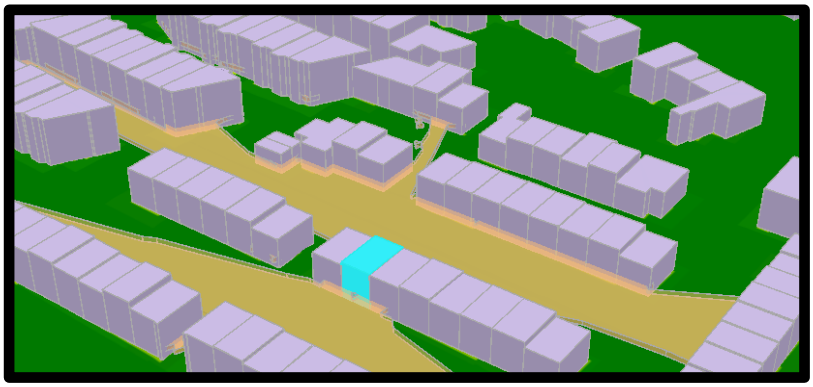
GIMA
Geographical Information Management and Applications

GIS based property valuation



Objectifying the value of view

Author: D.A.J.Oud
Study Area: Municipality of Alkmaar
Supervisor: Dr. H.J.F.M. Bouwmeester
Professor: Prof. Dr. Ir. P.J.M. van Oosterom
Date: August 2017

Universiteit Utrecht
TU Delft
UNIVERSITY OF TWENTE.
WAGENINGEN UR

Acknowledgements

In front of you lies my thesis written for the Master of Science ‘Geographical Information Management and Applications’. I genuinely hope I have been able to contribute to the body of knowledge on the crossroad of GIS analysis, statistics and real estate.

I would like to express my gratitude to Dr. Harry Bouwmeester for the support, continuous enthusiasm and constructive feedback throughout the entire research project. Right from the start he was genuinely interested in my ‘ambitious‘ plans. I would also like to thank dr. Peter van Oosterom for his detailed feedback, expertise and insights.

Without the database of NVM realtors the research could not have been conducted. Therefore I thank Frank Harleman for providing me with the required information and Dree Op ‘t Veld of Momentum Technologies for the additional recommendations and support regarding the project.

Finally, I would like to thank my family and friends, especially Arend, for their thoughts and loving support throughout the project.

Abstract

Automated valuation models evolved substantially since the ‘80s and are now the main practice for property valuation in The Netherlands, where they are used both for taxation purposes and the property trade market. Important factors for the value of one’s property are physical conditions of the house and the influence of the property’s location. The latter, however, is often insufficiently represented in an automated valuation model.

Incorporation of the spatial character of properties in property valuation can be pursued in two ways. On one hand the model can be improved in the data collection phase by inserting additional locational variables, on for example the quality of the surroundings, in the valuation model. On the other hand the modelling process itself can be improved by exploiting spatial statistical methods to specify the regression model. Much literature has been written on the two fields, though little is touching both.

The developments in the field of Geographical Information Systems eased the spatial approaches in automated valuation. GIS technologies offer the possibility to objectify information that was traditionally collected in a subjective manner, such as the view from a property. Furthermore, GIS technologies facilitate spatial regression models, that account for spatial errors. A main spatial error in regression analysis occurs when the property values are not functioning independent, since properties close to each other often show similar values.

This paper demonstrates the use of GIS applied to automated regression to estimate the value of a view on two clusters in a residential urban housing market. The outcomes of the study show that including the spatial variables on view, automatically computed with GIS, improve the property price predictions. Also, the spatial approach in regression modelling significantly improves the model fit. In the two clusters the best prediction model is the one that combines both spatial approaches.

Table of Contents

1. Introduction

The defining mantra of the real estate industry is considered to be *'location location location'*. Real estate properties are spatial goods of which the value is expected to depend upon its unique location. Despite the mantra was claimed to originate from the 1920s, property valuation models seemed to ignore the spatial character of real estate properties for decades. Omitting geography is often expressed in both the modelling techniques as well as the data input for the model.

In the Netherlands property values are the base for several taxations. Up until 1994 taxpayers could estimate the value of their property themselves. With the 1994 Valuation of Immovable Property Act this activity was shifted towards the local governments. Municipalities were informed they could use three different options for valuation, of which the third was Automated Valuation Modelling (AVM). Back then AVM was not popular, resulting that the two remaining manual options where most in use. By the time of 2008 the efficiency, objectivity and added value of AVMs was recognized, which changed periodical assessed mass valuations into annual assessed mass valuations (Francke, 2010).

The taxations based on the values of real estate properties are of great financial importance to the state, bringing in around 10 billion a year (Waarderingskamer, 2016). The valuations commissioned by the Municipalities reflect the market value of a property at the first of January of the previous year. Since this value is not considered to be an up to date market value, they are generally referred to as assessed values. Other uses of assessed values are determining the amount of mortgage loans and the level of insurance for real estate.

Another frequent user of property valuation models is the Dutch real estate agent. Real estate agencies aim to estimate a current market value. In contrary to Municipalities, realtors generally calibrate the objective model output with additional subjective information. To obtain a fair and up-to-date market valuation, realtors put emphasis on the quality and maintenance aspects of properties. They often perform physical inspections of the property to add this information to the valuation model output. They are also able to take into account up-to-date demand rate towards a specific neighborhood or location.

The both ways of using valuation models to estimate market values have two main differences, which are the time of assessment and model specification. Although time is predominantly a legal constraint, the model input is a rather technological constraint. Enriching the model input in other ways than physical inspections, is possible with digital data collection techniques such as GIS, which support performing advanced data analysis. The more advanced the input, the better the model is expected to perform. GIS technologies offer the possibilities to quantify information that traditionally had to be collected in a subjective manner, such as the type of view from a property.

Predicting property values requires two main practices, collecting the data and estimating the relationships between variables. For this research the strength of implementing GIS in valuation modelling is tested by including the influence of the view range in square meters around a property in the model. The visible space around a property is calculated by taking into account other properties as visibility obstacles in a 3D environment. Also the quality of the

view, based on land use types, will be considered. To estimate the strength of the relationships between the property value and its characteristics, a regression model is defined. This regression model will be specified within a GIS environment, considering the property's coordinates.

In the subsequent chapter the research framework is presented by explaining the research questions and study area. The concluding research methodology outlines the chosen approach to answer the research questions. The third chapter discusses the theory on the property valuation concepts. The main concepts are spatial analysis, spatial statistics, automated property valuation and viewshed analysis. In the fourth chapter the data for the regression analysis is prepared, discussing tools and techniques used to automatically quantify viewsheds. In the fifth chapter the results of the regression analysis are presented, those results show how certain variables influence the property value. In the sixth and seventh chapter the research questions are answered and recommendations for further research are made.

2. Research Framework

In this chapter the framework is set for the scope of the research. The discussed subjects are respectively the research objectives, the research questions, the study area, definitions, the considerations and the methodology.

2.1 Objectives

The primary objective of this research is to use an explicit spatial methodology in conjunction with a basic spatial regression model to test the significance of geographic variables on residential property prices. To generate the spatial valuation model this research aims to explore the potential of Geographical Information Systems (GIS). Central to the implementation of a spatial valuation model stands the purpose of the current Dutch valuation system, which proclaims that the model stays automated handling objective information (general purpose of assessed valuation), while improving in accuracy by including case-specific information (general purpose of market valuation).

The literature review will explore to what extend GIS-based tools and techniques can improve the current models of property valuation in the Netherlands. The emphasis will lie on two components:

- The first component is data enrichment. The potential of spatial data acquisition as input to the valuation model will be considered and tested.
- The second component is the regression analysis. Improvement in accuracy of the statistical model will be tested by comparing the model's predictive accuracy of traditional regression techniques and spatial regression techniques.

The outcomes of this research will be of relevance for all parties involved in the valuation of properties, since improved insights on both the explanatory factors of property prices as well as the efficiency of the workflow is essential for fair valuation. Applying regression analysis furthermore gains understanding of the willingness to pay for intangible goods, which contributes to justified decision making in the field of urban planning. New insights on the possibilities of GIS-based valuations will increase awareness for the needs of base registries such as the Cadastre to support the automated valuations.

2.2 Questions

In order to achieve the objectives of this research, the main research question that shall be answered is: *To what extent can spatial techniques improve the accuracy of automated property valuation models?* This main question is divided into the following sub-questions:

1. *What are the needs for current property valuation models?*
   This question tries to identify what the needs for current models are, focusing on both the needs for information and the needs for improved valuation methods.

2. *To what extent will model accuracy improve when including spatial variables?*
   Assuming that the location of a property weights in its value, this question tries to identify whether including locational variables improve estimating the property price. A model with merely physical variables will be compared with a model with both physical and locational variables as provided by NVM. Those locational NVM variables are collected by hand.

3. *How can subjective spatial information be quantified for automated valuation?*
   This question seeks to find an automated workflow for the quantification of information that is currently physically collected by an inspector. Could GIS for example quantify the view of a property?

4. *To what extent will model accuracy improve with the use of GIS?*
   As an addition to question 2 this question evaluates the influence of spatial variables that are automatically computed using GIS instead of collected by hand. Furthermore, this question reviews the implementation of local GIS regression techniques in predictive modelling.

2.3 Study area

The choice of study area depends on several factors. Since this research focusses on spatial autocorrelation and spatial heterogeneity of the real estate market, the area should be large and varied enough to measure those spatial phenomena. Therefore the focus will not be on just one neighborhood, but on an entire city.

Furthermore the city should have a real estate market that is representative for the Netherlands. Cities such as Amsterdam do not reflect the general market conditions since this capital city has an enormous pull-factor and therefore an excessive demand group. Cities such as Volendam as well have a distinct demand group since their community is relatively closed.

For this research the chosen study area is the Dutch municipality Alkmaar. Alkmaar is part of the G14, which are the 14 medium sized cities in the Netherlands. It is predicted that the real estate market of Alkmaar region will develop similar to the national market up to 2025, with an annual increase in real estate prices of 2% (EBZ, 2015). According to the Dutch statistics the municipality of Alkmaar contained at the end of September 2016 a total of 50.456 residential properties (CBS, 2016).

Alkmaar obtained city rights in 1254 and still holds its historic city centre, counting over 1000 monuments. In summer times the authentic cheese market attracts many tourists. Alkmaar is the 10th biggest shopping city of the Netherlands. The cities Alkmaar, Oudorp and Koedijk have a surface area of 3220 square kilometers with a housing density of 1533 homes per square kilometer (CBS, 2016).

2.4 Definitions

The use of the two terms *property value* and *models* in this research can cause confusion because of their ambiguity and are therefore explained in more detail.

For this research the value of property is separated into assessed values, market values and transaction values. The transaction value is very straightforward the amount that is paid for a property. The assessed values, '*WOZ-waarden'*, are in the Netherlands produced by tax districts that calculate their real estate taxes upon a percentage of this value. The value is estimated by a statistical model and generalized for all Dutch properties, taking into account objective property characteristics and transaction values of nearby houses with similar characteristics. The market values *'taxatiewaarden'* are produced by realtors that try to approximate the transaction value of a property in an open market as close as possible. Therefore they take into account subjective case-specific characteristics. The difference in property valuation termination could be confusing for Dutch readers since the direct translation for market values incurred by realtors is 'taxation values', even though the calculated market values are not used for taxations, those are the assessed values.

This research refers to two types of models, mathematical models representing phenomena and digitally drawn (3D) models representing real world features. The first definition applies to the regression model, of which an equation models the phenomena of property valuation. Also the viewshed model represents a mathematical workflow to calculate the view of a property. The latter definition applies to the 3D models that were used to draw the digital version of Alkmaar on which the viewshed model is based.

2.5 Considerations

As to each project with a demarcated time span, concessions in the scope of the research are made. This research is demarcated on the following elements:

- The research is limited down by merely focusing on residential properties in two cluster groups.
- The research is performed in one Dutch city, although this area has a heterogenic real estate market, it is possible that the research is not entirely representative for the Dutch market.
- Within the city of Alkmaar two smaller cluster groups are evaluated.
- The potential of GIS is targeted to Desktop GIS, the potential of online GIS is not included in this research.
- Because of technical restrictions, apartments could not be analyzed.
- The researcher has affiliation with Esri, QGIS, Geoda and SPSS software. The tools and techniques used in this research can all be found within those software packages.

2.6 Methodology

This study seeks to review the potential of geo-information in property valuation modelling. To test the presumed potential of geographical information systems, a spatial property valuation model will be developed and tested. The ultimate goal of this project is to determine the influence of view in the valuation of property prices. The project will be executed following a quantitative approach, using spatial analysis and spatial statistical regression techniques.

The approach for determining the property value is hedonic. A hedonic valuation model is based on the assumption that a homebuyer values the characteristics of the property, rather than the property as a whole. This means that the property prices reflect the prices of the property characteristics, including the locational variables that homebuyers consider in their purchase. When using a regression model the value of each characteristic can be determined. In the figure below the conducted workflow of this research is presented.

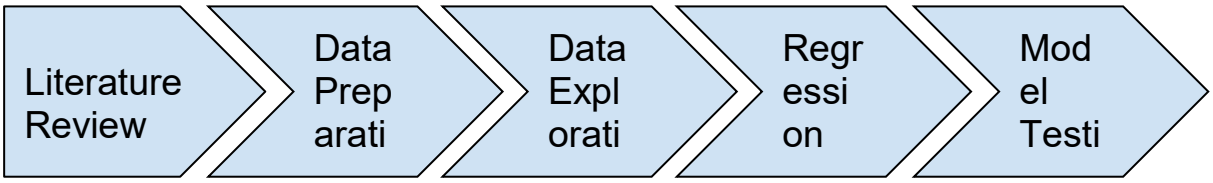


Figure 1: Research methodology

Literature Review
The research starts with a literature review to identify the state-of-the-art of property valuation. This chapter discusses both the variables related to the value of a property as well as the tools and techniques for spatial analysis and spatial statistics. The secondary information is complemented with primary information out of interviews and conversations.

Data preparation
As a result of the literature review, the main variables that are important to the value of a property are selected, cleaned and georeferenced to insert in the valuation model. Cleaning is performed in SPSS, georeferencing in Esri's ArcGIS Pro. The additional GIS-based variables viewshed and view quality are prepared by using spatial analysis techniques of the ArcGIS Pro software package.

Data Exploration
In between the data preparation and the modelling phase it is important to thoroughly understand the data. Summarizing statistics, mapping the data and testing model assumptions will give a deeper understanding of the data and provide better informed decisions on the specification of the model.

Regression Modelling
To determine whether the prepared variables actually correlate with property values, a regression model is prepared to estimate their influence. First a non-spatial global model is executed in GeoDa (SPSS also suffices). Then spatial statistical methods are implemented to consider the spatial character of properties, by using a local regression model in ArcGIS Pro.

Model Testing

The best fitting models are tested on their predictive accuracy by using a test set. The test set contains 10% of the cluster that were not used for modelling purposes. Estimating the property values of the test cases based on the proposed model will directly show the usability of the prediction model.

3. Theoretical Framework

In this theoretical framework the spatial approach in both valuation modelling and data acquisition will be discussed. This chapter will be the backbone of the model input that will be discussed in the subsequent data preparation chapter. After a brief introduction on the global discourse on spatial property valuation, the focus is set towards Dutch valuation needs. Subsequently, one need is selected for further research, which is viewshed analysis. Techniques in the field of spatial analytics are proposed to fulfill the need for incorporating viewshed variables in the valuation model. The last section discusses the regression methods suitable to predict the property values, resulting in an analysis framework for explicit spatial modelling.

3.1 Developments in the field of property valuation

This section outlines the main concepts of property valuation within the approach of automated and hedonic modelling. The two main research field supporting this are spatial econometrics (3.1.2) and spatial analytics (3.1.3).

3.1.1 Automated and Hedonic Modelling

The market value of a residential property is the price one should expect in an arm's length transaction between informed and willing buyers and sellers. This value depends on the property's structural and locational characteristics. The objective characteristics can be obtained easily, while subjective indicators will require a physical inspection of the property. When the characteristics are collected, they are put into a model to predict the property's market value.

Most Automated Valuation Models (AVM) only take objective property characteristics and transaction information to fit a statistical prediction model. This method is less detailed than one with additional subjective information, but highly cost efficient. The main input for automated models are structural property characteristics such as size, age and type of property. In AVM minor attention is paid towards locational and quality characteristics since its methodology lacks physical inspections.

Central to property valuation stands the hedonic modelling approach. The method treats the good as a composition of its characteristics. Each of the characteristics contribute to the eventual price of the good, which makes it able to infer willingness-to-pay of certain characteristics. Hedonic modelling is particularly applicable to heterogeneous goods such as real estate properties, since homebuyers are assumed to value each characteristic of the house separately. Ridker and Henning (1967) are the first ones to apply this method to real estate properties, quantifying the intangible influence of air pollution.

The first years after the work of Ridker and Henning much research has been conducted to the hedonic method itself, though not to the application of this method in estimating the value of other marginal characteristics. This trend was observed by McLeod in 1985. He stated that

any property characteristic that differs across the housing stock and weights in the buying process, has the potential to influence the housing price pattern. In that case an hedonic regression model can be used to value these characteristics.

In response to the work of McLeod many aspects for property valuation have been considered, varying between locational, environmental and neighborhood characteristics. Examples are studies that examine the influence on property values related to proximity to certain amenities (Dekkers & Koomen, 2008), proximity to open space (Luttik, 2000; Irwin, 2002), presence of noise pollution (Lake et al, 1998), and the neighborhood income levels (Cavailhes, 2009). The studies show significant relationships between the spatial variables and the property values.

The above-mentioned studies indicate that incorporating spatial factors in automated valuation models is essential to acquire fair and accurate mass valuations, especially since the use of automated modelling is becoming common practice for property valuation. This automated mass valuation requires mass data acquisition techniques, which are not possible without the use of computer assisted tools and digital maps, since physical data collection of spatial variables is highly time-consuming.

Geographical Information Systems (GIS) have the strength to handle large sets of georeferenced spatial data in a digital manner. The four components of a GIS are data capturing & preparation, data management, data manipulation & analysis and data presentation. GIS have rapidly evolved since the late 70's, both in their technical and processing capabilities (Huisman & de By, 2001).

Due to major technical developments in the fields of econometrics and geo information, both feeded by GIS (figure 2), spatial automated valuation modelling now has the tools and techniques available for its two core components; spatial data analysis and spatial regression. Though in practice, those GIS-based techniques are not used to their full potential, often lacking one or both spatial approaches.

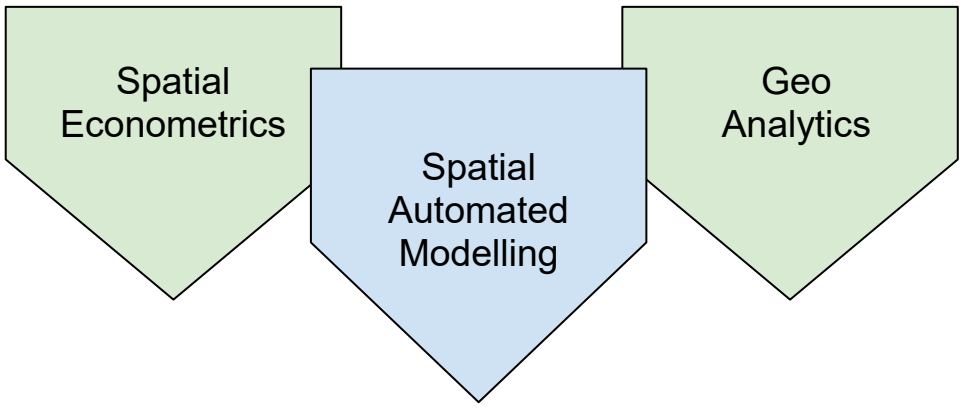


Figure 2: Spatial econometrics and Geo analytics as drivers for spatial automated modelling.

3.1.2 Geo Analytics with GIS

Geo analytics is the heart of GIS, where data is transformed into information. Functionalities of analysis have evolved for the past decades and are still growing. The most common analytical functionalities can be divided into analysis on a single dataset or on multiple datasets. With a single dataset, retrieving data by selections and measurements are common (figure 3 image A). With multiple datasets overlay analyses are used to acquire information on one location, neighborhood analyses are used to acquire information on the surrounding area of a location and connectivity analysis are used to acquire information over a network. Those analysis functionalities will be explained below with examples on property valuation analysis.

When starting spatial property valuation analysis at least a dataset table with the property's location is required. When working with location information on this single data layer, point features containing the attribute information will be displayed. In case of apartments, when the height of the property is known, the points can be extruded in space. Often the geographical data is supported by non-spatial attributes such as price, size and age of the property. Basic analyses on this layer can be retrieving all feature locations from a certain building year. When there are building features instead of points a basic analysis is to measure the the property size or volume.

