SPATIAL DETERMINANTS OF REAL ESTATE APPRAISALS IN THE NETHERLANDS: A MACHINE LEARNING APPROACH

Master thesis for Business IT, specialization Data Science & Business.

Author: B. E. Guliker

June 25, 2021

Supervised by:
dr. ir. E.J.A Folmer, University of Twente, BMS
dr. ir. M.J. van Sinderen, University of Twente, EEMCS
R. Rops, Stater N.V.

UNIVERSITY OF TWENTE.

Preface

First of all, thank you for taking your time to read my thesis. This research is the culmination of 6 months of hard work, mostly done from home due to the ongoing Corona crisis. At first, trying to model appraisal values seemed like a daunting task. Some of the earliest iterations of the model had a large 10+% deviation, which was pretty demotivating. However, as cliché as it sounds, nothing is perfect on the first try. As a perfectionist, it is sometimes hard to let go of that notion. Luckily, there is a saying among statisticians which helped me stay focused on experimenting: "All models are wrong, but some are useful".

After all, for a complex problem like predicting house prices, no statistical model is perfect, as they all rely on chance. You just have to keep on trying different things, seeing what works and what does not and improving upon the things that do. This is, all in all, the essence of basic science. As someone once said, "the only difference between science and screwing around, is writing it down." (Adam Savage). I hope this provides some perspective and motivation for those working on their own large projects. A big part of the journey is about improving yourself and the end product in steps. In the end, your result will be the culmination of all the effort you put in.

Besides the lessons I picked up along the way for myself, I hope this thesis teaches you more about how open data can play a role in modelling appraisal values. I have always tried my best to write it in a way that is approachable even for those that are not familiar with the specific models mentioned in this paper. At one point in life, each and everyone of us will probably rent/buy a place of their own. So even if the technical details are not as interesting to you, this paper provides some insight into the characteristics that we value when buying a home. Many of the models were new for myself, so I have provided a collection of information I wish I had before I started this project.

Finally, I would like to thank both of my university supervisors Erwin Folmer and Marten van Sinderen, as well as, Roy Rops, my supervisor from Stater N.V.. Our monthly meetings together provided me with the necessary input and motivation to keep on going. When I felt like I was getting stuck, the feedback provided me with new motivation to approach the problem from a different angle. Furthermore, I would like to thank the people close to me. If you are reading this and feel like you contributed in any shape or form, I thank you. In this time, during which we keep mostly to ourselves, I was lucky to have support from the people around me. This research would not have been completed without you.

Management summary

There is a growing need for better localised value predictions for mortgage collaterals within the financial sector. Money lenders know the value of a house through an appraisal once the mortgage is approved. However, 20 years later, it is unknown how much the house is increased in value without conducting another appraisal. Still, money lenders are mandated by the Authority for the Financial Markets (AFM) to make a proper risk analysis of their portfolios. Currently, at Stater N.V., the Kadaster regional index is used to index appraisal which give a value indication for a mortgage collateral. This generalises the price increase for all types of housing to the same regional price index. The goal of this thesis is to find out if external data sources allow for more localised predictions of appraisal values by answering the following research question:

"How can hedonic price models, based on location and intrinsic characteristics of real estate, serve as an alternative to price indexation, in order to more accurately valuate the collateral (house) of Stater's mortgages in Netherlands?"

In the literature review, four types of hedonic pricing models are identified to model houses prices. These models are: Linear Regression (LR), Geographically Weighted Regression (GWR), Multi-scale GWR (MGWR), and Extreme Gradient Boosting (XGBoost). Chapter 3 (Methodology) outlines the solution design approach of the thesis, which is based on an application of the Design Science Methodology. Using a 5-step approach, three models are realised (LR, GWR and XGBoost) to model the appraisal values for five unique municipalities: Amsterdam, Amersfoort, Eindhoven, Groningen, Rotterdam.

The second contribution lies in the collection of public datasets to describes all houses in the Netherlands and the neighbourhoods they are located in. All in all, 33 variables are used, as seen in the variable overview of A.3.7. This includes intrinsic characteristics about each house from the Kadaster, sociodemographic variables from CBS, and energy labels from 'Rijkdienst voor ondernemend Nederland' (RVO).

In the end, the XGBoost model is able to model a large subset of the houses with a better accuracy than indexation. For the five municipalities, a single XGBoost can explain 83% the variance with a RMSE of €65,312, a MAE of €43,625 and MAPE of 6.35% (Table 5.5). The two most important variables in the model are the total living area (vbo_oppervlakte, from Kadaster) and WOZ-Waarde (from CBS) (Table 5.5) . As shown in the comparison between indexation and XGBoost for predicting the appraisal values of 2000 for the current year, the XGBoost model is able to take into account the different housing types (Figure 5.4). The downsides of the XGBoost model are the larger outliers than the conservative indexation method, as well as the extra effort needed to keep the data of the models up-to-date. However, in return for this extra effort, XGBoost can make more localised predictions for the entire Netherlands to valuate Stater's mortgage collaterals.

List of Figures

List of Tables

Contents

1 Introduction

1.1 The important role of real estate appraisals

Buying a house marks an important milestone in the lives of many. As most current and potential future homeowners know, the value of one's house plays an important role in the many aspects of home ownership. Not only is the price important for home buyers and sellers. It also plays a role in mortgage and insurance applications, as well as property taxes. The insurance companies and mortgage lenders need to determine the premium for the risk they are taking on. Furthermore, local governments estimate the values of property for capital gains or property tax. All these different parties rely on an indication of the true value of the house. Their desires for either a low or high valuation clash, which can lead to over- or undervaluation.

When a house is *overvalued*, the value is appraised to be higher than the true market value of the house. House owners want a high valuation when they sell their house to make a larger profit. On the other hand, house buyers want the price to be as low as possible. However, after the house is sold, the new home owner also wants a high valuation for his house so he can get take on a large enough mortgage. These are two drivers behind the risk for overvaluation. Overvaluation is a risk to the buyer and mortgage lender. During an unforeseen foreclosure, the homeowner will be left with an outstanding debt if the house is sold for a much lower price than the borrowed sum.

On the other hand, an *undervalued* house leads to less borrowing power for a home buyer. Under-appreciation is less of a risk for the mortgage lender as the lent sum will simply be lower when the value of the collateral is undervalued. Despite this, one can say it is advantageous to the homeowner that his house is undervalued, as this leads to lower insurance premiums as well as less property tax. For home insurance this is still a risk, since the pay-out for damages can be much lower than the actual damage done during an accident. Overall, over- and undervaluation both bear risks, as well as benefits, depending on the desires of the party involved. In the end, what matters most is a truthful valuation to ensure a fair deal between both parties. As such, appraisals are traditionally conducted by an unbiased third party, called an appraiser.

An appraiser visits a house to evaluate its condition as well as compare sale prices of houses with similar characteristics. The intrinsic characteristics of the house determine a large part of its price; examples include: number of bedrooms, amount of living space, presence of a garden or garage, presence of solar panels. By weighing all these factors, the appraiser tries to make an objective estimation of the property value.

In the Netherlands, it is mandatory to get an appraisal by a certified appraiser when taking out a mortgage. Further requirements, as mandated by the Authority Financial Markets (AFM) since 2018, are that the borrowed sum for a mortgage can never be more than the value of the property [1]. In mortgage lending the ratio between the borrowed sum and the collateral value is called *Loan-to-Value*.

Together with *Loan-to-Income*, these two ratios form the most important indicators of how much money can be borrowed and serve as a good indicator for the risk of the mortgage lender [2]. These indicators need to prevent people from taking on a mortgage they cannot afford. Accurate house price appraisals play an important role in this process, but as it turns out, these appraisals can be biased.

1.2 Traditional vs. model-based appraisal

In 2018, the Dutch national bank, 'De Nederlandsche Bank', released a critical report about the quality and independence of Dutch housing appraisals [3]. Their conclusion was that there is a structural over-appreciation by appraisers, based on 95% of all appraisals being equal to or higher than the sale price. All parties involved (buyer, seller, lender, estate agent) want the house to be sold, causing appraisers to be pressured into giving a higher appraisal. This, in turn, drives up the prices for housing even further.

The costs of these appraisals was another concern, as an appraisal can cost €500,- on average. This is much higher compared to the costs of a model-based estimation, which is closer to €50,-. The higher cost leads to believe that a model-based appraisal, or hybrid appraisal done by both a model and appraiser might be beneficial for potential house owners.

These model-based estimations are already being used in practise as an alternative to the traditional appraiser. In the Netherlands, a famous example is the *WOZ-Waarde*. The *WOZ-Waarde* serves as an indication of value for the property, which is later used during taxation. It is simply impossible to appraise every single house in person on an annual basis. Many insurance companies and mortgage lenders are in the same boat: the costs to conduct an official appraisal for each and every house in their portfolio is simply too high.

However, the financial sector needs to comply with international regulations such as the *Basel II accords* [4], which state that financial institutions need to ensure that capital allocation is more risk-sensitive. As such, many mortgage lenders and insurance companies opt to adjust the house values in their portfolio with national indices to re-evaluate the house prices.

The drawback of indexation is that it still generalises different types of houses into a single index. Consequently, houses can still be over- or undervalued if there are differences between for example the types of houses used in the index, or different price growth rates for different cities. Instead, a local model can more fairly estimate the regional differences in housing type and location. Such a model ensures trust between both the customers and the financial sector since estimations aim to be unbiased by being based on quantitative data.

Many of these models are so-called *Hedonic Pricing Models*, which estimate the house on quantitative data about the house characteristics, location and the supply versus demand, similar to the role of an appraiser. Literature has shown that for many cities, e.g. London [5], Rotterdam [6], Leipzig [7] and Singapore [8], the house prices can be estimated using these types of models. However, many of these models focus on a single city within a country. Studies which compare local models across cities have yet to be explored.

1.3 The goal: towards more localised prediction of house prices

As introduced above, an accurate estimation of house prices is beneficial for both the financial sector as well as the home buyers. An accurate and transparent house price reduces the risks for both parties by quantifying the value using actual data. Furthermore, a fair system ensures trust between the home buyer and the financial sector, which is beneficial to society. Finally, there is a growing need for prediction models for house prices which are not bounded to a single region or city, but that can estimate the prices for houses across an entire country more cost-efficiently than a traditional appraiser.

This thesis is supervised by Stater N.V., a mortgage service provider in the Netherlands. Currently, for risk allocation, the values of mortgage collaterals are estimated by indexing the original appraisal with the housing index of the Dutch Kadaster. This index is based on the average sale price for each of the twelve provinces in the Netherlands. This is a large generalisation, which assumes that prices in the entire province, for all types of housing, have risen at the same rate. A more accurate estimation using hedonic price models would be beneficial for the risk management of Stater and their clients, since it allows them to make a better estimate for the Loan-to-Value of a mortgage. From this business motivation arises the goal for the final thesis. The goal is formulated below as a design problem, according to the Design Science Methodology [9]:

"Improve the accuracy of automated collateral value estimations of Stater N.V., by designing a model that valuates the collateral (i.e. house) based on location and intrinsic characteristics, instead of price indexation, to facilitate better portfolio risk management."

To be specific, the house price estimation refers to the value determined by the appraiser, as this is the value of the house that is considered when taking out a mortgage. This can differ from the final (market) sale price of the house as well as the *WOZ-Waarde*. Note that the *WOZ-Waarde* is only semi-publicly available, as such it cannot be used as an indicator for large datasets of houses. From here on out, house value and house price will be used to specifically refer to the appraisal value of a house. Furthermore, the model accuracy is based on quantitative metrics including R^2, RMSE and MAPE of the model for a separate test set. In addition to this, the run-times and implementation times of the models will be considered when choosing the best model for Stater.

To realise this goal, this research investigates if modern machine learning techniques can help make more localised estimation for house prices, specifically if price differences between and within cities can be modelled using the both public location and housing characteristics, as well as the data of Stater. It is crucial to discover if the effects are similar between cities, or if separate local models need to be trained for every city. Therefore, the main research question of this report is defined as follows:

"How can hedonic price models, based on location and intrinsic characteristics of real estate, serve as an alternative to price indexation, in order to more accurately valuate the collateral (house) of Stater's mortgages in Netherlands?"

To answer the research question, the report starts off with a literature review to provide more background information into a selection of state-of-the-art models and features used for modelling house prices. This literature review is guided by answering the following two questions:

Q1.1: What state-of-the-art machine learning models are used to model house prices in existing literature and in practise?

Q1.2: To what extent do relationships exist between location characteristics and housing characteristics which explain the differences in house prices?

1.4 Scope

For a good model-based predictions, many data points about the house and its unique location are required. During initial research of the data of Stater N.V., it was discovered that the larger cities also had the most transaction data. A challenge presented in this research is the sparsity of the data for some regions. This is due to a mortgage often only being taken out once every 30 years. With yearly changing housing prices, each individual year only has a limited sample of houses from the entire population size.

As such, the final thesis is scoped around five large municipalities spread across the Netherlands, namely Rotterdam, Amsterdam, Eindhoven, Amersfoort and Groningen. These five municipalities where most prominently represented in the market value dataset of Stater. The cities in this dataset contains at least 25 thousand market values spread out over 20 years (2000-2020). The regions are all located in different parts of the Netherlands. The assumption is made that this dataset provides sufficient variety to train the model for any particular city in the Netherlands.

1.5 Contributions

Academic relevance

The aim of this research is to provide more evidence for whether house appraisals can be modelled using intrinsic and spatial characteristics, but most importantly, it aims to fill the gap in research if the characteristics have different or similar influences between cities. The novel contribution to the field of existing house price models by specifically comparing multiple cities across an entire country, instead of just focusing on training a model for a localised area. Finally, the research seeks to provide an overview of features which are important for building reliable house appraisal models and how the chosen models can play a role in achieving this goal.

Societal relevance

This research hopes to pave the way for better and more reliable appraisals and accurate estimations for house prices by exploring how data-driven machine learning models can help better estimate housing prices. As highlighted in the introduction, accurate model-based estimations of housing prices are both beneficial to homeowners as well as the financial sector, including companies such as mortgage lenders and insurance companies. A transparent and fair market value for a house ensures trust between the financial sector and homeowners. Additionally, quantifying which factors have a bigger impact on house prices allows local policy makers to make better informed decisions for new housing development projects. All in all, it is clear that more accurate estimations of house prices are beneficial to society.

1.6 Thesis Outline

The report is structured into six chapters, starting with this chapter which provides an introduction to the main research question and the problem motivation. Chapter 2: background, provides a literature review on state-of-the-art models for predicting house prices and commonly used data features. Based on the results of the literature review, four potential models are identified as a solution to the main question: (1) linear regression, (2) geographically weighted regression (GWR), (3) multi-scale GWR (MGWR), (4) extreme gradient boosting (XGBoost). Chapter 3: methodology, outlines how the Design Science Methodology is applied to formulate an approach for building and evaluating three models. Here is decided that if the results of GWR are satisfactory, a more specialised MGWR model will be explored. Otherwise, the XGBoost algorithm is implemented. The methodology concludes with a 5-step approach used throughout the remainder of the thesis. Chapter 4: solution design, outlines the gathering of external variables and iterative creation of the models. The results of the final iteration of each of the models are listed in chapter 5: results. Ultimately, XGBoost was chosen in favour of MGWR due to only the small improvement of GWR over LR. The chapter finishes by comparing the models to the current approach of indexation. Finally, chapter 6 provides the final answer to the research question by answering each of the four sub-questions defined in the Methodology. It discusses the reliability and concludes with a recommendation for Stater and areas for future work.

2 | Background

The goal of this literature review is threefold. Firstly, it discusses the benefits and limitations of two approaches for estimating house prices: price indices and hedonic pricing models. Simultaneously, the Kadaster price index and other house price indices of the Netherlands are explored, to show developments in the Dutch housing market. Secondly, the review evaluates both two practical models as well as four state-of-the-art models commonly used in literature for hedonic price models: linear regression (LR), geographically weighted regression (GWR), multi-scale GWR (MGWR) - an improvement upon GWR and extreme gradient boost (XGBoost).

Finally, the review concludes with an overview of features which are commonly used in hedonic price models for house prices. This overview is divided into three categories: market characteristics, location characteristics and intrinsic characteristics of the house. This literature review provides the foundation for building the model to predict the house prices for five municipalities in the Netherlands.

2.1 Dutch house price indices and the repeat-sales model

Price indexation is a method for calculating a normalised average price increase for different types of goods. Four common methods to calculate an index are: (1) Paasche index, (2) Laspeyres index, (3) Lowe index and (4) Fisher index. Every index aims to give a good indication for the price change during a specific interval of time. A price index is often used to estimate the present value using a historic known value, this process is called *indexation*. In the case of house prices, the current value of a house can be estimated by using a sale price from the past and indexing it using a house price index.

For the Netherlands, a notable house price index is calculated by the *Kadaster*. The Kadaster is the Dutch land registry and mapping agency. They maintains the official registry of properties and land ownership in the Netherlands. This registry is called the *Base-registry Addresses and Buildings* (BAG). The house price index, together with other statistics related to the Dutch housing market, are presented in a publicly available dashboard which is updated every month.

There exist additional house price indices for the Netherlands. NVM, the largest Dutch association of real estate agents, publishes a house price index based on all the transactions handled by its members [10]. Funda, the largest online Dutch real estate marketplace, publishes indices related to consumer confidence and their willingness to buy real estate [11]. These indices are an aggregate index, meaning not only the sale price but also other data points play a role in the calculation of this index. For the goal of estimation, these aggregated indices are less suitable due to the compounded factors influencing the index.

Lastly, the Central Bureau of Statistics (CBS) also publishes house prices indices.

Among others, they publish data on the average sale price of houses per municipality (see Figure A.1.1). From this figure it can be seen that there exist large differences between municipalities. This supports the need for more local models for home appraisals. The graphs of the CBS and their indices are however based on the same data provided by the Kadaster [12]. In the end, the Kadaster is the source of truth used by all the municipalities for real estate. It includes all official real estate transactions. As such, the Kadaster index is the most truthful index for calculating house prices for the Netherlands.

The Kadaster index is calculated using a *weighted repeat-sales model* [13]. The four aforementioned methods for calculating price indices require multiple sales of the same good, in the desired time span, for an accurate index. This means multiple sales of the same good per year for a yearly based index. However, this is not the case for houses, which often do not get traded for decades. The repeat-sales model is developed to specifically circumvent this issue.

The repeat-sales model averages the change in sale price for a single good between two different moments in time [14]. In case of house prices, it averages the change in price for the same house which has been sold in separate years. Inevitably, a prerequisite for this model is the need for at least two separate sales dates for every unique house. The repeat-sales model is not only used to calculate house prices, but other infrequently traded goods such as collectables (e.g. pieces of art). The *weighted* repeat-sales model expands on the model by having more frequently traded houses contribute less to the total average than houses traded over a larger span of time. This avoids bias towards more frequently traded houses.

Besides giving a national index for house prices in the Netherlands, the Kadaster house price index consists of two unique refinement levels: one is for the different provinces of the Netherlands (Figure A.1.2), the other for six different types of housing (Figure A.1.3). Both indices are based on all real estate transactions of the last twenty years (2001-2020), with 2015 as base year. For the sake of clarity, Figure A.1.2 only shows six out of the twelve provinces. While the house prices follow the same trend, the small differences over many years add up to a significant differences between over time [13]. The largest increase is seen in *Noord-Holland*, where prices have risen up to 76.70%, twice as high as compared to 38.16% in *Limburg* (as seen in Table A.1.1). For housing types, the difference is also statistically significant as proven in [13]. Considering these facts, it can be concluded that additional factors are needed in order to model the house prices on a more localised scale for the Dutch housing market.

In the end, indexation provide a reasonable estimation for house prices but only on a global scale. In a local model, when one wants to estimate the current value of a specific house, an index is likely to give a 'good enough' estimation. The most that can be said using an index for a specific house is that the price has increased or decreased if its a high positive or negative value. Including different factors to compose more indices

improves the accuracy for more local models. Despite this, the biggest downside still remains. Indices rely on large samples of the total transactions to be reliable. Hedonic price models, in this case, are a valid alternative for explaining the variances in house prices that do not rely on large samples through the use of regressions.

