GIMA
Geographical Information Management and Applications

Influence of population demographics on real estate prices in Zuid-Holland

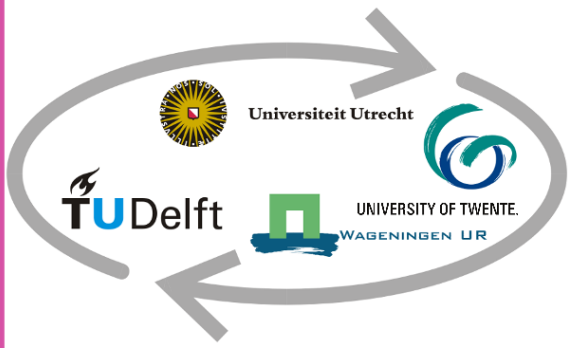
MSc Thesis

Vic Bensdorp

v.r.j.bensdorp@students.uu.nl

Supervisor: Dr. Egbert van der Zee

Responsible professor: Prof. dr. Stan Geertman (UU)



Abstract

Because the Dutch population is aging, a change in housing preferences can be expected. According to the lifecycle theory older people have accumulated more wealth, and therefore should be able to spend more on their housing. This thesis aims to model the real estate market between 2009 and 2016 in Zuid-Holland, comparing the results of linear regression to random forest regression while trying to incorporate the development of local population change in people over 65. Data from the Dutch national real estate broker association (NVM) is being used, enriched with publicly available neighborhood statistics. Analyses have been performed for each year, for both models. Results show the relations of structural, locational, and neighborhood variables on the recorded transaction price per square meter. From these results, conclusions have been drawn on the effectiveness of the linear regression in relation to the forest regression, as well as the performance and impact of the inclusion of age dynamics into hedonic price modeling. It was shown that the inclusion of age dynamics added a very slight value to the adjusted R-squared, however this variable was not consistently significant. Furthermore, the random forest regression has shown to consistently outperform the linear regression.

Contents

Abstract 2
Introduction 6
Economic impact of the housing market 6
Modelling of the housing market 6
Age structures and the housing market 7
Demography in relation to the housing market 7
Theoretical framework 9
Modelling real estate prices 9
What affects real estate prices 9
Structural variables 10
Locational variables 10
Neighborhood variables 11
Methodology 14
Detailed location description: Zuid-Holland 14
Data sources 15
Brainbay listing data 15
Additional data by CBS 15
RIVM noise pollution 16
Data preparation 16
Brainbay data 16
Creating dummy variables 17
Joining data sources 17
Adding temporal population dynamics 18
Variables included 19
Linear regression 21
Hedonic price models: Linear regression 21
Random forest regression 24
Hedonic price models: Random forest classification and regression 24
Comparing of model outcomes 25
Model fit 25
Variable influence 26
Results 27
Model accuracy compared 27

Research question 1 30
Structural variables 30
Locational variables 32
Neighborhood variables 33
Research question 2 35
Research question 3 36
Discussion 39
Research question 1 39
Linear regression models 41
Research question 2 43
Outcomes in context 43
Age development variable 43
Research question 3 44
Outcomes in context 44
Conclusions 46
Research questions 46
Implications 47
References 48
Appendices 52
Appendix A: description of NVM data records as delivered 52
Appendix B: Descriptive statistics on variables of interest 55
Appendix C: Description of joined data sources for yearly data 61

List of figures

Figure 1 conceptual model of the housing market 13
Figure 2 population development of Zuid-Holland over the past 15 years 15
Figure 3 map of the study area, including all (158 852) geocoded listings before cleaning 17
Figure 4 schematic overview of the data preparation steps taken 18
Figure 5 example of multiple different decision trees used alongside each other. Source: Edureka (2021) 25
Figure 6 comparison of a linear model estimation versus a classified decision tree estimation. Source: Deepnote (2021) 25
Figure 7 R-squared values for all different models through the years 27
Figure 8 Standardized Residual vs Predicted values Plot for the linear regression of 2009 29
Figure 9 map of the distribution of standardized residuals for the linear regression of 2009 29
Figure 10 coefficient and significance of floor area over time 30
Figure 11 coefficient and significance of number of rooms over time 30
Figure 12 coefficient and significance of availability of a garage over time 31
Figure 13 coefficient and significance of Detached and semi-detached properties over time (compared to terraced houses) 31
Figure 14 coefficients and significance of properties period over time (compared to properties built between 1970 and 1990) 32
Figure 15 coefficients and lowest significance of properties classified with a type of special view over time (compared to properties without) 32
Figure 16 coefficient and significance of a property's proximity to a supermarket 33
Figure 17 coefficient and significance of income level within the neighborhood 33
Figure 18 coefficient and significance of noise within the vicinity 34
Figure 19 coefficient and significance of the number of primary schools within 3 kilometers 34
Figure 20 coefficient and significance of crime rate within the neighborhood 34
Figure 21 R-squared values for both created models, including the difference between them 35
Figure 22 yearly R-squared values for the complete linear and forest regression models 36
Figure 23 relative influence of each variable per model type for 2009, ranked 0 (lowest influence) to 19 (highest influence). 37
Figure 24 standardized residuals plotted against the predicted values for the year 2015 40
Figure 25 map of the distribution of standardized residuals for the linear regression of 2009 (same as Figure 9) 42
Figure 26 spatial distribution of residuals of the linear regression model (left) and the random forest regression model (right) 45

List of tables

Table 1 all variables to be included based on the literature 13
Table 2 structural variables to be included 19
Table 3 Locational variables 20
Table 4 neighborhood variables 21
Table 5 variables used to explain the transaction price per square meter 23
Table 6 output of the linear regression model of 2009 28
Table 7 continuation of Linear regression output with diagnostics 28
Table 8 influence scores in both models per variable 37
Table 9 linear regression model diagnostics for 2009 without age (same as Table 7) 41
Table 10 point counts for the spread of residual values shown in Figure 26 45

Introduction

Economic impact of the housing market

Housing is a basic need for families, and the construction and regulation of housing is frequently a topic of political discussions with high societal relevance. The supply and demand of property can be used as a key indicator for the state of the national economy as demand for housing triggers activity through several other sectors such as construction, transport and related services (Chin & Chau, 2003). In addition, investing in real estate is regarded as a stable and relatively safe investment as well as a way to spread the risks of any large sum of money when available (Sulamoyo, 2016). Because of these reasons, the housing market can and has been studied extensively before.

Although many figures and theories can be used to draw general conclusions on a national scale, the housing market does not behave as a uniformly spread phenomenon. On a regional scale, the conclusion has been drawn that the west of the Netherlands is urbanizing whereas the north and more rural regions are in decline. One of the causes of the growth in the west and decline in the north has been identified as the aging population combined with younger people leaving for the cities. This is reflected on real estate prices in these respective regions (Mulder, 2016). Current developments within the Dutch housing market cause rising prices, and people having difficulty to find a suitable property (van Geuns, 2020). This raises the question what makes up the price of a house and how does it develop over time?

Modelling of the housing market

To gain understanding of previous events as well as plan for future policy and building projects, the housing market is constantly being observed. The goal of close observation is the development of accurate models that will aid in decision making and planning processes. The technique frequently used for this research is called hedonic price modeling. The discipline of hedonic price modeling is built on a strong foundation of theoretical and empirical observations. The assumption that (housing) prices are built up of individual value-adding components is the central idea within the existing body of knowledge (Bisello, Antoniucci, & Marella, 2020; Chin & Chau, 2003; Evangelista, Ramalho, & Andrade e Silva, 2020; Sopranzetti, 2010). Previous research has shown that wen valuing a property, the explaining variables can be divided into three categories: structural variables, the physical surroundings of the location, and the demographic surroundings of the location. (Chin & Chau, 2003; Helbich, Brunauer, Vaz, & Nijkamp, 2014). The existing literature allows for several conclusions to be drawn (Chin & Chau, 2003; Li et al, 2017; Lytvynchenko, 2014; Yao & Stewart Fotheringham, 2016):

- Most studies only consider either the spatial or the temporal aspect of varying house prices, only a few studies look at both (Li et al., 2017; Yao & Stewart Fotheringham, 2016).
- Recent studies attempt to model the relationships of variables more accurately by theorizing that not all relationships can be considered linear. The most commonly used non-linear method is the random forest approach using a non-linear fitting of the model (Li, 2019; Yoo, Im., & Wagner, 2012).
- Almost all studies use a set of “standard” explanatory variables, in addition to their unique variables related to their research topic. The selection of variables in most studies is based on previous research and empirical findings rather than a theoretical foundation.

This study aims to address these research gaps, and in addition expand upon the current body of knowledge by exploring the influence of local demographic structures on the housing market by comparing two different methods of modelling. To expand upon the current body of knowledge local demographics will be explored next to the “standard” variables. Local age structures are not (widely) included in these analyses. However, the change of age structures might have an influence on the supply and demand, causing fluctuations in prices.

Age structures and the housing market

Expanding upon the standard collection of variables, this study will investigate the influence of local age structures on transaction prices. Current research has focused little on the interactions between different generations and their behavior on the real estate market. Due to the combination of higher life expectancy and the retirement of the "baby boom" generation, born shortly after the second world war, the population in the Netherlands is experiencing a rise in people over the age of 65, and a decline in people aged between 20 to 65. Between 2020 and 2040 the number of people between 20 and 65 is expected to decline by about 7.5% whereas the population above 65 years old is expected to increase by 41.5% (CBS, 2020).

A shift in the age pyramid of a region can have far-reaching consequences on several socioeconomic aspects of society (An & Jeon, 2006; Vishnevsky & Shcherbakova, 2018). One of these areas of influence is the real estate market. Financial freedom as well as changing living standards of a generation can be of influence on the properties they acquire. The effects of house characteristics on the age groups of potential buyers have been observed before (Hennadige, 2016; Silva, 2017). This indicates that different features are attractive to different demographic groups.

Demography in relation to the housing market

The effects of shifting demographics on the real estate market can be placed within the framework of the life cycle theory. According to the life cycle theory (Ando & Modigliani, 1963), households will save during their working years and spend their accumulated wealth during their retirement years. More specifically this is broken down into the age groups of 20-39 and 40-64. The former being considered as borrowers, people that acquire a property with a loan or mortgage. The second age group is viewed as actively accumulating wealth and assets without external support (Hennadige, 2016). The remaining group older than 65 is expected to dissipate wealth. When applied to the real estate market, the lifecycle theory would mean that when the working population is large, so is the demand for their most favored or best obtainable property. As a result, the prices of this type of real estate should increase. Alternatively, when the population is aging, the supply of (former) high-demand types of real estate increases, and logically the price should decline.

When combining the effects on supply and demand by the life cycle theory, with care for non-working populations through taxation and social security mechanisms, a higher age dependency ratio should leave the working population with less financial capabilities and can raise the question of "who is going to buy the boomer assets?" (Hennadige, 2016). Similarly, the market could adjust itself down due to people dying or moving out of their current residencies, raising supply. In a more general way, the effects of a changing age structure on the real estate market require closer studying.

This thesis aims to explore the effect of changing age structures on the local real estate valuation within the province of Zuid-Holland. From a theoretical perspective, an aging population is likely to impact housing prices in various ways, and this effect is expected to work out differently in different locations. This study therefore aims to investigate the effect of aging populations on housing prices and to explore different methods to model spatial relationships more accurately (Fory, 2014; Silva, 2017; Zhang, Jin, Xiao, & Gao, 2020). The three identified research gaps have been transformed into the formulated research objective.

In order to best examine the interactions between population demographics and real estate prices, the following research question has been formulated:

"To what extent can the changing local age structure, next to structural and locational characteristics, be used to explain the evolution of local real estate prices, and to what extent does the conventional method suit spatial relationships"

To answer this question three sub-questions have been formulated that will each be answered in order to reach a conclusion:

- How do structural and locational characteristics influence the price of real estate and does this change over time?

- To what extent does including developing local age structures lead to an improvement of the estimation of real estate pricing?
- To what extent does the inclusion of non-linear variable fitting enhance the analysis result?

In the following chapter, the existing theory and previous studies will be covered, as well as their implications. This is followed by the methodology, explaining the conducted analysis as well as the data sources and variable selection. The results will be presented before drawing conclusions in relation to the research questions. Finally, this research will be discussed and placed in context of the existing literature that it is built upon. Detailed information on the data is included in the appendices.

Theoretical framework

There are few other ways to valuate real estate other than its monetary value. In addition to its importance as an indicator, the need for regulated external agencies such as a realtor and a notary make the real estate market one that is relatively easy to monitor for governing bodies. Next to the economic activity that is sparked by providing people with a house, housing is a basic need for the buyers of a property. In accordance with a buyers budget and stage of life different demands are to be met to find the right property (Hennadige, 2016; Zhang et al., 2020). To maintain price levels, the supply and demand of properties in general as well as specific features of properties should be roughly equal. To monitor both these aspects individual properties can be divided into several variables aiming to create the most accurate depiction of a property and derive its current monetary value. This chapter will go over the existing research and theories behind the modelling of real estate prices. First common modelling techniques will be discussed, before going through different types of variables and deciding upon the variables that will be included in this research.

Modelling real estate prices

In search of the model that fits reality most accurately different approaches have been tested before. To place a different emphasis on factors that might influence the transaction price, the data can be approximated by several mathematical functions. The most basic of these is a linear fitting, assuming that every addition of variable X adds an equal amount of value. Especially when taking into account geographical influences other methods might be preferred. One of these is the geographically weighted regression, which decreases the 'importance' of observations further removed from the studied location. Spatial error and spatial lag models are expansions of this that determine a maximum number of neighbors to consider for each study location. By limiting the modelling of an observation point to a certain number of neighbors, a separate local variation of the general model is created. The general model is called a global model, that then encompasses several local model variations (Osland, 2010; von Graevenitz & Panduro, 2015). With increased computational power and availability of data, a method was developed to approximate linear trends within small parts of the same variable, allowing for the modelling of price functions that cannot necessarily be defined by an mathematical function. This technique is called the random forest method and will be explained in greater detail in the methodology chapter.

The following paragraphs will look into the most relevant variables to be examined in this study.

What affects real estate prices

Kohlhase (1991) has previously concluded that the influence of variables can vary and change between nations and time periods. It is therefore that the selection of these variables has been heavily based on the practices of several recent studies conducted in the Netherlands, or that had a very similar methodology (Steegmans & Hassink, 2017; Visser et al., 2008; Yao et al, 2016). The literature is consistent in dividing variables into three main categories: structural, locational and neighborhood variables (Chin & Chau, 2003; Do, Wilbur, & Short, 1994; Yoo et al., 2012). This assumes that the state of any property is affected by its own characteristics, its location relative to other locations, and the state of its surroundings. Therefore, an accurate depiction of a property will include variables from each of these categories. The following paragraphs will further elaborate on these categories, and the variables they include. The value to be modelled is nearly always the transaction price. Also existent within the literature are the use of average neighborhood price (Yao & Stewart, 2016), the use of the logged total transaction price (Clark & Herrin, 2000; Steegmans & Hassink, 2017), or the price per square meter (Visser et al., 2008). To be able to compare differently sized properties, the transaction price per square meter will be used as dependent variable, which is to be modelled.

Structural variables

Structural variables make up any physical features of a property, both interior as well as exterior. Structural variables are not subject to dynamic changes without the interference of an owner. A property cannot spontaneously develop an additional room, in a way that for example crime rates can emerge and decline. Either can be influenced by the owner of a property, however without active efforts, a neighborhood can change on its own whereas the physical features of a property will remain equal. It was observed by Randeniya et al. (2017) that people primarily value the structural characteristics in a property, rather than the neighborhood or locational features. A possible explanation for this is the ability to alter certain structural features as opposed to the inability to change the neighborhood statistics.

When looking at different studies within the area of hedonic price modeling the structural variables can roughly be divided into two categories, the first being space-related variables. Space-related structural variables describe the functional area that is being offered. Variables include total floor area, number of rooms, and the type of property as well as parking spaces and garden specifics (Randeniya et al., 2017). There are many types of properties, and in accordance with the literature a distinction will be made only between corner house, semidetached and detached properties (Steegmans & Hassink, 2017; Visser et al., 2008). Other structural variables describe the existing facilities within the structure (Chin & Chau, 2003). These facilities can provide a range of information on the state or modernness of the property by describing the number of showers or bathtubs, specifics on the isolation or heating systems in place. In recent years a growing number of studies feature variables related to various aspects of sustainability: Energy efficiency (Bisello et al., 2020), energy certification (Bottero et al., 2019) and finally sustainable construction (Lorenz, Lützkendorf, & Trück, 2007).

Present in all studies is the property floor area. The general consensus is that people value more space. This is also reflected in the variables measuring the number of bed- and bathrooms (Chin & Chau, 2003; Visser et al, 2008; Yusof & Ismail, 2012). Theories behind this are that either large families create a higher demand for more spacious properties, or that richer people are willing to spend more on a bigger place (Garrod & Willis, 1992).

One variable that appears to be negatively correlated to housing prices is building age. As a building ages its design and construction become outdated. This leads to higher maintenance costs as well as decreased practicalities in room arrangements. On the other hand, studies have found that buildings with a special historical status or a vintage status can attract people and increase its value (Chin & Chau, 2003; Journal, Statistical, & Jun, 2009). Related to age and design is the architect. Chau, et al. (2001) note that buildings designed by renowned architects do not lose their value over time. Most studies consider the absolute age in years in their analysis. Another approach used by Visser et al (2008) is to divide the properties into age classes, ranging between 20 and 40 years. This allows for the detection of a change in valuation of houses of a certain age rather than these canceling each other out and the age variable not turning out significant.

Literature is divided on whether or not apartments should be directly compared to other types of housing. Visser et al (2008) note that the demand for apartments differs from that of single-family houses but that to some extent they can be comparable. Because apartments play a big role in the theory of people searching for different types of property throughout their lifetime, apartments could be included as property type.

Locational variables

Locational variables can be seen as the structural surroundings of any property. Most locational variables relate to measures of accessibility. Accessibility measurements are traditionally related to distances to community hotspots such as a central business district or town hall (Chin & Chau, 2003). Accessibility can be measured in multiple ways, the most conventional being straight line distance or distance by road. Instead of travel distance or time, Benson et al (1998) have shown that out of sight or not can be used for certain characteristics that are near. By considering a line of sight or view, the structural attribute of building stories is being incorporated. So, Tse, & Ganesan,

(1997) find that there is a large correlation between view and floor level. Generally, a higher floor level has a less obstructed view and is therefore valued higher.

Throughout the literature mixed results have been observed for a broad range of variables: although a shorter distance to a school or business district can have benefits, it also comes with potential crowdedness or noise disturbance. A good example of the devaluation of immediate access is the process of suburbanization. Within the suburbanization process, buyers are effectively raising their transportation costs to balance them with lower costs of a property. This allows people to live in a bigger place or for a lower price in a quieter area for a reduction, but still present, accessibility to a range of facilities (Chin & Chau, 2003; Journal et al., 2009). Locational variables aim to map out to what extent which services are near, and secondly how well can additional services be reached. In addition to services or features related to the ease of living, Visser et al (2008) note that the availability of jobs is an important factor for people to value their location of living. This is contrasted however by the research of Edmonds (1984) who observes that when an employer will reimburse his employees for their travel costs, people are willing to travel more to get to their work. So et al (1997) describe that transportation can be divided into four categories:

- Availability of transport
- Transportation costs
- Travel time
- Convenience of transport

When transportation costs are covered by the employer, the transportation costs can be neglected. This leaves the remaining categories concerning the absolute distance, and the ways of transport being offered. Under the assumption that a high (varied) availability of transport possibilities will increase the convenience as well. This divides transportation into the absolute distance and the number of transportation possibilities. This does not take into account the possibility of switching jobs. When in need of a new place to work the advantage of compensated traveling might be lost, therefore the availability of jobs can be of importance in accordance with Visser et al (2008). To incorporate these factors into this research, the number of services is included in the analysis as any offered service offers potential employment.

The process of suburbanization shows a tradeoff being made between accessibility and cost of living. People will not move away to the cheapest location regardless of the distance to their occupation. This balance indicates a nonlinear relationship between multiple factors that have an optimum somewhere. The same can be argued for schools or city centers, which people would like close enough, but not within a proximity that causes them nuisance (Long & Wilhelmsson, 2020; Shin, Shin, & Lee, 2019). To accurately include this into this research, the proximity to schools as well as the proximity to potential sources of noise will be included into the analysis. A comparison between different modeling methods will be made to investigate the importance of these proximity relationships.

Neighborhood variables

Neighborhood variables are non-physical elements that are of influence on the valuation of a property. These socio-economic statistics can be related to social class or neighborhood religion (Chin & Chau, 2003). In relation to local variables, neighborhood statistics can provide a more in-depth insight into the amenities identified by the local variables. An example of this is the study conducted by Ketkar (1992) showing that test results or the expenditure of budget per pupil at a school can be positively related to local property prices. The explanation could be that people are willing to spend more to live close to a good school instead of just weighing the proximity to any school.

Another tested indicator is the local crime rate. Both Clark & Herrin (2000), as well as Li & Brown (1980), have proven different increasing crime statistics to negatively impact the local real estate market. Another connection is made there to the local age demographics, indicating a relation between the share of residents aged between 16 and 21 and various types of crime rates. Especially

since this study is focusing on interactions with age groups, the local crime statistics should be included in the analysis.

This study will, in addition to standard indicators, seek to find added value in the dynamic changes in the local age structure. As theorized before, in accordance with the life cycle theory, elderly people should have money to spend in their (new) neighborhood, whereas younger people do not have the same financial freedom (Zhang et al., 2020). According to Iwona Fory (2014), this can better be connected to significant life events rather than age. Examples of significant events in the model of Fory are the birth of a first child and the death of a spouse later in life. These are reasons to change a household's housing needs. This does follow the general model of households gradually needing larger housing until a later age when the household declines in the number of members and therefore size.

To express the (im)balance in the age structure, the status of a population age structure can be expressed as the 'age dependency ratio'. The age dependency ratio is a measure that represents the ratio between people of working age (20-65) and people outside of working age (0-20 and 65+). When the age dependency goes up, every working person must contribute more to support the non-working population. This mechanism is a topic of political discussion and public policy (Ford & Jennings, 2020). Although the effect of the age dependency ratio is a national statistic, this thesis explores whether this mechanic can have an effect on the local housing situation. What is yet to be conducted is a detailed study of both city and rural areas combined over a longer period of time. Local comparisons within a regional study are unique and provide more detailed information on where certain dynamics are most present.

Important to realize is that the effects of the life cycle theory should only occur during a change in neighborhood demographics. When no people are moving, there is no gap to be filled by someone that might not be financially capable of doing so. Therefore, to examine the theoretical discrepancy between elderly people moving out and downsizing their property and its effects, the relevant statistic is not the share of young and older people in an area but the change in demographic share (Hennadige, 2016; Zhang et al., 2020).

In addition to local income and the availability of schools, the proximity to sources of noise and the change in people over the age of 65 will be included in the analysis. According to the life cycle theory, an increase in elderly people should result in an increase in real estate prices, and vice versa. This study questions whether the available housing options suit the needs of a new generation of house owners with a different family structure (Silva, 2017; Zhang et al., 2020). In addition to this it is investigated weather a different modelling technique is able to more accurately depict housing prices based on empirical findings of locational variables.