To find answers on the influence of locational factors on property values, the required information can be obtained by combining two or more datasets. The information in the aforementioned researches on the influence of proximity to amenities or open space, noise pollution, and the neighborhood income levels, can be digitally retrieved with the use of GIS analysis. Noise pollution statistics are retrieved by overlaying noise measurements, on for instance roads, with the property's location (figure 3 image B, overlay analysis). Proximity to open space can be determined by selecting all unbuilt land use within an euclidean distance range by using buffers (figure 3 image C, neighborhood analysis). Proximity to parks can be determined over a network by including a third road network dataset (figure 3 image D, connectivity analysis).

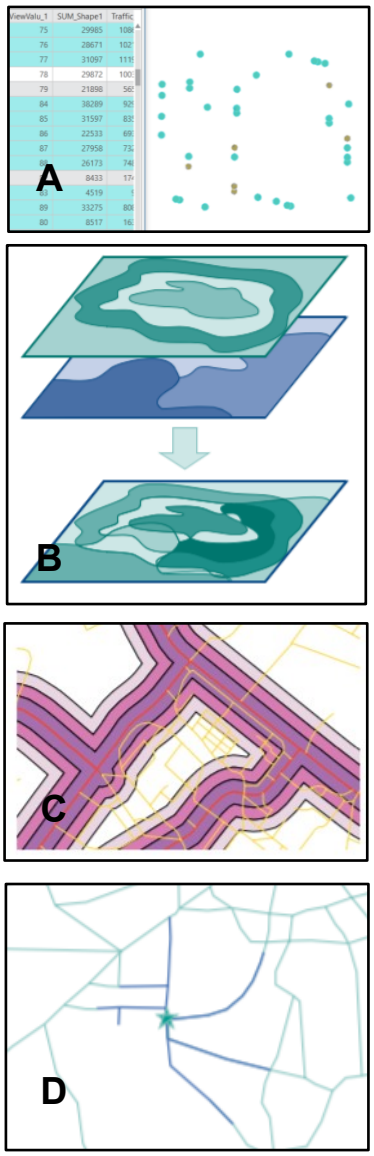


Figure 3: Spatial analysis techniques. Source: Huisman & de By, 2001

The above-mentioned data analyses describe how locational information can be digitally retrieved without performing physical inspections. Another benefit of using GIS is the ability to perform this in an automated manner. Data transformations can be performed for a large group of entities at the same time. When a sequence of data transformations is required, this can be expressed in a model that will process all transformations automatically.

Although GIS makes data analysis more accessible and time efficient, collecting and updating the input data for the analysis can still be difficult and immensely time consuming. The

mentioned park locations and unbuilt land use have to be collected before they can be transformed. Fortunately, in recent years there is a growing trend of releasing open data in the Netherlands. The amount of data in authentic registries is growing vastly, 11 active registers provide data on topics as topography, the subsurface and cadastral information. The strength of authentic registries is that collection, quality management and maintenance is centralized. Public sector institutions increasingly provide their base datasets as open data at online portals, even private parties are picking up this trend (Welle-Donker et. al., 2016).

Parallel to the growing amount of open data, there is a rapid rise in the use of open standards, facilitating seamless data interoperability. Web services are frequently used to directly communicate through open standards with other servers to request data, without the need to download the data. In recent years, new releases of open data are becoming more sophisticated, providing advanced datasets such as up-to-date volunteered geo-information (VGI), real-time sensor data and three-dimensional point clouds. With such data, advanced spatial analysis are gaining ground in the field of GIS.

3.1.3 Spatial Econometrics with GIS

The collected information on a property can be used to predict the value of the property, all characteristics increase or decrease the value with a certain weight. Valuation techniques and automated prediction models for mass appraisal began developing during the 1970s and 1980s (Anselin, 1998). The main employed regression technique for valuation models is Ordinary Least Squares (OLS). This functional form draws a best fitting regression line between two variables, assuming that variation in the errors terms are random.

However, when estimating spatial goods, often the errors are subjected to spatial patterns. Which means that for instance the *size* of a house is in some areas a stronger indicator for the property's value than for other areas. In OLS the mean weight is taken for the whole region, flattening the unique weight indicators related to *size* for different areas. For real estate valuation, this spatial dimension influencing the characteristics was found to be the main distinguishing characteristic, making it a contribution to the creation of a separate field of study (Rodriguez, 1995). This separated field of study is the intersection of econometrics and spatial analytics, since 1979 referred to as 'spatial econometrics' (Anselin, 2010).

Spatial econometrics strives to take into account the spatial character of certain goods in regression analysis. Although the stage was set, the application in the field lingered. Back in 1998 Anselin, one of the principal developers in the field of spatial econometrics, stated that *"despite widespread recognition by both theorists and practitioners of the complex roles of location and spatial interaction and the resulting geographically segmented nature of real estate markets, an explicit ''spatial'' treatment of these markets in empirical real estate research is still in its infancy".* This belief is shared by many other frequently cited researchers at that time (Pace et. al. 1998; Dubin et. al. 1999; Fotheringham et. al. 2002).

In 2010 Anselin again reflects upon the developments in the field of spatial econometrics and concludes that it has grown from the margins to the mainstream, showing a somewhat exponential growth. Since the turn of the 21st century also the access to software on spatial econometrics was no longer an impedance to spatial statistics, since proprietary and and open

source toolboxes for regression analysis within GIS environments became widely accessible (Anselin, 2010).

When regression analysis is performed within a GIS environment, all properties have a unique location in space, based on a geographical coordination system. A spatial weight matrix tries to capture the spatial structure of the data. This weight matrix is used as an overlay on the variables to indicate at which places a certain weight should be enhanced or reduced. In section 3.4 the methods for spatial regression are further elaborated upon.

3.2 Property valuation in the Netherlands

3.2.1 The Dutch valuation framework

Since the annual assessed valuations of 2008, AVMs are common practice in the Netherlands. The Dutch valuation models are used for both taxation and market valuation purposes. Municipalities and real estate agents often use the services of companies to perform the actual valuation or cooperate with those companies (Isikdag et al, 2015). In 2011 the OTB research institute of Delft University of Technology identified the models of 9 of the 15 companies active in modelling. Of those companies, 5 were specialized in taxation valuations, 2 in market valuations and 2 were specialized in both (Bouwmeester et. al., 2011).

To perform valuations both building information and cadastral information are required (Isikdag et. al, 2015). The valuation offices maintain their own building information models and request upon the Cadastre's legislative information on boundaries and ownership. This construction depicts that not all valuation models rely on the same set of data. The Dutch Council for Real Estate Assessment (Waarderingskamer) set up a list of quality standards that should be met to obtain fair valuation. They explain that a fair valuation needs at least the basic physical characteristics, which are location, type of property, building year, size of the building and size of land.

Literature on assessed valuation models and market valuation models suggest that the required characteristics are often enriched with additional characteristics such as energy label, number of rooms, and annexes to the house. In the Netherlands realtors are involved in the transactions of around 90% of all residential properties. The main main Dutch realtor organisation is *The Netherlands Organisation of Real Estate Brokers (NVM-realtors)* with a market share of around 76% of the residential properties in 2016 (NVM, 2016). NVM-realtors maintain a rich collection of data over several years that is used for the valuation models. De database depends on the input of real estate agents. Newly added properties can directly be reviewed by other agents and the information can directly be used for comparison analysis. In the next section their dataset is presented.

3.2.2 Dutch valuation needs

For the purpose of this research NVM-realtors provided their property valuation dataset, of which the variables are shown on the next page (figure 3). The variables presented in **bold** are the variables required for a fair valuation, as described by the Dutch Council for Real Estate Assessment. The table confirms that besides the required variables, the NVM-realtors building model contains a broad range of additional information.

Considering this dataset is used to generate a regression model for property valuation, the variables can be subdivided into explanatory and descriptive variables. Both the spatial and physical explanatory variables are similar to so called independent variables, which are able to explain the estimated market values of properties. The descriptive variables can be seen as a framework supporting the dependent variable 'property value'. Within this framework there is information about legal issues and there is administrative information on the temporal, spatial and terminology matters concerning the dependent variable.

Explanatory Variables			Descriptive Variables	
Spatial Information	Physical Information		Legal Information	Administrative Information
Fine location Heavy traffic In centre	Category **Property type** **Building period** Rooms Garage Furnished Volume Inbuilt Garage New estate Isolation Basement Quality Elevator Monumental **Size (m^2)** Balconies Toilets Floors	Inside maintenance Outside maintenance Fireplace Parking **Parcel size** Shed Apartment Roof type Housing type Garden size Heating Living shape Living size Basic attic Attic fixed stairs Garden position	Investment Leasehold Partly Rented Buyer condition Sales condition Status	Entry date Closing date Duration Initial listing price Initial listing price m^2 Ultimate listing price Ultimate listing price m^2 Transaction price Transaction price m^2 Neighborhood **House number** **House letter** Postal code 4 **Postal code 6** Postal code **Street name** District **Place of residence**

Figure 3: NVM realtors database (delivered by NVM). Required information in bold.

The explanatory variables are used to specify the prediction model. From the dataset it can be concluded that the spatial information corporated in a valuation model is limited. Only a few variables, 'situated in the centre', 'road with heavy traffic' and 'fine location', are taken into account as spatial explanatory variables. The description of the classes within the two latter variables show subjective indicators of the property's location. The distinguished classes are *quiet road, busy road, no information* and*, at forest, at water, at park, open view, no information.* A third and main drawback is that the spatial variables are obtained during physical inspections and are subjected to change, therefore these variables are not useful for automated and repeated valuations.

To satisfy the needs of both the assessed valuations and the market valuations, a model with objective and measurable spatial information is desired. Information on traffic can be automated and quantified by connecting digital sensor information about traffic along the property's location, although this information is not yet publicly available at street level. Fine

location can be automated by digitally determining the property's surroundings. Information on surroundings, such as land use types, is in the Netherlands provided as open data based on the authentic register on topography. This makes determining 'fine location' in a quantified manner a suitable variable for automated property valuation. In the next section the possibilities for including viewshed in the valuation model are discussed.

3.3 Viewshed valuation

In this section quantifying the variable 'fine location' as the view from a property is discussed by first focusing on the international body of knowledge, followed by an overview of current technologies.

3.3.1 Lessons Learned

To determine the surroundings of a property, simple buffer and overlay techniques on two datasets will suffice. Research on property valuation show positive results when evaluating the proximity to open space, especially when the open space is directly surrounding the property (Dekkers & Koomen, 2008). However, in those researches open space is taken into account as a buffer around the property, not as the actual view from the property. For two neighboring properties, their values may differ because one can actually see a certain amenity while for the other the view is blocked. Defining the viewshed of a property is different from the traditional adjacency calculations since it requires a third dataset, the obstructions.

Early literature on incorporating view in property valuation in a quantitative manner dates from 1985 when McLeod measured the influence of river view by using a dummy variable. The data was manually collected in the field, visiting 270 properties. Measured in a regression model, the research pointed out that properties that did have a river view were valued higher than similar properties without the view.

Since 1993 researchers started to consult visibility analysis in GIS, mostly for land use planning such as measuring the visual impact of placing wind farms on certain locations (Howes & Gatrell, 1993). The first research to include visibility analysis in property valuation was Lake et. al. in 1998, where the visibility of roads was calculated using surrounding buildings as obstacles to the view. After the view was calculated, an overlay with land use types was performed to extract road surface (Figure 4). This research showed that visibility of road has a negative impact on the value of properties. Most interesting to this research was not the outcome but its methodology, Lake et. al. showed that it was possible to measure the influence of view in a quantitative manner without performing on-site data collection, but using base datasets.

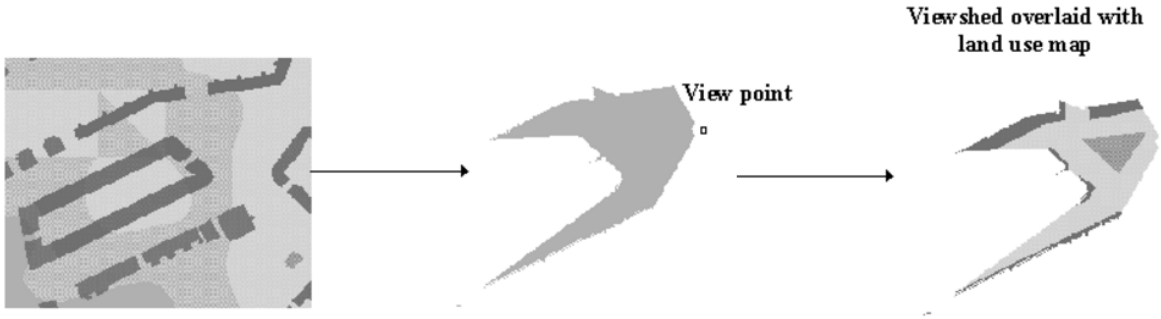


Figure 4: Non-automated viewshed valuation based on point features. Source: Lake et. al. 1998

Where the previous two studies used dummy variables to indicate whether there is a view of a certain amenity or not, Paterson & Boyle (2002) included different types of view. They differentiated four types of land use; development, agriculture, forests and surface water. This study compared the visibility of land use types with the presence of land use types around the property, showing interesting outcomes of positive correlation with the presence of forests while there was a negative correlation with the visibility of forests. This suggests that view is truly a consideration of property purchasers.

The research of Yu et. al. (2007) introduced the 3D-GIS approach to property-based visibility analysis, which makes it possible to capture both the horizontal sight and vertical sight from a 3D point, which could indicate an observer at a certain floor or view from a window. This is useful when properties are situated in high rise buildings, and thus able to overlook certain obstacles. The increase in highrise buildings enhances the need for 3D valuation. Recent literature is investigating the possibilities of a 3D cadastre and 3D building models that support automated property valuation (figure 5; Tomic et.al., 2012; Isikdag et. al., 2015).

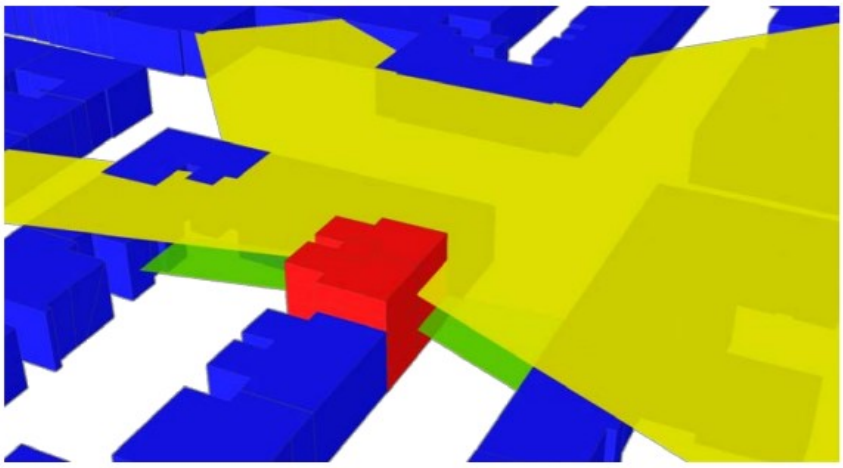


Figure 5: Non-automated 3D viewshed valuation based on point features at different heights. Source:Tomic et. al. 1012

The lessons learned from previous surveys show that property-based visibility analysis with a 3D-GIS approach is possible, albeit for a limited set of 3D observers. None of the consulted surveys provides a workflow for view calculated in an automated manner. In this research the analysis focusses on automated mass valuations, using large observer datasets.

Central to the visibility analysis stands information on the observers and the obstacles. When addresses of properties are known, point features can be used to indicate the observer position. The main obstacles in a built environment are other buildings. Buildings in the Netherlands are open data and can be linked to address information. However, when using the address point as observer and buildings as obstacles, the building that the observer is situated in, will also be taken into account as obstacle. The point features will be enclosed by the building, resulting in a biased viewshed of merely the inside of the building (figure 6 image A and B).

For visibility analysis with only one observer, deleting the building enclosing the observer address point solves the problem, this is the case in the example of Tomic et. al. (figure 6 image C). When there are multiple observers of which the visible area should be calculated in an automated manner this method cannot be used, since the building of a certain observer should be taken into account as an obstacle to other observers.

As a workaround for mass valuation, the outlines of the building can be used as observers instead of the address point. The line features than overlap the building outlines, solving the problem with the enclosed point feature (figure 6 image D). A drawback to this solution is that most visibility analysis functionalities are based on point features as observer input, including 3D functionalities such as variable observer z-values and visibility volume calculations.

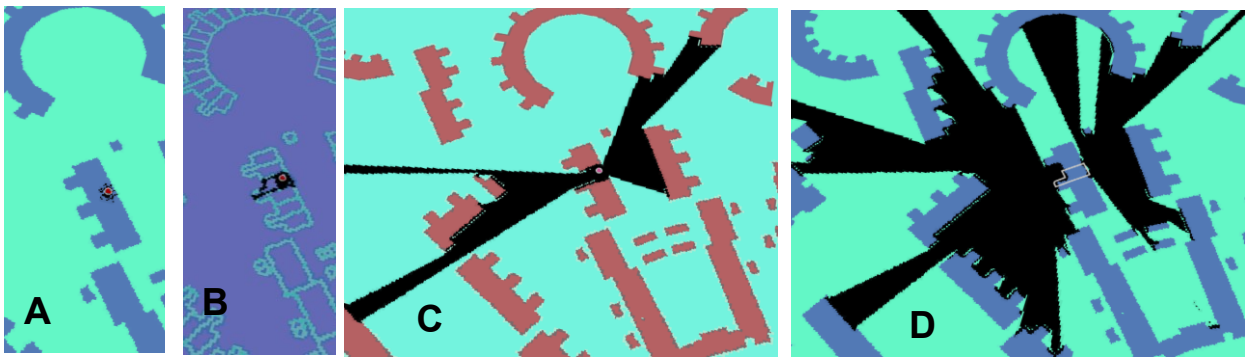


Figure 6: A & B. enclosed point problem; C. deleted house problem; D. line feature solution. Source: author's own.

The ArcGIS Pro visibility tool *'viewshed'* that is able to work with line features as the observers input takes a fixed observer height for all observers, which means that visibility cannot be calculated for properties at differing heights, such as apartments. Although the environment in which the viewshed is calculated can use three-dimensional information, the output will determine for each raster cell whether it is visible or nonvisible to the observer, lacking information on the elevation of the visible area.

3.4 Spatial Hedonic Regression

This section discusses the methods to obtain a regression analysis for property valuation. The models are divided into three groups: non-spatial global models calculated for an entire area, spatial global models taking into account adjacency of observers and spatial local models taking into account the geographic location of the observer.

3.4.1 Global non-spatial model

The general purpose of regression analysis is to determine the relationship between the dependent variable (Y) *'value of a property',* and several known independent variables (X), based on the weight parameter of each of these variables (β). In case all parameters are estimated, the unknown dependent variable can be calculated in a predictive model.

Developing a spatial statistical model is an iterative process. After the specification and preparation of the independent variables, the parameters will be estimated for the first time. All variable combinations used for the model should be tested on indicators such as significance, redundancy and performance. Often a re-specification of the model is required to improve the model fit. The functional form of the traditional linear regression model is the non-spatial Ordinary Least Squares model. Within the framework of multiple regression this model can be described as follows:

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \ldots + \beta_n * X_{ni} + \varepsilon$$

Y_i = dependent variable at location *i*
β_0 = intercept or constant parameter
$X_{1i,2i,\ldots ni}$ = independent variables at location *i*
$\beta_{1,2,\ldots n}$ = slope parameter (weight)
i = location
ε = error term

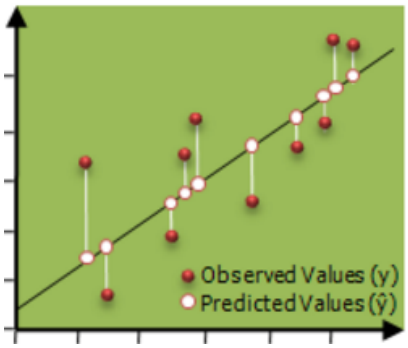


Figure 7: Best fitting line in regression analysis. With multiple regression, there is a best fitting plane. Source: Esri, 2017

The estimated parameters are obtained by drawing a best fitting plane through the known values (figure 7). Since a global model is estimated, a single equation is used to represent the process. Independent values related to unknown property prices can be filled in the equation.

3.4.2 Global spatial model

As noticed by Quigley (1979) properties have a unique combination of features; they have a fixed position in space, are heterogenous goods, bring high costs for change, and have high durability. Due to this, locational effects are an integral part of the way the housing market functions. Tobler's first law of geography, *"Everything is related to everything else, but near things are more related than distant things.",* explains the spatial dependence, or autocorrelation, that is influencing regression models (Tobler, 1970). It occurs that prices in one location are correlated with prices in nearby locations. When using a non-spatial regression method for spatial goods, the model fails to capture the true effect of the independent variables. After the OLS model is estimated, it can be tested on the presence of spatial autocorrelation. A proper spatial approach in property valuation is both the recognition of the importance of spatial effects and their implications for spatial statistics (Anselin, 1998)

When spatial autocorrelation turns out to be apparent in the OLS model, alternative functional forms should be examined to explain price variation in properties. Spatial global models that correct for spatial dependence by including a neighborhood weights matrix based on adjacency, putting increased weight on a set amount of neighboring observations. Adding weights to neighboring observations has proven that spatial dependence is accounted for (Bidanset & Lombart, 2014; Anselin, 1988).