2.2 Hedonic price models

Hedonic pricing states that the price for a product is an aggregation of prices which a buyer is willing to spend for individual characteristics of the product. For a house, these characteristics range from intrinsic characteristics (e.g. number of rooms), to location characteristic (e.g. access to amenities), as well as market characteristics (e.g. supply of houses in the area) [15]. Correspondingly, house prices reflect macro-economical changes in the wishes and values of society. As such, house prices play a versatile role in quantifying the price of intangible goods such as clean air [5], presence of green space [16] and accessible infrastructure.*Hedonic price models* use different types of regression models to estimate the price and weight of each characteristic. The four types of regression models used in recent research for hedonic house price estimations are: (multi) linear regression, geographically weighted regression (GWR), multi-scale GWR (MGWR) - an improvement upon GWR and extreme gradient boost (XGBoost).

2.2.1 Linear regression (LR)

Linear regression (LR) models the change in a dependent variable based on a linear relationship to one or multiple independent variables. Using ordinary least squares, the influence of each feature is described by a single coefficient. Research successfully shows linear relationships exist between house prices and the living surface area of a house [17]. Furthermore, many other intrinsic characteristics such as the number of bedrooms [18] and the amount of garden space [16] show an underlying linear contribution to the price of a house. The advantage of the linear regression model lies in its simplicity to have the same response for all data points. As a result, linear regression models are generally less prone to over-fitting the dataset.

Conversely, the simplicity of linear regression models is also their downfall when it comes to modelling more complex phenomena such as house prices. In practise, many other factors that play a role in house prices also show non-linear relationships [6]. For example, an additional room has a larger influence on the value of an apartment than it has for a detached home. This can be resolved by breaking down the non-linear relationship into a linear relationship by including another feature, in this case the type of house. However, it is often the case that the non-linear relationships simply cannot be broken down into linear relationships through the inclusion of additional features.

Finally, linear regression models are argued to not be a good estimator for house

prices due to the lack of modelling a spatial component [18]. House prices for the same type of house in Amsterdam vary wildly from those in Groningen [12]. Both on a national level, as well as city level, the price of the same house is often different. This is because of spatial heterogeneity, meaning the value of a variable varies across space. Not considering spatial heterogeneity in the model causes spatial non-stationarity. Spatial non-stationarity is the name [19] for the situation in which a global model, such as linear regression, is unable to accurately predict the outcome due to location playing a role.

One way to mitigate the spatial non-stationarity problem, is to group observations through the use of a dummy variable, such as the inclusion of zip-codes [20] or distance to the centre of the city [21]. Furthermore, it is argued that through quantifying enough features, it is possible to distinguish regions [22]. Nevertheless, the downside of this method is that it is a very data intensive to make reliable distinctions. Despite all this, the model will still ignores the spatial dependence of nearly located houses which has been proven to be statistically relevant when modelling house prices. All in all, the lack of spatial component and subsequent decrease in model accuracy might not be significant when looking only at the individual characteristics of houses in a neighbourhood or city.

2.2.2 Geographically weighted regression (GWR)

Geographically weighted regression (GWR) is a parametric model based on traditional linear regression but also takes into account the spatial heterogeneity to avoid the problem of spatial non-stationarity. Similar to linear regression, GWR gives each independent variable an estimated coefficient, however the coefficient varies spatially depending on near data points [19]. Which points are considered near enough and the weight each point gets assigned is defined through a kernel function. GWR has proven beneficial for better accuracy based on both intrinsic characteristics [6] and location characteristics [7].

For spatial analysis such as GWR, it is important to know about spatial auto correlation. Spatial auto correlation is most famously described in a quote by Tobler, also known as the First Law of Geography: "everything is related to everything else, but near things are more related than distant things" [23]. More formally, spatial auto correlation is the correlation between data points of nearby locations in space. Commonly used statistics for determining spatial auto correlations are Moran's I and Geary's C test statistics. Spatial auto correlation can be an indication of missing a dependent variable. Which in turns means a wrongly specified model, leading to results that can be statistically invalid.

The kernel function plays an important role in how the model weights each of the coefficients. Two main types of kernel functions exist: (1) *fixed*; which considers data points in a fixed radius, and (2) *adaptive*: which considers a fixed amount of

neighbours. An adaptive function will automatically adjusts its bandwidth to always include the same number of data points. This makes it ideal for spatial datasets which are not uniformly distributed spatially. The most commonly used kernel function across the identified literature in real estate pricing is the adaptive Gaussian kernel, which considers all observations but the weight tends towards zero the farther away an observation is [6], [7], [8], [24]. The kernel function of the GWR model can be optimised through usage of the Golden search method and cross-validation. The step of kernel function optimisation is crucial as a randomly chosen kernel function decreases the accuracy of the model.

The most discussed downside of the GWR model is the fact that the kernel function is forced to have the same bandwidth for all variables. The bandwidth is the amount of data points that are weighted in the kernel function. The consequence of the same bandwidth is that each data point influences the for all given variables. This is not necessarily the case in practise. Some effects might only be related to influences of other houses in the same neighbourhood, while others are globally influenced by all data points in the city. This simplification of reality sparked the creation of a new variation upon GWR which does include variable bandwidths, called multi-scale geographically weighted regression.

2.2.3 Multi-scale geographically weighted Regression (MGWR)

Multi-scale geographically weighted regression (MGWR) introduces variable bandwidths for each of the coefficients [25]. Despite the first publication in 2017, this model has seen fewer studies than GWR, both overall, as well as in the context of house price estimations. This can be due to the fact that popular spatial analysis tools, such as ArcGis, do not yet have a build-in MGWR analysis, only for GWR. The recent release together with no major support of spatial analysis tools makes it that less research has yet been conducted on MGWR as compared to GWR.

Nevertheless, research has shown MGWR always offers an improvement over GWR [25]. The total improvement however varies across studies. These differences are sometimes too small to be statistically significant. As seen in [26], the explained variance (r^2) papers show a minor increase of 0.05 (10% improvement) in explained variance. Furthermore, a recent study into prices of AirBnB rental prices also had a 0.10 improvement with the use of MGWR versus GWR [27]. Overall, research [26], [27] agrees that the different local and global influences of variables are the main benefit of MGWR over GWR.

2.2.4 Regression Trees and Extreme Gradient Boosting (XGBoost)

Although with (M)GWR the coefficients can vary spatially to model both positive influences in one location, and negative influences in another location, they still rely on linear relationships to perform regression analysis. An alternative to this is a decision tree model, which is able to model non-linear behaviour. Commonly used for classification, decision trees can also be used for regression, often called regression trees in that scenario. Gradient Boosting is a technique that uses ensemble learning of many weak prediction models to make better prediction then using a single tree. Finally, Extreme Gradient Boost (XGBoost) is a library that implements this gradient boosting for tree models in a way that is fast and efficient.

XGBoost also sees applications in literature for predicting house prices. It has been used to model the Boston housing dataset with a mean absolute percent error of less than 5%) [28]. This dataset is a popular dataset for Kaggle competitions to compare the performance of various machine learning models. Similar to the Boston dataset, most other applications of XGBoost also focus on modelling house prices based on intrinsic characteristics of the house itself [29]. Overall, this makes XGBoost another prime candidate for a hedonic pricing model that can also capture non-linear relationships.

2.3 Applications of hedonic price models in the Netherlands

In the Netherlands, a well-known practical example of a hedonic price model comes from the *WOZ-waarde*. The WOZ-waarde indicates the value of a property, which is used for taxation. At its core, the WOZ-waarde comes from matching the sale prices of houses with similar characteristics [30]. Just like a hedonic model, it weights the characteristics and location of the house against to make a prediction. This data comes from official registries from the Kadaster, such as the BAG and their sales registry. In actuality, the model is more complex than a hedonic price model. It uses many extra layers for improving and validating the accuracy of the model [31]. For example, they conduct samples of physical appraisals for very unique houses to ensure validity. In addition, satellite pictures are used to check for physical difference, like when somebody has built a house extension or swimming pool, which increases the property's value. In most municipalities, a house owner is able to get a report about his WOZ-waarde which shows which similar houses are used as a comparison.

A commercial example of (hedonic) house price model is Calcasa [32]. Calcasa puts itself in the market with their own valuation model, that is certified by ratting bureaus such as Moody's, Fitch Ratings and Standard & Poor's. They target insurance companies and mortgage providers to provide model-based appraisals for their portfolios. Unfortunately, as this is their business model, it is unclear what exact model they run, just like the WOZ-waarde. For €28,- a consumer can get an online valuation.

Fortunately, they provide an example valuation report. This report appears very similar to the one provided with the WOZ-waarde.

Firstly, the report consists of a valuation of your house based on sold houses with similar housing characteristics together with a score to indicate the reliability. It also includes a few pages on market characteristics such as developments of the housing price in the Netherlands and the amount and type of house sales in your neighbourhood. Finally, it concludes with a page on neighbourhood characteristics such as average income, price per living area, type of household as well as nearby schools and transport options. While it is not stated explicitly, it appears the data mostly consists of data publicly available through the Central Bureau of Statistics (CBS).

Other notable smaller examples include tools from Kadaster-Data [33], or Hypotheker [34]. These web services both provide free online estimation for house prices. They are very explicit in the fact that it is not an official appraisal, but a mere estimation. Besides using public data, it is unique that these services use data collected through survey questions. To get a free online estimation, you are required to fill out a survey about the characteristics of your house. Part of their business model is to store this data to make better predictions for house prices.

All in all, from these examples it can be seen that there definitely exists a market for house price models in the Netherlands. All these models seem to rely on systems that try to match sale prices of similar houses based on their characteristics. This sales data is the key starting point for all models. If enough sales data is present, the most difficult challenge is collecting as much accurate data about a house as possible. The main physical characteristics, as well as neighbourhood characteristics, are publicly available through the Dutch Kadaster and CBS respectively. In the end, whoever has the most, but also accurate, data will ultimately be able to make the best prediction.

2.4 Features for house price estimations

Based on the analysed studies and practical applications for hedonic pricing models, a list of characteristics is identified and divided into three categories: market characteristics, location characteristics and intrinsic characteristics of the house. The two most important categories are the intrinsic and location characteristics of the house, since the market characteristics are global influences impacting all houses. Nevertheless, the market characteristics have been included for the sake of completeness. This overview is based on the overview of hedonic model variables of Zhou et al. [18]. This overview however focuses mainly on variables that have also been included in geographically weighted regression models.

The market characteristics are identified as global influences on the entire housing market. One large market influence are national policies, such as the recent abolition

(January 2021) of transfer tax for starters in the Dutch housing market. These national policies often have an equal impact on the price for all housing [22]. Another global influence is the mortgage interest rate. A lower interest rate leaves the home buyer with more money to spend. As a result, this often drives up house prices. Since market characteristics are global influences, it does not explain the spatial variance in house prices. As such, these variables do not belong in a geographically weighted regression model. Nevertheless, they play a crucial role in explaining the temporal difference in houses prices, as they do play a role when looking at the growth of house prices on a yearly basis.

In contrast, intrinsic characteristics are the biggest differentiating factors for house prices. As such, they are also by far the most used variables for hedonic pricing models. Not only in literature, but also in the hedonic price models such as the from practise these were stated as the heaviest influences for house prices. The largest influences are naturally the living area and volume, commonly followed by the amount of garden space. Amenities such as a garages and multiple bathrooms also contribute to higher house prices. The build year can serve as a moderate indicator of energy efficiency and state of maintenance, however it does not always depict the true condition of the house. Old houses are likely renovated once in their life span, so other features such as an energy label are needed. Furthermore, older buildings can also be cultural heritage, which can result in higher prices for older buildings due to their significant historic value as stated in [6]. The complete overview of all variables is given in Table 2.1.

The largest downside of these intrinsic characteristics is that the data is especially hard to come by. Most intrinsic characteristics are part of advertisements of real estate agencies, which are not publicly available for any random house. This can be attributed to privacy concerns, as well as, the fact that collecting all this data takes time and effort. As a result, many parties do not want others to use their valuable data. Despite this, good public sources for house characteristics do exist. In the Netherlands, the Kadaster, provides basic information about every house including year of construction and living area.

In literature, the majority of the GWR models for house pricing focus on modelling only intrinsic characteristics based on data gathered from real estate marketplaces or real estate agencies [6], [35], [36], [37]. However, research [5], [8], also show that the location characteristics can be reliably be used when only the surface area and location is known about the property it self. According to [5], the location or neighbourhood accounts for 15% to 50% of the total house price. As such, even when little data is available about each specific house, a more local estimation can still be performed using location characteristics.

Identified intrinsic characteristics influencing house prices		
Characteristic	*Influence*	*Sources*
Year of construction	Positive/Negative	[6][18][38]
Living area	Strongly positive	[6][15][18]
Type of housing	Positive	[6][15][18]
Garden space / presence of garden	Positive	[15][18]
# of rooms (bedrooms, bathrooms)	Positive	[15][18]
Presence of facilities (shower, lift, swimming-pool, garage)	Slightly positive	[15][18]
Furnished	Slightly positive	[15][18]
Energy Efficiency	Slightly positive	[6]
Sustainability measures (solar panels & better insulation)	Slightly positive	[6]

Table 2.1: Source: author's summary

Location characteristics are features derived from the type of neighbourhood and the presence of nearby buildings. Nearby access to convenience stores, recreation, parks all have positive influences on house prices. This agrees with *bid rent theory*, which states that rent for housing gets higher, the closer the house is to the central business district.

Similarly, accessibility plays another role in the price of a house. Travel time to certain locations such as the central business district can be a better indicator than the distance. However not all forms of transport are a positive influence. The proximity of highways have a larger detrimental effects. The effect of the noise disturbance is greater then the impact on better accessibility of other cities. Views also play a part, outlook on a river, lake or sea can have positive influences whereas wind mills and high-rise buildings have detrimental effects.

Lastly, there are socio-economic indicators for a neighbourhood that also relate to house prices. A higher average household income is most often found in areas with more expensive housing. Crime rate often has a negative impact on house prices. An important thing to keep in mind is that these characteristics do not necessarily mean there exists a causal relationship. Overall, the location characteristics have a less pronounced effect than most intrinsic characteristics, as the value associated with each of them varies on a personal basis, yet they can still provide large insights into why certain houses have higher house prices than others. A summary of the variables is given in Table 2.1.

Identified location characteristics influencing house prices		
Characteristic	*Influence*	*Sources*
Household income	Strongly positive	[8][19]
House shortage / Surplus	Strongly positive	[39]
Notable view (sea or lake)	Highly positive	[37]
Time to travel (foot, bike, bus) or distance to city centre	Highly positive	[16][20]
Proximity to place of worship	Positive/Negative	[6][40]
Distance to highway	Negative	[41]
Distance to heavy industry	Negative	[41]
Presence to high rise / view obstruction	Negative	[18]
Crime rate	Negative	[20]
Unemployment rate	Slightly negative	[19]
Population density	Positive	[39]
Presence of cultural landmarks	Slightly positive	[19]
Birth surplus	None	[40]

Table 2.2: Source: author's summary

2.5 Conclusion

The literature review offers three contributions to this research. First of all, the literature explored the differences between estimating house price through index calculations and hedonic models. While indexation is useful for global predictions, it does not get close to the accuracy offered by more localised hedonic models. The Kadaster Price Index gives a clear overview of global developments in the Netherlands, showing significant differences in price developments for both housing types and regional difference, further suggesting spatial heterogeneity and the necessity of local models. The indices can, however, serve as a benchmark for evaluating the performance of the more refined local hedonic models later explored in the review.

The second contribution of this literature review is the exploration of the benefits and limitations of three models for creating a hedonic price model. This part highlights the necessity of local models that include a spatial component when modelling house prices. A global linear regression model is a poor choice unless a dummy location variable is included to account for spatial heterogeneity. Geographically weighted regressions (GWR) on the other hand specifically account for spatial variations which influence the coefficients of variables. Furthermore, a further improvement on GWR is the multi-scale geographically weighted regression (MGWR). MGWR is an improvement over GWR in all identified cases, since it allows for a flexible bandwidth of the kernel function. This means the scale of local effects can differ between variables, which generally leads to a slightly better performing. Finally, XGBoost is identified as a potential model for modelling the more complex non-linear relationships of location characteristics. In the end, the four models offer a trade off between additional layers of complexity which can in turn result in better predictions.

The review further explored (hedonic) house price models in practise. There already exist applications of such pricing models in the Netherlands, both for commercial purposes (Calcasa) as well as governmental purposes (WOZ-waarde). Despite this, not much is known about the exact type of models these business and organisations use. Further highlighting the academic relevance of this research. From the discussed examples, it can be seen that there exists a market for house price models in the Netherlands. All these models rely on systems that match sale prices of similar houses based on their characteristics. The hardest part is getting a large enough sample of house sales. Then, the only hurdle that remains is to gather as much physical / neighbourhood characteristics, which is commonly done through public sources as well as surveys. In the end, whoever has the most, but also accurate, data will ultimately be able to make the best prediction.

Finally, the third contribution of this review is an overview of characteristics for house prices. The three characteristics are divided into three categories: market

characteristics, location characteristics and intrinsic characteristics of the house. The market characteristics mostly explain temporal variances between different years and usually have a global effect on all houses. As such, they can be excluded from the model when modelling house prices for only the current year. Furthermore, intrinsic characteristics are more commonly used for modelling house prices as they often contribute the most to the sales price. The location characteristics are less important and often have more complex non-linear relationships. However, the lack of intrinsic variables can be compensated to, sometimes, even make better estimations as long as sufficient location characteristics are modelled. Whether or not this is the case for Stater's dataset of the Netherlands, is ultimately what this research aims to discover.

3 | Methodology

This chapter outlines how the design science methodology is applied within this research, resulting in a 5-step approach. Furthermore, additional sub-questions for the final thesis are formulated based on the literature review. As a reminder, the main research question of the thesis is defined as follows: *"How can hedonic price models, based on location and intrinsic characteristics of real estate, serve as an alternative to price indexation to more accurately valuate the collateral (house) of Stater's mortgages in Netherlands?"*

3.1 Application of Design Science Methodology

The Design Science Methodology (DSM) proposed by Hevner et al. [9] presents a set of guidelines for design science research within the discipline of information systems. Their original model can be seen in Figure A.2.1. This methodology fits this research, as the problem of this thesis is inherently a design problem. The artefact that is designed is in this case the prediction model. Below in Figure 3.1, the author's application of the Design Science Methodology is summarised.

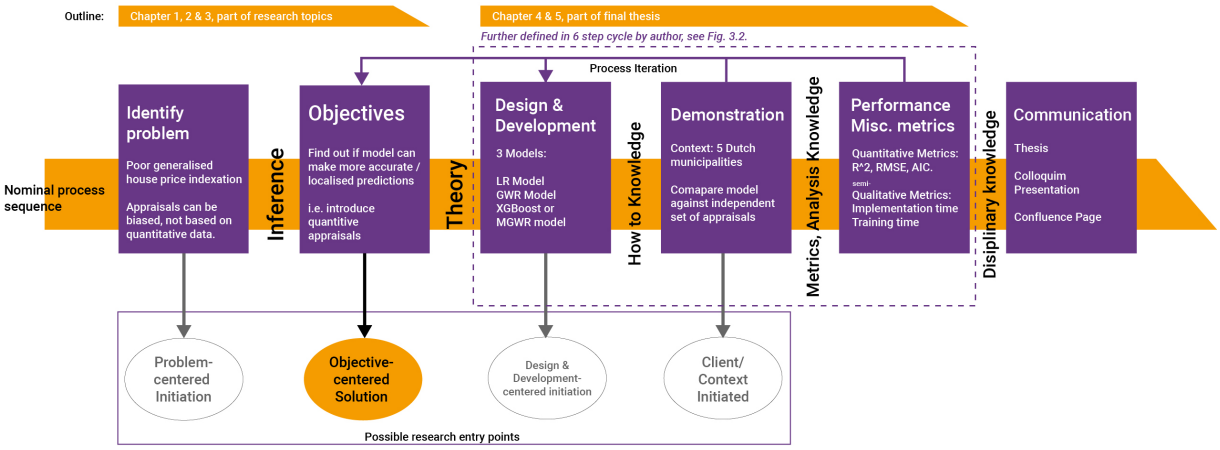


Figure 3.1: Author's application of the Design Science Methodology [9].