To conclude, relevant literature has been closely examined and the most fitting variables have been selected. Variables have been selected to represent the structural, locational and neighborhood circumstances for the listings to be modeled. The methodology section will explain how these variables will be analyzed in further detail. A conceptual model has been constructed to visualize the theoretical relations within the housing market and is shown in Figure 1. This research will examine what happens when the steady flow of people through this process is altered through an aging population. In theory this should disturb the balance between supply and demand, causing a readjustment in prices. Table 1 provides an overview of the most used variables throughout the literature.

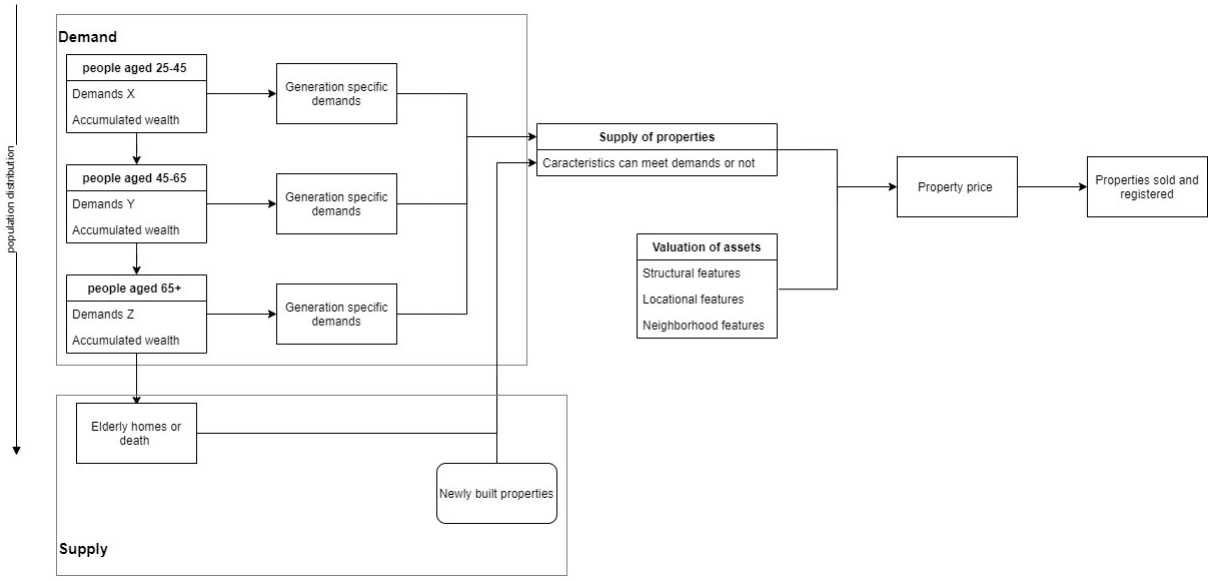


Figure 1 conceptual model of the housing market

Table 1 all variables to be included based on the literature

Variable type	Variable	Most prominently based upon
Structural variables	Number of rooms	(Yao & Stewart Fotheringham, 2016)
	Floor area of property	
	Garage	(Garrod & Willis, 1992)
	Property type	(Steegmans & Hassink, 2017; Visser et al., 2008)
	Age	(Visser et al., 2008)
locational variables	Urban environment	(Steegmans & Hassink, 2017; Visser et al., 2008)
	Views	(Chin & Chau, 2003; Steegmans & Hassink, 2017)
	Services	(Visser et al., 2008)
neighborhood variables	Neighborhood wealth	(Steegmans & Hassink, 2017)
	Potential noise pollution	(Visser et al., 2008)
	Availability of schools	(Yao & Stewart Fotheringham, 2016)
	Crime statistics	(Clark & Herrin, 2000)
	Age structure	(Hennadige, 2016; Zhang et al., 2020)

Methodology

All three research questions combine two aspects: First, insight is sought on a specific area of presumed influence, in this instance the combination of structural and locational features of a house on its price. Secondly, this relation was observed throughout time, to observe a temporal aspect. The first two research questions have the goal of creating a working, and accurate linear regression model as a basis. To answer the final sub-question, the combined results of the first two sub-questions should be compared to the random forest classification model. The third question investigates what can be observed in other relationships than linear, as can be the case when the variables have a spatial impact. To answer the main research question and its sub questions, prices for real estate properties have been modelled per year and compared to their actual values to compare the accuracy of the created models. Multiple data sources have been combined to produce the highest quality data, which was then used in two different models to examine their different workings.

Sub questions one and two are relatively similar, where the question 'To what extent does including local age structures lead to an improvement of the estimation of real estate pricing?' is an extension of 'How do structural and locational characteristics influence the price of real estate and does this change over time?', adding the theory on local demography. Research question three then uses the same data as questions one and two, however, applied to a model that accounts for the nonlinear distribution of the data. Essentially these questions examine different parts of the regression model and evaluate its effectiveness in terms of the ability to accurately model the transaction prices of real estate.

Detailed location description: Zuid-Holland

Real estate sales data for one province of my choosing has been made available for this study by the Dutch national real estate broker association (NVM). The chosen province is the province of Zuid-Holland. Due to its varied landscape of highly urbanized regions as well as rural areas, data for the province of Zuid-Holland was observed. Zuid-Holland is currently both the most highly as well as the most densely populated province in the Netherlands with 3.7 million inhabitants (CBS, 2020).

This province borders the provinces of Utrecht, Gelderland, and Brabant to the east. This area covers a part of the rural landscape that is called the green hart. Largely enclosed within the surrounding urban regions, the green hart is attempted not to be filled with large-scale urbanization (de Gans, H. A., & Oskamp, 1992). The largest part of this area is used for agricultural or recreational purposes on the reclaimed peatlands. The southern part of the province houses the estuaries of the Maas and the Schelde, the Maas facilitating the port of Rotterdam. This port is the largest seaport in Europe and provides 385 thousand jobs and contributes about 6.2% of the national GDP (Port of Rotterdam, 2020). Furthermore, the old levees and river runs provide a landscape that houses several different forms of agriculture as well as populated areas.

Although the Hague is the province's capital, it is Rotterdam that is the biggest city with currently almost 588 thousand inhabitants. However, these two major cities can be viewed as a larger metropolitan region containing multiple population centers throughout the region and containing some 2.4 million inhabitants within the Rotterdam-The Hague metropolitan area. This is a densely populated area with a lot of ethnic and cultural diversity (MRDH, 2020).

Over the past years, the total population of the province has been steadily increasing (see Figure 2), and so has the need for (affordable) housing. The amount of sold properties in the 4th quarter of 2020 was 41,184 where the average price had increased by almost 12% compared to the year before ("Marktcijfers koopwoningen | NVM," 2021).The provincial government aims to stimulate and balance the local economy, infrastructure, and the surrounding environment ("Missie en visie - Provincie Zuid-Holland," 2020). These topics are heavily impacted by the questions "where do people live?" and "how do people live?". To formulate effective policies in the right places the provincial government benefits from insights into the local housing market and its spatiotemporal development. Therefore it is not only beneficial for the local government to know what factors influence the housing market in what way, but relating this to developing demographics can also help anticipate future developments.

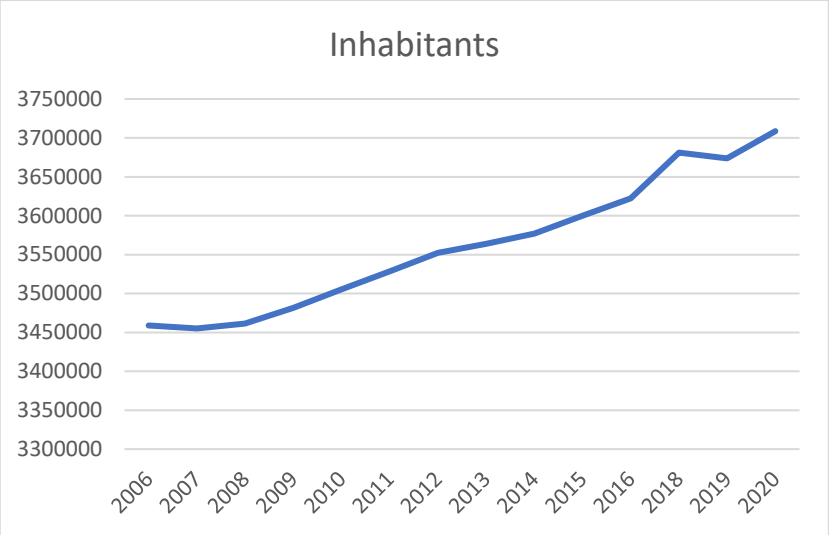


Figure 2 population development of Zuid-Holland over the past 15 years

Data sources

Brainbay listing data

Relevant data was provided by the NVM through Brainbay that record and facilitate the (online) advertising and selling of about 70% of the Dutch real estate market (Steegmans & Hassink, 2017). The data provided by the NVM includes 69 columns describing the structural features of every listing. A description of all variables can be found in Appendix A: description of NVM data records as delivered. The provided data by Brainbay constituted of data on one province for a period of 11 years (data from 2006 until 2016). The value of this research to Brainbay consists of a better understanding of their data as well as insights into potential future trends within the world of real estate concerning the aging Dutch population.

Additional data by CBS

For the collection of additional data, the Dutch bureau of statistics (CBS) has primarily been used. This bureau publishes yearly reports and datasets on several scale levels. To work with consistently defined and publicly available data, the decision was made to use the CBS data instead of performing my own locational analyses out of multiple data sources. Especially for comparing many different time series, it is useful to use the same format and source of data to ensure comparability and operability.

Data provided by the CBS is available on several scale levels. The smallest level is squares of one hectare. While these provide the most detailed information, for privacy concerns, the CBS does not include statistics for areas with a population density under 'a certain' threshold. Rather large parts of the case study do not meet this threshold value and thus cannot be included in this study. The second most detailed scale is squares of 500 by 500 meters, these provide a more complete surface of aggregated data. When looking from the perspective of provincial analysis and policymaking, no legislation can be made at a 500 by 500-meter scale, therefore it is more logical to use any administrative area. The most detailed administrative level on which the CBS publishes its

data is the neighborhood level. Therefore, the largest scale level at which this analysis would be sensible is neighborhood (Dutch: buurt) level.

On a neighborhood level, the CBS offers both a shapefile, presented in the yearly “wijk- en buurtkaart” as well as data presented in a .csv table format called the “Kerncijfers wijken en buurten”. These two sources do not always contain the same information; therefore, both were used to complement each other where necessary.

The only selected metric not to be included in the neighborhood datasets is the crime rate in the area. In order to include crime as a factor of influence, a third CBS source is used. The “Geregistreerde criminaliteit per gemeente, wijk en buurt, 2010-2015” offers a wide variety of statistics on neighborhood crimes in the categories of theft, destruction, and violence either in absolute numbers or standardized to the number of incidents per 1000 neighborhood inhabitants.

All these data sources were combined through an overlay analysis with the individual locations of house listings, thereby not aggregating the data to any predefined area. The demographic statistics were then fed into the regression model for various consecutive years to analyze its effect.

RIVM noise pollution

In addition to the CBS one other source is being used to enrich the data. The National Institute for Public Health and the Environment (RIVM) is an independent institute that provides research and advice on population health and security. This institute monitors many different factors that might impact the general wellbeing of the Dutch population. The relevant data required for this study is a map containing levels of noise pollution from 2016. This data is publicly available at a raster size of 10x10 meters and accounts for noise produced by roads, trains, air travel, industries, and wind turbines. As previously mentioned, this research aims to use as much openly available data as possible. Next to the advantage of saving time and effort in performing a custom analysis that might be more accurate, the use of publicly available data ensures the data quality as well as the reproducibility of the research.

Data preparation

Brainbay data

The received data contains about 170 000 records, of which almost 160 000 could successfully be geocoded and therefore be used for analysis as seen in Figure 3. Most of the 10 000 listings that could not be coded did not have a valid address, in most cases these were parcel indicators for newly built properties making it impossible to assign them x and y coordinates.

For correct computation, all dates have been transformed to date formats. The obtained data is the raw output of manually inserted values by many different contributors. Because of this data formats are not consistent, or always complete. Records that cannot be deciphered have been removed.

After all records have been dated, the main dependent variables were cleaned, and their descriptives are included in Appendix B: Descriptive statistics on variables of interest. Both the columns transaction price, as well as transaction price per m2, were checked for null or unrealistic values. Unrealistic values being the highest outliers of 999,999,999 or 99,999,999. These values could be viewed as missing data, but (likely due to manual entries) there is no consistency in labeling no data values. A lower realistic price limit has been set at 100,000 to ensure listings to be actual properties instead of for instance parking spaces. After this selection 122.340 records remain. These records have been split per year on the date that they have been removed as a listing. Splitting the data ensured matching to the correct additional data. A conscious decision has been made to only remove unrealistic values. Similar values like 999999 can still be present but blended in between other values. There is no way to tell which values were inserted truthfully or just to fill the field. This shows that the data quality can almost never be perfect. Outliers that were not obviously unrealistic have not been removed for two reasons:

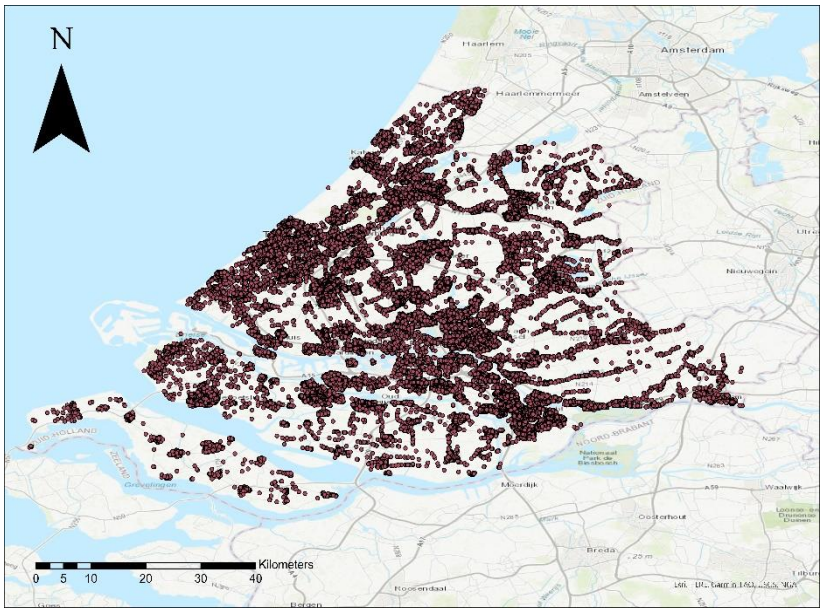


Figure 3 map of the study area, including all (158 852) geocoded listings before cleaning

1. Values that are very high compared to the average value might still be clustered together and tell something about these locations by their inclusion.
2. Due to the relative transparency and available comparison possibilities, property prices are expected to be properly founded. If they were not, they would not have been sold, or the transaction price would have been lowered until they were. The goal of this research is to investigate what factors are of influence in this process and therefore a broad range of properties is included.

The possibility that a value was wrongfully entered can never be fully eliminated, this will have to be taken into account when interpreting the results.

Creating dummy variables

Within the Brainbay data, several categories consisted of coded ordinal or nominal values. These values were not necessarily all fit for analysis as data was missing, irrelevant, or to be combined. Therefore, columns would be reclassified, and eventually each class would be transformed into its own binary column, creating a dummy variable. An important part of the reclassification process is to ensure that the resulting classes do not differ greatly in their number of values to prevent skewness. The creation of dummy variables was performed in python with the aid of the pandas package before storing the data as CSV tables. The use of the comma-separated value format ensured easy interoperability between different analysis and editing programs for the following steps.

Joining data sources

The geocoded Brainbay data has been assigned an x and y location and can therefore be plotted on a map. This process as well as the further data analysis was performed in ArcGIS Pro version 2.7. The corresponding neighborhood statistics should then be added to the points within

that neighborhood; however, the individual points contained no more location information than their coordinates and address. To properly join these different tables, a spatial join is performed between the points representing individual transactions and the shapefile of each respective year's neighborhoods. It is important to use the appropriate year of neighborhood geometries to account for any possible change in neighborhood compositions. Next to the statistics on the area included in the shapefile, a neighborhood code was added to each point indicating in which neighborhood this point was in this year. This code consists of the letters "BU" followed by 8 numbers. This code was used as common field upon which to join further neighborhood statistics. This is especially important as only the shapefile had a spatial reference and acted as the combining element between the transaction point data and additional locational statistics located in tables only. The BU code was used to join the neighborhoods to the "Kerncijfers wijken en buurten" as this table offered the most complete data. This process was repeated for the crime statistics, which also shared the BU code. The CBS has occasionally revised their offered numbers, usually expanding their range of statistics, however, in some instances, a particular statistic was missing for one or more years. In the case of missing statistics, the closest following year has been used to fill the data, to ensure that different analyses have the same interacting variables. Appendix C: Description of joined data sources for yearly data shows what key data sources have been used for each relevant year.

To complete the data aggregation, the tool "Extract Values to Points" was used to obtain the specific noise pollution score from the raster and added to each point. This works similar to a spatial join, where the data at the location of the specific point feature was be attributed to this point feature.

The complete datasets were then checked for the right data formats. Due to the importation from a CSV file and the exploration of data in excel, fields were frequently read as text values by ArcGIS. This was solved by copying these values into the "double" data format. The complete and functional tables were then exported, keeping only the relevant fields in order to decrease the size and chance of error. These outputs are the final data products and have been analyzed in the following analyses. The process of data preparation is schematically shown in Figure 4 below.

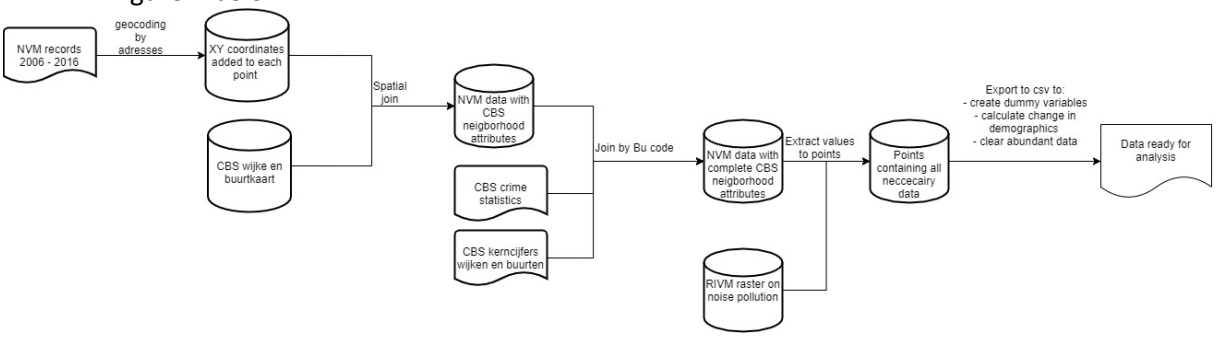


Figure 4 schematic overview of the data preparation steps taken

Adding temporal population dynamics

One of the research objectives of this project is to see what the impact of change in the age buildup of the local population is. Under the assumption that people over 65 have had more time to accumulate wealth, and do not have to provide for children living with them, they would have more financial means to spend on a house. According to supply and demand, an increase in this demographic would then lead to higher demand resulting in higher prices. To include this in the analysis, the percentual change in people over the age of 65 relative to the year before has been calculated and included as a parameter in the final data.

Variables included

An important part of building the hedonic model is the selection of proper dependent and independent variables. The dependent variable has been set at price per square meter, this allows for a fair comparison between houses of different sizes. To contribute to the space of living the availability of a garage will be included as dummy variable. Also, the property type has been identified and will be taken into account so as to differentiate between a small family house and a large free-standing mansion. A large range of subdivisions between property types can be made, however over specification can lead to misguided conclusions. Therefore, the types of properties will depend on the relation to its neighbors and can either be terraced, semi or fully detached, as is a common division (Steegmans & Hassink, 2017; Visser et al., 2008). As a further indication of the living space, the total number of rooms will be included.

Literature is divided on whether or not apartments should be directly compared to other types of housing. Visser et al (2008) note that the demand for apartments differs from that of single-family houses but that to some extent they can be comparable. Because apartments play a big role in the theory of people searching for different types of property throughout their lifetime, apartments could be included as property type. Unfortunately, the retrieved data did not include any type of apartment, and therefore it will not be taken into account in the analysis.

Kohlhase (1991) has previously concluded that the influence of variables can vary and change between nations and time periods. It is therefore that the selection of these variables has been heavily based on the practices of several recent studies conducted in the Netherlands, or that had a very similar methodology (Steegmans & Hassink, 2017; Visser et al., 2008; Yao et al, 2016). To make sure that the variables do not interfere with each other they will have to be observed more closely and tested for their usefulness in order to prevent the absence of data values or extreme outliers.

Table 2 structural variables to be included

Variable	Most prominently based upon	Adjusted or directly based on exiting research	measurement	Variable type	Data source
Number of rooms	(Yao & Stewart Fotheringham, 2016)	Based on Yao et al, however adjusted to point data instead of areal averages	Total number of rooms within the property	numerical	NVM
Floor area of property			Number of square meters of available floor area	numerical	NVM
Garage	(Garrod & Willis, 1992)	Not adjusted	Yes or no	Dummy	NVM
Property type	(Steegmans & Hassink, 2017; Visser et al., 2008)	Not adjusted, standard classification by NVM	Terraced house	Dummy	NVM
			Semi detached	Dummy	
			Detached	Dummy	
Age	(Visser et al., 2008)	Values are numerical. Based on Visser et al, they have been classified	Before 1906	Dummy	NVM
			1906 - 1944		
			1945 - 1970		
			1971 - 1990		
			1991 - 2000		
			After 2000		

To analyze the level of available services in the area, the decision has been made to include the variable “distance to services” within a given area rather than the number of nearby facilities, measured by supermarkets (Visser et al., 2008). The choice to measure distance rather than the density of services was made to prevent multicollinearity with the urban density. To prevent issues with multicollinearity between objects such as restaurants, only one has been chosen as many services such as restaurants and bars often occur alongside each other. Features of the direct surrounding environment will be included as established by the realtor. In addition to this, the view of any surrounding nature will be included as well as an indicator for the level of urbanization and business represented by the number of primary schools within 3 kilometers. The number of schools was chosen rather than the shortest distance to a school because the number of schools provides information on the local urban density. In rural villages with only one school, there are properties within proximity of this school, generating the same scores as densely urban areas when measuring the closest distance. In addition to this, having multiple schools within a proximity might be an extra source of traffic or activity that cannot be explained when looking at the closest school only. To exclude the influence of religious or quality motivations for choosing a school or not, the local availability was chosen. This aspect was deemed more significant for schools than for supermarkets, where people may have their preferences, but to prevent interference between the two, two different measures were chosen (Chin & Chau, 2003).

Table 3 Locational variables

Variable	Most prominently based upon	Adjusted or directly based on exiting research	measurement	Variable type	Data source
Urban environment	(Steegmans & Hassink, 2017; Visser et al., 2008)	Adjusted to not interfere with the intended OAD variable	Number of primary schools within 3 kilometers (as a replacement for OAD due to multicollinearity)	numerical	CBS neighborhoods
Views	(Chin & Chau, 2003; Steegmans & Hassink, 2017)	Adjusted to the available data classification	forest	Dummy	NVM
			Water		
			Park		
			Free view		
Services	(Visser et al., 2008)	Adjusted to the available data	Distance to closest supermarket (km)	Numerical	CBS neighborhoods

Thirdly, based on the existing literature the following neighborhood variables have been selected to complete the initial model to create the basis on which to expand upon in research questions two and three. Neighborhood wealth is measured as the average income per income recipient. In theory the income could be divided over the total inhabitants instead of recipients. When reasoning that a larger family has a lower average income per person, and therefore should be able to spend less of their income on housing, this could be used as measurement. However due to data availability and the literature standard, the income per recipient has been included.

