MASTER

What is the monetary value of urban green spaces?
Comparing different spatial econometric models

Lammes, Roelof J.

Award date:
2022

MASTER THESIS

What is the monetary value of urban green spaces?

Comparing different spatial econometric models

R.J. Lammes
1462830
05-08-2022

DEPARTMENT OF THE BUILT ENVIRONMENT

EINDHOVEN UNIVERSITY OF TECHNOLOGY

What is the monetary value of urban green spaces?
Comparing different spatial econometric models

Colophon

Title	What is the monetary value of urban green spaces?
Subtitle	Comparing different spatial econometric models

Student information

Name	Lammes, R.J. (Roelof)
Student ID	1462830
E-mail address TU/e	r.j.lammes@student.tue.nl

Research institution

Name	Eindhoven University of Technology
Faculty	Department of the Built Environment
Master	Architecture, Building and Planning
Graduation program	Urban Systems and Real Estate
Graduation project	7Z45M0
Study load	45 ECTS

Graduation committee

First supervisor:	prof.dr. T.A. Arentze t.a.arentze@tue.nl
Second supervisor:	dr. T. Feng t.feng@tue.nl
Third supervisor:	dr. I.V. Ossokina i.v.ossokina@tue.nl

This Master's thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Integrity.

Preface

This is a 45 ECTS master thesis written for the Urban Systems and Real Estate track of the Architecture, Building and planning program at the Department of the Built Environment, University of Technology Eindhoven.

In the context of my research and personal development, I would like to thank some people who contributed to this process. First, I would like to thank my first and second supervisors Theo Artenze and Tao Feng for their time, effort and willingness to share their constructive feedback, expertise and knowledge to guide me through this process. Since this thesis was written during the Covid-19 pandemic, we did not have any single physical thesis meeting at the University. Despite this, I always felt that they both were involved in the process.

Secondly, I would like to thank my third supervisor Ioulia Ossokina for sharing her expertise about hedonic pricing and modelling spatial dependencies. Her constructive feedback did not only help me to understand these difficult topics better, but it also helped me to improve the structure of this thesis.

Special thanks goes out to Peter van der Waerden for helping me handle the large amount of data and providing the opportunity to calculate some variables in the software program TransCAD. Although he was not involved in my research as a supervisor, he did not hesitate to share his knowledge, expertise or time to help me further in the process, which I am very grateful for.

At last, I would like to thank the Nederlandse Coöperatieve Vereniging van Makelaars en Taxateurs (Dutch Associations of Realtors), Husqvarna and Overstory for their willingness to share their data to make this research possible. Special thanks go out to the staff of Overstory, who put quite some effort into ensuring that the data was provided in the proper structure for this thesis.

I hope you will enjoy reading my thesis.

Roelof Lammes
Eindhoven, 05-08-2022

Abstract

This thesis applied the Hedonic Pricing method to examine how the implicit value of urban green spaces (UGS) is reflected in the transaction price of dwelling in the Dutch cities Almere and Eindhoven. Four (spatial) hedonic models were estimated to retrieve the marginal values associated with proximity to different UGS amenities and the density of trees, grass and water in the 6-digit zip code area where the dwelling is located. Special attention was paid to classifying urban green spaces, given the research question of whether different UGS types have other price effects. The models use a sample of 19,461 dwelling transactions in Almere and Eindhoven between 2014 and 2018, provided by the NVM Netherlands association of real estate agents and appraisers. Additionally, this study used land coverage data, provided by HUGSI, which is expected to be more accurate than land use data, to account for the density of urban green. In this thesis land coverage data allowed to distinguish water, trees and grass from other land use types for a 10 by 10 grid, providing very accurate UGS data. This type of data is promising for future research and conceivably has many other applications like distinguishing more types of UGS or for collecting data on solar panels and green roofs.

We estimate the ordinary least square model, as well as three spatial econometrics models (SLX, SEM and SDEM) to control for the spatial dependency in the data. No statistically significant differences in outcomes or overall fit were found. The results suggest further that continuously operationalized proximity variables are less suitable for spatial econometric models, also because interpreting the lags of these variables does not seem to have much practical sense. Findings indicate that dwelling transaction prices increase by 0.9% with a 10 percentage point increase of land covered by water in the 6-digit zip code area where the dwelling is located. A 10 percentage point increase in the area covered by trees goes together with a 0.21% increase in the transaction price while a 10 percentage point increase in grass leads to a 0.2% decrease in the transaction price. Further, the results indicate that the UGS proximity impact on housing prices differs between UGS types. The social relevance of this study can be found in the awareness-raising concerning the economic values of UGS in the Netherlands. The academic relevance lies mainly in a contribution to the discussion about modelling spatial dependencies and exploring new methods to operationalise UGS in a hedonic pricing model.

Keywords: Land coverage data, Hedonic pricing method, Urban green spaces, Spatial autocorrelation, Spatial interaction effects

Table of content

List with figures

List with tables

List with abbreviations

AVG	Algemene Verordening Gegevensbescherming
CBS	Centraal bureau van statistiek
CVM	Contingent valuation method
EP	Experienced preference
ESA	European Space Agency
GIS	Geographic information system
GNS	General nesting spatial
GWR	Geographically weighted regression
HP	Hedonic pricing
HUGSI	Husqvarna Urban Green Space Index
IDW	Inverse distance weighted
LM	Lagrange Multiplier
LR	Likelihood ratio
NVM	Nederlandse Coöperatieve Vereniging van Makelaars en Taxateurs in onroerende goederen
NWB	Nationaal Wegen Bestand
OGS	Open green space
SAR	Spatial autoregressive
SAC	Spatial autoregressive combined
SEM	Spatial error term
SDM	Spatial Durbin
SDEM	Spatial Durbin error
SLX	Spatial lag of X
PS-RS	Pseudo-Repeat Sales
UGS	Urban green spaces
VIF	Variation Inflation Factor
WDC-RSAT	World Data Centre for Remote Sensing of the Atmosphere

List with model equations

Nr.	Model	Equation
1	**Standard HP model**	$Y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i$
2	**Standard HP model UGS studies**	$Y_i = \beta_0 + \sum_{j=1}^{J} a_j S_{ij} + \sum_{m=1}^{M} \beta_m L_{im} + \sum_{n=1}^{N} \gamma_n U_{in} + \varepsilon_i$
3	**GNS model**	$Y_i = \beta X + \rho W y + \theta W X + \mu$ $\mu = \lambda W \mu + \varepsilon$
4	**Moran's I equation**	$I = \frac{n}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}$
5	**Standard SAR model**	$Y_i = \beta X + \rho \mathrm{W} y + \varepsilon$
6	**Standard SEM model**	$Y = \beta X + \mu$ $\mu = \lambda W \mu + \varepsilon$
7	**HP thesis**	$\ln(T_i) = \beta_0 + \sum_{j=1}^{J} a_j S_{ij} + \sum_{m=1}^{M} \beta_m L_{im} + \sum_{n=1}^{N} \gamma_n U_{in} + \sum_{k=1}^{K} x_k C_{ik} + \varepsilon_i$
8	**SLX model thesis**	$\ln(T_i) = \beta_0 + \sum_{j=1}^{J} a_j S_{ij} + \sum_{m=1}^{M} \beta_m L_{im} + \sum_{n=1}^{N} \gamma_n U_{in} + \sum_{k=1}^{K} x_k C_{ik} + \theta W X + \varepsilon_i$
9	**SEM model thesis**	$\ln(T_i) = \beta_0 + \sum_{j}^{J} a_j S_{ij} + \sum_{m}^{M} \beta_m L_{im} + \sum_{n}^{N} \gamma_n U_{in} + \sum_{k}^{K} x_k C_{ik} + \mu_i$ $\mu_i = \lambda W \mu + \varepsilon_i$
10	**SDEM model thesis**	$\ln(T_i) = \beta_0 + \sum_{j}^{J} a_j S_{ij} + \sum_{m}^{M} \beta_m L_{im} + \sum_{n}^{N} \gamma_n U_{in} + \sum_{k}^{K} x_k C_{ik} + \theta W X + \mu_i$ $\mu_i = \lambda W \mu + \varepsilon_i$
11	**SAR model thesis**	$\ln(T_i) = \beta_0 + \sum_{j}^{J} a_j S_{ij} + \sum_{m}^{M} \beta_m L_{im} + \sum_{n}^{N} \gamma_n U_{in} + \varepsilon_i$
12	**SDM model thesis**	$\ln(T_i) = \beta_0 + \sum_{j}^{J} a_j S_{ij} + \sum_{m}^{M} \beta_m L_{im} + \sum_{n}^{N} \gamma_n U_{in} + \sum_{k}^{K} x_k C_{ik} + \rho W y + \mu_i$ $\mu_i = \lambda W \mu + \varepsilon_i$

CH 1 Introduction

1.1 Background

1.1.1 The benefits of urban green spaces

The construction of more urban green spaces is one of the sustainable development goals that should be reached by 2030 (United Nations, 2015). Urban green spaces (UGS) are proposed as a spatial intervention that can help to reduce and control for the negative consequences of urbanization and climate change. The World Health Organization (2017) defines urban green space as follows:

‘All urban land covered by vegetation of any kind. This covers vegetation on private and public grounds, irrespective of size and function, and can also include small water bodies such as ponds, lakes or streams.’

In the last decades, many positive social, health, ecological and environmental externalities of urban green spaces were found by studies of different academic disciplines. Since the nineties, urban studies have been interested in the impact of urban green space on the living environment. Different resilience effects in urban areas are attributed to urban green spaces. Studies found that urban heat islands can be eased by a reformation of the amount and design of the land cover, mainly when more room is allocated to urban green space (Zhou et al., 2011; Lin et al., 2015). Urban green space cooling effects of 1.7 to 4.5 ° during average peak hours were found, which underline the critical role of urban green spaces in reducing urban heat island effects (Santamouris et al., 2017; Aram et al., 2019). In addition to the urban green space cooling effect, more positive environmental impacts are linked to urban green spaces. Different studies found that urban green spaces contribute to preserving CO_2 and storing carbon, help to reduce air pollution and improve the air quality and play an important role in biodiversity protection by being the habitat for a diversity of flora and fauna (McPherson et al., 1998; Nowak, 2002; Kuhn et al., 2004; Cornelis & Hermy, 2004; Myeong et al., 2006; Depietri & McPhearson, 2017;).

Next to environmental and ecological externalities, studies discovered different health benefits of urban green spaces. The presence of green spaces reduces psychological stress and has a positive effect on the mental health (Feng & Kolt; 2013, Black & Richard, 2020; Oude, 2020). In addition, it is proven that the presence of urban green spaces increases the physical activity of residents living nearby (Lennon et al, 2017). The presence of urban green spaces also leads to social-cultural benefits as urban green spaces provide an important role within communities to function as a meeting place for people (Vries de et al., 2013; O’ Brien et al., 2014). However, a few negative externalities of urban green spaces are also mentioned. Not all trees are suited for urban places because some are bad for the local air quality or cause adverse health effects (Henniger, 2011). In addition, urban green spaces are sometimes (perceived as) unsafe and can be a location with high crime rates (Sanesi & Chiarello, 2006). However, with this knowledge in mind, urban planners and decision makers should be able to account for these disadvantages and construct urban green spaces that only bring social, ecological, environmental and health benefits.

From an economic perspective, the advantages and disadvantages of urban green spaces are more intertwined. Nur Syafigah et al. (2020) mention that urban green spaces can bring economic benefits to their surroundings like; providing an attractive business location, boosting social and community development and creating new jobs. In addition, different studies found that the presence of urban green spaces leads to higher property values. A positive price effect of 5 to 27 percent has been found between the proximity of urban green and the transaction prices of houses (Tajimia, 2003; Czembrowski & Kronenberg, 2016; Daams et al., 2019; Du & Zhang, 2020). Therefore, studies argue that urban green spaces do not only bring economic benefits but also instigate new dynamics of exclusion and segregation in the form of eco-gentrification.

The NYC’s High line project is a frequently used example of eco-gentrification in academic studies. Different hedonic pricing studies show a tremendous increase in transaction prices from houses surrounding this newly built UGS.

Value increases between 55 and 103 percent are mentioned (Darren, 2013; Loughran, 2014; Angeulovski, 2018; Jo Black & Richards, 2020). Higher house prices and rents caused by spatial transformation led to more negative externalities like the displacement of local businesses and residents. With that in mind, it should be mentioned that urban greening is a double-edged sword policy makers must carefully wield (Anguelovisk et al., 2018; Jo Black & Richards, 2020).

1.1.2 Monetary value of urban green spaces?

Regarding the previous section, it can be stated that UGS bring many benefits. However, a lack of an uneven distribution of urban green spaces can also result in gentrification with all its disadvantages. To be prepared for these unwanted disadvantages, it seems important to predict the price effect of urban green spaces as accurate as possible. Therefore, this thesis elaborates on how urban green spaces are valued. Urban green spaces are public goods with considerable interest in integrating into systems of national account. But because there exists no-market based valuation for public goods, it is hard to put an exact monetary value on urban green spaces. Therefore, economists refer to other methods that can indicate an estimation of the monetary value. Looking at former ecosystem valuation studies, it can be concluded that three main approaches are used to put a monetary value on urban green spaces. The stated preference approaches experienced preference approaches and revealed preference approaches (Krekel & Kolbe, 2020).

1) With the stated preferences approach, hypothetical cases are used to ask respondents what direct financial value they would give. Examples of these approaches are discrete choice models and contingent valuation methods. A drawback of these approaches is that bias can lead to incorrect values because of dishonest, strategic or socially desirable answers.

2) The experienced preferences approach uses an OLS regression to measure the effect of different factors associated with well-being, including UGS, for example (Shi et al., 2019). The impact of UGS on the subjective well-being is then evaluated against income, providing a monetary value.

3) The revealed-preference approach also follows an OLS regression, but instead of well-being data, it uses real estate transaction data to account for the dependent variable. This method is called hedonic pricing. With the help of hedonic pricing, changes in real estate prices around public goods are used to infer a monetary value. According to Krekel & Kolbe (2020), hedonic pricing is the dominant strategy for valuing the effect of urban green spaces in a monetary way.

The literature study of this thesis agrees with the statement of Krekel & Kolbe (2020). Table 1 shows that most studies use hedonic pricing as the valuation method of choice to put a monetary value on urban green spaces (Luttik, 2000; Conway et al., 2008; Hoshino & Kuriyama, 2009; Saphores & Li, 2012; Panduro & Veie, 2013; Kolbe & Wüstemann, 2014; Herath et al., 2015; Sander & Zhao, 2015; Zygmunt & Ghuszak, 2015; Czembrowski et al., 2016; Daams et al., 2016; Helgers & Vastman, 2016; Xu et al., 2016; Engström & Gren, 2017; Liebelt et al., 2017; Votis, 2017; Trojanek et al., 2018; Daams et al., 2019; Czembrowski et al., 2019; Du & Zhang, 2020; Jo Black & Richards, 2020; Samad et al., 2020).With the hedonic pricing method, the transaction prices of residential real estate objects are predicted by different variables, including (the presence of) urban green spaces as one of the characteristics. There are also two studies that choose subjective well-being data or self-reported life satisfaction instead of transaction prices (Ambrey & Flemming, 2013; Krekel et al., 2015). And a number of studies use the CVM method to determine the value of urban green spaces (Jim & Chen, 2006; Alex & Jim, 2010; Dumenu, 2013; Latinopoulos et al., 2016; Adegun, 2017; Gelo & Turpie, 2021).

However, the hedonic pricing method does not take spatial dependencies into account, while two observations in space might be spatially correlated. This could lead to heteroskedastic issues or bias coefficient estimates (Hilger, 2018). To overcome this problem, urban researchers found a way to extend the hedonic pricing model with the help of spatial econometrics.

Since 2008, studies have used different forms of models to account for spatial dependencies. However, it is remarkable that so many different models are used, but no dominant models can be distinguished for this specific research subject. According to Elhorst (2017), choosing between the different models and specifications of the spatial weight matrix '*W*' is by far the greatest problem in empirical research. Especially when no reference to specific economic theories can be made. Krekel (2020) recognizes this, and states that the methodological discussion of ecosystem valuation is currently focused on the benefits and disadvantages of the respective method. According to Krekel, this discussion should move towards practical guidance on which (combination of) method(s) is preferable under which circumstances. This would be a major step towards including a common methodology for monetary valuation of intangibles in systems of national account. So far, no research appears to focus on what valuation methods and which spatial econometric model is favored when the research focus is on the monetary value of UGS.

As shown in Table 1, different methodologies and models are used to put a monetary value on urban green spaces. Approaching a goal in different ways has benefits. This leads to new perspectives and increases the chance that all views and insights are respected, reducing the possibility of an unambiguous analysis. However, in the case of putting a monetary value on UGSs, it seems that there is a need for a common methodology and a transparent, straightforward approach. There seem to be ambiguities about the different spatial econometric models regarding which model is the most suitable one for this specific topic of interest.

Reference	Year	Country	Valuation method	Spatial extension	Specification of W	Time extension
Adegun	2017	South Africa	CVM	-	-	-
Alex & Jim	2010	Hongkong	CVM	-	-	-
Ambrey & Fleming	2013	Australia	EP	-	-	-
Conway et al.	2008	USA	HP	SAR	Queen-contiguity	-
Czembrowski et al.	2016	Poland	HP	SDEM	Nearest-five-neighbours	-
Czembrowski et al.	2019	Sweden	HP	SAC	Nearest-five-neighbours	-
Daams et al.	2016	Netherlands	HP	-	-	-
Daams et al.	2019	Netherlands	HP	-	-	-
Dekkers & Koomen	2008	Netherlands	HP	SAC + SEM	Unknown	
Du & Zhang	2020	USA	HP	-	-	-
Dumenu	2013	Ghana	CVM	-	-	-
Engström & Gren	2017	Malmö	HP	SAR, SEM & SDEM	Unknown	-
Gelo & Turpie	2021	Uganda	CVM	-	-	-
Helgers & Vastman	2016	België	HP	-	-	-
Herath et al.	2015	Austria	HP	SLX, SEM & SDM	Distance band & K-nearest neighbours	-
Hoshino & Kuriyama	2009	Japan	HP	SDEM	Max-min distance-band	-
Jim & Chen	2006	China	CVM	-	-	-
Jo Black & Richards	2020	USA	HP	-	-	Triple-difference
Kolbe & Wüstemann	2014	Germany	HP	-	-	-
Krekel et al.	2015	Germany	EP	-	-	-
Latinopoulos et al.	2016	Greece	CVM	-	-	-
Liebelt et al.	2018	Germany	HP	-	-	-
Luttik	2000	Netherlands	HP	-	-	-
Panduro & Veie	2013	Denmark	HP	-	-	-
Samad et al.	2020	Malaysia	HP	GWR	-	-
Sander & Zhao	2015	USA	HP	SEM & GWR	500 meter distance-band	-
Saphores & Li	2012	USA	HP	SDEM & GWR	-	-
Trojanek et al.	2018	Poland	HP	-	-	-
Votsis	2017	Finland	HP	SEM, SDM & SEM	Rook-contiguity	-
Xu et al.	2016	Bejing	HP	-	-	-
Zygmunt & Gluszak	2015	Poland	HP	SAR & SEM	1 km distance-band	-

Abbreviations: CVM = Contingent valuation method EP = Experienced preference, HP = Hedonic pricing, GWR = Geographically weighted regression, SAR = Spatial autoregressive, SAC = Spatial autoregressive combined, SEM = Spatial error term, SDM = Spatial Durbin, SDEM = Spatial Durbin error.

Table 1: Overview studies with different UGS valuation methods

1.2 Research problem

Different academic disciplines found that urban green spaces contribute to a better living environment, counteract negative externalities of climate change and provide many more ecological, social, environmental and health benefits. Therefore, there seems to be enough evidence that urban green spaces should be more interwoven into the urban landscape. However, it is difficult to demonstrate the importance of urban green spaces as they are public goods that lack a monetary value. Resulting in the fact that urban green spaces are often neglected by decision-makers and urban planners in the decision-making process as they give priory to other public services like public transport, education, health or energy infrastructure. In addition, studies show that unfair distribution of urban green spaces could lead to undesired phenomena like eco-gentrification.

Many international studies about the value of urban green spaces have been conducted. However, with only four Dutch studies, little is known about the monetary value of urban green spaces in the Netherlands. It is stressed that a standard hedonic pricing model does not account for spatial dependencies. To account for these spatial effects, a hedonic pricing model can be extended with the help of spatial econometric modelling. However, past research shows that different models and spatial weight matrices are used and a common methodology is lacking while the usage of different models and spatial weight matrices influence the estimated price effects.

1.3 Research aim and questions

This thesis will be focused on the value of urban green spaces in the Netherlands as only a few studies have done this before. As discussed, the dominant method to put a monetary value on urban green is hedonic pricing. By involving spatial econometrics to account for the spatial dependencies in this urban research, this thesis tries to predict the value of urban green spaces as accurate as possible. In doing so, this thesis aims to contribute to the discussion about what spatial econometric model and spatial weight matrix are the most appropriate to use for urban green space valuation.

On the forehand, an extensive study will be done to find out how UGS should be classified and operationalized into the model. As urban green space is an umbrella term that is used to indicate different types of UGS. Studies define UGS partly on their research question but also by the available data and technology. This thesis aims to include UGS as completely as possible in the analysis by learning from previous study results and exploring new data sources.

Concluding, this thesis aims to get a better understanding of the value of UGS in the Netherlands by measuring the marginal impact of UGS on residential housing prices and while doing so, explore new methods to operationalize UGS and involve spatial econometric modelling to account for spatial dependency, in order to contribute to the general discussion about the valuation of UGS by hedonic pricing.

The main research question is based on the research aim stated above and follows:

'What is the impact of urban green space on the market value of residential real estate in the Netherlands?'

A few sub-questions are formed to help answer the main research question. The sub-questions are as follows:

'*How can urban green spaces be operationalized and included in a hedonic pricing model?'*

'What spatial econometric extension to the hedonic pricing is most suitable for taking spatial dependencies into account in UGS valuation research?'

'What specification of the weight matrix is the most suitable for UGS valuation research?'

The first question above elaborates on the concept of UGS and hedonic pricing, with the main objective of gaining insight into those concepts. The second and third questions aim to better understand the spatial econometric modelling concepts in relation to UGS valuation.

1.4 Research relevance

Research relevance can be divided into two types, scientific and social relevance. Scientific relevance discusses the direct or indirect contribution of practice-based research to a determined research field, while social relevance debates the link between research and practice (Verschuren & Doorewaard, 2015). This section will first discuss the social relevance before elaborating on this thesis's academic relevance.

1.4.1 Social relevance

First of all, this thesis aligns with the United Nations sustainable development goals to construct more UGS before the end of 2030. This thesis tries to contribute to this goal by understanding what effects UGS have on residential housing. Capturing this value effect can help urban planners and policymakers argue for UGS instead of other land use types. Compared to industrial sites, residential areas or, for example, parking space, UGS could have a backlog because it does not have a clear monetary value expressed in direct financial gains. In addition, estimating the price effect for UGS can provide other interesting insights as it might help detect negative externalities like (eco)-gentrification effects. This information can help decision-makers with creating policies on forehand that prevent the displacement of residents for example or contributes to arguing for a fair distribution of UGS. In addition, this thesis thinks it relevant to distinguish different types of UGS, as it expects that they will have a different effect on the housing prices. Not only for policymakers but also for developers or investors, it can be valuable information to know if the house price will be different when it is located next to a cemetery instead of a city park.

Concerning the interest of this study in the UGS operationalization and (spatial) modelling process, another argument should be mentioned. A recent social matter in the Netherlands that impacted thousands of Dutch individuals illustrates why it is so essential that the chosen research methodology and model choices are correct. In 2019 an advisor committee with delegates from TU Delft published an advisory report on behalf of the Ministry of Economy Affairs and Climate. This report contains an analysis of different models that could be used to calculate the amount of financial compensation to each affected individual in the province of Groningen for the value decrease of their dwelling due to earthquakes caused by the extraction of natural gas. This report shows that minor modelling differences lead to great differences in outcome, what could have tremendous impact for individuals, as this matter resolve around large sums of money that are at interested for a great number of different stakeholders. Although the modelling process in this thesis has no direct impact on different individuals, this example underlines how important the modelling process is. Therefore, this study tends to contribute in getting a better understanding about the impact modelling spatial dependencies.

1.4.2 Scientific relevance

Besides its social relevance, distinguishing different UGS types also has a scientific contribution. A well-known argument in UGS valuation studies is that urban green spaces are often handled as a homogenous good, while they differ in many ways. By taking different types of UGS into account, this thesis hopes to contribute to the academic discussion about operationalising UGS. New (open) data sources will be explored to find what types of UGS can be distinguished, and the research results should point out if it is relevant for future research to distinguish different UGS types. In addition, this thesis will explore the possibilities with 'land coverage' data obtained by deep learning techniques. So far, as this thesis is concerned, this new type of spatial data has not been used yet for the operationalisation of urban green in a hedonic pricing model.

At last, this thesis hopes to contribute to the discussion about modelling spatial dependency in UGS valuation studies. So far, there is no common methodology regarding which models and spatial weight matrixes should be used. This thesis hopes to contribute to this academic discussion.

Research results should not only point out if it is relevant to distinguish different types, but the operationalisation part of this research should also point out how much effort, in terms of time, it costs.

1.5 Research area

As far as this thesis is concerned, only a few hedonic pricing studies with interest in determining the price effect of urban green spaces have been conducted in the Netherlands. Most of these studies were focussed on large cities in the Randstad region (Luttik, 2000; Dekkers & Koomen, 2008; Daams et al., 2016; Daams et al., 2019). All these studies argued that the largest cities in the Randstad region have the highest urban pressure and therefore are an obvious choice to use as a study case. This thesis uses another part of the Netherlands as a region of study. In the last decade, other cities like Almere and Eindhoven have experienced a high degree of urbanisation. The city of Eindhoven is located in the southern part of the Netherlands, while Almere is also part of the Randstad region. With the predicted population growth until 2030 (Almere (22.8%), Eindhoven (16,4%)), it can be expected that both areas will gain even more urban pressure in the future (CBS,2019).

The city of Eindhoven is divided into 20 districts comprising 116 neighbourhoods and houses 235.691 residents. Almere contains six districts with 69 neighbourhoods and has 214.715 residents. Eindhoven is the 5th city in the Netherlands, and Almere is ranked as number eighth, considering the number of inhabitants (CBS, 2021). According to the urban green space index, Eindhoven is the second greenest city in the Netherland with 46% area of urban green space divided by the total urban area of the city, while Almere is ranked fourth with 44% of urban green space (HUGSI, 2021). Figure 1 shows the distribution of urban green spaces in Almere and Eindhoven. Light green areas indicate grass, dark green indicates trees, blue indicates water and grey indicates other land uses (HUGSI, 2021).

Figure 1: Research areas (left: Eindhoven, right: Almere) and their distribution of UGS.

1.6 Structure of the thesis

Following this introduction chapter, this thesis is structured as follows: the **second chapter** provides the literature review focussed on; exploring the concept of UGS, including how UGS is defined, classified and operationalised in previous research, the findings of former UGS studies in the Netherlands and the theoretical framework of this research. In **chapter three**, the methodology used in this research is explained. First, the concept of hedonic pricing in relation to UGS valuation is explained, followed by the theoretical foundation of spatial econometric modelling. Including an explanation of its core concepts, how spatial interaction effects can be captured with the help of lag models, what the differences are between local and global models and how the spatial weight matrix can be specified. Finally, the last section of this chapter discusses the spatial dependency statistics that will be used to explore if spatial autocorrelation is present in the dataset. **Chapter four** specifies the weight matrix, OLS and spatial model. In the first section, the motivation behind the specification of the HP function, including the choices for its functional form and the explanatory variables, is given.

This is followed by a motivation for the specification of the weight matrix and the specification of the six different spatial lag models. The last section includes an explanation of the method that will be used to determine which spatial model(s) should be used to interpret the results. **Chapter five** follows with the data that is used in this thesis. First, the content of the derived four data sets is explained. In the second section, the UGS classification used in this thesis is given and explains how the UGS data is transformed into workable variables, followed by a discussion of what variables should be included in the hedonic pricing model and what price effects are expected. After an outlier and multicollinearity analysis, an overview of the included variables with their descriptive statistics is given. In **chapter six**, the results are discussed. First, the spatial dependencies test results are discussed, followed by the result of the Lagrange Multiplier (LR) tests. These results inform what kind of spatial dependency is present in the data and what spatial model should be used for interpretation. Next, the results of the estimated models are given, first, the general model results will be briefly discussed, followed by an elaborate discussion about the results of the individual variables. The outcomes of the UGS will be transformed into a monetary value. Finally, **chapter seven** concludes the thesis by answering the research questions, discussing the finding of this thesis, explaining its limitations and giving recommendations for future research.

CH 2 Literature review

This chapter provides a review of related research about the valuation of urban green spaces. First, the concept of urban green spaces is explored. In this section, the definition, classification, and operationalization of urban green spaces are discussed, ending with innovative recommendations from previous studies for UGS valuation research. This chapter's second section provides an overview of the findings of previous Dutch UGS valuation studies, followed by the final section of this chapter that provides the theoretical framework for this research.

2.1 Concept of UGS

This first section explores the concept of urban green spaces. First, the definition of green spaces is explored. The second section provides an overview of how urban green spaces can be classified. The third section explains how urban green spaces can be operationalized, while the last section elaborates on some innovative recommendations from previous studies.

2.1.1 Defining UGS

Before elaborate research can be conducted on how UGS is valued, UGS must be defined. The definition of UGS handled by the World Health Organisation (2017) is already mentioned in the background of this study. This definition is in line with another definition that spatial planners often use in urban green literature:

"Urban green space is referred to as any vegetation that exists in the connection between urban and nature. It includes parks, open green spaces, street trees, residential gardens, golf courses and other vegetation involved in the urban environment (Pietsch 2012; Nur Syafiqah et al. 2018; Samad et al. 2020)".

Among urban green space valuation studies, the exact definition of UGSs differs. For example, Daams et al. (2016) use a broad definition of urban green spaces that includes all features that could be perceived as nature like parks, recreation areas, forests, agricultural areas and water bodies. Some studies follow this definition and include water bodies as urban green spaces (Panduro & Veie, 2013; Czembrowski et al., 2019; Daams et al., 2019). At the same time, other studies exclude water bodies and only focus on green natural areas on land (Czembrowski et al., 2016).

No standard definition of urban green spaces is used in all studies. However, all definitions have in common that they include some form(s) of nature in urban areas. Which forms are included depends on the same research goals, available datasets and operationalization method(s). Therefore, it is relevant for this research to investigate how urban green spaces can be classified and how these different classifications were derived.

2.1.2 UGS Classification

The concept of 'urban green spaces' lends itself to being understood in various ways. Over the years, different classifications of urban green spaces have been used to determine the effect of urban green spaces on real estate prices. According to the type of Alex & Jim (2010), urban green spaces should be open spaces that consist of greenery. This included areas like parks, sports stadia and playgrounds. At the same time, the study from Eom et al. (2019) classifies urban green spaces as parks, rivers or forests. Most classifications overlap to some extent, but in former studies, no classification was found that has been used consistently in other research.

However, urban green space classifications always contain a combination of the following categories (Conway et al., 2008; Alex & Jim, 2010; Panduro & Veie, 2013; Kolbe & Wünstermann, 2014; Sander & Zhao, 2015; Czembrowski et al., 2016; Daams et al., 2016; Xu et al., 2016; Engström & Gren, 2017; Czembrowski et al., 2019);

- Parks
- Sport fields
- Playgrounds
- Golf courses
- Agricultural areas
- Forests
- Water bodies
- Cemeteries
- Farrow land
- Lawns
- Nature reserves

Different arguments for the choice of classification can be found. The studies of Kolbe & Wüstermann (2014) and Xu et al. (2016) used the classification from the land use maps provided by governmental agencies. The European Urban Atlas argues that green urban spaces are public areas that are predominantly used for recreational purposes, while the land use map of Beijing that was used by Xu et al. (2016) is based on China's land use classification system. More studies use a land use map as a GIS database for their research, but Xu et al. (2016) point out a serious drawback of this data selection. Most of these maps are based on the local governmental land use classification system. This means that it contains categories like commercial, residential land and (urban) green spaces. However, these 'first-order' land use categories like residential areas are also 'second-order' categories like residential land plots, parking spots and green spaces. Unfortunately, the land use data only accounts for the 'first-order' categories, while a residential area might consist of 25 percent of green spaces. In addition, the choices made in drawing up the land use map may also affect the representativeness of the data. For example, the dataset from EUA only accounts for parks larger than 25 hectares. Keeping these drawbacks in mind is important because these data quality issues could result in wrong estimations.