This research is approached from an objective centred solution entry point. Chapter 1, the introduction, identifies the problem and objective of the thesis. The objectives of the solution (model) can now be refined based on the results from the literature review in Chapter 2. In the literature review, four models are identified that can be used to model house appraisals: (1) LR, (2) GWR, (3) MGWR and (4) XGBoost. If the results of the GWR model are promising, the more specialised MGWR model will be implemented. Otherwise, XGBoost will be implemented as an alternative approach. If GWR is not an improvement over LR, it means that MGWR will likely also not be an improvement, due

to the similar reliance on linear relationships. This means that in the end, three models are developed: LR, GWR and either MGWR or XGBoost. The chosen models will be compared against the current approach of price indexation, according to the specified solution metrics. The highlighted box in Figure 3.1 outlines the focus of the final thesis. This includes the development, demonstration and evaluation cycle of the models. The cycle is further specified in the 5-step approach of 3.3. Based on the new knowledge from the literature review, the following four sub-questions are defined as part of the solution objectives to answer the main question:

> Q2.1: How do the three models compare in terms of bias and variance when estimating for the five selected municipalities?
>
> Q2.2: To what extent do the influences of the housing and location characteristics differ between the five municipalities?
>
> Q2.3: How do the three models compare in terms of bias and variance when estimating on a national scale when also including municipalities with less data points?
>
> Q2.4: Based on the solution metrics specified in 3.2, what are the disadvantages / advantages of the three models as compared to the current approach of indexing house prices?

3.2 Solution Metrics

The end goal is to discover if the house and location characteristics allow for reasonable predictions of appraisals, and if this is a better approach than Starter's current method of indexation. To formally evaluate indexation vs. the three models, some requirements have to be defined. The three models will both be evaluated using quantitative as well as qualitative metrics.

Quantitative metrics
The quantitative metrics are based upon common **accuracy performance metrics** for machine learning models. First, the R^2 as a measure for goodness of fit. Secondly, the prediction error is quantified by the root mean squared error, or **RMSE**. The RMSE weighs large errors more heavily than smaller ones by squaring them. This is the metric that is often used to optimise regression models. As additional reference, also the **MAE** is calculated, which is the absolute mean average error. The MAE is always lower or equal to the RMSE, as it does not put a heavier weight on larger absolute errors. Finally, the mean absolute percent error, or **MAPE**, gives the relative error. This is helpful as house prices range from €150,000 up to over a million, and as such a

relative error might give a better indication of the accuracy as expensive house will also have absolute larger errors.

Qualitative metrics
A slightly more accurate model is not necessarily a better model for Stater if the maintainability of the model has much higher costs. The qualitative metrics aim to provide better insight into the **operational costs** to implement the model and keep the model up-to-date. These requirements came forth from discussions with Stater. The two main metrics here are: (1) model implementation time: how much time / effort would it need to take to replace the current model, (2) model upkeep: how much time needs to be spent on keeping the model up to date and running (loading new data and training the model). All in all, with the two types of metrics, the models can objectively be compared and a substantiated conclusion can be drawn whether the proposed approach can replace the current approach.

3.3 Solution Design Approach

Based on the four sub-questions, a 5-step approach is formulated for the remainder of this research, as seen in Figure 3.2. These steps come from the design, demonstration and evaluation phases of the DSM, highlighted in the box Figure 3.1. The 5-step approach, as seen in Figure 3.2, outlines how the models are iteratively build and evaluated. The approach for this research is based on a similar 5-step process used in the research of Potuijt [21]. How each of the steps is applied within the context of the project is explained below:

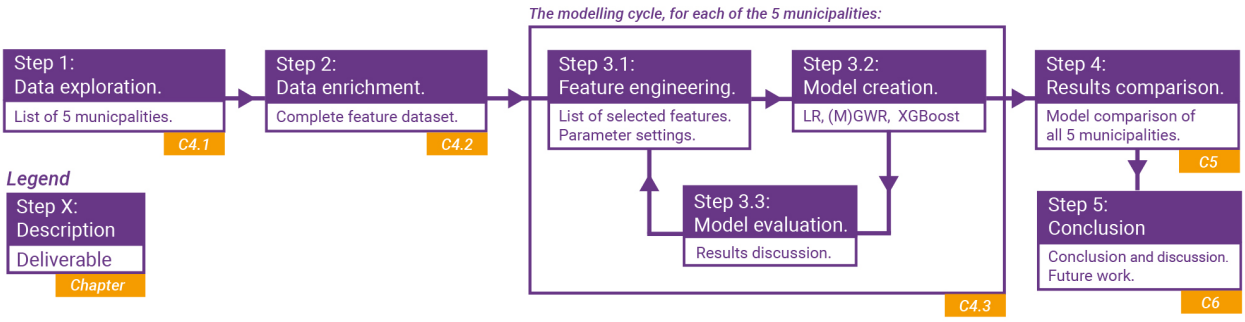


Figure 3.2: The 5-step approach for the solution design part of the thesis.

- **Step 1 - Data exploration:** In the first step, the independent variables, which are the house valuations records from Stater, are explored to identify which regions have sufficient market values to be considered potential areas of interest for creating the model.

- **Step 2 - Data enrichment:** The second step outlines collecting the dependent variables. First, a selection of overview of dependent variables from the literature review in Chapter 2 is made. This selection is based on which variables are publicly available for the entire Netherlands, such that the model can later be applied to different municipalities as well. The result is an overview of found sources per variable. Finally, the all found dependent data variables are joined with the independent data variables. This will result in a single feature dataset which contains all the features that can then be used as input for the model.
- **Step 3 - The modelling cycle** The third step is iteratively building the models for each municipality and collecting the results. First, for the LR model, the relevant features are selected. After building the model, redundant features are removed to improve accuracy. Finally, if the selected variables are satisfactory, the GWR will be developed. Finally as discussed in 3.1, if the results of the GWR model show a significant improvement, a MGWR model is developed. Else, the XGBoost algorithm will be developed. Additionally, each model is evaluated using different parameter settings. The cycle of step 3 is fully outlined in Figure 3.3.
- **Step 4 - Results comparison** The fourth step compares the results from each of the three models, and evaluates if the results show any statistically significant difference between municipalities. The hypothesis is that there indeed will be differences in the weights between each region.
- **Step 5 - Conclusion & Discussion** The fifth step concludes the research by answering the four sub-questions to finally give an answer to the main research question. Finally, it discusses the reliability of the results and recommends areas for future work.

3.4 Conclusion

The main contribution of this chapter is the 5-step approach presented in 3.3, which is centred around the four sub-research questions specified in this chapter. Each step corresponds to a section of the remaining chapters in this thesis, as is highlighted by the orange boxes. The 5-step approach is based on the Design Science Methodology. This approach is chosen in this thesis as the main research question is inherently a design science problem. Finally, the quantitative and qualitative metrics of 3.2 will help decide if the model is a suitable improvement over the current method of indexation at Stater. As specified in the 5-step approach, the next chapter, Chapter 4, outlines the data exploration and enrichment process, as well as the specification of each model and how they are realised.

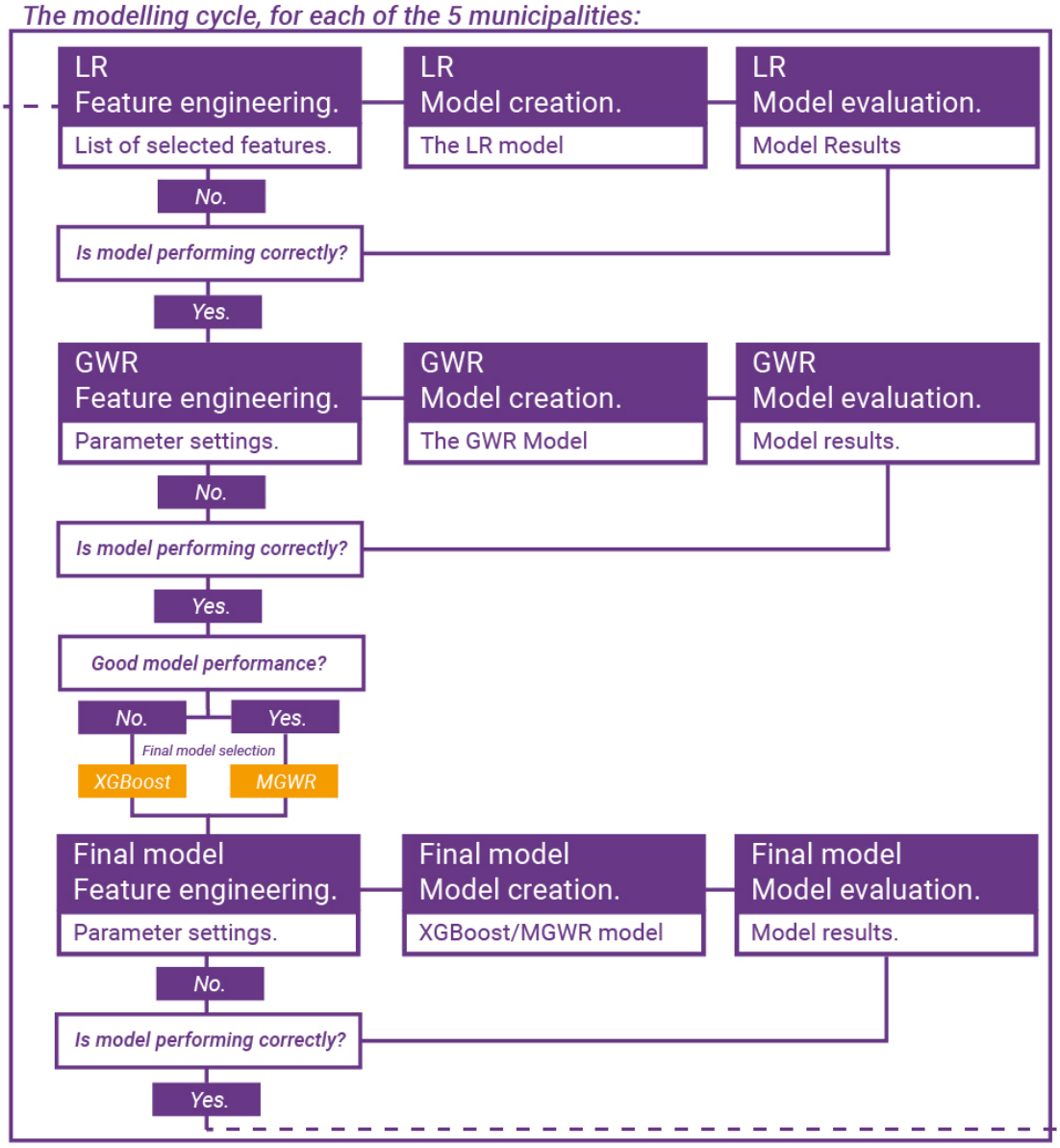


Figure 3.3: Zoom-in on step 3: the full modelling cycle in more detail.

4 | Solution Design

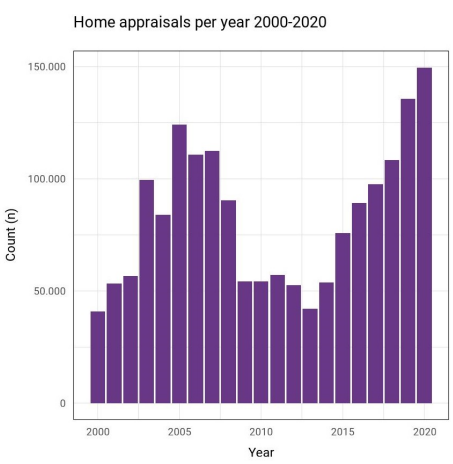
This chapter outlines the design of the final appraisal models. It covers the first three steps of the approach outlined in the Methodology (Chap. 3). The first step is the exploration and collection of appraisal values of the mortgages managed by Stater. Furthermore, this analysis motivates the decision for the two years and the five municipalities on which the models will be trained and tested. The second step is the enrichment of the appraisal values with external data sources. These variables are used as additional independent variables to build a better predictive model. Finally, the third step outlines the iterative process of building and improving the three models. These models are: linear regression, geographically weighted regression and XGBoost. The chapter concludes with an overview of the final models that have been realised.

4.1 Step 1: Data Exploration of the Appraisal Values

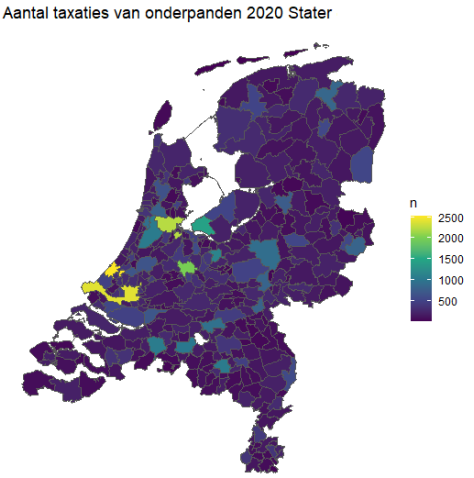
Each mortgage application in the Netherlands needs an official appraisal by a certified appraiser. This means that Stater has a home appraisal value for every mortgage application. The dataset used in this thesis is a combination of two different databases: the mortgage applications and accepted mortgages. These are stored separately as not every mortgage managed by Stater goes through their approval process. Some money lenders opt to handle the approval themselves and only employ Stater for the long-term management. Furthermore, a specific house might have several applications before finally being accepted. Thus, the two datasets are combined to have as much mortgage appraisals as possible. Any duplicates that arise are removed by only keeping the appraisal of the accepted mortgage. The result is a total of 1.135.896 appraisals.

The total number of real estate appraisals per year is given in Figure 4.1a. It highlights that the total amount of appraisals varies per year. For example, around the financial crisis of 2007–2008, there were a lot less mortgage applications. On the other hand, recent years have more mortgage applications due to the increasing demand on the Dutch Housing market. Additionally, Figure 4.1b shows that the number of appraisals varies per municipality. This appears to be roughly correlated with the population density of the Netherlands, where larger municipalities have more appraisals. Figure A.3.1 in the appendix shows that this distribution remains similar across years. In years with few mortgage applications, such as 2008, many smaller municipalities only have around 300 appraisals, which is only a small fraction of their total amount of houses. For these regions it is harder to make accurate predictions. Instead, the thesis focuses on five large municipalities, namely Rotterdam, Amsterdam, Eindhoven, Amersfoort and Groningen. If the models make predictions with good accuracy for these five regions, then they already cover a large percentage of Stater's dataset.

Additionally, these regions are specifically chosen to represent some of the different



(a) Appraisals per year, 2000-2020



(b) Spatial distribution for 2020

Figure 4.1: Number of home appraisals at Stater.

provinces of the Netherlands. The assumption is made that, if the models for these municipalities are accurate, then these models are likely more accurate predictions than using indexation for those specific provinces. If the models cannot make good predictions, then trying to build a model that also includes the smaller municipalities will yield no better results. In the end, Rotterdam, Amsterdam, Eindhoven, Amersfoort and Groningen were among the largest municipalities regarding number of appraisals (Fig A.3.1) and thus make for suitable test regions.

Municipalities	**# of appraisals (2018)**
Amersfoort	1494
Amsterdam	5084
Eindhoven	1845
Groningen	1160
Rotterdam	3011

Table 4.1: Number of appraisals for chosen municipalities (2018).

It is a common conception that house prices differ per province in the Netherlands, this notion is supported by data from the CBS in Figure A.1.1. Similarly, the average appraisal value in Stater's dataset also varies per municipality and also in time. For the appraisals values of 2000 and 2020, an increase in the number and average appraisal value can be seen between 2000-2020 (Figure 4.2). This means that a complete prediction model for appraisal values would need to discern the differences both in time and regional location. However, the goal of this research is not to explain the differences between years and predict future appraisal prices for houses, which is a more difficult

task requiring a different approach. In Stater's scenario, only the current value of a mortgage collateral is what matters the most. As such, it is not a problem to only train the models for a specific year. In this research, the models are trained on data from 2018 and 2020. 2020 is chosen because this is the most recent complete year. Additionally, 2018 is chosen to validate the model for a different year with less appraisals. There is an additional reason 2018 is chosen (instead of 2019) as this year has the most complete set of external independent variables, which is further discussed in 4.2. For 2018, the number of appraisals for these 5 municipalities is summarised in Figure 4.1. This concludes the overview and motivation behind the 5 chosen municipalities and two specific years for which the appraisal values will be modelled.

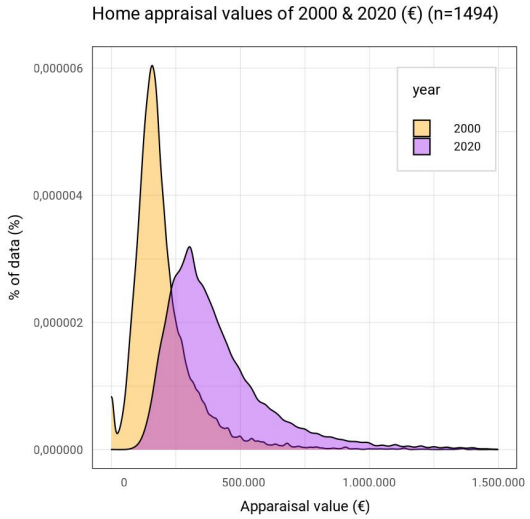


Figure 4.2: Increase in average real estate appraisal (€) between 2000 and 2020 for Amersfoort.

Besides the appraisal values, there are only a few variables in Stater's own dataset which describe the house belonging to a mortgage. Most variables such as purchase price, energy label, or living area contain largely missing values due to them not always being mandatory for the application process. The only relevant and complete variable is the housing type (code: 'kd_ondrpnd_oms'), shown in Figure 4.3. It appears that the more specific housing types such as a corner house, farm or free-standing house are almost never chosen. Most houses get generalised either under apartment or single-family house. There are further specifications of these two categories such as 'with garage' or 'with parking space'. Since the chosen models (LR, (M)GWR, XGBoost) only deal with numerical / ordinal variables, this categorical variable is transformed using one-hot encoding. One-hot encoding transforms a categorical variable to multiple new binary 0/1 variables. The end result is three new variables: 'is_gzns': to indicate if the house

is a family house (1) or apartment (0), 'garage': to indicate the presence of garage yes(1)/no(0), and 'parkeerplaats': to indicate the presence of parking space yes(1)/no(0).

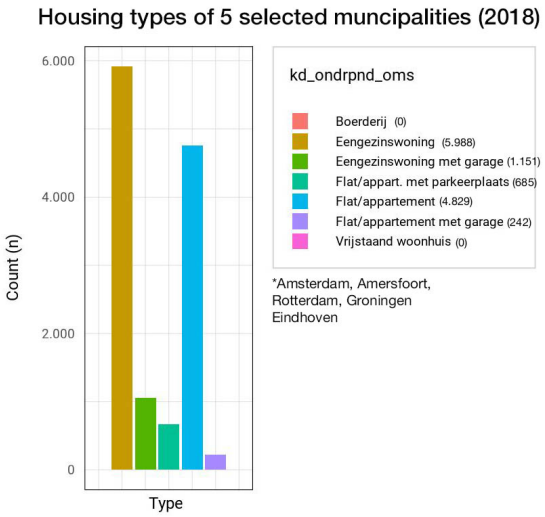


Figure 4.3: Different home-types

In conclusion, step 1 has provided a dataset of appraisal values for the 5 chosen municipalities. This is a combination of both the mortgage applications as well as accepted mortgages. The dataset consists of all appraisals from the year 2020, with appraisals from 2018 serving as an additional validation set. Additionally, from Stater's own data, three variables are derived which indicate if the appraisal belongs to a family home or apartment and if there is a garage or parking space present. This is not enough to explain the variation in appraisal values. Thus, step 2 discusses the enrichment of the dataset, using variables from external datasets.

4.2 Step 2: Data Enrichment of Independent Variables

This section presents the main novel approach of this research, where external data sources are combined to create a better understanding of the type of house and the neighbourhood it is located in, which allows for better predictions of the appraisal values. Currently, the training dataset only contains the appraisal value and three variables related to housing type. This is likely not enough to make accurate predictions. While Stater does have more information about each house in the official appraisal reports, this data is not usable for this research, as it is stored as unstructured data in PDF files. As such, additional information is collected from from outside sources.

The largest collection of public data in the Netherlands is 'Data Overheid'. This website, initiated by the government, is a place where publishers of open data can list their datasets. The platform is meant to be the centralised overview of all governmental

data in the Netherlands. However, one problem that is not solved is that municipalities still publish their own datasets separately. It is difficult to find a single dataset that can be used for all houses in the Netherlands, as many datasets apply to a specific city or province. However, large national datasets do exist. They are mostly published by large (semi-)governmental organisations. The four chosen datasets come from three parties: Kadaster, CBS and 'Rijkdienst voor ondernemend Nederland' (RVO). An overview of the datasets is given below in Figure 4.4.

Dataset name	Contents	Joined using	Source
BAG: 'Addresses and Buildings key register'	Geo-coordinates, build year, surface area	Address	Kadaster
DKK: 'Digital cadastral map'	Land lot area	BAG-VBO-ID	Kadaster
CBS Square statistics (NL: 'Vierkantstatistieken')	Variables for area's of 100x100m & 500x500m	Geo-coordinates	CBS
EP-Online	Energy labels	BAG-VBO-ID	RVO

Figure 4.4: External data sources for additional housing characteristics.

