10 km
Hospital exclusive 20 km	The amount of hospitals excl. external dep. within a radius of 20 km
Distance supermarket	The distance to the nearest supermarket
Supermarket 1 km	The amount of supermarkets within a radius of 1 km
Supermarket 3 km	The amount of supermarkets within a radius of 3 km
Supermarket 5 km	The amount of supermarkets within a radius of 5 km
Distance other daily shops	The distance to the nearest other daily shop
Other daily shops 1 km	The amount of other daily shops within a radius of 1 km
Other daily shops 3 km	The amount of other daily shops within a radius of 3 km
Other daily shops 5 km	The amount of other daily shops within a radius of 5 km
Distance department store	The distance to the nearest department store
Department store 5 km	The amount of department stores within a radius of 5 km
Department store 10 km	The amount of department stores within a radius of 10 km
Department store 20 km	The amount of department stores within a radius of 20 km
Distance pub	The distance to the nearest pub
Pub 1 km	The amount of pubs within a radius of 1 km
Pub 3 km	The amount of pubs within a radius of 3 km
Pub 5 km	The amount of pubs within a radius of 5 km
Distance cafeteria	The distance to the nearest cafeteria
Cafeteria 1 km	The amount of cafeteria within a radius of 1 km
Cafeteria 3 km	The amount of cafeteria within a radius of 3 km
Cafeteria 5 km	The amount of cafeteria within a radius of 5 km
Distance restaurant	The distance to the nearest restaurant
Restaurant 1 km	The amount of restaurants within a radius of 1 km
Restaurant 3 km	The amount of restaurants within a radius of 3 km
Restaurant 5 km	The amount of restaurants within a radius of 5 km
Distance hotel	The distance to the nearest hotel
Hotel 5 km	The amount of hotels within a radius of 5 km
Hotel 10 km	The amount of hotels within a radius of 10 km
Hotel 20 km	The amount of hotels within a radius of 20 km
Distance day care	The distance to the nearest day care
Day care 1 km	The amount of day cares within a radius of 1 km
Day care 3 km	The amount of day cares within a radius of 3 km
Day care 5 km	The amount of day cares within a radius of 5 km
Distance after school care	The distance to the nearest after school care
After school care 1 km	The amount of after school cares within a radius of 1 km
After school care 3 km	The amount of after school cares within a radius of 3 km
After school care 5 km	The amount of after school cares within a radius of 5 km
Distance elementary school	The distance to the nearest elementary school
Elementary school 1 km	The amount of elementary schools within a radius of 1 km
Elementary school 3 km	The amount of elementary schools within a radius of 3 km
Elementary school 5 km	The amount of elementary schools within a radius of 5 km
Distance high school	The distance to the nearest high school
High school 3 km	The amount of high schools within a radius of 3 km
High school 5 km	The amount of high schools within a radius of 5 km