Not all studies use the classification of their dataset. In most cases, the exact subject of study and methodological arguments determine the classification of urban green spaces that is used. The study of Engström and Gren (2017) was only interested in parks as the UGS unit of study. However, they excluded parks smaller than 40,000 m^2. They were aware that small patches of urban green could also provide value, but large parks provide more and different recreational usages, which should lead to a greater effect on house prices. Therefore, it would be easier to measure the effect and distinguish it from other neighbourhood amnesties. Others argue that water bodies should be added to the classification because, in most cases, water is closely connected to (urban) green (Panduro and Veie, 2013; Czembrowski et al., 2019). Daams et al. (2019) argued that agricultural areas should also be added to their UGSs classifications because real estate buyers might evaluate them as attractive although they are cultivated lands.

On the other hand, a few studies have tried to account for the heterogeneity of urban green spaces with alternate classifications compared to the list above. Panduro and Veie (2013) classified urban green spaces into eight categories. They assumed that UGSs are heterogenous based on the service that they provide. To distinguish the different categories, internal, external and social accessibility, maintenance and neighbouring land use were used. They found that urban green spaces with high recreational potential have a higher value impact on real estate prices. The study of Czembrowski et al. (2016) was interested in how the residents perceived urban green spaces and if the UGSs that are perceived as more likeable have a greater effect on the real estate prices as well. Therefore, they combined two datasets, the land use map of Lótz and a SoftGIS database that contained the perception of urban green spaces derived from a geo-questionnaire. They argued that combing these two datasets made it possible to classify UGSs in categories that represent how they are perceived, like 'high perceived value', 'no net preference' or 'low perceived value'. In addition, it allowed detecting urban green patches which are not included in the land use maps.

Remarkable is that these studies include different types of urban green 'areas' in their classification. But none of them included street trees, residential gardens or other street vegetation, which also fall within the definition of urban green spaces.

This can be explained by the fact that most studies use land use data that do not account for them. However, there are solutions to overcome this problem. Conway et al. (2010) used a georeferenced aerial photo to account for urban green. They manually digitalized all the green cover like lawns, sports fields, parks, cemeteries and trees. However, this technique is time-consuming and error sensitive. Another approach was used by Sander & Zhao (2015); in addition to a regular land use map, they added another database that encompasses all trees to account for three coverage in the study area. More hedonic pricing studies have been conducted that account for the effect of trees on the real estate price (Dombrow et al., 2000; Holmes et al., 2006). However, in these studies, the focus was on trees, particularly without a link to urban green spaces. Therefore, it should be noted that in most studies, trees or other single vegetation are not part of the urban green space classification. The omission of these components of urban green spaces within the most classification of UGS can certainly be seen as a deficiency within UGS valuation research. Because of these classifications, a noteworthy part of urban green spaces is not considered.

2.1.3 Operationalization of UGS

This section discusses how to control for urban green spaces in the hedonic price model. To measure the effect of urban green spaces on the transaction price of dwellings, the UGSs should be turned into a measurable variable. Therefore, this section describes how previous studies operationalized UGSs into a model variable.

Operationalization based on distance

Values of environmental amenities can be captured in hedonic pricing according to the proximity principle described by Crompton (2001). This proximity principle assumes that dwellings near environmental amenities like beaches, parks and golf courses are likely to be more expensive and sold for a higher price. Therefore, many studies use a distance measure to take UGS into account.

Most UGS valuation studies operationalize UGS as a continuous variable that makes use of the Euclidean distance (Daams, 2019; Czembrowski et al., 2016; Daams et al., 2016; Xu et al., 2016; Kolbe & Wüstemann, 2014; Panduro & Veie, 2013; Dekkers & Koomen, 2008). The Euclidean distance is a mathematical concept also known as the straight-line distance (Lu, Charloton & Fotheringham, 2011). It measures the distance between two objects using a straight line. However, there is also criticism on the use of Euclidean distance. It is hard to believe that the degree of connectedness between two points is optimally represented by a straight line (Lu, Charloton & Fotheringham, 2011). Especially in an urban context, where areas are divided by buildings, fences and rivers are connected internally by streets and bridges, a straight line does not seem to be a fair representation of reality. According to Engström & Gren (2017), 'street morphology' and the 'park entrances' determine the accessibility of green amenities in an urban context. Both are not taken into account when the Euclidean distance measure is used to calculate the distance. Therefore, Engström & Grenn (2017) discuss two alternatives to overcome this problem.

First, they discuss the 'pedestrian network' distance, also known as the 'walking' distance. This measure is the shortest distance between the property and destination, calculated along a street network. The pedestrian network distance seems to tackle the criticism of the Euclidean distance measure and is therefore more appropriate for UGS valuation research. Czembrowski et al. (2016) used the network distance to operationalize the walking distance from dwellings to the entrances of UGS amenities and clusters. At the same time, the study of Czembrowski et al. (2019) argued to use the network distance as they were interested in measuring the price effects of UGS benefits that could only be derived when the users are physically present in the UGS amenity. However, Czembrowski et al. (2016) argue that the Euclidean distance measure is a better choice for some research cases. For example, when the research is interested in measuring the price effect of 'passive externalities', from living in the proximity of a UGS amenity, like noise and smell pollution, air quality, or microclimate improvement, the Euclidean measure is a better fit to express the distance. At the same time, the pedestrian network measure is better to describe the relationship of people 'visiting UGS' for physical activity, social or recreational purposes. The second alternative, discussed by Engstrom & Gren (2017), is called the 'axial line step distance'.

This measure adds cognitive environment data to the analytical framework to make it a 'people-based' distance measure. It allows adding data or factors that could be perceived as important for the accessibility of UGS amenities like, 'is the UGS fenced or not?' or 'is the UGS perceived as a safe place to visit?'. This measure could help explain extremely low values of properties located close to a UGS, measuring accessibility and providing in-depth information that could be used for sustainable planning and design. However, to the knowledge of this thesis, this distance measure has not been used so far by urban green valuation studies. Probably, because the data availability for these specific topics, which are also site-specific, is low and this quantitative data collection takes a lot of time.

Remarkable is that only a few studies mention between what 'points' the distance is determined. For example, Czembrowski et al. (2016) measured the distance between the dwelling and the entrance of a UGS. They argue that using the entrances of UGSs instead of their centroids further improves the measurement accuracy. At the same time, another study by Czembrowksi et al. (2019) calculated the distance between the dwelling and the closest border of a UGS. The choice between which points the distance is measured obviously influences the value effects. Therefore, the modelling choices should be argued. But in most studies, it was not found between what points the distances have been calculated, let alone an argument for the choice of measurement.

The last method that could be used to take the proximity of UGS into account is using buffers with a distance radius. With the help of GIS software, like ArcGIS or QGIS, buffers with different radius can be drawn around each UGS, and apartments overlapping this buffer can be assigned with a value. This dummy variables can be made for various buffers, e.g., 100m, 500m or 1000m. Various studies made use of this operationalisation method (Dekkers & Koomen, 2008; Daams et al., 2016; Daams et al., 2019).

Operationalization based on density

Studies also take the coverage of UGS surrounding the dwelling into account to control for the price effect of UGS (Conway et al., 2008; Sander & Zhao, 2015; Kolbe & Westermann, 2014; Czembrowski et al., 2016; Daams et al., 2016; Krekel et al., 2016; Czembrowski et al., 2019). According to Daams et al. (2016), property buyers are not only interested in the distance to the nearest urban green space but especially take the amount of green surrounding the dwelling, which contains 'scenic value' into account. So, using a density variable instead of a distance variable can provide other insights into the data. The density can be operationalized by defining a certain distance radius around the dwelling and measuring the amount of urban green space within this radius (Krekel et al., 2016). Figure 2 shows a simple illustration of the operationalization.

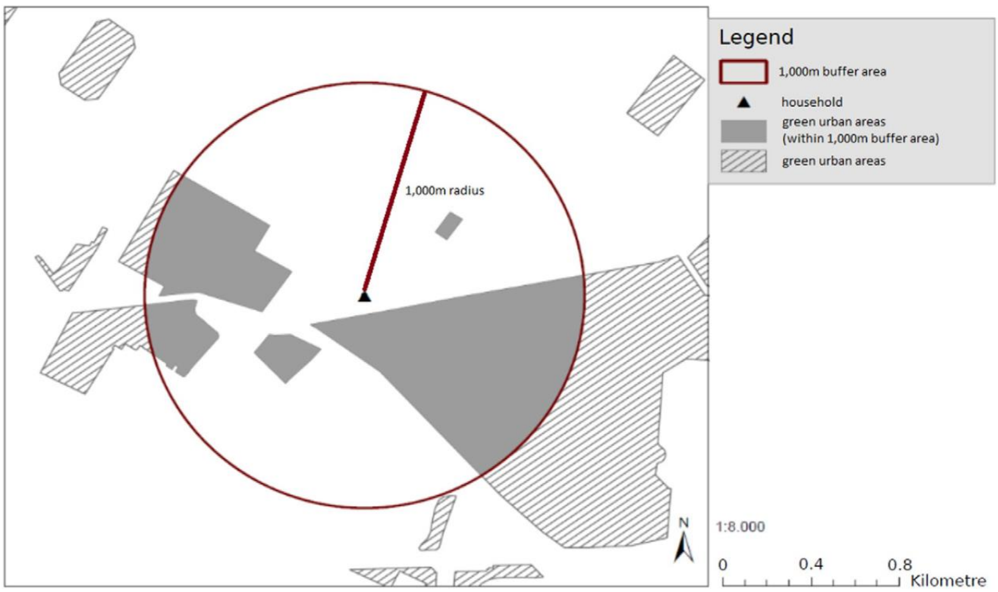


Figure 2: Definition of density (Krekel et al., 2016)

Kolbe & Wüstermann (2014) used distance variables and three density variables (buffers of 500, 1000 and 2000 metres) to account for urban green spaces. Their study found that both distance and the coverage variable showed a positive price effect on the transaction prices of real estate. This indicated that residents indeed do not only prefer to live close to urban green spaces but also like to be surrounded by urban green spaces. According to Krekel, Kolbe & Wüstemann (2016), both density and distance variables should be used for urban green spaces. It provides the opportunity to compare the outcomes of both variables. Therefore, it serves as a robustness check given that the assumption about density might be contrary to the distance outcomes.

2.1.4 Innovation in urban green space valuation research

Liebelt, Bartke & Schwarz (2018) state in their recommendations that it would be valuable to analyse different types of UGS instead of aggregating them into one category. As different UGS amenities provide other ecosystem services and thus different benefits for residents, that could have different impacts on the residential prices. They follow the suggestion of Panduro and Veie (2013), who stated that UGS are not a heterogeneous good. Since their suggestion, a few studies did try to account for heterogeneity. Kolbe & Wüstemann (2014) distinguished the three UGS types, parks, farmland and forest, in their research and found that parks and forests result in positive price effects while farmland showed an inverse relation. Xu et al. (2016) took various spatial characteristics like the richness, shape and fragmentation of UGS into account, while the studies from Daams et al. (2016; 2019) distinguished urban green spaces that are perceived as attractive. The study of Czembrowski et al. (2019) followed by categorizing Stockholm's urban green spaces according to their use values, social meanings and multifunctionality. Still, plenty of options are available to distinguish the different types of UGS.

Using the proximity to urban green spaces and accounting for the amount of UGS are straightforward approaches in the literature to control for UGSs. However, using new, different or additional methods might lead to other insights. Engström and Gren (2017) created a heatmap in addition to the standard analysis of their spatial econometric models. The heatmap contained information about the house prices of each neighbourhood. By overlaying the heatmap with a map of UGS, an in-depth view of the price effect from urban green spaces can be created. Therefore, the heatmap could provide evidence about which specific UGSs have a positive price effect on their surrounding dwellings and which do less or not.

Krekel & Kolbe (2020) suggest other innovative ideas that could give more profound insights into the price effects of UGSs. According to them, different methods exist that could be used to account for the arrangement and structure of UGS. The number of UGS patches or patch density (the number of patches divided by the total land area) could be calculated to account for the UGS arrangement. Structural details of UGSs can be taken into account by looking at the total edge (accounting for all UGS borders segments with other land use types) and edge density (dividing the length of all UGS border segments by the entire landscape area). From an urban planning perspective, this could provide interesting insights into what kind of UGS structures are preferred. However, two essential conditions must be met to put this into practice. The exact GIS coordinates from the UGS and dwellings must be available, while the fragmentation measures must be calculated separately. In addition, they plead for using satellite imagery on chlorophyll activity, which can be used to measure the health of urban green and provide the opportunity to take the quality of UGSs into account. Organizations like the World Data Centre for Remote Sensing of the Atmosphere (WDC-RSAT) and the European Space Agency (ESA) share satellite imagery data. According to Krekel & Kolbe (2020), it is preferable to use datasets that consist of 'land coverage' data instead of 'land use'. Datasets based on land cover provide information about the actual observable usage, while with land use data, the urban space classified as a park is also used as one. The downfalls of using 'land use' data are already stressed by Xu et al. (2016).

2.2 Dutch UGS studies

As stated in section 1.5, only a few Netherlands hedonic pricing studies took urban green space into account. Details about the studies can be found in Table 2. Luttik (2000) conducted the first hedonic pricing study, with an acceptable research size, in the Netherlands that was interested in the value effect of UGS on residential real estate. The study was not only focused on the urban green at a neighbourhood level, but it was also interested in the price effects of UGS surrounding the place of residence and attractive green at the edges of the neighbourhood. The cities of Emmen, Leiden, Zoetemeer, Veenendaal, Tilburg and the municipality of Utrecht were used as a research case. Dekkers & Koomen (2008) studied the value of regional open space in the Randstad region with the help of hedonic pricing. Although this study is not focused on urban green spaces, it has been included in the literature review as the concept of open spaces is closely linked to UGS. Open spaces are free of buildings and other proof of human presence and are used for recreational purposes (Dekkers & Koomen, 2008). In addition, the study operationalized open spaces by selecting 'open-land use types' that could also be indicated as UGS; agriculture, nature, water and recreation. A regular OLS model was used to estimate the price effects in Amsterdam, Leiden and het Gooi. However, an SAC and SEM model was also estimated to test for spatial dependency. The study does not mention what spatial weight matrix was used. As discussed in section 2.1.2, Daams et al. (2016) studied the price effect of 'attractive' (urban) green spaces on residential property prices. Their study used an OLS model with transaction data from 203.344 single-family properties throughout the Netherlands. They used sub-market fixed effects instead of a spatial econometric extension to control for spatial property price dependency. The study of Daams et al. (2019) focused on the impact of attractive UGS on the house prices in the urban core of Amsterdam. Instead of the standard hedonic pricing model, they used a Pseudo-Repeat Sales (PS-RS) model. This model uses spatial differencing to reduce bias from (spatial) omitted variables. It is similar to a standard hedonic pricing model with spatial fixed effects. However, it is slightly stricter. In the case of this study, the PR-SR model is applied to differentiated houses that are paired to the same neighbourhood level (e.g., zip code areas). Therefore, the model cancels out the local similarity in prices but also a local similarity in the unobserved characteristics of (paired) houses. Examples of these unobserved characteristics are the neighbourhood's architecture style or average maintenance level (Daams et al., 2019).

Luttik (2000) included variables to account for the view on different UGS amenities. With the help of a buffer of approximately one to two kilometres, the presence of parks or lakes was taken into the analysis. The study also contained variables that account for green outside the urban region. However, those are not discussed because they are beyond the scope of this research.
This study was not able to use GIS techniques like nowadays. Therefore, information about the environmental and locational factors was drawn from maps and by visiting each of the 3000 dwellings from the research sample. Dekkers & Koomen (2008) distinguished three different spatial scales. First, open spaces visible from home indicate if the object is located within or adjacent to UGS. This was operationalized by a dummy variable that indicates if the object is within a distance of 0 to 10 metres from an open space. Second, open spaces on a local level are operationalized with four distance dummy variables with a scale between 25 and 100 metres. This level indicates the presence of small or large patches of open space within walking distance that could be used on a daily base for walking the dog or something similar. The third spatial scale presents the regional level and indicates large open spaces, which should be at least 500 hectares or greater. It is operationalized by a variable representing the Euclidean distance to the nearest regional open space in kilometres. This spatial scale accounts for open spaces that can be used for outdoor activities and escape the urban landscape. Daams et al. (2016) used a Euclidean distance measure to account for UGS. The study observed that the mean distance was 4.63 kilometres and decided to discretise this in different distance interval variables from 0 to 8 kilometres. In addition, buffer variables that account for the amount of 'attractive' UGS and 'non-attractive' UGS were added. The study of Daams et al. (2019) used the same method to operationalize the UGS variables. However, a smaller distance interval was used between 0 to 1.25

All the UGS variables in the study of Luttik (2000) show a positive effect on the value of the residential real estate. Outstanding is the value increase of 12% due to the presence of a park.

The view of open space also leads to a great price increase of 8 to 10 %. Dwellings that are located next to water can account for a price increase of 6 to 12 %. However, considering the data gathering of this study, this study's results could be affected by researcher bias.

The study result of Dekkers & Koomen (2008) shows that real estate buyers appreciate living in or close to (small) patches of UGS. The dwelling price increased from 4 to 8 % when the object was located within 0 to 10 metres of a patch of UGS. Positive price effects have been observed for the presence of UGS within 50 metres from the dwelling. Remarkable is that a closer distance to regional open spaces seems to decrease the house price. This result applied to all study cases and all model forms. It could indicate that real estate buyers do not want to pay more for a house close to such a large green space. However, the high spatial error term also indicates that unobserved spatial characteristics might influence the price, like local noise disturbance or a high local crime rate. To find answers to these uncertainties, Dekkers and Koomen (2008) recommend further research to investigate local differences in UGS valuation. Daams et al. (2016) found that the proximity to 'attractive' UGS results in higher dwelling prices. The nearest proximity indicator, within 0.5 km, found a positive price effect of 15%. While the presence of UGS within 6 to 7 km still could result in an increase in property price by 1.5%. Compared with the study results of Dekkers & Koomen (2008), the results from Daams et al. (2016) confirm that property buyers are prepared to pay more for dwellings that are located close to (attractive) urban green spaces. However, in contrast, this study also shows that property owners also appreciate living near green space extended over a larger distance. This could be an important argument for spatial policymakers to invest in green spaces near residential areas (Daams et al., 2016). The study of Daams et al. (2019) also found that proximity to 'attractive' UGS gave higher dwelling values. For dwellings located within 0.25 kilometres from 'attractive' UGS, an increase of 7 to 9 % in the dwelling price was estimated. A rise of 2% was found for dwellings located within 0.75 to 1 kilometer from 'attractive' UGS. However, in contrast to Daams et al. (2016), this price increase stagnates after 1 kilometre.

Concluding, taking all study results together, it appears that especially urban green spaces in the direct surroundings of dwellings lead to a high price effect. The negative price effect for UGS at a distance seems to be negated by the other study results. However, the study of Daams et al. (2019) that has been carried out on a neighbourhood level states that the positive price effect stops after 1 kilometre.

Reference	Year	Study case(s)	Data dependent variable	UGS variable	Price effect
Daams et al.	2016	Netherland	Transaction data of 293.621 sales of single-family properties between 2009 – 2012 in the Netherlands	Distance to attractive UGS within 0. – 0.05 km	9 – 15 %
				Distance to attractive UGS within 0.5 – 1 km	6 – 10 %
				Distance to attractive UGS within 1 – 2 km	3 – 9 %
				Distance to UGS attractive within 2 – 3 km	2 – 8 %
				Distance to attractive UGS within 3 – 4 km	2 – 7 %
				Distance to attractive UGS within 4 – 5 km	2 – 5 %
				Distance to attractive UGS within 5 – 6 km	0 – 4 %
				Distance to attractive UGS within 6 – 7 km	0 – 1,6 %
				Distance to attractive UGS within 7 – 8 km	0 %
				% Attractive UGS within 7 km ring	0 %
				% non – attractive UGS within 7 km ring	0%
Daams et al.	2019	Amsterdam	Transaction data of 293.621 sales of single-family properties between 2009 – 2015 in the urban core of Amsterdam	Distance to attractive UGS < 0.25 km	7 – 9 %
				Distance to attractive UGS 0.25 – 0.50 km	2 %
				Distance to attractive UGS 0.5 – 0.75 km	0 – 1 %
				Distance to attractive UGS 0.75 – 1.0 km	0%
				Distance to attractive UGS > 1.25 km	0%
Dekkers & Koomen	2008	Amsterdam, Leiden, het Gooi	Transaction data of 36.848 sales of fourteen different house types between 1997 and 2001	View of UGS (0 – 10 metre)	4 – 8 %
				Presence of local open space (10 – 25 metre)	3 – 8 %
				Presence of local open space (25 – 50 metre)	0 – 6 %
				Presence of local open space (50 – 75 metre)	-3 – 2 %
				Presence of local open space (75 – 100 metre)	-2 – 1 %
				Distance to regional open space (km) (Euclidean)	-9 - -1 %
Luttik	2000	Emmen, Leiden, Veenendaal,Tilburg, Utrecht & Zoetemeer	Transaction data of 3.000 house sales between 1989 – 1992	View of green strip	4 – 5%
				View of park	7 – 8%
				View of canal	4 – 5%
				View of lake	8 – 10%
				View of open space	6 – 12 %
				Presence of park	12%
				Presence of lake	5 – 7 %

Table 2: Overview Dutch UGS valuation studies (Luttik, 2000; Dekkers & Koomen, 2008; Daams et al., 2016; Daams et al., 2019, own edit)

2.3 Theoretical framework

Literature shows that hedonic pricing is the dominant method for estimating the monetary value of ecosystem amenities like UGSs. According to the hedonic pricing theory, multiple housing characteristics determine the transaction price of real estate (Rosen, 1974). By estimating the coefficients for each characteristic in a multiple regression analysis, the marginal willingness to pay for each characteristic can be found (Engström & Gren, 2017). As the goal of this research is to put a monetary value on UGSs, this research is interested in the coefficients from the different UGS characteristics.

According to the proximity principle from Crompton (2001), many studies assume that real estate buyers are willing to pay more if the dwelling is located within the proximity of UGSs with recreational value. In other words; real estate buyers value a dwelling higher when they have access to an UGS that present high recreational value like parcs or golf courses. Other studies stress (Kolbe & Wüstermann, 2014; Krekel, Kolbe & Wüstermann, 2016) that UGSs do not only contain recreational value but also present scenic value for real estate buyers. Their research proved that real estate buyers are willing to pay more whenever the density of urban green surrounding the dwelling is higher.

However, in order to obtain valid estimates, other characteristics, next to the UGSs characteristics should be included into the analysis. Major part of the dwelling prices can be attributed to their structural characteristics (e.g., living area, number of rooms, plot size), while former research proved that locational characteristics (e.g., distance to transport hubs, number of available amenities) also influence the price. In addition, the model should control for spatial dependencies to obtain valid estimates. Derived from Tobler (1970) first law of geography it is suggested that data collected from different units in space are often correlated. To deal with these spatial dependencies the hedonic pricing model can be extended to a spatial lag model that also takes the value of neighbouring observations into account. This and the concept of hedonic pricing applied in UGS valuation research will be discussed in the next methodology chapter.

Concluding, this thesis is focused on putting a monetary value to urban green spaces by estimating a hedonic price function. During the operationalization process this thesis will account for the fact that UGS do provide recreational and scenic value that should be modeled differently. In order to obtain reliable estimates for the UGS variables, it is important to take the spatial dependency, structural and locational characteristics for each dwelling into account. The next methodology chapter will elaborate on this. The conceptual model of this thesis can be found in Figure 3.

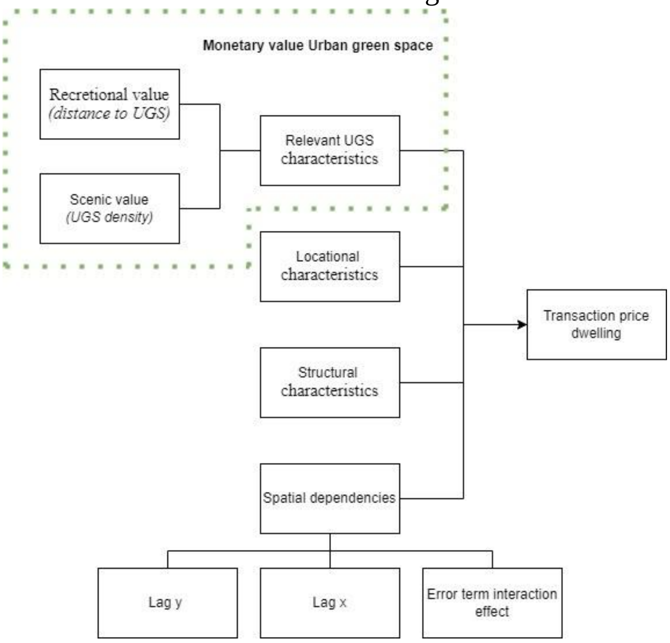


Figure 3: Conceptual model

2.4 Summary

This chapter first discovered the concept of UGS by evaluating its definition and how it is conceptualized. Next, it discussed the state of art concerning the operationalization of urban green space, where operationalization based on distance or density are the leading principles. According the proximity principle of Crompton (2001), UGS contain recreational value that residents get by visiting them, while Daams et al. (2016) state that real estate buyers mainly value the scenic side of UGS that surround the dwelling. The usage of heatmaps, ‘land coverage’ instead of ‘land use’ data and UGS ‘arrangement’ or ‘structure’ variables are suggested as possible innovations within the field. The second section contains a literature discussion about the findings of previous Dutch UGS valuation studies. In the third, and final section, of this chapter the finding of this literature review are summarized and convert into the theoretical framework of this research.

CH 3 Methodology

This chapter describes the methodology of this research. First, the concept of hedonic pricing in relation to urban green space valuation is explained and followed by an extensive explanation of the concept of spatial econometrics. This starts with a brief discussion of why it was deemed necessary to use spatial econometrics, followed by an explanation of its core concepts. Next, the concept of spatial interaction effects is discussed and how spatial lag models could capture these effects. Then the different types of lag models are discussed and the extent to which they capture local or global spillovers. The following sections explain what a spatial weight matrix is and how it can be used to capture local similarity, including a brief discussion about the different methods that can be used to specify the weight matrix. Finally, the last section of this chapter explains the spatial dependency statistics that can be used to determine if spatial autocorrelation is present in the data.

3.1 Hedonic pricing

Hedonic pricing is an old theory developed by Rosen (1974) based on the foundation of Lancaster (1966). The technique is frequently used in real estate and property research to study the impact of a variety of factors on housing prices. The idea is that a certain number of characteristics will determine the price of the (real estate) object. In other words, the price can be described based on the implicit price of the different characteristics. This represents the buyer's marginal willingness to pay for that specific attribute (Engström & Gren, 2017). Therefore, the hedonic pricing is interesting for ecosystem and urban green space valuation studies as well because the price of each characteristic can be determined by its marginal contribution to the overall price (Nappi-Choulet, Maleyere & Maury, 2007).

The hedonic pricing model is formulated as a multiple regression model and often appears in the following form:

1) $$Y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i$$

where Y_i represents the value of property i. Parameter β_0 stands for the constant while parameter β_j represents the independent variable j with value x_{ij} as the associated coefficient in a $N \times K$ matrix of explanatory variables. At last, ε_i stands for the error term that contains the price effect of omitted variables on property i. An important assumption is that the error term is independent across the observations and normally distributed with constant variance and a mean of zero (Fuerst, van de Wetering & Wyatt, 2013). This error term is the part of the model that could not be explained. The residual can be caused by missing variables, wrong model specification or errors in the variables (Cox, 2017).

In UGS valuation, the presence of urban green spaces is seen as one of the characteristics that determine the property's value. Besides that, no standard set of independent variables is used in every study. However, three main categories of characteristics can be distinguished that are used as independent variables in most UGSs valuation studies. Most studies account for (1) structural characteristics like floor area, number of (bath)rooms etc., (2) location characteristics like access to public transportation, distance to amenities etc. and (3) UGS characteristics. So, the general hedonic pricing used for UGSs valuation often appears in the following form (Czembrowski et al., 2019):

2) $$Y_i = \beta_0 + \sum_{j=1}^{J} a_j S_{ij} + \sum_{m=1}^{M} \beta_m L_{im} + \sum_{n=1}^{N} \gamma_n U_{in} + \varepsilon_i$$

where S_{ij} represents the structural variables, L_{im} the locational variables and U_{in} the urban green space variables with values α_j, β_m and γ_n as their respective coefficients for dwelling *(i)*. The other parameters follow the same specification as model 1.

Although hedonic pricing is the dominant method for UGSs valuation, it has some serious shortcomings. Hedonic pricing does not take spatial and time dependencies into account, which could lead to heteroskedastic issues or bias in the coefficient estimates (Hilgers, 2018). The following section will elaborate on the concept of spatial econometrics that will be used to account for spatial dependency in the data.

3.2 Spatial econometrics

In the section the concept of spatial econometric is discussed. First it explains why spatial econometric modelling is important for this research. The second section explains the spatial econometric core concepts while the third section explains how spatial interaction effects can be modelled. In the fourth section an overview of the different lag models is given while the fifth section explains the difference between local and global models. In the last section of this chapter the different spatial weight matrices are explained.

3.2.1 Why spatial econometric modelling

As Hilgers (2018) stated, the traditional hedonic pricing model does not take spatial and temporal dependencies into account, which could result in heteroskedasticity and research bias issues. Extending the traditional hedonic pricing model using spatial econometric modelling can help to take the spatial – and time dependencies into account. Elhorst (2017) defines a spatial econometric model as:

'A spatial econometric model is a linear regression model extended to include spatial interaction effects.'

For most studies and models, the spatial units of observation, e.g. individual houses for hedonic pricing, are likely to have differences in their background variables. Some of these differences are easy to take into account, like most structural characteristics of a dwelling. But, space-specific and time-invariant variables that affect the dependent variable are usually harder to measure. For example, one unit is located in a neighbourhood with high crime rates or low average income, while another spatial unit might differ substantially. This also accounts for time-related effects; the first sale of a unit might occur in an economic recession while the sale of another unit was in times of economic expansion. However, this study is focused on controlling for spatial effects.

The motivation behind spatial econometric models can be found in the first law of geography from Tobler (1970), which states; 'everything is related to everything else, but near things are more related than distant things.' From the modelling point of view, the first law of geography suggests that data collected across different units in space are often correlated, this thesis will call these 'spatial interaction effects'. The strength of such dependence is usually determined by the proximity of two units in space. Therefore, the use of spatial econometric models enables us to test whether there is an existence of spatial interaction effects and related to that spatial spillover effects (Elhorst, 2017). Spatial econometric models allow finding potential relationships or interactions between different spatial units like cities, regions or countries (Sarrias, 2020).

Spatial econometric models extend standard economic models like hedonic pricing to address spatial interaction and heterogeneity. Therefore, some UGSs valuation studies did use a spatial econometric extension to the hedonic pricing model. (Conway et al., 2008; Dekkers & Koomen, 2008; Hoshino & Kuriyama, 2009). Before explaining the different types of extensions, this research will elaborate further on the core concepts of spatial econometric modelling.

3.2.2 Core concepts of spatial econometrics

Before this thesis elaborates on spatial econometric modelling, it is important to understand the core fundamentals of spatial econometrics. Spatial econometrics is focused on dealing with nonstationary in the dataset, where stationary refers to a dataset that includes variable values that do not change in mean, variance and location dependency through space (Meik & Lawing, 2017). The definition of spatial econometrics explains the essence of this discipline (Anselin, 2006):

'Spatial econometrics is a subfield of econometrics that deals with spatial interaction (spatial autocorrelation) and spatial structure (spatial heterogeneity) in regression models for cross-sectional and panel data.'