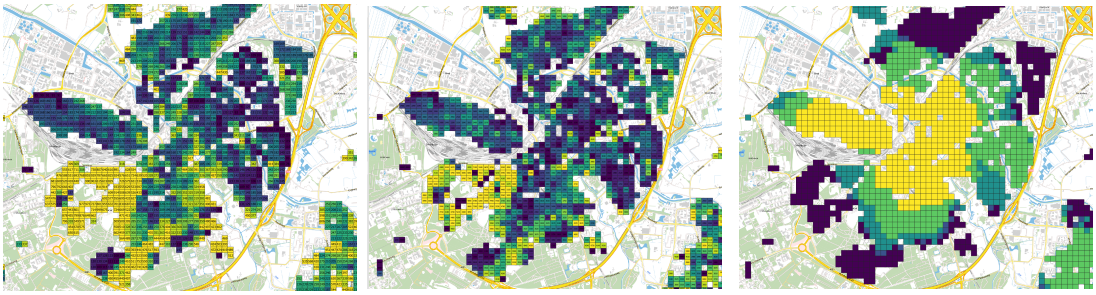
As mentioned in Chapter 2, the Kadaster maintains the central registry related to land ownership in the Netherlands [42]. The BAG registry contains data on all buildings in the Netherlands. Extracting the necessary information requires some work, as the data model is relatively complex to fully capture the situations that arise in practise. In this thesis, the only relevant types of buildings are houses, not offices or other types of buildings. In the BAG, these buildings have a so called 'living' designation. To get an address, with the corresponding building the following elements are joined: 'verblijfsobject' (EN: liveable space), 'pand' (EN: building), 'nummeraanduiding' (EN: adres), 'openbare ruimte' (EN: public space, e.g. street), 'woonplaats' (EN: city) and 'woonplaats-gemeente relatie' (EN: city-municipality relation). The relationship between these elements in the BAG data model is visualised in Figure A.3.2. The difference between a 'verblijfsobject' and 'pand' is made clear when looking at an apartment building. There is a single building ('pand') and multiple individual apartments which are all 'verblijfsobjecten'. In Stater's own dataset, some houses only have a zip code with a house number. The BAG allows for joining the corresponding city, municipality and province for each unique combination of house number and zip code. Furthermore, the BAG provides with extra information about each house, namely the house's build year and living area (total floor area, not only ground floor area), which are important characteristics of the house itself.

Besides information about the actual houses, the Kadaster also has information about the boundaries of all land lots in the Netherlands, stored in the DDK [43]. As literature has shown, lot area is less important than the living area but still of influence on house prices. Especially in the city centres, more garden space is especially valuable.

Finding the corresponding land lot that belongs to a house, is not a simple join operation between the 'verblijfsobject' of the BAG and land lots of the DDK. In practice, houses can belong to multiple land lots. When a house is sold, all corresponding land lots are sold with it, thus all land lots contribute to the house price and appraisal value. However, in the DKK, not every land lot is directly related to an address or BAG object, as land lots are usually only registered by ownership (name), which is private information. Additionally, even with the required ownership info, an owner can own multiple houses, which does not solve the problem of finding all related land lots belong to a house. The solution to this problem is the LKO table of the Kadaster.

For this research, the Kadaster has provided the 'Locatie Kadastraal Object' (LKO) table, which is a dataset that links land lots from the DDK to the buildings from the BAG. The most recent version of this table is not publicly available, but the Kadaster has provided a special extract specifically for this research. In this table, the 'verblijfsobject id' from the BAG can be linked to a land lot from the DDK. However, there is an additional field specifying the type of join. This field has three possible values: 'Geographic', 'Administrative' or 'Both'. Geographic means the house ('verblijfsobject') physically overlaps with the land lot. However, due to small inaccuracies in the measurements of land lots, two houses side by side might overlap. As such, there is also the administrative relationship, which means an explicit relationship exists in the deed of the house. Based on the evaluation of some samples, the conclusion is drawn that only the relationships of type 'administrative' or 'both' should be considered valid. All in all, after joining and computing the combined surface area of all land lots, on average 69.3% of all family homes have an associated land lot area. For all apartments that are missing a land lot, as a zero is filled in as apartments generally do not have a land lot. A scatter plot of the Kadaster variables is given in Figure A.3.3, which shows a strong relationship between the appraisal value for both the living area and the land lot area. Finally, the overall percentage of missing records for this variable is summarised in Figure 4.2 under 'Land lot area'.

The next dataset, is called 'vierkantstatistieken' (EN: 'square statistics') and comes from the CBS [44]. The CBS publishes many sociographic and demographic variables about the entire Netherlands. They publish this data for different levels of resolution. From highest resolution to lowest resolution, the following sets are published: full postal code (PC6), 100x100m tiles, 500x500m tiles, 4 character postal code (PC4), and neighbourhoods & city blocks (illustrated in A.3.4). The most detailed level is PC6, this data is grouped by the entire postal code, e.g. AAAA11. However, this dataset is behind a paywall, so it will not be used in this research. Thus, the next highest resolution is the 100x100m tiles from the 'vierkantstatistieken' dataset. One of the main advantages of the tile dataset is that their size and geographical position remains constant throughout the years. On the other hand, city neighbourhoods and even municipalities can merge,



(a) WOZ-Waarde [€1k] (b) Electricity usage [kWh] (c) Nearest cafe [km]

Figure 4.5: Various CBS 100x100m statistics (Amersfoort, 2018)

split or change borders. This means the values of the variables describing this area change, and as such they are meaningless for answering questions about the changes between years. This makes the 100x100m tiles dataset a good choice. Figure 4.5 gives an example of three variables for Amersfoort (2018).

Joining the tile dataset is possible using the geo-coordinates that have been collected from the BAG. However, not every house lies within a tile. The main reason is that tiles with less than 5 households have their values censored due to privacy reasons. This problem was mainly an issue with demographic variables, such as the number of people aged between 0-14 years, 15-24 etc. and the average WOZ-Waardes (automated indication of house value used for taxation). For variables that refer to amounts, such as the number of people aged between 0-14 years, it is not possible to mix the 100m and 500m tiles datasets. On the other hand, for average values (such as average income, or energy usage), it is possible to use the 500m tiles instead of the 100m ones, since 500m tiles will just give a more generalised average of a large sample. For average income and energy usage, table 4.2 quantifies how large the subset of data is that has the missing values of 100m tiles replaced with 500m tiles, this is on average 5% of the total number of observations.

Furthermore, inside the CBS dataset there are many variables which list the distance to nearest 'X' or the amount of 'Y' within a certain radius of the tile. These are abbreviated respectively with 'AFS' (for 'Afstand', EN: Distance) and AV## (where ## specifies the radius in km.) The X and Y refer to places such as, grocery stores, cafes, swimming pools, hospitals, cinemas and more. The 'distance to' and 'amount within radius' variables that describe the same type of building, end up being highly correlated. As such, only the 'distance to ...' variables are included. To summarise, the total variable overview of A.3.7 in the appendix lists the descriptions of all variables and which tile set they use (variable names ending in _100 or _500).

Additionally, based on the geo-coordinates from the BAG, it is possible to calculate the distance to the city centre for each house. The coordinates of the city centres are

manually determined using Google maps. For the five municipalities in this research, this is still do-able by hand. However, for the entire Netherlands a different solution must be found. The resulting variable is called 'dist_centre'. In the end, the distance to the city centre variable turns out to also correlate with the CBS distance variables. Take for example distance to nearest cafe. As can be seen in 4.5c, there is a relationship between the distance to cafe and the distance to the city centre of Amersfoort. For linear regression, correlated variables have to be removed, else the model can become unstable.

Despite removing the 'amount within radius' variables, there still exists a correlation issue. Some of the 'distance to' variables, as well as the city centre distance, are correlated with each other, see the correlation plot in figure 4.6. The boxes highlighted in purple indicate a correlation factor of 0.75 or higher (strong correlation). The rest of the non-significant correlations is crossed out. As such, the following variables are removed: distance to daily necessities (in favour of distance to supermarket), distances to cinema, museum and podium (in favour of distance to nearest train-station), distance to hospital and pharmacy (in favour of distance to general practitioner), distance to cafeteria (in favour of distance to cafe) and finally, as outlined the paragraph before, distance to city centre.

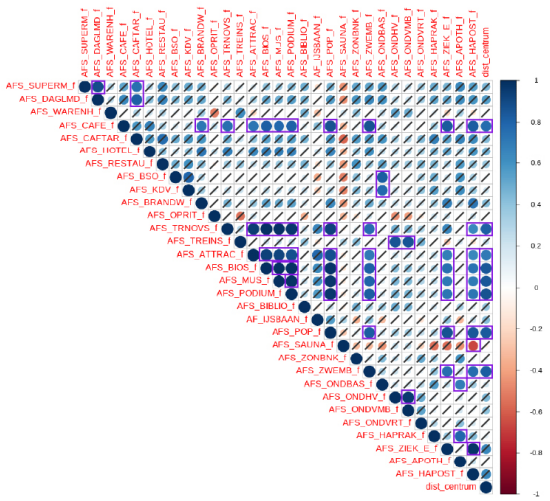


Figure 4.6: Correlation plot of 'distance to nearest ...' variables of CBS (Amersfoort, 2018).

Finally, the 'Rijkdienst voor ondernemend Nederland' (RVO) publishes a dataset containing all official energy label registrations in the Netherlands [45]. This data can be joined with the existing dataset using the 'verblijfsobject ID' from the BAG. This dataset also has its limitations, as not every house has an official energy label. In the past it

was not mandatory to have an energy label when selling a house. The RVO dataset only contains registrations, so not every house is present in this dataset. In addition to the energy label, the dataset also contains more detailed information on the house type and energy usage. However, due to many houses not being present in this dataset, the existing house type from Stater is used, as well as the average energy usage from CBS. In the end, the energy label is available for 70% of the houses (Table 4.2), for an example distribution see A.3.6.

The complete collection of variables is summarised in A.3.7. However, there still are variables that have missing values. As has been referenced before, the number of missing values are summarised in table 4.2. Here 'Distance' refers to the distance variables of the CBS dataset. The variables not included in this overview are 100% complete. For the CBS, a large number of missing variables were resolved by also including the 500x500m tiles, the number of records that uses values from the 500x500m dataset is summarised in 4.3.

An additional small issue concerns the fact that all variable are available for 2020. The most recent fully complete year is 2018. For 2020, some of the variables related to income and the 'distance to ...' are not yet available. However, it is safe to assume that most of these variables have only changed little in the last two years. As such, the dataset for 2020 can still be created by taking the missing variables from 2018.

# of missing records (% of original data), imputed using KNN (n=7).							
Municipality	*Build year*	*Land lot area*	*Address density*	*House -holds*	*Energy usage*	*Distance*	*Energy label*
Amersfoort	4 (0.27%)	451 (30.19%)	15 (1.00%)	16 (1.07%)	28 (1.87%)	15 (1.00%)	454 (30.39%)
Amsterdam	116 (2.28%)	731 (14.38%)	0	71 (1.40%)	127 (2.50%)	0	1391 (27.36%)
Eindhoven	16 (0.87%)	659 (35.72%)	0	93 (5.04%)	2 (0.11%)	2 (0.11%)	587 (31.82%)
Groningen	25 (2.16%)	382 (32.93%)	0	97 (8.36%)	15 (1.29%)	2 (0.17%)	312 (26.90%)
Rotterdam	1 (0.03%)	732 (24.31%)	0	39 (1.30%)	17 (0.56%)	0	948 (31.48%)

Table 4.2: Number of missing records for incomplete variables (2018).

Removing all the records with missing values is not an option, as a large portion of the records have at least one or two variables missing. The result would be a dataset consisting only of a few hundred records per municipality. Instead, the unknown values are imputed from similar records. This is done using 'k-nearest neighbours imputation'

# of observations taken from 500x500m instead of 100x100m		
Municipality	WOZ-Waarde	Income
Amersfoort	96 (6.43%)	74 (4.95%)
Amsterdam	398 (7.83%)	259 (5.09%)
Eindhoven	171 (9.27%)	88 (4.77%)
Groningen	138 (11.90%)	84 (7.24%)
Rotterdam	216 (7.17%)	101 (3.35%)

Table 4.3: Number of observations taken from 500x500m instead of 100x100m, for WOZ-Waarde & income variables (2018).

with 7 neighbours. The number of neighbours is based on the fact that appraisal reports commonly use around 5 houses as reference houses. Before imputing the values, first the variable columns are sorted from least missing values to most missing values. This is important, as the variables with the least missing variables need to be imputed first, since their imputed values are used for finding similar neighbours / records for the other missing variables. In the end, 'k-nearest neighbours' picks seven similar houses and uses their average to compute the missing value.

In conclusion, in step 2, four external data sources from Kadaster, CBS and RVO are used to gather a total of 31 usable variables. The total overview of variables is presented in figure A.3.7 in the appendix. The Kadaster mainly provides intrinsic characteristics about the house, while CBS provides the location characteristics about the neighbourhood. Additionally, RVO also provides the energy labels for a large percentage of all houses. However, not all available variables are used. Figure A.3.8 summarises the 22 variables that are not included because of high correlation with other variables or being used to derive other variables. Finally, there is the issue of missing values as shown in figure 4.2. The two largest variables with missing values are the land lot areas and energy labels, which have up to 30% missing values. The missing values are imputed using 'k-nearest neighbours' with 7 neighbours to prevent throwing away the majority of records. This complete dataset is used in step 3 to realise three prediction models.

4.3 Step 3: Modelling cycle

The result of step 2 is a dataset for the five chosen municipalities consisting of 31 features about the house, its location and its neighbourhood. Step 3 is the final step of this chapter which outlines the creation of three models: multiple linear regression, geographically weighted regression and XGBoost. Furthermore, it covers the steps which are taken to validate if the models are implemented correctly. The data for the

models is split into 75% training data, and 25% test data. The RMSE (root mean squared error), is the main metric used for comparing the performance of all three models. Finally, the chapter outlines how repeated k-fold cross-validation is used to optimise the hyper-parameters of the geographically weighted regression and XGBoost models.

4.3.1 Multiple linear regression model

The multiple linear regression model uses multiple independent variables to model a linear relationship. For a dataset consisting of n observations, the model is expressed as:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i + \varepsilon \qquad \text{for i = 1 ... n}$$

Where...

- y_i is the random independent / response variable.
- β_0 is the intercept.
- $\beta_1 ... n$ is the slope, or influence for each independent variable x_i .
- ε is the error term, also called random error.

Linear regression is fitted using ordinary least squares (OLS). OLS fits the model such that the squared error is minimised. The slope β_1 is estimated as:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \tag{4.1}$$

And the intercept β_0 is estimated as:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{4.2}$$

For regression analysis, there are four key assumptions that are made to allow for valid inference of the results:

- **Linear relationship -** The relationship between the dependent and independent variables can be modelled by a linear relationship. This means the relationship can be modelled as a straight-line, consisting of an intercept and slope.
- **Random sampling -** The dependent variable must be a random variable. This means the data sample must be drawn at random from the population.
- **Normal distributed error terms -** The error terms are normally distributed with a variance of ρ^2 and a mean error $\mu = 0$.

- **No heteroscedasticity -** No heteroscedasticity means there is no relationship between the predicted dependent variable x_i and the error terms. In this case, it means the error term ε does not grow larger for higher appraisal values.

The implementation
This thesis uses the 'lm' (linear model) method from the 'stats' R-package. When building linear regression models with many variables, it is important to check for multicollinearity. This means one or more features in the dataset can be used to predict one of the other features, due to a linear relationship between them. In step 2, many of these collinear variables were removed using a correlation scatter plot.

Besides the correlation matrix, there is also the variable inflation factor (VIF) which helps identify collinear variables. The VIF can only be computed after the model is built, as it indicates the ratio between the total variance explained by the model and the variance explained by a model only including that single independent variable. Thus, a higher VIF corresponds to a more collinear variable. A VIF of 2.5 is approximately equal to a correlation score of 0.6, which indicates moderate correlation. On the other hand, a VIF above 5 is an indication of high correlation to one of the other variables. Table A.3.2 shows the VIF scores of the variables that have been selected.

There are still a few variables (marked in bold) that are highly collinear, but the cause can be explained. *OAD*, the address density, is collinear with *STED*, the urbanisation factor. This makes sense, as city centres are often the most densely populated areas. The additional correlated variables include the demographic variables about age. The percentage of people over 65 naturally depends on the percentage of people between 0 and 14 years old. Finally, the percentage of incomes belonging to the bottom 20% is related to the percentage of high incomes. Taking out these final variables is necessary if one wants to explain the independent effects of these variables on the appraisal value. However, when the goal is purely predictive accuracy, then it is less of an issue. Besides, multicollinearity is only a potential issue for the linear regression model, not the GWR or XGBoost model.

Additionally, the first iterations of the linear model had relatively poor performance ((r^2) 0.7) with an average prediction error of 11.81% deviation (MAPE). This is largely caused by the most expensive houses (i.e. highest appraisal values), as can be seen in Figure 4.7. As an example, the sample X1244 shows a deviation of €400k. By limiting the model to only the majority of the houses below €750.000,- the model performance increase to (r^2 = 0.785) and MAPE: 9.86%. This large performance boost , due to only taking out less than 1% of the samples, shows that these unique houses can be considered outliers. The more expensive houses are harder to model due to less demand and, as such, there is more influence of the individual preference of buyers. The appraisal outliers above €750.000,- are therefore excluded from further models, as no other variable is able to account for the variance of these houses.

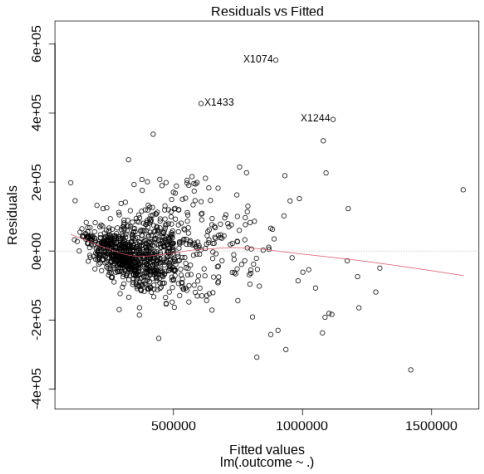


Figure 4.7: Initial LR model showing large deviations for high appraisal values.

The final iterations of the LR model used log-transformed variables. If the unexplained variance of the more expensive houses is due to personal preferences of the buyers, than perhaps the influence of living space and other variables diminishes. By taking the logarithm of either the appraisal values and/or the dependent variables, their effect becomes a percentage of influence (i.e. increase of 1% in appraisal value per square meter). In the end, neither log-transforming the variables, or predicting the appraisal value per m^2 improved the RMSE of the overall model.

4.3.2 Hyper-parameter optimisation using CV

Unlike LR, GWR and XGBoost have model parameters which can be optimised. This is done using *N times repeated k-folds cross validation*. With k-folds cross validation, only the training dataset (75% of original dataset) is used, which is split into is split into *k* parts. Each time, one part is chosen as the test set and the other parts as the training set for evaluating the parameters. This is repeated until every part has been used as the testing set. The data is the shuffled and split into new k-folds, doing the cycle all over again for N times. The performance metrics of all models is then averaged. This means that a total of N x k models are evaluated per combination of parameters. In this thesis, 4 folds (k=4) are repeated 10 times (N=10), due to the small sample size (1k training samples) per municipality. Each fold is thus approximately 750 samples for tuning the parameters and 250 for evaluating. Using (repeated) k-folds cross validation helps reduce over-fitting and a better picture of the real performance.

4.3.3 Geographically weighted regression model

Geographically Weighted Regression works by computing multiple regressions at different locations, for example using a grid or neighbourhood boundaries. The GWR model weights data points by their distance to the location for which the regression is being computed. Each area ends up with its own intercept and influences for each of the variables.

The global regression model is expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{m} \beta_j(u_i, v_i)x_{ij} + \varepsilon$$

Where...

- y_i is the random independent / response variable.
- β_0 is the intercept for some geographical location (u_i, v_i).
- $\sum_{j=1}^{m} \beta_j$ is the sum of the slopes, or influence for all weighted data points m, at some location (u_i, v_i), for each independent variable x_i.
- ε is the error term.