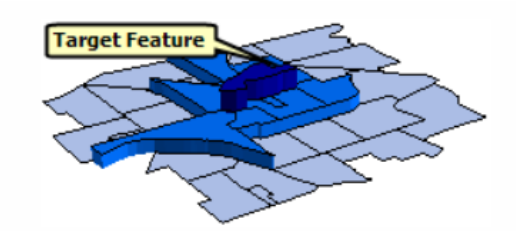


Figure 7: Neighborhood weighting
Source: Esri, 2017

3.4.3 Local spatial model

Spatial heterogeneity is a second common phenomenon in spatial modelling. Often the effect of spatial dependence does not have the same influence on the correlations in all locations of the study area. Also, relations between variables can fluctuate over space, this indicates that the model is subjected to spatial non-stationarity, or spatial heterogeneity (Fotheringham et al., 2002). Previously mentioned global models do not capture the effect of spatial heterogeneity, therefore models with a local focus are required.

In global models, taking into account five neighboring observations can signify that four observation points are actually close by, while the fifth point is a few kilometers away. A local regression method accounts for spatial heterogeneity by taking into account the actual location of the target observation and the geographical distance to the locations of nearby features, in order to produce local regression results for each unique location. Real estate properties are because of their locational immobility and heterogeneous nature assumed to be subjected to the locational effects of both spatial autocorrelation and spatial heterogeneity, making property valuation a proper candidate for the local regression method.

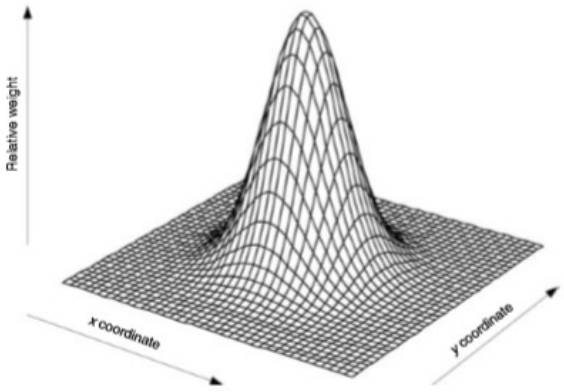


Figure 8: Gaussian Kernel interpolation.

Geographically Weighted Regression is a local regression method described by Brunsdon et. al. in 1996. The method operates by assigning weights to all observations depending on their distances to a geographical focal point. The weight system is based on distance decay, using a kernel function that reduces the influence of distant observations and emphasizes the influence of nearby neighbouring observations. The functional form is as follows:

$$Y_i = \beta_0(i) + \beta_1(i)*X_{1i} + \beta_2(i)*X_{2i} + \ldots + \beta_n(i)*X_{ni} + \varepsilon$$

The difference with the former OLS equation is that in this case the parameters differ per location, making a unique equation for each observer location.

The weights in GWR are mainly assigned using the Gaussian kernel (figure 8). In practice it matters little when other kernels than the Gaussian one are used, as long as the kernel is '*Gaussian-like*' (Fotheringham & Charlton, 2009). Important in this method is the choice of the kernel's bandwidth. The bandwidth parameter can be adjusted to either widen or narrow the shape of the kernel, this will depend on the density of the observed points and the fit of the model (Borst & McCluskey, 2008). An interpolated raster of each of the variable parameters shows the patterns of the spatial heterogeneity of the variable across the study area (figure 9).

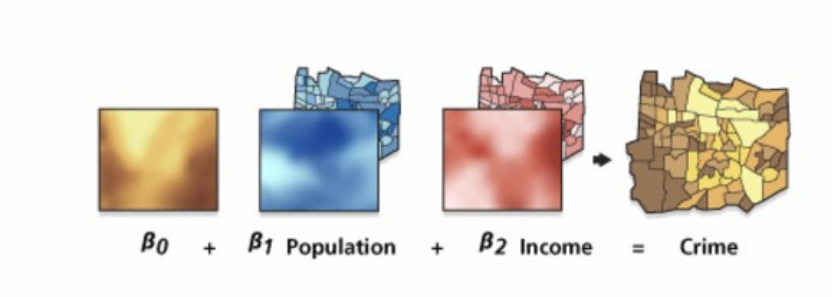


Figure 9: Patterns of parameter intensity over space. Source: Esri, 2017

3.4.4 Spatial Hedonic analysis framework

Based on the consulted national and international literature on spatial hedonic analysis of property valuation, the analysis framework below (figure 10) is proposed. Starting on the left, stepwise tools and techniques should be exploited to pursue an integral spatial approach to hedonic property valuation. The proposed OLS estimation does include the spatial variable 'view', but the observers are not spatially weighted. In the GWR prediction, the variables are spatially weighted. The OLS estimation is required to calculate the model fit.

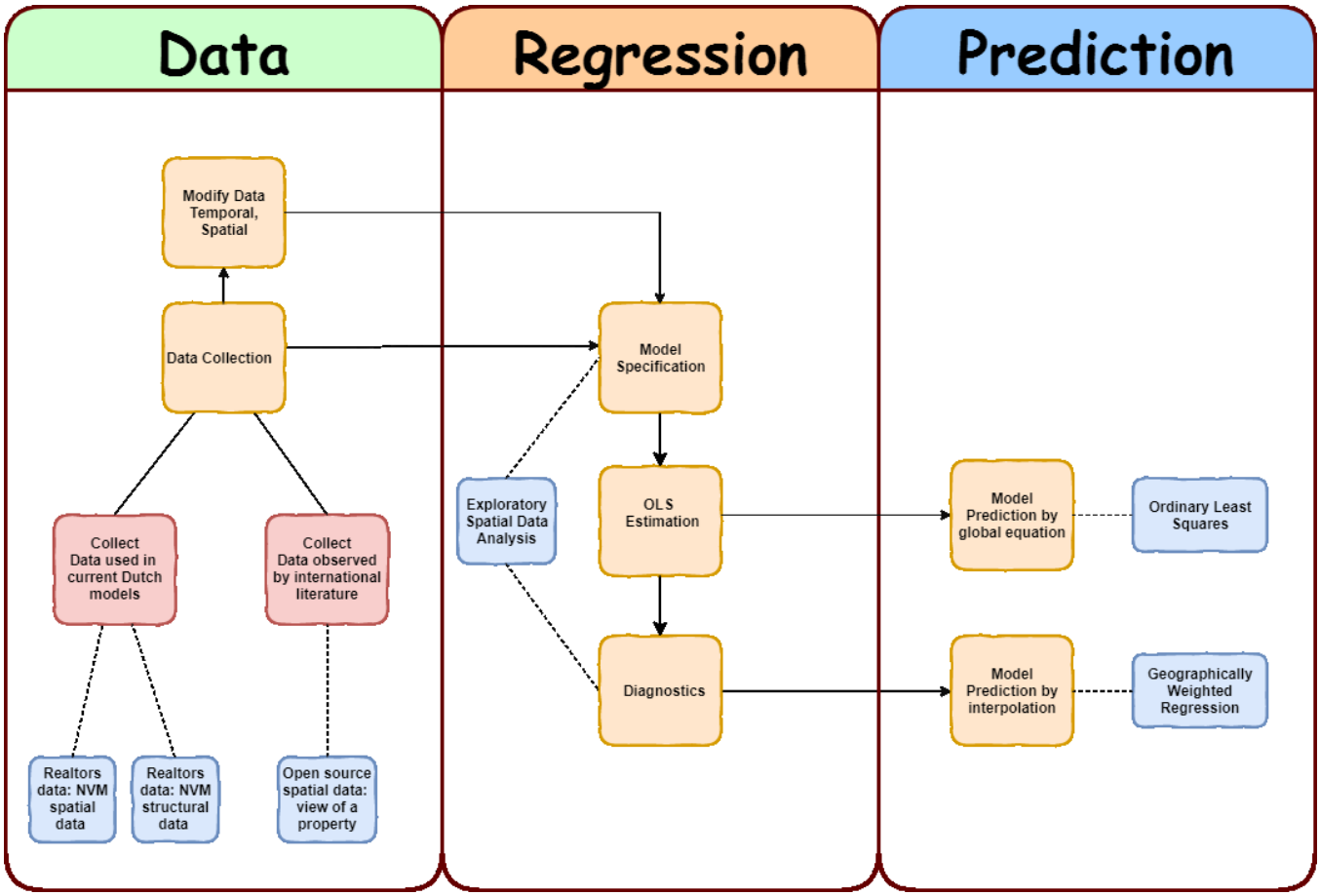


Figure 10: Spatial Hedonic Analysis Framework Source: author's own.

4. Preparatory data analysis

As discussed in the literature study spatial influence on property prices is generally explained by a hedonic model, in which the property value is explained by a set of property characteristics. Within a hedonic regression model the correlation (relationship) between the characteristics and property value can be determined with a training dataset. Each correlation is calculated with a certain degree of confidence (significance), this information is used to built the prediction model.

The property characteristics to be tested in the regression analysis are obtained in two ways. The first part is the provided valuation dataset of NVM-realtors. The NVM dataset contains mainly physical characteristics of the property. To answer the research question "*To what extent can spatial techniques improve the accuracy of property valuation models?",* this set of characteristics will be expanded with additional spatial data generated in this project using GIS techniques.

This chapter will outline what steps are taken to come to the design of the hedonic prediction model. The process can be divided into four stages (figure 11). The first stage is the cleaning and clustering of the NVM-realtors dataset (section 4.1). In the second stage new variables are generated through the use of GIS data analysis (section 4.2). When all data is prepared stage three requires an exploration will be performed on suitability for regression, followed by the last stage, the design of the regression models (section 4.3).

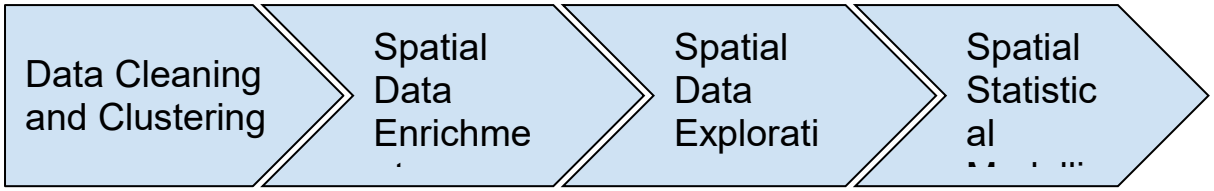


Figure 11: Four stages of preparatory data analysis. Source: author's own.

4.1 Data Cleaning & Clustering

The modelling of the housing prices was conducted with the use of data from the Dutch Organisation of Real Estate Agents, the NVM-realtors. The database consists of 17052 (houses) cases and 68 variables over the period 2005 - 2015 in the Municipality of Alkmaar. The NVM has a market share around 80% in the region of Alkmaar (NMA, 2012). In this chapter the transformations to the dataset are examined.

4.1.1. Variable specification

The variables in the NVM dataset are, for the purpose of this research, subdivided into explanatory and descriptive variables (section 3.2.2, figure 2). In this section the descriptive variables are employed to set the framework of the model. The explanatory variables are used to determine the model design.

The main consideration drawn from the *administrative variables* is what price the dependent variable 'property value' should be based on. The aim of the model is to estimate the market value of a property, by performing a realistic valuation of what the property is worth in a certain market at a certain point in time. This values is shown by transaction price, when obtained from free market negotiations. The listing price of the property is generally higher than the transaction price since the seller starts with an optimistic amount which gives him a buffer for negotiations. In the case of Alkmaar, listing prices are on average 3.9% higher than the actual transaction prices. Due to this, the market value model will be based upon transaction values.

The remaining administrative variables will be used to account for spatial and temporal issues. The date variable will be used to allow for the expected inflation over 10 years. The location variables will be used to create an 'address' variable on which the dataset can be geocoded. Geocoding is the process of adding real world coordinates to address text strings. Spatial analysis is only possible when the cases are geocoded.

Concerning the *legal variables* The Dutch Council for Real Estate Assessment pursues two main valuation rules for properties, which are full ownership and immediate obtainability. This means that rent, partial rent and leasehold cases should be filtered out of the dataset. Furthermore, the International Valuation Standard prescribes that transactions should be possible in a free market between a willing buyer and a willing seller. This results in filtering out properties that are for instance sold as investments or at an auction, of which the second is the consequence of forced sale.

Of the remaining *explanatory variables* a selection is used for the model design. They all have the potential to contribute to the explanatory power of the model. Although there is a rich dataset of 37 variables, the model should stay as simple as possible. There are several variables that are more fixed than others. The fixed variables will be used to make clusters of homogenous property groups. When the influence of marginal variables are tested within one homogenous group, their influence will be the focal point of the analysis. When the influence is tested in a global model it is possible that their influence is flattened because their impact only applies in some cases. Fixed variables used for the clusters are building period, category, property type and size (rooms). The explanatory power of the fixed variables will be presented by the intercept parameter in the model equation.

4.1.2. Data cleaning

In the previous section a framework was created for the use of variables. In this section the data that is not desired for the model is erased from the dataset or corrected for. The main variable is the dependent variable, transaction price. To make the dataset operable for the regression analysis, it is important that all cases contain data on the transaction price. Apart from being available, the provided price information should be reasonable. By considering low prices and missing values as input errors, the transaction prices lower than €50.000 are eliminated from the dataset. The limit is based on the minimum price range that can be consulted for Alkmaar at the main Dutch real estate website Funda.nl.

Reviewing the remaining transaction prices, it was observed that some high transaction prices did not seem to be reasonable since their transaction values were multiplied by the power of

10 compared to their listing prices, assuming a typing error in data entry (e.g. listing price of €279.000, transaction price of €2.550.000). Also, the price per square meters appeared improbable high compared to others in their segment. In response to the evidence those transaction prices are corrected.

Furthermore, undesired transactions mentioned in the previous section, determined by the legal variables, are eliminated from the dataset. Those transactions include rent, partial rent, leasehold, investments and auctions. Next to this, cases covering multiple addresses and essential missing values on the fixed variables are removed. The revised dataset comprises 9537 valid cases for regression analysis.

4.1.3 Temporal corrections

Alongside the explanatory variables provided by the NVM, economical market conditions have their influence on property prices. Especially the past 10 years the property market has experienced major fluctuations. Starting 2008 the global financial crisis struck the Dutch market, resulting in a fast decline of property prices. This drop lasted until 2013, after which the market increased slowly. Due to the market fluctuations, the exact same properties have had different market prices at different points in time.

To be able to predict property prices based on transaction prices from the past, the prices are adjusted to the present current market conditions (Op 't Veld et. al., 2008). All cases are corrected to the price level of 2016, based on the most recent Price Index of Existing Properties provided by the Dutch Statistics (CBS, 2016b). The indices of both the Netherlands and the province North-Holland are shown in figure 12 and indicate the tipping points of 2008 and 2013. The disaggregated local index of North-Holland is used for the price corrections in Alkmaar.

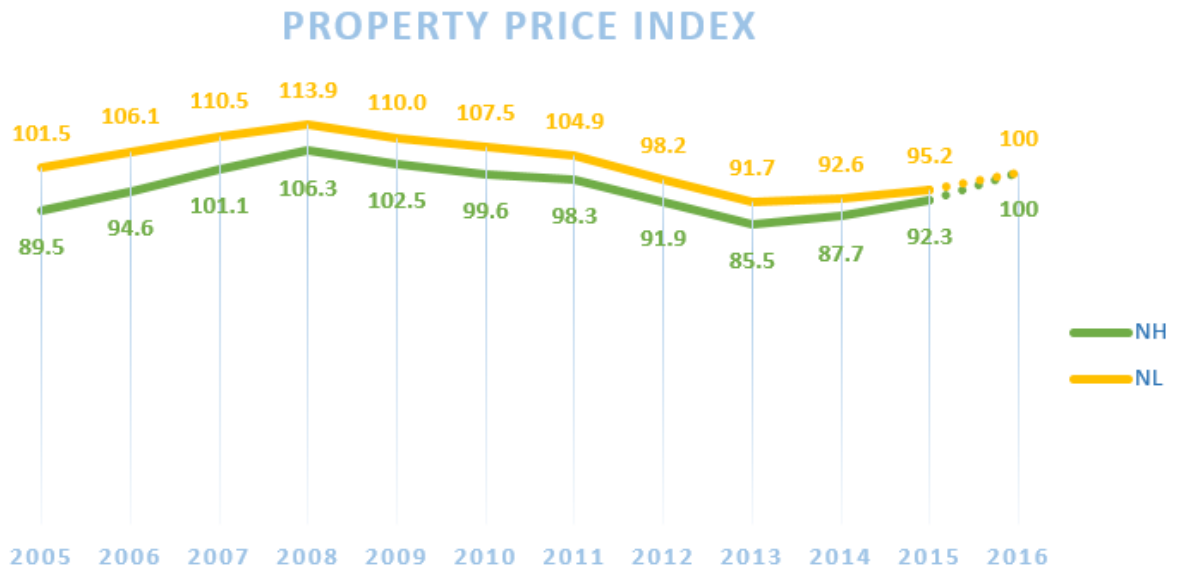


Figure 12: Property price index developments based on 2016

In figure 13 the blue line indicates the development of the actual transactions prices under their market condition at specific points in time. The orange line indicates what the transaction

price would have been when the property was sold in 2016. In line with the index, in the years 2007, 2008 and 2009 the properties have been sold for more than they would bring up under current market conditions. For the other years the opposite applies, there they brought up less than they would in 2016. With the price adjustments, all property prices are corrected for inflation in the real estate market.

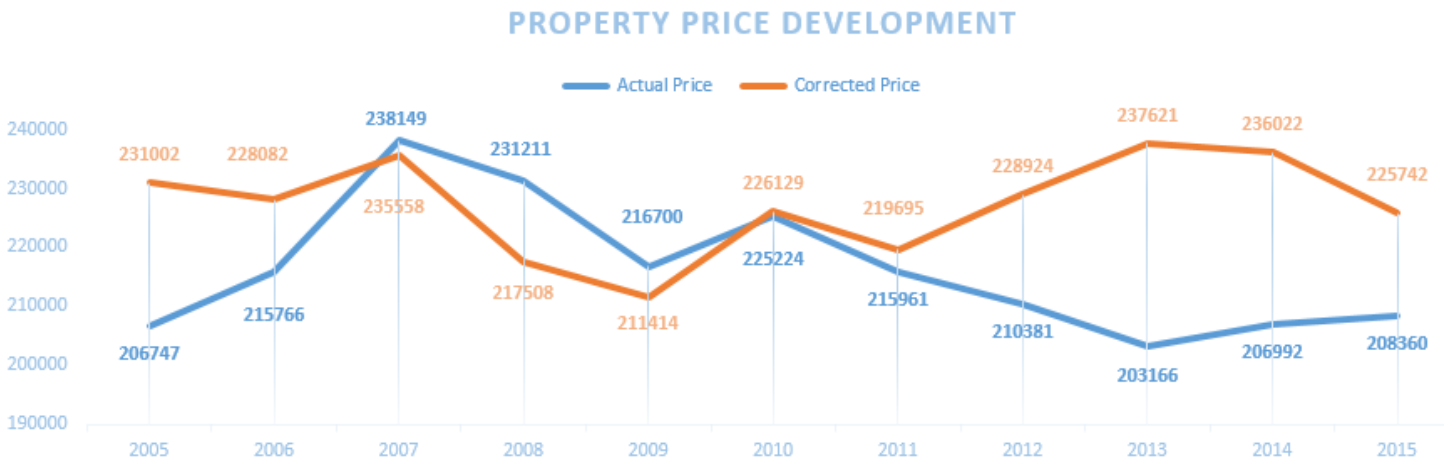


Figure 13: Property Price developments, actual price and corrected price to 2016 market conditions.

4.1.4 Spatial entities

Since a spatial regression is desired, a third round of cleaning is subjected to the geocoding process. With geocoding, the properties become spatial entities, by linking their addresses to coordinates on a map. The process was performed using the Dutch Registration of Buildings and Addresses (BAG). The addresses of the NVM do not follow the same semantic rules as the BAG addresses, which required additional data cleaning. The geocoded cases do all have a unique location on the map, which qualify them for the spatial data analysis. A map of the geocoded properties (7954) can be found below (figure 14). The deceased amount of properties is due to unclear addresses in the NVM database.