Table 4 neighborhood variables

Variable	Most prominently based upon	Adjusted or directly based on exiting research	measurement	Variable type	Data source
Neighborhood wealth	(Steegmans & Hassink, 2017)	Not Adjusted	Average neighborhood income	Numerical	CBS neighborhood statistics
Potential noise pollution	(Visser et al., 2008)	Not adjusted	Distance to railroads, highways, airports and industry	Numerical	RIVM noise pollution
Availability of schools	(Yao & Stewart Fotheringham, 2016)	Yao et al consider distance to university	Amount of primary education schools within 3 kilometers	numerical	CBS neighborhood statistics
Crime statistics	(Clark & Herrin, 2000)	Instead of murder rate, theft is used	Number of thefts per year per 1000 inhabitants	Numerical	CBS neighborhood statistics

Linear regression

All three research questions combine two aspects. First insight is sought on a specific area of presumed influence, in this instance, this is the combination of structural and locational features of a house on its price. Secondly, this relation is observed throughout time, to observe a temporal aspect.

A property is made up of all its individual assets, however, it can only be obtained altogether. This research question was intended to establish an empirical base model of property prices that will later be expanded upon. This model is intended to identify the added value of certain features based on transaction records, as is the core of hedonic price modeling (Chin & Chau, 2003; M. M. Li & Brown, 1980). This model will observe all available records for the province in a year. The second aspect of this question relates to the temporal aspect. To observe the change in valuation of certain aspects a model was built for each year between 2009 and 2016. Then consecutive models reflect changes in valuation for each individual aspect. To conclude, this question has provided the baseline model that will be compared against more elaborate models built later.

Hedonic price models: Linear regression

To find any relation between multiple points with common attributes in a graph, a line can be fitted through them and the line can be defined by a formula. This formula should approximate the relationship between the attributes. To identify the effects of structural aspects of a property the known characteristics of all properties will be compared to each other. When property A has a feature that the otherwise similar property B does not have, a conclusion can be drawn on the effect on the transaction price. This method is called regression analysis when applied to find the relations between multiple observations and a fixed or dependent variable (Edmonds, 1984; Grömping, 2009). When applying this to many properties sold in the same timeframe a trend can be observed, and the difference between consecutive trends can be attributed to the changing valuation of different structural characteristics.

A regression model can be summarized by a formula for the estimated value of the dependent variable, expressed in each independent variable and their relative influence factor. The most simplified form of this formula is: $Y = \beta X + \alpha$ with X being the value of one independent variable that is of influence on Y by the factor β. In other words, β expresses the slope of Y. The other part of the function is α, a set value-added to the variable outcome of the equation. α is included to "fit" the model to its observations. In regression modeling, the Y values are known and sought to be explained by the observations of variables of X. For each different variable (another X is included in the equation) a value of β is sought that best approximates Y for all observations. Any difference between Y and βX is attributed to potential errors or external factors at play and compensated with α, therefore a more accurate model will have a lower α value (Long & Wilhelmsson, 2020)

The hypothesis under which this model will be tested is: "Transaction prices per square meters are dependent on locational, structural and neighborhood characteristics."

With H_0: $Y = \beta X_{locational} + \beta X_{structural} + \beta X_{neigborhood\ (including\ age)} + \alpha$

H_a: $Y = \beta X_{locational} + \beta X_{structural} + \beta X_{neigborhood\ (including\ age)} + \beta X_{place} + \alpha$

In these hypotheses, the X factors will contain a bundle of variables, where the local age structure was included in the neighborhood variables

Although any regression model aims to simulate the real-world circumstances as closely as possible; it is not necessarily more accurate or reliable to include as many variables as possible. Conflicting issues such as multicollinearity and misspecification of variables might decrease the effectiveness of an overly complicated model, as noted by Chin and Chau (2003).

To perform the linear regression, the tool "Generalized Linear Regression" was used. This tool is one of the least complex ways of performing linear regression, as it does not consider location or clustering and searches for linear relationships between all inserted independent variables and the dependent variable. Within this tool, the gaussian model type is used to fit a numerical dependent variable, as opposed to dependent variables being binary or counts of specified met criteria.

Comparing fitted values to the recorded values in the data allows for the calculation of the model precision, primarily measured by the R^2 score. This score is based on the amount that the model estimation is off from the recorded values. If the model were to simulate all values exactly, the R^2 would be 1, meaning that the model is 100% accurate. A lower score means lower accuracy, but a 100% accuracy generally means that the model is too specific or precise to be applicable to other situations (Chin & Chau, 2003).

The datasets per year were run individually with the explanatory variables shown below in Table 5. The variable PROCENTVERSCHIL_65_VOORGAAND is only included in the second run to show the added value of this statistic. Descriptive statistics on the complete dataset can be found in appendix C. The life cycle theory and the general economic theory of supply and demand were tested through the results of this question. As supply and demand cannot be directly measured, derivative indicators based on population demographics were used.

To capture the temporal aspect as well as the dynamic motion of market fluctuations, demographic variables will be observed relative to each other. The assumption in this question is that a decrease in one age group will increase the supply of their properties in that location. Naturally, this means that an increase in this age group rises the demand and is expected to drive prices up. To represent this in the model the created variable was the relative change in people over 65 compared to the year before. All variables previously discussed stayed unchanged, with the additional change in absolute presence of people over 65 in relation to the previous year.

Table 5 variables used to explain the transaction price per square meter

Variable name (year 2009)	Meaning of variable	Variable measurement
OBJ_HID_M2	Number of square meters	numerical
OBJ_HID_NK	Number of rooms	numerical
F1906___1944	Built between 1906 and 1944	0= no 1= yes
F1945___1970	Built between 1945 and 1970	0= no 1= yes
F1991___2000	Built between 1990 and 2000	0= no 1= yes
PRE__1906	Built before 1906	0= no 1= yes
AFT__2001	Built after 2001	0= no 1= yes
GARAGE_AAN	Garage available	0= no 1= yes
DETACHED	House type is detached	0= no 1= yes
SEMI_DETAC	House type is semi-detached	0= no 1= yes
FOREST_VI	Property has a view of the forest	0= no 1= yes
FREE_VIEW	Property has a free view	0= no 1= yes
PARK_VIEW	Property has a view of a park	0= no 1= yes
WATER_VIEW	Property overlooking water	0= no 1= yes
GELUIDSNIV	Level of noise pollution	Classified into levels 1, 2, 3, 4 or 5
TOTAAL_CRIME_NR	Total number of crimes per 1000 inhabitants	numerical
AV3_ONDBAS	Number of primary schools within 3 minutes	numerical
AF_SUPERM	Distance to nearest supermarket	numerical
INK_ONTV	Neighborhood income level (per income recipient)	Numerical (x1000)
OAD	Building density (degree of urbanization)	numerical
PROCENTVERSCHIL_65_VOORGAAND	(percentual change in population aged over 65 relative to the year before)	percentage

When running the consecutive years of transaction entries, there appeared to be an issue when reading the data of the year 2008. As it had gone through the same preparation process as the other years, it is unclear what is causing this problem. Due to time pressure and the continuity of the results the decision was made to look at the timeframe from 2009 until 2016. After running the regression model, the following outputs were produced:

- A graph of variables plotted against each other, to determine relationships between all variables individually.
- A table showing variable coefficients and significance, as well as several model performance indicators.
- A map of the distribution of standardized residuals.
- Standardized Residual vs Predicted values plot.

When studying the outputs of these regressions, the variance inflation factor (VIF) of the OAD and lesser so for the AV3_ONDBAS variables would in some cases go up to 5 and above. In the literature, some argue that scores over three, and others claim that a score above ten, indicates significant multicollinearity. There is no consensus on any type of strict upper limit, and therefore it was decided to leave out the OAD as it had the highest scores (Thompson, Kim, Aloe, & Becker, 2017). A degree of interference between the number of primary schools and the degree of urbanization seems logical: When the number of people in an area is higher, there can be an expected higher need for more schools to educate the children of those people.

These metrics and the full extent of their implications will be discussed in the results and conclusions sections.

Random forest regression

The focus in the first two research questions was on the temporal development of key characteristics of a property. The outcome of these questions will be used in this final research question to include spatial variation and distribution of the previously examined factors.

Next to the spatial features or services that a location can offer which are accounted for in the locational variables, housing markets do also depend on interactions with their location (Hochstenbach & Arundel, 2020; Inchauste, Karver, Kim, & Abdel Jelil, 2018; So et al., 1997; Yao & Stewart Fotheringham, 2016). This question used a different approach that should in theory be able to more accurately approach nonlinear relationships to increase the model accuracy. This relatively new method is called random forest classification. This method will be explained in further detail in the following paragraph.

Hedonic price models: Random forest classification and regression

The random forest regression algorithm creates many outcome estimates for the same value by randomly looking at its different features for each estimate. The definitive estimate is composed of all different predictions. The aggregation of many different predictions helps to mediate extreme values. Each individual estimation is the result of a series of 'if' statements leading to new 'if' statements until a uniform selection remains. This is called a (decision) tree, and many trees make up a forest (Breiman, 2001; Ouedraogo, Defourny, & Vanclooster, 2019).

An example of different trees can be seen in Figure 5, here can be seen that each tree uses different combinations of criteria to reach a conclusion. In this example this is a buy or not buy decision, however in the case of forest regression, the outcome will be an estimate based on the linear approximation of all the values in that final classification. To complete the random forest regression, the estimates or outcomes of these three trees will be combined and either averaged for numerical values, or the most frequent predicted outcome is used in case of categorical outcomes.

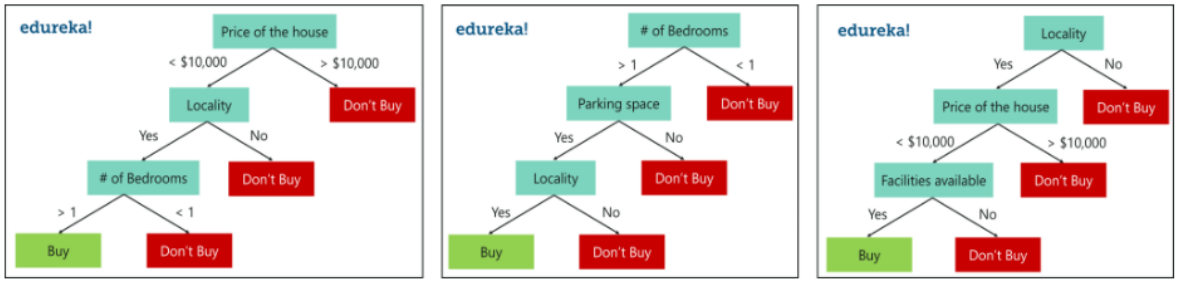


Figure 5 example of multiple different decision trees used alongside each other. Source: Edureka (2021)

Figure 6 shows how narrowing down on the data can improve the estimation capabilities of the produced model. By being able to fit a linear approximation to a more specific subset of the data, more details can be modeled. The risk of this method is overfitting, whereby each point of the training data fits the model perfectly, however this model would not be applicable beyond its training data. To combat multiple different trees are used, and each tree uses a different random sample of the training data. The performance of a random forest regression is measured by the R^2 statistic, just like a linear regression model (Yuchi et al., 2019).

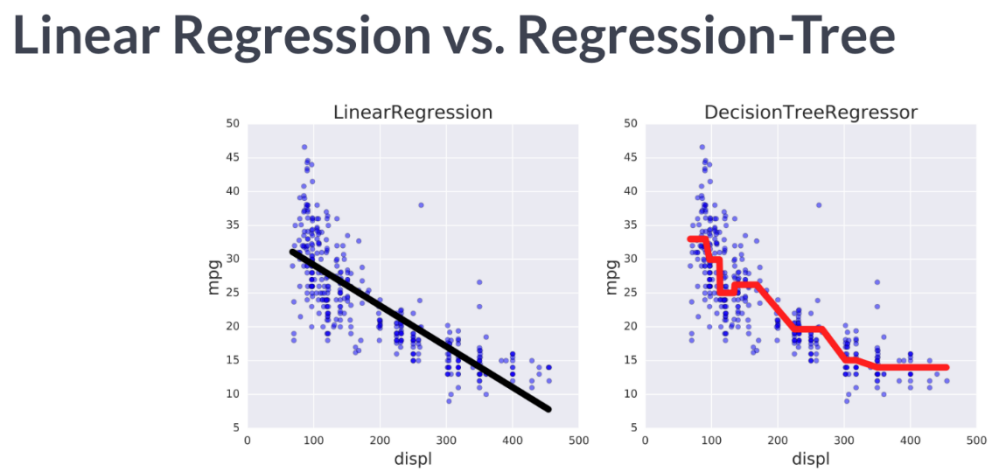


Figure 6 comparison of a linear model estimation versus a classified decision tree estimation. Source: Deepnote (2021)

By using a large volume of different trees, a broad estimate of the target value can be made as the final outcome will be an aggregation of all individual trees, thereby aiming to cancel out extreme values. This model is not limited to linear relationships and therefore was expected to depict more accurately some of the unique locational variables.

No changes were made to the inputs when running the forest-based classification and regression tool. The forest-based classification and regression is used to test the geographical influence on the linear regression model. For this specific research question, the difference between incorporating demographic change or not was no longer the topic of interest. Because of this, all runs have included the PROCENTVERSCHIL_65_VOORGAAND variable. Again, outputs were generated with both the transaction price, as well as the transaction price per square meter for comparison purposes before deciding to use the price per square meter.

This machine learning model uses a randomly selected portion of the complete data as training data to fit a model. This model can then be evaluated on its prediction capabilities by feeding it the remaining portion of the data without including the dependent variable. The model performance can then be measured by the accuracy of its predicted transaction prices versus the actual prices. However, as this research was aimed at explaining the dependent variable instead of predicting it, the trained model R-squared has been used to measure its performance against the linear regression.

For each run, ten iterations of 500 decision trees were used that used all training data as potential input. Due to the “randomness” of the different decision trees in the forest, all ten iterations had a different outcome. The median R-squared of these runs will be compared to the outcome of the linear regression outcome.

Comparing of model outcomes

Model fit

The primary tool to measure the performance of both model types is the R^2 value. The linear regression will have one consistent value for each run, where the random forest regression will slightly differ based on the created decision trees. To minimize the influence of created output

outliers, the median R^2 value over ten runs will be compared to that of the linear regression, as opposed to the average of the ten outputs (Cai et al, 2020). In addition to gaining understanding of past events, models can be used to create predictions of future events. To measure the capability of a fitted model to make predictions, some of the known data is not used to fit the model, but it will be used as input for predictions. As this data is known, the predicted value can be compared to the actual value, and another residual score can be calculated. However for the purpose of this research, only the ability of the models to fit existing data accurately will be examined.

Variable influence

To gain a further understanding of the respective models, the general accurateness provides little information. To see how the models estimate their outcomes, the workings of individual variables will have to be examined. It is difficult to compare the modelled coefficients from the linear regression to the variable importance score of the generated random forest regression. Both produce a numerical value for each variable; however, these are not directly translatable for comparison. Because of this the relative contribution to their respective model will be used to draw conclusions on the results (Grömping, 2009). A higher variable coefficient will produce larger deviations in the linear model outcome, whereas a higher importance in the random forest regression indicates that a variable is to a greater extent responsible for the estimation outcome. Because of this, to compare the workings of both models, all variables will be ranked by their coefficient, or importance value and compared. This will allow for observations on weather a variable contributes the similar to the linear regression as it does to the random forest and vice versa.

Results

Model accuracy compared

Combing the linear analysis with and without age variable, and the random forest regression for transaction price per square meter, a total of 3 different analyses have been performed, generating a total of 24 separate outputs. Full output reports will be included in a separate folder with this report Labeled “Analysis_Results”. Figure 7 shows a comparison of explained variance values for each model through the years. This figure shows that the added value of this age statistic is negligible as the lines cover each other. Furthermore, the forest regression shows to explain between 15% and 30% more variance. This chapter will provide a more general overview of the outcomes upon which the conclusions can be based in the following chapter. As described in the theoretical framework, the price per square meter was used as the dependent variable. Important to realize in the outputs is that the coefficients have not been normalized, and therefore their meaning is the added value to the price per square meter of a one-step increase in the respective explanatory variable.

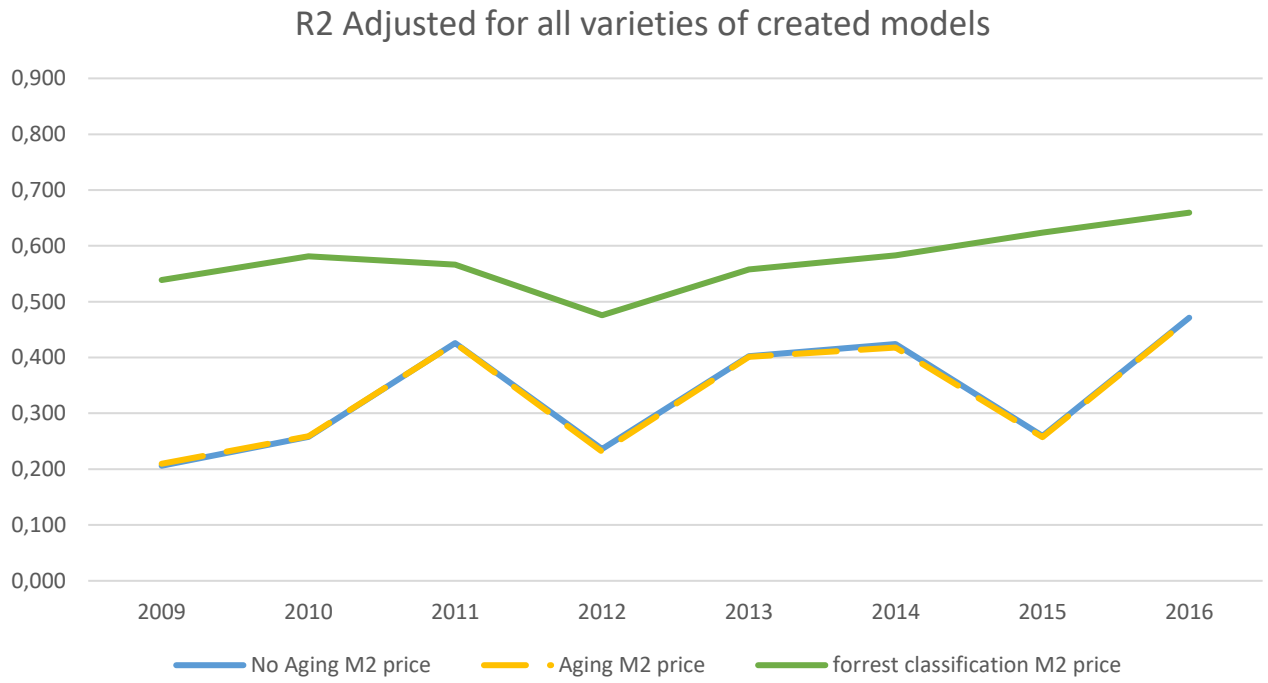


Figure 7 R-squared values for all different models through the years

For the evaluation of the eight linear regression models several outputs are of relevance:
The model output log (

Table 6 and

- Table 7)
- The distribution graph of the standardized residuals (Figure 9)
- Standardized Residual vs Predicted values Plot (Figure 8)

The most prominent statistic on a statistical model is the explained variance measured by the R^2. The robust R^2 over the different years can be seen in Table 7. As opposed to the regular R^2 the

adjusted R^2 will take into account the number of independent variables and decrease when variables are included that do not or barely add to the model accuracy (Leach et al., 2007).

Table 6 output of the linear regression model of 2009

Variable	Coefficient [a]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	1644,114	55,287	29,738	0,000*	--------
Number of square meters	0,325	0,376	0,864	0,388	2,654
Number of rooms	33,160	11,784	2,814	0,005*	2,229
Built between 1906 and 1944	336,749	20,496	16,430	0,000*	1,473
Built between 1945 and 1970	157,308	26,887	5,851	0,000*	1,350
Built between 1991 and 2000	200,574	15,578	12,875	0,000*	1,277
Built before 1906	446,178	38,113	11,707	0,000*	1,249
Built after 2001	289,749	22,480	12,889	0,000*	1,339
Garage availible	147,277	27,357	5,384	0,000*	1,152
Detached property	982,801	71,873	13,674	0,000*	1,262
Semi detached property	189,808	13,833	13,721	0,000*	1,084
Forrest view	592,443	125,999	4,702	0,000*	1,022
Free view	76,510	21,616	3,539	0,000*	1,054
Park view	62,636	32,210	1,945	0,052	1,027
Water view	217,637	41,319	5,267	0,000*	1,091
Noise pollution	-30,770	9,150	-3,363	0,001*	1,103
Crime number	0,881	0,215	4,092	0,000*	1,111
Change in elderly population	0,000	0,000	1,142	0,253	1,152
Availability of schools	0,000	0,000	0,271	0,787	2,897
Services	0,000	0,000	0,549	0,583	3,002
Neighborhood wealth	0,000	0,000	-0,734	0,463	1,154

Table 7 continuation of Linear regression output with diagnostics

Input Features:	Complete_2009__XYTableToPoint	Dependent Variable:	OBJ_HID__TRANSAC_M2
Number of Observations:	9362	Akaike's Information Criterion (AICc) [d]:	148535,4687
Multiple R-Squared:	0,207844	Adjusted R-Squared [d]:	0,206148
Joint F-Statistic:	122,543791	Prob(>F), (20,9341) degrees of freedom:	0,000000*
Joint Wald Statistic:	1577,368544	Prob(>chi-squared), (20) degrees of freedom:	0,000000*
Koenker (BP) Statistic:	116,551375	Prob(>chi-squared), (20) degrees of freedom:	0,000000*

Jarque-Bera Statistic:	17056230,76	Prob(>chi-squared), (2) degrees of freedom:	0,000000*

The second output is the standardized residual values plotted against the predicted values. If all residuals would be normally distributed, the predicted value would be the mean probability of the predicted value or the value that is on the fitted line through the data. Surrounding this line are different data observations, that should be decreasing in likelihood and therefore occurs when the residuals are normally distributed. Figure 8 shows the standardized residual for the predicted values of the dependent variable. This means that the predicted values have been standardized at 0 with a standard deviation of 1. According to a rule of thumb, about 99 percent of all data in a normal distribution should be within twice the standard deviation of the mean, and therefore should in this case form a stroke with uniform width around the 0 value. Next to some high outliers, this graph shows two other situations.

First of all, an empty triangle can be found in the bottom left corner of the graph. Between Y=0 and X≈2200. This means that cases at the lowest price level that turn out to be cheaper than estimated by their attributes are non-existent. Secondly, the graph appears to gradually be widening as X increases, indicating that the certainty of prediction decreases for higher estimated floor area prices.