The implementation
GWR model calibration uses an iterative back-fitting procedure; thus, the computational overheads are high when handling a large number of observations. This makes it less ideal when the model needs to be trained for the entire Netherlands. However, training it on only 5 municipalities takes less than a few minutes. This thesis uses the 'gwr.model' from the 'GWmodel' R-package. The most important aspect of the GWR model is the kernel function which decides weights of the data points which are considered when computing the regression.

There are three parameters related to the kernel function that are fine-tuned. The kernel function itself, the kernel bandwidth and 'adaptive' setting. The kernel function determines the shape of the kernel, an example is given in Figure 4.8. Gaussian, boxcar and bi-square (more dense than Gaussian), were most commonly used in the literature [27][46]. For the bandwidth, the R-package provides a function which utilises golden-section search to iteratively work towards the most optimal bandwidth.

Finally, there is the 'adaptive' setting. The bandwidth of an adaptive kernel depends on a percentage of data points instead of a fixed distance. This means the bandwidth can be larger for regions with few samples. For regions with samples that are spatially not spread out, this usually leads to better results. In the end, the cross validated results show that the adaptive Gaussian kernel also performed better in all cases for this thesis. Although it should be noted that, in case of Amersfoort and Eindhoven, a fixed

kernel could provide equal results. Interestingly, Groningen has a larger bandwidth. One possible explanation is that Groningen has the least samples of all five municipalities, as such it would need a larger bandwidth to have the same number of samples for a regression.

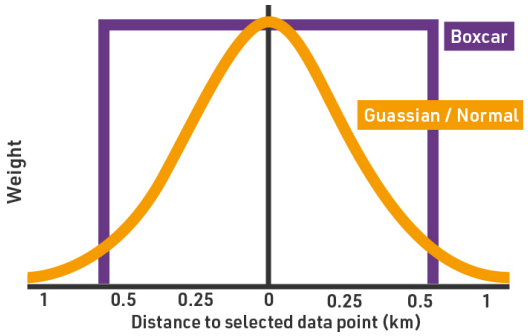


Figure 4.8: Gaussian and boxcar, two examples of possible kernel functions for GWR.

Best kernel settings per municipality (2018)		
Municipality	*Kernel (Bandwidth)*	*Adaptive/Fixed*
Amersfoort	Gaussian (0.28)	Adaptive
Amsterdam	Gaussian (0.19)	Adaptive
Eindhoven	Gaussian (0.27)	Adaptive
Groningen	Gaussian (0.43)	Adaptive
Rotterdam	Gaussian (0.25)	Adaptive

Table 4.4: Best kernel settings for GWR model (2018).

4.3.4 Extreme gradient boosting model (XGBoost)

After the parameter optimisations of GWR, it is clear that it is an improvement over LR (as can be seen in the results of Chapter 5). Despite this, it is decided to implement the XGBoost model in favour of MGWR. MGWR uses the same approach as GWR, thus it is likely to only give a small improvement over GWR. On the other hand, XGBoost uses a different approach that could potentially be better suited to deal with the nonlinear relationship of the CBS variables.

Extreme gradient boosting is a fast implementation of regularised gradient boosting. Gradient boosting creates an ensemble of discussions trees. These trees can be weak predictors on their own, but with an ensemble of trees tuned to various weights, good predictions can be made. This also allows gradient boosting to capture non-linear relationships.

The 'extreme' part comes from the many optimisations algorithms used to optimise the gradient boosting for large datasets: among others this includes XGBoosts unique decisions trees, weighted quantile sketches, sparsity-aware split finding and parallel leaning. An advantage of XGBoost over GWR is that it is possible to train the model using the GPU, which drastically increases performance. This thesis does not go further into the details of XGBoost, but rather discusses which parameters are optimised in this application of XGBoost.

The implementation

In this thesis, the official 'xgboost' R-package is used. The parameters that are optimised are the learning rate (eta) and the max tree depth. A higher learning rate means the model takes larger steps towards a minimum of the loss function. The tree depth determines the number of levels / splits in each tree, a larger value results in a more complex model that can capture more complex relationships, but is also more likely to over-fit. Each parameter is a numeric value which can take on a range of values. The default settings from the official documentation are used as a starting point for the model. A grid of hyper parameters is constructed by defining a range of possible values for each of the parameters, as is listed below. This grid contains a total of 155 rows, corresponding to the 155 unique parameter combinations.

- Eta (η) - Learning rate, factor to which the output of each new tree is scaled. (default=0.3, min=0.1, max=0.4, step=0.01)
- Max_depth - Max number of levels per tree. (default=3, max=7, min=3, step=1)

The other parameter settings that are fixed are the number of rounds (n=2000) and the objective: minimise RMSE (using 'regression:squarederror'). The number of rounds specifies the number of boosting iterations. Additionally, using the early_stopping_rounds parameter set to 250, the model stops optimising if no improvement is made after 250 rounds. Using 4 fold cross-validation repeated 10 times, the final parameter settings of the XGBoost model are determined. The final best settings are: $\eta = 0.15$ and max depth = 6.

Each of the five models has $0.13 < \eta < 0.17$, as such, they were averaged to 0.15, since the end goal is to create a single model for the entire Netherlands. This had a negligible impact on the RMSE. Similarly to the tree depth, 4 out of 5 models performed best with a tree depth of 7. However, this only improved the test RMSE slightly, while greatly improving the training set RMSE. As such, to prevent over-fitting, a slightly lower tree depth of 6 is chosen. This reduces the test RMSE by only €1015,-. This concludes the construction of the three models (LR, GWR and XGBoost) and the optimisation of the hyper-parameters.

5 | Results

This chapter summarises the results for the final LR, GWR and XGBoost models that are trained as outlined in 4.3. Each of the models is evaluated according to the quantitative and qualitative metrics from 3.2. First, the unique models for each municipality are evaluated for 2018 and 2020. The most important table is 5.4, which shows the averaged performance over each of the five municipalities, for each model type. Second, a single XGBoost model is evaluated that is trained on all five municipalities. Finally, a comparison is made between indexation and the five unique models, where they predict the appraisal values of collateral's belonging to mortgages from 2000.

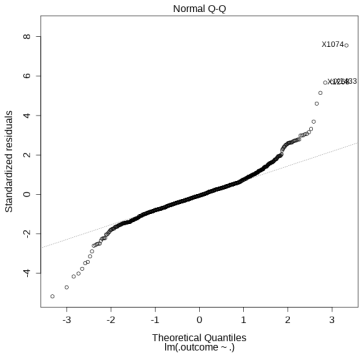
For the multi linear model, the initial model provided a poor fit mainly due to the high variance of high appraisals values. As discussed in Step 4.3, all appraisal value above €750.000 are filtered out to give a better performing model for the majority of the samples. The comparison of the quantile-quantile plots of Figure 5.1 shows that the model with the filtered appraisals contains less severe outliers and that it has a better fit to the normal distribution. The high appraisal values are most likely not good representatives of the total population of houses, and thus they are excluded as they have a large influence on the prediction accuracy.

Additionally, as an additional alternative approach, the appraisal values were logged to model a diminishing influence of the living space. Sadly, both the log-linear model with logged appraisal values and the linear-log model with logged living spaces did not improve model accuracy. In the end, the best performing LR model is the one with the filtered out appraisal values. The LR model has a RMSE €85.628,- and R^2 of 0.785, that is overall an adequate fit. Since the appraisal values vary wildly from €50.000 to €750.000, it is also worth looking at the relative absolute percentage error (MAPE) and simply the mean average error (MAE). These correspond to an average error of 9.61% and €56,219 respectively.

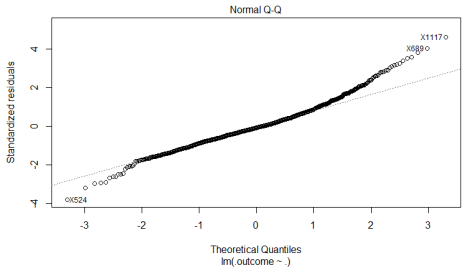
Linear model results (Amersfoort, 2018).				
Metric	R^2	*RMSE*	*MAE*	*MAPE*
LR (all appraisals)	0.709	€150,211	€72,391	11.81 %
LR (appraisals <€750.000)	0.785	€85,628	€56,219	9.61%
LR-LOG	0.768	€89,136	€63,577	10.62%

Table 5.1: Results for linear models (Amersfoort 2018).

It is not surprising the LR performance is adequate at best. Many of the CBS variables do not show a strong linear relationship with the appraisal value. Still, due to the inclusion of the living area (variable name: perceeloppervlakte) and WOZ-Waarde, an adequate model with less than 10% deviation can still be made for Amersfoort. Figure A.3.9 shows that these two variables are by far the two most important factors,



(a) All appraisals, poor fit.



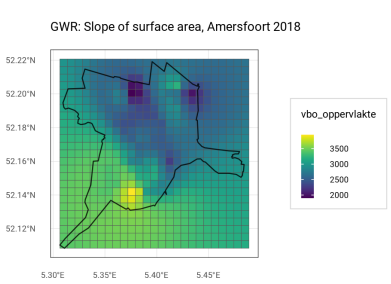
(b) Appraisals <€750.000, adequate fit.

Figure 5.1: Q-Q plot showing impact on overall fit for including all appraisals (Amersfoort, 2018).

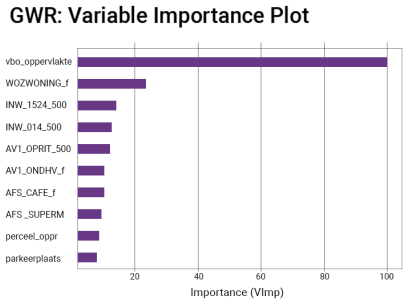
followed by the variable describing high incomes (P_HINK_HH), people aged between 15-24, and build year.

The geographically weighted regression (GWR) provides a better fit than the LR model, as is summarised in table 5.2. As outlined in Step 3, the GWR is trained using an adaptive Gaussian kernel function with varying bandwidths per municipality. Table 5.2 provides a performance overview for each of the municipalities. For Amersfoort, the top 10 most important variables and an example of the spatial influences of the living area, are plotted in Figure 5.2. For the complete overview of all variables and their spatial influences, see Figures A.3.10 and A.3.11.

The most important variable is, again, the living area, followed by the WOZ-Waarde. The variable importance plot appears to have a similar shape as the one for the linear regression (Figure A.3.9). This time, also some of the distance variables such as distance to nearest supermarket and cafe make an appearance. While the influence of the other variables appears to be minor, without their inclusion the R^2 would be lowered by 0.09, resulting in a less good fit with a MAPE of again 10%. The final GWR manages to model the appraisal values with only 7.67% deviation on average. More important is the larger reduction of the R^2 and RMSE, indicating less severe outliers. The worst performing municipality is Groningen, likely due to it having the least samples. Rotterdam on the other hand performs especially well, perhaps due to the larger percentage of apartments in this dataset. On average, the apartments have a smaller prediction error (6.98%) than the family homes (7.41%). This in turn can be attributed to the lower average appraisal value of apartments and lower appraisals having more reference points. The results for 2020 are summarised in Table A.3.3, in the appendix. The results for 2020 show a slight decrease in predictive accuracy, but not significant.



(a) The influence of living area



(b) Variable importance

Figure 5.2: Plots describing the GWR model (Amersfoort, 2018).

GWR model results (2018)				
Metric	R^2	*RMSE*	*MAE*	*MAPE*
Amersfoort	0.822	€61,459	€48,393	7.42%
Amsterdam	0.831	€60,213	€53,671	7.31%
Eindhoven	0.812	€62,942	€54,103	8.01%
Groningen	0.789	€83,233	€55,213	8.61%
Rotterdam	0.861	€56,431	€47,312	6.99%

Table 5.2: Results for GWR models (2018).

The final model is the XGBoost model. As discussed in Step 4.3, the optimal parameter settings are eta = 0.15, tree depth = 6 and nr. of rounds = 2000, for each of the five municipalities. After 39 boosting rounds on average, no major improvements are made and after 159 rounds the performance starts to deteriorate slightly (Figure A.3.12). The fit of the XGBoost model has the best overall fit (R^2 = 0.848) with the lowest RMSE scores (€58,374). A summary of the performance metrics is given in Table 5.3. Figure 5.3 shows the predicted vs actual appraisal values for Amersfoort 2018. The other municipalities are shown in Figure A.3.13. The first tree of the final model for Amersfoort is shown in Figure A.3.15. The living area and WOZ-Waarde are again the most important variables, as seen in Figure A.3.14. Even with the appraisals above €750.000 excluded, there is slightly more unexplained variance in the high appraisal values. Some outliers still exist, but, overall, the XGBoost model provides accurate predictions with only 5% deviation on average.

Finally, a single XGBoost model is trained for all 5 municipalities using the same parameter settings (Figure 5.5). This model includes the municipality name as an additional variable. It is not surprising that the resulting model has less performance than the five individually trained models. The model's prediction error increases slightly to 6%. Furthermore, the RMSE increase substantially more than the MAE,

XGBoost model results - 2018				
Metric	R^2	*RMSE*	*MAE*	*MAPE*
Amersfoort	0.851	€57,391	€34,283	5.38%
Amsterdam	0.845	€57,964	€35,258	5.50%
Eindhoven	0.838	€57,385	€36,192	5.62%
Groningen	0.829	€59,832	€38,241	5.88%
Rotterdam	0.871	€56,144	€34,831	5.45%

Table 5.3: Results for GWR models (2018).

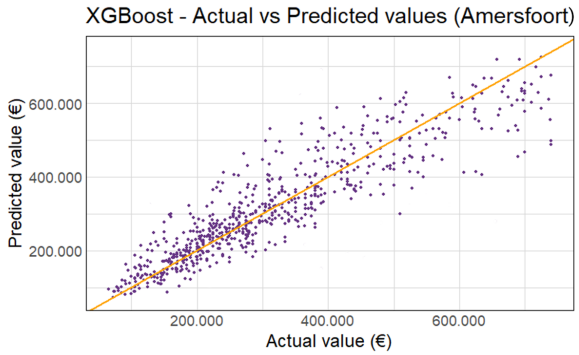


Figure 5.3: XGBoost Predicted vs Actual Values (Amersfoort, 2018).

suggesting that while the overall performance only decreased slightly, the model is worse at capturing outliers. The municipality name ends up becoming the third most important variable. While the model performance is slightly worse, it still outperforms the individually trained GWR models.

All in all, when looking at the quantitative performance metrics, the XGBoost models outperforms the linear regression and GWR models. The final qualitative metrics are the implementation time and model upkeep. In this research, the most effort went into gathering all the variables and preparing the dataset. As such, in practise, this is also expected to require the most maintenance. The BAG can be routinely updated using API request, however the RVO and CBS datasets both use an extract that does not have an API endpoint, which means it needs to be downloaded manually.

Additionally, there is the consideration of training time. LR is simple and fast, for many millions of records this is rarely an issue on a modern computer. On the other hand, the GWR computes regressions for a grid. In case of the municipality Amersfoort, a 100x100m tile grid for Amersfoort (roughly 10km x 10km) equals 100x100 tiles = 10k tiles = 10k unique regressions that are computed. On the hardware of Stater, this takes less than 5 minutes. For a national scale, the grid needs to be much larger in both dimensions, thus the required computing power increases exponentially. Fitting the

Averaged model performance for the 5 municipalities.								
Year	**2018**				**2020**			
Metric	R^2	*RMSE*	*MAE*	*MAPE*	R^2	*RMSE*	*MAE*	*MAPE*
LR	0.725	€97,232	€67,814	10.55%	0.734	€94,927	€62,871	10.23%
GWR	0.822	€64.856	€51,738	7.67%	0.809	€65,826	€52,237	7.92%
XGBoost	0.848	€58,374	€35,761	**5.89%**	0.852	€61,028	€35,451	**5.76%**

Table 5.4: Averaged model performance for the 5 municipalities, for each model type.

Single XGBoost model (2018)				
Metric	R^2	RMSE	MAE	MAPE
XGBoost	0.832	65,312	43,625	6.35%

Table 5.5: Single XGBoost model trained on all five municipalities (2018).

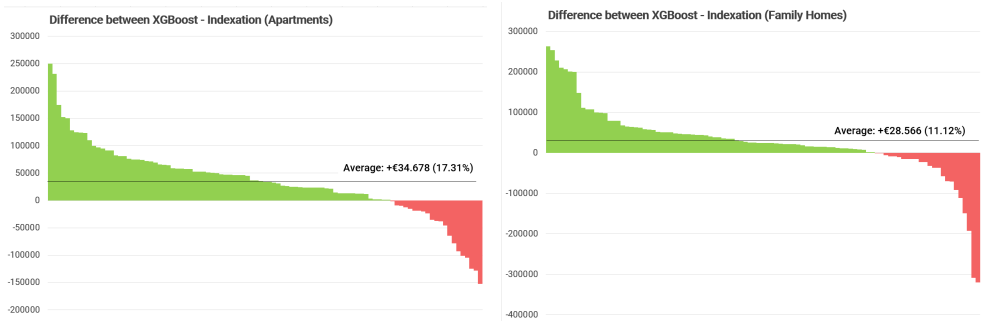
regression for the entire Netherlands likely takes a day, instead of a few minutes.

Unlike GWR, XGBoost also has a GPU implementation. In this thesis, the sample sizes for one year per municipality were relatively small, so even using only the CPU resulted in a good fit in less than 10 minutes using XGBoost. By using the GPU, XGboost is faster than the GWR model when training for the entire Netherlands. Model training time is something that does not costs much time of an employee. In the end, gathering the data and creating the dataset remains the most active time-consuming task, which takes an equal effort for all three models.

Finally, this section compares the current approach of indexation and the single XGBoost model. The current approach at Stater uses the Kadaster regional house price index (Figure A.1.2). The comparison with XGBoost is made for the appraisals from 2000. Figure 5.4 shows the difference between predictions by subtracting the indexed appraisal value from appraisal value predicted by XGBoost. The differences have been separated by the housing type as they appear in Stater's dataset: one lists the differences for all family homes and the other for all apartments. In both cases, XGBoost predicts higher appraisal values than the indexation method, on average €34.678 for the apartments (+17.31% higher than the index) and €28.566 (11.12%).

Two observations can be made from Figure 5.4. First, the XGBoost predictions for the apartments show less deviation from the index as compared to the predictions for the family homes. One explanation for this is the higher variance in the appraisal values of family homes as compared to apartments. The model is more likely to make a poor prediction for a family home than for an apartment as indicated by the larger outliers (rarely a large difference of €250k+).

Second, the difference between apartments and family homes corresponds to the other Kadaster index for housing types (Figure A.1.3). From this index, it can be



(a) Difference for Apartments, XGBoost predicts 17.31% higher

(b) Difference for Family homes, XGBoost predicts 11.12% higher

Figure 5.4: Differences XGBoost and indexation method using Kadaster regional price index (green = XGBoost predicts higher).

seen that apartments have increased almost an additional 20% over the family homes across the entire Netherlands (2000-2020). The XGBoost model is able to account for this, whereas the regional index is not. This supports the main conclusion that the XGBoost model can be a better alternative to price indexation. An ideal index for the Kadaster would discern both region and house type. This could be a relatively simple improvement over the current method of indexation. All in all, this provides additional support to the conclusion that the model approach can be an improvement over indexation, as it is able to account for housing type.

6 | Conclusion & Discussion

The introduction of this thesis outlines the need for more localised predictions of mortgage collaterals within the financial sector. Money lenders know the value of a house through an appraisal once the mortgage is approved. However, 20 years later it is unknown how much the house is actually still worth. Still, money lenders are mandated by the Authority for the Financial Markets (AFM) to make a proper risk analysis of their portfolios. Currently, at Stater N.V., the Kadaster index is used to index the appraisal value to give a value indication for a mortgage collateral. This generalises the price increase for all types of housing to the same regional price index. The goal of this thesis is to find out if external data sources allow for more localised predictions of appraisal values by answering the following research question:

"How can hedonic price models, based on location and intrinsic characteristics of real estate, serve as an alternative to price indexation, in order to more accurately valuate the collateral (house) of Stater's mortgages in Netherlands?"

In the literature review, four types of hedonic pricing models are identified to model houses prices. These models are: Linear Regression (LR), Geographically Weighted Regression (GWR), Multi-scale GWR (MGWR), and Extreme Gradient Boosting (XGBoost). Chapter 3 (Methodology) outlines the solution design approach of the thesis, which is based on an application of the Design Science Methodology. Using a 5-step approach, three models are realised (LR, GWR and XGBoost) to model the appraisal values for five unique municipalities: Amsterdam, Amersfoort, Eindhoven, Groningen, Rotterdam.