Anselin (2006) distinguishes two spatial phenomena; spatial autocorrelation and spatial heterogeneity. This thesis is focused on spatial autocorrelation and dealing with spatial interaction between the dwellings, which are the unit of analysis. Related to spatial autocorrelation is spatial dependency. Spatial dependency points out the degree of spatial autocorrelation between individual values observed in a geographic area (Crawford, 2009). Miller, Franklin & Aspinall (2007) defined spatial dependency as:

'the property of variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for randomly associated pairs of observations'

While spatial autocorrelation is defined as (Haining, 2001):

'*Spatial autocorrelation is the term used to describe the presence of systematic spatial variation in a variable and positive spatial autocorrelation is the tendency for areas or sites that are close together to have similar values, which is most often encountered in practical situations '*

Without spatial autocorrelation, it can be assumed that the spatial allocation of the observation is random (Bouayad Agha & Bellefon, 2018). But whenever spatial autocorrelation is present, it is very likely that the OLS estimates are biased and violate the assumption of homoscedasticity. Global autocorrelation statistics can be used to check for spatial dependency in a geographically referenced dataset. The Moran's I (*I*) test measures the overall spatial autocorrelation in a dataset (Hijmans & Ghosh, 2021). A positive significant test statistic indicates positive autocorrelation in the data, while a negative outcome informs that negative autocorrelation is present. While the Moran's I test is a more general statistic, that only informs that spatial autocorrelation in the dataset is present, the Lagrange-Multiplier (LM) tests can be used to diagnose spatial dependence in the data causes the spatial autocorrelation. The LM-error test tests for spatial correlation in the error term, while the LM-lag test can be used to explore the presence of spatial correlation in the lagged values of the dependent variable (Hijmans & Ghosh, 2021).

The practical utility of spatial autocorrelation can be found in the fact that it helps to outline processes that cause spatial structuring of activities but are hard to quantify or even unobservable. The following section will explain how adding 'spatial interaction effects' to the OLS model can help control for spatial autocorrelation.

3.2.3 Modelling spatial interaction effects

When spatial autocorrelation is present in the data but not controlled for in the regression model, then it is possible that the coefficient estimated will be biased. Therefore, the standard hedonic pricing model specification should be extended with a mathematical expression that deals with the proximity of units across space. Spatial lag models are often used to account for spatial autocorrelation, while the geographically weighted regression (GWR) model can be used to control for spatial heterogeneity. According to Elhorst (2017), the primary purpose of using a lag model is to account for the existence of spatial interaction effects and related to that, spatial spillover effects.

Spatial spillover effects are defined as (Elhorst & Vega, 2017):

'*Spatial spillover effects are the marginal impacts of a change to one explanatory variable in a particular cross-sectional unit on the dependent variable values in another unit.'*

In the case of hedonic pricing, it is plausible to think that a change in the explanatory variable of a specific observation (i) does not only affect the value of that particular observation (i) but also the value of other neighbouring observations ($\neq i$). But it is also known that real estate brokers often determine the transaction price of a dwelling by looking at the past transaction prices of neighbouring properties (Björklund, Dadzie & Wilhelmsson, 2006). This indicates that a change in the dependent variable (the transaction price) of dwelling *(i)* leads to a price increase in the neighbouring properties. Additionally, omitted variables could also cause a price increase in neighbouring observations. In order to add a parameter that accounts for spatial interaction between the units of analysis to the model, it is necessary to create a spatial weight matrix (W) which defines the spatial structure in the dataset. Elhorst & Vega (2017) define spatial weight matrices as:

"Spatial weights matrix (W) symbolises the spatial arrangement of the cross-sectional units in the sample."

The spatial weight matrix describes the proximity of two observations in space, different approaches can be used to specify this relationship. Before these approaches are discussed, the next section will first elaborate on the different types of spatial interaction effects that can be added to the model.

3.2.4 Lag models

Spatial interaction effects can occur in three ways, known in spatial econometrics literature as an endogenous -, exogenous – or error term interaction effect (Elhorst, 2017).

1) Endogenous interaction effects, also known as lag *y*, refer to the spatial dependencies in the dependent variable. With the endogenous interaction effect, the correlation between the dependent variable of an observation (i) and the dependent variables from neighbouring observations (j) is measured.

2) Exogenous interaction effects, also known as lag x, refer to spatial dependencies in the independent variables. It measures the effects of independent variables from neighbouring observations (j) on the dependent variable of observation (i).

3) Error term interaction effects, also known as spatial interaction effects, measure if the residuals or error term of neighbouring observations (j) affect the dependent variable of observation (i). It indicates if there is a correlation between observations given possibly similar unobserved characteristics.

Different lag models can be made by adding spatial interaction effects to the standard OLS regression model. The general nesting model (GNS) includes all three forms of spatial interaction effects. However, in practice, the GNS model is hardly used because it suffers from overfitting. In addition, the model is criticised because it lacks formal proof under which conditions the model's parameter should be identified (Elhorst, 2017). The mathematical equation of the GNS model takes the following form:

$$
\begin{gathered}
Y = \beta X + \rho W y + \theta W X + \mu \\
\mu = \lambda W \mu + \varepsilon
\end{gathered}
\tag{3}
$$

Where Y_i denotes an *N X 1* vector consisting of one observation on the dependent variable (i.e. transaction price) for every observation (i.e. dwelling) in the data sample (i = 1, …, N). X represents an *N x K* matrix of explanatory variables associated with the parameters β contained in a *K x 1* vector. The variable Wy denotes the interaction effects among the dependent variables, WX the exogenous interaction effects among the explanatory variables and $W\mu$ the interaction effects among the error term of neighbouring units. The scalar parameters p and λ measure the strength of the dependence between the dwellings, while θ, like β is a K x 1 vector for the response parameters (Elhorst & Halleck Vega, 2013).

However, for the reasons stated above, in practice, empirical studies only use models that include one or two interaction effects. Table 3 gives an overview of the added interaction effects and mathematical equation for the different lag model types. The spatial autoregressive (SAR) model only includes the endogenous interaction effects by adding ρWy to the model. Where parameter ρ shows the size of the spatial interaction effect, W defines the neighbours and y stands for the dependent variable from neighbouring observations. So, in the case of hedonic pricing, the SAR model accounts for interaction effects between transaction prices of neighbouring units.

The spatial lag of X model (SLX) only accounts for the exogenous effect through including θWX in the model. Parameter θ accounts for the coefficients of the exogenous interaction effects and X for explanatory variables. Adding exogenous interaction effects to a hedonic pricing model allows the average value of the characteristics of neighbouring units ($\neq i$) help to explain the transaction price of dwelling (i). This also allows interpreting the marginal effects of individual characteristics of neighbouring houses.

It is called a spatial error model (SEM) when $\lambda W\mu$ is added to account for the error term interaction effects. Parameter λ stands for the coefficients of the error term interaction effects and μ for the neighbouring error terms. Including this in the hedonic pricing model allows controlling for the interaction effect of omitted variables in neighbouring units. Socio-economic factors, e.g. crime rates or education levels, are often neglected in HP models, while it is expected that they influence housing prices. They can also vary across locations, indicating that they should be controlled for.

Model	**Lag x**	**Lag y**	**Error term**	**Mathematical equation**
General nesting spatial (**GNS**)	x	x	x	$Y = \beta X + \rho Wy + \theta WX + \mu + \varepsilon$ $\mu = \lambda W\mu + \varepsilon$
Spatial autoregressive (**SAR**)		x		$Y = \beta X + \rho \mathrm{W}y + \varepsilon$
Spatial lag of X (**SLX**)	x			$Y = \beta X + \theta WX + \varepsilon$
Spatial error term (**SEM**)			x	$Y = \beta X + \mu$ $\mu = \lambda W\mu + \varepsilon$
Spatial autoregressive combined (**SAC**)		x	x	$Y = \beta X + \rho \mathrm{W}y + \mu$ $\mu = \lambda W\mu + \varepsilon$
Spatial Durbin (**SDM**)	x	x		$Y = \beta X + \rho \mathrm{W}y + \theta WX + \varepsilon$
Spatial Durbin error (**SDEM**)	x		x	$Y = \beta X + \theta WX + \mu$ $\mu = \lambda W\mu + \varepsilon$

Table 3: Overview lag models

Furthermore, different models exist that account for a combination of two interaction effects in the same specification. The spatial autoregressive model (SAC), also known as the Cliff – or SARAR model, includes the endogenous and error term effect.

In the spatial Durbin model (SDM), the endogenous and the exogenous interaction effects, are included. At the same time, the spatial Durbin error model accounts for the endogenous and error term effect. Their mathematical equations are reported in Table 3.

3.2.5 Global and local models

Except for the GNS model, it is proven that the parameters of other models can be estimated (Elhorst, 2017). However, the other models also face limitations that should be mentioned. To understand the first limitation, it is necessary to understand the difference between global and local models. Local models allow measuring the effects of characteristics from neighbouring regions on the dependent variable. This is accomplished by adding the average of the explanatory variables from neighbours to measure their influence on the dependent variable, while global models have endogenous spillover effects that are used to model global influences (Lesage, 2008). Global models are the SAR, SAC and SDM models that include a lag y, while the SLX, SEM and SDEM are seen as local models.

With local spillovers, a modification in variable *X* at location *i* will only affect those locations that are connected with location *(i)* by the weight matrix. In contrast, with global spillovers, a modification in variable *X* at location *(i)* will be transmitted to all other locations, regardless of whether the locations are connected according to the weight matrix or not. This is caused by the involvement of inverse matrices in global models. Another difference between the two model types is that global models contain feedback effects that pass through neighbouring units (from location *(i)* to location *(j)* to location *(k)*, e.g.) and finally influence the original location (*i*) where the change of *X* started (LeSage and Pace, 2011). Figure 4 portrays the feedback effect process in a global model. Whenever the transaction price of a dwelling (T_i) increases, due to an increase in one of its explanatory variables (*x2*), this spills over to its neighbouring dwelling through parameter ($\rho W y$) that accounts for average transaction prices of its neighbours. The increase of this parameter causes an increase in the transaction price of dwelling *(j)* *(*T_j*)*, which starts an increase in ($\rho W y$) for dwelling (*i*), where the process started. Then again, the transaction price of dwelling (*i*) will increase another time due to the feedback effect. According to Elhorst & Halleck Vega (2013), using global models for empirical studies is hard to justify due to these feedback effects. They affect the magnitude of the spatial spillover outcomes, which makes the individual lag estimates biased and unsuitable for prediction. Therefore, the usage of the SAR, SAC or SDM model is only limited to a small range of spatial research cases.

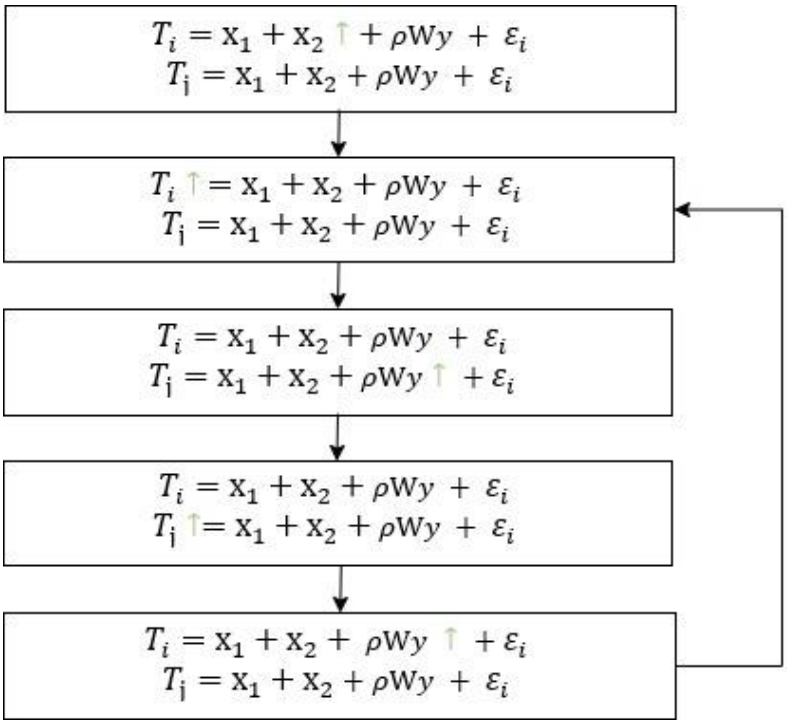


Figure 4: Feedback effect on global spillovers (own edit, 2022)

A well-known limitation of the SEM model is that the spillover effects are set equal to zero by construction. In other words, the model only allows to draw conclusions from the unobserved characteristics but not from the effects of the explanatory variables from the neighbouring dwelling.
Other models that include an exogenous effect (SLX, SDM and SDEM) have flexible effects that allow changes in observations that also impact neighbouring observations. However, these local spillovers are generally overlooked (Elhorst and Vega, 2017).

3.2.6 Spatial weight matrices

To account for spatial autocorrelation, location similarity must be captured so that this can be linked with attribute similarity. Therefore, an index must be constructed from the data, which is called a spatial autocorrelation statistic. To capture the locational similarity, spatial weight matrices, denoted as W, are used. Spatial weight matrices are a representation of the spatial structure of the data and inform the model when two spatial units interact. It is a quantification of the spatial relationships that exist among the units in the dataset. From a conceptual point of view, the spatial weight matrix is an N ×N table with one row and one column for each observation, where N represent the number of units in the dataset. The spatial relationship between the given row and column combination is quantified by the weight in each cell value. Usually, the spatial weight matrix is row-normalized. Row-normalizing is just a rescaling of the weights, it rescales the weight such that a row sum of the weight is one. As can be seen in Figure 5, the second row consists of two neighbours.

$$Row - normalizing \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} gives\ W = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$$

Figure 5: Example of row-normalized binary contiguity matrix for N=3 (Elhorst, 2018, own edit)

In statistical terms, the spatial weight matrix has element (w_{ij}) at locations (i) and (j) for (N) observations. This reflects the spatial influence of location (i) on location (j) or the other way around (Suryowati, Bekti & Faradila, 2018). There are many ways to define the weights depending on the research subject, study area, and desired spatial structure. Two main approaches to define spatial weights can be distinguished: contiguity and distance weights.

Binary contiguity weights, also known as the neighbouring matrix, identify 'neighbours' as units with common boundaries. So, whenever location (i) and location (j) share a border, then the spatial weight (w_{ij}) will be 1. Because weights only account for zero or one, this approach is called binary contiguity. To account for common borders sounds simple. However, there are different methods to interpret or define boundaries.

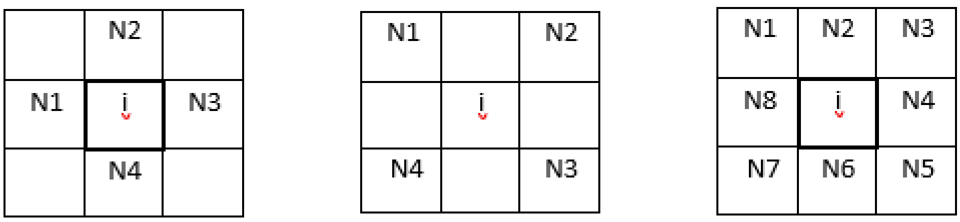


Figure 6:Rook contiguity, bishop contiguity and queens contiguity (Anselin 2002, own edit)

Anselin (2020) explains the different definitions of contiguity with the help of a standard grid, where i stands for the spatial units for which neighbours are defined as n = number of neighbours (see Figure 6). Rook contiguity defines neighbours by having a common edge between two spatial units. Bishop contiguity defines neighbours as spatial units that share a common border, while queen contiguity accounts for both. The definitions of contiguity determine the number of neighbours. Figure 6 shows that the rook – and bishop contiguity account for four neighbours, while the queen contiguity accounts for eight.

In practice, the precision of GIS can also influence the difference between the rook – and queen contiguity. However, the contiguity weights sometimes fail to capture a potentially significant relationship. For example, according to the administrative boundaries, islands are neighbouring.
But based on the polygons in a map, they do not share a common border or edge, so they are not counted as neighbours. To tackle this problem, instance-based weights can be used (Suryowati, Beti & Faradila, 2018).

In the case of Distance-based weights, the spatial weight matrix is constructed with the help of a distance measure (w_{ij}) between region (*i*) and region (*j*). (w_{ij}) is usually computed as the distance between the centroids of (*i*) and (*j*). Like the contiguity weights, there are different approaches to define the neighbours based on distance. The most common method is the distance band weight, also known as the threshold weight. It specifies that unit (*i*) is the neighbour of unit (*j*) if the distance between them is less than a specified maximum distance (band). Different distances can be set to define how many units are considered neighbours. However, if the critical distance is set too narrow, it could lead to units that are not assigned to a neighbour. To avoid that a unit gets isolated, the distance must be chosen such that each location has at least one neighbour. Therefore, the max-min distance can be applied. It specifies the threshold distance as the largest of the nearest neighbour distance. In other words, the distance between the two points in the dataset that are the furthest apart is chosen as critical distance. By following the max-min criteria, it can be ensured that every unit in the analysis has at least one neighbour and that isolated islands are avoided (Anselin & Rey, 2014). A drawback of handling this criterion is that it often leads to too many neighbours for units that are clustered.

With the *K*-nearest neighbour criterion, the problem of isolates can be overcome. This criterion explicitly limits the number of neighbours to a certain number. Instead of setting a distance threshold, a number is set for how many 'nearest' neighbours each unit takes into account. Besides avoiding the problem of isolates, the *K*-nearest neighbour criterion solves another issue of the distance band weight. The distance band weight implies that there is a symmetric relation between unit (*i*) and unit (*j*). At the same time, the fact that unit (*i*) is the nearest neighbour to unit (*j*) does not imply that unit (*j*) is the nearest neighbour to unit (*i*). Maybe unit (*k*) is the nearest neighbour from unit (*j*). These spatial relations can be accounted for with the *K*-nearest neighbour criteria because it does not require a symmetric relation.

Inverse distance weighting (IDW) assumes that things that are close are more alike than those further away. Therefore, IDW predicts a value considering the values surrounding the prediction unit. The closer values influence the prediction value more than the observations further away. By giving greater weights to points closest to the prediction locations, IDW takes into account that each observation has an influence that diminishes with distance. However, due to calculating power restrictions, it is a common practice to exclude the more distant points that only have a minor influence on the predicted value. This is done by specifying a neighbourhood. Commonly, a circle is used to determine what values will be taken into the calculation. However, other shapes to determine the neighbourhood could be used as well. The use of IDW also has disadvantages. First of all, the output of IDW is sensitive for clustering and outliers. In addition, IDM does not give a standard error of the estimate, it is not possible to measure the accuracy of the predictions (Suryowati, Beti & Faradila, 2018). This is a serious drawback that questions the justification of the use of this weight measure and therefore, the IDW measure is actually not used for empirical research.

Other distance weights discussed in spatial econometric literature are radial distance, exponential distance and double distance weight (Anselin, 2002; Suryowati, Beti & Faradila, 2018). However, these will not be discussed further as they are beyond the scope of this research. Different ways of weighting do give different spatial autocorrelation outcomes (Suryowati, Bekti & Feradila, 2018). Therefore, it is essential that a study makes well-deliberate choices in determining the spatial weight matrix. Contiguity weights are commonly used when dealing with spatial data on a larger scale like zip codes, counties or states, while distance-based weights are often used in HP analysis because the distances between dwellings can be calculated easily (Liu, 2018). In the next chapter, the choices regarding the specification of the weight matrix used in this research are discussed.

3.3 Spatial dependence statistics

As discussed in the previous section, the Moran's I and Lagrange multiplier tests can be used to find out if spatial autocorrelation is present in the data. This section will briefly discuss the mathematics behind both statistics and how their results should be interpreted.

3.3.1 Moran's I

Moran's I (*I*) is a commonly used statistic that measures the overall spatial autocorrelation in a dataset (Hijmans & Ghosh, 2021). The coefficient for a given dataset represents the actual presence of positive or negative spatial autocorrelation. The equation for the Moran's I statistic can be written as:

4) $$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

where n represents the number of observations, w_{ij} the spatial weight matrix, y is the transaction price, and $\bar{y}$ stands for the mean value of y (Hijmans & Gosch, 2021).

The null hypothesis for the Moran's I static states that the attribute being analyzed, is randomly distributed among space. This hypothesis can be rejected whenever the p-value is statistically significant ($p < 0.05$). A positive value of I indicates a positive spatial autocorrelation and a negative test value negative spatial autocorrelation. Positive spatial autocorrelation refers to the clustering of observations with similar values, while negative autocorrelation points to the clustering of dissimilar values (Bouayad Agha & Bellefon, 2018).

3.3.2 Lagrange-Multiplier tests

The statistical Lagrange-Multiplier (LM) tests are often used to diagnose where spatial autocorrelation is present in the dataset. While the Moran's I test is a more general statistic, the LM tests are more specific. The LM-error test tests for spatial correlation in the error term, while the LM-lag test can be used to explore the presence of spatial correlation in the lagged values of the dependent variable (Hijmans & Ghosh, 2021).

Lagrange multiplier test for spatial lag dependence (LM-lag)

The LM-lag statistic is used to determine the presence of spatial correlation in the dependent variable. In the case of this study, it will test for spatial correlation in the dwelling prices. For the Lm-lag test, a SAR model is written:

5) $$Y_i = \beta X + \rho \mathrm{W} y + \varepsilon$$

to test the hypothesis that $\rho = 0$. Whenever this null hypothesis is rejected, spatial correlation can be found in the dependent variable, and the OLS model should be extended. The variables and parameters in the SAR model follow the same specification as model 3.

Lagrange multiplier test for spatial error dependence (LM-error)

The LM-error tests if the spatial correlation is present in the error terms. The following SEM model is used to test the hypothesis:

6) $$Y = \beta X + \mu$$
$$\mu = \lambda W \mu + \varepsilon$$

With the LM-error test, the null hypothesis states that $\lambda = 0$, there is no spatial autocorrelation in the error term if it cannot be rejected. The variables and parameters in the SEM model follow the same specification as model 3.

3.4 Summary

This chapter first explained the concept of hedonic pricing and that structural, locational and UGS characterises are often used when a study is focused on UGS valuation. However, hedonic pricing models are criticised because they do not account for spatial dependencies that could lead to heteroskedastic issues or bias in the estimates of coefficients. Therefore, the second section of this chapter discussed the concept of spatial econometrics, as extending the HP model with the help of spatial econometric modelling can help account for the spatial dependencies. First, the spatial econometric core concepts; spatial dependence, spatial autocorrelation, spatial interaction effects and spatial spillovers, are explained. With the help of the spatial weight matrix, the proximity between observations in the data sample is described. The spatial weight matrix can be used to model endogenous, exogenous or error term interaction effects that account for spatial autocorrelation in the dependent variable, explanatory variables or error term. This chapter described the different lag models that can be created by adding these interaction effects to the standard hedonic pricing model. The different methods of how the spatial weight matrix can be specified are also discussed. In the last section of this chapter, Moran's I and Lagrange multiplier tests statics are explained that can be used to find out if there is spatial autocorrelation in the dataset.

CH 4 Model Specifications

The previous chapter explained the methodology of this research. This chapter discusses the motivation behind the specification of the hedonic pricing model, weight matrix and spatial models. The first section provides the motivation for the specification of the hedonic price function, including the choices for the functional form and explanatory variables. This is followed by the specification process for the spatial weight matrix and the specification of the spatial models. The last part of this chapter will discuss the approaches that could be used to determine which spatial model results should be used for the interpretation of the individual variable estimates.

4.1 Specifying the hedonic price function

When specifying a regression function, several considerations and choices must be made. First, it should be decided what set of explanatory variables should predict the value of the dependent variable. As described in the methodology chapter, the hedonic pricing function determines the property price defined by a function with the transaction price as the dependent variable and a set of dwelling characteristics as the explanatory variables. In this thesis, it is decided to follow the hedonic specification from Czembrowski et al. (2019), which includes three vectors that represent sets of structural, locational and UGS characteristics, to predict the transaction price (T_i). It is decided to add a fourth vector to this equation, used to present the explanatory variables that are added to control for other factors than the mentioned housing characteristics, which have an impact on the transaction price, e.g., transaction year.

Next, the functional form of the HP model is determined. In this thesis, it is decided to specify the HP function as a semi-log function, where the dependent variable will be transformed into the natural log of the transaction price. The motivation to use a semi-log function is threefold. Firstly, using a natural logarithm helps reduce the skewness in the data distribution and normalize the variable when the assumption of normality is violated. Next, it helps reduce the variability in the data, reducing the impact of outliers, so they do not have to be removed (Field, 2009). Finally, an added advantage of the semi-log form is the straightforward interpretation of the estimated price coefficients. Coefficients from linear explanatory variables present the price change caused by one unit increase in the explanatory variable, while the coefficients from log-transformed explanatory variables can be interpreted as elasticities. In that case, the estimated coefficient indicates the percentage change in the explanatory that is associated with a 1% increase in the transaction price. This can be very useful for housing characteristics that do not have a linear relationship with the transaction price. Rosen (1974) explains that the transaction price of a dwelling can be linear and related to one of its characteristics. This implies a constant marginal price, which is not the case for all the dwelling attributes. Take the size of a dwelling expressed in m² as an example. A constant marginal price would indicate that an extra square meter of living area results in a standard price increase that is the same for all dwellings. However, it is known that this relationship is not linear, often a square meter increase in the case of a small dwelling will have a higher price effect than in the case of a large house. Therefore, expressing the relation between the transaction price and dwelling size as a price elasticity seems a better option, which can be achieved by using the log transformation for both variables.

Concluding, the specification for the hedonic price function that will be used in thesis can be written as:

$$\ln(T_i) = \beta_0 + \sum_{j=1}^{J} a_j S_{ij} + \sum_{m=1}^{M} \beta_m L_{im} + \sum_{n=1}^{N} \gamma_n U_{in} + \sum_{k=1}^{K} x_k C_{ik} + \varepsilon_i \tag{7}$$

in which the transaction price of the dwellings (Ti) is transformed into a natural logarithm. Some of the explanatory variables will be transformed to a logarithmic form so their coefficients can be interpreted as price elasticities, but most will enter in linearly. The constant is denoted as β_0, while S_i represents the dwelling's structural variables, L_i the locational variables, U_i the urban green space variables and C_i the control variables. The values, a_j, β_m, γ_n and x_k are their respective coefficients while ε_i presents the error term.

4.2 Specifying the spatial weight matrix

To specify the spatial model to account for spatial interaction effects, a spatial weight matrix (*W)* must be defined. There are no standard rules on how the spatial weight matrix should be specified (Anselin, 2017). The most common specification methods are discussed in section 3.2.5.

Contiguity weights are considered inappropriate to use for this thesis. First of all, this thesis will use point data, while contiguity weights are based on polygons that define adjacent borders between different spatial units. Thiessen polygons from point features could solve this problem (Thiessen, 1911). However, a more important consideration is that contiguity weights are not seen as a fair reflection of the spatial relation between dwellings, which are the units of analysis. As stated by Liu (2018), contiguity weights are appropriate to use when the study is interested in the interaction of units with clear spatial boundaries, but that does not apply to the units of analysis in this thesis.

According to Liu (2018), most HP analyses use distance-based specifications of the weight matrix, although the literature review indicated that quite some HP studies used the k-nearest neighbour's specification instead (Herath et al., 2015; Sander & Zhao, 2015; Czembrowski et al., 2016; Czembrowski et al., 2019). The appeal of this approach is that it creates a matrix with the same number of neighbours for each dwelling. However, the transaction data is geocoded at a 6-digit zip code as this dataset contains 19,641 houses located in 5,831 zip codes, it can be assumed that some dwellings share the same zip code and the same spatial indicator. An analysis of the frequency table of the '6-digit zip code' variable indicates that there is even one zip code that is assigned to 49 dwellings. This means that at least 49 neighbours should be assigned to each dwelling if a k-nearest neighbourhood matrix is made with this dataset. Otherwise, computational errors occur because those 49 dwellings are geocoded in the exact same area. However, this would lead to uneven distributed 'neighbourhood areas' as some of the dwellings will be assigned to 48 neighbours from the same 6-digit zip code areas, while others will be assigned to 48 houses from different 6-digit zip code areas. This results in a matrix that consists of very different spatial areas.

Therefore, this study follows Liu (2018) and uses a distance-based weight to create the spatial weight matrix. Using an inverse distance matrix sounds interesting as it embodies Tobler's first law of geography (1970). However, due to its drawbacks discussed in section 2.2.4, it was not considered an option. The spatial weight matrix can only be specified if there are no 'isolates and each observation is assigned to at least one neighbour (Anselin & Rey, 2014). Therefore, it was decided to use the 'max-min distance weight'. The main criticism of using this max-min distance is that it generates many neighbours for clustered units (Anselin & Rey, 2014). Looking at table 19, this also seems to apply in this case. However, as the mean number of neighbours is close to the average number of houses in each neighbourhood (#observations (19,641) / total number of neighbourhoods in Almere and Eindhoven = 105) and a radius of 330 metres around each dwelling sound as a fair reflection what people consider as their direct neighbours, this specification of the spatial weight matrix seems justified.

4.3 Specifying the spatial models

This section explains the specifications of the spatial models. First, the specifications of the local models (SLX, SER & SDEM) are given followed by the specification of the global models (SAR, SDM).

4.3.1 Spatial lag of X model

8) $$\ln(T_i) = \beta_0 + \sum_{j=1}^{J} a_j S_{ij} + \sum_{m=1}^{M} \beta_m L_{im} + \sum_{n=1}^{N} \gamma_n U_{in} + \sum_{k=1}^{K} x_k C_{ik} + \theta WX + \varepsilon_i$$

this variant on the OLS models allows the logged transaction price of each dwelling ($\ln(T_i)$) to depend on its own characteristics, plus the same characteristics averaged over the neighbouring dwellings (θWX). Parameter θ accounts for the exogenous interaction effects, X stands for the matrix with all explanatory variables, while W is the weight matrix defined in the previous section. The other variables and parameters are defined as in model 7.

4.3.2 Spatial error term model

9)
$$\ln(T_i) = \beta_0 + \sum_j^J a_j S_{ij} + \sum_m^M \beta_m L_{im} + \sum_n^N \gamma_n U_{in} + \sum_k^K x_k C_{ik} + \mu_i$$
$$\mu_i = \lambda W\mu + \varepsilon_i$$

This SEM model extends the OLS model with μ_i that contains the error term (ε_i) of location *i* but also the averaged error term effect of its neighbours ($\lambda W\mu$). Parameter λ stands for the error term interaction effects, μ for neighbouring error terms and W for the weight matrix.

The other variables and parameters are defined as in model 7.

4.3.3 Spatial Durbin error model

The SDEM model combines the added interaction effects from the two previously defined models into the following equation:

10)
$$\ln(T_i) = \beta_0 + \sum_j^J a_j S_{ij} + \sum_m^M \beta_m L_{im} + \sum_n^N \gamma_n U_{in} + \sum_k^K x_k C_{ik} + \theta WX + \mu_i$$
$$\mu_i = \lambda W\mu + \varepsilon_i$$

The other variables and parameters are defined as in model 7.

4.3.4 Spatial autoregressive model

With the SAR model, the endogenous interaction effect (ρWy) is added to the OLS model.

11)
$$\ln(T_i) = \beta_0 + \sum_j^J a_j S_{ij} + \sum_m^M \beta_m L_{im} + \sum_n^N \gamma_n U_{in} + \sum_k^K x_k C_{ik} + \rho Wy + \varepsilon_i$$

The parameter (ρ) measures the strength of dependence in the transaction data (y) of neighbouring dwellings (W). The other variables and parameters are defined as in model 7.

4.3.5 Spatial Durbin model

The SDM model adds the endogenous interaction and error term interaction effect to the OLS model as specified in model 9 and model 11. The other variables and parameters are defined as in model 7.

12)
$$\ln(T_i) = \beta_0 + \sum_j^J a_j S_{ij} + \sum_m^M \beta_m L_{im} + \sum_n^N \gamma_n U_{in} + \sum_k^K x_k C_{ik} + \rho Wy + \mu_i$$
$$\mu_i = \lambda W\mu + \varepsilon_i$$

4.4 Method for model selection

In the previous section, different spatial models have been specified. However, it still needs to be decided which one fits the best to the dataset and should be used for interpretation. According to Elhorst (2017), the greatest problem in empirical research is choosing the right *W* and lag model(s) specification. Especially if there is no reference to a specific economic theory. Therefore, many studies choose their model by following data-analytic considerations, often leading to the SAR or SEM model. On top of that, most studies lack guidance on how the spatial weight matrix should be specified. Even when they provide a well-founded background for determining the right spatial interaction effects. Krekel (2020) agrees with the challenges described by Elhorst. According to Krekel, the methodological discussion of ecosystem valuation is currently centered around the benefits and disbenefits of the respective method. This should move towards practical guidance on which (combination of) method(s) is preferable under which circumstances. It would be a major step toward finding a standard methodology for monetary valuation of intangibles systems of national accounts. This study hopes to contribute to this methodological discussion by looking for the proper method to model urban green spaces. Therefore, it is essential to understand what (statistical) methods are used to determine the right model. Among spatial econometric experts, there is a discussion of what statistical criteria should be used to determine the selection of models. First, the traditional and main approaches will be discussed. Afterwards, some relatively new and innovative approaches will be highlighted and finally the method used in this thesis will be explained.