Variable Name	Variable Description
High school 10 km	The amount of high schools within a radius of 10 km
Distance pre-vocational education	The distance to the nearest pre-vocational education
Pre-vocational education 3 km	The amount of pre-vocational education within a radius of 3 km
Pre-vocational education 5 km	The amount of pre-vocational education within a radius of 5 km
Pre-vocational education 10 km	The amount of pre-vocational education within a radius of 10 km
Distance pre-university education	The distance to the nearest pre-university education
pre-university education 3 km	The amount of pre-university education within a radius of 3 km
pre-university education 5 km	The amount of pre-university education within a radius of 5 km
pre-university education 10 km	The amount of pre-university education within a radius of 10 km
Distance library	The distance to the nearest library
Distance swimming pool	The distance to the nearest swimming pool
Distance artificial ice skating rink	The distance to the nearest ice skating rink
Distance cinema	The distance to the nearest cinema
Cinema 5 km	The amount of cinemas within a radius of 5 km
Cinema 10 km	The amount of cinemas within a radius of 10 km
Cinema 20 km	The amount of cinemas within a radius of 20 km
Distance sauna	The distance to the nearest sauna
Distance solarium	The distance to the nearest solarium
Distance theme park	The distance to the nearest theme park
Theme park 10 km	The amount of theme parks within a radius of 10 km
Theme park 20 km	The amount of theme parks within a radius of 20 km
Theme park 50 km	The amount of theme parks within a radius of 50 km
Distance firehouse	The distance to the nearest firehouse
Distance highway	The distance to the nearest highway
Distance railway station	The distance to the nearest railway station
Distance transfer railway station	The distance to the nearest step over railway station
Employment opportunities 10 km	The amount of employment opportunities within a radius of 10 km
Employment opportunities 20 km	The amount of employment opportunities within a radius of 20 km
Employment opportunities 50 km	The amount of employment opportunities within a radius of 50 km
Agriculture, forestry and fishery empl. opp. 10 km	The amount of agriculture, forestry and fishery employment opportunities within a radius of 10 km
Agriculture, forestry and fishery empl. opp. 20 km	The amount of agriculture, forestry and fishery employment opportunities within a radius of 20 km
Agriculture, forestry and fishery empl. opp. 50 km	The amount of agriculture, forestry and fishery employment opportunities within a radius of 50 km
Industry and energy empl. opp. 10 km	The amount of industry and energy employment opportunities within a radius of 10 km
Industry and energy empl. opp. 20 km	The amount of industry and energy employment opportunities within a radius of 20 km
Industry and energy empl. opp. 50 km	The amount of industry and energy employment opportunities within a radius of 50 km
Commercial services empl. opp. 10 km	The amount of commercial services employment opportunities within a radius of 10 km
Commercial services empl. opp. 20 km	The amount of commercial services employment opportunities within a radius of 20 km
Commercial services empl. opp. 50 km	The amount of commercial services employment opportunities within a radius of 50 km
Non-commercial services empl. opp. 10 km	The amount of non-commercial services employment opportunities within a radius of 10 km
Non-commercial services empl. opp. 10 km	The amount of non-commercial services employment opportunities within a radius of 10 km
Non-commercial services empl. opp. 10 km	The amount of non-commercial services employment opportunities within a radius of 50 km

Table 3.5 Index Table for Table 3.6

Variable Number	Variable Name
1	The density of addresses
2	The amount of high schools within a 3 km radius
3	The amount of high schools within a 5 km radius
4	The amount of pre-vocational education within a 3 km radius
5	The amount of pre-vocational education within a 5 km radius
6	The amount of pre-university education within a 3 km radius
7	The amount of pre-university education within a 5 km radius
8	The amount of doctors within a 3 km radius
9	The amount of doctors within a 5 km radius
10	The amount of pubs within a 3 km radius
11	The amount of pubs within a 5 km radius
12	The amount of cafeterias within a 3 km radius
13	The amount of cafeterias within a 5 km radius
14	The amount of restaurants within a 3 km radius
15	The amount of restaurants within a 5 km radius
16	The amount of day cares within a 3 km radius
17	The amount of day cares within a 3 km radius
18	The amount of after school cares within a 3 km radius
19	The amount of after school cares within a 5 km radius
20	The amount of elementary schools within a 3 km radius
21	The amount of elementary schools within a 5 km radius
22	The amount of supermarkets within a 3 km radius
23	The amount of supermarkets within a 5 km radius
24	The amount of other daily shops within a 3 km radius
25	The amount of other daily shops within a 5 km radius
26	Distance to nearest hospital inclusive external departments
27	Distance to nearest hospital exclusive external departments

Table 3.6 Correlations Variables Amount of Daily Facilities Within 3-5 km.