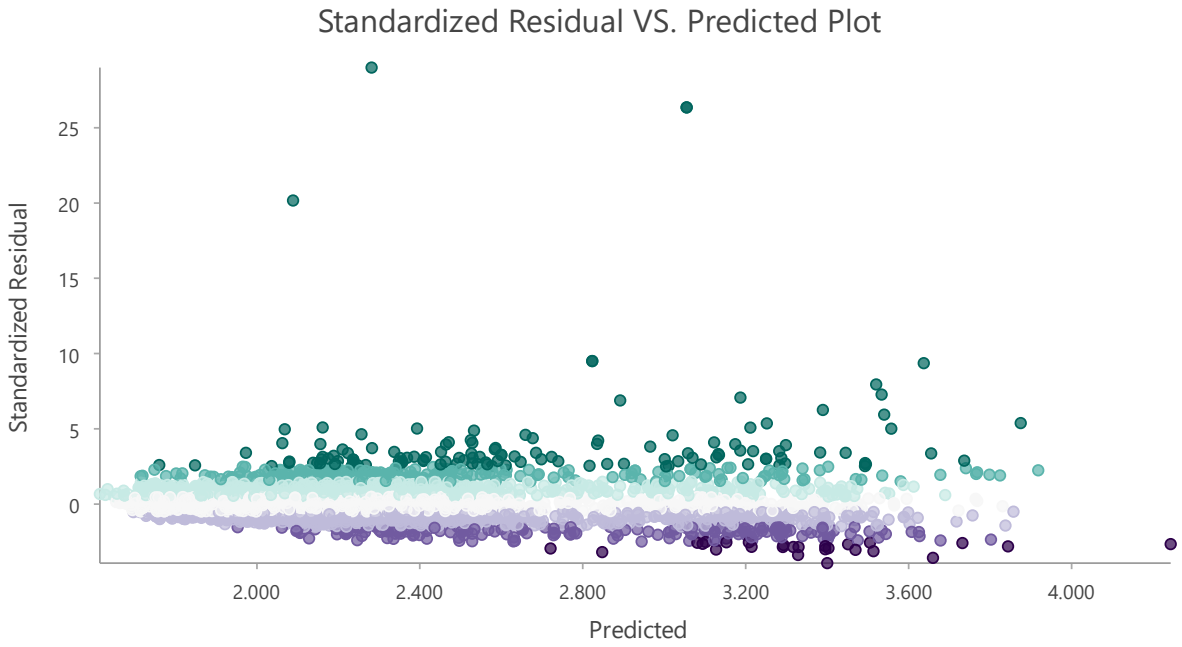


Figure 8 Standardized Residual vs Predicted values Plot for the linear regression of 2009

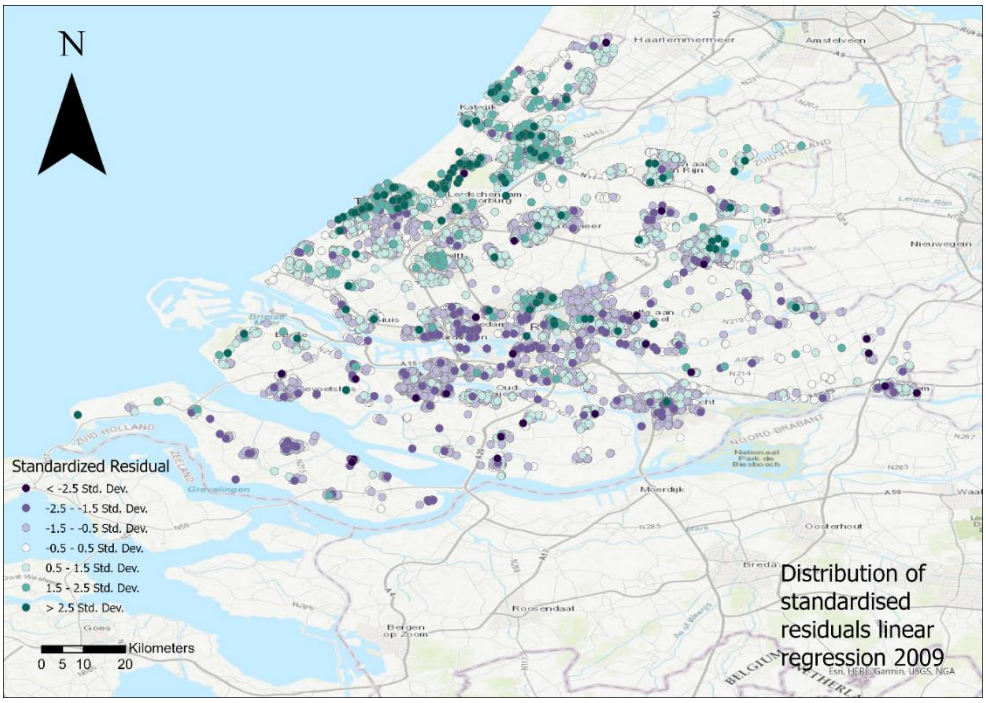


Figure 9 map of the distribution of standardized residuals for the linear regression of 2009

Figure 9 shows a diagonal divide between over-estimation to the northwestern part of the province, and underestimation towards the south and east. This is remarkable as the urban centers Rotterdam and the Hague are observed to behave differently. The following paragraphs will look into the results in further detail, aiming to provide answers to the three research questions.

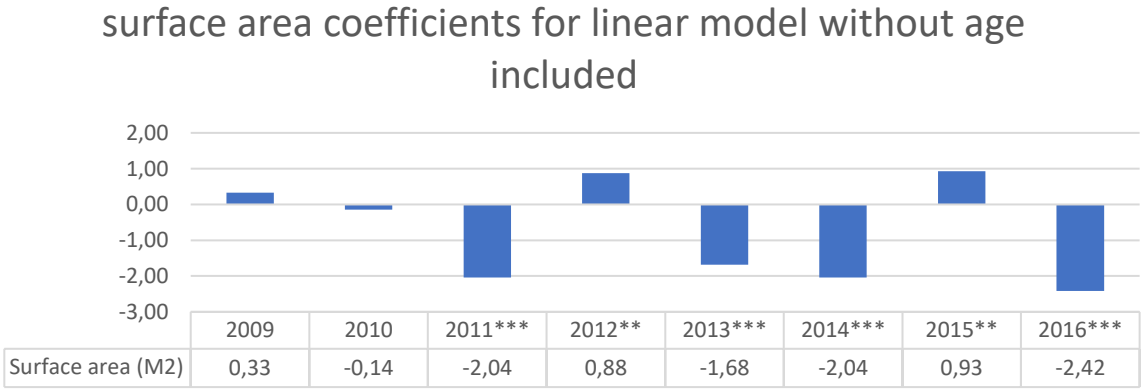
Research question 1

How do structural and locational characteristics influence the price of real estate and does this change over time?

To answer this question, the transaction data without the population age statistics have been analyzed. For this, the coefficients and their probability per variable are of importance. A probability below 0.05 means that the explanatory variable is statistically significant, indicating that the influence of this variable on the dependent variable is strong enough to discard the assumption that this effect can be random and should therefore be attributed to the explanatory variable.

Structural variables

Floor area

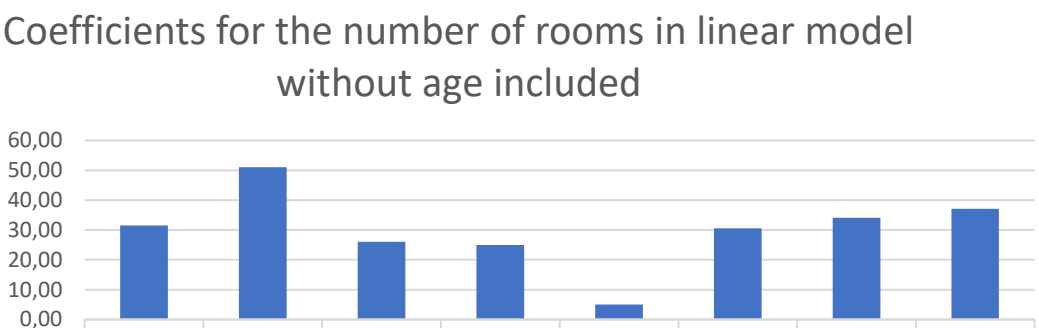


	2009	2010	2011***	2012**	2013***	2014***	2015**	2016***
Surface area (M2)	0,33	-0,14	-2,04	0,88	-1,68	-2,04	0,93	-2,42

Figure 10 coefficient and significance of floor area over time

Figure 10 indicates the added value in euros to the price per square meter, per added square meter to a property. When significant, this indicator is mostly negative although resulting only in small changes in the price per square meter.

Number of rooms

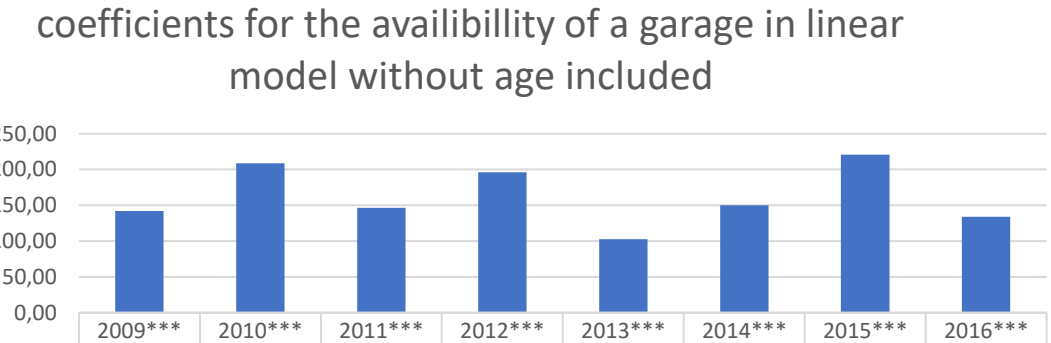


	2009**	2010	2011***	2012***	2013	2014***	2015***	2016***
Number of rooms	31,55	51,04	26,10	24,93	4,97	30,52	34,14	37,15

Figure 11 coefficient and significance of number of rooms over time

Figure 11 shows that the number of rooms of a property is in most years of significance and has a positive impact on the resulting price per square meter of the property.

Garage

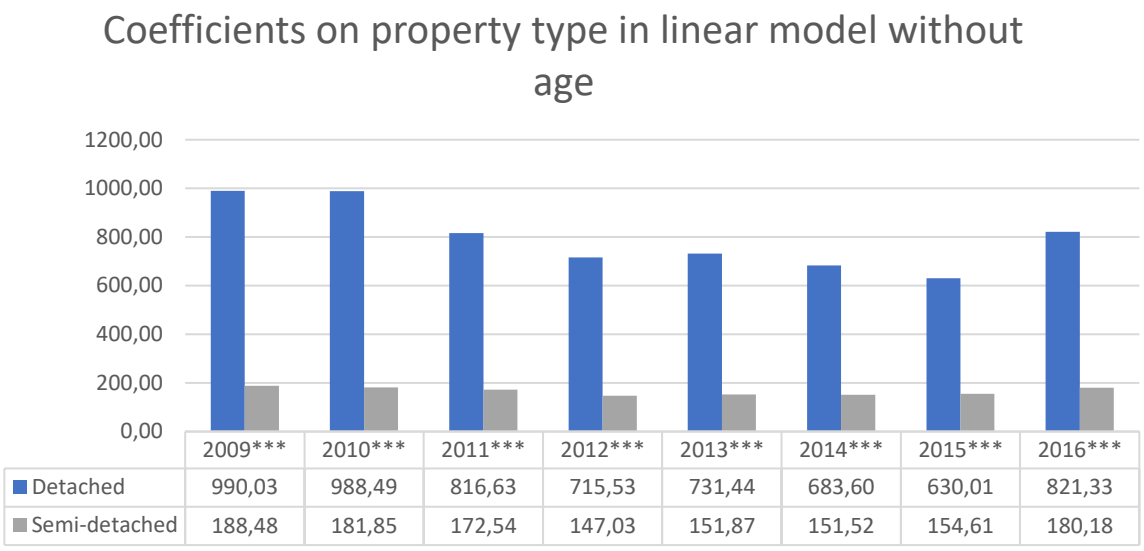


	2009***	2010***	2011***	2012***	2013***	2014***	2015***	2016***
Garage availible	141,99	208,85	146,65	195,98	102,62	150,28	220,71	134,25

Figure 12 coefficient and significance of availability of a garage over time

Figure 12 shows that in all years of analysis, the availability of a garage is significant and shows to add value to a property.

Property type

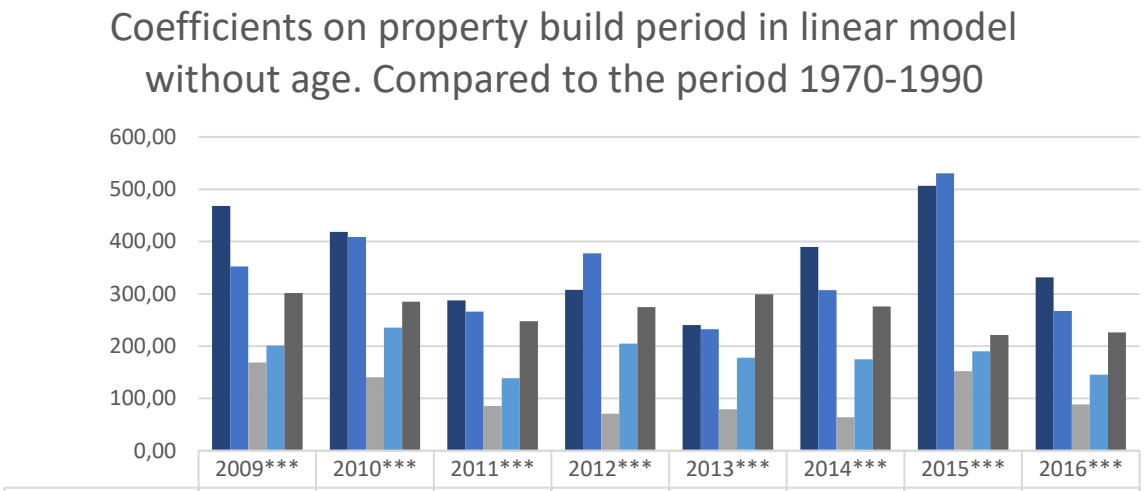


	2009***	2010***	2011***	2012***	2013***	2014***	2015***	2016***
Detached	990,03	988,49	816,63	715,53	731,44	683,60	630,01	821,33
Semi-detached	188,48	181,85	172,54	147,03	151,87	151,52	154,61	180,18

Figure 13 coefficient and significance of Detached and semi-detached properties over time (compared to terraced houses)

Figure 13 shows that both property types show a statistically significant addition of value when compared to the reference category that is terraced houses. A small decline in added value can be noted over the years, ending with a slight rise in added value.

Property age



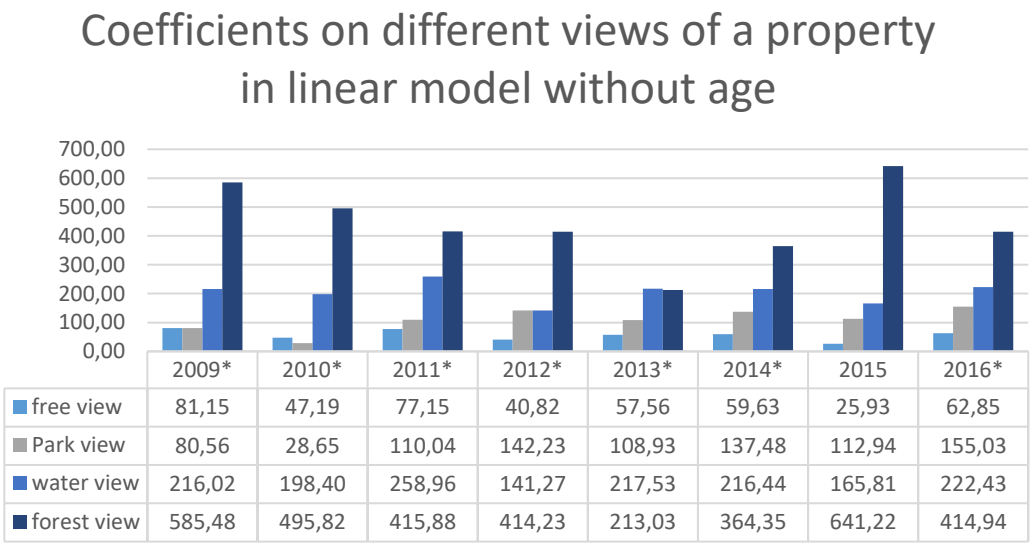
	2009***	2010***	2011***	2012***	2013***	2014***	2015***	2016***
Before 1906	468,07	418,57	287,54	307,56	240,42	389,67	506,80	331,39
Aged 1906-1944	352,03	408,81	265,92	377,40	232,66	306,87	530,13	267,06
Aged 1945-1970	168,61	140,50	85,51	70,56	79,18	63,91	152,00	88,47
Aged 1991-2000	201,03	235,36	138,96	204,74	177,97	175,15	190,22	145,63
After 2001	301,48	285,02	247,94	274,50	298,87	275,93	221,42	226,36

Figure 14 coefficients and significance of properties period over time (compared to properties built between 1970 and 1990)

The reference category for the variables in Figure 14 is the properties built between 1971 and 1990. The results show that for every timeframe, each variable is different enough relative to the reference category to be statistically significant and have a positive coefficient. Furthermore, the influence of these variables does not seem to change over the observed years. There is however a large difference between several age categories, where buildings in the categories before 1945 add more than twice the value that a property from the category 1945-1970 does. This observation is visible throughout all years.

Locational variables

Views

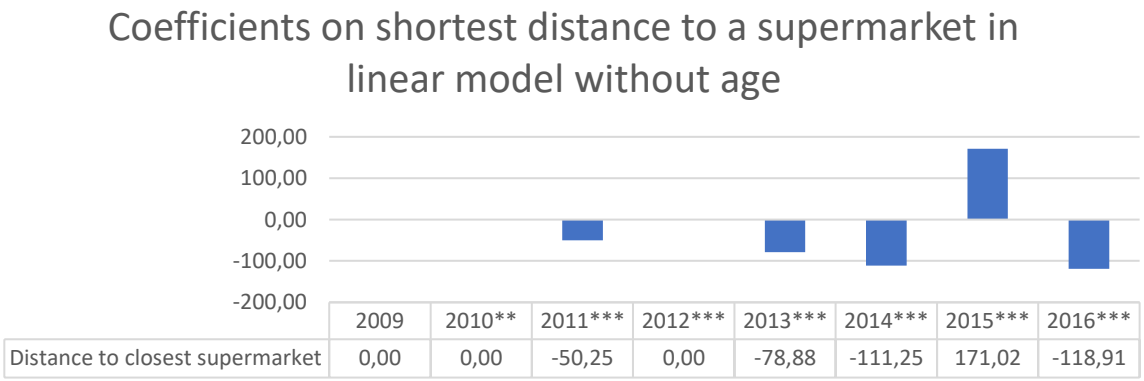


	2009*	2010*	2011*	2012*	2013*	2014*	2015	2016*
free view	81,15	47,19	77,15	40,82	57,56	59,63	25,93	62,85
Park view	80,56	28,65	110,04	142,23	108,93	137,48	112,94	155,03
water view	216,02	198,40	258,96	141,27	217,53	216,44	165,81	222,43
forest view	585,48	495,82	415,88	414,23	213,03	364,35	641,22	414,94

Figure 15 coefficients and lowest significance of properties classified with a type of special view over time (compared to properties without)

The outcomes of the variables on views indicate that in most years, all types are significant, and add value, as shown in Figure 15. Most years are more significant than $P \leq 0.05$, however for the accuracy of this graph, the highest level of significance that all variables share was used. The variable on the distance to the closest supermarket is significant in all years, except for 2009. In Figure 16 a predominantly negative trend can be observed, with a big positive influence shown in the year 2015. The high values for forest views can partly be caused by the low number of values this category has, making it more vulnerable to extreme values, as can be seen in Appendix B: Descriptive statistics on variables of interest.

Services



	2009	2010**	2011***	2012***	2013***	2014***	2015***	2016***
Distance to closest supermarket	0,00	0,00	-50,25	0,00	-78,88	-111,25	171,02	-118,91

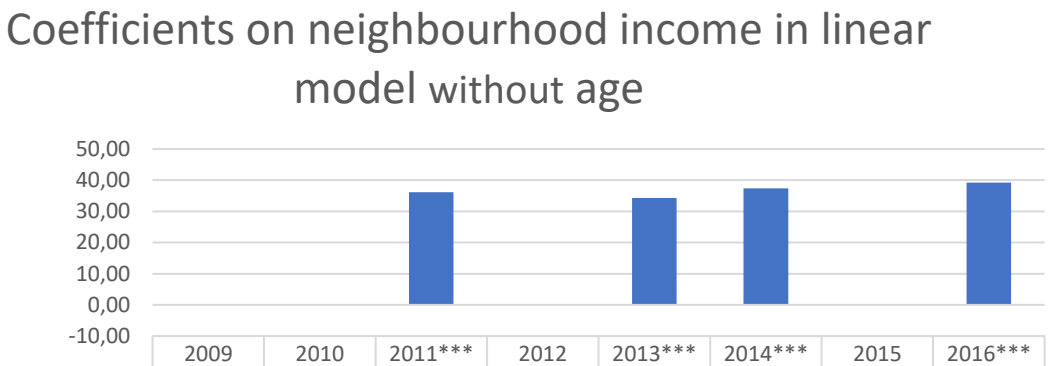
Figure 16 coefficient and significance of a property's proximity to a supermarket

Neighborhood variables

The average wealth of the neighborhood surrounding a property is showing mixed outcomes in Figure 17. For the years where this variable has shown to significantly impact the property price per square meter, a steady positive relationship has been established. Figure 15 seems to indicate a relation to the distance to supermarkets variable shown in Figure 16. Where a shorter distance to supermarkets decreases the property price, the price increases when the neighborhood income lever is higher. In Figure 18 it can be observed that sources of noise in the vicinity negatively influence the transaction price per meter. Although its influence seems to not be

very consistent throughout the years. When significant, the number of primary schools within an area seems to positively influence the dependent variable however, this variable is not significant during most years, as seen in Figure 19. Figure 20 shows that the result of the variable including crime statistics for the neighborhood is significant for all analyzed years. However, a higher number of crimes can be expected to lower the price of a property, whereas the analysis outcome shows the opposite.