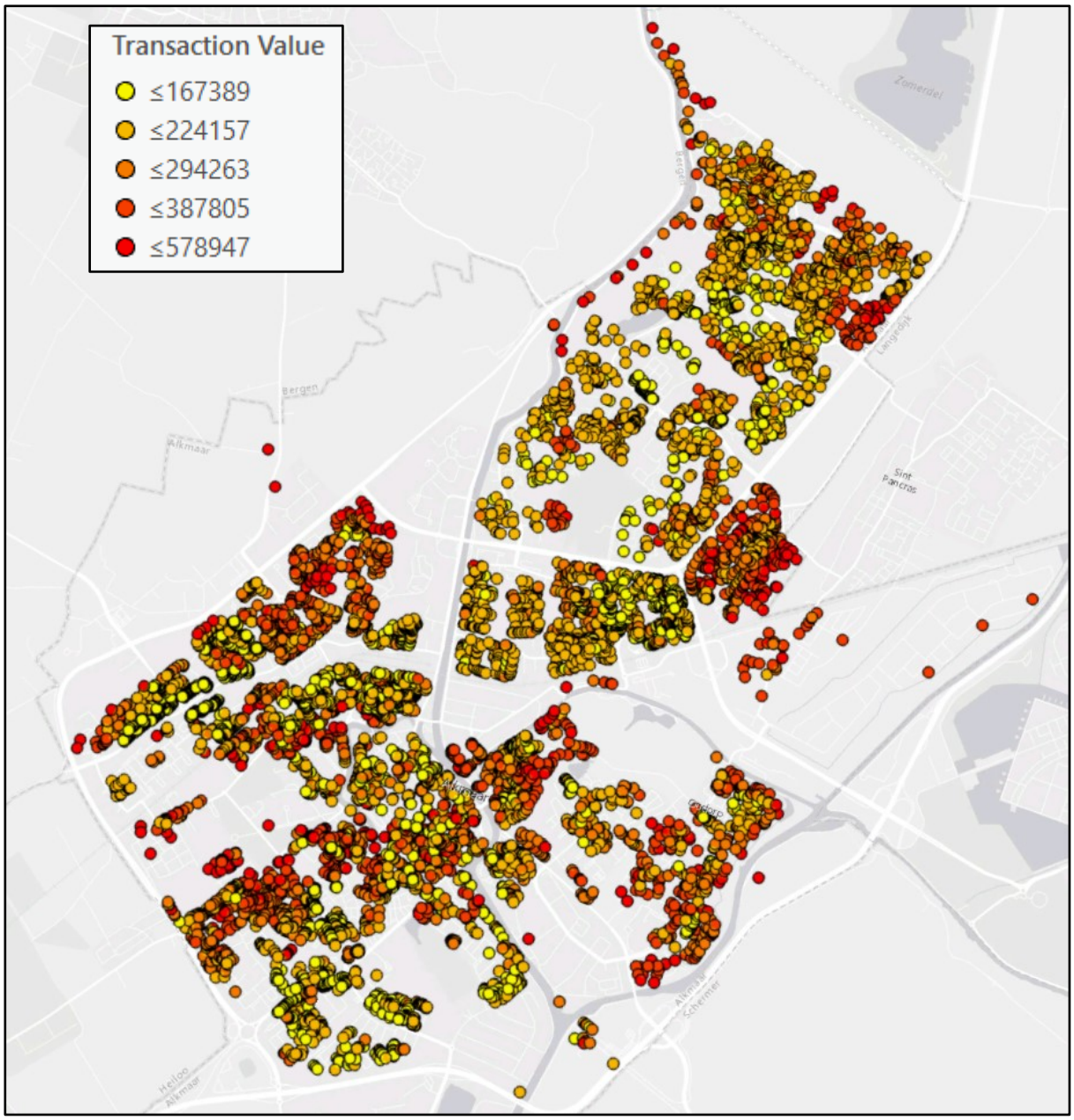


Figure 14: Geocoded properties, classified on property values.

4.1.5 Cluster selection

The study area Alkmaar Municipality locates the city Alkmaar and the villages Koedijk and Oudorp (appendix I). According to the most recent BAG database Alkmaar municipality contains 54504 unique addresses. NVM provided their database including property transactions and information over the period of 10 years. After data cleaning and geocoding, information on 7954 transactions is available for analysis.

Within the dataset 5692 cases are categorized as houses and 2262 as apartments. Since the viewshed analysis is not yet suitable for observer height specification in mass viewshed analysis, the apartments are not taken into account while clustering. The predominant housing type in Alkmaar is single family housing (4816 cases). Other main characterizing variables are building period and size of the property (Momentum Technologies, 2017). To calculate the influence of specific variables on property prices, clusters of similar groups are made to create a large constant, which will highlight parameters of marginal variables.

The clustering method is presented in appendix II. Two major clusters are identified that together represent 23% of the NVM houses in Alkmaar Municipality (figure 15). The clusters are both 5-room single family houses. The first cluster contains houses built between 1971 and 1980, covering 656 properties. The second cluster contains houses built between 1981-1990, covering 634 properties. The main difference between both clusters is their building year.

ID	Category	Type	Size	Building Year	Properties
1	House	Single Family	5 Rooms	1971-1980	656
2	House	Single Family	5 Rooms	1981-1990	634

Figure 15: Housing clusters based on property type, size and building year.

When mapping the building year variable in space, the spatial character of this variable becomes apparent (appendix III). The oldest buildings are situated in the centre of the city. The more recently build houses are situated at the outskirts of the city. This means that a comparison of both clusters will show how variables are acting in different environments. It can be assumed that most variation between both clusters is therefore based on location. In case an explicit spatial model is created, including all possible locational and physical characteristics of the properties, one prediction model should be able to capture both clusters.

4.1.6 Splitting the dataset

A last step in data transformations is the division of the dataset. Since predictive regression models will be developed and compared, the models should be tested on their predictive accuracy. When the design and tests are performed within the same dataset, the model may overfit. A common division between a build and test set is respectively 90% - 10%. With the known transaction values of the build dataset the parameters of the model are specified and put into a model. With the 10% test set the predictions are performed. The predictions will be compared with the actual transaction values to verify the accuracy of the model.

4.2 Viewshed Data

As GIS alternative to the NVM ‘fine location’ variable, the view from a location is used to measure the surroundings of a property in an automated manner. Comparing the NVM ‘fine location variable with the GIS ‘view’ variable in a regression model, will show the strength and importance of both variables in calculating property valuation. The view is calculated using Esri’s viewshed tool. The output of this tool is a visible raster area. When converted to a polygon, an overlay of the the visible area and land use in this area is performed to be able to measure the influence of distinct types of view.

For the Netherlands, it is possible to perform the analysis entirely on open data. The two input layers required for the viewshed analysis are 3D elevation as obstacle layer and the cluster buildings as observers. Since march 2013 the government of the Netherlands provides an open data elevation set (AHN). This makes it possible to perform the visibility analysis using 3D obstacles. The land use dataset is employed after the viewshed is measured.

In figure 16 below, the workflow describing data layers and tools is presented, as suited for the case of Alkmaar. The red blocks are initial data layers, the blue blocks are data layers made through analysis and the yellow blocks are the output data layers. In the following sections the workflow is discussed in more detail.

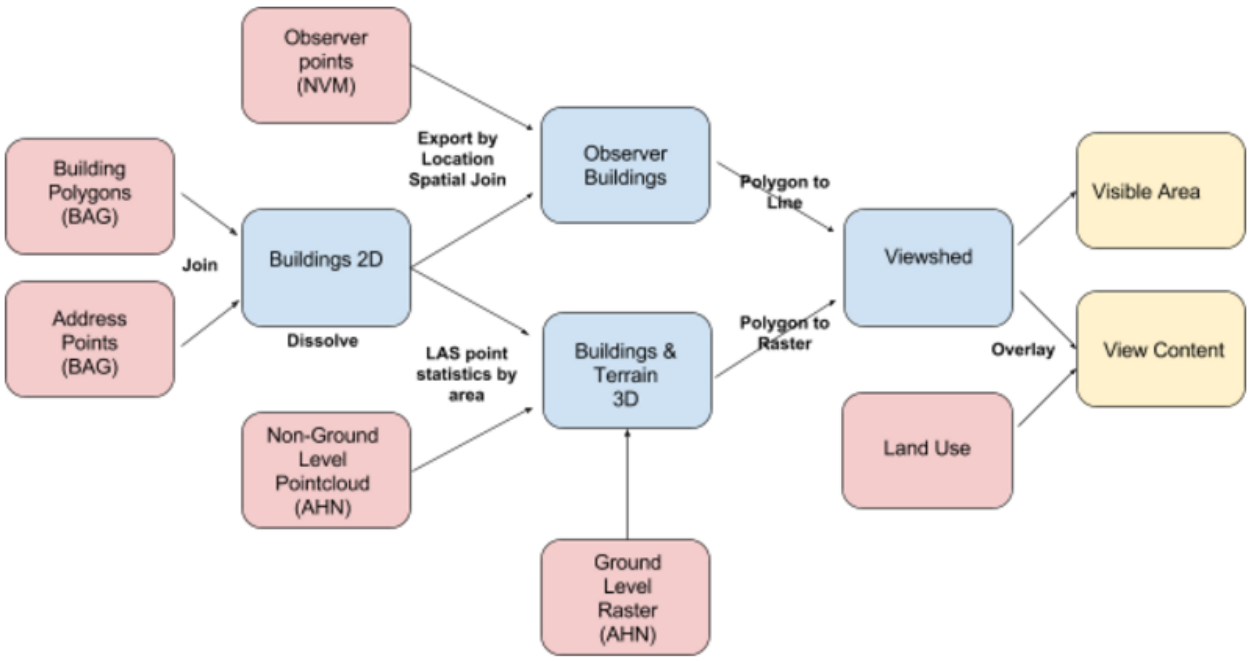
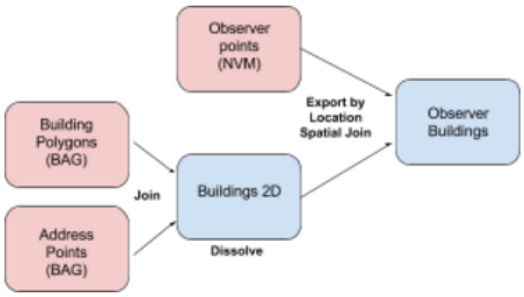


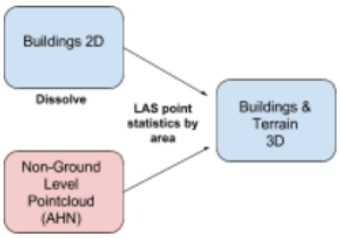
Figure 16: Workflow for automated view analysis

4.2.1 Observer buildings



NVM-realtors provided a dataset with properties in Alkmaar containing address information. In the Netherlands the base register for buildings and addresses (BAG) is publicly available and holds point features for addresses and polygon features for buildings. Joining both tables, a 2D building layer containing buildings with addresses is obtained (figure 18). Based on the point location of the observers the observer buildings can be extracted from the dataset as a separate layer.

4.2.2 Buildings in 3D



The second main datasets for viewshed analysis is the obstacles layer. When the obstacles have height information the visibility can be performed in more detail. Within the Netherlands an open data elevation model (AHN2) is available (figure 17). The datas is collected using LiDAR technology. This dataset can be downloaded as point cloud data or raster data. To extrude the buildings to 3D the point cloud data containing height values is used as summary statistics on each 2D building footprint. For each building the mean height value is extruded (figure 19).

As an alternative to this method, the topographic base register top10NL 3D building dataset can be used in which the height attributes of the AHN2 are already added to the features. However, a few drawbacks indicate that combining BAG and AHN is at the moment a better method. BAG webservices are updated daily, while Top10NL is based on aerial photos that were collected on average 2 years earlier. For a large part of the Netherlands the AHN3 is already available, making for those areas the combination of BAG and AHN3 a better option than using top10NL 3D. Also, since BAG focusses on buildings and top10NL on overall topography, building contours in BAG turn out to be more detailed. Furthermore, top10NL does not provide address information at property level, which makes a connection to the BAG data inevitable.

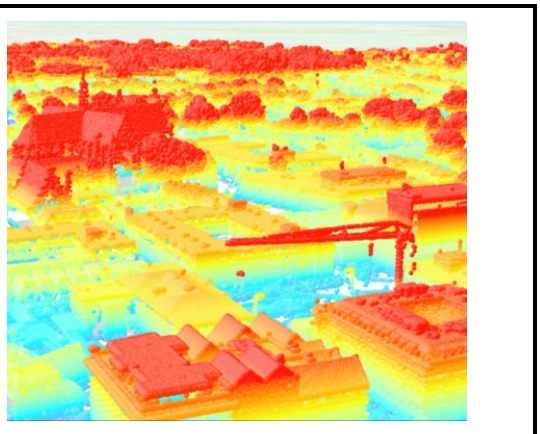


Figure 17: Elevation of non-ground level features in point cloud format.

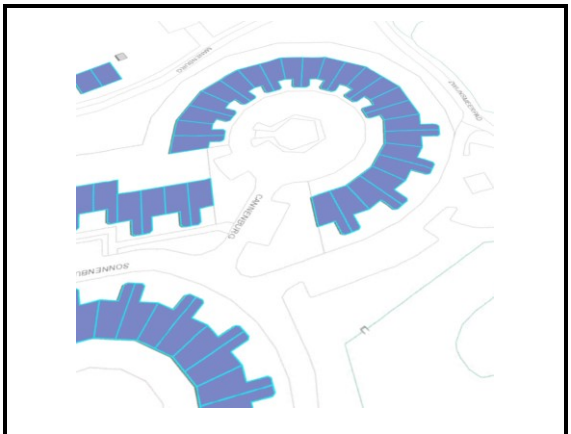


Figure 18: BAG building footprints in 2D.

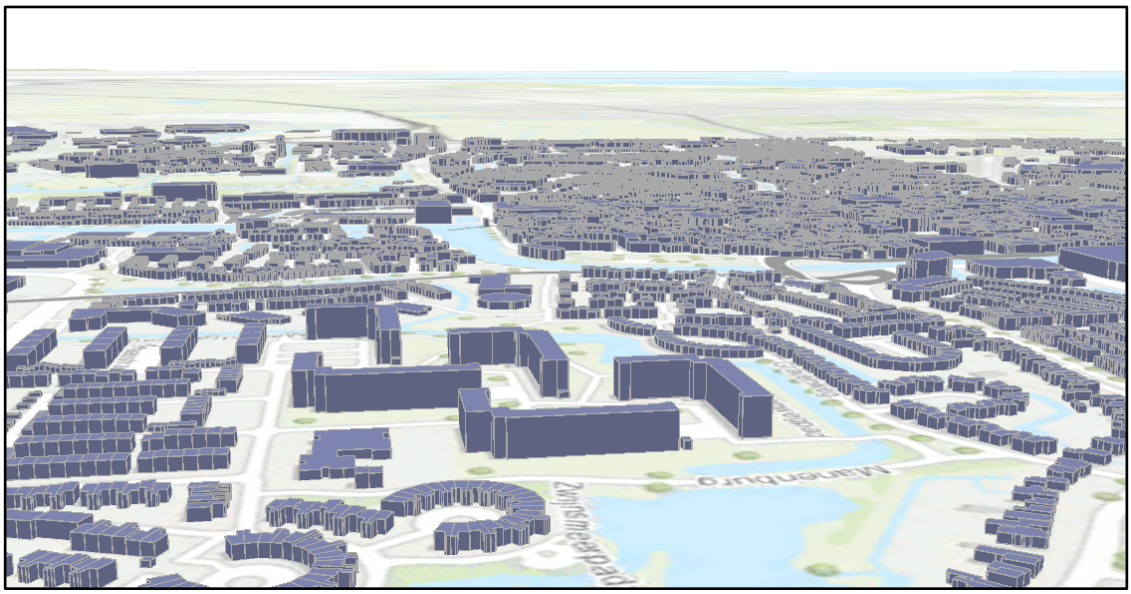
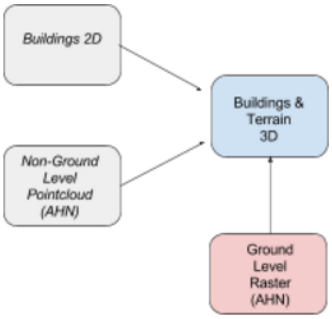


Figure 19: Extruded BAG buildings using AHN elevation data.

4.2.3 Digital Surface Model



Although the 3D buildings elevation will suffice as surface layer for the viewshed analysis (figure 20), adding ground elevation could improve the view analysis, especially in areas with significant relief. The AHN database provides elevation data sets in multiple formats. The elevation of the ground level is available in raster format, in which non-ground level features are filtered out (figure 21). Combining this data with the 3D building model generates a more realistic elevation model (figure 22).

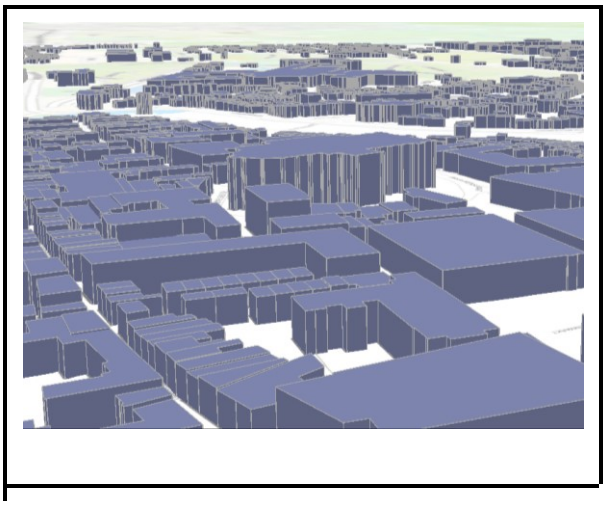


Figure 20: Digital Building Model

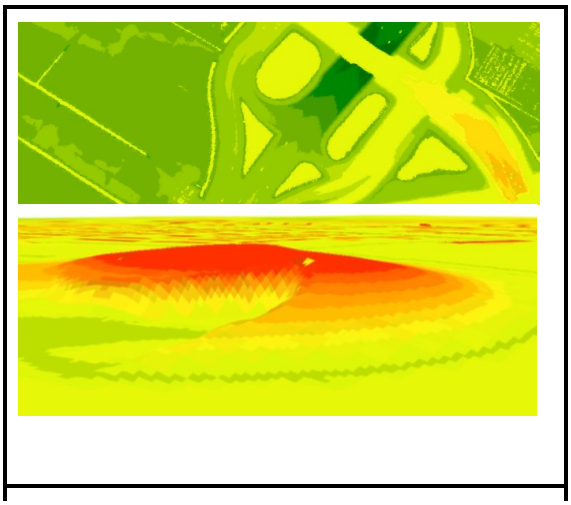


Figure 21: Digital Surface Model

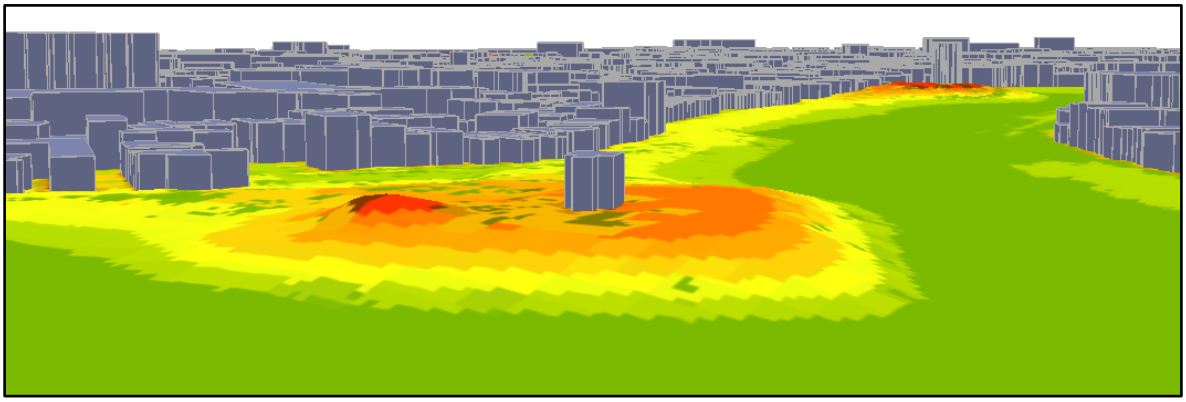
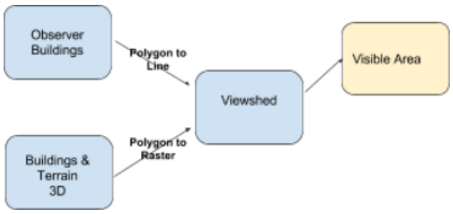


Figure 22: Digital Elevation Model (combination of building model and surface model).

4.2.4 Viewshed - Visible area

The viewshed tool requires observers and obstacles as input data. To make the analysis fit for mass valuation, only observers based on lines can be used as input data. The obstacle layer is defined by the surrounding area of the observers; the elevation model. The tool only accepts raster data, thus the 3D buildings are converted into a raster layer, the Digital Elevation Model (figure 23).



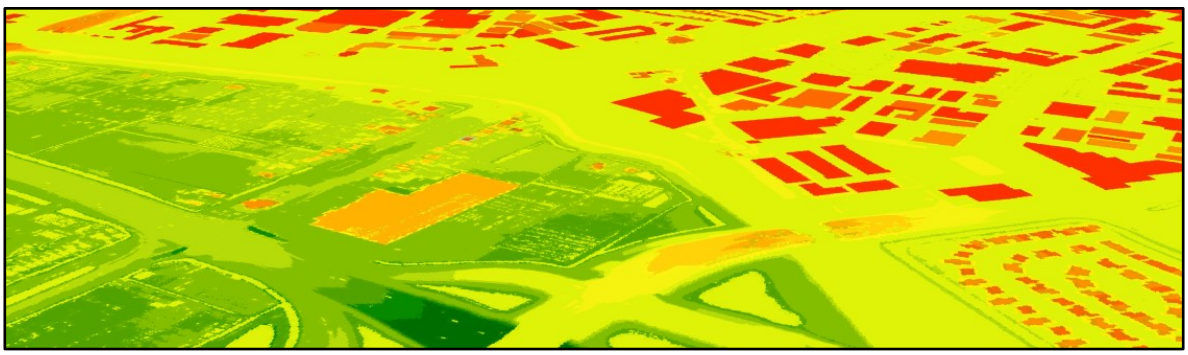


Figure 23: Raster representation of DEM (combination of building model and surface model).