4.4.1 Traditional approaches

The first method is called the *bottom-up* approach, developed by Anselin et al. (1996). This approach assumes that the neighbourhood matrix is known and that the explanatory variables are exogenous (Floch & Saout, 2016). This approach uses a non-spatial model, in other words, just a standard OLS model, as a starting point. With the help of a Lagrange multiplier test, it can be determined if the non-spatial model should be extended to a SAR or SEM model. Compared with the other approaches, the *bottom-up* approach requires little computer power because the tests are based on the residuals of the non-spatial model. Therefore, this approach was the most popular option till the 2000s. Figure a in Appendix I developed by Florax et al. (2003) shows the procedure of this approach.

The second method is called the *top-down* approach. This approach handles the same assumptions as the *bottom-up* approach. The rapid development of IT made it possible to estimate more complex models. Lesage and Pace (2009) argued that you should not start with a non-spatial model but instead with the Durbin spatial model. Based on the outcomes of the first Lagrange multiplier test, the model can be reduced to the SAR, SLX or SEM model. With the second Lagrange multiplier test, the model can even be reduced to a linear model. Figure b in Appendix I, developed by Lesage and Pace (2009), shows the procedure of this approach. Compared to the *bottom-up* approach, the *top-down* approach provides the opportunity to see if it is more appropriate to use a SDM, SLX, SAR, SEM or OLS model and not only a SAR SEM or OLS model. Whenever the first test gives strong statistically significant coefficients for all the spatial parameters, then the SDM model is the right choice. If not, a few restrictions can be made to test whether a simpler model suits the data better.

The third method is called the *combined* approach and was developed by Elhorst (2010). Just like the *bottom-up* approach, this method starts with an OLS model. However, as opposed to the other two models, it allows exogenous interaction effects to be integrated into the analysis. Therefore, the first Lagrange multiplier test determines if the SLX or SDM model is more appropriate. From that point, the likelihood radio and LM tests allow testing if the model should be reduced to SAR, SEM or OLS model. Whenever there is any doubt, the model that appears to be the most robust should be chosen. Figure c in Appendix I, developed by Elhorst (2010), shows the procedure of this approach. Beware that calculating the likelihood of the spatial models costs a substantial amount of computer power, especially when the number of observations is high.

Although the above three approaches have been widely used within spatial econometrics, some experts question their fundamentals.

According to Elhorst (2017), the Lagrange multiplier test is not a helpful statistic to determine the appropriate model because the Lagrange multiplier test does not take the exogenous interaction effects into account, while the models that do include an exogenous effect allow to check for local spatial effects. In addition, it is well-known that the GNS, SAR and SAC models have additional flaws, as discussed in CH 2.3. Finally, the traditional approaches assume that the neighbourhood matrix is known, while Lesage and Pace (2014) show that the specification is of vital importance because the specification of W determines the value and the significance level of the interaction parameter depends on the specification of the weight matrix. Therefore, the direct and indirect effects are sensitive to fundamental changes of *W*. Finding the correct specification of the spatial weights is a basic identification problem in applied econometric research, just like choosing the right spatial econometric model. When the specifications of *W* do not follow from theory, it is important that studies investigate if the model results are sensitive to the specification of *W*. Therefore, Elhorst (2017) argues that two other approaches are more promising to overcome the identification problem and to find the right spatial weight matrix and model.

4.4.2 The Bayesian comparison approach

The Bayesian comparison approach was developed by Lesage (2014). This approach emphasised that the first step is to choose a model with local or global spillover effects based on theoretical aspects of the research problem. With the Bayesian comparison approach, only two types of models are considered. If the theory implies that the spillover effects are global, then the SDM is the model of choice. Whenever the spillover effect is perceived as local, the SDEM model should be chosen (Elhorst, 2017).

Whenever theoretical arguments for one of the models are lacking, the models should be tested against each other. However, instead of the Lagrange multiplier test Lesage (2014) argues for another approach. First, because this test does not account for exogenous interaction effects moreover, the log-likelihood function values from the SDM and SDEM models are, in general, much closer to each other than those of the SAR and SEM models. And because the spatial spillover effects are often comparable numerically, while their interpretation is completely different (Elhorst, 2017). According to Lesage (2014), the posterior model probabilities of SDM and SDEM for different weight matrices should be determined. The log marginal likelihood of the model can be acquired by integrating out all parameters of the model over the entire space on which they are defined. The strength of this method is that the models' performances are compared on the whole parameter space. Unfortunately, this approach lacks a straightforward guideline on which statistical test should be used.

4.4.3 The SLX approach

According to Vega & Elhorst (2015), the SLX approach is a better approach to use whenever the research is interested in measuring spillover effects. They argue that the spatial model of choice should at least account for exogenous interaction effects. The SAR, SAC and SEM models do not contain the exogenous interaction effect and are therefore less appropriate to use.
Therefore, if the research is interested in measuring spillover effects, it should consider the SLX, SDM and SDEM as a model of choice. Because the SLX model is the simplest model with flexible spatial spillover effects, this should be the model to start with. As the SDM and SDEM are extensions to the SLX model, the Bayesian comparison approach can be used to determine if the model needs to be extended to one of those two models.

With the traditional approaches, it is assumed that the specification of the spatial weight matrix is known. On the contrary, the SLX approach allows to adopt a parametrized spatial weight matrix for more flexibility. This allows testing if different spatial weight matrices result in other noteworthy price effects, while the SLX model does not account for spatial econometric endogenous variables (Vega & Elhorst, 2015). It can still be used to test for endogenous effects using non-spatial econometric techniques. A combination of the Hausman test for endogeneity with tests for the validity of the instruments to assess whether they satisfy the relevance and exogeneity criteria could be used (Elhorst, 2017).

4.4.4 Thesis approach

Above, different procedures are discussed that help determine what model should be estimated and interpreted to answer the research question. This thesis follows the arguments of Elhorst (2017), that state that the traditional approaches should not be used to determine the most appropriate model. The SLX approach states the SLX model should be the starting point. However, if spatial dependency is present in the transaction data, error term and explanatory data, it makes more sense to start with a model that controls for at least two of the three. This is in line with the Bayesian comparison approach from Lesage (2014), which states that only the SDEM or SDM model should be considered. Unfortunately, this approach lacks a straightforward guideline of which tests should be used.

Therefore, this thesis decided to follow the scheme shown in Figure 7, which combines the different discussed approaches. First, the Moran's I and Lagrange Multiplier tests should determine if spatial dependency is present in the data. If the test points out that there is only spatial dependency in the transaction data, then the SDM model should be the point of departure whenever the test outcome shows only spatial dependence in the omitted variables, then the SDEM model is the one to start with. However, it is also possible that spatial autocorrelation will be found in the transaction data and omitted variables. Therefore, Elhorst & Halleck Vega (2013) suggest to rely on theory and the specific context of the research rather than statistical test as the main guide in selecting the model. Theory should point out if a local model (SDEM) or global model (SDM) should be the point of departure. After deciding the starting point, likelihood ratio (LR) tests should determine if the model should be restricted to a lower-level model.

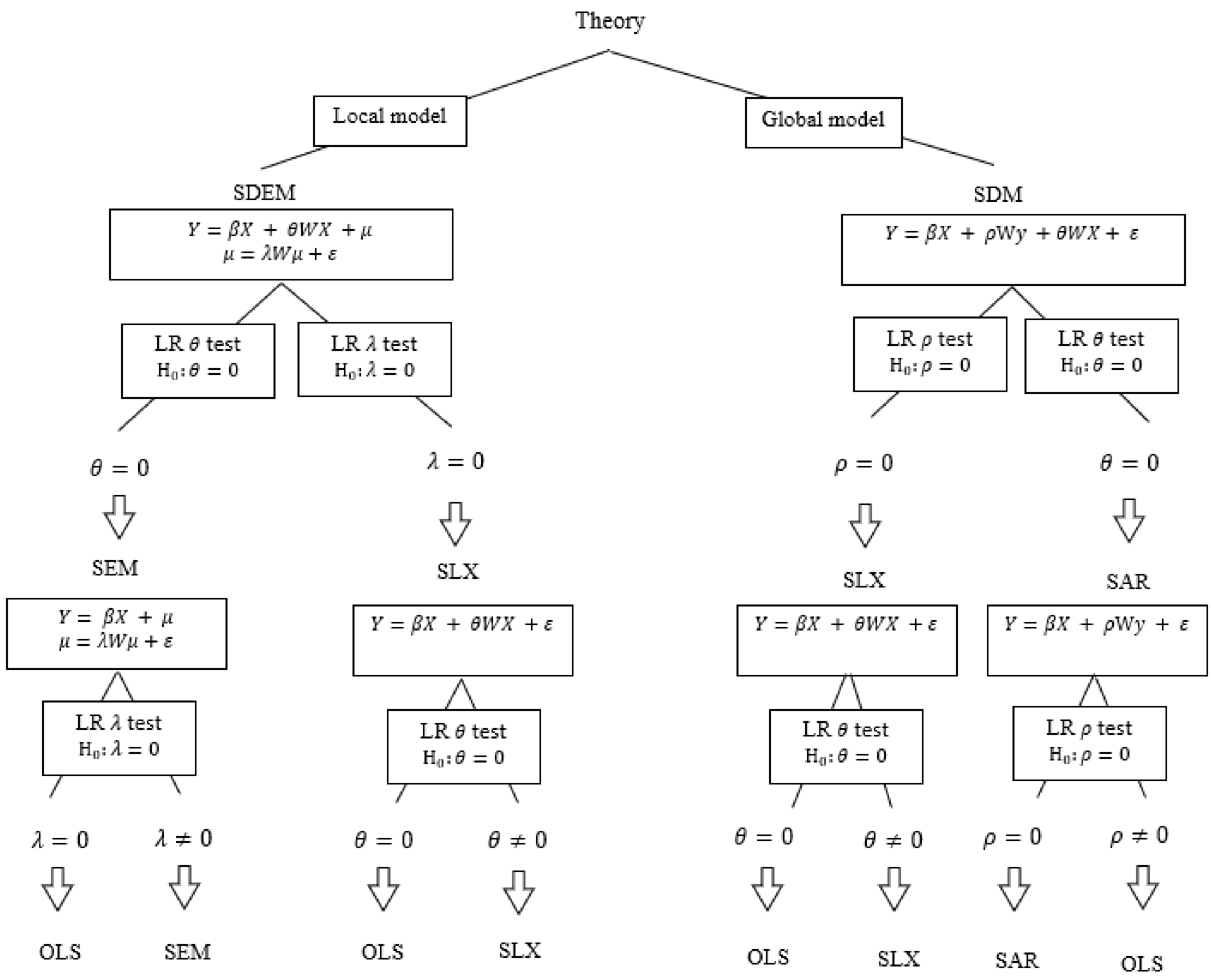


Figure 7: Model selection approach (own input, 2022)

4.5 Summary

This chapter has described the process of specifying the hedonic price function, spatial weight matrix and spatial models. The motivation to use a semi-logged form is given and it is explained that the explanatory variables will consist of control variables and structural, locational and UGS characteristics. Next, it has been explained why the 'max-min distance' is used to specify the spatial weight matrix and the specification of the spatial SLX, SER, SDEM, SAR and SDM models were given. At last, the approaches that can be used to select which spatial model(s) should be used for interpretation are explained. This thesis will use a combination of the Bayesian comparison and SLX approach. The Moran's I and Lagrange Multiplier tests shall first be used to determine if there is spatial dependency in the data. If spatial dependence is present in both the transaction data as the omitted variables, then the theory should point out if the SDM or SDEM model should be the point of departure. The likelihood ratio test outcomes will help decide if the model should be simplified to a lower-level model.

CH 5 Data

This chapter elaborates on the dataset that is being used to estimate the models that are specified in the previous chapter. First, the consulted data sets are explored, followed by the UGS classification handled in this thesis. Next, it is explained how the provided UGS data is transformed into workable variables using QGIS and TransCAD. This includes a brief explanation of how the proximity variables are made for different spatial scales and a longer description of how the Euclidean and Network distances are calculated to account for the proximity of UGS. The third part of this chapter describes the theoretical choices for the individual variables that should be included in the hedonic pricing function, followed by an explanation of the outlier and multicollinearity analysis which determine the final data set and set of variables that will be used in the hedonic pricing model. Subsequently, a summary of the variables that enter the model with their descriptive statistics is given. In the final section, heatmaps are provided and analyses for the transaction data and some of the UGS variables.

5.1 Data sources

Data have been collected from several sources to account for the structural, locational and UGS characteristics of each sold dwelling. This first section of the chapter describes the content of each dataset.

5.1.1 Transaction data and structural characteristics (NVM dataset)

The Nederlandse Coöperatieve Vereniging van Makelaars en Taxateurs in onroerende goederen (NVM) provided a dataset that contain all the sale transactions of residential properties in Almere and Eindhoven between 2014 and 2018. The raw dataset consists of 22,025 transactions in total. From this total, 10,196 transactions are from Almere and 11,829 from Eindhoven. The transaction data can also be categorized into dwelling types. From the total dataset, 15,561 are residential houses, while 6,463 are apartments. The data is geocoded at a 6-digit zip code level.

In addition, the NVM provided data on a variety of property attributes. Table a in Appendix II shows the available structural characteristics of the properties in the NVM dataset. The first rows contain continuous variables about the living area in m², plot size in m² and the volume of the dwelling in m³. NVM provided two variables that account for the living space in m². It was chosen to use the one that contained corrected values whenever the obtained value was perceived as unreliable (NVM, 2021).

5.1.2 Locational characteristics (CBS dataset)

The Centraal Bureau van Statistiek (CBS) publishes a public dataset with key figures on district, neighbourhood and zip code (4, 5 and 6 digits) levels every year. These datasets contain statistical information at a low regional level about different social-economic topics like; population, employment, crime, energy, healthcare, business establishments and public facilities. All these datasets are publicly accessible on the webpage from the CBS and were obtained as a shape or CSV file. This dataset has been consulted to account for the locational variables.

Because the NVM dataset contains data for the period between 2014 and 2017 and the HUGSI dataset consists of 2019 and 2020 data, it was chosen to use the CBS dataset with key figures of 2017. First, because this year was some were in the middle, but foremost because the statistics of 2017 are available on the lowest regional level (6-digit zip code), while newer datasets are incomplete and not freely available for the most detailed spatial level (5- and 6-digit zip code level). It was desired to obtain the data at a 6-digit zip code level as this provides the most detailed information. Still, it also allowed merging the dataset easily with the HUGSI data, which was provided on the same spatial level.

The dataset of the CBS consists of 129 variables in total. However, not all the data is relevant for this thesis (e.g., man/female ratio, the average age of residents) and is disregarded in advance. The relevant information about locational characteristics for this thesis can be found in Table b in Appendix II. The CBS uses two measurement methods to account for the presence of a local amenity.

First, by considering the average amount of amenities within a certain distance by road for all the residents in an area. What distances are used differs per amenity. But also by taking the average Network distance calculated by road for all residents in an area to the nearest amenity. The distances are calculated with a road network that only includes paved roads used by cars, cycle paths and footpaths were excluded. The calculation did not take traffic restrictions like one-way traffic and no-entry zones into account (CBS, 2020). The distance to the nearest amenity is calculated for each type, while not all the classes are provided with the average amount of amenities within a particular area. For most amenities with daily services, 1, 3 and 5 kilometres are taken as distance barriers to calculate the average amount of amenities. While, for amenities that are not used daily, like hospitals, distance barriers of 5, 10 and 20 kilometres are used.

CBS has the statistical information for squares of 100 by 100 metres and linked the information to the geometry of the zip code surfaces derived from the ESRI Netherlands BAG addresses. It should be stressed that the CBS variables take the average distance and number of facilities for the entire zip code into account. In other words, the data shows the average distance to – and several nearby facilities for all the dwellings in the 6-digit zip code. This dataset divides the municipality of Eindhoven into 5.592 6-digit level zip code areas and the city of Almere into 5.694 6-digit level zip code surfaces (CBS, 2020).

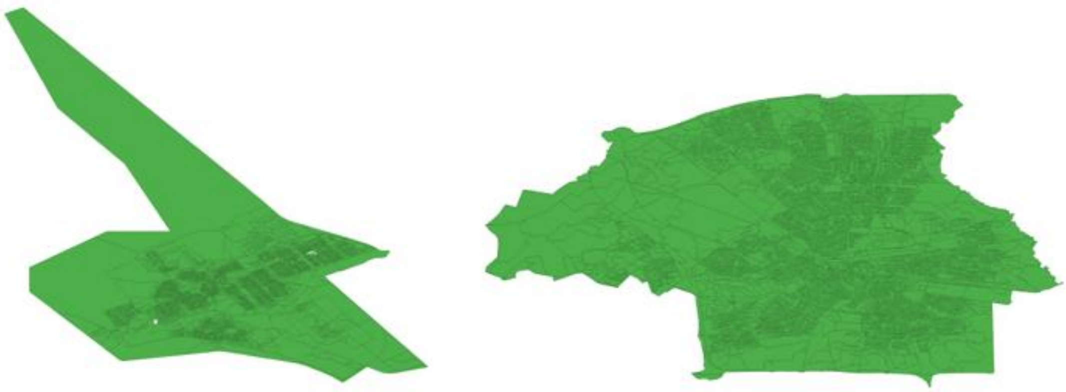


Figure 8: Almere & Eindhoven on 6-digit zip code level (CBS, 2021)

5.1.3 UGS characteristics (Multiple datasets)

The goal of including urban green space variables is to estimate the marginal willingness of real estate buyers to pay for different UGS amenities. These estimates represent the monetary value of UGS. The literature research found that UGS variables can be operationalised based on the proximity principle or density. Additionally, it was found that it is encouraged to take the heterogeneity from UGS into account and use 'land coverage' instead of 'land use' data as it probably provides a more accurate reflection of reality.

Therefore, three data sources are derived to create UGS variables. HUGSI provided land coverage data that is used to generate density variables. The land use data from OpenStreetMap allows for distinguishing different types of UGS. In combination with Road Network data from Nationaal Wegenbestand (NWB), the distances to the UGS amenities can be calculated and used as proximity variables. More details about this process will be given in the next section. First, a brief explanation of the content of the derived UGS data sources is provided.

<u>UGS Density (Husqvarna Urban Green Space Index)</u>

Husqvarna Urban Green Space Index (HUGSI) is developed by Husqvarna in collaboration with Overstory. They created an index to compare 155 cities from 60 countries on the level of greenness. To do so, they applied deep learning algorithms on satellite images from European Space Agency (ESA) to determine the amount of urban green space in each city. The algorithms were trained to distinguish urban green from other types of land coverage. In addition, the algorithm can determine different forms of urban green.

All vegetation above 1 meter is classified as trees, while the vegetation below 1 meter is seen as grass. The algorithms also recognise land that is occupied by water. When sunlight strikes vegetation, a certain amount is adopted while the rest is reflected. Healthy vegetation reflects most of the NIR light and absorbs most of the visible light, while unhealthy vegetation absorbs more NIR light. Based on these assumptions, the algorithm can recognise the vegetation's health by the light they reflect (HUGSI, 2021). With the help of deep learning techniques, they were able to measure the amount of each type of urban green on a 10 by 10 grid.

HUGSI provided a dataset that contains information about Eindhoven for the years 2019 and 2020 and Almere for the year 2020. The raw dataset consists of 11.176 geo-coded points representing a six-digit zip code with the average variable values of the 10 by 10 grids that are part of the six-digit zip code area. Table 4 gives an overview of the variables in the HUGSI dataset, all the variable values are provided on a zero to one scale.

Variable	**Description**	**Measurement level**
Grass	Indicates the amount of land covered with vegetation ≤ 1m	Continuous from 0 to 1
Trees	Indicates the amount of land covered with vegetation > 1m	Continuous from 0 to 1
Water	Indicates the amount of land covered with water	Continuous from 0 to 1
UGS	Indicates the amount of land covered with grass, trees or water	Continuous from 0 to 1
Other	Indicates the amount covered with other land uses than UGS	Continuous from 0 to 1
Health	Indicates the health of the urban green space	Continuous from 0 to 1

Table 4: Overview variables provided by HUGSI. (HUGSI 2021, own edit)

The percentages variables represent a value between 0 and 1. For each variable counts that 0 stands for a land coverage of zero percent while 1 stands for a land coverage of a hundred percent. The summed values of trees and grass equal the value of urban green. The variable '% other' indicates every other land coverage except trees, grass and water. The health variable is also a value between 0 and 1, where zero stands for worsts health and one indicates the best health.

It should be stressed that the usage of this UGS land coverage data is unique. As far as this thesis is aware, land coverage data derived with satellite images and machine learning have not been used yet for UGSs valuation studies. At the same time, it provides an opportunity to model the amount of urban green space in a spatial area very precisely.

UGS Proximity (OpenstreetMap and Nationaal Wegenbestand data)

OpenStreetMap (OSM) is an open data source that provides crowdsourced geo-data under an open database license. It allows downloading data in a layered GIS format for each province in the Netherland. It consists of several shapefiles containing various features like city locations, land use types, traffic infrastructure and amenities like police stations and post offices. For this thesis, the data for Flevoland and North Brabant was derived, this thesis used the OSM data layers related to land use types and road networks. An overview is added to Table c in Appendix II.

With the help of the 'clip function' in QGIS, a new vector layer was created to distinguish the city of Eindhoven and Almere from the rest of the provinces. The CBS shapefile was used as a base layer. According to the OSM classification, Almere and Eindhoven are divided into 21.723 surfaces that belong to one of the 21 land use classes from the water and land use layer. A majority of these land use classes fall within the UGS definition. The following section will discuss which exactly are used in combination with road network data to operationalise the proximity of dwellings to different UGS amenities.

The Nationaal Wegenbestand (NWB) is consulted to account for accurate road network data. The NWB (2020) dataset contains all the Dutch roads under the administration of the National government, the provinces, the municipalities or the water authorities. The dataset consists of 158.000 hecto-points and digitalized roads of almost the length of 158.000 kilometres in total. Included are all the roads that are provided with a street name. This includes loose, unpaved, cycling and walking roads as long as they are related to a street name (Rijkswaterstaat, 2021).

5.2 Data transformation

A relation between each dwelling and UGS amenities should be established to account for urban green spaces in a HP function. As discussed in the literature, proximity and density are the standard methods to establish this relation. Both methods will be used to account for UGS in the hedonic price function of this thesis. The previous section discussed the data that is derived, this section will elaborate on the choices and techniques used to create the UGS variables. However, first, the classification of urban green spaces used in this thesis is outlined.

5.2.1 Classification of UGS

As stressed in the literature research, there does not exist a common classification for UGS in the literature. However, remarkable is that most studies do not account for trees or other vegetation in their classification while these classes are part of the UGS definition. This is probably caused by the fact that most studies account for urban green spaces with the help of land use data. Unfortunately, land use data is not accurate enough to take all green land coverage into account, as stressed by Xu et al. (2016).

The HUGSI dataset is consulted to overcome this problem, this data provides the exact percentage of green land cover. Therefore, with the help of this data, it can be studied if the total amount of land covered with trees, grass or water in an area affects the real estate price. However, the HUGSI does not allow to control for other UGS features that arise from its definition. The OSM data allowed to add different UGS amenities to this classification. This helps to test the statement from Liebelt, Bartke & Schawarz (2018) that it is valuable to analyse different types of UGS, as they provide other ecosystem services and thus different benefits for residents. The OSM layer provided data for the following UGS categories; allotments, beaches, cemeteries, farmlands, forests, farmyards, heaths, grass, meadows, orchards, nature reserves, recreational grounds, riverbanks, vineyards and water. On the forehand, the recreational ground does not sound like UGS, but OSM describes it as an open green space for recreational purposes. Therefore, the recreational grounds in the OSM dataset will be seen as open green space within this thesis. However, it was decided that not all OSM categories will be used for the UGS classification of this thesis. This thesis is interested in green within urban areas; therefore, forests are not considered as the cities of Almere and Eindhoven do not have forests within their urban areas. This also counts for nature reserves and heaths. As only one riverbank and one vineyard exist in this dataset, both in Almere, these were dropped from the classification. The UGSs classification used in this thesis can be found in Table 5.

UGS Class	Definition	Data source
Allotments	Area that is characterised by a concentration of small private gardens assigned to individuals or families. Also known as community gardens	OSM
Beaches	Loose geological form of land along a body of water	OSM
Cemeteries	Cemetery or graveyard	OSM
Farmland	Area of land used for agriculture	OSM
Grass	Grass and vegetation from one meter or lower	HUGSI
Meadow	Area of land primarily vegetated by grass, possibly used for grazing cattle	OSM
Open green space	An open green space for general recreation often owned by the municipality	OSM
Orchard	Area used for intentional planting trees or shrubs used for food production	OSM
Park	Area of open space for recreational use. Usually designed in semi-natural state within an urban area	OSM
Trees	Trees and vegetation higher than one meter	HUGSI
Scrub	Area of uncultivated land covered with shrubs, bushes and trees	OSM
Water	Area covered with water, sea or ocean	HUGSI

Table 5: Urban green space classification. (HUGSI, 2021 ; Openstreetmap, 2022, own edit)

5.2.2 UGS density variables

As discussed, HUGSI provides the percentage of land covered by grass, trees, water or other land uses in a 6-digit zip code area. Because the data is provided on such a detailed geographical scale, it was decided to calculate the percentage from higher spatial scales to test which spatial scale gave the most significant result.

With the help of the ‘flash fill’ function in excel, the provided ‘6-digit zip codes’ were broken down to 5-digit and 4-digit zip codes and added as new variables. The ‘Group-stat plugin’ in QGIS allowed to add the land coverage values by calculating the averages from the 6-digit areas located within these two larger postal code areas. With the help of the ‘linking attributes with field values’ function, these new variables were added to the existing database. These steps were carried out separately for each land coverage variable. Determining the average land coverage for other spatial scales (5 and 4 digit zip code areas) makes it possible to see at later stadia which ones provide the most significant outcomes.

5.2.3 UGS proximity variables

To control for the proximity of UGS amenities in the HP function, the distance to each nearest amenity must be calculated for each individual dwelling. Section 2.1.3 of the literature chapter discusses that three different distance measures, Euclidean, Network and Axial line, can be used to account for the distance. The Axial line distance requires additional qualitative (cognitive environment) data about each individual UGS, which is not available and requires a lot of time. Therefore, this thesis only uses the Euclidean and Network distance, this section discusses the steps and choices that were made to generate the variables.

The Euclidean distance is the easiest to compute, however, for many UGS amenities count that they present value because real estate buyers want to visit them, the Network distance seems more appropriate. Especially in a city like Almere, with a whole bunch of canals and bridges (see Figure 9), this measure appears to be a better choice. However, keeping in mind that the proximity of some amenities, like farmland, might cause noise or smell nuisance, it was decided to also calculate the Euclidean distance. The following section will discuss for each amenity what distance measure is the most appropriate to use.

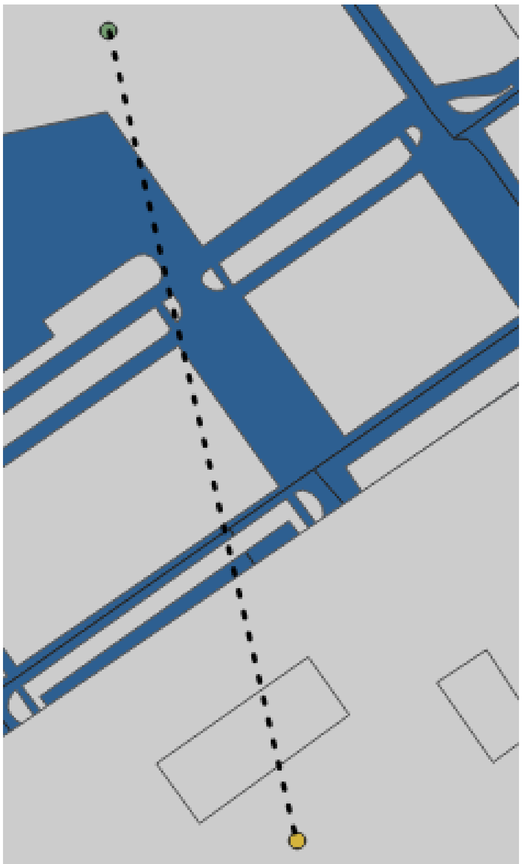


Figure 9: Euclidean distance between random dwelling and parc in Almere

Before the distances can be calculated, it should first be determined between what points the calculation should be done. Unfortunately, with the available data, it was impossible to derive the exact location of each sold residential real estate object. The 6-digit zip code is the most accurate geographical indicator for each dwelling. Because of convenience and time considerations, it was chosen to use the given x – and y coordinate for each zip code from HUGSI to present a point for each dwelling location. The transaction data from the NVM and HUGSI were merged in QGIS 3.16 based on the corresponding 6-digit zip code variable.

The OSM polygon layers are used to determine the location of each UGS amenity. Figure 10 shows how the different amenities are spread over Eindhoven and Almere, while Table 6 contains the numbers of each amenity. According to Czembrowski et al. (2016), assigning the points to the entrances of each UGS amenity improves the measurement accuracy.

However, the exact amenity entrances are unknown, and it would take tremendous effort to find these for each individual amenity. In addition, this manual method is also prone to errors, especially when the researcher is unfamiliar with the area of interest. Therefore, it was decided to use the geometric centroids from each UGS polygon. With the help of the 'geometric tool' function in QGIS 3.16.4, the geometric centroid for each polygon was determined and transformed into a point layer. Next, the 'field calculator' in QGIS allowed assigning each point with the corresponding longitude and latitude value. Finally, the Euclidean distance between the dwelling and amenity points was calculated with the help of the 'distance to the nearest HUB' function.

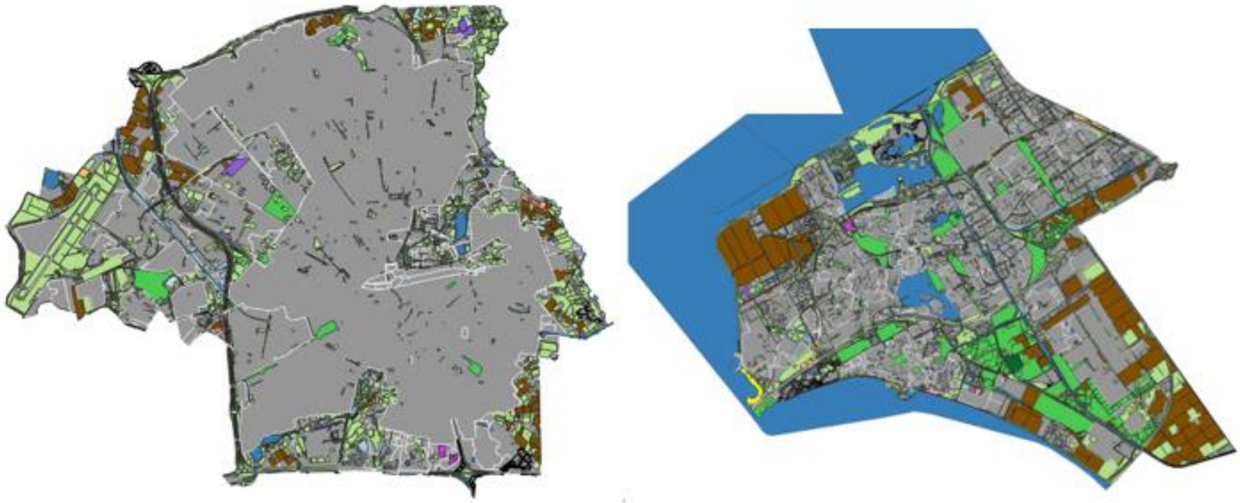


Figure 10: Overview land coverage UGS amenities in Almere and Eindhoven

Allotments (46)	Beach (18)	Cemetery (36)
Farmland (291)	Orchard (87)	Open green space (OGS) (14)
Meadow (47)	Park (180)	Scrub (158)

Table 6: Number of UGS amenities in Almere and Eindhoven (OSM, 2021)

The software program TransCAD was used to calculate the Network distances. To calculate the Network distance, an additional step was needed. Besides, the data points, the road network must also be defined. As the (Pedestrian) Network distance represents the walking distance, a network must be chosen that gives a fair representation of the routes that could be used to walk. On the forehand, it was determined to use the OSM road layer, as this dataset consists of many classes representing different types of roads (see; Table c in Appendix II). This would allow to exclude roads not used by pedestrians, like highways and main roads, or even to specify a network that only includes footpaths and pedestrian streets, for example. Unfortunately, the OSM network proved to be unsuitable for calculating the shortest routes, as many of the starting and ending points did not connect properly to the data points from the urban green space amenities and dwellings in TransCAD.