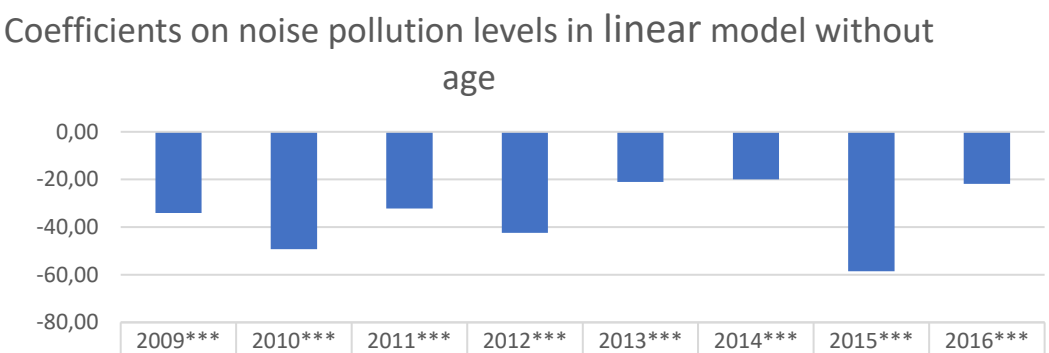
Neighborhood wealth



	2009	2010	2011***	2012	2013***	2014***	2015	2016***
Neighbourhood income	0,00	0,00	36,12	0,00	34,30	37,35	0,00	39,27

Figure 17 coefficient and significance of income level within the neighborhood

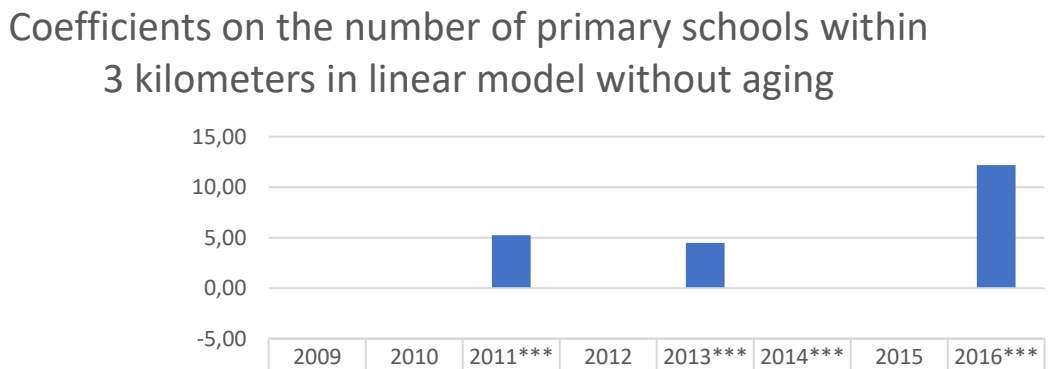
Potential noise pollution



	2009***	2010***	2011***	2012***	2013***	2014***	2015***	2016***
Noise pollution level	-34,09	-49,22	-32,20	-42,38	-21,09	-19,99	-58,54	-21,91

Figure 18 coefficient and significance of noise within the vicinity

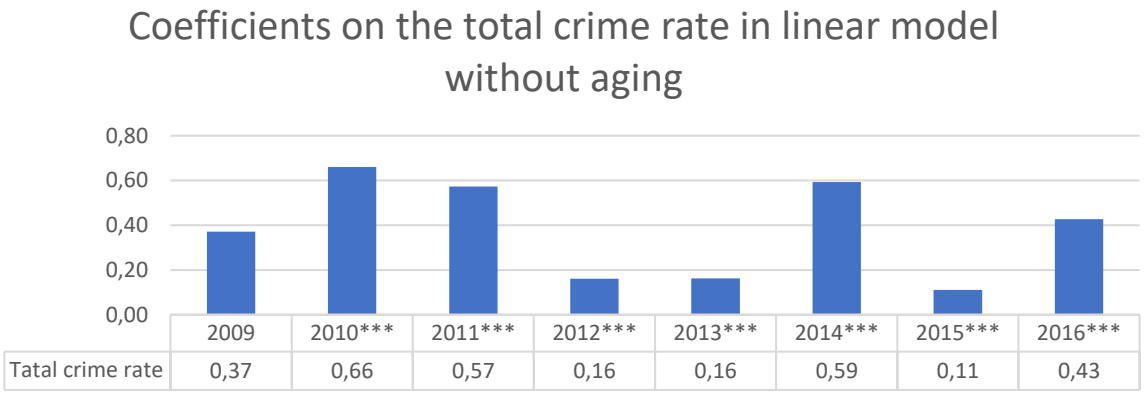
Availability of schools



	2009	2010	2011***	2012	2013***	2014***	2015	2016***
Primary schools within 3 kilometers	0,00	0,00	5,24	0,00	4,49	0,00	0,00	12,18

Figure 19 coefficient and significance of the number of primary schools within 3 kilometers

Crime statistics



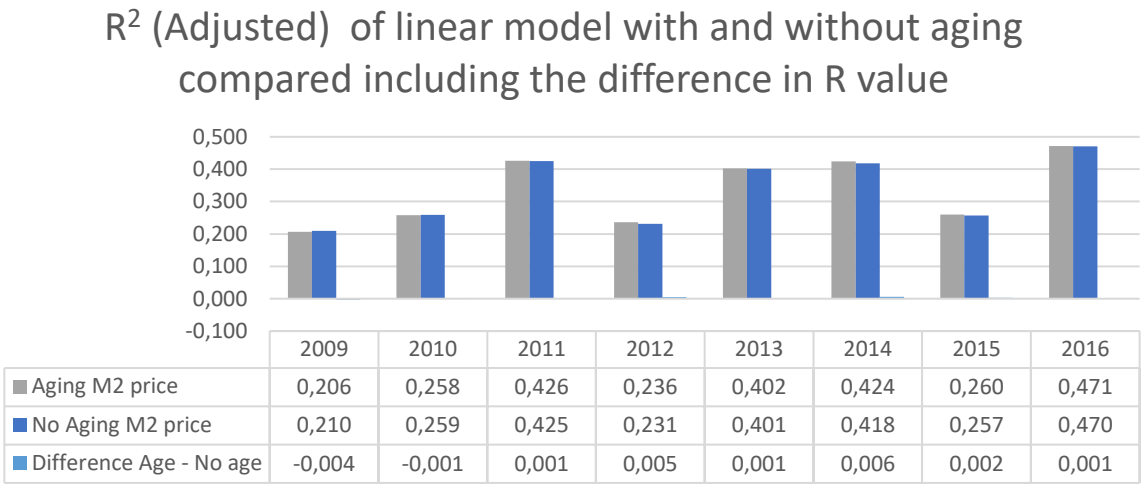
	2009	2010***	2011***	2012***	2013***	2014***	2015***	2016***
Tatal crime rate	0,37	0,66	0,57	0,16	0,16	0,59	0,11	0,43

Figure 20 coefficient and significance of crime rate within the neighborhood

Research question 2

To what extent does including local age structures lead to an improvement of the estimation of real estate pricing?

The answer to this research question includes two perspectives. First, a comparison between the model fit was made between the created linear regression models, one without age variable, the other including this variable. Secondly, the behavior of the variable coefficients over time was examined.



	2009	2010	2011	2012	2013	2014	2015	2016
Aging M2 price	0,206	0,258	0,426	0,236	0,402	0,424	0,260	0,471
No Aging M2 price	0,210	0,259	0,425	0,231	0,401	0,418	0,257	0,470
Difference Age - No age	-0,004	-0,001	0,001	0,005	0,001	0,006	0,002	0,001

Figure 21 R-squared values for both created models, including the difference between them

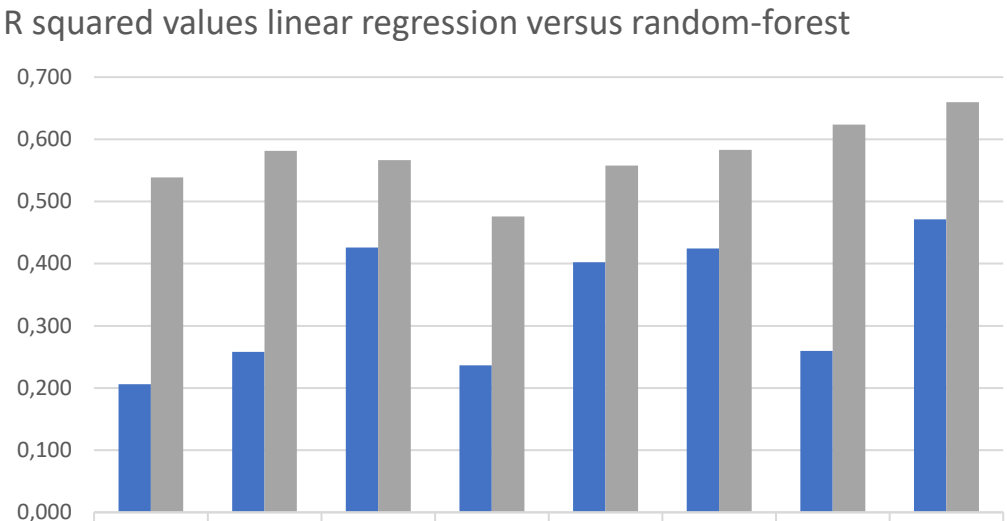
When looking at the adjusted R-squared statistic indicating the overall model fit the two created models can be compared. Figure 21 shows that not including the aging variable results in a slightly higher model fit for the years 2009 and 2010. Between 2011 and 2016, the opposite is true, although the increase in model precision is only fractions of percentages. This is in contrast to the theoretical basis for the assumption that including local age statistics can help explain the local property prices. The discrepancy between theory and practice can have three causes:

- The theory is incorrect
- The model variable is invalid
- The model itself is not capable of accurately modeling this variable

The discussion will look further into these possible explanations and what might create better results in the future.

Research question 3

To what extent does the inclusion non-linear variable fitting enhance the analysis result?



	2009	2010	2011	2012	2013	2014	2015	2016
Linear regression	0,206	0,258	0,426	0,236	0,402	0,424	0,260	0,471
forest classification	0,539	0,582	0,566	0,476	0,558	0,583	0,624	0,659

Figure 22 yearly R-squared values for the complete linear and forest regression models

The forest-based classification model was used to identify higher-order relationships instead of applying all data to the same fitting within a variable. Figure 22 shows that the created forest-based models for each year consistently score higher R-squared values than the general linear regression models.

To gain further understanding of the differences between these two models, their detailed outcomes should be compared. The forest regression calculates a value of importance for each variable as an absolute value, whereas the linear regression calculates both the coefficient and significance for each variable. These two values cannot directly be compared, however they are both related to their impact on the dependent variable. Because of this, an ordinal comparison was made to compare the outcomes of each model. Variables were ranked and sorted by coefficient and importance after which the lowest value would receive 0 points, adding one point to each more influential variable up to 19. These values have no meaning other than to indicate their influence on the dependent variable, where higher means more influence. Figure 23 shows the resulting graph. Showing that there is a difference in how both model types asses the same data.

Table 8 shows the actual coefficient and importance values. These cannot be compared between the models however, this table can be used to compare the relations between variables within the two models. For example, the variables “water view” and “aged 1991-2000” show comparable values in the linear model, however the importance factor of “aged 1991-2000” is almost

three times as high as “water view” in the forest model. This is an example of the different internal workings of both models.

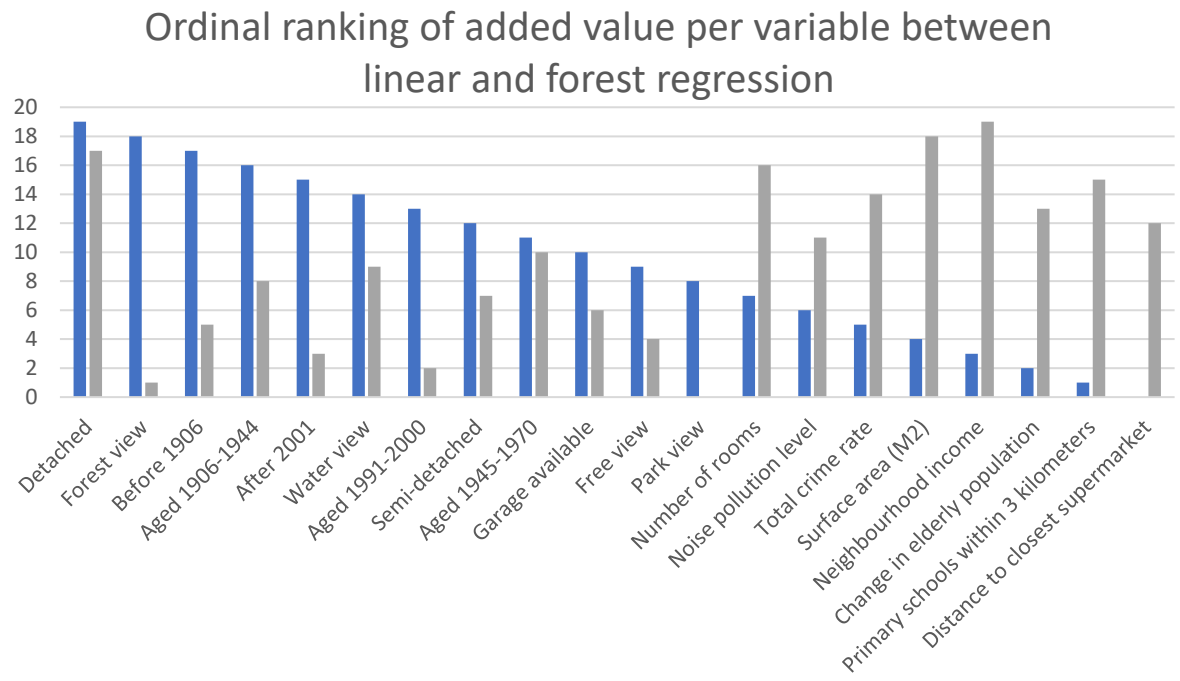


Figure 23 relative influence of each variable per model type for 2009, ranked 0 (lowest influence) to 19 (highest influence).

Table 8 influence scores in both models per variable

Variable	Linear coefficient	Forest importance value
Detached	982,80	414 640 693
Forest view	592,44	25 337 339,37
Before 1906	446,18	45 734 013,23
Aged 1906-1944	336,75	71 150 880,82
After 2001	289,75	35 151 496
Water view	217,64	77 413 920,33
Aged 1991-2000	200,57	28 336 083,01
Semi-detached	189,81	67 492 130,74
Aged 1945-1970	157,31	84 554 769,2
Garage availible	147,28	56 478 312,47
Free view	76,51	44 669 837,57
Park view	62,64	18 521 691,81
Number of rooms	33,16	364 295 661,2
Noise pollution level	-30,77	210 309 432,1
Tatal crime rate	0,88	307 408 972,4
Surface area (M2)	0,33	746 062 087,2
Neighbourhood income	-1E-06	963 844 126,2
Change in elderly population	0	301 002 537,9
Primary schools within 3 kilometers	0	330 615 558,7
Distance to closest supermarket	0	248 629 967,5

What can be observed from both

Table 8 and Figure 23 is that both models value different variables differently. What is interesting is that the random forest regression model consistently appears to value geographical variables higher than the linear regression. This had been part of the theorized effects beforehand and will be elaborated upon further in the discussion. Where in a lot of instances such as distance to supermarket or neighborhood income the results are contrasting, the building age period is valued relatively similar. In addition to this, the property type is also valued important in both types of model. This goes to show that although the influence of some variables can be disputed, there are several common factors that share their relevance between the two models.

Discussion

In general, this study has produced three models to represent the housing market. This was successfully done, and the explained variance is in line with existing research. An attempt to expand upon this model was made by looking into the local dynamics of the aging population, however this showed mixed results. This was further examined by applying a different method, showing interesting differences in results. A difference emerged between the relative contribution of geographical variables to the model as well as a higher explained variance. A timeframe of six consecutive years was examined. This would have allowed for the analysis of developing trends for different factors of influence. However no clear trends were discovered within the used data. Perhaps a larger timeframe or different ways of measuring will create more insights. Although the outcomes might not be perfect, a similar methodology, as well as comparable data to earlier studies, have been tested. This section will evaluate each variable outcome and seek to relate this to the existing literature.

Research question 1

How do structural and locational characteristics influence the price of real estate over time?

Structural variables

In general, and specifically for structural variables that are not related to living space, it is important to realize that preferences and valuations might vary over time and between nations (Chin & Chau, 2003). Because of this, the most value is placed on the comparisons between similar research projects within the Netherlands. As mentioned before, it has been found that floor area can contribute up to 40% of the explained variance within hedonic pricing models (Visser et al., 2008). However, the floor area used in this research is different in the way that it effectively explains the added value of another square meter on the price per square meter. This is providing a less outspoken influence as the absolute area of living is no longer taken into account. Conceptually it seems logical that as a property grows, the added value of each extra squared meter decreases, as it is adding less extra space relatively. This ties together with the number of rooms. Although the added value is relatively small, all years indicate a positive relationship to the transaction price per area. This does not directly coincide with the research of Visser et al (2008). However, that research specifically used bedrooms as a variable. It seems logical that adding more and more bedrooms to a household of a limited number of people does not necessarily add more value once every individual has one bedroom. The research of Yao (2016) that properties with three or fewer rooms are negatively correlated whereas seven to nine rooms was positively correlated, compared to the reference category of four to six rooms. This seems in line with the produced findings here, as well as the literature overview provided by Chin & Chau (2003) who conclude that any form of functional space will add value to a property. The argument made here in relation to Visser et al relates to functional space specifically, after a threshold, more bedrooms are no longer of use. It should be noted that although significant, both the floor area and the number of rooms do not add a lot of value when compared to other variables.

The availability of a garage does prove to consistently add a good amount of value to the property valuation. This can be related to the concept of functional space mentioned before, and is undisputed throughout the literature. The same goes for the type of property. The produced results on detached and semi-detached houses are fully in line with existing research (Chin & Chau, 2003; Visser et al., 2008).

For the variables on building age, all are significant for all years, indicating that any house is expected to be more valuable when it was not built between 1970 and 1990. This is not in line with either Visser at al. (2008) or Steegmans & Hassink (2017). Based on the literature it seems unlikely that properties from between 1970 and 1990 have the lowest prices. What might be skewing the results for the older properties is the possibility that only the ones of (historical) value are worth preserving leaving only more expensive properties to fill these samples.

Locational variables

The attributed views by the realtor appear to be significant and positive in various quantities. This does not differ from the observations by Chin & Chau (2003). However, the quality of this variable can depend on the input quality. Any entry into the Brainbay database is manual, however especially on values that require the valuation of the submitter, it is practically impossible to check the accuracy of any input and raises questions such as “is the vision of a creek a view on water or not?”.

The second locational variable: the distance to supermarkets largely shows a logical trend that implies people like necessary services to be close. The year 2015 shows an odd significant but positive trend, indicating that as properties were placed further away from the closest supermarket, they were valued higher per square meter.

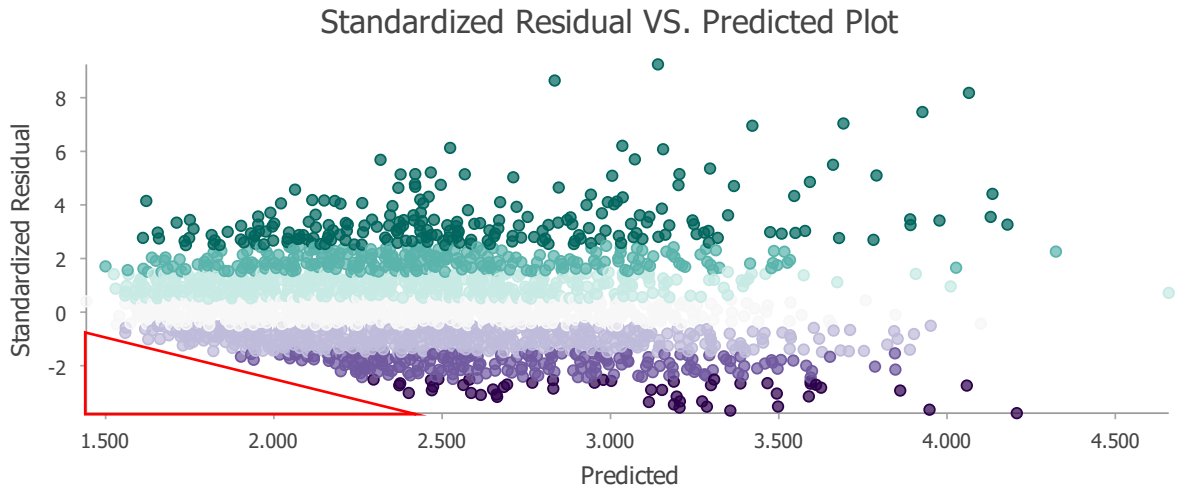


Figure 24 standardized residuals plotted against the predicted values for the year 2015

When observing the Standardized Residual vs Predicted graph for this year, many outliers, or a general lack of structure can be seen. A possible scenario in which this outcome could be expected is when in this period a lot of expensive, secluded properties have been sold. Despite being more remote, the price will still increase. Geographic modeling is expected to account for this on a local level, and thereby not skew the entire outcome due to a local cluster of global outliers.

For the measure of ease of access, the number of public transport stops within 400m would have made a good addition to the model. 400 meters is a reasonable estimation of when a public transport stop would still be considered for traveling as this is roughly a 5-minute walk (Bolten, 2020). However due to the absence of available data for any period this variable has not been taken into account.

Neighborhood variables

The outcome of the neighborhood wealth variable shows mixed results. In theory, it makes sense to assume that people with a high income would live in more expensive places. This is also the observation by the existing studies (Chin & Chau, 2003; Steegmans & Hassink, 2017; Visser et al., 2008). This is however not the picture drawn by our data. Part of the outcome can be explained by invalid data. Although the Brainbay data has been cleaned from unrealistic and null values as good as possible, due to privacy concerns the CBS will not publish their data when an area is too scarcely populated, causing occasional data absence. Another possible error could have occurred when joining the CBS data to the Brainbay listings. Either the spatial join or the regular join could have resulted in mismatches. On a large scale, this however seems unlikely, as this would have caused errors in the analyses due to null values on a large scale. A somewhat logical graph can be seen when only looking at the significant years, in line with existing findings however, no explanation can be found for the low values and insignificance of the remaining years.

The variable on noise pollution has produced clear and logical results, producing negative coefficients. Although Visser et al mentioned that proximity to rail or highways can be viewed as both

beneficial in terms of accessibility, as well as a source of disturbance. An interesting point referenced to in the article by Chin & Chau is that the proximity to churches also plays a part in this as a source of sound and activity. This could be examined in further research (Do et al., 1994).

The number of primary schools within three kilometers shows to be significant and adding positive value about half the time. The mixed results can theoretically be explained by the limited influence that schools have on for example elderly. In addition to this, proximity to schools might come with added activity and disturbance. Clark & Herrin (2000) have shown that the quality of schools (in California) is of great importance to the impact a school can have. This study does not classify schools in any way, and therefore cannot be expected to show a similar effect. By interfering with the building density variable, the number of schools within 3 kilometers is also an indicator of the local degree of urbanization. After all, a small village of 100 inhabitants has no place for multiple schools, just like a dense urban district needs to be able to offer education to all inhabitants. Ideally, the OAD for the degree of urbanization would have been directly for this, however as mentioned before this variable caused the most interference with the other data and therefore was removed.

The final variable outcome is that of the crime statistics. Logically speaking, people would be thought to avoid areas where there is a higher risk of any type of criminality. The literature analyzed by Chin and Chau (2003) indicated that there are evident negative relationships between criminality statistics and property price. However, a more recent paper by Yao & Stewart Fotheringham (2016) manages to find significance in only one of the nine observed years. The outcome produced in this research is different than previous findings. A possible explanation for the positive observed relationship is that more expensive neighborhoods are more vulnerable to (certain types of) crimes. The linear regression statistics only indicate the existence of a relationship but does offer any insights into causality or any further connections that might explain this observation. It should be noted that certain topics such as crime or quality of schools do not follow globally uniform descriptions and therefore literature on other regions might show different relationships.

Linear regression models

This section will evaluate the quality of the models produced for this research question and seek to explain any irregularities within the presented results. Because all three models have produced outcomes with highly similar implications, not each one will be shown individually here. All individual outputs can be found in the folder delivered with this document.

The decision to use the robust R^2 is supported by the significance of the “Koenker (BP) Statistic”. This statistic is used to measure the likelihood that the modeled relationships are consistent across the study area. The fact that the Koenker statistic is significant indicates that not all variables behave the same throughout the area.