For all observers, the view is calculated with a radius of 150 meter. This threshold is based on literature indicating that the for first 100 meters the variables form significant relationships, between 150-300 meters only a few relationships stay significant (Cavailhès et. al., 2009; Dekkers and Koomen, 2008). The observer offset is a fixed position of 1.8m above ground level, an approximation of height of adults (Paterson and Boyle, 2002), applied to Dutch circumstances.

Since the viewshed tool has the functionality to process only a few observers at the time, the tool is incorporated into a model that loops through each observer one by one. The model design is presented in appendix IV. While running the model, the observer table with information on each property is enriched with an extra field ‘viewshed volume’. The model will automatically fill this new field with square meter view values for each observer. Besides the enriched table, the model saves each id-based viewshed polygon in a geodatabase. The viewshed polygons can be used for testing, visualisation and further analysis. Both outputs are shown in figure 24.

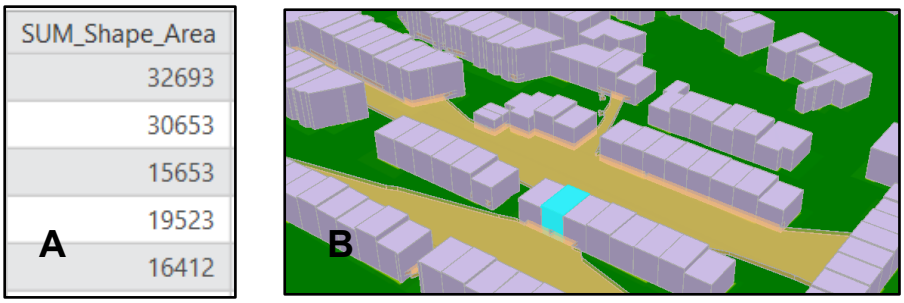
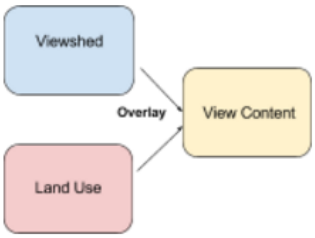


Figure 24: Automatically generated model outputs; A: Enriched data table; B: Unique viewshed polygons



However, when the viewshed polygons were evaluated for testing purposes, it became visible that at some locations the the elevation raster enhanced with the DSM gives biased outputs on surfaces when having minor elevation differences of 0.5 meter. Although the lower parts are not visible, they do not block the view, the observer is still able to overlook this area. The same property returns a viewshed output of 2316m2 when buildings and surface are included (figure 26), and a viewshed of 5336m2 when only the buildings are included (figure 25). Since Alkmaar is not subjected to major relief differences, the viewsheds for the Alkmaar clusters are calculated using only the buildings as obstacles.

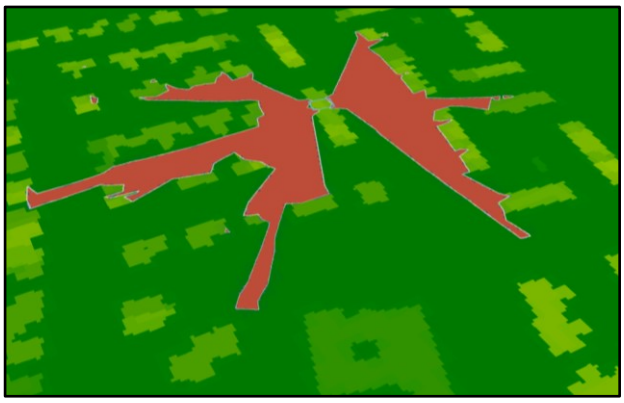


Figure 25: Digital Building Model: 5336m2 open view

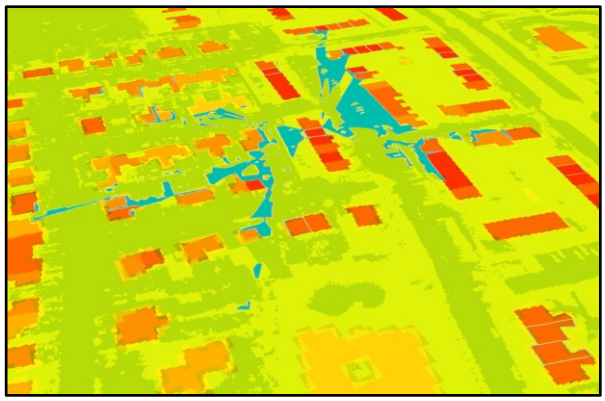


Figure 26: Digital Surface Model: 2316m2 open view

4.2.5 Viewshed - View content

The NVM data contains information about the land use types surrounding a property, making a distinction between water, forest, parks and traffic. The surrounding areas of a property can also be observed with GIS. In the Netherlands, the landuse dataset is provided as open data by the CBS. The dataset is detailed, making a distinction between 39 landuse types. The variables are recoded into 5 main classes; surface water, green, agriculture, developments and roads. An overlay with the generated viewshed polygon gives detailed information on landuse types in square meters (figure 27).

B

SUM_Shape_Area	Traffic	Developments	Water	Green
32693	693	26841	2900	2259
30653	375	25222	2812	2244
15653	367	11887	1166	2232
19523	168	17134	106	2115
16412	506	12985	813	2108

Figure 27: A: Intersection of landuse and viewshed; B: Automatically enriched data table.

4.2.6 Viewshed limitations

The generated model makes it possible to calculate viewsheds in an automated manner, entirely relying on open data sources. Although this is a huge step forward, there are still some limitations to the current methodology.

The main constrain is the limited possibilities for a 3D approach. The obstacle data is inserted in 3D, however the observers can only be extruded to a fixed height (2.5D), which makes it impossible to involve observer buildings on different floors or certain window points. Also the tool scans the area in a 3D space, including visibility of higher buildings that can be seen looking over another obstacle, though the output is given in 2D, lacking height information.

A limitation to the current model is the age of the datasets. The BAG building dataset is updated frequently, therefore the building footprints were up to date for this research. However, the elevation data coming from the AHN dates from 2011, therefore some new buildings could not be extruded. The latest version of the landuse BBG data was released in 2015, containing landuse information of the year 2012. Alternatively, other landuse registers, such as Top10NL could be used.

Furthermore, the model requires large computational time. For clusters 1 and 2 (656 and 634 properties) the model runned for respectively 10h03’ and 8h17’, since each raster output is stored individually. The computational time highly depends on the view radius settings, when testing the first cluster on a 1 kilometer radius, computational time took over a week.

4.3 Exploratory Data Analysis

In this section an overview of the model variables will be given by summarizing their statistics. Before starting a linear regression analysis, two preconditions should be met, which are the *quantitative data condition* and the *linearity condition.* If the conditions are explored, the clusters are ready for model specification.

4.3.1 General statistics

The predominant housing cluster in Alkmaar Municipality is the *5-room single family house built between 1971 and 1980*. The second largest housing cluster is *5-room single family houses built between 1981 and 1990*. This second cluster will be used in the comparative model. In the table below the general statistics of the two data clusters are shown alongside the statistics on the overall dataset.

	Cluster 1 *Build*	Cluster 2 *Build*	Total dataset
Properties	593	573	7762
Average value	€229401	€208481	€219654
Minimum value	€127416	€122937	€76023
Maximum value	€564424	€489182	€578947
Standard deviation	57800	44977	79401

Figure 28: Summary statistics of the dependent variable.

The maps visualising the distribution of property values (presented in appendix V), show a certain amount of spatial clustering. For the first cluster a hotspot with higher values can be found on the west side of the map. The second cluster shows less clear patterns, showing clustering in the South and North of the map.

The explanatory variables of both clusters have similar values for *inside maintenance, property type* and *size.* In the first cluster, more houses have inbuilt garages. In both clusters the view consists mainly of surrounding developments, agricultural land is in both cases limited. For the first cluster a good part of the view is covered with green, only a small part of the view is water. While in the second cluster water is more present than green (Appendix VI).

4.3.2 Quantitative data and linearity conditions

The dependent variable *'transaction value'* is determined in euro's, which is a ratio variable. When the dependent variable is of ratio level and explained by multiple independent variables, multiple linear regression analysis can be performed. MLR requires the independent variables to be of dichotomous (dummy), interval or ratio level. Variables that have other measurement levels, namely nominal or ordinal, should be recoded into dichotomous, interval or ratio variables.

The main independent variables for the identified clusters are considered to be housing type, maintenance, size and garage (Momentum Technologies, 2017). The initial basic model will test the explanatory power of those variables. To determine the influence of space, spatial variables are tested in a second model containing the NVM information on the property's surroundings and in a third model containing the objective GIS-based visibility information.

Some of the selected variables are of nominal and ordinal nature, making them unfit for MLR. Ordinal variables can be treated as interval data if they have distances between groups that are considered equal and meaningful (i.e. if a change from 1 to 2 is roughly equivalent to a change from 3 to 4). This is considered to be the case for the variables *housing type, garage* and *outside maintenance.* The variable *Fine location* obtains values that cannot be treated as interval data. Therefore, the unique values are recoded into dummies. The types and values for the independent variables are shown in figure 22, the detailed overview is shown in Appendix VII. For all variables it is possible to fit a linear regression line.

Independent Variable	Housing type	Inside Maintenance/ Outside Maintenance	Size	Garage	Fine location & Busy road	View (GIS)	View content (GIS)
Type	Interval	Interval	Ratio	Interval	Dummy	Ratio	Ratio
Values	1. Terraced 2. Linked 3. Corner 4. Semi-detached 5. Detached	1. Poor 2. Poor to moderate 3. Moderate 4. Moderate to adequate 5. Adequate to good 6. Adequate to good 7. Good 8. Good to excellent 9. Excellent	M2	0. Unknown 1. Not Inbuilt 2. Inbuilt	At forest At water At park Open view Busy road	M2 Open space	m2 Water M2 Roads M2 Developments M2 Green M2 Agriculture

Figure 29: Independent variable types and values

4.3.3 Model specification

The models will be specified in such a way that they are able to answer the two research questions: *To what extent will model accuracy improve when including basic spatial variables?* and *To what extent will model accuracy improve with the use of GIS?* For the first of these questions a non-spatial model and a spatial model are created (without GIS-based view). A third model will be tested containing the GIS-based view variables. To answer the second question, all models will be estimated with both the non-spatial OLS method and GIS- based spatial GWR methods.

When a model is estimated, a regression report is provided. This report gives information on the fit of the model by providing outcomes on multiple performance tests. The R-square gives the overall model fit, indicating the amount of variance that is explained by the model. Furthermore information on the error terms, or residuals, related to the regression line is gathered. With this information model assumptions can be verified by diagnostic tests. The three assumptions to be tested are:

- Normality of Error Distribution
- Independence of Errors (no spatial autocorrelation)
- Homogeneity (global model)

In case the assumptions are met, the OLS model suffices. When the assumptions are violated, re-specification of the functional form of the model is required. The best performing functional form for linear property valuation is considered to be Geographically Weighted Regression, since this method deals with both spatial autocorrelation and spatial heterogeneity. In the next chapter the results of the model estimations are provided.

1. Non-spatial OLS model (NVM variables)
2. Spatial non-GIS OLS model (NVM variables)
3. Spatial GIS-based OLS model (NVM + viewshed variables)
4. Geographically Weighted Regression model (NVM + viewshed variables)

The best fitting OLS regression model is the basis for the prediction model. For each cluster, 10% of the data is kept apart to validate whether the prediction model functions properly. The validation will be performed by multiplying the parameters with the values of the independent variables. The GWR model is tested within the GIS environment, using a prediction tool.

5. Results

In this section the results of the regression models are presented. For both clusters the four models; non-spatial, spatial non-GIS, spatial GIS and GWR, are compared. The first part gives a general overview of the outcomes. Subsequently the results of the models are explained per cluster. A conclusion on the results is presented in the last part.

5.1 Results overview

In the table below the overall results of the model performance tests for both clusters are presented. In both cases the GWR model turned out to be the strongest prediction model for property valuation.

Models	Model results			
Cluster 1	R2*	AIC**	Moran's I***	Error
OLS non-spatial	0.5077	14274		
OLS spatial non-GIS	0.5354	14241		
OLS spatial GIS	0.5617	14210	0.248	12.5%
GWR	0.5553	14076	0.165	10.9%
Cluster 2	R2	AIC	Moran's I	
OLS non-spatial	0.6680	13280		
OLS spatial GIS	0.6757	13268	0.341	18.6%
GWR	0.7027	13219	0.018	18.5%

Figure 30: Regression model output.

*R2: To compare global models, the adjusted R-square is consulted. In comparison to the normal R-square, this value takes into account the complexity of the model when extra variables are added.

**AIC:To compare local and global models, Akaike's Information Criterion is consulted. This value measures the relative quality of a model, taking into account the change in degrees of freedom when working with local models. A decrease of at least 3 points signifies a better model fit.

***Moran's I: The Moran's I value calculates the amount of spatial autocorrelation on a scale from 0 to 1.

In the table presented on the next page the variables of the best fitting OLS model are shown, which is in both clusters the spatial GIS model. Since for GWR the coefficients are calculated locally for each specific property, no global coefficients can be presented in the table. The interpolation raster of the coefficients shows the regional patterns of each variable in appendix VII. For the second cluster, all variables were used in the GWR model, while for the first cluster 2 variables *inside maintenance* and *garage* could not be used in the GWR model due to multicollinearity issues. Multicollinearity is the result of redundancy or limited values per variable class.

	Cluster 1			Cluster 2		
Variabele	OLS spatial GIS Coefficient	Price impact	In GWR	OLS spatial GIS Coefficient	Price impact	In GWR
Intercept	-76442 (p=0.000)			14206.9 (p=0.245)		
M2	1515.95 (p=0.000)	63.9%	yes	1193.42 (p=0.000)	71.9%	yes
Inside Maintenance	14537 (p=0.000)	22.7%	no	4941.92 (p=0.008)	12.7%	yes
Property Type	14991.1 (p=0.000)	8.3%	yes	12768.5 (p=0.000)	12.8%	yes
Garage	16220.5 (p=0.000)	2.3%	no	35856.6 (p=0.000)	2.8%	yes
Water	4.1441 (p=0.000)	0.2%	yes			
Roads	-4.23285 (p=0.000)	-0.2%	yes			
Developments	0.770282 (p=0.017)	2.7%	yes			
Agriculture	2.1173 (p=0.020)	0.1%	yes			
Green				0.91711 (p=0.000)	0.7%	yes

Figure 30: Coefficients and price impact of best fitting OLS models.

The coefficients indicate whether the relationship between the explanatory variable and property value is positive or negative. The associated p-value shows the significance of the variable. If the p-value is smaller than 0.05, it is statistically significant that the coefficient of the variable is most likely not 0, and thus a meaningful explaining variable of the model. The coefficient value indicates the strength of increase or decrease of the property price for one point increase or decrease of the variable value. Since the units are different for each variable, the relative price impact is calculated for each variable.

Price impact is calculated by taking the mean value of each variable (as presented in appendix VI) multiplied with the coefficient. This gives the absolute price impact of the variable. To make the impact comparable between both clusters the relative price impact is divided by the average property price of its cluster minus the constant. This gives the relative price impact of each variable. The values show that square footage is as expected the main explaining variable for property prices. In cluster 1, garage turned out to be an even smaller indicator of property prices than the view on developments surrounding a property. This could be related to the probable positive impact of having facilities in developed areas.

5.2 Model results cluster 1

The regression results of all four models of the first cluster are discussed below. The regression outputs are presented in appendix VII. The prediction outputs are presented in appendix VIII.

5.2.1 The non-spatial model

The first predicted model is the spaceless model without any spatial variables. For proper model specification, an OLS model is prepared including all variables to test their significance. The explanatory variables for the initial model are *house type, size, inside maintenance and garage. Outside maintenance* turned out to be insignificant and was therefore excluded from the model.

The R-squared fit of the initial model is 0.5077, indicating that a good 50% of the variance in the residuals is explained by the model. All variables have statistically significant positive relationships with the property value. The strongest indicator of the property value is the size of the property in square meters. An increase of 1 square meter of usable area will increase the property value with an average of €1688.-.

5.2.2 The spatial non-GIS model

In the spatial non-GIS model the geographical information provided by NVM is included. The dataset contains information on the presence of open space, water, parks and roads. Since open space can consist of the latter three elements, those variables are redundant and should be tested in separate models to prevent a biased model.

The presence of open space is first tested as additional information on the initial model. The information is binary, only showing properties that are or are not surrounded by open space. The variable influences the value of a property in a positive manner. However, the relationship is not statistically significant, due to the high standard deviation of the error term. The fit of the model does not increase when adding this variable.

When evaluating the remaining three variables, water, parks and heavy traffic, the model shows that presence of water is statistically significant. Whether water is present or not, impacts the value of a property on average with 15.5%. Including this variable in the model increases the R-squared fit of the model up to 54%.

5.2.3 The GIS-based spatial model

The third OLS model reviews the influence of the spatial information collected using GIS-techniques in an automated manner. The values collected are more specific than the binary ones, showing the amount of square meter visible area around a property and the content of this view. The view's content consists of water, green, roads, developments and agriculture. This model shows that view, similar to open space, is not a significant explanatory variable for changes in property values. The variable shows again a high standard deviation.

When testing the view content variables, apart from the view on green, all variables are statistically significant. In this model water is again significant, showing an increase of €4.11 per additional square meter. View on roads is negatively influencing the property value, decreasing the value with €4.23 per square meter. Developments and agriculture show a positive significant relationship. Based on the calculated price impact, view on other developments is the strongest explaining variable of the view content, influencing on average 2.7% of the property price. The fit of this model is 56.17.

Since this OLS model performs, the prediction power of this model is tested. The following prediction equation is based on the coefficients determined with the build set:

$$\hat{Y}_i = -76442.0949 + X_{1i} * 1515.953295 + X_{2i} * 14536.95455 + X_{3i} * 14991.068667 + X_{4i} * -4.232846 + X_{5i} * 0.770282 + X_{6i} * 2.117304 + X_{7i} * 4.144103 + X_{8i} * 16220.463297$$

in which:

X_1 = Size of the property in M2
X_2 = Maintenance level of the property
X_3 = Property Type
X_4 = View on roads in M2
X_5 = View on developed area in M2
X_6 = View on agricultural land in M2
X_7 = View on water in M2
X_8 = Type of garage

This model is tested on the 10% test set. Two examples below demonstrate the prediction process. The predicted property values are compared with the known transaction values, showing a average deviation in prediction of €29051, leaving 12.5% of the property value unpredicted (Appendix VIII).

Example 1:
A 145m2 good maintained terraced house without garage and a view of 5655m2 road, 12901m2 development, and 957m2 water:

*-76442.0949 + 145 * 1515.953295 + 5 * 14536.95455 + 1 *14991.068667 + 5655 * -4.232846 + 12901 * 0.770282 + 0 * 2.117304 + 957 * 4.144103 + 0 * 16220.463297 = 221014*

The predicted property value is €221014.- The actual property value was €215428.-, giving a prediction residual of €5586.-

Example 2:
A 130m2 good maintained detached house with external garage and a view of 1371m2 road, 16493m2 development, and 4823m2 water:

-76442.0949 + 130 1515.953295 + 5 * 14536.95455 + 5 * 14991.068667 + 1371 * -4.232846 + 16493 * 0.770282 + 0 * 2.117304 + 4823 * 4.144103 + 1 * 16220.463297 = 311383*

The predicted property value is €311383.- The actual property value was €280196.-, giving a prediction residual of €31187.-

Results of the model tests are presented in appendix VII. The significance of the Jarque-Bera test indicates that the errors are biased for not being normally distributed, which is shown to some extent by the plot of the residuals against the predicted values (figure 30). The Moran's Index of 0.248 confirms that the residuals are subjected to global spatial autocorrelation. The significance of the Koenker test indicates inconsistency amongst the modelled relationships, due to spatial heterogeneity. Since the assumptions of spatial stationarity and independence are violated, the GWR model is required.

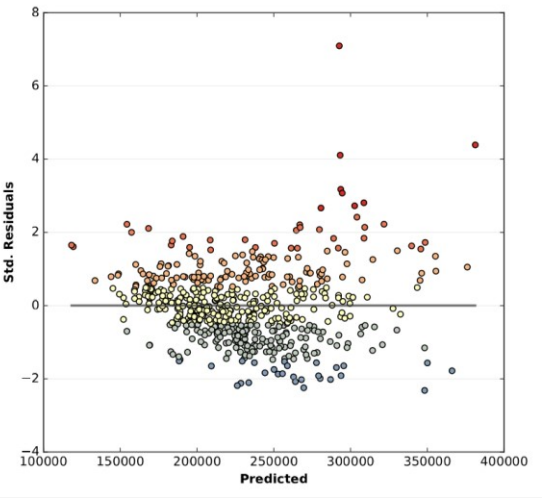


Figure 31: Plot of residuals against predicted values.