The second contribution lies in the collection of public datasets to describes all houses in the Netherlands and the neighbourhoods they are located in. In the end, 33 variables are used, as seen in the variable overview of A.3.7. This includes intrinsic characteristics about each house from the Kadaster, sociodemographic variables from CBS, and energy labels from 'Rijkdienst voor ondernemend Nederland' (RVO).

Finally, the methodology outlines four sub-questions to evaluate the three models and support the main research question. Each sub-question is answered in the next sections, followed by the conclusion to the main research question. The final section outlines some of the limitations of the research and provides further areas of research.

How do the three models (LR, GWR, XGBoost) compare in terms of bias and variance when estimating for the five selected municipalities?
The main quantitative results for each of the models are presented in Table 5.4. The models in this overview are tested on appraisal values below €750.000. The 5% of samples with higher appraisal values have very high variance due to the stronger influence of the buyers individual preferences. For 2020, XGBoost is able to best

explain the variance of the appraisal values with an average R^2 of 0.852. This is a statistically significant improvement over GWR (R^2 = 0.809) and LR (R^2 = 0.734). For XGBoost for, the mean RMSE of the five models is €61.028 and the MAE is €35.451. Paired with Figure 5.3, it can be seen that the larger appraisal values have a larger variance than the lower appraisal values. Thus, some outliers are present in the made predictions. On average, the mean absolute percentage error (MAPE) is 5.89%. For the average housing price of €450,000, this corresponds to an error of about €27.000,-. Overall, XGBoost is thus a good fit for modelling appraisal values.

To what extent do the influences of the housing and location characteristics differ between the five municipalities?
The two most important variables in all three model types are the total living area (vbo_oppervlakte, from Kadaster) and WOZ-Waarde (from CBS). Additionally, the other top-5 most important variables in the XGBoost model consisted of: the x-coordinate (horizontal position), percentage of incomes belonging to 20% highest incomes in the Netherlands, electricity usage and distance to nearest cafe. These variables make sense, as the western part of the Netherlands generally has higher appraisal values. Also, rich people tend to live in more expensive neighbourhoods. The variables that had the least influence were some of the other distance variables and variables with more missing values such as energy labels. The missing variables were imputed using k-nearest neighbours (k=7), which contributes to the fact these variables have less influence.

Finally, the living area and WOZ-Waarde account for at least 70% of the explained variance, while the other variables increase the explained variance by 7%. In this research, the WOZ-Waarde is the average of all WOZ-Waardes within a 100x100m tile. This can lead to inaccuracies when an expensive house is surrounded by cheaper apartments. For Stater, the individual WOZ-Waardes are available within the appraisal report. Unfortunately, these cannot be used directly, since this information is stored within a PDF. WOZ-Waardes normally cannot be collected in bulk. Stater could have legal grounds to request WOZ-Waardes for their portfolio from 'de Waarderingskamer' who determines the WOZ-Waarde. The underlying reason is that it allows Stater to better manage credit risk to protect their customers, leading to a more stable financial market. An alternative is the use of five WOZ-Waarde classes, for which a single experimental dataset is available [47]. This solves the previous issue of a possible inaccurate average value. In the end, the use of these classes improved the XGBoost model for Amsterdam significantly by lowering the MAPE from 6% to 4.5%.

How do the three models compare in terms of bias and variance when estimating on a national scale when also including municipalities with less data points?

The five municipalities were specifically chosen as they represent unique provinces in the Netherlands. Together with the fact that these municipalities contain over 40% of all appraisal values in Stater's dataset, it is fair to assume this provides a good indication if a model can be trained for the entire Netherlands. The LR and GWR were both disregarded, as XGBoost provided a large improvement over both models. In the end, a single XGBoost model is trained for all five municipalities using the same parameter settings, since the results for the individual models were similar (as seen in Table 5.3). All in all, this XGBoost model performs only marginally worse, with only a 0.02 reduction for the R^2 and a 0.48% increase for the MAPE, when compared to the individual models (see Figure 5.5). It can thus be concluded, that it is highly likely that XGBoost is also able to model the appraisal values for all municipalities.

Based on the solution metrics specified in 3.2, what are the disadvantages / advantages of the chosen model as compared to the current approach of indexing house prices?
Finally, looking at the quantitative metrics R^2 (model fit), RMSE and MAPE, the XGBoost model is the clear winner out of the three trained models. However, it also performs well in terms of training time performance compared to GWR. XGBoost comes with the advantage that it can run on the GPU, whereas GWR runs into problems when computing regressions for large grids for entire countries. The training time of XGBoost is thus not an issue when training models for all appraisal values. The largest time consumption, compared to indexation, lies in keeping the model data up-to-date, which is equally time consuming for all three models. Only the Kadaster data is easily accessible through various APIs. The CBS and RVO dataset have to be downloaded manually.

A quantitative comparison between XGBoost and indexation is made by comparing the predictions for appraisal values from 2000. As a reminder, the regional price index of Kadaster can been seen in A.1.2. The predictions are discerned in two categories: apartments and family homes, as these categories appear in Stater's dataset. In both cases, the XGBoost model makes higher predictions than the index: +17.14% for apartments and +11.12% for family homes (Figure 5.4). This is not surprising as the index is a more conservative estimate of the price increase by taking the average of many real estate prices. The predictions of the XGBoost model are in line with a different Kadaster index, namely the one for housing types (Figure A.1.3). This index shows a larger increase of 70% in apartment prices, as compared to only 50% for family homes (2000-2020). This supports that the XGBoost model is able to account for differences in price development for apartments and family homes. Finally, it should be noted that for the family homes, the XGBoost model also has a few outliers in its predictions. However, based on the training results for 2018, it can be concluded that the XGBoost model can be more reliable than indexation.

6.1 Answering the main research question

In the end, the XGBoost model is able to model a large subset of the houses with a better accuracy than indexation. For the five municipalities, a single XGBoost can explain 83% the variance with a RMSE of €65,312, a MAE of €43,625 and MAPE of 6.35% (Table 5.5). The two most important variables in the model are the total living area (vbo_oppervlakte, from Kadaster) and WOZ-Waarde (from CBS) (Table 5.5). As shown in the comparison between indexation and XGBoost, the XGBoost is superior to indexation, as the model takes into account different housing types (Figure 5.4). The remaining unexplained variance of 17% is likely due to a missing variable that explains the quality of the house itself. Information specific to the house from the official appraisal reports can help alleviate this variance, as they contain more information about the house itself.

The downsides of the XGBoost model are the larger outliers compared to the conservative indexation method, as well as the fact that the model currently predicts for an entire year and does not account for monthly changes. This can partially be mitigated by ensuring the model gets retrained every month, replacing the appraisals of the oldest month with the new month. Finally, it takes extra effort to keep the data of the models up-to-date. However, in return for this extra effort, XGBoost can make more localised predictions for the entire Netherlands to valuate Stater's mortgage collaterals.

6.2 Recommendations for Stater & Future work

There are numerous improvements that could potentially lead to a better performing model. The accuracy of 6.35% (XGBoost) is not accurate enough to hold the model on equal standard to the actual appraiser itself. However, as identified in the literature review, the appraisal models are certainly a booming business. There are numerous parties, such as Calcasa, offering their own appraisal services. Money lenders are required by the AFM to provide insights into credit risk. As such, money lenders often buy these appraisals to valuate their own portfolios. These models are most certainly better than the XGBoost model presented in this thesis. However, it highlights a business opportunity for Stater for extra services to their own money lenders. The data from appraisal reports is likely to solve the current gap of the remaining 17% of unexplained variance. With the inclusion of this data, Stater could create their own valuation model which will end up being a competitive alternative to other valuation models.

Finally, the author has the following recommendations for future research areas centred around modelling real estate using open data and XGBoost:

- **The application of XGBoost or GWR to other housing related problems.** For example, the ground sinkage map from TU Delft provides an interesting use case for looking at real estate portfolio risk factors. Ground sinkage is a real problem in the Netherlands, especially in Groningen. As a result of the gas exploitation, the property values are reduced drastically in the region. This poses a clear risk to the mortgage owner and the money lender. Another problem for many houses is foundation rot, perhaps risk areas can be identified by combining sinkage data with ground compositions. Additionally, based on a use case about identifying solar panels using image recognition at Stater's datalab, one can look into the effect of solar panels or other energy savings methods on the house prices, to ultimately determine which saving method is the most cost-effective.

- **The exploration of time-based differences in appraisal values.** This thesis focuses on making accurate predictions for a given year, usually the current year. Yet, for most variables used in this historic data is available. The GWR and LR model are less suited for this as they cannot properly model temporal changes, since they rely on a variable either having a positive or negative influence. However, XGBoost or a temporal variant of GWR called MGTWR [25], can potentially be used to create price indices for other features besides region and housing type.

- **The lack of a feature to model house quality.** The remaining unexplained variance of 17% is likely due to a missing variable that explains the quality of the house itself or other location characteristics. The average WOZ-Waarde has helped slightly in the XGBoost model, but the only other intrinsic variables used in this research are the living area and land lot space. In the literature review, most identified papers [8][17][27] mainly focus on modelling only intrinsic characteristics, such as number and types of rooms. As outlined in the previous section, this information can be found in official appraisal reports. This will help paint a better picture of the house itself.

- **Computing other location variables.** The datasets from CBS and RVO are used in this thesis since they are the most complete, ready for instant use and can be directly related to a house. There are many other potential datasets available that can be used to derive new features, two examples being PDOK.nl and data.overheid.nl. Here one can find information about locations of all high voltage cables, wind mill parks, highways and more. The challenge lies within transforming this geo-data to information related to the house, for example proximity to a highway influencing the house price due to noise disturbance. Ultimately, more potential variables for the model will help Stater build a more accurate valuation model, that can be competitive with other models on the market.

Bibliography

[1] AFM. (2018). "Hypotheek in relatie tot waarde huis (ltv)," [Online]. Available: https://afm.nl/nl-nl/consumenten/themas/producten/hypotheek/hoeveel-lenen/maximale-hypotheek (visited on 02/02/2021).

[2] M. Lozej and A. Rannenberg, "The macroeconomic effects of the ltv and lti ratios in ireland," *Applied economics letters*, vol. 25, no. 21, pp. 1507–1511, 2018.

[3] D. N. Bank, *De kwaliteit en onafhankelijkheid van woningtaxaties, occasional studies*, 2019.

[4] B. for International Settlements. (2004). "Basel ii: International convergence of capital measurement and capital standards: A revised framework," [Online]. Available: https://bis.org/publ/bcbs107.htm (visited on 01/17/2021).

[5] L. M. Brander and M. J. Koetse, "The value of urban open space: Meta-analyses of contingent valuation and hedonic pricing results," *Journal of environmental management*, vol. 92, no. 10, pp. 2763–2773, 2011.

[6] T. Potrawa, "Hedonic pricing model for rotterdam housing market," 2020.

[7] V. Liebelt, S. Bartke, and N. Schwarz, "Hedonic pricing analysis of the influence of urban green spaces onto residential prices: The case of leipzig, germany," *European Planning Studies*, vol. 26, no. 1, pp. 133–157, 2018.

[8] K. Cao, M. Diao, and B. Wu, "A big data–based geographically weighted regression model for public housing prices: A case study in singapore," *Annals of the American Association of Geographers*, vol. 109, no. 1, pp. 173–186, 2019.

[9] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[10] N. V. van Makelaars en Taxateurs. (2021). "Market information real estate," [Online]. Available: https://nvm.nl/wonen/marktinformatie/ (visited on 01/21/2021).

[11] Funda. (2021). "Funda index, de stand van wonend nederland," [Online]. Available: https://funda.nl/funda-index/ (visited on 01/16/2021).

[12] CBS. (2021). "Prijzen koopwoningen," [Online]. Available: https://cbs.nl/nl-nl/reeksen/prijzen-koopwoningen (visited on 01/16/2021).

[13] S. Jansen, P. de Vries, H. Coolen, C. Lamain, and P. Boelhouwer, "Developing a house price index for the netherlands: A practical application of weighted repeat sales," *The Journal of Real Estate Finance and Economics*, vol. 37, no. 2, pp. 163–186, 2008.

[14] J. P. Harding, S. S. Rosenthal, and C. Sirmans, "Depreciation of housing capital, maintenance, and house price inflation: Estimates from a repeat sales model," *Journal of urban Economics*, vol. 61, no. 2, pp. 193–217, 2007.

[15] A. C. Goodman, "Hedonic prices, price indices and housing markets," *Journal of urban economics*, vol. 5, no. 4, pp. 471–484, 1978.

[16] J. Luttik, "The value of trees, water and open space as reflected by house prices in the netherlands," *Landscape and urban planning*, vol. 48, no. 3-4, pp. 161–167, 2000.

[17] S. Farber and M. Yeates, "A comparison of localized regression models in a hedonic house price context," *Canadian Journal of Regional Science*, vol. 29, no. 3, pp. 405–420, 2006.

[18] J. Zhou, H. Zhang, Y. Gu, and A. A. Pantelous, "Affordable levels of house prices using fuzzy linear regression analysis: The case of shanghai," *Soft Computing*, vol. 22, no. 16, pp. 5407–5418, 2018.

[19] A. Fotheringham, C. Brunsdon, and M. Charlton, "Geographically weighted regression: The analysis of spatially varying relationships," *John Wiley & Sons*, vol. 13, Jan. 2002.

[20] Y. Gong, "The spatial dimension of house prices," *A+ BE| Architecture and the Built Environment*, no. 4, pp. 1–186, 2017.

[21] L. Potuijt, "Een voorspelling van onbekende waarden uit de basisregistratie adressen en gebouwen," 2019.

[22] M. Tomal, "Modelling housing rents using spatial autoregressive geographically weighted regression: A case study in cracow, poland," *ISPRS International Journal of Geo-Information*, vol. 9, no. 6, p. 346, 2020.

[23] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic geography*, vol. 46, no. sup1, pp. 234–240, 1970.

[24] S. Sims, P. Dent, and G. R. Oskrochi, "Modelling the impact of wind farms on house prices in the uk," *International Journal of Strategic Property Management*, vol. 12, no. 4, pp. 251–269, 2008.

[25] C. Wu, F. Ren, W. Hu, and Q. Du, "Multiscale geographically and temporally weighted regression: Exploring the spatiotemporal determinants of housing prices," *International Journal of Geographical Information Science*, vol. 33, pp. 1–23, Nov. 2018. DOI: 10.1080/13658816.2018.1545158.

[26] S. Zhang, L. Wang, and F. Lu, "Exploring housing rent by mixed geographically weighted regression: A case study in nanjing," *ISPRS International Journal of Geo-Information*, vol. 8, no. 10, p. 431, 2019.

[27] Z. Shabrina, B. Buyuklieva, and M. K. M. Ng, "Short-term rental platform in the urban tourism context: A geographically weighted regression (gwr) and a multiscale gwr (mgwr) approaches," *Geographical Analysis*, vol. n/a, no. n/a, DOI: https://doi.org/10.1111/gean.12259.

[28] M. Shahhosseini, G. Hu, and H. Pham, "Optimizing ensemble weights for machine learning models: A case study for housing price prediction," in *Smart Service Systems, Operations Management, and Analytics*, H. Yang, R. Qiu, and W. Chen, Eds., Cham: Springer International Publishing, 2020, pp. 87–97, ISBN: 978-3-030-30967-1.

[29] J. Avanijaa *et al.*, "Prediction of house price using xgboost regression algorithm," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 2, pp. 2151–2155, 2021.

[30] Waarderingskamer. (2020). "Hoe de woz-waarde tot stand komt.," [Online]. Available: https://waarderingskamer.nl/klopt-mijn-woz-waarde/totstandkoming-woz-waarde/ (visited on 03/01/2021).

[31] M. Joeman, "De woz-waarde en de realiteit," Aug. 2019. [Online]. Available: http://hdl.handle.net/2105/48004.

[32] Calcasa. (2020). "Wox-waarde," [Online]. Available: https://calcasa.nl/wox-online (visited on 03/02/2021).

[33] KadasterData. (2020). "Woningwaarde," [Online]. Available: https://kadasterdata.nl/woningwaarde (visited on 03/02/2021).

[34] D. Hypotheker. (2020). "Wat is mijn huis waard test," [Online]. Available: https://hypotheker.nl/zelf-berekenen/wat-is-deze-woning-waard/ (visited on 03/02/2021).

[35] C. Brunsdon, J. Corcoran, and G. Higgs, "Visualising space and time in crime patterns: A comparison of methods," *Computers, Environment and Urban Systems*, vol. 31, no. 1, pp. 52–75, 2007, Extracting Information from Spatial Datasets, ISSN: 0198-9715. DOI: https://doi.org/10.1016/j.compenvurbsys.2005.07.009.

[36] E. R. de Wit, P. Englund, and M. K. Francke, "Price and transaction volume in the dutch housing market," *Regional Science and Urban Economics*, vol. 43, no. 2, pp. 220–241, 2013, ISSN: 0166-0462. DOI: https://doi.org/10.1016/j.regsciurbeco.2012.07.002.

[37] D. C. Wheeler and A. Páez, "Geographically weighted regression," in *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, M. M. Fischer and A. Getis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 461–486, ISBN: 978-3-642-03647-7. DOI: https://doi.org/10.1007/978-3-642-03647-7_22.

[38] J. J. McCluskey and G. C. Rausser, "Hazardous waste sites and housing appreciation rates," *Journal of Environmental Economics and Management*, vol. 45, no. 2, pp. 166–176, 2003, ISSN: 0095-0696. DOI: https://doi.org/10.1016/S0095-0696(02)00048-7.

[39] B. W. AMBROSE, P. EICHHOLTZ, and T. LINDENTHAL, "House prices and fundamentals: 355 years of evidence," *Journal of Money, Credit and Banking*, vol. 45, no. 2-3, pp. 477–491, 2013. DOI: https://doi.org/10.1111/jmcb.12011.

[40] F. Fuerst, P. McAllister, A. Nanda, and P. Wyatt, "Does energy efficiency matter to home-buyers? an investigation of epc ratings and transaction prices in england," *Energy Economics*, vol. 48, pp. 145–156, 2015.

[41] F. De Vor and H. L. De Groot, "The impact of industrial sites on residential property values: A hedonic pricing analysis from the netherlands," *Regional Studies*, vol. 45, no. 5, pp. 609–623, 2011.

[42] Kadaster. (2021). "Bag, addresses and buildings key register," [Online]. Available: https://kadaster.nl/zakelijk/registraties/basisregistraties/bag (visited on 01/15/2021).

[43] K. /. E. Nederland. (2021). "Ddk: Land lot dataset," [Online]. Available: https://arcgis.com/home/group.html?id=eb452ccc59e0431c8b42b06c7e7a6fee#overview (visited on 03/05/2021).

[44] CBS. (2021). "100x100 vierkantsatistieken," [Online]. Available: https://cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken (visited on 03/01/2021).

[45] RVO. (2021). "Energielabels database," [Online]. Available: https://www.ep-online.nl/ (visited on 03/01/2021).

[46] A. Fotheringham, W. Yang, and W. Kang, "Multiscale geographically weighted regression (mgwr)," *Annals of the American Association of Geographers*, vol. 107, pp. 1247–1265, Aug. 2017. DOI: 10.1080/24694452.2017.1352480.

[47] K. Labs. (). "Woz-waarde," [Online]. Available: https://data.labs.kadaster.nl/experiment/woz/ (visited on 03/01/2021).

[48] Kadaster. (2021). "Vastgoed dashboard, prijsindex," [Online]. Available: https://kadaster.nl/zakelijk/vastgoedinformatie/vastgoedcijfers/vastgoeddashboard/prijsindex (visited on 02/15/2021).

A | Appendix

A.1 C2: Housing market prices Figures

A.1.1 CBS: Spatially varying house prices across the Netherlands.