Therefore, the road network data from NWB was used instead. Unfortunately, the NWB does not provide such an extensive selection of road types, as OSM. Therefore, all the roads were included, this could mean that calculation is done over roads that are not accessible for pedestrians, like highways e.g. In addition, the NWB datasets only include roads that are provided with a street name while footpaths apply that they are often not linked to an official street name, while it could be the most efficient path to take, this especially accounts for land uses were no address information is needed like UGSs. Figure 8 shows that the OSM (black) road network includes more footpaths than the NWB (red) road network. However, despite these facts, the NWB is still considered as a network that could be used to give a fair representation of the Network pedestrian distance.

TransCAD could link the dwelling data with 4.542 network nodes and in the case of the UGS, 180 amenity points with 160 different network nodes.

In other words, not every dwelling point starts at a unique network node meaning that the same distances are calculated for some of the dwellings that might be located in different six-digit zip codes. And some of the unique amenity points share the same network node. In contrast to the 'distance to the nearest HUB' function in QGIS, TransCAD does not only provide the distance to the nearest UGS as an output. Instead, the distances from every transaction point to all the UGS points are calculated. As this research is only interested in the distance to the nearest UGS, excel was used to separate the distances to the nearest UGS from the other data.

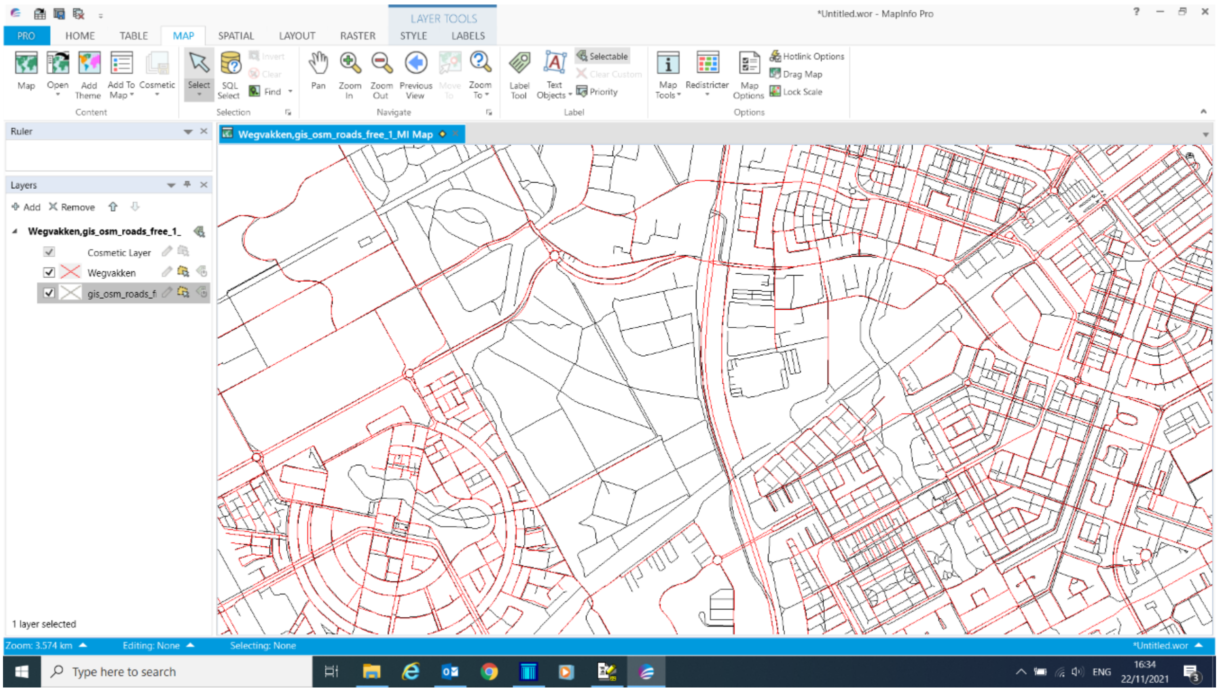


Figure 11: Comparison of the OSM and NWB road network (own figure, 2021)

With the help of different software programmes, the data from HUGSI, OSM and the NWB is transformed into different UGS (proximity) variables available for use in the OLS model (see Table 7). Table d in Appendix II shows the descriptive statistics for the Euclidean and Network distance variables. Remarkable is that for a few amenities, the observed average Network distance is smaller than the Euclidean distance, while the Euclidean distance is the shortest measurable distance between two points. This can be explained by the fact that in TransCAD, the transaction and UGS amenity points have been linked to network nodes that might be located closer to each other. This should be kept in mind when interpreting the results.

UGS class	Measurement types	Measurement level
Allotments	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
Beach	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
Cemetery	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
Farmland	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
Grass	The amount of land covered with vegetation ≤ 1m. Available on 4, 5, 6-digit zip code level.	Continuous from 0 to 1
Orchard	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
OGS	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
Meadow	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
Park	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
Scrub	Straight line distance to the nearest allotment (Euclidean) Distance to the nearest allotment calculated by road (Network)	Continuous in km
Trees	The amount of land covered with vegetation > 1m. Available on 4, 5, 6-digit zip code level.	Continuous from 0 to 1
Water	The amount of land covered with water. Available on 4, 5, 6-digit zip code level.	Continuous from 0 to 1

Table 7: Overview available UGS variables

5.3 Variable selection

This section discusses the theoretical motivation to include a variable in the HP function. In total, 108 explanatory variables are available, that possible all could affect the transaction price of a dwelling. However, including all the variables leads to less precise coefficient estimates due to multicollinearity, as it can be expected that some of the variables will be highly correlated with each other. Therefore, these theoretical choices serve as a first round to exclude some available variables. Later, the multicollinearity analysis will determine if more variables should be excluded. This section also describes if a positive or negative estimate is expected.

5.3.1 Structural variables

In general, most of the transaction price will be explained by the structural characteristics of the dwelling. Therefore, 25 of the 34 available structural variables were seen as relevant and are supposed to enter the HP model to account for its structural characteristics. These can be found in Table 8.

The property's living area in m^2 is included as a log-transformation in the model. This is done because it allows interpreting the estimates as a percentage change in size associated with a percentage change in the transaction price. The plot size in m^2 and volume of the dwelling in m^3 also entered the model as a log-transformation. An additional step for the log-transformation is needed for the plot size variable. There are many observations (6,463), all apartments, that have a plot size of zero in the dataset. As calculating the log of zero gives a mathematical error, a constant of one is added to all plot size values before calculating the log.

No (reliable) data is available for the architecture style, building materials or the maintenance of the dwellings. This makes the building year an indicator for multiple aspects and therefore an important variable. It is expected that the oldest buildings are valued higher due to their historical and architectural value, while newer buildings have fewer maintenance costs because most building materials degrade with age increasing their transaction price. In the NVM dataset, indoor and outdoor maintenance variables are available. However, for both variables, over 80 percent is labelled as poor maintenance or worse. This data's reliability is questioned, therefore, these two variables are excluded from the model.

In order to include the building year of the dwelling, some data transformation is conducted. The NVM dataset consists of nine-time intervals that indicate the construction period of the dwelling. However, during the data exploration, it was found that the categories in the early years were underrepresented compared to the other groups. It is decided to transform the data into five different time intervals, with the buildings built before 1960 as a reference category. The other building periods are included as dummy variables and can be found in Table A in appendix III. For the post-war buildings, that are constructed between 1960 and 1980, a negative price effect is expected. While dwellings built in other construction periods probably show a positive sign.

Variables that contain data about the number of stories, rooms and insulation measures are included in their original numeric form. It is expected that an increase in numbers results in a higher transaction price. This was also expected for the variable indicating the number of toilets. However, during the data exploration, it was found that over 11,000 observations have ‘five or more toilets’, including an outlier with ’13 toilets’. Due to the uncertainty in the interpretation of this data, it was decided to exclude this variable from the model

Other numerical variables, like the number of dormers, rooftop terraces, pantries and balconies are included as a dummy variable that indicates the presence of this characteristic. It is assumed that it would make more sense to include them as a dummy variable instead of numerical. Because it is expected that the price effect between one or four rooftop terraces is neglectable while it makes sense that the presence (or rather the absence) of one of those attributes does have a high impact on the transaction price. Also, for the bathroom and kitchen variables, it was chosen to transform them from numeric into dummy variables. It could be argued that the presence of a second bathroom or kitchen increases the price evidently.
However, only a small percentage of the data has two or more of both attributes and as this thesis is mainly focused on the effect of the UGS variables, it was decided that an indicator for the presence of one or more of both attributes is sufficient.

During the data exploration, it was found that the data distribution of the heating system variable is disproportional. Ten percent of the observation has no heating system. In contrast, 89,9% of the observations have a boiler and the other 0,2% are distributed over the remaining ‘gas or coal stove’ and ‘air or solar-driven heating’ categories. Because of this distribution, it was decided to transform the data into a dummy variable that just indicates the presence of a heating system. Regarding parking space, the NVM provided two variables, one of them indicates if one of the five parking space types is present, while the other variable account for the presence of an indoor parking space. Analysing the data, it was found that indoor parking space corresponds with categories in the other variable. Therefore, it was decided to combine and transform them into two different dummy variables. One that indicates the presence of a parking space and one that indicates the presence of a garage. For both variables, a positive price effect is expected, while it is likely that a garage has a greater impact on the price.

The NVM dataset contains four variables with information about the location of the dwellings. Unfortunately, data about the location of the dwelling in relation to the road, city centre and scenic surroundings are missing from over 10,000 observations. This, combined with the fact the CBS data provide other options to account for the locational characteristics, is used as an argument to exclude these variables from the model. However, the variable that indicates the garden’s orientation in relation to the sun is used. Flemming et al. (2018) found that dwellings with gardens located at a southern orientation point have higher transaction prices than dwellings with a garden located at other orientation points. Therefore, the orientation variable with nine categories was transformed into two different dummy variables, the presence of a garden located at a southern orientation point and the presence of a garden located at another orientation point.

Variable	Description	Measure	Expected impact
Structural characteristics			
Living area (log)	Log of living area in m²	Continuous	+
Plot size (log)	Log of plot size in m²	Continuous	+
Volume (log)	Log of volume in m³	Continuous	+
Building year 1960 - 1980	Building year (0/1)	Binary	-
Building year 1981 – 1990	Building year (0/1)	Binary	+
Building year 1991 – 2000	Building year (0/1)	Binary	+
Building year 2001 ≥	Building year (0/1)	Binary	+
Rooms	Number of rooms	Categorical	+
Stories	Number of stories	Categorical	+
Insulation measures	Number of insulation measures	Categorical	+
Kitchen	Dwelling has a kitchen (0/1)	Binary	+
Bathroom	Dwelling has a bathroom (0/1)	Binary	+
Dormer	Dwelling has a dormer (0/1)	Binary	+
Rooftop terrace	Dwelling has a rooftop terrace (0/1)	Binary	+
Pantry	Dwelling has a Pantry (0/1)	Binary	+
Balcony	Dwelling has a Balcony (0/1)	Binary	+
Basement	Dwelling has a basement (0/1)	Binary	+
Heating system	Dwelling has a heating system (0/1)	Binary	+
Parking space	Dwelling has a parking space (0/1)	Binary	+
Garage	Dwelling has a garage (0/1)	Binary	+
Garden	Dwelling has a garden not located in a South direction (0/1)	Binary	+
Garden South	Dwelling has a garden located in a South direction (0/1)	Binary	+
Monument	Dwelling has a monumental status (0/1)	Binary	+
Ground lease	Dwelling has a ground lease construction (0/1)	Binary	+

Table 8: Structural variables

5.3.2 Locational variables

The location of a dwelling and the available amenities in its surrounding are considered important for many real estate buyers. Therefore, it is essential to take different neighbourhood characteristics into account. For some of the amenities, CBS provided more than one variable (see Table b in Appendix II) that could be added to the model. However, to avoid multicollinearity, it is preferred only to use one variable for each different amenity. In addition, not all amenities might be relevant for real estate buyers. Therefore, this section discusses the process of selecting locational variables. Table 9 shows the 22 locational variables intended to be included in the model.

Shopping

It is expected that real estate buyers are particularly interested in the distance to the nearest supermarket and not the number of supermarkets in their neighbourhood. Obviously, people have different supermarket preferences, however, in general, most supermarkets offer the same products as their competitors. Therefore, it is chosen to take the variable with the closest distance to the supermarket into account and drop the variables that show the average number of supermarkets within the surrounding of the dwelling.This argumentation also applied to the decision to use the variable that shows the distance to the nearest department store. Additionally, the distance measure is more attractive as it could also be used as a proxy for the distance to the city centre, which indicates the distances to central facilities.

In the case of the ‘other stores for daily groceries’ variable, the argumentation goes the other way around. As described in Table b in Appendix II, this variable consists of a variety of store types. Therefore, it is thought that using a variable that only takes the distance to the ‘closest daily groceries store’ will give meaningless results. For example, this could indicate the distance to the nearest bakery for one observation, while it might show the distance to the closest tabaco store for another observation. Instead, the variable that considers the number of stores can be seen as a good neighbourhood indicator that reflects the diversity in retail supply. It is chosen to use the variable that shows the average number of stores within 3 kilometres from the dwelling, as it is expected that most real estate buyers will not travel any further for their daily groceries.

Education and out of school care

When selecting a school, it is expected that people emphasize not only the distance but also other details like education type, religious background and education quality. Therefore, it is thought that it makes more sense to add the variables representing the number of primary schools and high schools in the neighbourhood and drop the variables that show the distance to the nearest school. A real-estate buyer probably, values having a choice when selecting a school. It was decided to leave out the differences between HAVO/VWO and VMBO schools and just look at the high school in general, as these details go beyond the scope of this research. According to the CBS (2018), the average child must travel 1.2 kilometres to elementary school and 2.4 kilometres to high school. Therefore, the average number of elementary schools within 1 kilometres and an average number of high schools within 3 kilometres are taken into the model. For daycare centres and after-school care also counts that it is expected that parents make their decisions on availability more than distance. Because it is a service used to spare time, the average numbers of both amenities within 3 kilometres are considered.

Medical care and emergency facilities

Regarding the medical amenities, it is also chosen to consider the number of available amenities. Because it is expected that the choice for a particular healthcare provider is made with more considerations than just the distance, other essential aspects like room for treatment, does health insurance covers care from the care provider, health care quality, e.g., are expected to be just as important as the distance when selecting a medical amenity.

Regarding the medical amenities, it is also chosen to consider the number of available amenities because it is expected that the choice for a particular healthcare provider is made with more considerations than just the distance. Other aspects like room for treatment, does health insurance covers care from the care provider and health care quality, e.g., are expected to be just as important as the distance when selecting a medical amenity. In general, hospitals have a large catchment area. Therefore it was decided to take the average amount of hospitals within a distance of 20 kilometres into the model. The data exploration found that the hospital variable including outdoor clinics has a higher variance (6) compared to the variable only including hospitals without an outdoor clinic (4). Therefore, it was decided to use the variable including all the hospitals. The average number of community health centres and general practitioner clinics, with the largest distance of 5 kilometres, are also taken into the model.

The CBS (2021) dataset provides a variable that considers the distance to the nearest pharmacy. This variable was taken into the model as most pharmacies offer similar services. It is decided to disregard the distance to the closest fire department as it is expected that the average household does not consider this when buying real estate.

Accessibility to mobility

Economists and planners generally accept that dwelling values tend to increase when it is located in the proximity of (public) transportation hubs because the access to other locations is improved (Kanelleas et al., 2018). Therefore, it is expected that this category, within the locational variables, will show the highest positive price effects. However, negative aspects like noise, pollution, traffic or the safety of living close to a transport hub might negatively affect the transaction prices. Still, a positive price effect is expected for all three 'mobility' variables; distance to the nearest central station, distance to the nearest train station and distance to the nearest highway entrance. Unfortunately, there was no data available on bus stops since this is a good indicator for mobility within the city. The indicators now included are focused on mobility to locations outside the city.

Culture, entertainment, restaurants and other leisure activities

As shown in Table b, in Appendix II, the CBS included a lot of different amenities within the cafeteria, pub, theatre and restaurant variable. Too many differences within this category make the variable inappropriate to use, representing the distance to the nearest amenity. For the first three, the number of amenities within a distance of 3 kilometres are present in the model. While for the theatres the distance of 20 kilometres is used.

Only distance data is available for the following amenities; library, cinema, swimming pool, tanning salon and sauna. This is reviewed as a proper measurement and included in the model. The data supplied for the presence of hotels was not used as it is expected that real estate buyer is not interested in the number of hotels within a radius of 20 kilometres from the property they are considering buying. It is also decided not to use the data indicating the distance to the nearest ice rink as in the data exploration, a variance between 19.1 and 33.3 kilometres is found for the observations in Almere, which indicates that there is no ice rink in Almere. The distance to the nearest pop podium has been left out as this amenity is also part of the 'theatre variable' classification.

Locational characteristics			
Supermarket	Distance to nearest supermarket in metres	Continuous	-
Department store	Distance to nearest department store in metres	Continuous	-
Daily groceries	Number of stores for daily groceries within 3 km from the dwelling	Categorical	+
Primary schools	Number of primary schools within 1 kilometre from the dwelling	Categorical	+
High schools	Number of high schools within 1 kilometre from the dwelling	Categorical	+
Day care centre	Number of day care centres within 3 kilometres from the dwelling	Categorical	+
Afterschool care	Number of centres within 3 kilometres from the dwelling	Categorical	+
Hospital	Number of hospitals within 20 kilometres from the dwelling	Categorical	+
General practitioner	Number of general practitioners within 5 kilometres	Categorical	+
Pharmacy	Distance to nearest pharmacy in metres	Continuous	-
Central train station	Distance to nearest central train station in metres	Continuous	-
Train station	Distance to nearest train station in metres	Continuous	-
Highway	Distance to highway in metres	Continuous	-
Cafeteria	Number of cafeterias within 3 kilometres	Categorical	+
Pub	Number of pubs within 3 kilometres	Categorical	+
Restaurant	Number of restaurants within 3 kilometres	Categorical	+
Theatre	Number of theatres within 20 kilometres	Categorical	+
Library	Distance to nearest library in metres	Continuous	-
Cinema	Distance to nearest cinema in metres	Continuous	-
Swimming pool	Distance to nearest swimming pool in metres	Continuous	-
Tanning salon	Distance to nearest tanning salon in metres	Continuous	-
Sauna	Distance to nearest sauna in metres	Continuous	-

Table 9: Locational variables

5.3.3 Urban green space variables

The specification and inclusion of structural and locational variables are done to specify a correct model that gives valid estimates. In contrast, the UGS variables that outline the relationship between real estate and UGS are the main interest of this thesis. This section will elaborate on the variables that describe the relationship between the dwellings and urban green amenities. A description of each variable will be given, the expected impact of each variable and why they are included in the model. In total, 13 UGS variables are intended to be incorporated into the model and can be found in Table 10.

Scenic value (density variables)

This study follows the suggestion of Kolbe & Wüstermann (2014; 2016), that state that UGS also consist of scenic value for real estate buyers. To test this suggestion, this thesis will add a continuous variable to the model that indicates the share of land surrounding the dwelling covered by waters, trees or grass. An increase in land covered by water, trees or grass is assumed to result in a higher transaction price. Expected is that trees and water contain more scenic value and will have a greater price effect than grass.

Unfortunately, it is impossible to make clear distinguishments between quality differences within the trees, grass and variables. At the same time, this thesis is aware of the fact that those are no heterogeneous goods. And there might be poorly maintained grass fields or polluted waters within the data set. Still, this thesis assumes that the vast majority of the water, grass and tree coverage brings a scenic value that real estate buyers apricate. However, at last, there is another variable that does help to say something about one of the characteristics of urban green. A variable that indicates the health of urban green, based on NIR radiation, is added. In section 5.3.1, the concept of NIR-radiation is explained. An increase in health scores is expected to lead to a higher transaction price.

A backward regression should point at what spatial scale the density variables will be added (6 digit-zip code, 5 – digit zip code or 4-digit zip code).

Proximity principle (distance variables)
As explained in the previous section, the Euclidean and Network distance between the dwellings and the nearest urban green amenities are derived. This section will elaborate on the choice of distance measure for each different urban green amenity.

Allotments
Allotments are community gardens available for individuals to grow food plants. They are often cultivated by neighbouring residents that do not have their own garden or do not have enough space for growing crops. Approximately 35% of the observations in the dataset do not have a garden. Therefore, real estate buyers are expected to appreciate living at a location with allotments within a travelable distance. The Network distance variable is used in the model to account for this relationship. Only two studies took the distance to allotments into account with opposite findings. In 2016, Czembrowski & Kronenberg took the walking distance to the nearest entrance of a cemetery or allotment into account. Their study results suggest that allotments are perceived as unwelcome by real estate buyers as the apartment price increases whenever the distance increases. However, this study used a variable that covered the distance to the nearest entrance of an allotment or a cemetery, while both amenities might be perceived differently.

In addition, Czembrowksi & Kronenberg (2016) stress that the political and historical context of allotment gardens in their study case (Łódź) might be an important cause for their results. Another study used allotments with forests, parks, woods and cemeteries to represent UGS as one single variable in their hedonic pricing study (Liebelt, Bartke & Schwarz, 2017). They argued that all these land cover types provide recreational services to some extent and found a positive relationship. This thesis also expects to see a positive price effect; whenever the distance to the allotment increase, the transaction price will decrease.

Beaches
Beaches are defined by Openstreetmap (2022) as a landform along a coast or other water body that consists of sand, gravel pebbles, cobblestones or shell fragments. In Eindhoven, only two land plots are assigned as 'beaches' at the outskirt of the city near a lake. The city of Almere consists of more 'beaches' located at water bodies at the city borders. This fact might explain the high average distances to the beaches compared with the other UGS amenities, as shown in Table 9. Despite the large distances, it is still expected that beaches are a green amenity valued by real estate buyers due to their multiple recreational purposes. It can be visited for a stroll or to hang out, while it is also used for different (water) sports and some of the beaches in Almere even contain hospitality operators. Therefore, it is decided to use the Network distance in the model to test if real estate buyers in Almere and Eindhoven take the recreational value of beaches into account while buying a dwelling. A positive price effect is expected.

Cemeteries
Studies have shown that cemeteries are unwanted green amenities that negatively impact real estate prices (Tse and Love, 2000; Anderson & West, 2006; Czembrowski and Kronenberg, 2016). The sentiment linked to these places, which is established through its particular purpose, might cause people not to appreciate living close to them. However, most of the cemeteries in this study, especially in Almere, are spacious, green and quiet, which could be valued to use for a daily stroll e.g.
The Network distance is used to consider that cemeteries are often green, quiet locations that might be used for recreational walks. However, this study is unsure whether to expect a positive or negative price effect.

Farmland
Farmlands are areas that are used for agricultural purposes. Kolbe & Wüstemann (2014) found that an increase in the Euclidean distance to the nearest farm led to the rise in price, while also the dwelling price decreases when the density of farmland surrounding the farm increases.

This is explained by the fact that using fertiliser or pesticides is associated with a negative impact on the living environment. This thesis also expects a negative effect from Farmland on the transaction price of residential real estate. Therefore it makes more sense to use the Euclidean distance as noise and (air) pollution nuisance do not travel by road.

<u>Orchard</u>
Orchards are areas used for the intentional planting of trees and shrubs maintained for food production (OpenStreetMap, 2022). As far as this thesis is concerned, there are no studies yet that tested the price effect of Orchards on their surrounding real estate. As Orchards are quiet and calm places that use fewer fertilizers and pesticides than agricultural farmlands, it is expected that real estate buyers do appreciate Orchards in their surroundings. Therefore, the Network distance is used to test this relationship.

<u>Open green spaces</u>
As stated in section 3.3.1, the OGS are open green spaces used as a general recreation ground, which could include formal or informal pitches, nets and other recreation equipment (Openstreetmap, 2022). Because urban planners design these places for recreational purposes, it is expected that OGS have the highest impact on the transaction prices compared to other green amenities. The Network distance will be used as the distance measure.

<u>Meadow</u>
Meadows are areas of land primarily vegetated by grass and other non-woody plants, mainly used for grazing by cattle or hay (Openstreetmap, 2022). Nilsson & Pia (2008) assessed a hedonic pricing study to find the impact of meadows on house prices in a rural setting. They found that houses located in the immediate surrounding meadows were on average 2.6% more expensive than other properties. This thesis is interested in finding out if the proximity of meadows in urban areas also leads to higher transaction prices. On one hand, it could be argued that the presence of cattle could bring smell nuisance. However, it is expected that real estate buyers see the recreational value that these areas bring, as they are peaceful environments compared to crowded urban areas. The Network distance is chosen to use in the hedonic price analysis.

<u>Park</u>
Many studies proved that the proximity of parks leads to higher transaction prices. Most of the studies attribute this to the recreational nature of the parks. This study also expects that the proximity of parks leads to an increase in the real estate price, the Network distance will be used to take the proximity into account.

<u>Scrub</u>
Scrubs are areas of uncultivated land covered with shrubs, bushes or stunted trees (Openstreetmap, 2022). A few studies in the USA tested if living close to a scrub area increases house prices. Doss & Taff (1996) found a positive relationship, while Mahan, Polasky & Adam (2000) found a negative price effect for living close to scrub areas. However, both studies are old and conducted in Wetland areas, while this thesis is focused on urban areas. This thesis will use the Network distance to test the effect of the proximity of scrub on real estate in an urban area. It is difficult to appoint if the proximity to scrub brings a positive or negative impact. On one hand, it could be argued that uncultivated land improves biodiversity and is good for local nature. However, it is also conceivable that people do not apricate scrubs close to their dwelling, despite a place to walk it offers no other recreational activities and some people might interpret these places as unfinished and messy areas that should be developed.

UGS characteristics			
Allotments	Network distance to nearest allotments in metres	Continuous	-
Beach	Network distance to nearest beach in metres	Continuous	-
Cemetery	Network distance to nearest cemetery in metres	Continuous	+/-
Farmland	Euclidean distance to nearest farmland in metres	Continuous	+
Orchard	Network distance to nearest orchard in metres	Continuous	-
Open green space	Network distance to nearest OGS in metres	Continuous	-
Meadow	Network distance to nearest meadow in metres	Continuous	-
Park	Network distance to nearest park in metres	Continuous	-
Scrub	Network distance to nearest scrub in metres	Continuous	+/-
Grass	Share of land covered with grass within the 6-digit zip code from the dwelling	Continuous	+
Trees	Share of land covered with trees within the 6-digit zip code from the dwelling	Continuous	+
Water	Share of land covered with water within the 6-digit zip code from the dwelling	Continuous	+
Health	Health of the vegetation within the 6-digit zip code from the dwelling based on NIR-radiations	Continuous	+

Table 10: UGS variables

5.3.4 Control variables

The rest of the explanatory variables included in the model are added to control for other factors, from which it is expected that they influence the transaction price. As discussed, the transaction data consist of different sub-markets. Besides the sales of the apartments, it contains different residential houses that are divided into the following subcategories; terraced houses (8.880), semi-detached houses (773), corner houses (2.832), duplex houses (1.774) and detached houses (1.302). As discussed in section 3.1, mixing different housing markets will not lead to the actual marginal value of any of the markets but to all markets involved. Nevertheless, it is decided not to estimate different models for each individual sub-market but run one model for the different types of real estate. In the end, this thesis is interested in putting a monetary value on urban green spaces and explaining the value effects of urban green spaces on real estate instead of explaining the transaction prices of a certain sub-market. However, to control for the different markets, it was decided to add a dummy variable for each housing type where apartments are used as the reference category.

The data provided by the NVM contains dwelling transactions that were conducted between 2014 and 2018. According to CBS (2021), there has been an upward trend concerning the transaction prices of houses in the Netherland since 2013. In this period, the average house prices increased by 26% in Almere and 20% in Eindhoven (CBS, 2021). To control for this general price increase, it was decided to use dummy variables that indicate the transaction year. The first transaction year, 2014, is taken as the reference category, it is expected that the transaction year variables show an ascending positive price effect equal to the inflation trend from the CBS.

In total, 9 control variables, shown in Table 11, are intended to enter the hedonic price function.

Control variables			
Terraced	Building type (0/1)	Binary	-
Semi-detached	Building type (0/1)	Binary	+
Corner	Building type (0/1)	Binary	+
Duplex	Building type (0/1)	Binary	+
Detached	Building type (0/1)	Binary	+
2015	Transaction year (0/1)	Binary	+
2016	Transaction year (0/1)	Binary	+
2017	Transaction year (0/1)	Binary	+
2018	Transaction year (0/1)	Binary	+

Table 11: Control variables

5.3.5 Overview explanatory variables

Table A in Appendix III shows the selected 25 structural, 22 locational, 13 UGS and 9 control variables that are intended to be included in the final OLS model. However, the variables first need to be tested on multicollinearity to ensure that all the explanatory variables can be included within the model without a high correlation between them. But first, before the multicollinearity analysis is conducted, the following section discusses the outlier analysis.

5.4 Outlier analysis

Linear regression is sensitive to outliers and therefore, it is important to check for their presence of them. Outliers are observations in the data that are well separated from the rest of the data (Field, 2009). Some general steps were taken to remove strongly deviant observations. First, observations (20) with a transaction price of '€1' were deleted. Also, the observations (281) with a missing value for the living area (m²) have been removed.

The locational data from the CBS was missing for 112 objects, of which 74 are located in Eindhoven and 38 in Almere, spread across 32 different 6 digit-zip code areas. According to CBS (2020), the average statics are only calculated when the statistics of at least 90% of the dwellings within a zip code area are known. As the map in Figure 12 shows, the missing values are pretty clustered. Looking at Almere, the areas are located on the outskirts of the city. This, in combination with the fact that most of these transactions date from 2018, could indicate that these are newly built houses, of which data from 2017 is missing. No logical explanation could be found concerning the dwellings in Eindhoven because they are clustered in the city centre.

Figure 12: 6-digit zip code areas with missing values for the CBS locational data (own input, 2021)

The transaction data shows an average transaction price of €226,796 with a standard deviation of €104,152. It was decided to remove the strongly deviant transactions to ensure that the properties belong to the same housing market. As a rule of thumb, it was decided to remove the transactions with a value greater than three times the standard deviation. In total, 416 transactions were removed with a price greater than €539,265. Concerning the plot size, observation with a surface larger than 1000 m² (66) was deleted. This cut-off value is based on the plot-size distribution and what could be considered normal but still somewhat arbitrary.

5.4.1 Spatial outlier analysis

During normal data exploration, four 'extreme' outliers were found for the variables that indicate a distance. All four outliers were located in the same 6-digit zip code in Almere (see orange point in figure 13). With additional spatial analysis, the extreme values are easily explained. All four transactions were done in one of the zip code areas with a huge surface at the north edge of the city boundaries. However, the geographical boundaries of this zip code area give a biased reflection of the real situation. In fact, most of the surface in this area is part of a large lake, het Markermeer.

The sold dwellings are probably located at the border of this zip code area, but due to the geo-tag process, the data is presented if the houses are located far in het Markermeer. These outliers were removed from the data set.

Further spatial analysis noticed other observations that were found inappropriate to keep in the analysis. As this research is interested in the urban environment, observations located in an area with a non-urban land use area were deleted. Most of these observations could be found at the border regions of the municipalities. Due to the geo-tag process, some observations are located in an area where an urban green space land use is allocated. Because this also gives a distorted picture of the distances between the dwelling and UGSs, it was decided to drop those observations from the analysis. Based on these assumptions, the spatial analysis deleted 141 observations spread across 22 six-digit zip codes in Eindhoven and 18 in Almere. These are indicated as the red points in Figure 13.