5.2.4 Geographically Weighted Regression model

The specified OLS model with computed spatial variables was found to be significant with a good R-squared value of 0.5553 (p=0.000), but violated the essential assumptions of OLS modelling. The Geographically Weighted Regression model solves the spatial issues by taking into account geographical distance to the neighbors to correct for spatial autocorrelation. However, this method is, due to local selections of properties, sensitive to multicollinearity (redundancy) amongst the explanatory variables.

The proposed OLS model could not entirely be used in GWR, because of local multicollinearity. Therefore the variables *garage* and *inside maintenance* were excluded from the model. Lacking the two extra variables, GWR still performs a better model showing an increased adjusted R-square value of 0.6553 and a decreased AICc of 14076. The bandwidth was fixed at the size of a 1256 meter kernel.

Out of all OLS models, the GIS-based spatial model turns out to perform best (R2=0.5617; AIC=14210). The following functional form of the prediction model is based on the coefficients determined in the build set;

$$\hat{Y}_i = \beta_0(i) + \beta_1(i) * X_{1i} + \beta_2(i) * X_{2i} + \beta_3(i) * X_{3i} + \beta_4(i) * X_{4i} + \beta_5(i) * X_{5i}$$

in which:

X_1 = Size of the property in M2
X_2 = Property Type
X_3 = View on roads in M2
X_4 = View on agricultural land in M2
X_5 = View on water in M2

Since the parameters depend on the location of the property, a general model cannot be performed. The model can only be predicted within a GIS environment that creates a best fitting model for each property. The distribution of parameter values over space is presented in appendix VII. This model is tested on the 10% test dataset. The predicted property values are compared with the known transaction values, showing a average error in prediction of €25391, an improvement of €3660 with the OLS model (Appendix VIII). This indicates that the

GWR model leaves only 10.9% of the price unpredicted, compared to 12.5% of the OLS model.

Although the Moran's I decreased to 0.165, there is still evidence of spatial autocorrelation between the residuals. A plot of the prediction errors shows spatial clustering (Appendix VIII). This means that the model is misspecified, lacking essential information of explanatory variables. This can be related to the variables that were specified by OLS but had to be removed for GWR, which were *garage* and *maintenance*, since these variables showed multicollinearity.

5.3 Model results cluster 2

The regression results of the second cluster are discussed in less detail. This cluster is used as a comparative study to review deviant results from the first cluster. The regression outputs are presented in appendix VII. The prediction outputs are presented in appendix VIII.

5.3.1 The OLS models

To get an overview of the relationship that the variables have with property value in this cluster several models are generated to explore the best model fit. The non-spatial model already shows a very good fit with an adjusted R-square of 0.6703, compared to 0.5077 in the first cluster.

Interesting in this cluster is that the adding the view value to the non-spatial variables is in this case significant, increasing the adjusted R-square slightly up to 0.6749. The best fitting model, with an adjusted R-square of 0.6757 is the GIS-based spatial model including the variables *M2, maintenance, property type, garage, green.* All explanatory variables are statistically significant, correlating positively with the value of the property. The AICc value of this model is 13265. The prediction model of this cluster is specified as follows:

$$\hat{Y}_i = 14206.945267 + X_{1i} * 1193.423498 + X_{2i} * 4941.923349\ X_{3i} * 12768.536702 + X_{4i} * 35856.615071 + X_{5i} * 0.91711$$

in which:

X_1 = Size of the property in M2
X_2 = Maintenance level of the property
X_3 = Property Type
X_4 = Type of garage
X_5 = View on green in M2

The prediction model is validated by the test dataset. The average deviation between the predicted value and the actual property value is €37747, leaving 18,6% of the property value unpredicted (appendix VIII).

The regression tests are presented in appendix VII. The Jarque-Bera test indicates that the errors are biased indicating that they are not normally distributed over space. A plot of the predicted values against the prediction errors shows a clustered structure (figure 32). The Moran's Index of 0.341 confirms that the residuals are subjected to global spatial autocorrelation. The significance of the Koenker test indicates inconsistency amongst the modelled relationships, due to spatial heterogeneity. The outputs of the tests are similar to the first cluster.

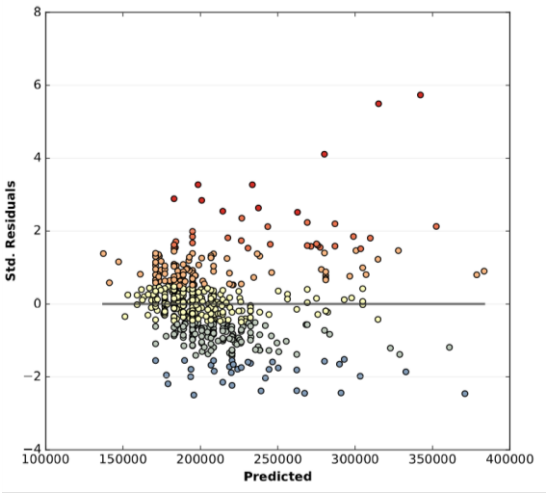


Figure 32: Plot of residuals against predicted values

5.3.2 Geographically Weighted Regression model

In contrary to the OLS of the first cluster, the OLS of this cluster can directly be tested in a GWR model, since no multicollinearity arises. The functional form of the equation is as follows:

$$\hat{Y}_i = \beta_0(i) + \beta_1(i) * X_{1i} + \beta_2(i) * X_{2i} + \beta_3(i) * X_{3i} + \beta_4(i) * X_{4i} + \beta_5(i) * X_{5i}$$

in which:

X_1 = Size of the property in M2
X_2 = Maintenance level of the property
X_3 = Property Type
X_4 = Type of garage
X_5 = View on green in M2

The parameter weight rasters are presented in appendix VII. The GWR shows a reduced AICc value of 13219 and the increased adjusted R-square of 0.7027, confirming a slightly improved performance of the model. The Moran's Index of 0.018 indicates that in this case the spatial autocorrelation is solved, confirming the strength of GWR modelling. The average deviation between the predicted value and the actual property value is slightly decreased to €37516, leaving 18,5% unpredicted. This high amount of error indicates that there are still essential variables missing in the model.

5.4 Regression results

The first cluster, building year 1971 - 1980 with properties more towards the centre, shows that predicting property prices by means of a spatial workflow does clearly generate improved outputs of the model. Adding GIS-based spatial variables in the model raised the initial 51% model fit to 56%. The strongest geographical relationships were found to be the positively related view on water and the negatively related view on roads. Significance was also found for the view on agriculture and developed area. The local GWR regression highly increased the model fit (65%), indicating that the dataset is was prone to strong spatial errors within the explanatory relationships influencing property value.

In the second cluster, building year 1981-1990 more at the outskirts of the city, the spatial workflow is only slightly improving the predictive power of the model. Including spatial variables on the view of the property increased the fit of the model with 0.05%. This suggests that the structural variables are already great explanators of property values. In contrary to the first cluster, the variable view on green was found to be a significant explanatory variable of changes between property values in this dataset. The local regression increased the model fit from 67% to 70%. Indicating that small spatial variations in explaining the property value were accounted for.

The variable view was in the first cluster not a direct explanator for the property value, due to the multiple significant variables related to the content of the view. Since certain landuse types correlated positive while others negative, the general variable view was not a good explanatory factor. For the second cluster, where spatial variables had a less strong effect on the overall fit of the model, the variable view had a significant correlation with property value, possibly due to the fact that the content of the view was significantly explained by only one variable.

Overall, the physical property variables have a higher impact on the property values than the tested locational variables. Though especially for the first cluster, the impact of the locational variables can not pass unremarked. Of all view variables, view on developments shows the highest impact in the first cluster, influencing the property price even more than the variable garage. Since developed areas can contain all kinds of amenities, further research on this variable is required.

Although the second cluster shows a better model based on the adjusted R-square and AIC values, the predictive power of this model is not impressive. The first model shows better predictions while having a weaker model fit. It is possible that for the second cluster the 10% test dataset is not entirely representing the build dataset. However, a closer look at the prediction errors in appendix VIII shows a constant overprediction, indicating that the problem can be found in the model design.

6. Conclusion

The primary objective of this research was to use an explicit spatial methodology in conjunction with a basic spatial regression model to test the significance of geographic variables on residential property prices.

What are the needs for current property valuation models?

Reviewing current property valuation models pointed out that spatial variables are underrepresented in the database. For assessed valuations only structural variables are required. Since assessed valuations are calculated annually in The Netherlands, only objective information is suitable for the automated prediction model. The database of the The Netherlands Organization of Real Estate Brokers indicates that a few variables on the property's surroundings are taken into account for taxation purposes, though they are collected in a subjective manner, which makes them unsuited for automated valuation. A quantification of the spatial variables would enrich the database for assessed valuations and improve efficiency of data collection for realtors.

To what extent will model accuracy improve when including basic spatial variables?

The subjective information that NVM includes in the property models is the presence of certain landuse types around a property. To measure the influence of those variables in a regression model, they should be recoded into binary variables indicating presence or absence of a certain landuse type. The regression model exhibited that including a dummy on the presence of water around a property is significantly improving the fit of the prediction model. The dummies *parks, heavy traffic* and *open space* do not show significant influence on the property value. Since information on the surrounding of the properties is often collected in the field, including presence of water in an automated model is considered highly inefficient.

How can subjective spatial information be quantified for automated valuation?

With the use of Geographical Information Systems and open data this could be quantified in an automated manner. Basic spatial analysis on the property's surroundings would concern buffer and distance calculations. A more advanced method is to take into account the landuse types that actually visible around the house, assuming that neighboring properties could differ vastly in value because one has an open view on attractive landuse types while the other has a blocked view by other properties.

The proposed automated model in this research makes it possible to calculate the view in square meters of each property at a fixed radius. An intersection with a landuse layer calculates the visible square meters of each landuse type. The model is set up in such a way that inserting a table with the property's location will automatically update the table with the desired information.

To what extent will model accuracy improve with the use of GIS?

The output of the viewshed model is, because of its quantitative nature, a suitable input to test in a regression model. Comparing two clusters with properties situated in different locations, the first is closer to the city centre, the second more towards the outskirts of the city, shows different outputs on the influence of spatial variables. In the first cluster the the views of a property are strongly improving the model fit. Especially water and roads have a major influence in explaining the property price. Also agriculture and developments show significant relationships. Since the variables impact the property price in different directions, the umbrella variable open space is not a significant explanatory variable. In the second cluster, views of a property have a minor impact on the value of the property. In this cluster open space is a significant factor in explaining the property value, this is related to its content that is only explained by one variable.

The case studies on the two cluster groups in Alkmaar both demonstrate that an explicit spatial approach when explaining property values, is significantly improving the predictive power of the models, proving that space should not be omitted in current property valuation models.

7. Recommendations

The outcomes of this study provide a base for further research. The proposed spatial workflow has proven to work, as tested on two property clusters within the city of Alkmaar. Within the two clusters, already major differences are found in the explanatory power of the spatial variables. Further research on other clusters and similar clusters in different cities will provide a meaningful extension to the body of knowledge on property valuation.

In the analysis of viewshed a higher level of detail could be considered by introducing more landuse subgroups to review the distinction between types of green, water, agriculture. Also the view should be placed in its context, how does the influence of view on certain landuse types differ from the proximity of these landuse, which thus includes non-visible adjacency.

At the moment the tools for a viewshed analysis that work with 3D outputs, are only suitable for point features. Since properties block their observers when inserted as points, the observers are converted into the line features of their building. When working with line features only a limited set of tools is possible. This study emphasized the need for extended viewshed tools that work with both line features and 3D outputs.

Further developed advanced tools and techniques, that allow automated point viewshed calculations, have the potential to demonstrate the view from specific observer locations in a building, such as the floor or window an observer is situated at. This is however expected to increase the computational time of the viewshed model. To explore possible efficiency improvements of the model, a compact rewrite of the python functions related to the geoprocessing tools should be considered.

Due to the increase in multilevel buildings that house several apartments, 3D viewshed analysis in an automated manner is a main focus for further research. Companies such as NVM consult the cadastre for information on boundaries and ownership. To facilitate this practice of 3D analysis, cadastral base registers including 3D geometries are a requisite, since ownership can then be related to height attributes.

References

Literature

Anselin, L. (1988) Lagrange Multiplier test diagnostics for spatial dependence and spatial heterogeneity, *Geographical Analysis,* 20, 1-17.

Anselin, L. (1998) GIS Research Infrastructure for Spatial Analysis of Real Estate Markets. *Journal of Housing Research:* 1998, Vol. 9, No. 1, pp. 113-133.

Anselin, L. (2007) Spatial Regression Analysis in R A Workbook, *Center for Spatially Integrated Social Science.*

Anselin, L. (2010) Thirty years of spatial econometrics. *Papers in Regional Science,* 89, 3-25.

Anselin, L. (2013) Spatial Econometrics: Methods and Models, *Studies in Operational Regional Science*, Springer.

Bitter, C., Mulligan, G., Dall'erba, S. (2006) Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method, *Munich Personal RePEc Archive,* no. 1379, 9.

Borst, R. A. and McCluskey, W. J. (2008) The Modified Comparable Sales Method as the Basis for a Property Tax Valuations System and its Relationship and Comparison to Spatially Autoregressive Valuation Models, *in Mass Appraisal Methods: An International Perspective for Property Valuers (eds T. Kauko and M. d'Amato)*, Wiley-Blackwell, Oxford, UK. doi: 10.1002/9781444301021.ch3.

Bouwmeester, H.J.F.M., Drentje A.C., Vries, P. de, (2011) Inventarisatie Modelmatige Waardebepaling in Nederland*, Research Institute OTB*, Delft.

Bidanset, P. E., and Lombard, J. R. (2014) Evaluating spatial model accuracy in mass real estate appraisal: A comparison of geographically weighted regression and the spatial lag model. *Cityscape,* 16(3).

Brunsdon, C., Fotheringham, S.A., Charlton, M.E. (1996) Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis,* 28, no. 4.

Cavailhès, J., Brossard, T., Foltête, J.-C., Hilal, M., Joly, D., Tourneux, F.P. (2009) GIS-based hedonic pricing of landscape. *Environ. Resour. Econ. 44,* 571–590

CBS (2016), Voorraad woningen en niet-woningen, *Statline Dutch Statistics.*

CBS (2016b) Bestaande koopwoningen; woningtype; verkoopprijzen, *Statline Dutch Statistics.*

Esri (2017) ArcGIS for desktop, *Tools*.

Dekkers, J.E.C., Koomen, E. (2008) Valuation of open space, Hedonic house price analyses in the Dutch Randstad region, *Research Memorandum, no. 2008-24*, Faculteit der Economische Wetenschappen en Bedrijfskunde, Amsterdam.

Dubin, R., Pace, K., & Thibodeau, T. (1999). Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature,* 7(1), 79-95. ISO 690.

EBZ (2015) Dynamiek in de steden, stilte aan de randen van het land De regionale huizenmarkten tot 2025, ING Economisch Bureau.

Fotheringham, A.S., Brunsdon, C., Charlton, M.E. (2002) Geographically weighted summary statistics - a framework for localised exploratory data analysis, *Computers, Environment and Urban Systems,* 26, 501–524.

Fotheringham, A.S., Charlton, M.E. (2009) Geographically Weighted Regression - White Paper.

Francke, M. (2010). Casametrie: de kunst van het modelleren en het voorspellen van de marktwaarde van woningen, Vol. 353. *Amsterdam University Press.*

Huisman, O., de By, R.A. (2009) Principles of Geographic Information Systems, *ITC,* Netherlands: Enschede.

Howes, D., Gatrell, A. (1993). Visibility analysis in GIS: Issues in the environmental impact assessment of windfarm developments. *Proceeding of the Fourth European Conference and Exhibition on Geographical Information Systems,* Italy: Genoa.

Irwin, E.G. (2002) The effects of open space on residential property values, *Land Economics*: 465-480.

Isikdag, U., Horhammer, M., Zlatanova, S., Kathmann, R., Oosterom, P. van (2015) Utilizing 3D Building and 3D Cadastre Geometries for Better Valuation of Existing Real Estate, *FIG Working Week 2015,* Bulgaria: Sofia.

Lake, R., Lovett, A.A., Bateman, I.J., Langford, I.H. (1998) Modelling Environmental Influences on Property Prices in an Urban Environment, *Computers, Environment and Urban Systems,* 22, 121-36.

Luttik, J., (2000) The value of trees, water and open space as reflected by house prices in Netherlands. *Landscape and urban Planning,* 48, 161-167.

McLeod, P.B., (1985) The Demand for Housing and Amenity Attributes: An empirical analysis of the Perth housing market, *Dept. of Economics, University of Western Australia,* W.A: Nedlands.

NMA (2012) Marktscan Woningmakelaardij, *Nederlandse Mededingingsautoriteit.*

NVM (2016) Analyse Woningmarkt van de bestaande koopwoningen 1e kwartaal 2016, *De Nederlandse Vereniging van Makelaars en Taxateurs.*

Op 't Veld, D., Bijlsma, E. and van de Hoef, P. (2008) Automated Valuation in the Dutch Housing Market: The web-application 'MarktPositie' used by NVM-realtors, *In: Kauko, T. and d'Amato, M. (Eds.), Advances in Mass Appraisal Methods.* Blackwell, Oxford, pp. 70-90.

Pace, R. K., Barry, R., & Sirmans, C. F. (1998). Spatial statistics and real estate. *The Journal of Real Estate Finance and Economics,* 17(1), 5-13. ISO 690.

Paterson, R.W., Boyle, K.J. (2002) Out of Sight, Out of Mind? Using GIS to Incorporate Visibility in Hedonic Property, Value Models, *Land Economics, 78-3,* 417-425.

Quigley, J. M. (1979) What Have We Learned About Urban Housing Markets? *In: Mieszkowski, P., Straszheim, M., Current issues in urban economics.* Baltimore, Johns Hopkins Univ. Pr., p. 391-429.

Ridker, R.G. & Henning, J.A. (1967) Determinants of Residential Property Values with Special Reference to Air Pollution, *Review of Economics and Statistics,* 49, 246-257*.*

Rodriguez, M., Sirmans, C., & Marks, A. (1995). Using geographic information systems to improve real estate analysis. *Journal of Real Estate Research*, ISO 690.

Tobler, W. R. (1970) A computer movie simulating urban growth in the Detroit region. *Economic geography, 46(1),* pp. 234-240.

Tomic, H., Roic, M., Ivic, S.M. (2012) Use of 3D Cadastral Data for Real Estate Mass Valuation in the Urban Areas, *3rd International Workshop on 3D Cadastres*, China: Shenzhen.

Waarderingskamer (2016) Jaarverslag 2015 Waarderingskamer, *Council for Real Estate Assessment.*

Welle-Donker, F., van Loenen, B., Bregt, A.K. (2016) Open data and Beyond *ISPRS International Journal of Geo-Information,* 5, 48.