Correlations															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	.816**	.676**	.805**	.643**	.762**	.664**	.850**	.668**	.758**	.610**	.856**	.697**	.841**	.709**
2	.816**	1	.805**	.978**	.779**	.945**	.774**	.920**	.777**	.887**	.756**	.907**	.782**	.875**	.757**
3	.676**	.805**	1	.776**	.984**	.704**	.952**	.825**	.919**	.790**	.957**	.783**	.940**	.754**	.940**
4	.805**	.978**	.776**	1	.767**	.906**	.743**	.875**	.745**	.864**	.733**	.881**	.760**	.832**	.715**
5	.643**	.779**	.984**	.767**	1	.678**	.934**	.778**	.875**	.755**	.940**	.735**	.908**	.685**	.894**
6	.762**	.945**	.704**	.906**	.678**	1	.753**	.838**	.635**	.857**	.695**	.806**	.631**	.823**	.672**
7	.664**	.774**	.952**	.743**	.934**	.753**	1	.778**	.828**	.818**	.954**	.735**	.840**	.754**	.908**
8	.850**	.920**	.825**	.875**	.778**	.838**	.778**	1	.865**	.876**	.762**	.948**	.843**	.919**	.822**
9	.668**	.777**	.919**	.745**	.875**	.635**	.828**	.865**	1	.744**	.866**	.814**	.962**	.775**	.920**
10	.758**	.887**	.790**	.864**	.755**	.857**	.818**	.876**	.744**	1	.811**	.896**	.731**	.917**	.774**
11	.610**	.756**	.957**	.733**	.940**	.695**	.954**	.762**	.866**	.811**	1	.713**	.877**	.709**	.914**
12	.856**	.907**	.783**	.881**	.735**	.806**	.735**	.948**	.814**	.896**	.713**	1	.832**	.954**	.792**
13	.697**	.782**	.940**	.760**	.908**	.631**	.840**	.843**	.962**	.731**	.877**	.832**	1	.762**	.933**
14	.841**	.875**	.754**	.832**	.685**	.823**	.754**	.919**	.775**	.917**	.709**	.954**	.762**	1	.818**
15	.709**	.757**	.940**	.715**	.894**	.672**	.908**	.822**	.920**	.774**	.914**	.792**	.933**	.818**	1
16	.811**	.891**	.863**	.890**	.861**	.786**	.809**	.870**	.782**	.844**	.819**	.864**	.835**	.791**	.779**
17	.649**	.786**	.948**	.785**	.952**	.657**	.861**	.786**	.892**	.750**	.912**	.765**	.929**	.680**	.857**
18	.867**	.819**	.783**	.797**	.738**	.691**	.715**	.874**	.779**	.733**	.677**	.886**	.827**	.852**	.799**
19	.714**	.730**	.865**	.691**	.813**	.602**	.783**	.825**	.931**	.665**	.765**	.806**	.934**	.792**	.919**
20	.889**	.843**	.741**	.828**	.701**	.766**	.725**	.907**	.768**	.780**	.678**	.862**	.766**	.861**	.774**
21	.706**	.740**	.866**	.705**	.818**	.636**	.820**	.842**	.943**	.708**	.802**	.798**	.916**	.790**	.914**
22	.894**	.864**	.723**	.827**	.668**	.839**	.747**	.911**	.735**	.850**	.681**	.893**	.720**	.929**	.773**
23	.714**	.758**	.876**	.719**	.827**	.698**	.863**	.817**	.904**	.758**	.840**	.789**	.889**	.825**	.950**
24	.869**	.918**	.803**	.889**	.756**	.855**	.795**	.956**	.814**	.938**	.764**	.972**	.811**	.967**	.820**
25	.702**	.781**	.965**	.755**	.935**	.683**	.924**	.830**	.946**	.778**	.938**	.799**	.963**	.783**	.974**
26	-.606**	-.505**	-.610**	-.485**	-.578**	-.411**	-.556**	-.585**	-.647**	-.475**	-.517**	-.613**	-.681**	-.616**	-.696**
27	-.586**	-.556**	-.725**	-.511**	-.672**	-.479**	-.701**	-.698**	-.796**	-.620**	-.704**	-.643**	-.752**	-.680**	-.814**

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3.6 Correlations Variables Amount of Daily Facilities Within 3-5 km.