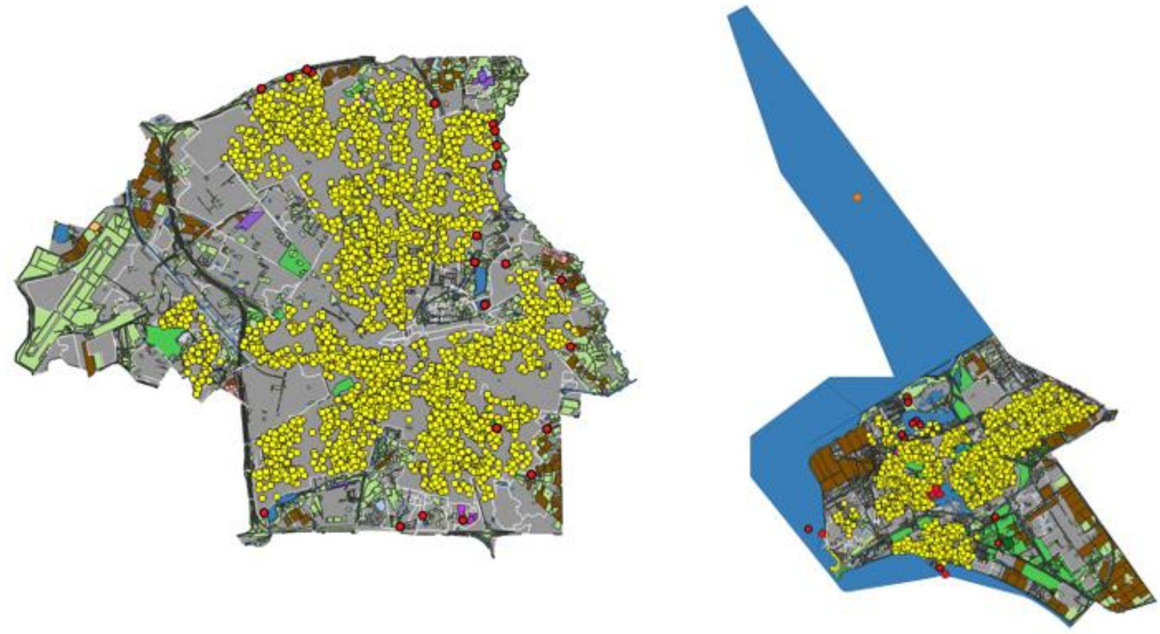


Figure 13: Transaction points (yellow = included transaction data, red and orange = deleted transaction data)

5.4.2 Cook's Distance

The (spatial) outlier analysis already helped to trim the data sample and remove the outliers with large residuals. An additional statistic is used to observe single observations and their impact on the model's ability to predict all cases. Cook's distance measures the overall influence of a single observation on the predicated model estimates (Field, 2009). It is not designed to detect outliers, but it measures the effect of deleting a specific observation from the overall sample. Observations with large residuals or high leverage may distort the outcome and accuracy of the regression. Therefore, Cook's distance can be used to detect single observations that should be reviewed. The Cook's distance statistic (D) is calculated by subtracting the predicted value of observation *j* when observation *i* has been omitted from the predicted value of *j* based on the total sample (Cook, 1979). SPSS has a function to automatically calculate the Cook's Distance for each observation when estimating an OLS regression. If the Cook's distance value for an observation is large, it indicates that deleting this observation will greatly impact the parameter estimations. An often used threshold to indicate if the Cook's distance for an observation is large is $D_{\mathrm{i}} > \frac{4}{N}$, where *N* represent the number of observations in the total sample (Field, 2009). Following this cut-off value, 1,262 observations were found with a large Cook's Distance.

	Total observations		$\boldsymbol{D_{\mathrm{i}} > \frac{4}{N}}$ **observations**	
City				
Almere	9,688	(46.7%)	363	(28.3%)
Eindhoven	11,035	(53.3%)	918	(71.7%)
Total	20,723	(100%)	1,281	(100%)
Building year				
≤ 1959	3,136	(15.2%)	436	(34.0%)
1960 – 1980	4,468	(21.6%)	257	(20.1%)
1981 – 1990	4,173	(20.1%)	173	(13.5%)
1991 – 2000	4,876	(23.5%)	168	(13.1%)
≥ 2001	4,070	(19.6%)	247	(19.3%)
Total	20,723	(100%)	1,281	(100%)
Transaction year				
2014	3,470	(16.8%)	271	(21.1%)
2015	4,045	(19.5%)	300	(23.4%)
2016	4,604	(22.2%)	271	(21.2%)
2017	4,442	(21.4%)	242	(18.9%)
2018	4,162	(20.1%)	197	(15.4%)
Total	20,723	(100%)	1,281	(100%)

Table 12: Most remarkable differences between total observations and observations with a large Cook's distance

Analysing the observations with a large Cook's distance found that these dwellings do not systemically differ from the complete data set (see Table 12). The biggest difference is that the houses with a large Cook's distance are disproportionality located in Eindhoven (71.7%), compared to (51.3%) in the complete data set. However, Figure 11 shows that observations in both sets are distributed similarly across the cities. Something else that stands out is that in the total data sample, 42.2% of the dwellings are terraced houses. At the same time, the observations with a large Cook's distance are more equally distributed over the different housing types, with only 27,7% of the observations appointed as a terraced house. No other notable differences were found.

Because removing the observations with a large Cook's distance does not lead to the underutilization of certain groups in the final sample and is justified regarding the total observations, it was decided to delete all 1,262 observations $D_{\mathrm{i}} > \frac{4}{N}$.

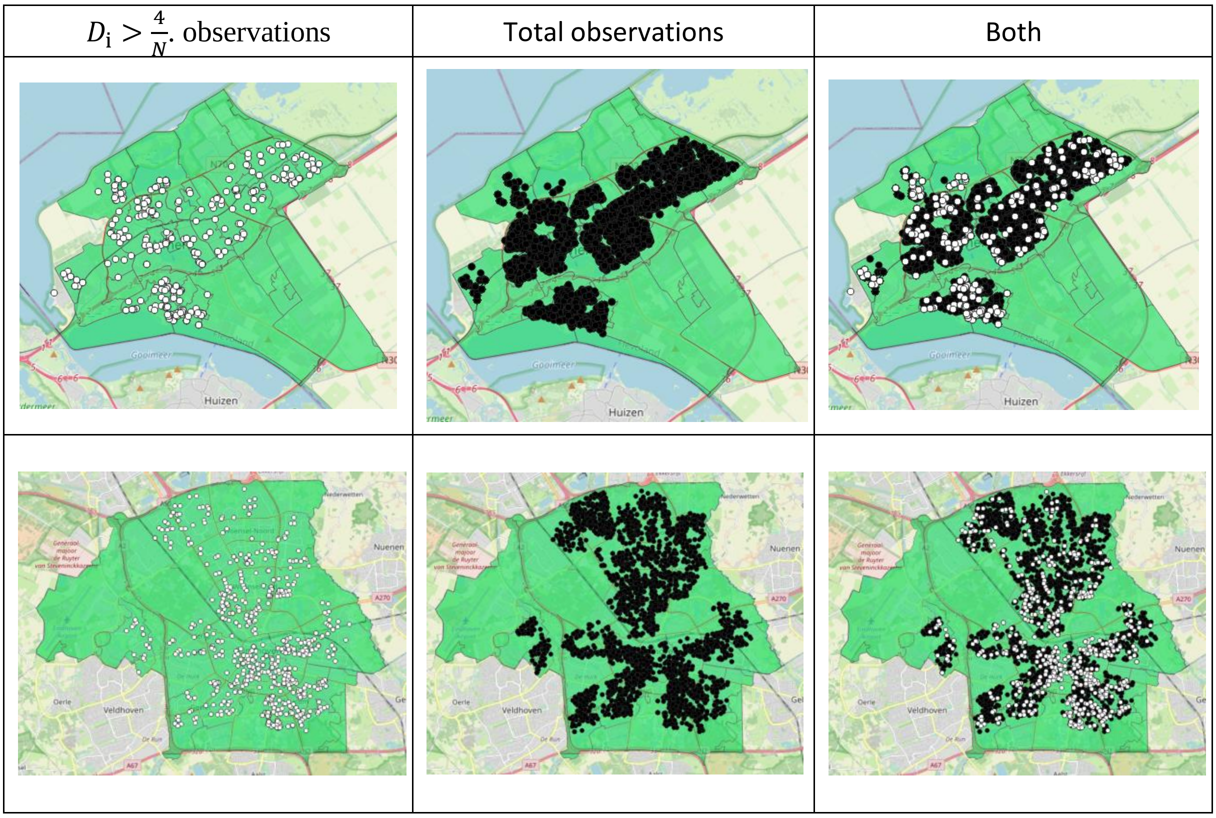


Figure 14: Comparison data distribution total observations and observation with a large Cook's distance

5.4.3 Final data sample

This resulted in a final data sample of 19,461 transactions, with 5,831 unique zip codes. Table 13 shows how the transaction data is distributed across time and over the different housing segments. Unfortunately, it is unknown if there are any repeated sales in the dataset as the NVM provided each sale with an unique property ID.

	Eindhoven				
	2014	**2015**	**2016**	**2017**	**2018**
Apartment	610 (34.8%)	749 (38.2%)	844 (37.5%)	741 (34.8%)	689 (33.5%)
Terraced house	685 (39.1%)	714 (36.4%)	863 (38.4%)	843 (40.0%)	866 (42.1%)
Semi-detached	62 (3.5%)	58 (3.0%)	62 (2.8%)	74 (3.4%)	75 (3.6%)
Corner house	207 (11.8%)	242 (12.3%)	279 (12.4%)	272 (12.7%)	245 (11.9%)
Duplex house	129 (7.4%)	151 (7.8%)	149 (6.6%)	149 (7.0%)	144 (7.0%)
Detached house	58 (3.4%)	46 (2.3%)	51 (2.3%)	45 (2.1%)	39 (1.9%)
Total	1,751	1,960	2,248	2,124	2,058
	Almere				
	2014	**2015**	**2016**	**2017**	**2018**
Apartment	354 (24.3%)	462 (25.5%)	453 (21.7%)	425 (20.6%)	391 (20.5%)
Terraced	661 (45.7%)	798 (44.1%)	1,012 (48.5%)	986 (47.9%)	936 (49.0%)
Semi-detached	56 (3.8%)	51 (2.8%)	82 (3.9%)	71 (3.4%)	64 (3.4%)
Corner	218 (14.9)	278 (15,3%)	295 (14.2%)	307 (14.9%)	283 (14.8%)
Duplex	119 (8.1%)	147 (8.1%)	170 (8.1%)	187 (9.2%)	153 (8.0%)
Detached	48 (3.3%)	77 (4.2%)	75 (3,6%)	83 (4.0%)	83 (4.3%)
Total	1,456	1,813	2,087	2,059	1,910

Table 13: Included transaction data observations across time and housing market-segment

5.5 Multicollinearity

An important OLS assumption is that the mutual correlation between the explanatory variables should be too strong, so the data must not show multicollinearity, which occurs when two or more independent variables are highly correlated with each other. Considering the nature of the data, it was expected that some of the UGS and locational variables would correlate.

Two methods are used to detect multicollinearity in our data. First, the correlation was checked between all pairs of x-variables. Pearson's correlation coefficient (*r*) is a statistical measure that indicates the extent to which two or more variables move together. The static goes from -1, indicating a negative correlation, to 1, showing a positive correlation. Positive correlations show that variables increase or decrease together, while negative correlations report that if one variable increases, the other decreases, and vice versa. This thesis used the index from Rumsey (2016) to interpret the strength of the linear relation between two variables (see: Table 14).

Strength relationship	**Pearson's correlation coefficient**
Perfect positive relationship	1
Strong positive relationship	0.7 < r < 1
Moderate positive relationship	0.3 < r < 0.7
Weak positive relationship	0.3 < r < 0
No linear relationship	0
Weak negative relationship	-0.3 < r < 0
Moderate positive relationship	-0.7 < r < - 0.3
Strong negative relationship	-0.7 < r < -1
Perfect negative relationship	-1

Table 14: Interpretation Pearson's correlation coefficient (Rumsey, 2016)

The variables that show a strong relationship should be reviewed. According to Field (2009), the correlation matrix should only be scanned for variables with a Pearson's correlation of 0.8 or higher. As there is no general cut-off score for what is too much correlation, this thesis also made use of another collinearity diagnostic.

The Variation Inflation Factor (VIF) indicates if a predictor has a strong relationship with other predictors. It does not only look at the bivariate relationship between the x variables, but it runs separate regressions where each explanatory variable is taken as the dependent variable to detect possible multicollinearity that could cause problems with the interpretation of the model or for the model fit (Field, 2009). To reduce multicollinearity, some variables could be transformed into different forms like logarithmic, but in some cases, it is better to delete one of the correlated variables from the model. A commonly used rule of thumb is that variables with a VIF value of 10 should be removed. However, just like the Pearson statistic, there is no general agreement on this threshold (O'brien, 2007).

Variable	**VIF**
Cafeteria	93.9
Restaurant	76.0
Health	64.2
Pub	57.6
Trees	50.6
Train station	47.2
Afterschool care	41.5
Theatre	39.6
Day care	33.4
Cinema	28.5
General practitioner	28.5
Hospital	22.6
Cemetery	15.2
Beach	11.4
Plot size (Log)	11.6
Terraced	11.1

Table 15: VIF scores ≥ 10

A backward regression with all the variables discussed in section 4.1 was used to find the optimal OLS model. Some explanatory variables like the distance to the nearest park, bathroom, garden and train station were not taken into the model because their coefficients estimates were insignificant. It was found that some of the variables were highly correlated, Table 15 shows all variables with a VIF score higher than 10, including some extreme cases. To deal with this multicollinearity and reduce the VIF scores to an acceptable level, the following decisions were made:

- The high VIF scores for the 'trees' and 'health' variables are caused by their mutual correlation (r = (1) 0.931). First, it was tried to combine both variables in different dummy variables, indicating the presence of healthy trees, unhealthy trees and so on. However, it was hard to make a proper distribution as the majority of the areas have no or a low percentage of trees and most areas score low on health, the majority of these dummy variables gave statistically insignificant estimates. Therefore, it was decided to drop the 'health' variable as this variable also had some high Pearson scores compared with other variables.
- The control variable that indicates if the dwelling is a terraced house seems highly correlated with the 'plot size variable'. Their mutual Pearson score is only (-) 0.441, but whenever both variables are taken into the model, it results in VIF values higher than 10 for both variables. As many studies prove that plot size greatly impacts the dwelling price, it was chosen to drop the 'terraced house' variable.
- Table B in Appendix III shows that the 'beach' and 'cemetery' variables had a VIF score higher than 10. As the UGS variables are the main focus of this thesis, it was essential to keep their correlation with other explanatory variables low. To reduce their VIF scores to an acceptable level, 11 (locational) variables were removed from the model.

This leaves 44 explanatory variables that enter the final OLS model, their descriptive statistics can be found in the next section. The final VIF scores of the included variables can be found in table b in Appendix III. Except for the plot size variable (5.9), all the VIF scores are lower than 5. A possible explanation for the relatively high VIF value for the plot size variable could be that it is probably highly correlated with the apartments, as all apartments have a plot size of zero. However, the Pearson coefficient could not be checked as the apartments were taken as a reference category.

Even though a large number of variables discussed in section 4.1 are omitted, it seems that the final selection is well balanced. Most structural variables are still in the model, while the different locational characteristics (mobility, shopping, e.g.) are still represented by at least one variable. Unfortunately, keeping the model's 'park' and 'health' variables in the model was impossible.

5.6 Summary included variables with descriptive statistics

Summarizing the previous sections, Table C in Appendix III shows the complete list of included explanatory variables that entered the OLS model with their description, measure type and expected price effect. Eventually, 20 structural variables, 8 locational variables, 11 UGS variables and 8 control variables entered the final OLS model. The descriptive statistics of the variables are shown in Table 16 below.

Variable	Mean (for dummy number of observations where *i = 1*)	Median (for dummy % of N where *i = 1)*	St. dev	Min	Max
Dependent variable					
Transaction price (log)	12.21	12.20	0.38	10.95	13.2
Structural characteristics					
Living area (log)	4.68	4.71	0.31	3.40	5.79
Plot size (log)	3.49	4.94	2.46	0.00	6.88
Building year 1960 - 1980	4,468	21.6%			
Building year 1981 – 1990	4,173	20.1%			
Building year 1991 – 2000	4,876	23.5%			
Building year 2001 ≥	4,070	19.6%			
Rooms					
Insulation measures					
Kitchen	18,027	87%			
Dormer	2,926	14.1%			
Rooftop terrace	1,976	9.5%			
Pantry	2,115	10.2%			
Balcony	4,784	23.1%			
Heating system	18,508	89.3%			
Parking space	3,381	16.%			
Garage	3,584	17.3%			
Garden	6,591	31.8%			
Basement	218	1.1%			
Monument	45	0.2%			
Ground lease	195	0.9%			
Locational characteristics					
Supermarket	0.66	0.60	0.42	0.0	2.9
Department store	2.22	2.10	1.14	0.0	7.5
Pub	4.16	0.70	9.15	0.0	70.2
Primary schools	2.07	2.00	1.26	0.0	7.0
Pharmacy	0.92	0.80	0.56	0.0	3.2
Central train station	2.57	2.20	1.46	0.1	6.3
Highway	2.05	1.90	0.92	0.2	4.2
Cinema	3.88	3.50	2.01	0.0	9.6
UGS characteristics					
Allotments	1.65	1.54	0.85	0.02	5.06
Beach	3.70	3.30	1.95	0.07	9.10
Cemetery	2.99	2.01	2.62	0.01	11.21
Farmland	1.55	1.34	0.89	0.05	4.51
Orchard	1.72	1.69	0.72	0.02	3.53
Open green space	2.16	2.09	0.99	0.02	5.88
Meadow	2.06	2.04	0.91	0.01	5.94
Scrub	2.22	1.61	1.63	0.01	7.24
Grass	0.02	0.00	0.07	0.00	0.77
Trees	0.14	0.07	0.18	0.00	0.98
Water	0.01	0.00	0.06	0.00	0.76
Control variables					
Semi-detached	723	3.5%			
Corner	2,771	13.4%			
Duplex	1,611	7.8%			
Detached	714	3.4%			
2015	4,045	19.5%			
2016	4,604	22.2%			
2017	4,442	21.5%			
2018	4,162	20.1%			

Table 16: Included variables with their descriptive statistics

5.7 Weight matrix

As discussed in chapter four, this thesis uses the max-min distance to specify the spatial weight matrix. With the dataset of this thesis, this leads to a max-min distance of 330 metres. Table 17 shows descriptive statistics of the spatial weight matrix. The main criticism of using this max-min distance is that it generates many neighbours for clustered units (Anselin & Rey, 2014). Looking at table 17, this also seems to apply in this case. However, as the mean number of neighbours is close to the average number of houses in each neighbourhood (#observations (19,641) / total number of neighbourhoods in Almere and Eindhoven = 105) and a radius of 330 metres around each dwelling sound as a fair reflection what people consider as their direct neighbours, this specification of the spatial weight matrix seems justified.

Descriptive statistics $W = distance \leq 330\ metres$	
Observations	19,641
Min neighbours	2
Max neighbours	268
Mean neighbours	107
Median neighbours	104

Table 17: Max-min spatial weight specification (authors elaboration using GeoDa)

5.8 Heatmaps

Engström & Gren (2017) used a heatmap for an in-depth analysis of the transaction data in relation to the UGS variables. Heatmaps are easy to construct and can be used for the identification of 'hotspots' and clustering of points. Therefore, this thesis decided to create heatmaps for the transaction data and the density of UGS (grass, trees and water). Figure 12 includes a map with all the individual spatial observations and three different heatmaps for both Almere and Eindhoven. The left column contains the maps for Almere, while the right column includes the Eindhoven maps. (Dark)red zones indicate clusters of observations with high values.

Looking at Table 18 on the next page, it can be found that clusters of areas with a high percentage of land covered with water can be found in the North-East part of both cities. In Eindhoven, a cluster of observations with a high percentage of land covered with grass can is located just north of the city centre, while in Almere, different clusters regarding the amount of grass are spread around the entire city. Considering the amount of land covered with trees, the observations in Eindhoven have the strongest cluster North of the city centre which is equivalent to the grass cluster. However, compared with the distribution of grass, it can be found that the clusters for trees are also located upper North and South East of the city, while the clustering for grass is also present in the South East. In Almere, the tree clusters show a similar pattern to the clusters for grass. At last, the heat map concerning the dwelling transaction prices shows where observations with high prices are clustered. In Eindhoven, the heatmap seems to have a similar pattern as the heatmap for tree coverage, while the transaction price clusters in Almere follow a similar distribution as both the tree and grass distribution. This enhances the expectation that the presence of trees and grass positively affects a house's transaction price.

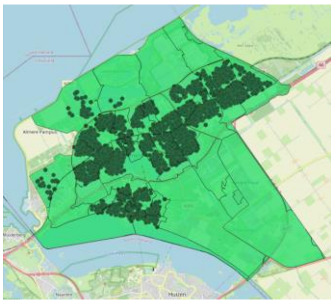
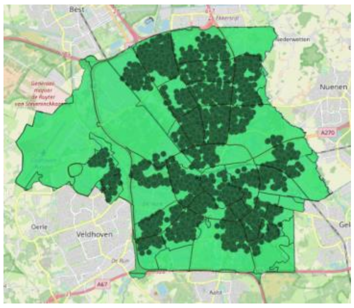
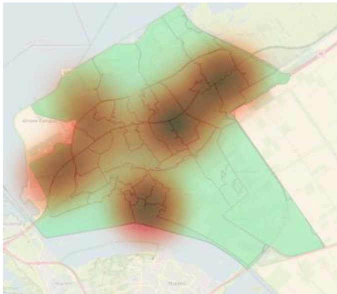
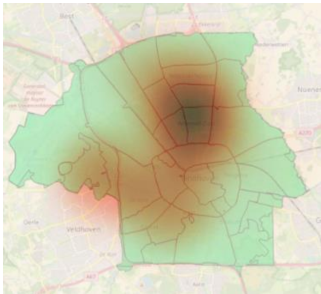
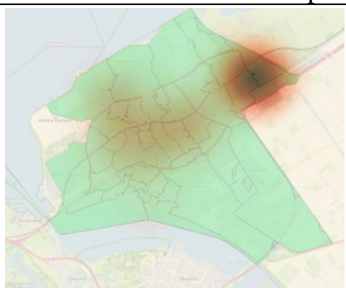
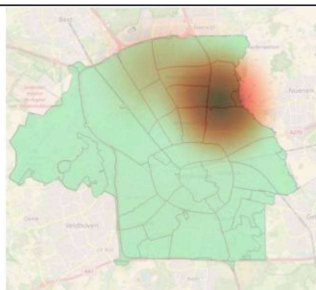
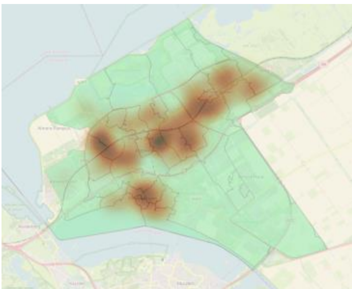
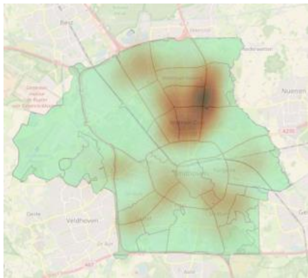
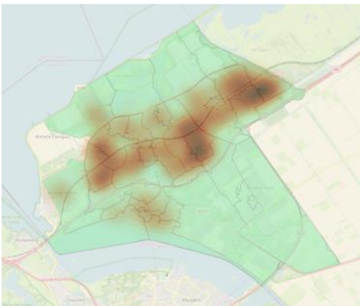
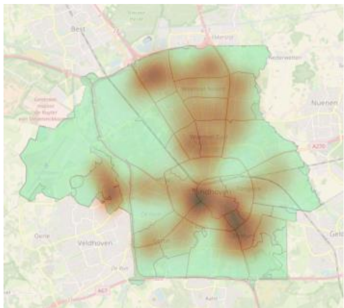
<table>
<tr><td colspan="2">Observations in Almere and Eindhoven</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">Heat maps % grass in Almere and Eindhoven</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">Heat maps % water in Almere and Eindhoven</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">Heat maps % trees in Almere and Eindhoven</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">Heat maps transaction prices in Almere and Eindhoven</td></tr>
<tr><td></td><td></td></tr>
</table>

Table 18: Heatmaps with the transaction data and UGS density variables in Almere & Eindhoven

5.9 Summary

This chapter has described the available data and (possible) variables for inclusion in the hedonic pricing function. First, the content of the five available data sources was described in detail. This is followed by an explanation of how the Euclidean and Network distance is calculated between the dwelling and UGS amenities. Next, it was described what variables were intended to enter the hedonic price function and what impact for each variable was expected. From a theoretical point of view, it was intended that 25 structural, 22 locational, 13 UGS and 9 control variables enter the hedonic price model. However, during the multicollinearity analysis, it was found that some variables were highly correlated. With the help of the VIF score and Pearson's correlation coefficient, the variables with the highest multicollinearity scores were detected and removed. Resulting in 20 structural variables, 8 locational variables, 11 UGS variables and 8 control variables that entered the final OLS model. Subsequently, an overview with a list of the included explanatory variables with their descriptive statistics was given. The chapter ends with the descriptive statistics of the estimated weight matrix and an analysis of heatmaps created from the transaction data and UGS density variables

.

CH 6 Results

This chapter describes the outcomes of the statistical tests followed by the model results. The first section of this chapter starts with the results of the Moran's I, Lagrange Multiplier and Likelihood ratio tests. The results of the OLS and spatial models are given in the second section. This includes a discussion about the model summaries and general model outcomes. Followed by a brief explanation of how the coefficients should be interpreted before all individual variable estimates are interpreted individually.

6.1 Test statistics

As discussed in chapter four, different statistical tests should be performed to determine which (spatial) model fits the data the best. The Moran's I, Lagrange Multiplier and Likelihood ratio tests have been conducted. Their results are described in the following sections.

6.1.1 Results spatial dependency tests

The OLS model specified in the previous chapter, and minmax spatial weight are used to run the statical tests for spatial dependency. The test results are reported in Table 19.

	OLS ($W = \min max$)	
Test statistic		**p–value**
Morans' I	116.5	0.00
LM-lag	1,391.8	0.00
LM-error	12,062	0.00
Robust LM-lag	393.61	0.00
Robust LM-error	11,063	0.00
SARMA	12,455	0.00

Table 19: spatial dependency tests results

The results in Table 20 show that the Moran's I statistic for the OLS regression is positive and statistically significant, indicating positive autocorrelation. Looking at the results of the LM-lag and LM-error tests in table 20, it can be found that both tests are statistically significant and that both hypotheses could not be rejected. This indicates the presence of spatial dependency in the error term and transaction prices. However, a problem with the LM-tests is that they both have power against each other alternative. So, the LM-error rejects the null hypothesis in the presence of a lag model, while the LM-lag rejects the null hypothesis in the presence of an error model. The Robust LM tests account for this by making an asymptotic adjustment to correct for this and should be used if both the LM-lag and LM-error are statistically significant and reject the null hypothesis (Anselin, 2017).

The results of the Robust LM-lag and Robust LM-error tests are also both statistically significant and reject the null hypothesis, again indicating spatial dependency in both the dependent variable – and the error term. The SARMA test is consulted to verify this statement as a final check. With the SARMA statistic, a joint hypothesis of $\rho = \lambda = 0$ to test for both y – and error dependence is tested (Anselin, 2017). Table 20 shows that the null hypothesis is rejected and the test is statistically significant, confirming the presence of spatial dependence in the error term and transaction data.

Concluding, all the tests are statistically significant. The result of the Moran's I test indicates the presence of positive spatial autocorrelation in the data, while the (Robust) LM and SARMA tests confirm spatial dependence in the transaction data and error term.

6.1.2 Results Likelihood ratio tests

From a theoretical point of view, it would be a logical choice to account for spatial dependency in the transaction data as real estate brokers often determine house prices by looking at the prices of neighbouring properties. However, even though the global models that account for endogenous interaction effects are state-of-the-art models, this thesis takes the criticism of the global models into account. As discussed in chapter three, the feedback effect affects the magnitude of the spatial spillover outcomes, making the individual lag estimates biased and unsuitable for prediction. Therefore, this thesis decided to use the local model and take the SDEM model as a starting point.

Likelihood ratio tests are used to compare the goodness of fit of two statistical models. In this test, two hierarchically nested models are compared to determine if adding more parameters (i.e. adding more complexity makes the model significantly more accurate. Hierarchically nested models simply mean that the complex model differs from the nested or simpler model by adding more parameters (Lesage & Pace, 2009). Following the spirit of Figure 7, the complex SDEM model first is tested against the simpler SEM and SLX models. Compared with the SLX model, the SDEM model adds $\lambda W\mu$ while in comparison with SEM, the SDEM model also contains the parameters θWX. This results in the following hypotheses:

$$\mathrm{H}_0 : \theta = 0$$

$$\mathrm{H}_0 : \lambda = 0$$

If the null hypotheses are rejected, it can be concluded that the complex model is significantly more accurate than the simpler models. The simpler models should be used if the tests fail to reject the null hypotheses. The test results are reported in Table 20, df shows the number of parameters that are restricted. So, in the Likelihood ratio θ test, all the lagged-x parameters are restricted, while in the Likelihood ratio λ test, only 1 parameter, the lagged error term is restricted. Both tests have a low p-value meaning the null hypothesis should be rejected and the SDEM model should not be restricted.

LR Test	Likelihood ratio	df	p-value
Likelihood ratio θ test (SDEM, SEM)	-257.65	62	<2.2 e-16
Likelihood ratio λ test (SDEM, SLX)	-1,250	1	<2.2e-16

Tabel 20: Likelihood ratio tests for spatial linear models (own edit, 2022)

Based on the Likelihood ratio tests, it can be concluded that there is no need to restrict the SDEM model to the simpler SLX or SEM model. However, the model choice should not only be decided from a statistical point of view. Aside from a solely data-driven approach, there can be conceptual reasons to choose a certain model. In our case, it is reasonable to add a lag error to account for the spatial dependency in the omitted explanatory variables. For example, this model does not contain any social-economic factors, while it is known that people tend to live close to like-minded people, with the same education level, for example.

Based on the statistic tests and the arguments why a local model is more appropriate for hedonic pricing analysis, the SDEM model seems the best choice. Therefore, we choose this model as the main workhorse. However, the SLX and SEM models will also be estimated to check if there are noteworthy differences between the estimates.

6.2 Model estimates (Model results)

The estimates of the OLS and spatial models are reported in Table 22 on the next page. Table 21 contains the model summaries.

Model summary	**OLS**	**SLX**	**SEM**	**SDEM**
Multiple r-square	0.888	0.897	0.895	0.904
Adjusted r-square	0.888	0.896		
F-statistic	2498 on 62 ***	1359 on 124 ***		
Degrees of freedom	19398	19336		
Wald-statistic			6142 ***	1630.4 ***
Number of observations	19461	19461	19461	19461
Significance levels	*0 (***), 0.001 (**), 0.01 (*), 0.05 (.)*			

Table 21: Summaries of the predicted models (own edit, 2022)

Looking at the R-square in table 22, it can be concluded that the explanatory power of the OLS model 0.888 is high. Moreover, the explanatory power increases slightly when the exogenous interaction effects of the neighbouring explanatory variables are added (0.897), the error term interaction effects (0.895) or both (0.905) are added. The increasing explanatory power proves that expanding the OLS with parameters that account for spatial interaction effects improves the model. At the same time, all three spatial models contain many statistically insignificant estimates, especially in the UGS variables. Remarkable is that most of the statistically insignificant estimates are coefficients by continuous variables based on distance. Except for a few lag-x variables in the SLX or SDEM model, all the structural and control variables are statically significant. In contrast, the locational and UGS variables show many statically insignificant results. A possible cause for the statically insignificant results could be the nature of these variables. Most of these variables offer the distance to the nearest (locational or UGS) amenity. Obviously, there is not much variance between the value of its own distance to the closet amenity and the average value of the neighbour’s distance to the amenity, as the distance is caused by their location, which is in the surrounding of each other. This might cause problems during the regression, resulting in statically insignificant variables. The operationalization process of this thesis might cause even less variance because individual dwelling observations share the same values in these distance variables when this is different in reality. This is explained in section 5.2.3. Therefore, the continuous proximity variables are might not suitable to include in a spatial econometric model. Additionally, adding the lag variant of this kind of variables seems not very useful for interpretation. As it is not expected that real estate buyers take the average distance to the nearest UGS amenity of their neighbours into account when buying a house.