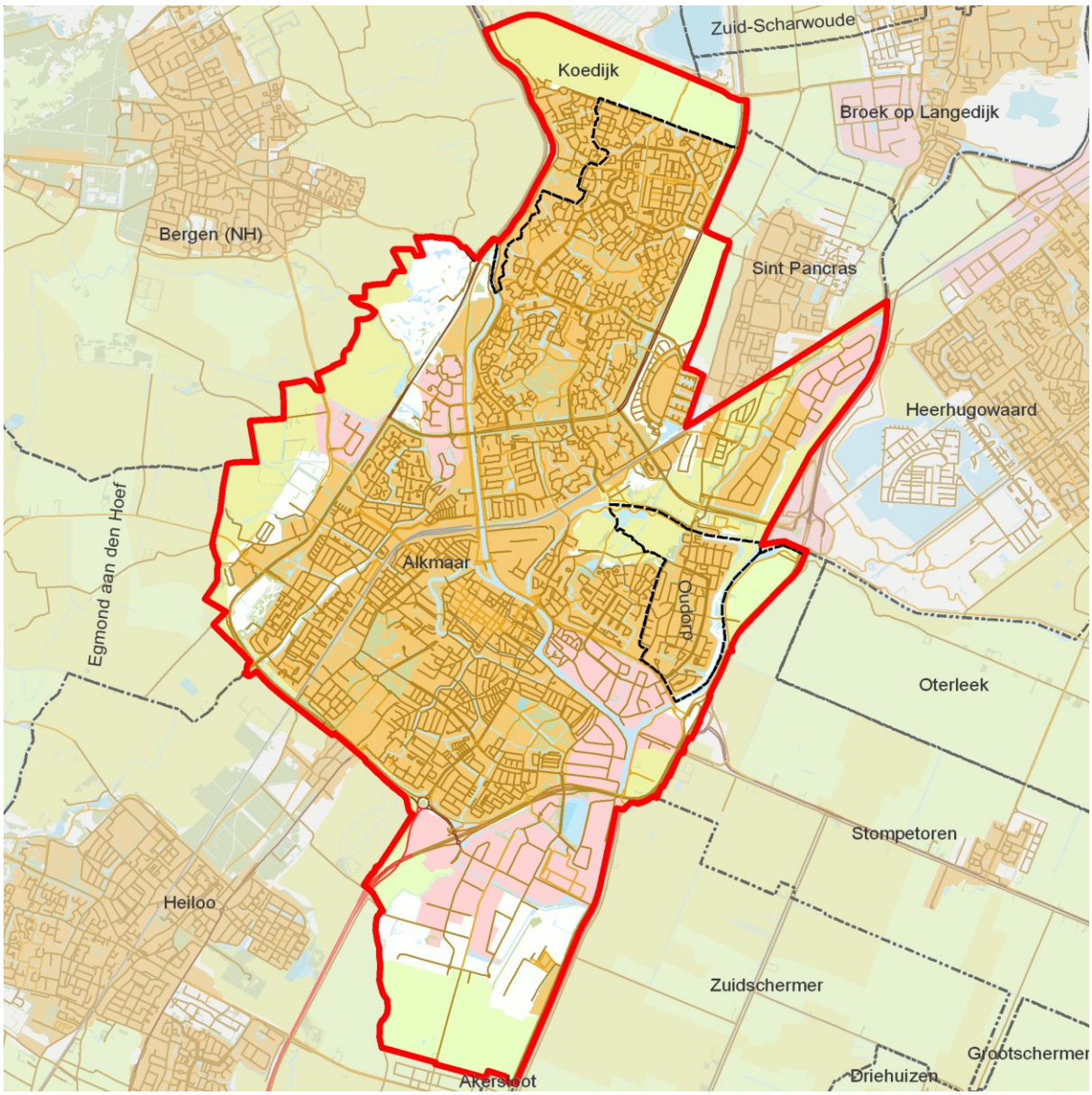
Yu, S.M., Han, S.S., Chai, C.H. (2007) Modelling the Value of View in Real Estate Valuation: A 3-D GIS Approach, *Environment and Planning B: Planning and Design,* Vol. 34, No. 1, 139–153.

Interviews and conversations

Momentum Technologies, Dree Op 't Veld (2017) *Theme*: *Property Valuation Modelling.*
Esri helpdesk (2017) *Theme*: *Viewshed Analysis.*

Appendices

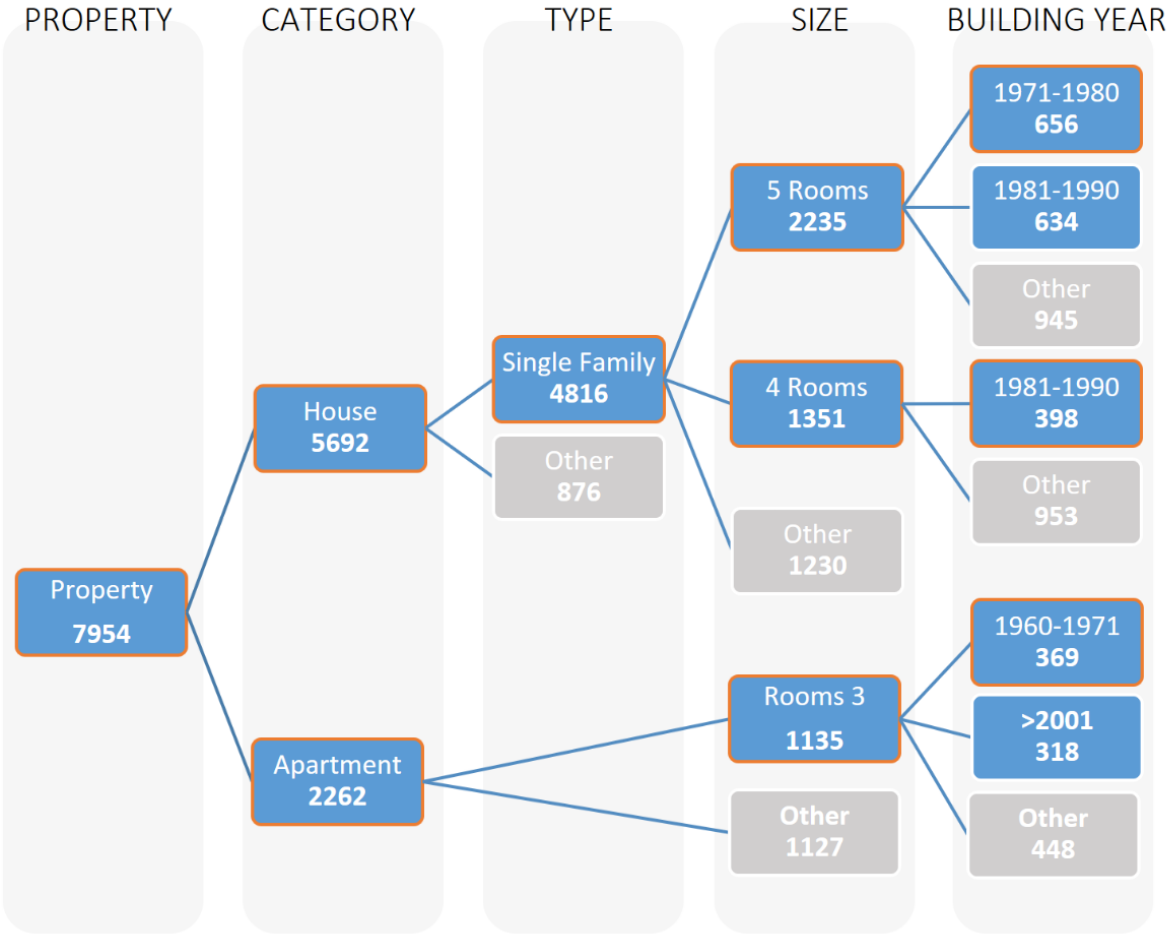
Appendix I - Study Area



(Source: Jan-Willem van Aalst based on Openstreetmap, 2011)

Alkmaar Municipality included up until 2015 the city Alkmaar and the villages Oudorp and Koedijk.

Appendix II - Clustering

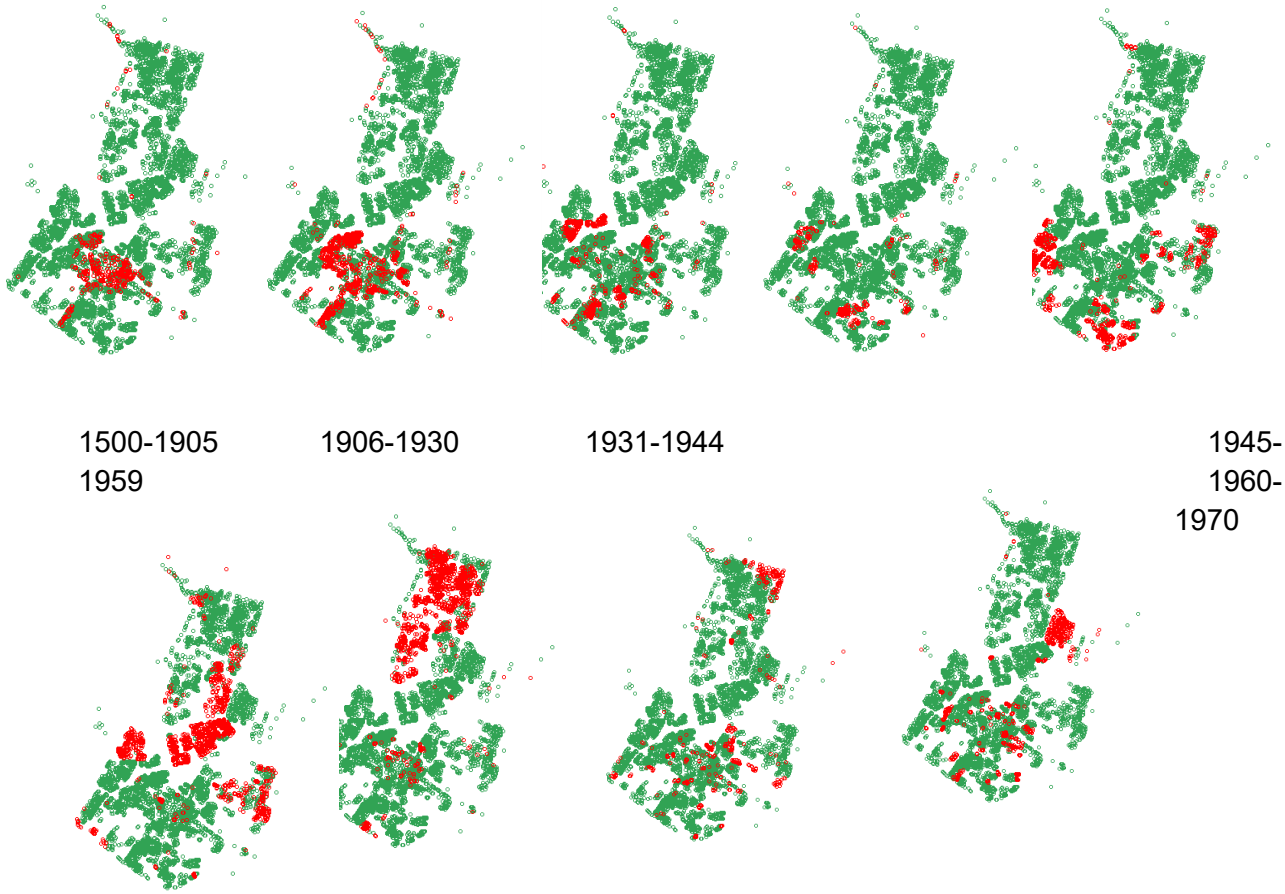


(Source: author's own)

Taking into account the spread of values of the dominant variables, the sequence of decision making is determined. The decision tree shows what paths are followed to come to the clusters. Variable classes are only taken into account if they represent more than 25% of their total. For instance the 5 and 4 rooms variable classes are taken into account within Single Family houses since they each represents more than 25% of the single family houses. The other Room classes are left out since none of them individually represents 25% or more of the Single Family houses.

Since Apartments will not be taken into account because of technical constraints, the two clusters are selected within the Single Family houses group.

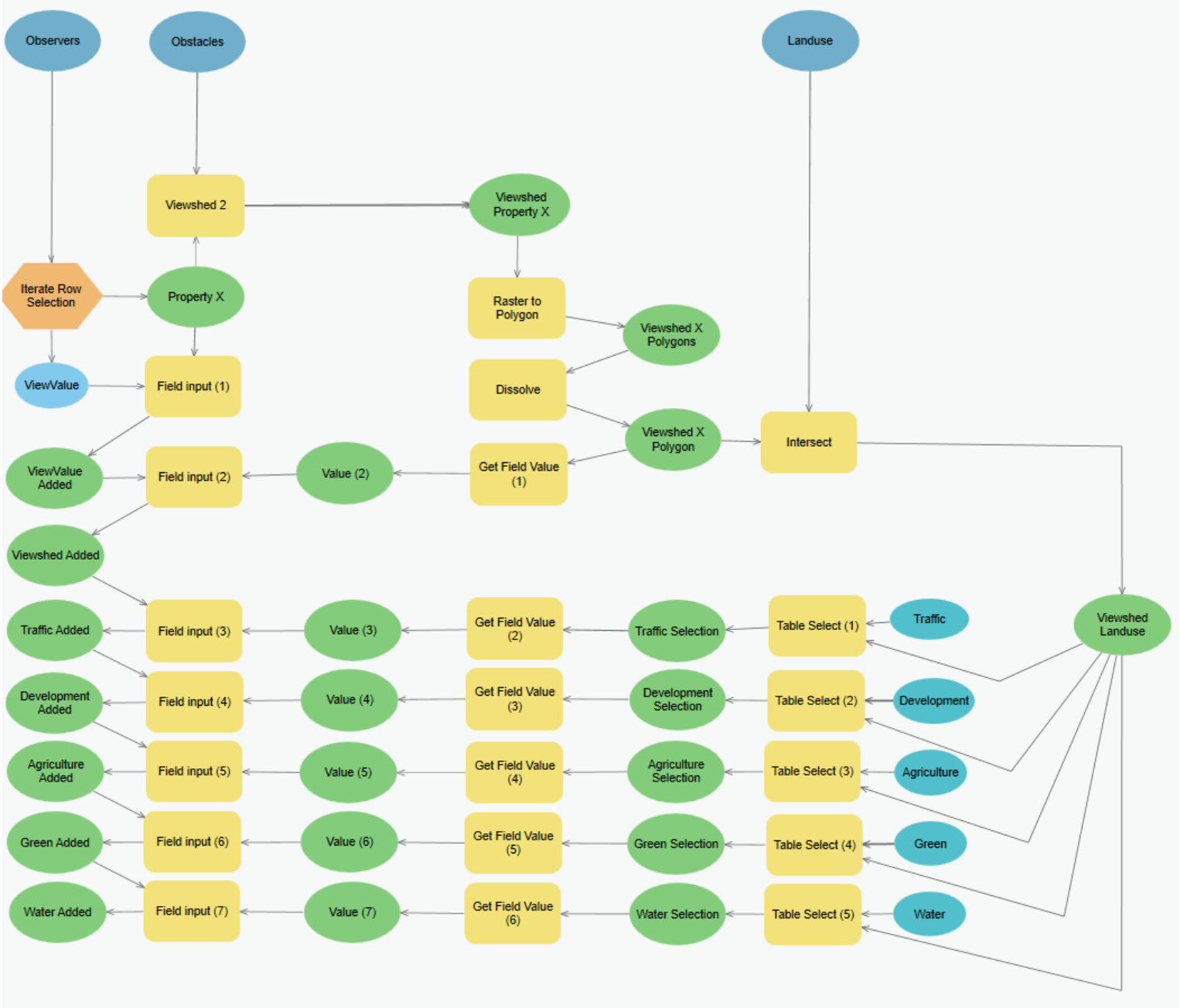
Appendix III - Spatial distribution of building year



(Source: author's own in GeoDa)

Testing with different building year clusters is related to testing in different geographical environments. The 5-room single-family properties constructed in 1971-1980 and 1981-1990 are used as clusters for this study.

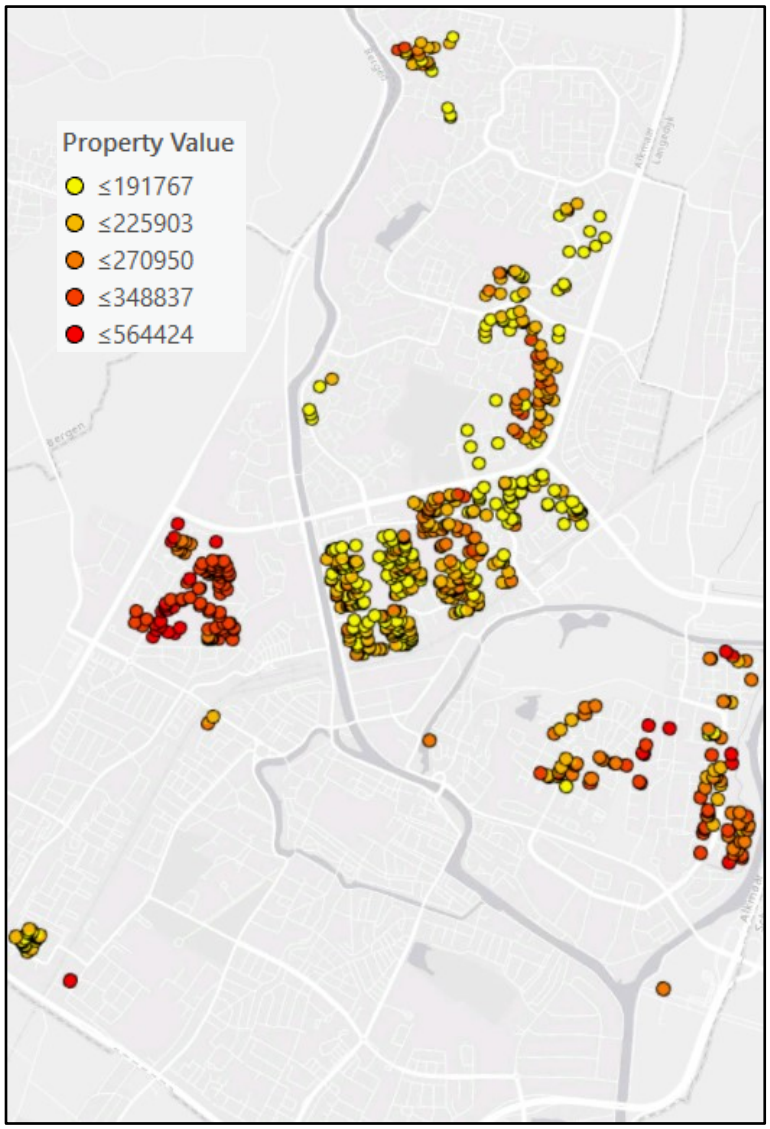
Appendix IV - Viewshed Model



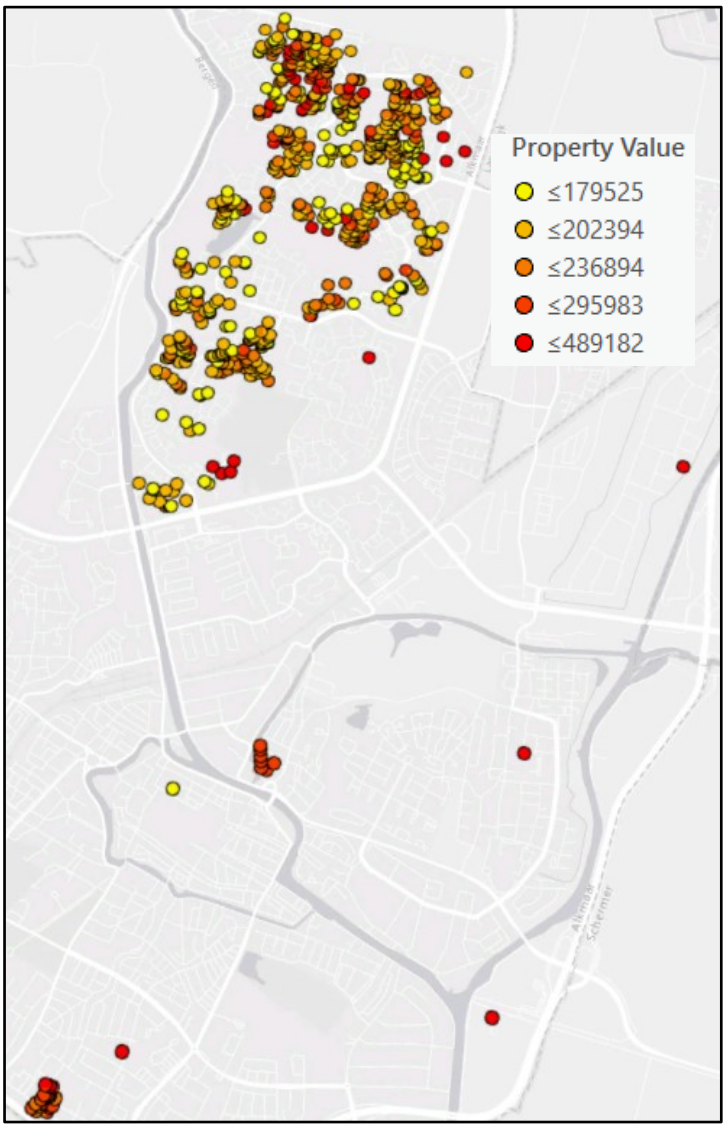
The viewshed per property is calculated with Esri's Model Builder, identifying a sequence of tools that are automatically activated on the 'run' button. The three databases on top are the input for the model. An iterator is added to handle each property separated. The view radius is set on 150m and the observer offset on 1.8m.

Appendix V - Spatial distribution of property values

A: Cluster 1 Property value distribution



B: Cluster 2 Property value distribution



The maps visualizing distribution of property values can be used to eyeball the effect of spatial clustering. In cluster 1 a clear clustering pattern can be identified. The West and East of the map locates clusters of high values. In the centre of the map lower values are clustered. Cluster 2 has less clear patterns, the few properties in the South of the map show high values, in the North of the map clustering of higher values takes place.

Note: Since both maps have different classifications, they should not be used for a comparison of the property values.