Table 9 linear regression model diagnostics for 2009 without age (same as Table 7)

Input Features:	Complete_2009__XYTableToPoint	Dependent Variable:	OBJ_HID__TRANSAC_M2
Number of Observations:	9362	Akaike's Information Criterion (AICc) [d]:	148535,4687
Multiple R-Squared:	0,207844	Adjusted R-Squared [d]:	0,206148
Joint F-Statistic:	122,543791	Prob(>F), (20,9341) degrees of freedom:	0,000000*
Joint Wald Statistic:	1577,368544	Prob(>chi-squared), (20) degrees of freedom:	0,000000*
Koenker (BP) Statistic:	116,551375	Prob(>chi-squared), (20) degrees of freedom:	0,000000*
Jarque-Bera Statistic:	17056230,76	Prob(>chi-squared), (2) degrees of freedom:	0,000000*

Table 9 indicates that the modeled relationships are not consistent throughout the area. This, in combination with the significant Jarque-Bera statistic that indicates that the model residuals are not normally distributed, suggests that the modeled relationships are not perfectly resembled by the current linear regression model. To further investigate this assumption the created Standardized Residual vs Predicted values Plot is examined in Figure 24.

The explanation for the empty triangle around x=2000 shown in red can most likely be found in the preprocessing. During the data preparation, records with a transaction price under a hundred thousand euros have been removed. Assuming these records do not have exceptional properties heavily increasing their price per m^2, these properties would have been able to contribute to the properties that would have been estimated around X=2000 but turned out to have a lower price. Therefore, eliminating this group causes the empty space near the origin of the graph.

The gradual widening of the value spread along the x axis can be explained by the logic that the properties that have the lowest price per square meter will generally be more uniform than properties of a higher value per area, which can present themselves as unique in more various ways.

What can also be of influence here is the effect of geographic location. Residual values can be expected to rise when all attributes of a property are equal, yet one is in a desirable location whereas the other is not. This can clearly be seen in Figure 25.

A possible explanation for the regional differences can be found in the different buildup of each city's internal market. A recent report by the Dutch national bank shows that the Hague contains much less social housing, and more privately owned properties as compared to Rotterdam. The share of privately owned properties can be of influence to the local housing price, explaining the higher prices in the Hague (Nijskens, Lohuis, Hilbers, & Willem, 2018). Another influence might be the international exposure of the Hague as a city with a lot of international institutions attracting more financially capable ex-pats.

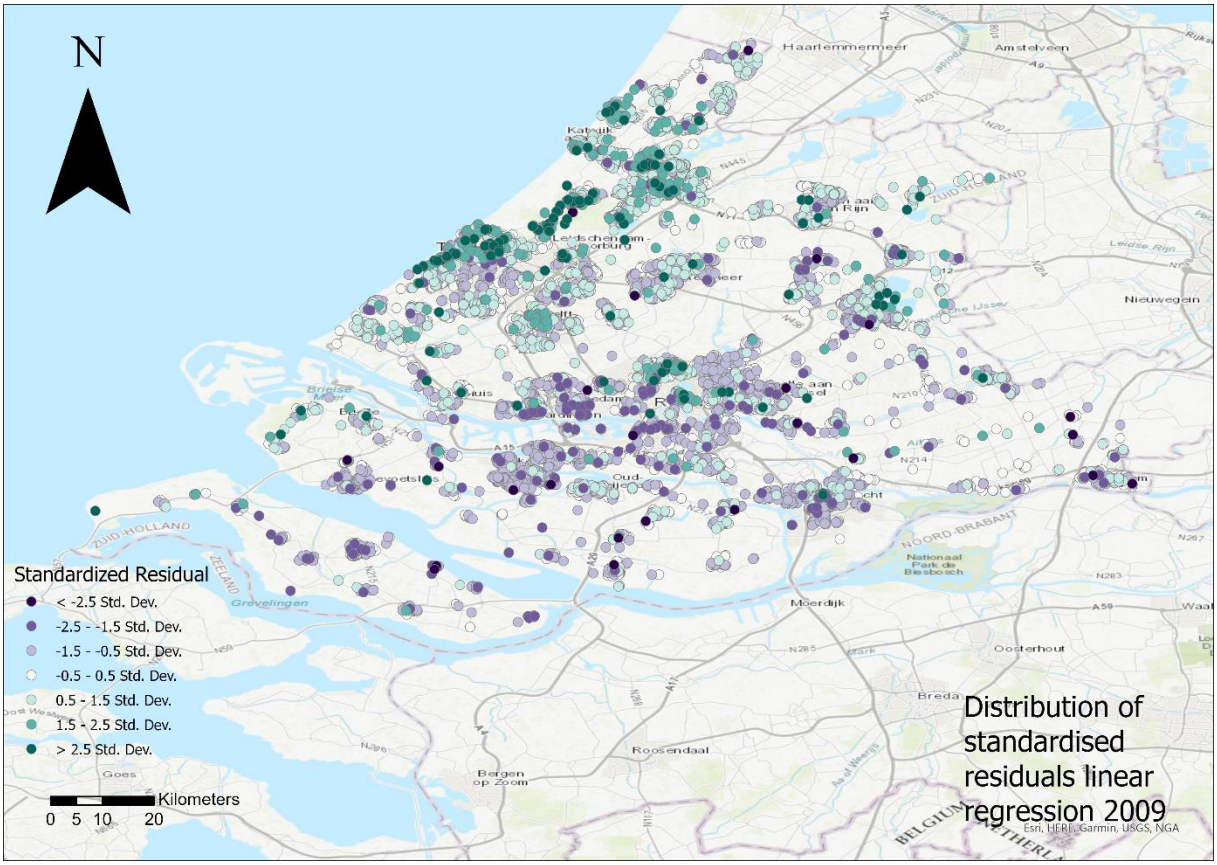


Figure 25 map of the distribution of standardized residuals for the linear regression of 2009 (same as Figure 9)

Research question 2

To what extent does including local age structures lead to an improvement of the estimation of real estate pricing?

Outcomes in context

In terms of overall model performance, the linear regression model that includes the variable on age dynamics has shown to add, although only slightly, a better fit. In this case, the application of the adjusted R^2 is noticed. When adding any variable to a model, no matter the explanatory capabilities, the R^2 will go up. This is no problem when all variables do indeed add to the quality of the model however when they do not, this does not accurately portray the model quality. As the adjusted R^2 is corrected for the interference of irrelevant variables, the two models can be compared to each other (Leach et al., 2007).

The behavior of the variable itself does indicate a different working than what was to be assumed based on the theory. When significant, mostly the variable is negatively related to the price per square meter. This translates to a lower transaction price when the increase of elderly in the neighborhood rises. Therefore, a declining elderly population would result in higher prices. This attribute has not been studied widely before, therefore no outcomes can be compared, except for the theorized mechanism that is not in line with the results produced here (Hennadige, 2016).

Age development variable

The variable itself has been constructed based on the existing theories on dynamics and implications of demographic changes. Conditions that would have to be considered based on the existing literature are:

- The demographic group of interest is part of the entire population, both of which are in motion. This means that if the elderly population doubles, but the entire population doubles as well, the percentage of elderly stays the same.
- For this topic, the movement of the elderly is more important than the total numbers at a given moment. When applying the principles of supply and demand, a single value does not carry meaning without a value to relate it to. In essence, this applies in the hypothetical situation where 70% of an area consists of people over 65, and the average price in this area is twice the global average. The conclusion can be drawn that the strong presence of elderly positively impacts the price. However, this situation does not provide any additional information on trends. What if the average price used to be quadruple the global average 10 years ago when the percentage of elderly was only 5%. Then the conclusion that the relative increase in the elderly population negatively influences the local transaction price seems more logical. This shows that a timeframe is needed for a necessary reference of ongoing trends.

Within the schedule of this project, no perfect measure has been found to depict these dynamics accurately and truthfully. For this project using the percentual change in relative demographic share was considered. However, Because of the abstract values of this measure and the uncommon method of working with percentages of percentages, to not lose the connection to the source data, this method was not pursued. Instead, the more basic method of calculating the percentual increase in population over 65 has been used. This incorporates the time frame; However, it does not account for the other demographics within the same area. Upon further research, a more suitable variable unit should be developed. As mentioned before, Fory (2014) notes that this process might be not so strictly related to age, rather than the more abstract “stage of life”. This can explain why this variable can be difficult to study.

Research question 3

To what extent does the inclusion of non-linear variable fitting enhance the analysis result?

Outcomes in context

The presented outcomes show that the random forest classification will consistently outperform a linear regression model. This indicates that the classification by decision trees in combination with the non-linear fitting methods does aid the explanation of local property prices. This is especially visible in the locational-related variables as seen in Figure 23**Fout! Verwijzingsbron niet gevonden.**. This makes sense as these variables are most likely to show nonlinear relationships. When making the tradeoff between access to a highway, and disturbance by this road, a parabolic relationship is more likely than a uniform sloping line that can be thought of when assuming that more space will virtually always lead to a higher valuation.

Many studies across different disciplines have compared the workings of multivariable linear regression to random forest classifications over time, with mixed conclusions. When applied to neuroscience, the linear regression proved to be more capable of accurately predicting values whereas the random forest classification had shown a higher accuracy when modeling hydrology (Ouedraogo et al., 2019; Smith, Ganesh, & Liu, 2013). Next to the fitting of known data, both methods can also perform differently when using them to predict values (Yuchi et al., 2019). This thesis only examined the capability to fit the data properly. Figure 26 shows a side-by-side comparison between both methods.

Where the distribution of the linear regression values has previously been discussed in the discussion on research question one, the observed divide appears to have decreased somewhat but is still evident in the random forest model outcome. The number of extreme values appears to have decreased, especially in the urban centers such as the city of Rotterdam and the Hague/Scheveningen. This can be explained by the observation that geographical factors are of more value than assumed by the linear model, as distances in densely populated areas can generally be expected to be lower than in the countryside. Therefore, an equal increase in distance will have a relatively much larger effect on parameters in city centers, whereas an extra kilometer in rural areas might not make a significant difference on the existing travel distances.

This question has shown that the traditional way of modelling real estate is not necessarily the best, and that a different approach can both create better results as well as new insights into the processes being modelled. In this case the difference in the importance of locational variables stands out and might be more important that previously assumed.

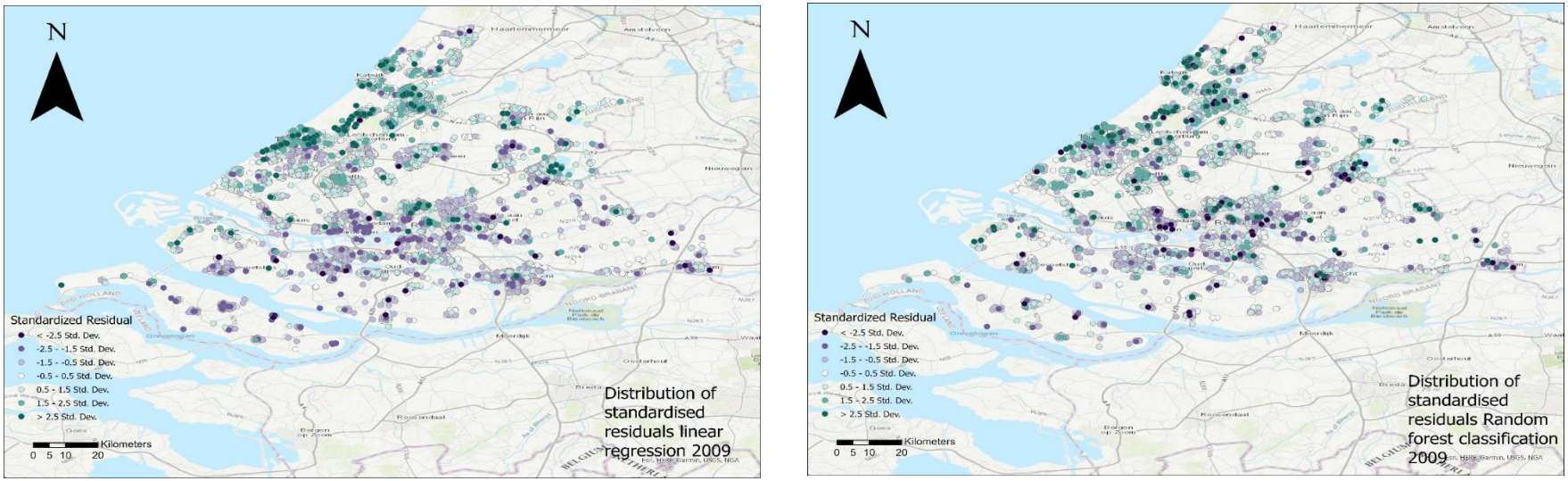


Figure 26 spatial distribution of residuals of the linear regression model (left) and the random forest regression model (right)

STANDARD DEVIATION RAN	Linear regression count	Random forest count
(-2,50 - -1,50 Std. Dev.)	195	172
(-1,50 - -0,50 Std. Dev.)	2.109	1.515
(-0,50 - 0,50 Std. Dev.)	5.168	6.088
(0,50 - 1,50 Std. Dev.)	1.446	1.189
(>1,50 Std. Dev.)	419	336
(< -2,50 Std. Dev.)	25	62

Table 10 point counts for the spread of residual values shown in Figure 26

Conclusions

Three research gaps were identified based on existing literature. First, most studies only investigate either spatial or temporal influences within hedonic price modelling. This study has done both, finding that the temporal dimension does not show a lot of development within this timeframe. However, this does not mean that there is no effect. The conclusion that lower priced properties have a higher certainty in the linear models makes for interesting new research opportunities, as housing prices have risen a lot in recent years. Secondly the spatial aspect of this research gap has shown to be relevant. The random forest model indicates that spatial variables might be more important that previously assumed in the linear regression. This relates to the second research gap, proving that linear relations are not necessarily a given in hedonic modelling. The random forest regression shows that especially locational variables profit from the decision tree based approach, improving overall model accuracy. The third observation from the body of knowledge consists of the approach to variable selection. This research has attempted to select variables based on theories, and proven by existing research. To expand on this a variable was added related to an aging population. The theory that this helps explain property prices shows little proof in practice. Further research is necessary to further explore the existence of this relationship.

Research questions

When looking at research question 1 "*How do structural and locational characteristics influence the price of real estate over time?"* The results show that the measured variables show different relationships to the transaction price per square meter. These findings are largely in line with existing research. Figure 23 shows the observed variables ranked from largest influence to lowest with their corresponding impact on the transaction price per meter for the year 2009. Performing the analysis over several years has not resulted in the identification of clear trends or other developments within the timescale.

Expanding upon the first research question "*To what extent does including local age structures lead to an improvement of the estimation of real estate pricing?"* has led to the conclusion that the theorized positive relationship between change in local age structure and the price of real estate has not been proven. The inclusion of this variable resulted in a marginally improved adjusted R^2 statistic. It appears that the relationship between change in the number of people over 65 and transaction price would more likely be negative than positive as was the expectation. The used variable has been found not to include all factors that would be of influence. There is literature supporting the theory that this factor can be of influence; therefore, further research is suggested. For further research, a more suitable measure for change in population buildup would have to be found.

In conclusion, when looking beyond the R^2 statistics, a small group of outliers as well as the (spatial) distribution of the data, cause the linear regression models to show that it might not be the most suitable method for modeling the price of properties per square meter. As not all assumptions can be met concerning the normal distribution of residuals and effects by outliers. In addition to this, Figure 9 shows that the linear regression analysis is unable to model local variations properly.

To answer the question "*To what extent does the inclusion non-linear variable fitting enhance the analysis result?"* the outcomes of the random forest classification to the linear regression model outcomes were compared. The explained variance of the random forest classification was significantly higher. In this case, the random forest classification has been shown to be more accurate in modeling the price per square meter throughout the province. When examining the results on the effect each variable has on the model, the two methods appear to value the independent variables differently as seen in Figure 23. This is reason to study both models to understand their inner workings and to decide on which might be more suitable for future use. Figure 26 However shows that although an improvement, the random forest regression still is not able to fully account for local deviations.

These observations have helped answer the overarching general research question ““To what extent can the changing local age structure, next to structural and locational characteristics, be used to explain the evolution of local real estate prices, and to what extent does the conventional method suit spatial relationships” be answered.

The first part of this question is about the explanatory value of the local changing age structure. This value is low, but provides a starting point for further research, but on the temporal aspect as well as the measuring of the data on population demographics.

The second half of the question, questions current methods and was tested by comparing the linear regression model to the random forest regression. It was shown that a linear regression analysis might not be the most suitable analysis for this data. As presented alternative method, the random forest classification has shown to be more effective in modeling transaction prices when compared to the linear regression model. This can be attributed to the different possible relationships that it can model next to solely linear relations drawn by the regression analysis. This appears to be most beneficial for the accurate modeling of variables that contain a geographical or distance aspect.

Implications

This research has shown that random forest modelling improves the accuracy of hedonic price modelling when compared to using a linear regression. And that locational variables have a larger part in this than shown under linear regression. This can help policy makers and planning institutions to better design plans that meet the local demand. Furthermore, increased transparency of the housing market can make it easier for people to find a living space. As a necessity in life, knowing about the financial implications of preferences and influences on your desired property makes it easier to know what to look for.

References

An, C. B., & Jeon, S. H. (2006). Demographic change and economic growth: An inverted-U shape relationship. *Economics Letters*, *92*(3), 447–454. https://doi.org/10.1016/j.econlet.2006.03.030

Ando, A., & Modigliani, F. (1963). The " Life Cycle " Hypothesis of Saving : Aggregate Implications and Tests. *The American Economic Review , Mar ., 1963 , Vol . 53 , No . 1 , Part 1 (Mar ., 1963), Published by : American Economic*, *53*(1), 55–84.

Benson, E. D., Hansen, J. L., Schwartz, A. L., & Smersh, G. T. (1998). Pricing Residential Amenities: The Value of a View Sinkholes and Residential Property Prices: Proximity and Density View project Retirement View project Pricing Residential Amenities: The Value of a View. *Journal of Real Estate Finance and Economics*, *16*(1). https://doi.org/10.1023/A:1007785315925

Bisello, A., Antoniucci, V., & Marella, G. (2020). Measuring the price premium of energy efficiency: A two-step analysis in the Italian housing market. *Energy and Buildings*, *208*, 109670. https://doi.org/10.1016/j.enbuild.2019.109670

Bolten, N. (2020). Equitable Network Modeling of Diverse Modes of Built Environment Pedestrian Navigation.

Bottero, M., Bravi, M., Marmolejo, C., Method, H. P., Model, A., & Model, S. E. (2019). *How the Impact of Energy Performance Certificates Differs in Two European Climatic Zones*. (July), 24–26.

Breiman, L. (2001). Random forests. *Random Forests*, 1–122. https://doi.org/10.1201/9780429469275-8

Cai, J., Xu, K., Zhu, Y., Hu, F., & Li, L. (2020). Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Applied Energy*, *262*, 114566. https://doi.org/10.1016/j.apenergy.2020.114566

Chau, K., Ng, F., Journal, E. H.-T. A., & 2001, undefined. (n.d.). Developer's good will as significant influence on apartment unit prices. *Search.Proquest.Com*. Retrieved from http://search.proquest.com/openview/f8e2fbe758b7854f64971914af490e97/1?pq-origsite=gscholar&cbl=35147

Chin, T. L., & Chau, K. W. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications*, *27*(2), 145–165.

Clark, D. E., & Herrin, W. E. (2000). The impact of public school attributes on home sale prices in California. *Growth and Change*, *31*(3), 385–407. https://doi.org/10.1111/0017-4815.00134

de Gans, H. A., & Oskamp, A. (1992). FUTURE HOUSING NEED IN THE RANDSTAD HOLLAND : A MATTER OF CONTINUING INDIVIDUATION? *Netherlands Journal of Housing and the Built Environment*, *7*(2), 157–178.

Deepnote. (2021). datascience 18: machine learning with tree-based models in python – deepnote. Retrieved February 22, 2021, from http://deepnote.me/2019/08/25/datascience-18-machine-learning-with-tree-based-models-in-python/

Do, A. Q., Wilbur, R. W., & Short, J. L. (1994). An empirical examination of the externalities of neighborhood churches on housing values. *The Journal of Real Estate Finance and Economics*, *9*(2), 127–136. https://doi.org/10.1007/BF01099971

Edmonds, R. G. (1984). A Theoretical Basis for Hedonic Regression: A Research Primer. *Real Estate Economics*, *12*(1), 72–85. https://doi.org/10.1111/1540-6229.00311

Evangelista, R., Ramalho, E. A., & Andrade e Silva, J. (2020). On the use of hedonic regression models to measure the effect of energy efficiency on residential property transaction prices: Evidence for Portugal and selected data issues. *Energy Economics*, *86*, 104699. https://doi.org/10.1016/j.eneco.2020.104699

Ford, R., & Jennings, W. (2020). The Changing Cleavage Politics of Western Europe. *Annual Review of Political Science*, *23*, 295–314. https://doi.org/10.1146/annurev-polisci-052217-104957

Fory, I. (2014). SELECTED DEMOGRAPHIC ASPECTS OF BUYERS '. 22(4), 92–104.

Garrod, G. D., & Willis, K. G. (1992). Journal of Environmental Management (1992) 34, 59-76 Valuing Goods' Characteristics: an Application of the Hedonic Price Method to Environmental Attributes G. D. Garrod and K. G. *Journal of Environmental Management*, *34*, 59–76.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *American Statistician*, *63*(4), 308–319. https://doi.org/10.1198/tast.2009.08199

Helbich, M., Brunauer, W., Vaz, E., & Nijkamp, P. (2014). Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria. *Urban Studies*, *51*(2), 390–411. https://doi.org/10.1177/0042098013492234

Hennadige, W. T. (2016). DEMOGRAPHICS AND ASSET PRICES IN AUSTRALIA.

Hochstenbach, C., & Arundel, R. (2020). Spatial housing market polarisation: National and urban dynamics of diverging house values. *Transactions of the Institute of British Geographers*, *45*(2), 464–482. https://doi.org/10.1111/tran.12346

Inchauste, G., Karver, J., Kim, Y. S., & Abdel Jelil, M. (2018). Living and Leaving. *Living and Leaving*, (November 2018). https://doi.org/10.1596/30898

Journal, S., Statistical, A., & Jun, N. (2009). Measuring the Value of Housing Quality Author (s): John F . Kain and John M . Quigley Published by : American Statistical Association Stable URL : http://www.jstor.org/stable/2284565. *Quality*, *65*(330), 532–548.

Ketkar, K. (1992). Hazardous waste sites and property values in the state of New Jersey. *Applied Economics*, *24*(6), 647–659. https://doi.org/10.1080/00036849200000033

Kohlhase, J. E. (1991). The impact of toxic waste sites on housing values. *Journal of Urban Economics*, *30*(1), 1–26. https://doi.org/10.1016/0094-1190(91)90042-6

Lateef, Z. (2021). Complete Tutorial On Random Forest In R With Examples | Edureka. Retrieved February 22, 2021, from https://www.edureka.co/blog/random-forest-classifier/

Leach, L. F., Henson, R. K., Finch, W. H., Fraas, J. W., Newman, I., & Walker, D. A. (2007). *Multiple Linear Regression Viewpoints Volume 33 • Number 1 • Fall*. *33*(1). Retrieved from http://www.glmj.org/archives/MLRV_2007_33_1.pdf#page=4

Li, L. (2019). Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. *Remote Sensing*, *11*(11). https://doi.org/10.3390/rs11111378

Li, M. M., & Brown, H. J. (1980). Micro- neighborhood externalities and hedonic housing prices. *Land Economics*, *56*(2), 125–141. https://doi.org/10.2307/3145857

Li, S., Ye, X., Lee, J., Gong, J., & Qin, C. (2017). Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective. *Applied Spatial Analysis and Policy*, *10*(3), 421–433. https://doi.org/10.1007/s12061-016-9185-3

Long, R., & Wilhelmsson, M. (2020). Impacts of shopping malls on apartment prices : The case of Stockholm.