In Table 23, it can be found that the spatial error variable, lambda, is large and statistically significant for both the SEM (0.854) and SDEM (0.747) models. This suggests that there are spatial effects among our residuals. Omitted variables that are common for the neighbours and influence the transaction price (e.g. crime rates, infrastructure quality, maintenance of public space etc.) could be the possible cause for this. The lower lambda score for the SDEM model compared to the SEM model also shows that part of these spatial effects can be attributed to the spatial dependency in the explanatory variables.

It is remarkable that, in general, the estimated values do not differ that much between the four models. The coefficients that do show differences will be discussed in the next section. However, due to the statically insignificant estimates in the spatial models, this thesis will mainly use the OLS estimates for interpretation. According to Mueller and Loomis (2008), the OLS estimates can still give reasonable estimates, even when spatial autocorrelation is present, which seems correct following the abovementioned findings.

Variables	**OLS**	Pr(>|t|)	**SLX**	Pr(>|t|)	**SEM**	Pr(>|t|)	**SDEM**	Pr(>|t|)
Intercept	3.815	<2e-16	3.579	< 2e-16	3.961	< 2.2e-16	3.504	< 2.2e-16
Structural characteristics								
Living area	0.590	<2e-16	0.551	< 2e-16	0.552	< 2.2e-16	0.552	< 2.2e-16
Living area (*l*)	-		*0.279*	< 2e-16			0.216	4.723e-06
Plot size	0.025	<2e-16	0.027	< 2e-16	0.027	< 2.2e-16	0.267	< 2.2e-16
Plot size (*l*)	-		-0.023	2.98e-14			-0.007	0.207
BY 1960 – 1980	-0.018	<2e-16	-0.017	2.98e-14	-0.017	8.882e-16	-0.018	< 2.2e-16
BY 1960 – 1980 (l)			-0.005	0.202			0.005	0.544
BY 1981 – 1990	0.008	2.54e-05	0.001	0.551	0.001	0.386	0.000	0.860
BY 1981 – 1990 (l)			-0.002	0.674			0.020	0.036
BY 1991 – 2000	0.044	<2e-16	0.036	< 2e-16	0.037	< 2.2e-16	0.036	< 2.2e-16
BY 1991 – 2000 (l)			-0.006	0.301			0.012	0.278
BY 2001 ≥	0.068	<2e-16	0.071	< 2e-16	0.071	< 2.2e-16	0.070	< 2.2e-16
BY 2001 ≥ (l)			-0.035	6.44e-08			-0.009	0.507
Rooms								
Stories								
Isolation measures								
Kitchen	0.034	< 2e-16	0.030	< 2e-16	0.030	< 2.2e-16	0.029	< 2.2e-16
Kitchen (l)			-0.015	0.234			-0.023	0.267
Dormer	0.009	2.57e-12	0.008	< 2e-16	0.008	5.563e-12	0.008	1.076e-12
Dormer (l)			0.014	0.551			0.019	0.078
Rooftop terrace	0.006	4.80e-06	0.009	6.54e-10	0.008	8.914e-10	0.008	4.107e-10
Rooftop terrace (l)			-0.009	0.082			0.011	0.277
Pantry	0.021	< 2e-16	0.019	2.40e-11	0.019	< 2.2e-16	0.019	< 2.2e-16
Pantry (l)			-0.004	0.551			-0.003	0.814
Balcony	0.022	< 2e-16	0.019	< 2e-16	0.019	< 2.2e-16	0.019	< 2.2e-16
Balcony (l)			-0.013	0.017			0.012	0.210
Heating system	0.077	< 2e-16	0.067	< 2e-16	0.066	< 2.2e-16	0.066	< 2.2e-16
Heating system (l)			0.119	2.45e-15			0.087	0.000
Parking space	0.025	< 2e-16	0.025	< 2e-16	0.025	< 2.2e-16	0.025	< 2.2e-16
Parking space (l)			-0.009	0.022			0.000	0.988
Garage	0.035	< 2e-16	0.033	< 2e-16	0.033	< 2.2e-16	0.033	< 2.2e-16
Garage (l)			-0.040	1.20e-10			-0.017	0.120
Garden South	0.004	1.06e-05	0.005	1.57e-07	0.005	4.925e-08	0.005	7.159e-08
Garden South (l)			-0.008	0.119			-0.001	0.952
Basement	0.012	0.0096	0.012	0.009	0.012	0.003	0.011	0.012
Basement (l)			0.010	0.702			-0.037	0.427
Monument	0.082	1.12e-09	0.062	1.73e-06	0.061	5.020e-07	0.061	1.084e-06
Monument (l)			0.254	0.001			0.043	0.732
Ground lease	-0.101	< 2e-16	-0.101	< 2e-16	-0.102	< 2.2e-16	-0.102	< 2.2e-16
Ground lease (l)			-0.009	0.551			0.031	0.280
Locational characteristics								
Supermarket	-0.005	0.0001	0.009	0.004	0.001	0.676	0.004	0.190
Supermarket (l)			-0.022	2.83e-08			-0.014	0.015
Department store	0.008	< 2e-16	-0.007	0.062	0.004	0.046	-0.003	0.393
Department store (l)			0.015	4.73e-05			0.011	0.008
Café	0.001	< 2e-16	-0.001	0.001	0.000	0.460	-0.001	0.003
Cafe (l)			0.002	8.50e-11			0.001	1.654e-05
Cinema	-0.012	< 2e-16	-0.176	2.64e-06	-0.018	< 2.2e-16	-0.019	4.723e-07
Cinema (l)			0.004	0.255			0.004	0.397
Primary schools	0.002	6.78e-07	-0.002	0.023	-0.001	0.219	-0.001	0.449
Primary schools (l)			0.004	0.000			0.001	0.743
Pharmacy	0.007	3.63e-14	0.008	0.017	0.005	0.052	0.008	0.011
Pharmacy (l)			-0.002	0.488			-0.003	0.478
Central train station	0.006	< 2e-16	-0.007	0.096	0.0025	0.218	-0.004	0.316
Central train station (l)			0.014	0.001			0.009	0.049
Highway	0.010	< 2e-16	-0.002	0.430	0.004	0.106	-0.002	0.542
Highway (l)			0.142	2.11e-05			0.013	0.002

UGS characteristics								
Allotments	0.002	0.009969	0.004	0.244	0.002	0.331	0.002	0.601
Allotments (l)			-0.003	0.356			0.001	0.783
Beach	0.005	< 2e-16	0.000	0.915	0.005	0.002	0.007	0.019
Beach (l)			0.005	0.144			--0.003	0.410
Cemetery	-0.004	< 2e-16	0.010	0.002	-0.0019	0.220	0.003	0.371
Cemetery (l)			-0.014	0.000			-0.006	0.114
Farmland	-0.003	6.44e-07	0.000	0.923	-0.003	0.207	-0.005	0.167
Farmland (l)			-0.006	0.098			-0.001	0.917
Orchard	0.004	1.41e-09	-0.003	0.403	0.002	0.386	0.000	0.930
Orchard (l)			0.009	0.006			0.003	0.412
Open green space	0.004	< 2e-16	-0.002	0.466	0.004	0.052	0.003	0.292
Open green space (l)			0.006	0.049			0.009	0.012
Meadow	0.003	1.05e-07	0.005	0.175	0.006	0.004	0.008	0.014
Meadow (l)			-0.004	0.206			-0.008	0.053
Scrub	-0.009	< 2e-16	-0.017	1.83e-07	-0.013	2.220e-16	-0.018	7.285e-08
Scrub (l)			0.009	0.008			0.008	0.035
Grass	-0.020	0.001	-0.002	0.791	-0.001	0.868	-0.003	0.635
Grass (l)			-0.065	0.000			-0.049	0.125
Trees	0.021	< 2e-16	0.021	8.19e-14	0.021	1.776e-15	0.021	3.109e-15
Trees (l)			0.008	0.269	-0.008		0.019	0.139
Water	0.090	< 2e-16	0.096	< 2e-16	0.095	< 2.2e-16	0.097	< 2.2e-16
Water (l)			-0.034	0.092			-0.050	0.198
Control								
Semi-detached	0.053	< 2e-16	0.049	< 2e-16	0.047	< 2.2e-16	0.047	< 2.2e-16
Semi-detached (l)			0.074	2.56e-09	-0.011		0.017	0.439
Corner	0.010	1.60e-14	0.010	< 2e-16	0.010	< 2.2e-16	0.010	< 2.2e-16
Corner (l)			-0.001	0.948	-0.002		-0.003	0.856
Duplex	0.059	< 2e-16	0.055	< 2e-16	0.054	< 2.2e-16	0.054	< 2.2e-16
Duplex (l)			0.052	1.84e-10	-0.014		0.030	0.024
Detached	0.123	< 2e-16	0.116	< 2e-16	0.116	< 2.2e-16	0.116	< 2.2e-16
Detached (l)			0.030	0.001	-0.069		0.002	0.881
2015	0.016	< 2e-16	0.016	< 2e-16	0.016	< 2.2e-16	0.016	< 2.2e-16
2015 (l)			-0.009	0.439	-0.011		-0.012	0.535
2016	0.046	< 2e-16	0.046	< 2e-16	0.046	< 2.2e-16	0.046	< 2.2e-16
2016 (l)			-0.008	0.471	-0.035		-0.010	0.596
2017	0.099	< 2e-16	0.099	< 2e-16	0.098	< 2.2e-16	0.099	< 2.2e-16
2017 (l)			0.018	0.093	-0.055		0.024	0.194
2018	0.147	< 2e-16	0.147	< 2e-16	0.147	< 2.2e-16	0.148	< 2.2e-16
2018 (l)			-0.025	0.029	-0.097		0.038	0.037
Rho ()								
Lambda ()					0.854	< 2.2e-16	0.747	< 2.2e-16
Significance levels 0.001 (***), 0.01 (**), 0.05 (*), 0.1 (.)								
** BY = building year l = lag*								

Table 22: Hedonic pricing regression results for all four models (own edit, 2022)

6.3 Interpretation

The models are specified as semi-log functions that affect how the coefficients should be interpreted. Coefficients from linear explanatory variables present the price change in percentage caused by one unit increase in the explanatory variable when the coefficient is multiplied by 100. For dummy variables count a similar interpretation as the coefficient shows the price change in percentage associated with the dummy going from 0 to 1. As earlier discussed, the coefficients from logged explanatory variables should be interpreted as elasticities, the coefficient indicates the percentage change in the explanatory variable associated with a 1 percent increase in the transaction price.

However, the main objective of this thesis is to identify the monetary value of UGSs. Therefore, a price change expressed in euros instead of percentages is desired. The coefficients can be used to calculate the marginal effect of a unit change in the independent variables on the price of an average dwelling in the data. To achieve this, an additional step is needed as the dependent variable was transformed to a logarithmic form and a re-transformation of the transaction price is required. $T = e^{ln(T)}$, shows the method to retransform the transaction price. From Table 18, it can be found that the average logged-transaction price of a dwelling is 12.21, applying the retransformation gives that the average apartment costs €200,336.

Table 22 contains the regression result of the four estimated models. The first chapter of this section gives a general interpretation of the results. This section will discuss all the estimated coefficients individually. The interpretation is mainly based on the OLS coefficients with the 0.001 level of statistical significance. It is mentioned when the significance level is smaller. Other models will be noted when considered relevant.

6.3.1 Structural variables

Regarding the structural variables, most of the coefficients have the expected sign. As in most hedonic price studies, the living area is the most influential variable with the highest impact on the transaction price. Looking at the estimate of the OLS model, it can be found that a 10% increase in the living area results in approximately a 5.9% increase in the transaction price. According to the spatial models, the impact of a 10% living area increase on the transaction price is slightly lower and between 5.1% and 5.2%. Following the OLS estimate, it is found that the transaction price increases by 0.25% if the plot size increases by 10%, which is not as high as expected. This might be caused because the garden variable is added to the model and probably overlaps with the plot size, capturing some of its value.

Besides the living area, a monumental status is the most influential factor for the transaction price. Suppose the dwelling has a monumental status, then the transaction price increases by 8.2%. Interesting is the lag estimate from the SLX model that indicates if neighbouring dwellings have a monumental status. According to this estimate, with the 0.001 level of statistical significance, the transaction price increases by 25.4% if the neighbouring dwellings have a monumental status. This could indicate that real estate buyers prefer to live near monumental buildings even more than living in one themselves. Obviously, living next to monumental buildings does bring aesthetic benefits for their surroundings, while owning one also has drawbacks like high maintenance costs, regulations e.g. The positive price effect found in this thesis is in line with the results from the study of Koster, Van Ommeren and Rietveld (2014). They found a positive price effect of 3.5% for dwellings that have a view on historical amenities.

The building year dummies are interpreted relative to dwellings built before 1960. The estimates align with the expectations stated in section 5.3 as only the post-war dwelling built between 1960 and 1980 are estimated to be sold for 1.8% less, while the other building year dummies show a positive price effect between 0.8% and 6.8%. More recently built dwellings have a higher transaction price, this is among other things caused by the fact that newer buildings have fewer maintenance costs and are more energy sufficient. Not surprisingly, the transaction price is highly affected by the presence of essential dwelling assets. The presence of a heating system increases the transaction price by 7.7%, while a kitchen results in a 3.5% price increase.

Real estate buyers are willing to pay for a parking space a premium equal to 2.5% of the transaction price and a premium of 3.5% for the garage. The coefficients by the lag variables in the SLX model suggest that the dwelling price decreases by 4% when the neighbouring houses have a garage, while the presence of neighbouring dwellings with parking spaces results in a reduction of approximately 1% from the transaction price. A possible explanation for this finding could be that garages are associated with noise nuisance, but this large impact is still remarkable. While parking spaces might use land that real estate buyers prefer to use for other activities e.g.

Surprisingly the transaction price only increases by 0.4% if the dwelling has a garden in the South. According to the OLS and spatial models, a ground lease transaction results in an approximately 10% decrease in the transaction price.

6.3.2 Locational variables

Regarding the locational variables, not all the estimated coefficients have the expected sign. The estimated coefficients for most locational variables (supermarket, department store, cinema, pharmacy, highway, central train station) show the price effect when the distance from the dwelling to the amenity increases by 1 kilometer. For each kilometer increase to the *supermarket*, the dwelling price decreases by 0.5%. This is in line with the intuition that real estate buyers value living close to a supermarket. On the contrary, proximity to *department stores* has a negative effect. If the proximity from the dwelling to the nearest department store decreases by 1 kilometer, the price decreases by 0.8%. The variable describing the proximity to the nearest *cinema* seems to have a higher positive impact than other locational variables as the coefficient implies a 1.2% increase in transaction price for each kilometer the dwelling is closer to the cinema. Surprisingly the proximity to the *pharmacy* coefficient is negative. Whenever the distance to the nearest pharmacy decreases by 1 kilometer, the price increases by 0.7%. Real estate buyers might not find it important to have a pharmacy within walking distance.

The estimated coefficients for the proximity to the central train station and highway are negative. Living a kilometer closer to the central train station leads to a 0.6% decrease in price, while a kilometer decrease in the proximity to the highway results in a 1% decrease. In section 5.3.2, it is already discussed that living in the proximity of a highway or train station also has disadvantages like noise nuisance and air pollution. The hedonic pricing study of Strand & Vågnes (2001) found that the view of a highway entrance or central train station has a negative price impact on residential real estate, while Seo, Golub and Kuby (2014) analysed the positive and negative relationship between house prices and the proximity to highways and light rail transits. They hypothesized that highway systems and light rail transits both bring positive (i.e. accessibility to transport) and negative (i.e. noise and air pollution) effects which are reflected in the market prices of nearby dwellings. These positive and negative effects were captured with multiple distance bands of 100 metres as dummy variables in a hedonic price function. Their study results show that the accessibility benefits adjacent to the station or highway exits are somewhat offset by the drawbacks associated with close proximity (see Figure 15). Regarding proximity to highway entrances, the disadvantages of close proximity disappear after 1.2 km, leaving the positive price effect associated with the accessibility to transport at a premium. This positive price effect decays after 6 kilometres. Concerning light rail stations, the disadvantages from close proximity decay after 0.8 km while the positive effects decay at 6.0 km (see Figure 15). This knowledge can help to explain this thesis's surprising negative price effects. Looking in Table 16 at the descriptive statistics of the variables from this thesis, it can be found that the mean value for the distance to the highway entrance is 2.05 with a standard deviation of 0.92. This means that a large part of the dwellings in the dataset is within the close proximity distance band of 1.2 km. However, our results might indicate that this distance band is even longer as the predicted estimate is negative. indicating that the drawbacks of proximity are stronger reflected in the market price than the positive effects. Regarding the central train station, it should not be compared with the results from the light rail station. However, the descriptive statistics in Table 16 also tell us that quite an amount of the dwellings in the dataset are located in close proximity to the central train station which could be the cause of the negative price effects.

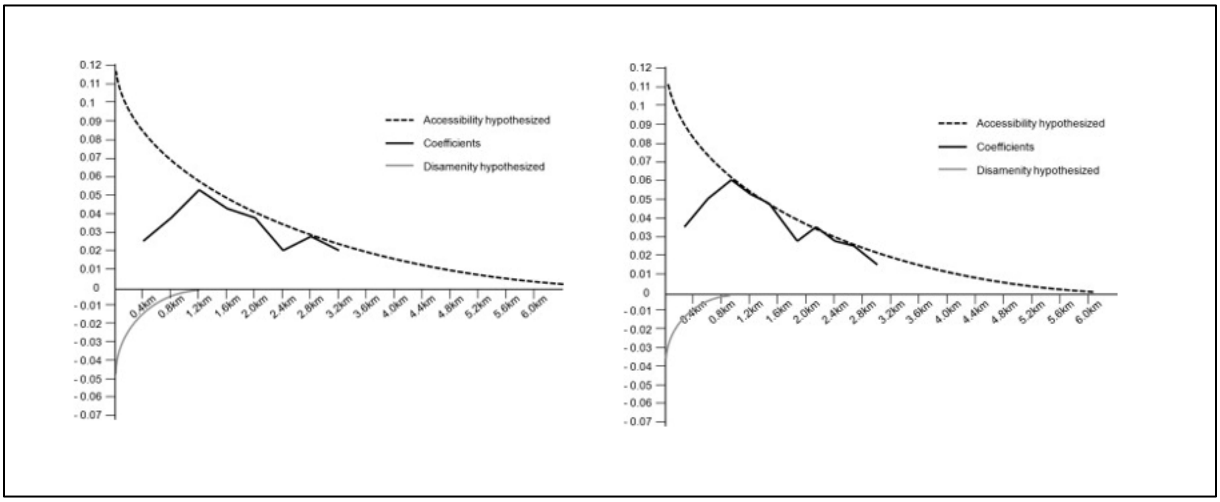


Figure 15: standardized coefficient of distance from highway entrances (left) and light rail stations (right) (Seo, Golub & Kuby, 2014)

The coefficients by the proximity to café and primary school variables need to be interpreted differently. The OLS coefficient indicates that the transaction price increases by 0.1% for each extra café located within 3 km of the dwelling, while the transaction price increases by 0.2% for each additional primary school located within 1 kilometer from the dwelling. These positive impacts are in line with the intuition.

6.3.3 Control variables

No surprising results are found regarding the estimates of the control variables. The transaction year dummies should be interpreted regarding the reference category for dwellings sold in 2014. As expected, an increasing price trend is found. Dwellings sold in 2015 are valued 1.6% higher, dwellings sold in 2016 raised the price by 4.6%, dwellings sold in 2017 by 9,9% and the transactions from 2018 are 14.7% higher. The graph in Figure 16 shows that dummy estimates follow the price trend of the entire Dutch housing market. Therefore, it seems that these dummies did serve their purpose of controlling for inflation and general price increases in the real estate market.

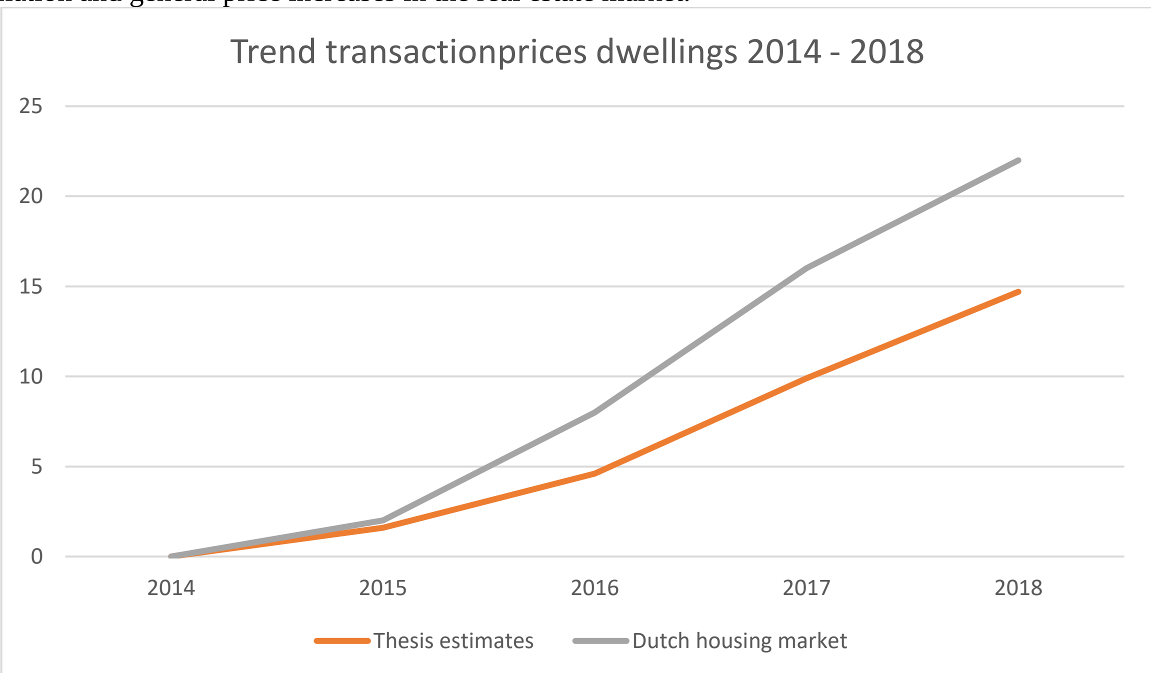


Figure 16: National dwelling price trend compared with transaction-year dummy trend line (CBS (2021), own edit)

As expected, all the dummies controlling for the dwelling type are positive and should be interpreted as price premium compared to an apartment, which is the reference category. Transaction of semi-detached houses are valued at 5.3% higher, Corner houses have a 1% higher price, duplex houses have a price premium of 5.9%, while detached houses have almost 12.3% transaction prices.

6.3.4 Urban green space variables

The UGS price effects are presented in Table 23. The average transaction price of approximately €200.000 was used to calculate the price effect expressed in euros.

Variable	Price effect in % per unit	Price effect in €
Allotments	-0.2% per km closer to nearest allotment	-€400 per km
Beach	-0.5% per km closer to nearest beach	-€1,000 per km
Cemetery	0.4% per km closer to nearest cemetery	€800 per km
Farmland	0.3% per km closer to nearest farmland	€600 per km
Orchard	-0.4% per km closer to nearest orchard	-€800 per km
Open green space	-0.4% per km closer to nearest OGS	-€800 per km
Meadow	-0.3% per km closer to nearest Meadow	-€600 per km
Scrub	0.9% per km closer to nearest Scrub	€1,800 per km
Grass	-2.0 % per 1% increase of grass	-€4,000 per %
Trees	2.1% per 1% increase of trees	€4,200 per %
Water	9% per 1% increase of water	€18,000 per %

Table 23: Estimated OLS effects of UGS variables on the average dwelling

<u>Allotments</u>

When the walking distance to the nearest allotment increases by one kilometer, the transaction price increases by 0.2%. This is in line with the findings from Czembrowski & Kronenberg (2016) that found a negative price effect if allotments are within a travelable distance from the dwelling. However, this thesis expected a positive price effect as the contextual causes, discussed in section 4.1.3.1, from the study case of Crzembrowski & Kronenberg (2016) do not seem to apply within the context of this thesis. A possible cause for the surprisingly negative price effect could be that allotments are often located in neighbourhoods with lower social-economic standards (e.g. no gardens, smaller homes) and therefore are surrounded by dwellings with lower transaction prices. However, it is also possible that real estate buyers refer to allotments with poorly maintained areas that attract crime or other negative externalities.

<u>Beaches</u>

This thesis expected that real estate buyers value living within walking distance to the beach. However, the coefficient estimates show a negative price impact for living in the proximity of the beach. For each kilometer decrease in distance to the beach, the transaction price decreases by 0.5%. The SEM model estimated the same negative price effect with a 0.001 level of statistical significance. A possible explanation for this unexpected outcome could be that not all the beaches in this research case do have recreational value. The beaches are defined as landforms along a coast or other water body consisting of sand, gravel, pebbles, cobblestones or shell fragments (Openstreetmap, 2022) and therefore might are not accessible for recreational purposes at all.

<u>Cemeteries</u>

The coefficient of the walking distance to the closest cemetery shows a negative price effect of 0.4%, indicating that the dwelling price increases whenever it is located a kilometer closer to the cemetery. This counters former study results (Tse and Love, 2000; Anderson & West, 2006; Czembrowski and Kronenberg, 2016) that found negative price effects from the proximity of cemeteries near residential real estate. As discussed in section 4.1.3, some of the cemeteries taken into the analysis are spacious locations with green scenic surroundings. The positive price effect that is found could indicate that people do not mind the negative sentiment associated with cemeteries and prefer to live close to them whenever they provide an attractive surrounding.

<u>Farmland</u>

Living in proximity to farmland was expected to have a negative impact on the real estate prices due to the noise and air pollution it might cause. Due to this expectation, the Euclidean distance was used to calculate the distances instead of the Network distance.

Nevertheless, the estimate related to farmland positively impacts the transaction price as the price increases by 0.3% for every kilometer that the dwelling is located closer to farmland.

Orchard
Following the estimated coefficient for the proximity to Orchard, a negative price effect of 0.4% is found for every kilometer the dwelling is located closer to the nearest Orchard. This suggests that real estate buyers do not associate orchards with areas that contain recreational value for them. As far as this thesis is concerned, no other studies tested this relationship and can be used to compare.

Open green spaces
Regarding open green spaces, a large positive price effect was expected as these areas were defined as; 'open green spaces used as a general recreation ground, indicating large recreational value (Openstreetmap, 2022)'. However, the coefficient shows that for each (kilo)meter decrease in the distance to OGS, the price decreases by 0.4%. A possible explanation for this unexpected result could be that public areas like OGS are open to everyone, what can also bring negative externalities like noise nuisance, criminality, feeling of unsafety, e.g..

Meadow
The meadow variable is estimated to have a negative price effect, as the transaction price decreases by 0.3% whenever the walking distance to the nearest meadow decreases by one kilometre. Suggesting that the positive price effect of meadow on the transaction prices of dwellings in rural areas found by Nisson & Pia (2008) does not apply in urban areas. Looking at the definition of areas identified as meadows, it can be found that cattle use the areas for grazing. A possible explanation for the negative price effect could be that cattle cause smell nuisance.

Scrub
Surprisingly, the largest price impact from the UGS proximity variables is caused by Scrub. Contrary to the expectations, real estate buyers tend to value living in the proximity of scrub areas, as the OLS coefficient shows an increase in transaction price of 0.9% for each kilometer decrease in distance to the nearest scrub. In the spatial models, all with the 0.001 level of statistical significance, the price effects are estimated even higher, between 1.3 and 1.8 percent. This is not the first study that found a positive relationship (Doss & Taff, 1996). Still, it was not expected that it would have the highest impact among the different UGS amenities. Scrubs are defined as uncultivated land covered with bush, scrubs and trees (Openstreetmap, 2022). A line of thought that might explain this unexpected result is that (urban) scrubs are often found in development areas. Development projects are often implemented with phased construction. Therefore, it could be that most of the newly built houses (that are often valued higher) are surrounded by scrub areas.

Grass
The coefficient for the density of land in a 6-digit zip area covered with grass shows a negative price effect for a 10 percentage point increase in land covered with grass the dwelling price decreases by 0.2%. This result was unexpected and is also opposed to the study results of Luttik (2000), that found a positive price effect of 4-5% whenever a green strip was located at a viewing distance from the dwelling. The results of this thesis suggest that grass and other vegetation lower than one meter do not contain a scenic value for the real estate owners. However, our findings align with study results from Li & Saphores (2011). They found that an increase in grassland coverage either on a multifamily building parcel or within a 200m boundary of the parcel did not lead to an increase in the transaction price. According to these results, they concluded that multifamily building owners in LA did not have a financial incentive to provide more grass areas on their property. The results of this thesis also suggest that real estate buyers are not willing to pay more for a house located in an area with high grass coverage.

Trees
Results show that a 10 percentage point increase in land covered by vegetation higher than one meter leads to an increase in the transaction price of 0.21%. This is in line with other studies that found a positive relationship between tree coverage and dwelling prices.

Li & Saphores (2011) found that an increase of land covered by trees within 200 metres from the dwelling leads to an increase in the property value, while an increase in the tree coverage at the property parcel leads to a decrease in price. The study from Sander, Polasky & Haight (2010) found that a 10% increase in tree cover within 100m from the dwelling increases the average home transaction price in Dakota county by approximately 0.5%. For the same increase of trees within 250m of the dwelling, a positive price effect of 0.3% was found. Increasing tree coverage beyond 250m did not contribute significantly to the transaction price. The hedonic pricing study from Plant, Rambaldi & Sipe (2017) found that property buyers in Brisbane, Australia, are willing to pay approximately 4% more for dwellings located in streets with at least 50% tree canopy coverage within 100 metres of the dwelling. Our study result show a comparable price effect as the study from Sander, Polasky & Haight (2010).

Water

The estimated coefficient for water shows that a 10 percentage point increase in land covered with water in a 6-digit zipcode leads to an increase in the transaction price of 0.9%. Our findings align with the results from Luttik (2000), that found a positive price effect for dwellings located next to the water of 6-12%, depending on the waterfront type. This is the opposite in the study results from Zhang & Dong (2018), they found a negative relationship between water land coverage and house transaction prices, indicating that Beijing's homeowners do not like to live close to water. However, they declared that this might be caused by the heavy pollution of water bodies in Beijing. Our results suggest that homeowners in Almere and Eindhoven prefer to live in areas with water. As this model took the water coverage of the dwellings surroundings into account, it is expected that the aesthetics of water ponds, rivers, e.g. contain scenic value for real estate buyers. However, it is conceivable that real estate buyers refer to water also with recreational activities, which underlies the price effect.

Regarding the UGS 'proximity' variables, it can be concluded that the results are unexpected. All the estimates are the opposite of the sign that was expected. Possible causes and explanations for the outcome of each variable have been discussed. In addition, some general remarks and suggestions should be made. Different amenities might have negative externalities for their direct neighbourhood while they only contain positive externalities for close dwellings outside their direct vicinity. This is known from other hedonic pricing variables based on distance, like the proximity to highways or light rail systems (Seo, Golub and Kuby, 2014). This is also conceivable for some UGS variables like allotments and might influence the predicted estimates.

Concluding, some unexpected price effects have been found, especially for the UGS 'proximity' variables. All the estimated coefficients for these variables were the opposite sign of what was expected. Possible causes and explanations for the outcomes of each individual variable have been discussed. In addition, some general remarks and suggestions should be made. First, different UGS amenities might have negative externalities for their direct neighbourhood while they only contain positive externalities for close dwellings just outside their immediate vicinity. This is known from other hedonic pricing variables operationalised on distance, like the proximity to highways or light rail systems (Seo, Golub and Kuby, 2014). It is conceivable for some of the UGS amenities that their proximity has both advantages and drawbacks and both are reflected in the estimates. Additionally, another possible explanation that could have caused the somewhat arbitrary results is that this thesis did not account for the distance decay effect during the operationalization process. The distance decay effect states that the interaction between two locations declines as the distance between them increases, once the distance is outside the activity space of two locations, their interaction shall decrease (Cheng & Bertolini, 2013). When the interaction between two spaces decreases, the price impact will also decrease. According to Daams et al. (2019), the price effect of (attractive) UGS on the transaction price decays with distance and is negligible after one kilometre. Table 9 shows that the distances to the UGS amenities used in the model of this thesis have maximums of approximately 3 to 8 kilometres. It could be an explanation for some of the surprising results that this thesis did not account for the fact the interaction between the dwelling and UGS amenity decreases whenever the distance increases.