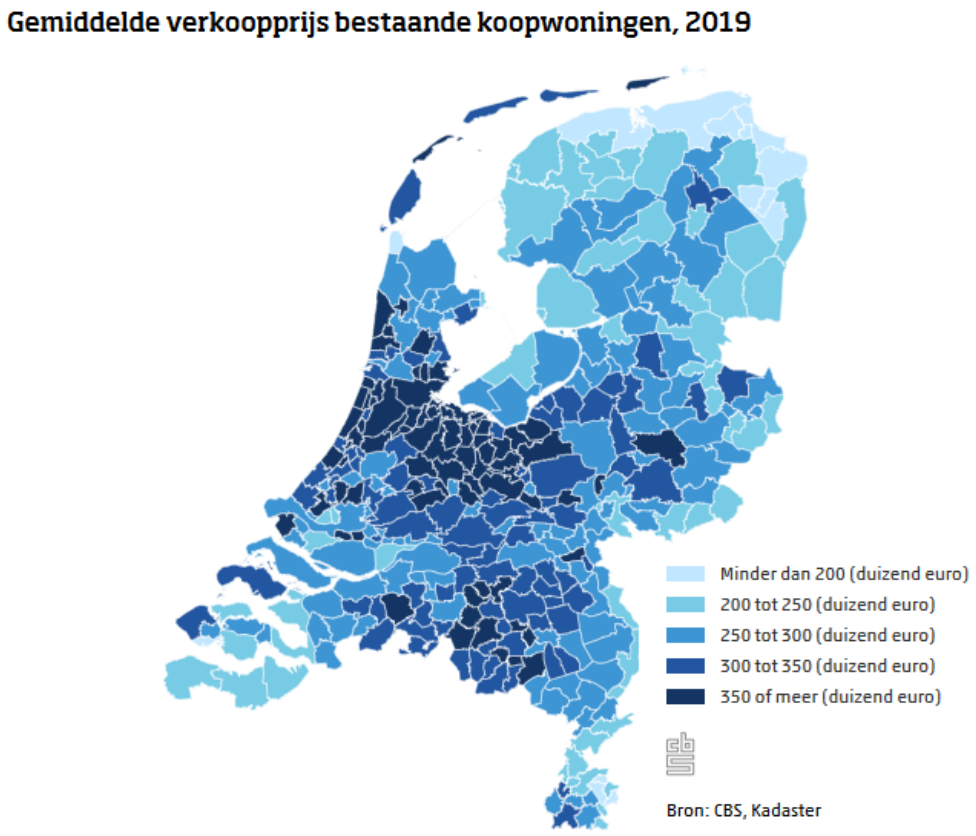


Figure A.1.1: Average house sale price per municipality in 2019 (thousand euros).

A.1.2 Kadaster house price index per region (% change per year)

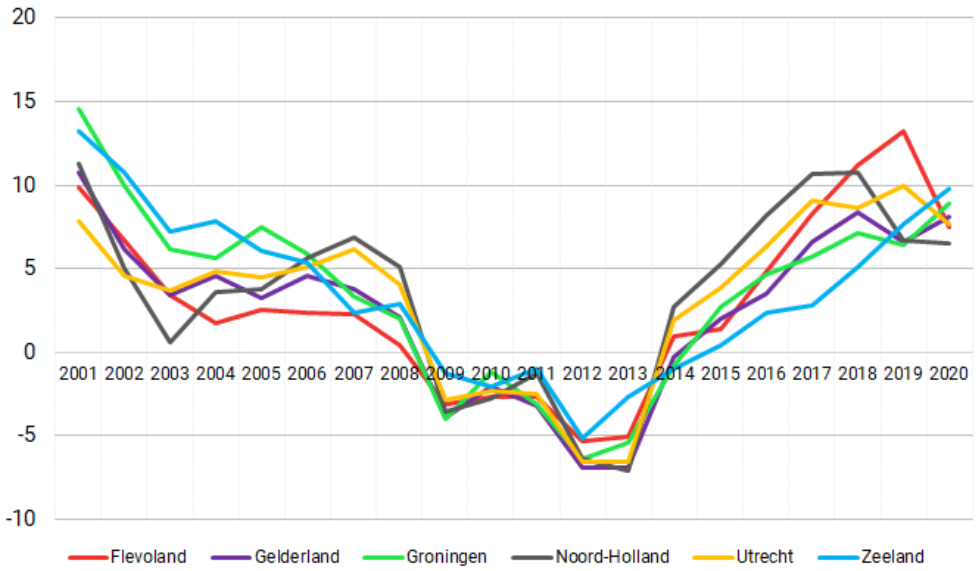


Figure A.1.2: Percent change in house prices compared to last year, for six provinces in the Netherlands, source: Kadaster [48].

Province	**% increase over 2000-2020**	**Province**	**% increase over 2001-2020**
Drenthe	56.35%	Noord-Brabant	42.90%
Flevoland	44.02%	Noord-Holland	**76.70%**
Friesland	55.34%	Overijssel	49.73%
Gelderland	45.05%	Utrecht	70.87%
Groningen	67.48%	Zeeland	74.83%
Limburg	38.16%	Zuid-Holland	52.36%

Table A.1.1: Cumulative % change in house prices between Jan. 2000 and Jan. 2020, for all twelve provinces of the Netherlands. (Based on Fig A.1.2)

A.1.3 Kadaster house price index per housing type (% change per year)

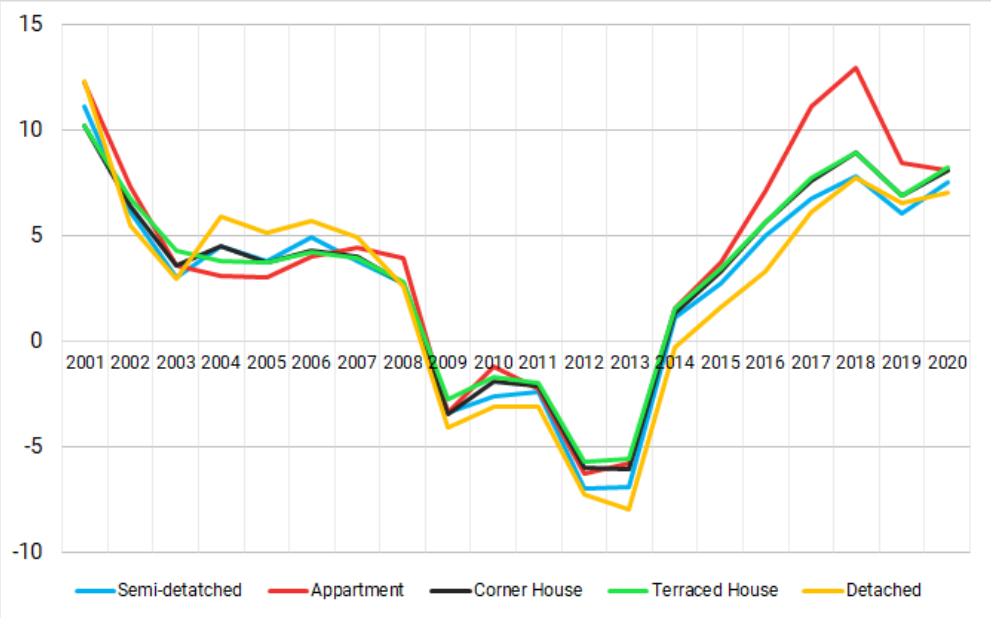


Figure A.1.3: Percent change in house prices compared to last year, for six housing types in the Netherlands, source: Kadaster [48].

Housing type	**% increase over 2001-2020**
Detached	54.4%
Semi-detached	51.2%
Terraced House	64.0%
Corner House	61.5%
Apartment	**75.3%**

Table A.1.2: Cumulative % change in house prices between Jan. 2000 and Jan. 2020, for six housing types in the Netherlands. (Based on Fig A.1.3)

A.2 C3: Methodology Figures

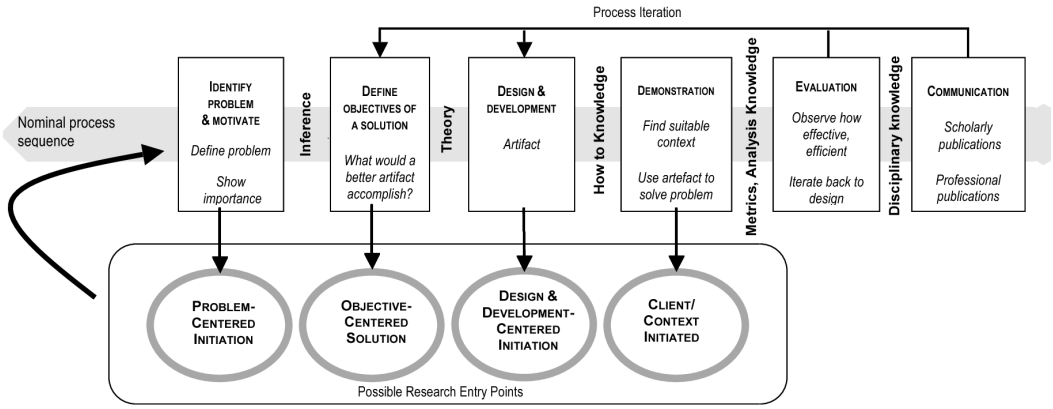


Figure A.2.1: Original Design Science Methodology diagram by Hevner et al. [9].

A.3 C4: Data & Model Figures

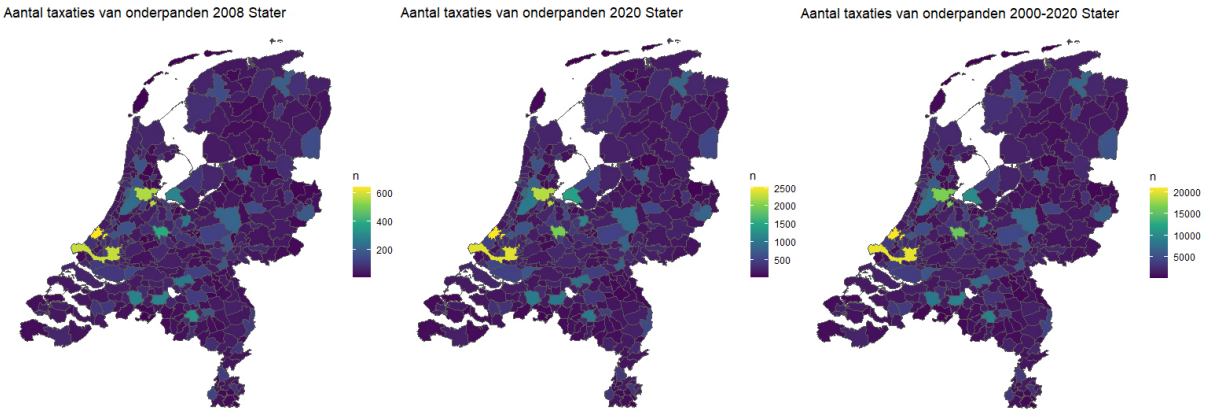


Figure A.3.1: Number of real estate appraisal values of State, left: 2008, middle: 2020, right: all appraisals from Jan. 2000 up until Jan. 2021. Showing a similar spatial distribution for each of the years.

Municipalities	# of appraisals
Rotterdam	17.757
Amsterdam	16.343
'S-Gravenhage	14.956
Almere	10.161
Eindhoven	8.936
Tilburg	8.558
Amersfoort	7.389
Groningen	7.215
Enschede	7.100
Breda	6.726
Haarlem	6.552
Arnhem	6.342
Apeldoorn	6.248
Dordrecht	5.917
Maastricht	5.899

Table A.3.1: Top 15 Largest number of appraisals per municipality (2000-2020).

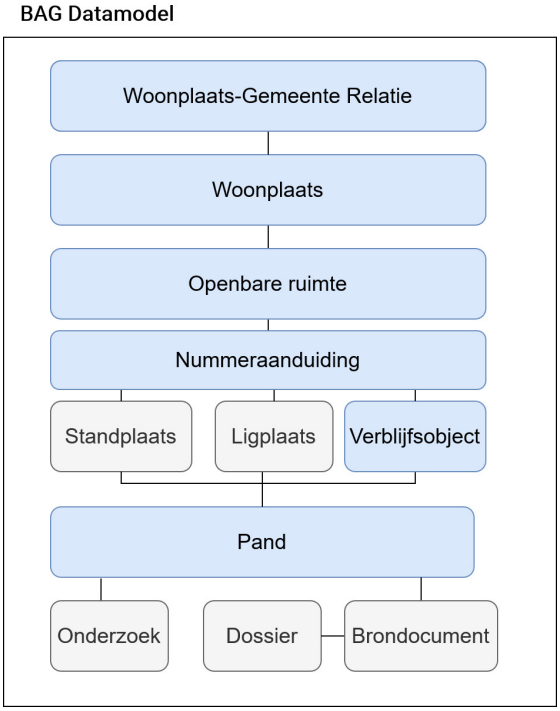


Figure A.3.2: BAG Data model, elements highlighted in blue are used in this research.

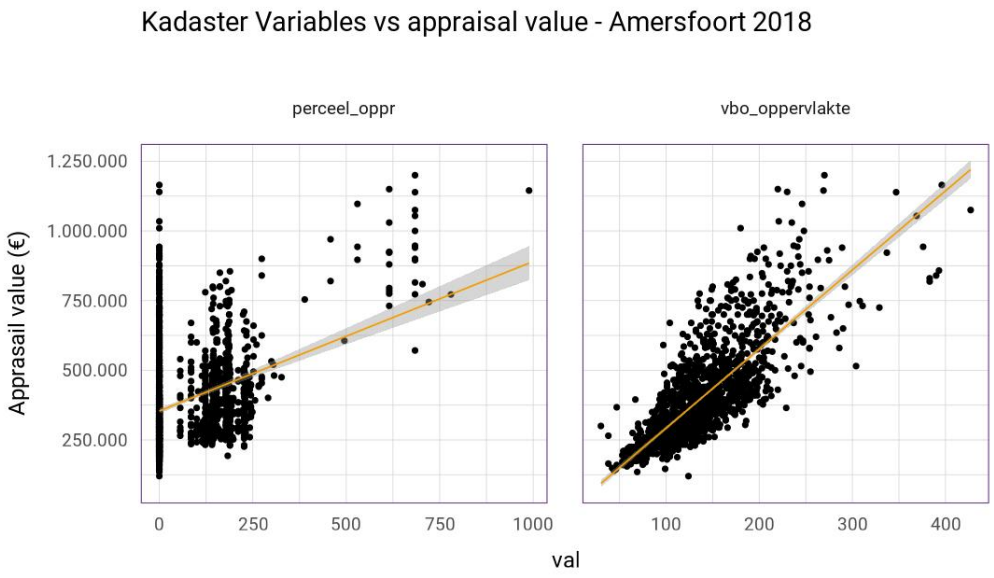


Figure A.3.3: Kadaster Variables vs. Appraisal Values - Amersfoort (2018)

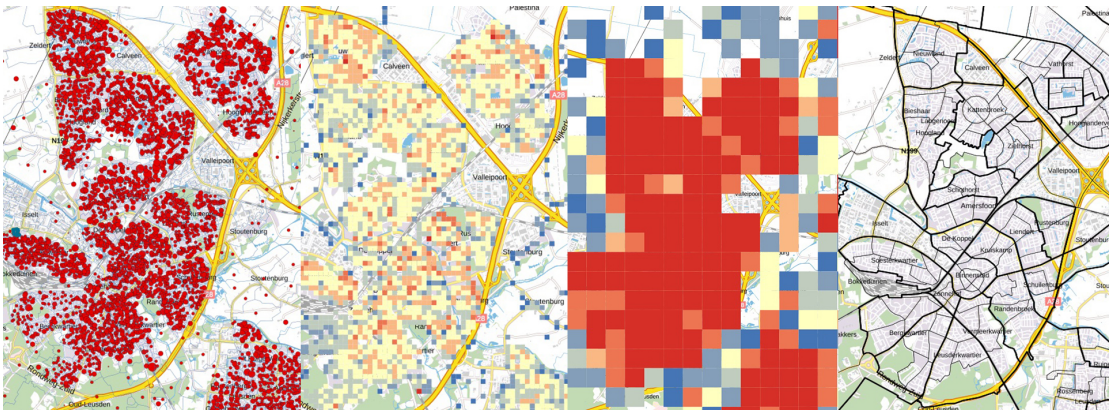


Figure A.3.4: Different resolutions of demographic variables from CBS. From left to right, 6 character postal code (1111AA), 100x100, tiles, 500x500m tiles, 4 character postal code.

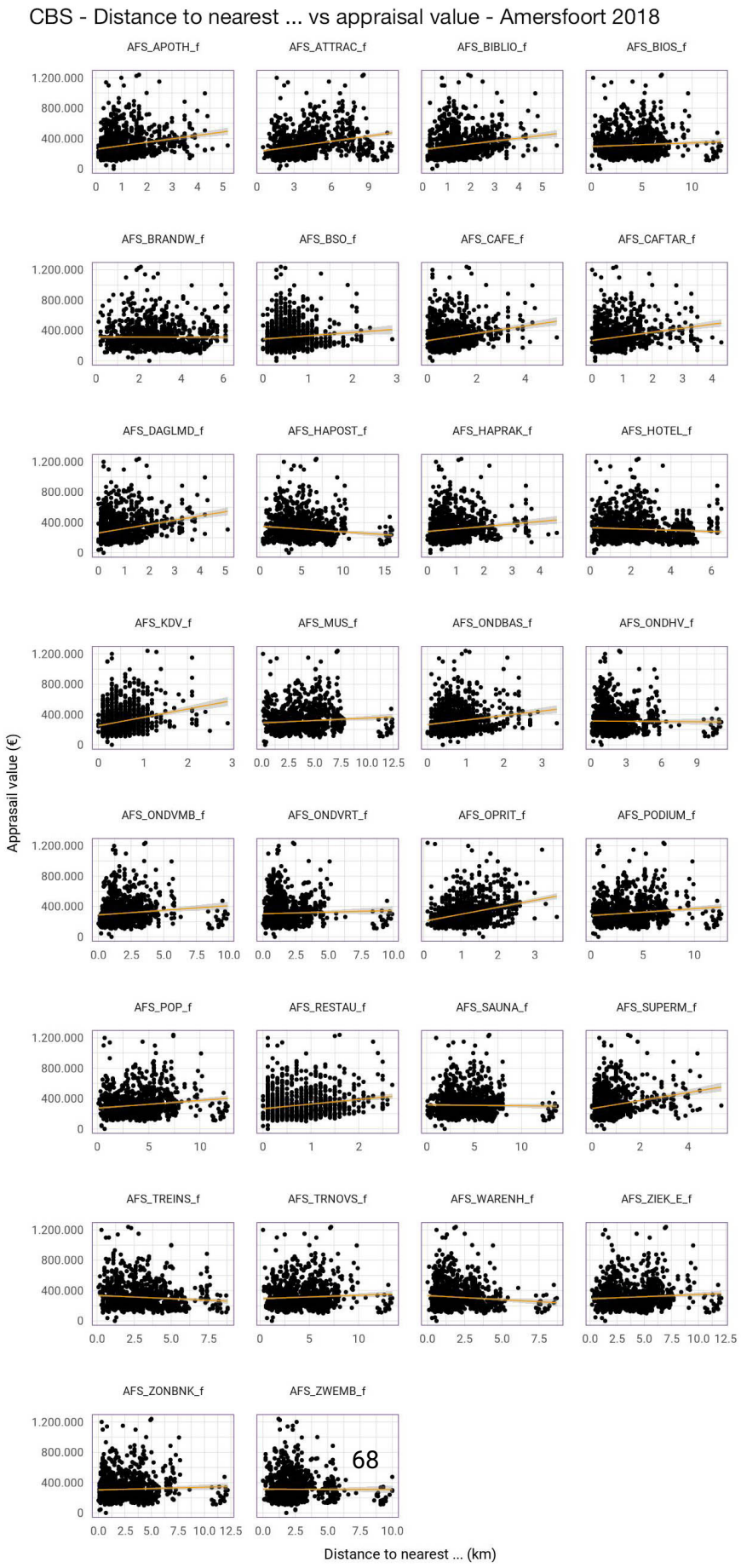


Figure A.3.5: CBS Distance to ... vs. Appraisal Values (Amersfoort, 2018)

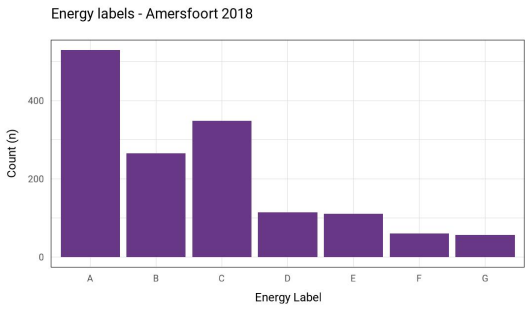


Figure A.3.6: RVO Energy Labels - (Amersfoort, 2018)

Abbreviation	Description	Source
Bedr_vov	The appraisal value of a house	Stater
Year	The appraisal year	Stater
Gemeentenaam	Municipality name	Kadaster
Geometry	Geocoordinates (point) of house. (verblijfsobject)	Kadaster
Is_gezinwng	Apartment (0) or Family home (1)	Stater
Garage	Presence of garage (yes/no = 1/0)	Stater
Vbo_oppervlakte	The total floor area of a house (m^2)	Kadaster
Energy label	Energy label / class (factor)	RVO
INW_014_500	% of people aged 0-14 (500m tiles)	CBS
INW_2544_500	% of people aged 25-44 (500m tiles)	CBS
INW_4564_500	% of people aged 45-64 (500m tiles)	CBS
INW_65PL_500	% of people aged 65+ (500m tiles)	CBS
TOTHH_EENP_500	% of single person house holds	CBS
TOTHH_MPZK_500	% of house holds > 1 and no children.	CBS
HH_EENOUD	% of one parent households with children	CBS
GEM_HH_GR_500	Average house hold size	CBS
WOZ_WONING	Average WOZ-Waarde (x1000 €)	CBS
P_KOOPWON	Percentage Owner-occupied home	CBS
G_ELEK_WON	Average Electricity Usage (kwH)	CBS
M_INKHH	Median income group (factor)	CBS
OAD_500	Address density (address/km^2)	CBS
STED_500	Urbanisation (factor)	CBS
WON_NBEW	% non-inhabited homes	CBS
P_LINK_HH	Percentage of households belonging to bottom 40% of national income	CBS
P_HINK_HH	Percentage of households belonging to top 20% of national income	CBS
AFS_SUPERM	Distance to nearest supermarket (km)	CBS
AFS_OPRIT	Distance to nearest provincial road or highway (km)	CBS
AFS_CAFE	Distance to nearest cafe (km)	CBS
AFS_BIBLIO	Distance to nearest library (km)	CBS
AFS_ONDVRT	Distance to nearest secondary education (km)	CBS
AFS_APOTH	Distance to nearest pharmacy (km)	CBS

Figure A.3.7: Overview of variables used in the final models.