Appendix VI - General Statistics

Variables			Cluster 1 ***Build set***				Cluster 2 ***Build set***			
Name	Type	Values	Count	Mean	Min	Max	Count	Mean	Min	Max
Property value	Ratio	In Euros	593	229401	127416	564424	573	208481	76023	578947
Housing type (House class)	Interval	1. Terraced house 2. Linked house 3. Corner house 4. Semi-detached house 5. Detached house	388 22 153 28 2	1.7	1	5	371 8 142 45 7	1.95	1	5
Inside maintenance	Interval	1. Moderate 2. Moderate to adequate 3. Adequate to good 4. Adequate to good 5. Good 6. Good to excellent 7. Excellent	1 25 20 0 527 1 19	4.81	3	9	2 0 17 8 520 4 22	5.00	1	7
Outside maintenance	Interval	1. Moderate 2. Moderate to adequate 3. Adequate to good 4. Adequate to good 5. Good 6. Good to excellent 7. Excellent	2 1 54 35 480 3 18	4.94	4	9	1 0 7 8 528 4 25	5.05	1	7
Size	Ratio	M2 usable area		129	78	220		117	75	183
Garage	Interval	0. None 1. Not inbuilt 2. Inbuilt	242 85 84	0.43	0	2	497 67 9	0.15	0	2
Subjective view	Dummy	At water At park Open view Heavy Traffic	49 15 59 5	0.08 0.03 0.10 0.01	0 0 0 0	1 1 1 1	87 29 60 3	0.15 0.03 0.10 0.00	0 0 0 0	0 0 0 0
Objective view	Ratio	M2 View M2 Roads m2 Development M2 Agriculture M2 Green M2 Water		16381 1689 10691 206 2431 130	1942 0 1933 0 0 0	54709 16308 39577 23805 21540 13933		14451 980 10333 220 1440 1480	1831 0 1380 0 0 0	45538 15211 39611 30696 35873 17216

Appendix VII - Regression Results

Cluster 1: Non-spatial OLS

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]
Intercept	-92077.80430	17243.508077	-5.339853	0.000000*
M2	1719.290065	113.103085	15.201089	0.000000*
INSIDEMAINTE	13656.661587	2213.184816	6.170592	0.000000*
TYPE	15896.756375	1659.220219	9.580860	0.000000*
GARAGE	14262.310400	2546.726820	5.600251	0.000000*

Input Features:	Cluster1_Building	Dependent Variable:	CLUSTER1_CORPRICE
Number of Observations:	593	Akaike's Information Criterion (AICc) [d]:	14274.911709
Multiple R-Squared [d]:	0.511043	Adjusted R-Squared [d]:	0.507717
Joint F-Statistic [e]:	153.640218	Prob(>F), (4,588) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	405.030355	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	60.430512	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Jarque-Bera Statistic [g]:	363.941509	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

Cluster 1: Non-spatial OLS

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]
Intercept	-91874.89119	16750.649665	-5.484855	0.000000*
M2	1692.598596	109.959861	15.392877	0.000000*
INSIDEMAINTE	13694.531808	2149.931863	6.369752	0.000000*
TYPE	15806.065537	1611.863332	9.806083	0.000000*
GARAGE	14968.584430	2476.720670	6.043711	0.000000*
DUMWATER	35392.704224	5889.522553	6.009435	0.000000*

Input Features:	Cluster1_Building	Dependent Variable:	CLUSTER1_CORPRICE
Number of Observations:	593	Akaike's Information Criterion (AICc) [d]:	14241.555604
Multiple R-Squared [d]:	0.539382	Adjusted R-Squared [d]:	0.535458
Joint F-Statistic [e]:	137.474724	Prob(>F), (5,587) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	478.710695	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	60.034672	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Jarque-Bera Statistic [g]:	413.401092	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

Cluster 1: Spatial GIS-based OLS

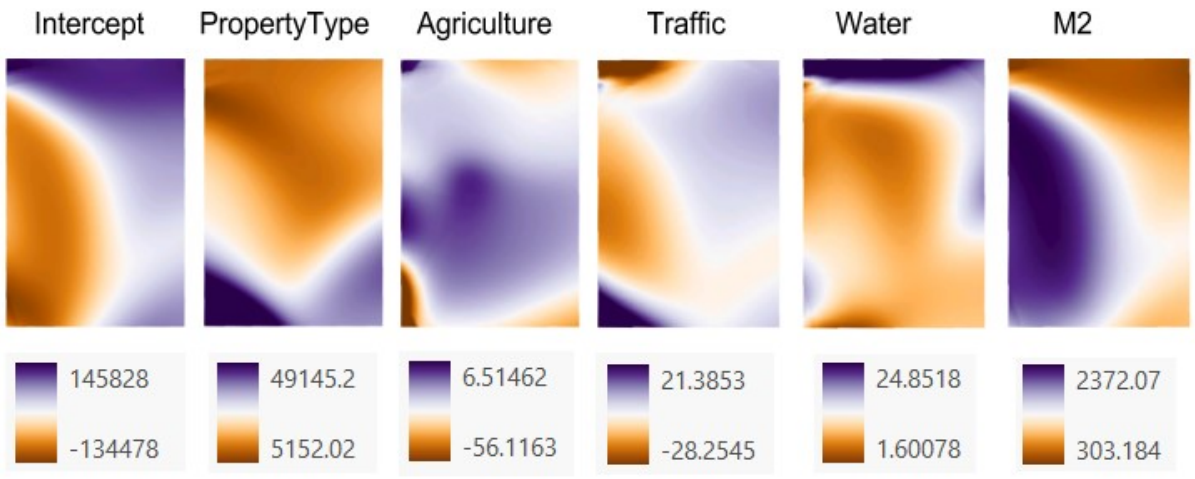
Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]
Intercept	-76442.09490	16784.107230	-4.554433	0.000008*
M2	1515.953295	109.450501	13.850583	0.000000*
INSIDEMAINTE	14536.954550	2096.545310	6.933766	0.000000*
TYPE	14991.068667	1597.492332	9.384126	0.000000*
TRAFFIC	-4.232846	0.613865	-6.895404	0.000000*
DEVELOPMENTS	0.770282	0.320737	2.401602	0.016621*
AGRICULTURE	2.117304	0.911054	2.324015	0.020452*
WATER	4.144103	0.747979	5.540401	0.000000*
GARAGE	16220.463297	2531.581457	6.407245	0.000000*

Input Features:	Cluster1_Building	Dependent Variable:	CLUSTER1_CORPRICE
Number of Observations:	593	Akaike's Information Criterion (AICc) [d]:	14210.289429
Multiple R-Squared [d]:	0.567574	Adjusted R-Squared [d]:	0.561650
Joint F-Statistic [e]:	95.814881	Prob(>F), (8,584) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	476.933917	Prob(>chi-squared), (8) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	66.608557	Prob(>chi-squared), (8) degrees of freedom:	0.000000*

Cluster 1: GWR model

Bandwidth : 1246.7236679573302
ResidualSquares : 651066064620.61389
EffectiveNumber : 28.649445023410795
Sigma : 33965.503071490028
AICc : 14076.043869326208
R2 : 0.67136514055679752
R2Adjusted : 0.6552642057764142

Parameter distribution over space per variable:



Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]
Intercept	11135.287048	12348.071980	0.901783	0.367540
M2	1209.252984	83.822153	14.426413	0.000000*
Maintenanc	5423.499656	1862.882459	2.911348	0.003746*
PropertyTyp	12849.348210	1111.272799	11.562731	0.000000*
Garage	35744.656396	3271.671857	10.925502	0.000000*

Cluster 2: non-spatial OLS

Input Features:	Cluster2_Building	Dependent Variable:	CL2LINE_CO
Number of Observations:	573	Akaike's Information Criterion (AICc) [d]:	13280.448611
Multiple R-Squared [d]:	0.670399	Adjusted R-Squared [d]:	0.668078
Joint F-Statistic [e]:	288.824103	Prob(>F), (4,568) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	504.722656	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	56.938099	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Jarque-Bera Statistic [g]:	941.041288	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

Cluster 2: spatial GIS-based OLS

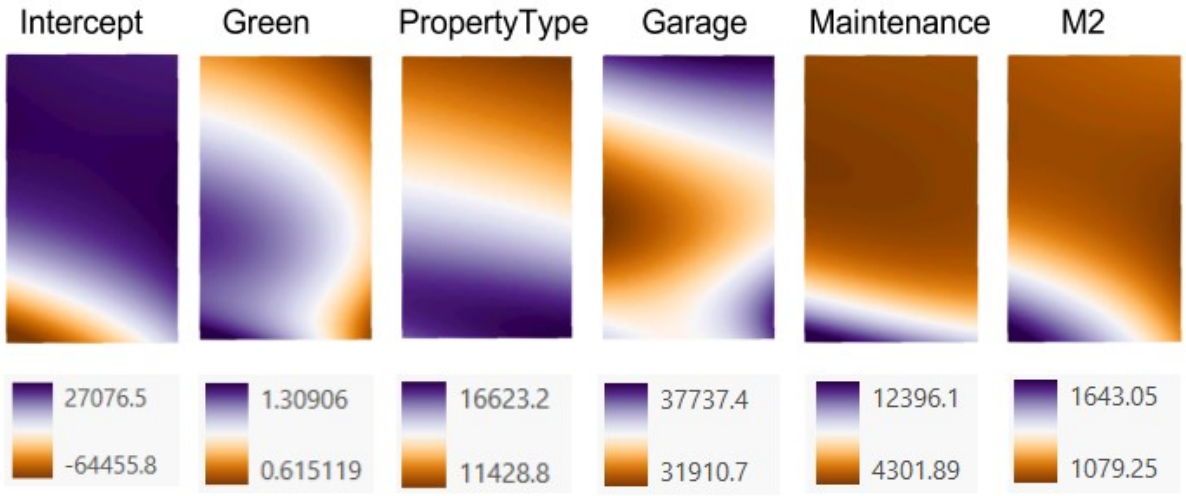
Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]
Intercept	14206.945267	12230.917230	1.161560	0.245901
M2	1193.423498	82.950167	14.387234	0.000000*
Maintenanc	4941.923349	1845.529601	2.677781	0.007622*
PropertyTyp	12768.536702	1098.533029	11.623262	0.000000*
Garage	35856.615071	3233.694071	11.088438	0.000000*
Green	0.917110	0.241110	3.803703	0.000168*

Input Features:	Cluster2_Building	Dependent Variable:	CL2LINE_CO
Number of Observations:	573	Akaike's Information Criterion (AICc) [d]:	13268.060605
Multiple R-Squared [d]:	0.678600	Adjusted R-Squared [d]:	0.675766
Joint F-Statistic [e]:	239.431685	Prob(>F), (5,567) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	628.668730	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	91.295614	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Jarque-Bera Statistic [g]:	412.745474	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

Cluster 2: GWR model

Bandwidth :	2894.7903212351748
ResidualSquares :	338171421474.35602
EffectiveNumber :	11.678118861764492
Sigma :	24544.96773381313
AICc :	13219.896427307351
R2 :	0.70826501994168334
R2Adjusted :	0.70271529722843318

Parameter distribution over space per variable:



Appendix VIII - Price Predictions

A. Model predictions of the first cluster

OLS model

Property Value	Predicted value	Residual	Deviation
179543	201460	-21917	21917
192389	238893	-46504	46504
314482	276614	37868	37868
237389	206452	30936	30936
305638	279646	25992	25992
242873	218702	24171	24171
251397	254670	-3273	3273
233520	213707	19812	19812
226340	210668	15672	15672
232936	238697	-5760	5760
242690	258841	-16151	16151
215428	221016	-5588	5588
137582	178168	-40585	40585
210976	206497	4478	4478
194732	225402	-30670	30670
190968	199211	-8243	8243
411647	367532	44115	44115
409313	324802	84511	84511
315560	243960	71601	71601
195846	160315	35531	35531
331355	225944	105411	105411
200317	213875	-13558	13558
237430	238953	-1523	1523
376283	289059	87224	87224
227498	259926	-32428	32428
198404	200490	-2087	2087

GWR model

Property Value	Predicted Value	Residual	Deviation
179543	183899	-4356	4356
192389	205839	-13450	13450
314482	284484	29998	29998
237389	222717	14672	14672
305638	297879	7759	7759
242873	230356	12517	12517
251397	259500	-8103	8103
233520	223689	9831	9831
226340	210333	16007	16007
232936	242764	-9828	9828
242690	232932	9758	9758
215428	228472	-13044	13044
137582	174996	-37413	37413
210976	212371	-1396	1396
194732	232790	-38058	38058
190968	218925	-27957	27957
411647	382242	29405	29405
409313	336625	72688	72688
315560	248144	67417	67417
195846	161426	34420	34420
331355	203235	128120	128120
200317	212479	-12161	12161
237430	208201	29229	29229
376283	282984	93299	93299
227498	234120	-6623	6623
198404	220725	-22322	22322

186891	236706	-49816	49816
217112	299494	-82382	82382
293798	266731	27066	27066
202811	213689	-10878	10878
223895	263093	-39198	39198
339429	279844	59585	59585
210359	271248	-60888	60888
183735	188681	-4946	4946
183580	170063	13517	13517
285141	289201	-4060	4060
264496	284894	-20397	20397
222787	163542	59246	59246
194972	224631	-29659	29659
183041	200279	-17238	17238
215084	218709	-3626	3626
225366	250731	-25365	25365
175978	220569	-44591	44591
215863	189026	26838	26838
237844	276254	-38411	38411
171038	205017	-33979	33979
183482	199493	-16012	16012
227519	201631	25888	25888
245614	211082	34532	34532
167854	213090	-45236	45236
184984	171770	13214	13214
231454	211965	19489	19489
240976	274883	-33907	33907
247883	259356	-11473	11473
237389	279183	-41794	41794
269006	254645	14361	14361
211672	214592	-2921	2921
213450	232881	-19430	19430

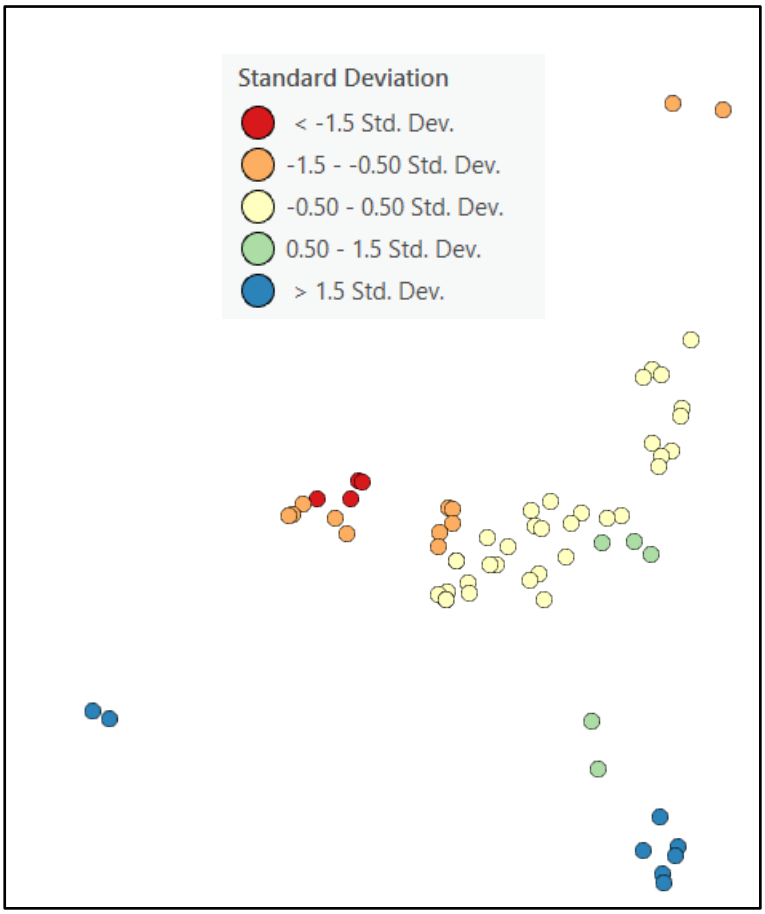
186891	223364	-36474	36474
217112	264990	-47878	47878
293798	292177	1620	1620
202811	214641	-11830	11830
223895	236547	-12653	12653
339429	302254	37175	37175
210359	238742	-28382	28382
183735	188504	-4769	4769
183580	199753	-16172	16172
285141	315513	-30372	30372
264496	312054	-47558	47558
222787	160594	62193	62193
194972	231087	-36114	36114
183041	200042	-17001	17001
215084	208095	6989	6989
225366	244530	-19164	19164
175978	218053	-42076	42076
215863	190871	24993	24993
237844	256723	-18879	18879
171038	206882	-35844	35844
183482	197100	-13618	13618
227519	206466	21052	21052
245614	192569	53045	53045
167854	206250	-38397	38397
184984	179854	5130	5130
231454	206173	25281	25281
240976	225303	15672	15672
247883	241394	6489	6489
237389	228851	8538	8538
269006	240075	28931	28931
211672	208014	3658	3658
213450	218622	-5172	5172

216581	232575	-15994	15994
182774	168333	14441	14441
167451	174099	-6649	6649
280196	311383	-31187	31187
174085	190452	-16366	16366
232192	**232814**		**29051**

216581	219193	-2612	2612
182774	175503	7271	7271
167451	201588	-34137	34137
280196	277689	2507	2507
174085	214713	-40628	40628
232192	**230110**		**25398**

The map on the right shows the distribution of the errors in the GWR prediction. A clear pattern of spatial clustering is observable, which means that the model did not manage to account for spatial dependency. This is confirmed by the Moran's I value of the model, which is 0.165, indicating spatial autocorrelation.

The biased model can be related to the fact that not all variables selected in the OLS model could be incorporated in the GWR model, which means that the GWR model is missing essential explanatory variables.



Cluster 2

OLS model

Property Value	Predicted value	Residual	Deviation
237538	250323	12785	12785
245984	248655	2671	2671
238901	249489	10588	10588
272008	212852	-59156	59156
190803	212852	22049	22049
167598	233758	66161	66161
168715	209734	41019	41019
209563	222399	12837	12837
198203	252952	54749	54749
170910	208434	37524	37524
192878	230753	37875	37875
206131	222894	16763	16763
168780	223112	54332	54332
184224	224141	39918	39918
163784	234472	70688	70688
178765	206924	28160	28160
215642	224786	9144	9144
177654	218819	41166	41166
237844	258731	20887	20887
220348	231343	10995	10995
202258	220752	18494	18494
238547	291919	53372	53372
172304	218821	46517	46517
166172	232692	66520	66520
178042	248664	70622	70622
223424	239361	15937	15937
191620	233970	42350	42350

GWR model

Property Value	Predicted Value	Residual	Deviation
237538	268922	31384	31384
245984	269776	23792	23792
238901	269036	30135	30135
272008	210642	-61366	61366
190803	208987	18184	18184
167598	232687	65089	65089
168715	207396	38681	38681
209563	218747	9184	9184
198203	250747	52544	52544
170910	206268	35357	35357
192878	226685	33806	33806
206131	220119	13988	13988
168780	220381	51601	51601
184224	221423	37199	37199
163784	233612	69827	69827
178765	204687	25922	25922
215642	221577	5934	5934
177654	216103	38449	38449
237844	254009	16165	16165
220348	228402	8054	8054
202258	218197	15939	15939
238547	287234	48687	48687
172304	216989	44685	44685
166172	229490	63318	63318
178042	245146	67104	67104
223424	236416	12992	12992
191620	232899	41279	41279

202960	243847	40887	40887
209693	238438	28745	28745
250235	304352	54116	54116
192878	240929	48051	48051
243650	292147	48497	48497
174634	250895	76261	76261
188199	214317	26117	26117
218101	228891	10790	10790
189759	236721	46962	46962
195531	249992	54461	54461
171776	206885	35109	35109
180761	218819	38058	38058
149577	183017	33440	33440
187206	232811	45605	45605
307263	352653	45390	45390
190275	212852	22577	22577
135366	218819	83453	83453
184724	232422	47698	47698
200846	228703	27857	27857
201788	221206	19418	19418
196100	237196	41096	41096
221261	223593	2331	2331
314540	305696	-8844	8844
196341	280213	83871	83871
204878	266610	61732	61732
190743	215239	24496	24496
244831	276633	31801	31801
269556	328784	59228	59228
214059	230753	16694	16694
186592	230753	44161	44161
163221	193084	29864	29864
201754	218819	17065	17065

202960	241528	38568	38568
209693	235880	26187	26187
250235	300741	50506	50506
192878	238041	45162	45162
243650	289502	45851	45851
174634	248854	74219	74219
188199	213378	25179	25179
218101	226666	8565	8565
189759	234825	45066	45066
195531	248087	52557	52557
171776	206420	34644	34644
180761	217738	36977	36977
149577	184559	34983	34983
187206	231238	44032	44032
307263	347786	40524	40524
190275	212855	22580	22580
135366	218482	83116	83116
184724	231711	46987	46987
200846	227477	26631	26631
201788	220967	19180	19180
196100	236237	40137	40137
221261	223497	2236	2236
314540	301236	-13305	13305
196341	279979	83638	83638
204878	267397	62519	62519
190743	215787	25044	25044
244831	276672	31841	31841
269556	326212	56656	56656
214059	230736	16677	16677
186592	230766	44174	44174
163221	195396	32175	32175
201754	219593	17839	17839

215863	244356	28493	28493
176533	250323	73791	73791
203117	**238924**		**38037**

215863	243373	27510	27510
176533	249077	72544	72544
203117	**238185**		**37516**

The map on the right shows the distribution of the errors in the GWR prediction. A random pattern is observable, which means that the model managed to account for spatial dependency. This is confirmed by the low Moran's I value of the model, which is 0.018, indicating that there is no sign of spatial autocorrelation.