Lorenz, D. P., Lützkendorf, T., & Trück, S. (2007). Exploring the relationship between the sustainability of construction and market value: Theoretical basics and initial empirical results from the residential property sector. *Property Management*, *25*(2), 119–149. https://doi.org/10.1108/02637470710741506

Lytvynchenko, G. (2014). Programme Management for Public Budgeting and Fiscal Policy. *Procedia - Social and Behavioral Sciences*. https://doi.org/10.1016/j.sbspro.2014.03.064

Marktcijfers koopwoningen | NVM. (n.d.). Retrieved April 12, 2021, from https://www.nvm.nl/wonen/marktinformatie/

Missie en visie - Provincie Zuid-Holland. (n.d.). Retrieved November 10, 2020, from https://www.zuid-holland.nl/overons/missie-visie/

Mulder, K. F. (2016). *Our Common City, het metabolisme van de stad*. 1–41. Retrieved from https://www.researchgate.net/publication/309904632_Our_Common_City_het_metabolisme_van_de_stad

Nijskens, R., Lohuis, M., Hilbers, P., & Willem, H. (2018). Hot Property: The Housing Market in Major Cities. In *Springer* (Vol. 206). Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-3-030-11674-3.pdf

Osland, L. (2010). An application of spatial econometrics in relation to hedonic house price modeling. *Journal of Real Estate Research*, *32*(3), 289–320. https://doi.org/10.1080/10835547.2010.12091282

Ouedraogo, I., Defourny, P., & Vanclooster, M. (2019). Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeology Journal*, *27*(3), 1081–1098. https://doi.org/10.1007/s10040-018-1900-5

Randeniya, T., Ranasinghe, G., & Amarawickrama, S. (2017). A model to Estimate the Implicit Values of Housing Attributes by Applying the Hedonic Pricing Method. *International Journal of Built Environment and Sustainability*, *4*(2), 113–120. https://doi.org/10.11113/ijbes.v4.n2.182

SAMENWERKEN MAAKT STERKER | MRDH. (n.d.). Retrieved November 9, 2020, from https://mrdh.nl/

Shin, M.-C., Shin, G.-M., & Lee, J.-S. (2019). The Impacts of Locational and Neighborhood Environmental Factors on the Spatial Clustering Pattern of Small Urban Houses: A Case of Urban Residential Housing in Seoul. *Sustainability*, *11*(7), 1934. https://doi.org/10.3390/su11071934

Silva, M. S. L. (2017). Study on Demographic Background of Potential Buyers: With Reference to Luxury Condominium Apartments in Colombo. *SSRN Electronic Journal*, (January). https://doi.org/10.2139/ssrn.2909747

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*, *220*(1), 85–91. https://doi.org/10.1016/j.jneumeth.2013.08.024

So, H. M., Tse, R. Y. C., & Ganesan, S. (1997). Estimating the influence of transport on house prices: evidence from Hong Kong. *Journal of Property Valuation and Investment*, *15*(1), 40–47. https://doi.org/10.1108/14635789710163793

Sopranzetti, B. J. (2010). Hedonic Regression Analysis in Real Estate Markets: A Primer. In *Handbook of Quantitative Finance and Risk Management* (pp. 1201–1207). https://doi.org/10.1007/978-0-387-77117-5_78

StatLine - Bevolking; kerncijfers. (n.d.). Retrieved August 24, 2020, from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37296ned/table?fromstatweb

StatLine - Regionale prognose 2017-2040; bevolking, intervallen, regio-indeling 2015. (n.d.). Retrieved August 23, 2020, from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83491NED/line?dl=F8E9&ts=1598197553620

Steegmans, J., & Hassink, W. (2017). Financial position and house price determination: An empirical study of income and wealth effects. *Journal of Housing Economics*, *36*, 8–24. https://doi.org/10.1016/j.jhe.2017.02.004

Sulamoyo, D. S. (2016). Preparing for the future. *Community Action Leaders: Rooting Out Poverty at the Local Level*, 211–236. https://doi.org/10.4324/9781315563497

Thompson, C. G., Kim, R. S., Aloe, A. M., & Becker, B. J. (2017). Extracting the Variance In flation Factor and Other Multicollinearity Diagnostics from Typical Regression Results. *Basic and Applied Social Psychology*, *39*(2), 81–90. https://doi.org/10.1080/01973533.2016.1277529

Toegevoegde waarde en werkgelegenheid | Haven van Rotterdam. (n.d.). Retrieved November 9, 2020, from https://www.portofrotterdam.com/nl/onze-haven/feiten-en-cijfers/feiten-en-cijfers-over-de-haven/toegevoegde-waarde-en-werkgelegenheid

van Geuns, R. (2020). Worden de rijken rijker en de armen armer? *Sociaal Bestek*, *82*(2), 4–7. https://doi.org/10.1007/s41196-020-0659-0

Vishnevsky, A., & Shcherbakova, E. (2018). A new stage of demographic change: A warning for economists. *Russian Journal of Economics*, *4*(3), 229–248. https://doi.org/10.3897/j.ruje.4.30166

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

von Graevenitz, K., & Panduro, T. E. (2015). An alternative to the standard spatial econometric approaches in hedonic house price models. *Land Economics*, *91*(2), 386–409. https://doi.org/10.3368/le.91.2.386

Yao, J., & Stewart Fotheringham, A. (2016). Local Spatiotemporal Modeling of House Prices: A Mixed Model Approach. *Professional Geographer*, *68*(2), 189–201. https://doi.org/10.1080/00330124.2015.1033671

Yoo, S., Im., J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, *107*(3), 293–306. https://doi.org/10.1016/j.landurbplan.2012.06.009

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., … Allen, R. W. (2019). Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environmental Pollution*, *245*, 746–753. https://doi.org/10.1016/j.envpol.2018.11.034

Yusof, A., & Ismail, S. (2012). Multiple Regressions in Analysing House Price Variations. *Communications of the IBIMA*, *2012*, 1–9. https://doi.org/10.5171/2012.383101

Zhang, Y., Jin, H., Xiao, Y., & Gao, Y. (2020). What are the Effects of Demographic Structures on Housing Consumption?: Evidence from 31 Provinces in China. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/6974276

Appendices

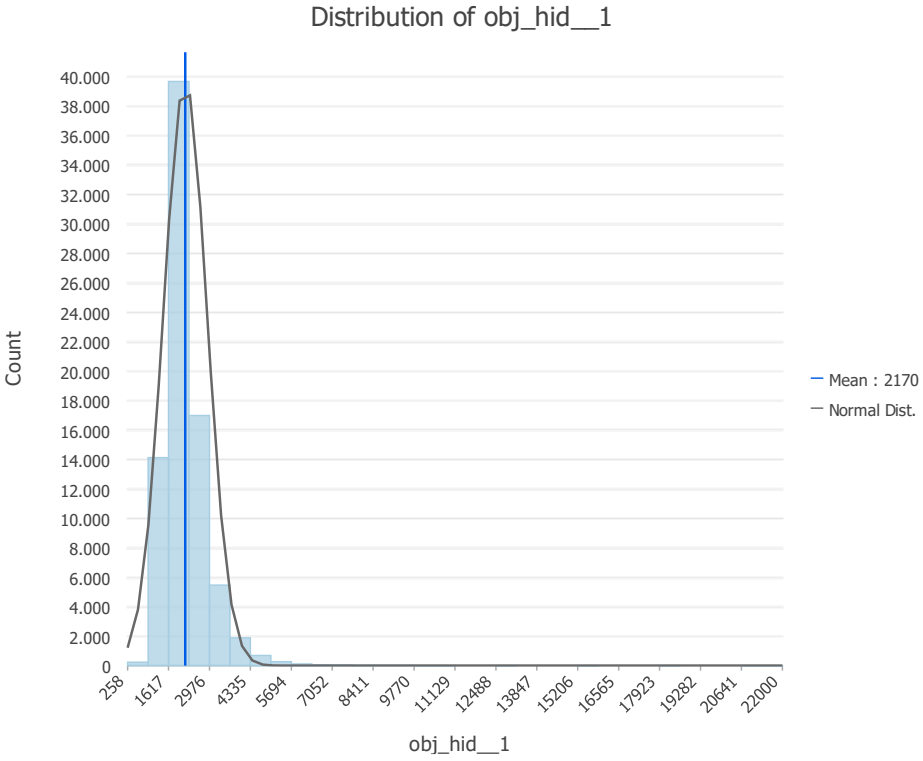
Appendix A: description of NVM data records as delivered

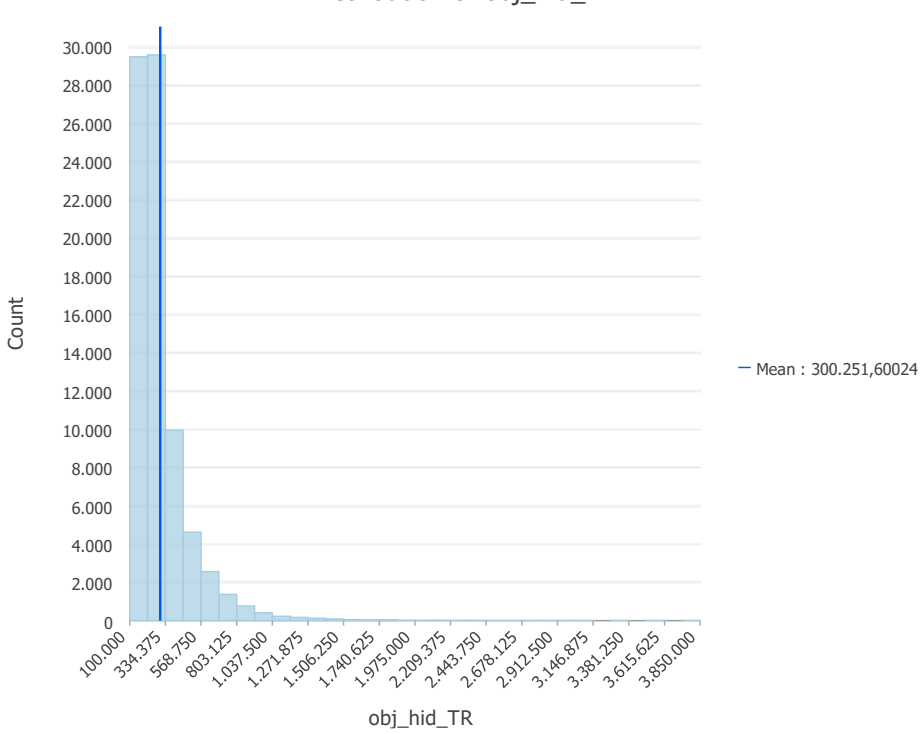
name	omschrijving	Variabelenaam
BUURT	Geeft de buurt id aan.	obj_buurt_ID
BWPER	De bouwperiode van de woning.	obj_hid_BWPER
CATEGORIE	Geeft aan of het object een woonhuis, appartement, bouwgrond of garagebox.	obj_hid_CATEGORIE
DATUM_AANMELDING	De datum dat het object is aangemeld	obj_hid_DATUM_AANMELDING
DATUM_AFMELDING	De datum waarop het object is afgemeld.	obj_hid_DATUM_AFMELDING
DATUM_LAATSTVRKOOPPR	Sinds wanneer geldt de laatst bekende vraagprijs.	obj_hid_DATUM_LAATSTVRKOOPPR
ERFPACHT_TONEN	Geeft de erfpacht aan.	obj_hid_ERFPACHT_TONEN
GARAGE	Geeft het type garage aan.	obj_hid_GARAGE
GED_VERHUURD	Geeft aan of de woning gedeeltelijk wordt verhuurd	obj_hid_GED_VERHUURD
GEMEUBILEERD	Geeft aan of de woning gemeubileerd is.	obj_hid_GEMEUBILEERD
HUISNUMMER	Bevat het huisnummer van het object.	obj_hid_HUISNUMMER
HUISNUMMERTOEVOEGING	Bevat de huisnummertoevoeging van het object.	obj_hid_HUISNUMMERTOEVOEGING
InDB_AANMELDING	De datum dat het object in de database is aangemeld.	obj_hid_INDB_AANMELDING
InDB_AFMELDING	De datum waarop het object in de db is afgemeld.	obj_hid_INDB_AFMELDING
INHOUD	De inhoud van de woning.	obj_hid_INHOUD
INPANDIG	Geeft aan of de woning een inpandige parkeergelegenheid heeft.	obj_hid_INPANDIG
ISBELEGGING	Geeft aan of de woning beleggingsobject is.	obj_hid_ISBELEGGING
ISNIEUWBOUW	Geeft aan of de woning nieuwbouw is.	obj_hid_ISNIEUWBOUW
ISOL	Geeft het aantal soorten isolatie aan.	obj_hid_ISOL
KELDER	Het type kelder bij een woning.	obj_hid_KELDER
KOOPCOND	Geeft de koopconditie aan.	obj_hid_KOOPCOND

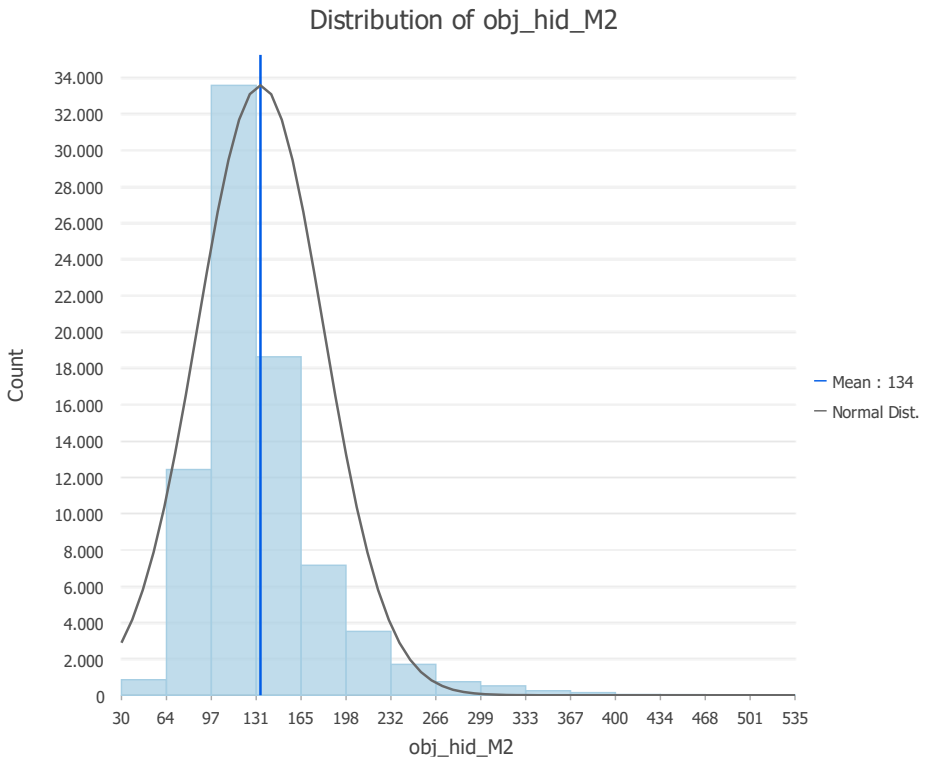
KWALITEIT	Wat is de kwaliteit van het appartement.	obj_hid_KWALITEIT
LAATSTVRHUURPR	Geeft de laatste verhuurprijs aan	obj_hid_LAATSTVRHUURPR
LAATSTVRKOOPPR	Geeft de laatste vraagprijs aan	obj_hid_LAATSTVRKOOPPR
LAATSTVRKOOPPRM2	Geeft de laatste vraagprijs per m2 aan.	obj_hid_LAATSTVRKOOPPRM2
LIFT	Heeft het appartementencomplex een lift.	obj_hid_LIFT
LIGCENTR	Geeft de ligging ten opzichte van het centrum aan.	obj_hid_LIGCENTR
LIGDRUKW	Geeft de ligging van de woning ten opzichte van de weg aan.	obj_hid_LIGDRUKW
LIGMOOI	Geeft de ligging van de woning aan.	obj_hid_LIGMOOI
LOOPT	Geeft de looptijd (in dagen) van een object aan	obj_hid_LOOPT
M2	De gebruiksoppervlakte van de woning, gecorrigeerd als het opgegeven woonoppervlakte niet betrouwbaar is, in vierkante meter.	obj_hid_M2
MONUMENT	Geeft aan of de woning een monument is.	obj_hid_MONUMENT
MONUMENTAAL	Geeft aan of de woning monumentaal is.	obj_hid_MONUMENTAAL
NBALKON	Geeft het aantal balkons aan.	obj_hid_NBALKON
NKAMERS	Het aantal kamers van de woning.	obj_hid_NKAMERS
NVERDIEP	Het aantal verdiepingen van de woning.	obj_hid_NVERDIEP
NVMCIJFERS	Woningklasse bij NVM-cijfers	obj_hid_NVMCIJFERS
NWC	Geeft het aantal wc's aan (vermenigvuldigd met 3).	obj_hid_NWC
ONBI	Geeft de staat van binnen onderhoud aan.	obj_hid_ONBI
ONBU	Geeft de staat van buiten onderhoud aan.	obj_hid_ONBU
OORSPRVRKOOPPR	Geeft de oorspronkelijke vraagprijs aan	obj_hid_OORSPRVRKOOPPR
OORSPRVRKOOPPRM2	Geeft de oorspronkelijke vraagprijs per m2 aan	obj_hid_OORSPRVRKOOPPRM2
OPENH	Geeft aan of de woning een openhaard heeft.	obj_hid_OPENH
PARKEER	Geeft het soort parkeergelegenheid aan.	obj_hid_PARKEER
PERCEEL	Het perceel oppervlakte van de woning.	obj_hid_PERCEEL

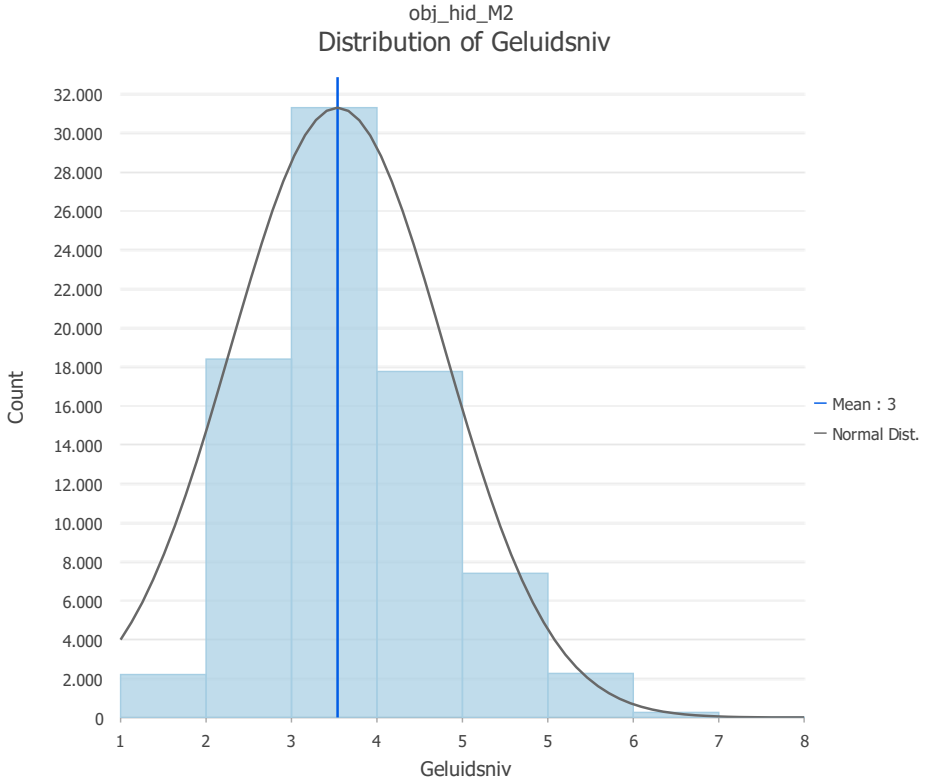
POSTC 4	Bevat het ID van de 4-cijferige postcode van het gebied waar de woning zich bevindt.	obj_pc4_ID
POSTC 6	Bevat het ID van de 6-cijferige postcode van het gebied waar de woning zich bevindt.	obj_PC6Co de
POSTC ODE	Bevat de postcode van het object.	obj_hid_PO STCODE
PROCV ERSCHIL	Het procentuele verschil tussen de vraagprijs en de transactieprijs.	obj_hid_PR OCVERSCHIL
SCHUU R	Geeft het type schuur aan.	obj_hid_SC HUUR
SOORT APP	Het soort appartement.	obj_hid_SO ORTAPP
SOORT DAK	Het soort dak.	obj_hid_SO ORTDAK
SOORT HUIS	De soort woning in geval van een huis.	obj_hid_SO ORTHUIS
SOORT WONING	Het soort woning.	obj_hid_SO ORTWONING
STATU S	Geeft de status van het object aan.	obj_hid_ST ATUS
STRAA TNAAM	Bevat de staatnaam van het object.	obj_hid_ST RAATNAAM
TRANS ACTIEPRIJS	Geeft de verkoopprijs of verhuurprijs aan	obj_hid_TR ANSACTIEPRIJS
TRANS ACTIEPRIJSM2	Geeft de verkoopprijs of verhuurprijs per m2 aan.	obj_hid_TR ANSACTIEPRIJSM2
TUIN_ OPP	De oppervlakte van de tuin.	obj_hid_TU IN_OPP
TUINLI G	Geeft de ligging van de tuin aan.	obj_hid_TU INLIG
TYPE	Het type woning in het geval dat het een huis is.	obj_hid_TY PE
VERKO OPCOND	De verkoopconditie van de woning.	obj_hid_VE RKOOPCOND
VERW	Geeft het soort verwarming aan.	obj_hid_VE RW
VLIER	Geeft aan of de woning een vliering heeft.	obj_hid_VLI ER
WIJK	Geeft de wijk id aan	obj_wijk_ID
WOON KA	Geeft het soort woonkamer aan.	obj_hid_W OONKA
WOON OPP	De gebruiksoppervlakte van de woning.	obj_hid_W OONOPP
WOON PLAATS	Bevat de woonplaats van het object.	obj_hid_W OONPLAATS
ZOLDE R	Geeft aan of de woning een zolder heeft.	obj_hid_ZO LDER