Correlations												
	16	17	18	19	20	21	22	23	24	25	26	27
1	.811**	.649**	.867**	.714**	.889**	.706**	.894**	.714**	.869**	.702**	-.606**	-.586**
2	.891**	.786**	.819**	.730**	.843**	.740**	.864**	.758**	.918**	.781**	-.505**	-.556**
3	.863**	.948**	.783**	.865**	.741**	.866**	.723**	.876**	.803**	.965**	-.610**	-.725**
4	.890**	.785**	.797**	.691**	.828**	.705**	.827**	.719**	.889**	.755**	-.485**	-.511**
5	.861**	.952**	.738**	.813**	.701**	.818**	.668**	.827**	.756**	.935**	-.578**	-.672**
6	.786**	.657**	.691**	.602**	.766**	.636**	.839**	.698**	.855**	.683**	-.411**	-.479**
7	.809**	.861**	.715**	.783**	.725**	.820**	.747**	.863**	.795**	.924**	-.556**	-.701**
8	.870**	.786**	.874**	.825**	.907**	.842**	.911**	.817**	.956**	.830**	-.585**	-.698**
9	.782**	.892**	.779**	.931**	.768**	.943**	.735**	.904**	.814**	.946**	-.647**	-.796**
10	.844**	.750**	.733**	.665**	.780**	.708**	.850**	.758**	.938**	.778**	-.475**	-.620**
11	.819**	.912**	.677**	.765**	.678**	.802**	.681**	.840**	.764**	.938**	-.517**	-.704**
12	.864**	.765**	.886**	.806**	.862**	.798**	.893**	.789**	.972**	.799**	-.613**	-.643**
13	.835**	.929**	.827**	.934**	.766**	.916**	.720**	.889**	.811**	.963**	-.681**	-.752**
14	.791**	.680**	.852**	.792**	.861**	.790**	.929**	.825**	.967**	.783**	-.616**	-.680**
15	.779**	.857**	.799**	.919**	.774**	.914**	.773**	.950**	.820**	.974**	-.696**	-.814**
16	1	.899**	.876**	.738**	.837**	.732**	.799**	.730**	.875**	.830**	-.554**	-.592**
17	.899**	1	.749**	.823**	.697**	.822**	.655**	.803**	.768**	.926**	-.595**	-.676**
18	.876**	.749**	1	.847**	.895**	.789**	.853**	.771**	.862**	.795**	-.703**	-.648**
19	.738**	.823**	.847**	1	.795**	.959**	.764**	.930**	.798**	.922**	-.797**	-.810**
20	.837**	.697**	.895**	.795**	1	.834**	.939**	.805**	.914**	.789**	-.618**	-.698**
21	.732**	.822**	.789**	.959**	.834**	1	.797**	.951**	.828**	.940**	-.714**	-.848**
22	.799**	.655**	.853**	.764**	.939**	.797**	1	.830**	.948**	.767**	-.638**	-.716**
23	.730**	.803**	.771**	.930**	.805**	.951**	.830**	1	.836**	.948**	-.750**	-.861**
24	.875**	.768**	.862**	.798**	.914**	.828**	.948**	.836**	1	.829**	-.601**	-.701**
25	.830**	.926**	.795**	.922**	.789**	.940**	.767**	.948**	.829**	1	-.678**	-.815**
26	-.554**	-.595**	-.703**	-.797**	-.618**	-.714**	-.638**	-.750**	-.601**	-.678**	1	.780**
27	-.592**	-.676**	-.648**	-.810**	-.698**	-.848**	-.716**	-.861**	-.701**	-.815**	.780**	1

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4.1 Independent Variables for Regression Analysis

Independent Variables	
Characteristics of Dwelling Itself:	*Location Factors:*
Quality luxury	Hospital inclusive 5 km
Maintenance condition	Hospital inclusive 10 km
Appearance	Hospital inclusive 20 km
Volume	Hospital exclusive 5 km
Residential layers	Hospital exclusive 10 km
Amount rooms	Hospital exclusive 20 km
Amount bathrooms	Distance pub
Amount bedrooms	Distance restaurant
Standalone house	Hotel 5 km
Semi-detached house	Distance library
Row house	Distance swimming pool
Corner house	Distance sauna
Apartment	Distance theme park
Size garden	Distance firehouse
Bay window	Distance highway
Dormer	Distance railway station
Balcony	Amount of facilities within 10 km
Basement	Amount of facilities within 20-50 km
Garage	Amount of facilities within 3-5 km 1
Barn	Amount of facilities within 3-5 km 2
Garden shed	Amount of facilities within 3-5 km 3
Garden apartment building	Distance to leisure and high schools
Built before 1900	Cinemas and transfer railway stations
Built between 1900-1929	
Built between 1930-1949	*Circumstances of Transaction:*
Built between 1950-1959	Transaction year 2015
Built between 1960-1969	Transaction year 2016
Built between 1970-1979	Transaction year 2017
Built between 1980-1989	
Built between 1990-1999	*Shortage:*
Built between 2000-2009	Shortage indicator
Built after 2009	
Environmental Factors:	
Percentage sport terrain	
Percentage green	
Percentage recreation terrain	
Neighbourhood type 1	
Neighbourhood type 2	
Neighbourhood type 3	
Neighbourhood type 4	
Neighbourhood type 5	
Neighbourhood type 6	
Distance to daily facilities	
Distance to school facilities children 0-11	