Abbreviation	Description	Source
Dist_centre	Distance to city center	Self-computed
UITKMINAOW	Income from AOW	CBS
INWONER	Inhabitants at start of year	CBS
AANTAL_HH	Number of households. (used to convert other variables to percentage)	CBS
HH_TWEEOUD	% of two parent households with children	CBS
WONING_500	Number of houses.	CBS
P_NW_MIG_A	Percentage of inhabitants (non-western)	CBS
P_HUURWON	Percentage of rented homes	CBS
G_GAS_WON	Average Gas Usage (m^3)	CBS
Pand_gebouwtype	Home type (more specific than Stater)	RVO
Pand_subtype	Home subtype	RVO
AV1/5/10/20 vars	Collection of 'Amount of X within radius 1/5/10/20 km' (hospitals, stores, schools etc.)	CBS
Other AFS vars	attraction parks, restaurants, hotels, hospital and others.)	CBS

Figure A.3.8: Variables excluded due to high correlation with other variables.

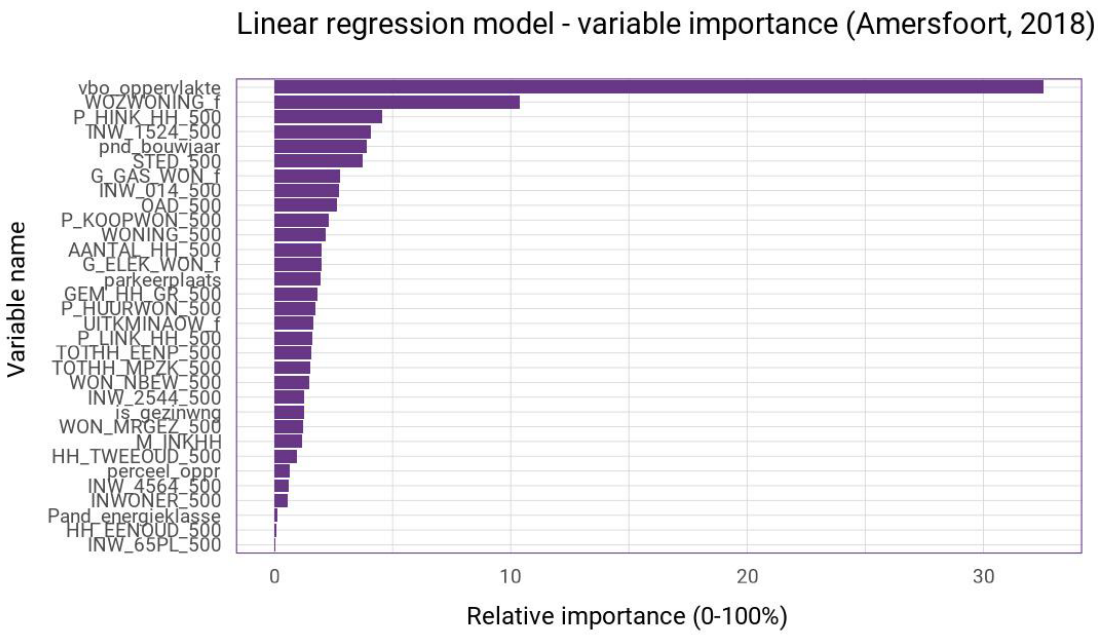


Figure A.3.9: Variable importance for the LR model of Amersfoort (2018). All 5 municipalities have similar results.

Variable Inflation Factors (Amersfoort, 2018)	
Abbreviation	*VIF*
vbo_oppervlakte	1.82
is_gezinwng	1.91
parkeerplaats	1.43
pnd_bouwjaar	2.31
perceel_oppr	1.8
Pand_energieklasse	2.69
INW_014	**15.05**
INW_1524	1.53
INW_2544	**11.47**
INW_4464	**7.67**
INW_65PL	**10.91**
TOTHH_EENP	4.12
TOTHH_MPZK	4.78
HH_EENOUD	4.68
WON_MRGEZ	4.4
WON_NBEW	1.9
OAD	3.87
STED_500	**6.12**
P_KOOPWON	3.43
WOZWONING	3.15
G_ELEK_WON	2.1
P_LINK_HH	**13.12**
P_HINK_HH	**14.45**
AFS_SUPERM	3.22
AFS_OPRIT	2.48
AFS_BIBLIO	2.27
AFS_ONDVRT	1.77
AFS_APOTH	2.08

Table A.3.2: Variable inflation factors, highlighted are the high collinear variables, VIF>5 (Amersfoort, 2018).

```
   ***********************************************************************
   *                       Package   GWmodel                             *
   ***********************************************************************
   Program starts at: 2021-06-04 00:39:45
   Call:
   gwr.basic(formula = bedr_vov ~ ., data = as(train, "Spatial"),
    regression.points = grid_sp, bw = 1354, kernel = "gaussian",
    adaptive = F, dMat = DM)

   Dependent (y) variable:  bedr_vov
   Independent variables:  .
   Number of data points: 1053
   ***********************************************************************
   *                    Results of Global Regression                     *
   ***********************************************************************

   Call:
    lm(formula = formula, data = data)

   Residuals:
    Min      1Q  Median      3Q     Max
-222507  -38385   -4559   32066  296026

   Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
   (Intercept)         2.677e+06  8.852e+05   3.024 0.002554 **
   vbo_oppervlakte     1.964e+03  6.371e+01  30.821  < 2e-16 ***
   is_gezinwng         1.192e+04  7.376e+03   1.616 0.106416
   parkeerplaats       5.553e+04  1.230e+04   4.513 7.13e-06 ***
   pnd_bouwjaar       -4.741e+02  1.020e+02  -4.649 3.77e-06 ***
   INW_014_500        -2.046e+05  1.333e+05  -1.535 0.124978
   INW_1524_500        2.885e+07  1.341e+07   2.152 0.031637 *
   INW_2544_500       -4.649e+04  8.361e+04  -0.556 0.578266
   INW_4564_500       -1.758e+05  8.079e+04  -2.176 0.029780 *
   INW_65PL_500       -2.157e+05  7.740e+04  -2.787 0.005413 **
   TOTHH_EENP_500      1.069e+05  4.885e+04   2.187 0.028933 *
   TOTHH_MPZK_500      1.265e+05  7.952e+04   1.590 0.112103
   HH_EENOUD_500      -4.814e+04  1.121e+05  -0.430 0.667605
   WON_MRGEZ_500      -1.398e+04  1.831e+04  -0.764 0.445094
   WON_NBEW_500        2.734e+05  1.079e+05   2.534 0.011426 *
   OAD_500             2.105e+01  6.298e+00   3.342 0.000861 ***
   STED_500            1.349e+04  5.544e+03   2.434 0.015105 *
   GEM_HH_GR_500       3.125e+04  2.252e+04   1.387 0.165640
   WONING_500         -7.924e+00  2.082e+01  -0.381 0.703637
   P_KOOPWON_500      -9.629e+02  8.142e+02  -1.183 0.237201
   P_HUURWON_500      -8.826e+02  7.849e+02  -1.124 0.261095
   M_INKHH             6.991e+03  2.689e+03   2.600 0.009453 **
   WOZWONING_f         4.736e+02  3.728e+01  12.703  < 2e-16 ***
   G_GAS_WON_f         1.391e+01  7.466e+00   1.863 0.062819 .
   G_ELEK_WON_f       -1.071e+01  5.011e+00  -2.137 0.032850 *
   HH_TWEEOUD_500      3.213e+01  7.341e+01   0.438 0.661705
   UITKMINAOW_f        2.814e+01  3.157e+01   0.891 0.372915
   P_LINK_HH_500       1.810e+03  8.355e+02   2.166 0.030537 *
   P_HINK_HH_500       1.498e+03  5.620e+02   2.666 0.007806 **
   Pand_energieklasse  3.892e+03  2.042e+03   1.906 0.056957 .
   perceel_oppr       -3.931e+00  2.658e+01  -0.148 0.882458
   coords.x1           9.678e-01  2.123e+00   0.456 0.648613
   coords.x2          -4.448e+00  1.839e+00  -2.419 0.015730 *

   ---Significance stars
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   Residual standard error: 62690 on 1020 degrees of freedom
   Multiple R-squared: 0.8412
   Adjusted R-squared: 0.8143
```

Figure A.3.10: Variable weights and significance tests for GWR (Amersfoort, 2018)

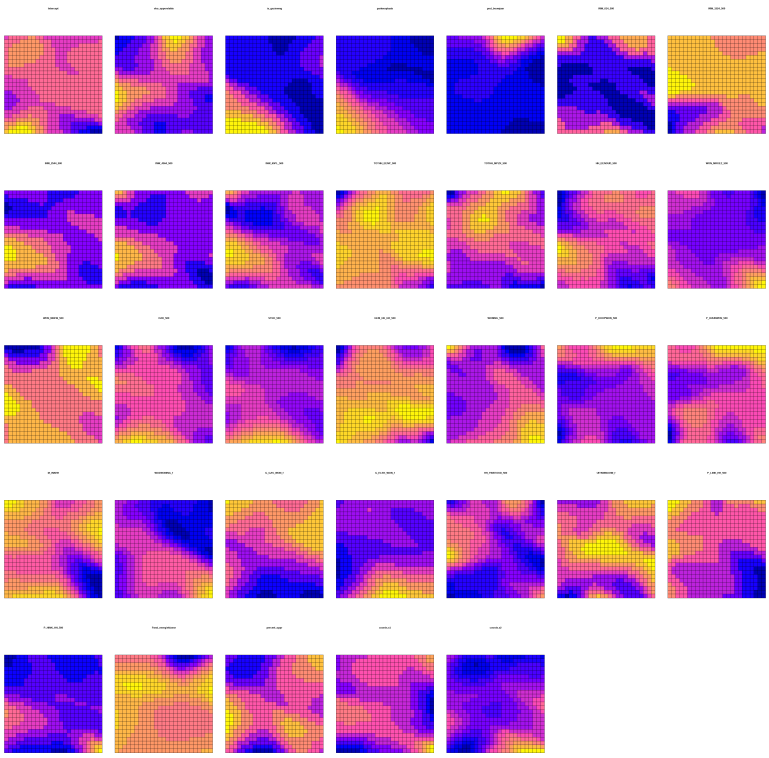


Figure A.3.11: Overview of spatial influences of all variables in GWR (Amersfoort, 2018)

GWR model results - 2020				
Municipality	R^2	*RMSE*	*MAE*	*MAPE*
Amersfoort	0.810	€61,928	€50,177	7.51%
Amsterdam	0.822	€62,596	€52,183	7.40%
Eindhoven	0.815	€62,942	€54,631	7.98%
Groningen	0.821	€79,192	€54,131	8.29%
Rotterdam	0.837	€58,561	€49,287	7.25%

Table A.3.3: Results for GWR models (2020).

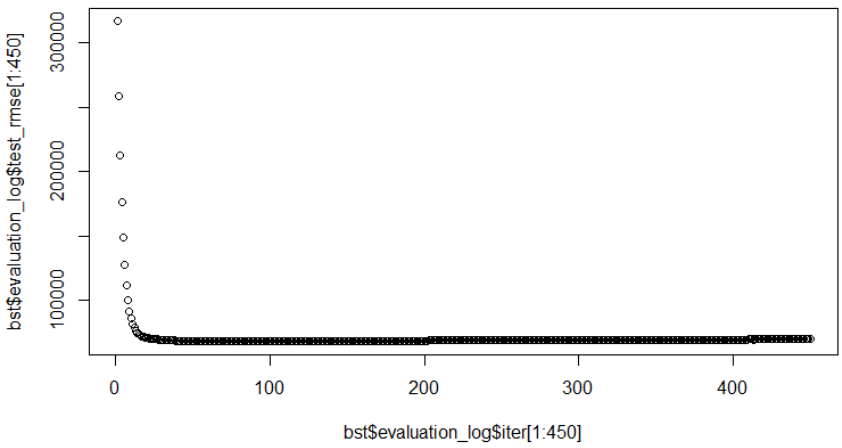


Figure A.3.12: XGBoost: Test set RMSE vs Number of boosting rounds (2018)

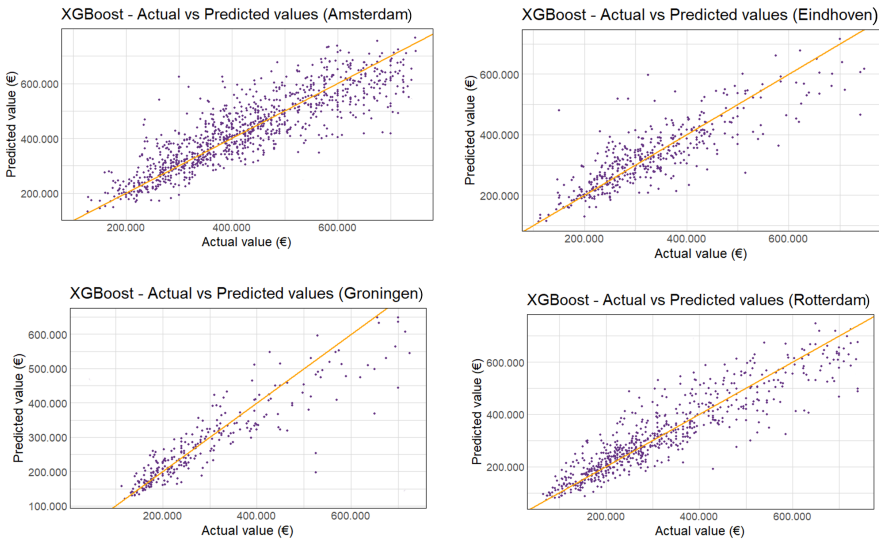
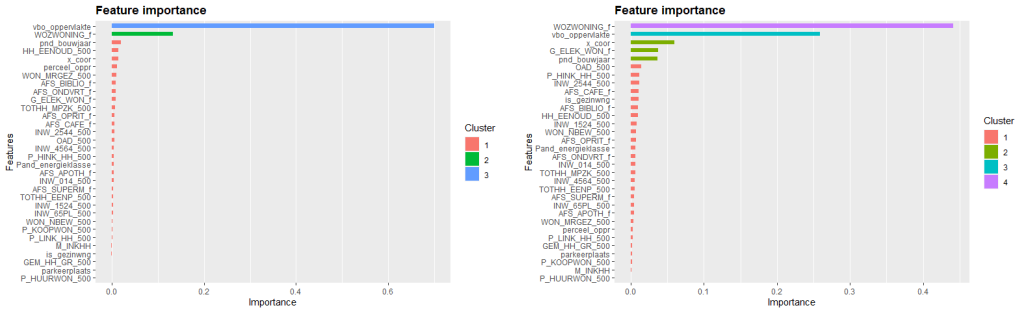


Figure A.3.13: Model fit of XGBoost models for Amsterdam, Eindhoven, Rotterdam, Groninningen (2018), (orange line is y=x)



(a) Amersfoort (b) Amsterdam

Figure A.3.14: XGBoost Variable Importance of Amersfoort & Amsterdam (2018)

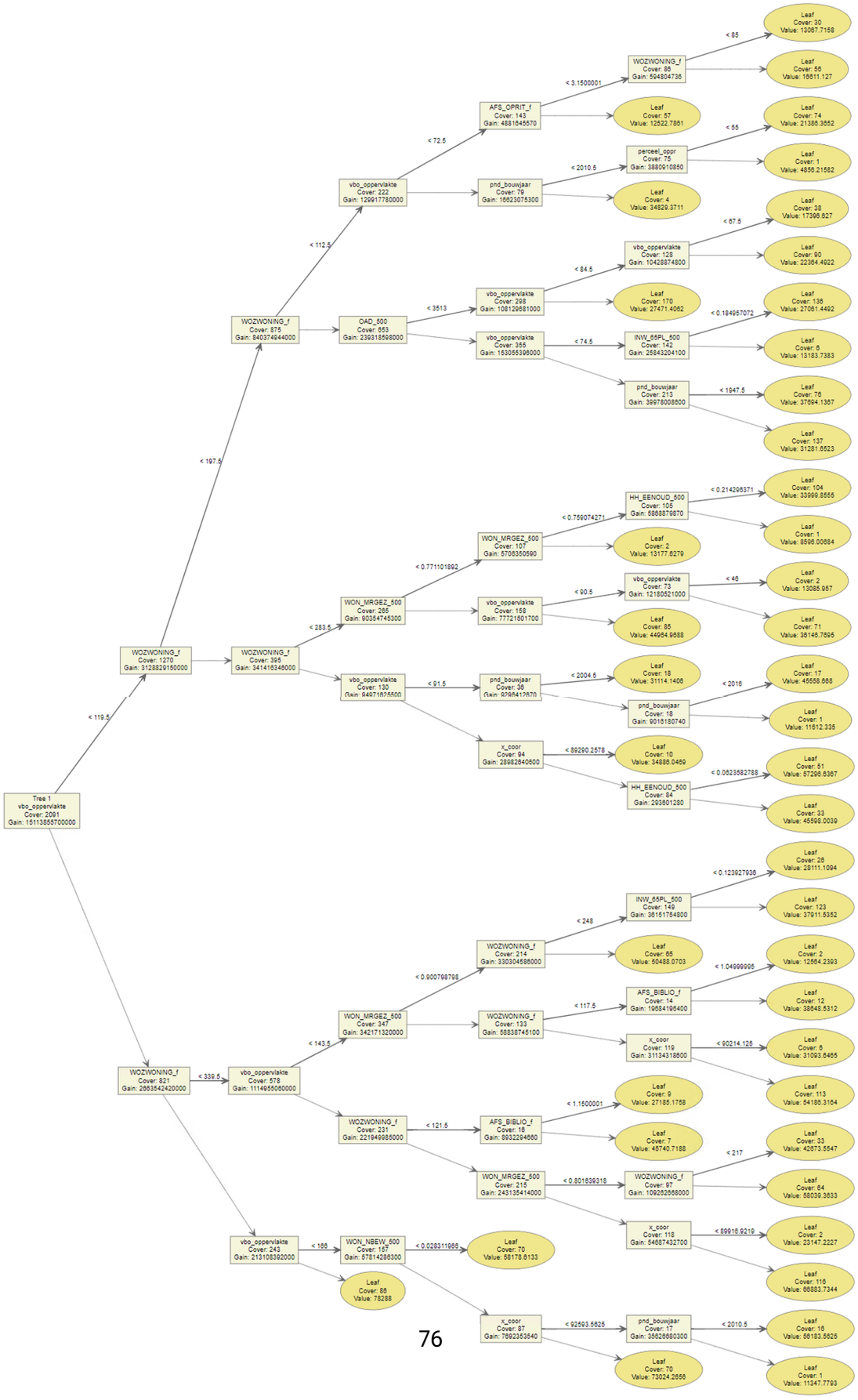


Figure A.3.15: First decision tree of final XGBoost model.