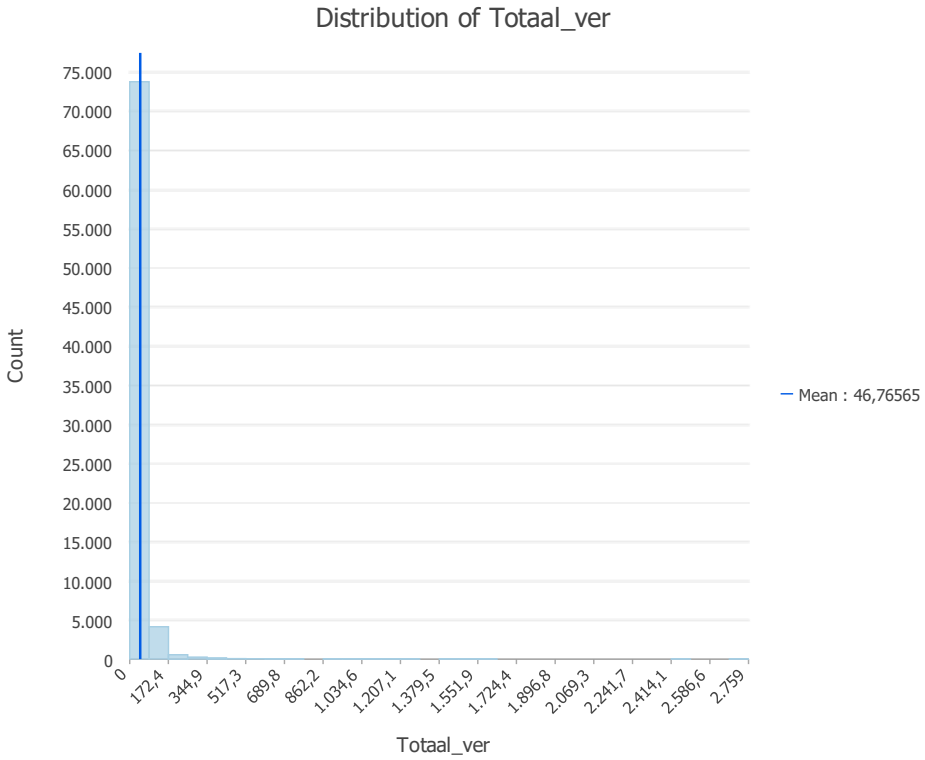
Appendix B: Descriptive statistics on variables of interest

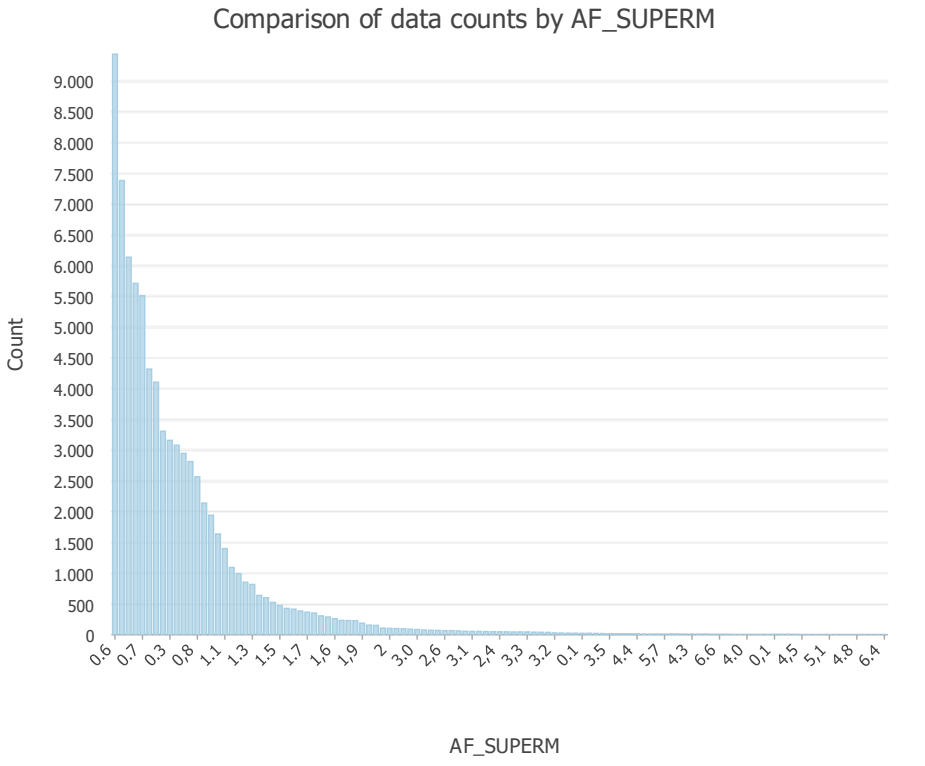


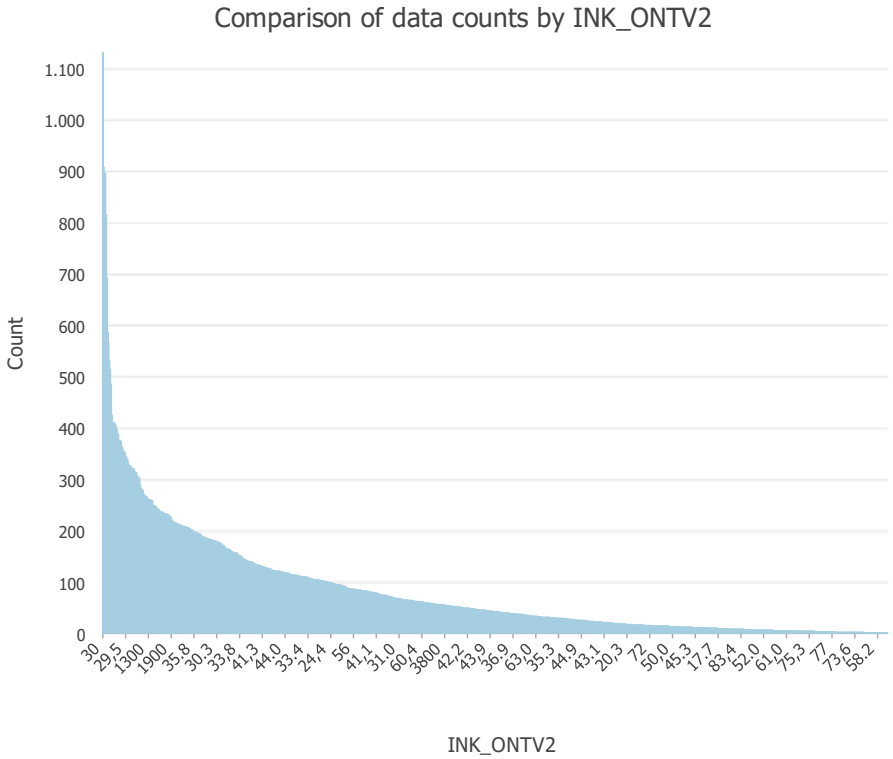


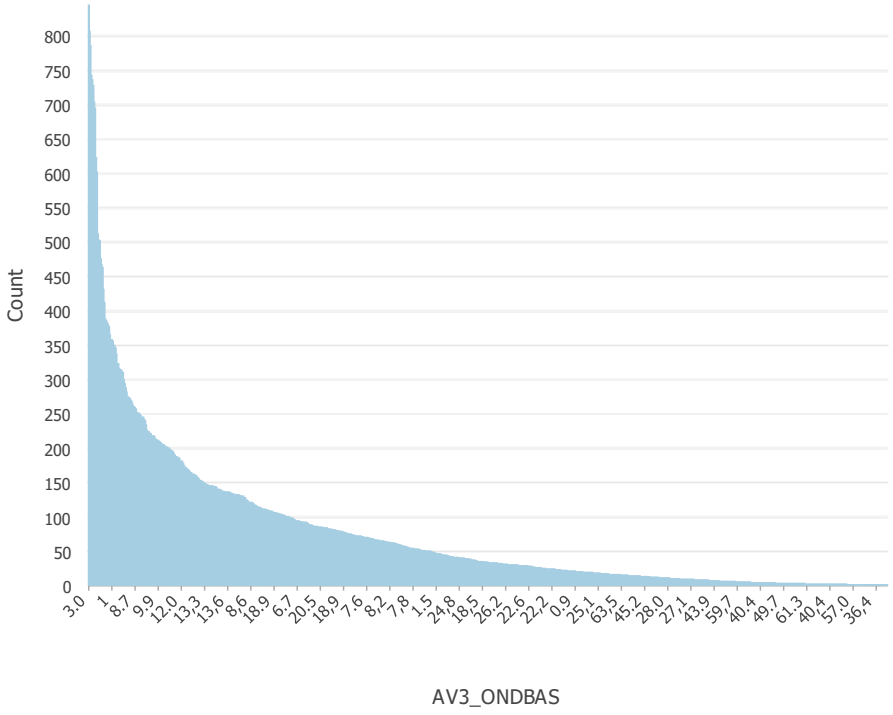

Distribution of obj_hid_M2
Count
34.000
32.000
30.000
28.000
26.000
24.000
22.000
20.000
18.000
16.000
14.000
12.000
10.000
8.000
6.000
4.000
2.000
0
30
64
97
131
165
198
232
266
299
333
367
400
434
468
501
535
obj_hid_M2
Mean : 134
Normal Dist.

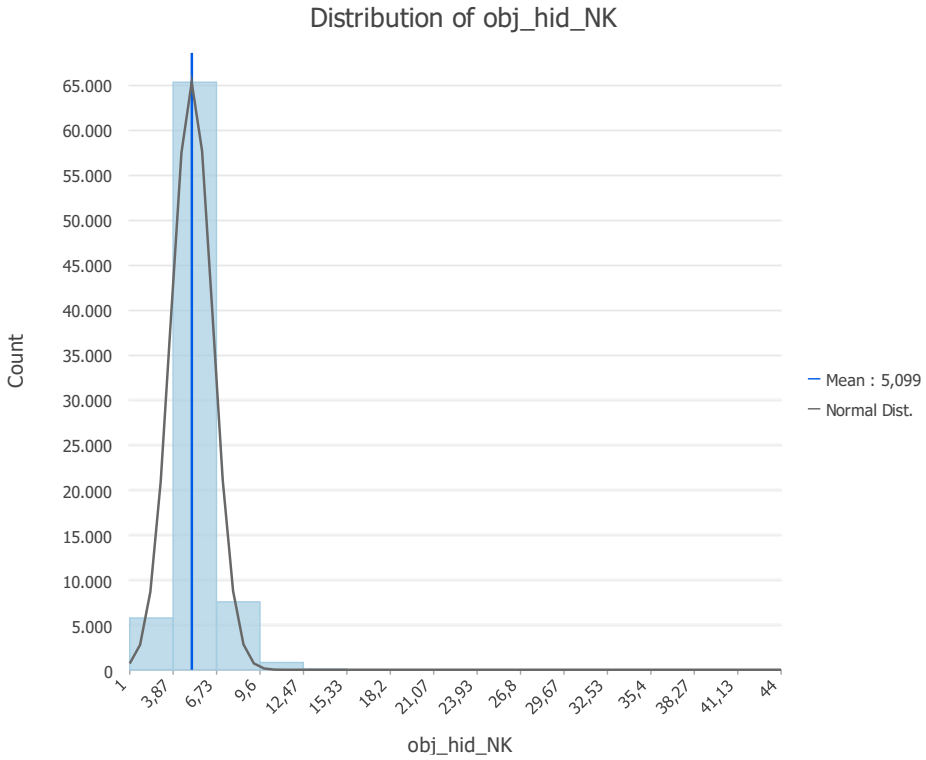

Distribution of Geluidsniv
Count
32.000
30.000
28.000
26.000
24.000
22.000
20.000
18.000
16.000
14.000
12.000
10.000
8.000
6.000
4.000
2.000
0
1
2
3
4
5
5
6
7
8
Geluidsniv
Mean : 3
Normal Dist.

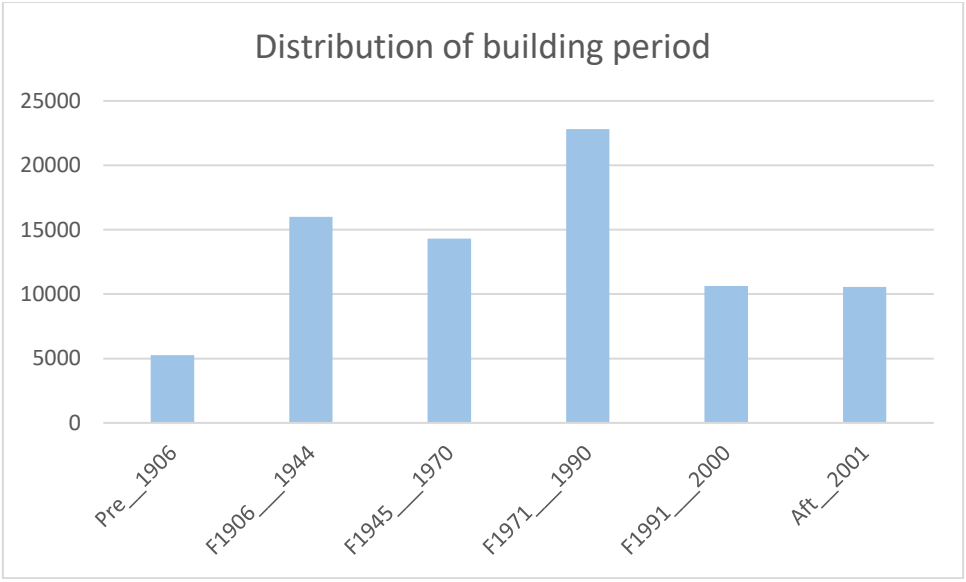

Distribution of Totaal_ver
Count
75.000
70.000
65.000
60.000
55.000
50.000
45.000
40.000
35.000
30.000
25.000
20.000
15.000
10.000
5.000
0
0
172,4
344,9
517,3
689,8
862,2
1.034,6
1.207,1
1.379,5
1.551,9
1.724,4
1.896,8
2.069,3
2.241,7
2.414,1
2.586,6
2.759
Totaal_ver
Mean : 46,76565

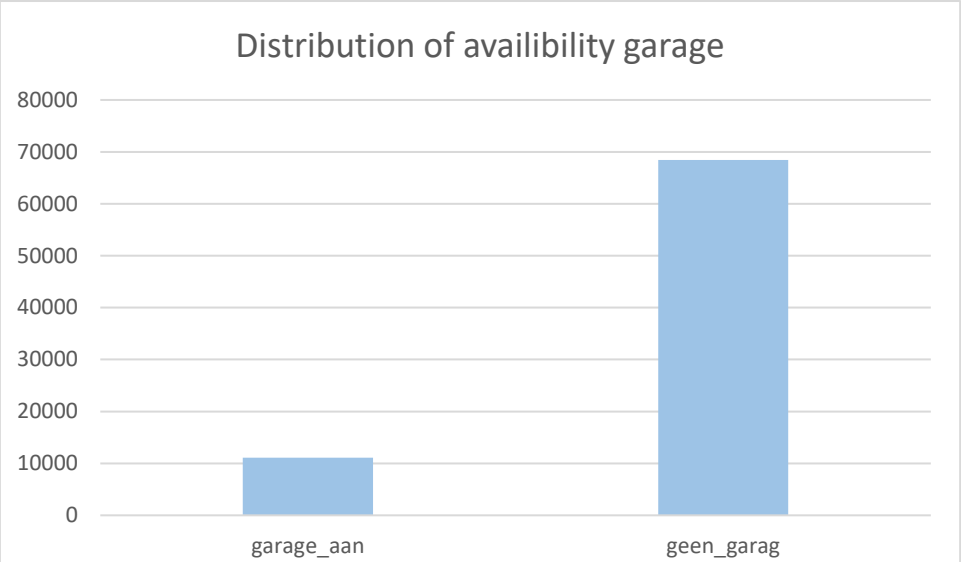

Comparison of data counts by AF_SUPERM
Count
9.000
8.500
8.000
7.500
7.000
6.500
6.000
5.500
5.000
4.500
4.000
3.500
3.000
2.500
2.000
1.500
1.000
500
0
0.6
0.7
0.3
0.8
1.1
1.3
1.5
1.7
1.6
1.9
2
3.0
2.6
3.1
2.4
3.3
3.2
0.1
3.5
4.4
5.7
4.3
6.6
4.0
0.1
4.5
5.1
4.8
6.4
AF_SUPERM

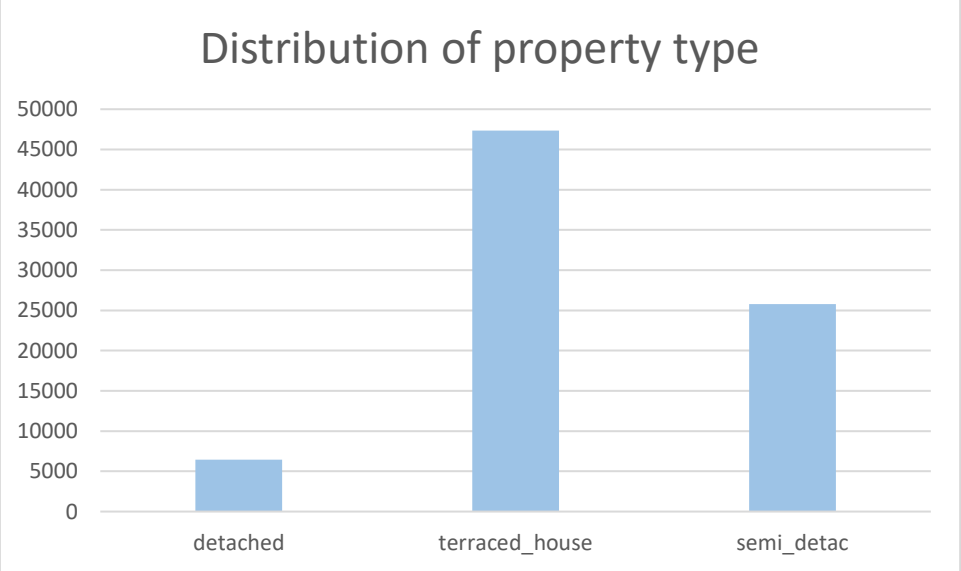

Comparison of data counts by INK_ONTV2
Count
1.100
1.000
900
800
700
600
500
400
300
200
100
0
30
29,5
1300
1900
35,8
30,3
33,8
41,3
44,0
33,4
24,4
56
41,1
31,0
60,4
3800
42,2
43,9
36,9
63,0
35,3
44,9
43,1
20,3
72
50,0
45,3
17,7
83,4
52,0
61,0
75,3
77
73,6
58,2
INK_ONTV2

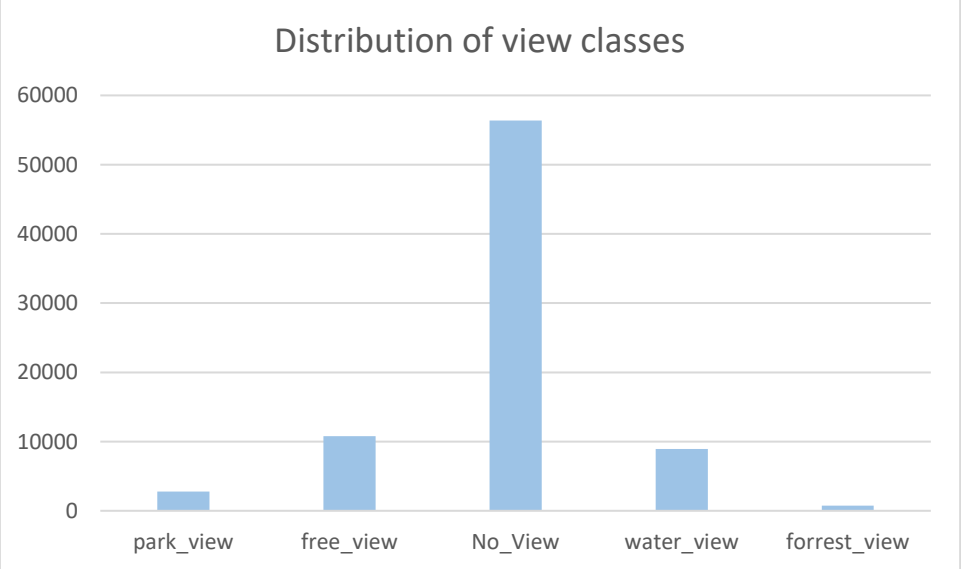

Comparison of data counts by AV3_ONDBAS
Count
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
3,0
1
8,7
9,9
12,0
13,3
13,6
8,6
18,9
6,7
20,5
18,9
7,6
8,2
7,8
1,5
24,8
18,5
26,2
22,6
22,2
0,9
25,1
63,5
45,2
28,0
27,1
43,9
59,7
40,4
49,7
61,3
40,4
57,0
36,4
AV3_ONDBAS


Distribution of obj_hid_NK
Count
65.000
60.000
55.000
50.000
45.000
40.000
35.000
30.000
25.000
20.000
15.000
10.000
5.000
0
1
3,87
6,73
9,6
12,47
15,33
18,2
21,07
23,93
26,8
29,67
32,53
35,4
38,27
41,13
44
obj_hid_NK
Mean : 5,099
Normal Dist.


Distribution of building period
25000
20000
15000
10000
5000
0
Pre__1906
F1906__1944
F1945__1970
F1971__1990
F1991__2000
Aft__2001


Distribution of availibility garage
80000
70000
60000
50000
40000
30000
20000
10000
0
garage_aan
geen_garag


Distribution of property type
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
detached
terraced_house
semi_detac


Distribution of view classes
60000
50000
40000
30000
20000
10000
0
park_view
free_view
No_View
water_view
forrest_view

Appendix C: Description of joined data sources for yearly data

measured variable >	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	Dummy from NVM data	Dummy variables classified by NVM	Urbanisation	Distance to clostest supermarket	Average income per income recipient (x1000)	IVM map of noise pollution 2016	Number of primary schools within 3 km	otal number of crimes per 1000 inhabitants
2006	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	dummy	Dummy	OAD2006	AF_SUPERM 2008	INK_ONTV 2006	patial join to noise raster 2016	AV3_ONDBAS 2006	otal number of crimes 2010
2007	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	dummy	Dummy	OAD2007	AF_SUPERM 2008	INK_ONTV 2007	patial join to noise raster 2016	AV3_ONDBAS 2007	otal number of crimes 2010
2008	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	dummy	Dummy	OAD2008	AF_SUPERM 2008	INK_ONTV 2009	patial join to noise raster 2016	AV3_ONDBAS 2008	otal number of crimes per 2010
2009	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	dummy	Dummy	OAD2009	AF_SUPERM 2009	INK_ONTV 2 2009	patial join to	AV3_ONDBAS 2009	otal numb

										noise raster 2016		er of crimes per 2010
2010	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	ummy	ummy	OAD2010	AF_SUPERM 2010	INK_ONTV2 2010	patial join to noise raster 2016	AV3_ONDBAS 2010	otal number of crimes 2010
2011	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	ummy	ummy	OAD2011	AF_SUPERM 2011	INK_ONTV2 2011	patial join to noise raster 2016	AV3_ONDBAS 2011	otal number of crimes per 2011
2012	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	ummy	ummy	OAD2012	AF_SUPERM 2012	INK_ONTV2 2012	patial join to noise raster 2016	AV3_ONDBAS 2012	otal number of crimes per 2012
2013	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	ummy	ummy	OAD2013	AF_SUPERM 2013	INK_ONTV2013	patial join to noise raster 2016	AV3_ONDBAS 2013	otal number of crimes per 2013

2014	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	ummy	ummy	OAD2014	AF_SUPERM 2014	INK_ONTV 2014	patial join to noise raster 2016	AV3_ONDBAS 2014	otal number of crimes per 2014
2015	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	ummy	ummy	OAD2015	AF_SUPERM 2015	INK_ONTV 2015	patial join to noise raster 2016	AV3_ONDBAS 2015	otal number of crimes per 2015
2016	obj_hid_M2	obj_hid_NK	Garage_Ava	Detached/Semi_Detac	ummy	ummy	OAD2016	AF_SUPERM 2016	g_ink_po 2016	patial join to noise raster 2016	AV3_ONDBAS 2016	otal number of crimes per 2015