6.4 Summary

First, the results of the Moran's I test and LM-tests were given, showing that spatial autocorrelation is present in both the transaction data as omitted variables. Next, the LR test shows that the local SDEM model should not be restricted. However, it was decided that SLX and SEM models were going to be estimated as well to find out if any differences in outcomes could be found. Unfortunately, most of the locational and UGS variables estimates were not significant. Before the individual variable estimates are discussed, the general model outcomes are reviewed. Guidance was given to interpreting the coefficients before all the individual coefficients of the OLS model were interpreted one by one. Results from the spatial models were discussed if appropriate. Most of the UGS coefficients had not the expected sign and possible causes have been discussed in detail.

CH 7 Discussion and conclusions

This chapter answers the research question stated in the introduction, discusses what can be learned from the research findings and provides recommendations for future research. In addition, the limitations of this research are clarified.

7.1 Conclusions

The objective of this thesis was to measure the effect of urban green spaces on the transaction price of dwellings in the Netherlands. With a final cleaned dataset of 19,461 property transactions in Almere and Eindhoven over the period of 2014 to 2018, a Hedonic regression and three different spatial models that account for spatial dependency were estimated to study the price effects of the UGS variables. The NVM provided transaction data including, the dwelling structural characteristics, while locational characteristics were derived from CBS and the UGS data was supplied by HUGSI and Openstreetmap. From the 108 available variables, 20 structural, 8 locational, 11 UGS and 8 control variables entered the final model.

The first research question in this thesis investigated the operationalization process of urban green spaces in a hedonic pricing model. During the literature research, it was found that urban green spaces impact the dwelling price, due to the scenic and recreational value they contain. This recreational value is taken into the regression model by creating variables according to the proximity principle (Crompton, 2001), while the scenic value is captured with variables that contain UGS density in a certain area. Land use data, generally based on governmental land class systems, are often used as a UGS data source. According to Xu et al. (2016), the limitation of land use data is that it does not provide an accurate reflection of the real amount of UGS as it neglects the urban green in areas that are appointed as other land-use types.

This thesis tackled this problem by using ‘land coverage’ data provided by HUGSI, to account for the density of grass, trees and water. Their data allowed to distinguish, water, vegetation higher than 1 metre (trees) and vegetation lower than 1 metre (grass) from other land use types for a 10 by 10 grid, providing very accurate UGS data. Our results show that people appreciate living in areas covered by vegetation higher than one meter and water as the increase of these land coverages results in higher transaction prices. On the contrary, the dwelling price decreases when the grass coverage in the surrounding area increases. Indicating that people do not derive scenic value from vegetation lower than one metre while the scenic value from water and vegetation higher than one metre is appreciated.

Another limitation mentioned in the literature research is that many UGS valuation studies do not account for the fact that UGS is a heterogeneous good (Panduro & Veie, 2013). This thesis followed the suggestion of Liebelt, Bartke & Schwarz (2018) by classifying different types of UGS to account for their heterogeneity and test if they have different impacts on the transaction price. For each dwelling, the distance to the nearest UGS amenity was calculated. Euclidean distance is used for the farmland, while the Network distance was used for the other 8 UGS types.

The second research question was focused on the most suitable method to model spatial dependencies in UGS valuation research. Three different interaction effects can be added to the hedonic equation to account for spatial dependence in the dependent variable, explanatory variable or omitted variables. It is impossible to account for all three interaction effects in one model as it leads to overfitting (Elhorst, 2017). However, six ‘spatial’ models that account for one or two interaction effects can be made where the models that account for interaction effects in the dependent variable are called global models. The Moran’s I test appointed the presence of spatial autocorrelation in the dataset and the LM test identified that spatial autocorrelation was present in both the transaction data as the omitted variables. Although modelling endogenous interaction effects, which refer to the spatial dependencies in the dependent variable, is a state-of-the-art method, it was chosen to not use these global models due to their feedback effects. Instead, it was chosen to estimate all local models (SLX, SER and SDEM), although the likelihood ratio tests showed that the SDEM model should not be restricted. Still, it was chosen to estimate all three local models so that their outcomes could be compared.

Unfortunately, most of the UGS variable estimates were statistically insignificant for all three models. Presumed is that there is too little variance within these variables as most of them share (almost) the same values for these variables. Results also indicate that continuously operationalised proximity variables seem unsuitable for spatial econometric models, also because interpreting the lagged variant of these variables does not seem to have practical utility. Even though the spatial model contains statistically insignificant results, it was remarkable that many statistically significant variable estimates shared (almost) the same value as the OLS results. This indicates that the assumption of Mueller and Loomis (2008) that the OLS model can still give reasonable estimates even when spatial autocorrelation is present is correct.

The third research question investigated what specification of the weight matrix is the most suitable for UGS valuation research. Different options such as contiguity, distance, k-nearest neighbours and inverse distance weights were explained in the methodology chapter. Weights based on contiguity seem more appropriate if counties, states or land with clear spatial boundaries are the unit of analysis and therefore were abandoned as an option for this thesis. The inverse distance weight matrix looks like an interesting option as it embodies Tobler's first law of geography (1970). Unfortunately, it still has too many (statistical) drawbacks and therefore is not used for empirical research (Suryowati, Beti & Faradila, 2018). As the geographical nature of our dataset made it impossible to use the k-nearest neighbour specification, this thesis used the max-min distance approach to specify the spatial weight matrix. This resulted that all dwellings within a radius of 330 metres from a dwelling being accounted as neighbours. In our case, this seemed to be the correct specification. However, the k-nearest neighbour or other distances could also be interesting options; this is further discussed in the research limitation section 7.3.

Finally, the last and main research question addressed in this thesis investigated the impact of UGS on the transaction price of residential dwellings in the Netherlands. The variables accounting for the proximity of UGS amenities show different results. The dwelling price decreases between 0.2 – 0.5% for each km that the dwelling is located closer to the nearest allotment, beach, orchard, open green space or meadow. Indicating that these UGS amenities do not contain a recreational value for real estate buyers. For a location near a cemetery, farmland or scrub, real estate buyers are willing to pay a price premium between 0.3-0.9 % for each km that the dwelling is closer to the amenity. They indicate that real estate buyers like to be in proximity of these UGS amenities and they contain a recreational value for them. Land covered with water and vegetation higher than one metre in the dwelling six-digit zip code area was found to have scenic value for real estate buyers. A price premium of 0.9% for water and 0.2% for vegetation higher than one meter was found for a 10 percentage point increase in land coverage. Additionally, a decrease of 0.21% in transaction price was found for each a 10 percentage point increase in land covered with vegetation lower than one meter. The results indicate that the proximity of different UGS types has various impacts on house prices. In contrast, diverse urban green vegetation surrounding the dwelling shows other effects on the house price.

Although additional research is necessary to gain further insights into the impact of urban green spaces on dwelling prices, this study gave new insights into the effects of urban green spaces in the Netherlands. This was done by investigating the housing markets of other Dutch cities and operationalizing UGS differently than previous studies in the Netherlands that found price premiums for the view on UGS amenities (Luttik, 2000), evaluated the price effect of the proximity of urban green spaces in Amsterdam (Dekkers & Koomen, 2008; Daams et al., 2016; Daams et al., 2019) or only accounted for UGS that are perceived as attractive (Daams et al., 2016; Daams et al., 2019). Moreover, to the best of this author's knowledge, this is one of the first studies that distinguished this many types of UGSs in a hedonic price function and used land coverage data to account for the density of water and green vegetation.

Regarding the spatial weight matrix specification and choice of the right model, this thesis agrees with Elhorst (2017) that this is a complex problem in empirical research. Krekel & Kolbe (2020) stated that practical guidance is needed about which method or combination of methods is required under which circumstances.

Unfortunately, this thesis encountered different problems, which make its contribution to this discussion small. However, a notable result is the insignificant estimates of the locational and UGS variables based on a continuous distance measure in the spatial models. This is probably due to the low variance within the lagged variants of these variables. It might suggest these variables are not suitable to be used in a spatial model. This thesis also questions the practical utility of interpreting these lag variants. Additionally, this thesis found no major differences between the statistically significant estimates of the OLS model and the spatial models. This indicates that Mueller and Loomis (2008) are correct with their observation that the OLS outcomes can still give reasonable estimates even when spatial autocorrelation is present.

7.2 Discussion

This section encompasses a short discussion about some aspects of the research process and the interpretation of the results. It briefly discusses what policy and theory makers can learn from the results, followed by academic recommendations on what future studies can learn from the research process and study results.

7.2.1 Recommendations for urban planners and policy makers

Our study results show that real estate buyers in Almere and Eindhoven are willing to pay extra for a house located in an area with vegetation higher than one meter. Vegetation lower than one meter seems undesired as it does not add value to the dwelling transaction prices, indicating that it does not provide scenic value for real estate buyers. Still, these findings align with the trend of more nature-inclusive construction. The knowledge that real estate buyers are willing to pay more for dwellings located in areas with more vegetation higher than one meter could be a financial incentive for developers to account for more trees in their development plans.

Moreover, houses located in areas with water are even more desired by real estate buyers. Our findings confirm the findings from Luttik (2000), that real estate buyers tend to live close to waterfronts due to their aesthetics and are willing to pay for it. However, the study results from Zhang & Dong (2018) in Beijing show that it can also be reversed. Whenever the water is heavily polluted, it can also result in negative price effects indicating the real estate buyers do not want to live close to them. For policymakers, developers and urban planners, it is valuable information to know how the presence of water corresponds to the real estate prices. Our results and the results of Luttik (2000) show that locations near water are desired in the Dutch housing market. However, foreign study results show that water quality also plays an important role here (Zhang & Dong, 2018). Therefore, policymakers should be aware that the water quality influences the attractiveness of an area as a place to live.

Studies from other academic disciplines already showed that UGS has diverse external positive (health) effects on its surroundings. Therefore, it is a logical consequence that building more UGS is included in the sustainable development goals (United Nations, 2015). However, the results of this thesis show that developing new UGS should be done with thoughtful thinking. According to our results, we could say that UGS certainly has a substantial impact on house prices, so with the misplacement or unequal distribution of UGS, there are certainly unwanted effects like eco-gentrification lurking. Therefore, this thesis would advise policymakers and urban planners to equally spread urban green vegetation within the city and ensure that UGS amenities are accessible to everyone.

7.2.2 Academic recommendations

This thesis made use of land cover data that HUGSI obtained by machine learning techniques on satellite images to account for urban green vegetation like grass, trees and water. This data type is promising for urban research and may have many other applications for future research. Conceivably, it could be used to distinguish more types of urban green vegetation. But it would also be interesting to be used for data collecting on solar panels and green roofs.

This thesis calculated both the Network distance and Euclidean distance between the dwellings and the UGS amenities.

Our findings indicate that the differences in outcomes are minor, which could also be caused by some flaws in the methodology of this study that will be discussed in the next section. However, this thesis would advise that future research only uses the network distance if a few conditions are met. First, it should theoretically make sense to use the Network distance instead of the Euclidean distance. Next, it is expected that the distance outcomes between both measures differ significantly. Finally, proper network data including walking paths should be available. Otherwise, this thesis thinks it is not worth the time and effort as calculating the Network distance requires more specific software knowledge.

7.3 Research limitations

Even though this thesis was carried out carefully and provided several interesting results, some limitations can be identified and should be discussed. This section will evaluate the methodology, discuss the areas of improvement and point out decisions that could have led to research bias in the results.

First of all, this thesis did not properly test what specification of the spatial weight matrix fits this dataset the best. Mainly because it took too much time to compute different spatial weight matrices and run them in different models. This is the downfall of using big datasets, which require more computing power and cause software programs like R to become slow. Additionally, the spatial resolution of the transaction data was unfortunately not perfect. This influenced the specification process of the spatial weight matrix as it prevented the k-nearest neighbour specification could be used for this research, while from a theoretical point of view, this would be the right method to specify the spatial relation between the dwellings in this thesis. Concluding, the quality of this thesis could have been improved if the performances of different spatial weight matrices had been tested.

Secondly, it can be concluded that there probably is a discrepancy between the calculated distances to the UGS amenities and the actual distances in reality. Two causes can be appointed for this. First of all, the exact geo-coordinates of each dwelling were unknown and therefore, each dwelling was assigned with a random x and y coordinate from its 6-digit zip code area that HUGSI provided. Secondly, the used road network from NWB missed most of the footpaths that might be used to visit the UGS amenities. Therefore, the Network distance might be calculated over a road that is not or poorly accessible for pedestrians. Finally, some of the unique UGS amenity points shared the same network node in TransCAD and therefore the same distances. Still, quite accurate measurement was done, but it should be mentioned that the reasons stated above might cause measurement bias. It will be an improvement if future research ensures that all single observations have a unique geographical indicator and suitable road network data is available.

Thirdly, there are likely several omitted variables in the estimated HP function. For example, the model might have performed better if socio-economic factors like crime rates and education level had been added to the model. These kinds of neighbourhood characteristics are not present in the model, while they probably influence the transaction price and could also play a role in the perception of UGS.

Fourthly and last, considering land covered by water, the research case Almere might be outstanding from the average Dutch city due to a large number of canals and lakes within the city and the fact that is it located between three major freshwater lakes. This must be considered if the outcomes of this thesis are generalized to other Dutch cities.

7.4 Recommendations for future research

Building up on some of the abovementioned limitations as well as the finding mentioned in the discussion and conclusion, the following directions for further research are suggested.

This research measured to what extent real estate buyers are willing to pay a transaction price premium for proximity to UGS amenities and UGS land coverage. This is different from the willingness of tenants to pay a rental premium for UGS, while they also present a large group in the Dutch real estate market, as approximately 40% of the dwellings in the Netherlands are rental houses (Rozing, 2020). Academically speaking, it is interesting to see if the preferences of tenants reflected in the housing rents are comparable to those of the real estate buyers. This would also provide interesting information for developers, investors and social housing corporations about the UGS preferences of tenants.

Most UGS studies in the Netherlands focus only on Amsterdam and the Randstad region, while smaller cities in more rural areas are neglected. The regional economic market differs between cities and regions and affects the housing demand and willingness to pay between locations. This thesis already shifted to another part of the Netherlands by taking Eindhoven into account. However, Eindhoven, part of the city network Brabant, is the second-largest urbanized area in the Netherlands (Provincie Brabant, 2022). It would be interesting if future research would investigate if the UGS price effects differ in less urbanized areas like Groningen or Maastricht, for example.

Instead of the entire housing market, further research could focus on in-depth case studies with only one single housing market segment in a certain city. It is conceivable to think that people active in housing markets that have often no or less access to a garden (e.g. apartments) are willing to pay more for access to public UGS. In the Netherlands, the studies of Daams et al. (2016; 2019) already focused on single-family houses.

This thesis followed the recommendation of Liebelt, Bartke & Schwarz (2018) to account for UGS heterogeneity and analyze different types of UGS instead of aggregating them into one category. Our results show that the price effects differ for each type. While Xu et al. (2016) proved that the spatial characteristics of UGS influence their price impact, Daams et al. (2016; 2019) found that UGS that are perceived as attractive have higher price effects and the finding of Czembrowski et al. (2019) showed that the impact of UGS is also influenced by their use-value, social meaning and multifunctionality. Still, there are other options for how the heterogeneity of UGS could be interwoven in the hedonic pricing model. Therefore, this thesis advices future research to explore other aspects of UGS like UGS arrangement, structural UGS details or UGS health as advised by Krekel & Kolbe (2020). The land coverage data of HUGSI also provides a health score based on chlorophyll activity, which could be an interesting source for future research to explore.

Interesting would be to take the land coverage of trees, grass and water into account on another spatial level than the 6-digit zip code. Using a band width of a certain distance in metres enables an equally large area to be taken into account for each dwelling. In addition, this can help to find to which distance real estate buyers in the Netherlands take the coverage into account, which could be helpful for future academic studies. The study from Sander, Polasky & Haight (2010) did this in LA and found that the price effects disappear after 250 metres. This would also be interesting information for both developers and policymakers.

References

A. Samad, N. S., Abdul-Rahim, A. S., Mohd Yusof, M. J., & Tanaka, K. (2020). Assessing the economic value of urban green spaces in Kuala Lumpur. Environmental Science and Pollution Research, 27(10), 10367–10390. https://doi.org/10.1007/s11356-019-07593-7

Ambrey, C., & Fleming, C. (2013). Public Greenspace and Life Satisfaction in Urban Australia. Urban Studies, 51(6), 1290–1321. https://doi.org/10.1177/0042098013494417

Anderson, S. T., & West, S. E. (2006). Open space, residential property values, and spatial context. Regional Science and Urban Economics, 36(6), 773–789. https://doi.org/10.1016/j.regsciurbeco.2006.03.007

Anselin, L., Bera, A. K., Florax, R., & Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. Regional Science and Urban Economics, 26(1), 77–104. https://doi.org/10.1016/0166-0462(95)02111-6

Anselin, L. (2010). Local Indicators of Spatial Association-LISA. Geographical Analysis, 27(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Anselin, L., Rey, S. J., & Li, W. (2014). Metadata and provenance for spatial analysis: the case of spatial weights. International Journal of Geographical Information Science, 28(11), 2261–2280. https://doi.org/10.1080/13658816.2014.917313

Anselin, L. (2017). Spatial Regression: 10. Specification Tests [Slides]. Spatial.Uchicago.Edu. https://spatial.uchicago.edu/sites/spatial.uchicago.edu/files/10_specification_tests_2_slides.pdf

Anselin, L. (2020). Spatial weights (Introduction to spatial data science) [Slides]. University of Chicago.

Aram, F., Higueras García, E., Solgi, E., & Mansournia, S. (2019). Urban green space cooling effect in cities. Heliyon, 5(4), e01339. https://doi.org/10.1016/j.heliyon.2019.e01339

Astell-Burt, T., Feng, X., & Kolt, G. S. (2013). Mental health benefits of neighbourhood green space are stronger among physically active adults in middle-to-older age: Evidence from 260,061 Australians. Preventive Medicine, 57(5), 601–606. https://doi.org/10.1016/j.ypmed.2013.08.017

Bellefon De, M. P., & Floch, J. M. (2018a). Geographically Weighted Regression. Handbook of Spatial Analysis, 232–252. https://www.insee.fr/en/information/3635545

Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Blanchet, G., Carvalho, M., Christensen, B., Chun, Y., Dormann, C., Dray, S., Gómez-Rubio, V., Krainski, E., Legendre, P., Lewin-Koh, N., Li, A., Millo, G., Mueller, W., Ono, H., . . . Yu, D. (2022). Package 'spdep: Spatial Dependence: Weighting Schemes, Statistics'. CRAM. https://cran.rproject.org/web/packages/spdep/index.html

Bivand, R., Keitt, T., Rownlingson, B., Pebesma, E., Summer, M., Hijmans, R., Baston, D., Rouault, E., Warmerdam, F., Ooms, J., & Rundel, C. (2021). Package 'rgdal'. CRAN. https://r-forge.r-project.org/projects/rgdal/

Bivand, R., Piras, G., Anselin, L., Bernat, A., Blankmeyer, E., Chun, Y., Gómez-Rubio, V., Griffith, D., Gubri, M., Halbersma, R., LeSage, J., Li, A., Li, H., Ma, J., Mallik, A., Millo, G., Pace, K., Peres-Neto, P., Rüttenauer, T., Tiefelsdorf, M. (2021). Package 'spatialreg'. CRAN. https://r-spatial.github.io/spatialreg/

Björklund, K., Dadzie, J.A. & Wilhelmsson, M. (01-08-2006). *Offer price, transaction price and time-on-market.* Retrieved from https://www.emerald.com/insight/content/doi/10.1108/02637470610671631/full/html?casa_token=jg6

2dm4Q6SUAAAAA:3BbnVQWKocIt1ACAYp09TsHK1GdmfpybwMm14AWAEZ6FuTnBoNgZVfP_LHV9X2MFSJy0RnOohn4zZWpugMxo5A3EN1jqlyLqNeRDcG2rfnX0_-MOW5c

Bouayad Agha, S., & Bellefon De, M. P. (2018). Spatial autocorrelation indices. Handbook for Spatial Analysis, 52–68.

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically Weighted Regression. Journal of the Royal Statistical Society: Series D (The Statistician), 47(3), 431–443. https://doi.org/10.1111/1467-9884.00145

CBS. (2021). Regionale kerncijfers Nederland. https://opendata.cbs.nl/#/CBS/nl/dataset/70072ned/table?searchKeywords=Eindhoven

CBS. (2022, April 22). Bestaande koopwoningen: verkoopprijzen prijsindex 2015 = 100. https://opendata.cbs.nl/#/CBS/nl/dataset/83906NED/table

Chen, V. Y. J., Deng, W. S., Yang, T. C., & Matthews, S. A. (2012). Geographically Weighted Quantile Regression (GWQR): An Application to U.S. Mortality Data. Geographical Analysis, 44(2), 134–150. https://doi.org/10.1111/j.1538-4632.2012.00841.x

Cheng, J., & Bertolini, L. (2013). Measuring urban job accessibility with distance decay, competition and diversity. Journal of Transport Geography, 30, 100–109. https://doi.org/10.1016/j.jtrangeo.2013.03.005

Conway, D., Li, C. Q., Wolch, J., Kahle, C., & Jerrett, M. (2008). A Spatial Autocorrelation Approach for Examining the Effects of Urban Greenspace on Residential Property Values. The Journal of Real Estate Finance and Economics, 41(2), 150–169. https://doi.org/10.1007/s11146-008-9159-6

Cook, R. D. (1979). Influential Observations in Linear Regression. Journal of the American Statistical Association, 74(365), 169–174. https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481634

Cornelis, J., & Hermy, M. (2004). Biodiversity relationships in urban and suburban parks in Flanders. Landscape and Urban Planning, 69(4), 385–401. https://doi.org/10.1016/j.landurbplan.2003.10.038

Cox, K. (2017). Het effect van duurzaamheid op de huurprijs en de vertaling naar de waarderingen van kantoren in Nederland. Universiteit van Utrecht. https://dspace.library.uu.nl/bitstream/handle/1874/354122/DEFINITIEF2_Masterhesis_KarlonCox_juli2017.pdf?sequence=2

Crompton, J. L. (2001). The Impact of Parks on Property Values: A Review of the Empirical Evidence. Journal of Leisure Research, 33(1), 1–31. https://doi.org/10.1080/00222216.2001.11949928

Czembrowski, P., & Kronenberg, J. (2016). Hedonic pricing and different urban green space types and sizes: Insights into the discussion on valuing ecosystem services. Landscape and Urban Planning, 146, 11–19. https://doi.org/10.1016/j.landurbplan.2015.10.005

Czembrowski, P., ŁAszkiewicz, E., Kronenberg, J., Engström, G., & Andersson, E. (2019). Valuing individual characteristics and the multifunctionality of urban green spaces: The integration of sociotope mapping and hedonic pricing. PLOS ONE, 14(3), e0212277. https://doi.org/10.1371/journal.pone.0212277

Daams, M. N., Sijtsma, F. J., & van der Vlist, A. J. (2016). The Effect of Natural Space on Nearby Property Prices: Accounting for Perceived Attractiveness. Land Economics, 92(3), 389–410. https://doi.org/10.3368/le.92.3.389

Daams, M. N., Sijtsma, F. J., & Veneri, P. (2019). Mixed monetary and non-monetary valuation of attractive urban green space: A case study using Amsterdam house prices. Ecological Economics, 166, 106430. https://doi.org/10.1016/j.ecolecon.2019.106430

de Oude, S. (2020). The association between proximity to urban green and mental health in Utrecht, the Netherlands. The association between proximity to urban green and mental health in Utrecht, the Netherlands, 19–20. https://dspace.library.uu.nl/handle/1874/399930

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.

Depietri, Y., & McPhearson, T. (2017). Integrating the Grey, Green, and Blue in Cities: Nature-Based Solutions for Climate Change Adaptation and Risk Reduction. Theory and Practice of Urban Sustainability Transitions, 91–109. https://doi.org/10.1007/978-3-319-56091-5_6

Dombrow, J., Rodriquez, M., & Simans, C. F. (2000). The Market Value of Mature Trees In Single-Familiy Housing Markets. Appraisal Journal, 1–5. http://www.actrees.org/files/Research/mature_trees_single_family_housing

Doss, C. R., & Taff, S. J. (1996). The Influence of Wetland Type and Wetland Proximity on Residential Property Values. Journal of Agricultural and Resource Economics, 21(1), 120–129. https://www.jstor.org/stable/40986902?seq=1#metadata_info_tab_contents

Du, M., & Zhang, X. (2020). Urban greening: A new paradox of economic or social sustainability? Land Use Policy, 92, 104487. https://doi.org/10.1016/j.landusepol.2020.104487

Elhorst, J. P. (2010). Applied spatial econometrics: raising the bar. Spatial Economic Analysis, 9–28.

Elhorst, J. P., & Halleck Vega, S. M. (2017). The SLX model: Extensions and the sensitivity of spatial spillovers to W. Papeles de Economia Espanola, 34–50.

Elhorst, J. P. (2018). Spatial Econometrics: The basics. [Slides]. University of Groningen. https://www.diw.de/sixcms/media.php/17/gc_ws1819_Elhorst_presentation.pdf

Engström, G., & Gren, A. (2017). Capturing the value of green space in urban parks in a sustainable urban planning and design context: pros and cons of hedonic pricing. Ecology and Society, 1–13. https://www.jstor.org/stable/26270097?seq=1#metadata_info_tab_contents

Eom, Y. S., Kim, A. S., Kim, S. G., & Kim, J. O. (2019). A Hedonic Valuation of Urban Green Space in Seoul. Environmental and Resource Economics Review, 61–63.

Fan, P., Xu, L., Yue, W., & Chen, J. (2016). Accessibility of public urban green space in an urban periphery: The case of Shanghai. Landscape and Urban Planning, 165, 177–192. https://doi.org/10.1016/j.landurbplan.2016.11.007

Field, A. (2009). Discovering Statistics using IBM SPSS statistics (Third edition). SAGE Publications.

Fleming, D., Grimes, A., Lebreton, L., Maré, D., & Nunns, P. (2018). Valuing sunshine. Regional Science and Urban Economics, 68, 268–276. https://doi.org/10.1016/j.regsciurbeco.2017.11.008

Floch, J. M., & Saout Le, R. (2018). Spatial econometrics - common models. Handbook of Spatial Analysis, 149–177.

Florax, R. J. G. M., & van der Vlist, A. J. (2003). Spatial Econometric Data Analysis: Moving Beyond Traditional Models. International Regional Science Review, 26(3), 223–243. https://doi.org/10.1177/0160017603254937

Ford, C. (2018). Interpreting Log Transformations in a Linear Model. University of Virginia Library. https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/

Fruth, E., Kvistad, M., Marshall, J., Pfeifer, L., Rau, L., Sagebiel, J., Soto, D., Tarpey, J., Weir, J., & Winiarski, B. (2019). Economic valuation of street-level urban greening: A case study from an evolving mixed-use area in Berlin. Land Use Policy, 89, 104237. https://doi.org/10.1016/j.landusepol.2019.104237

Fuerst, F., van de Wetering, J., & Wyatt, P. (2013). Is intrinsic energy efficiency reflected in the pricing of office leases? Building Research & Information, 41(4), 373–383. https://doi.org/10.1080/09613218.2013.780229

Goddard, M. A., Dougill, A. J., & Benton, T. G. (2010). Scaling up from gardens: biodiversity conservation in urban environments. Trends in Ecology & Evolution, 25(2), 90–98. https://doi.org/10.1016/j.tree.2009.07.016

Halleck Vega, S., & Elhorst, P. (2013). On spatial econometric models, spillover effects, and W. Faculty of Economics and Business, University of Groningen. http://www-sre.wu.ac.at/ersa/ersaconfs/ersa13/ERSA2013_paper_00222.pdf

Halleck Vega, S., & Elhorst, J. P. (2015). THE SLX MODEL. Journal of Regional Science, 55(3), 339–363. https://doi.org/10.1111/jors.12188

Henniger, S. (2011). Urban Green Areas: Lots of Benefits, but some Drawbacks. REAL CORP, 1–9. https://www.corp.at/archive/CORP2011_122.pdf

Herath, S., Choumert, J., & Maier, G. (2015). The value of the greenbelt in Vienna: a spatial hedonic analysis. The Annals of Regional Science, 54(2), 349–374. https://doi.org/10.1007/s00168-015-0657-1

Hijmans, R. J., & Ghosh, A. (2021). Spatial Data Analysis with R. Geographic Information Analysis.

Hilgers, B. A. J. (2018). Automated valuation models for commercial real estate in the Netherlands traditional regression versus machine learning techniques. Eindhoven University of Technology. https://research.tue.nl/en/studentTheses/automated-valuation-models-for-commercial-real-estate-in-the-neth

Holmes, T. P., Murphy, E. A., & Bell, K. P. (2006). Exotic Forest Insects and Residential Property Values. Agricultural and Resource Economics Review, 35(1), 155–166. https://doi.org/10.1017/s1068280500010121

HUGSI. (2021). Ranking. HUGSI. https://www.hugsi.green/ranking/

Jo Black, K., & Richards, M. (2020). Eco-gentrification and who benefits from urban green amenities: NYC's high Line. Landscape and Urban Planning, 204, 103900. https://doi.org/10.1016/j.landurbplan.2020.103900

Kolbe, J., & Wüstemann, H. (2014). Estimating the value of Urban Green Space: A hedonic pricing analysis of the housing market in Cologne, Germany. Humboldt University Berlin, 2015(002), 1–21. https://www.econstor.eu/bitstream/10419/107911/1/815374305.pdf

Koster, H. R. A., Van Ommeren, J. N., & Rietveld, P. (2014). Historic amenities, income and sorting of households. Journal of Economic Geography, 16(1), 203–236. https://doi.org/10.1093/jeg/lbu047

Krekel, C., & Kolbe, J. (2020). New ways of valuing ecosystem services: big data, machine learning and the value of urban green spaces. A Research Agenda for Environmental Economics., 159–183.

Kühn, I., Brandl, R., & Klotz, S. (2004). The flora of German cities is naturally species rich. Evol Eco research, 749–764. http://www.evolutionary-ecology.com/abstracts/v06/1629.html

Lancaster, K. J. (1966). A New Approach to Consumer Theory. Journal of Political Economy, 74(2), 132–157. https://doi.org/10.1086/259131

Lennon, M., Douglas, O., & Scott, M. (2017). Urban green space for health and well-being: developing an ‘affordances’ framework for planning and design. Journal of Urban Design, 22(6), 778–795. https://doi.org/10.1080/13574809.2017.1336058

Lesage, J. P. (2008). An Introduction to Spatial Econometrics. Revue D’Économie Industrielle, 123(3), 19–44. https://journals.openedition.org/rei/3887?file=1

LeSage, J., & Pace, R. K. (2009). Introduction to Spatial Econometrics. Revue d’économie Industrielle. https://doi.org/10.1201/9781420064254

Lewandowska-Gwarda, K. (2018). Geographically Weighted Regression in the Analysis of Unemployment in Poland. ISPRS International Journal of Geo-Information, 7(1), 17. https://doi.org/10.3390/ijgi7010017

Li, W., & Saphores, J. D. (2011). A Spatial Hedonic Analysis of the Value of Urban Land Cover in the Multifamily Housing Market in Los Angeles, CA. *Urban Studies*, *49*(12), 2597–2615. https://doi.org/10.1177/0042098011429486

Liebelt, V., Bartke, S., & Schwarz, N. (2017). Hedonic pricing analysis of the influence of urban green spaces onto residential prices: the case of Leipzig, Germany. European Planning Studies, 26(1), 133–157. https://doi.org/10.1080/09654313.2017.1376314

Lin, W., Yu, T., Chang, X., Wu, W., & Zhang, Y. (2015). Calculating cooling extents of green parks using remote sensing: Method and test. Landscape and Urban Planning, 134, 66–75. https://doi.org/10.1016/j.landurbplan.2014.10.012

Lo, A. Y., & Jim, C. (2010). Willingness of residents to pay and motives for conservation of urban green spaces in the compact city of Hong Kong. Urban Forestry & Urban Greening, 9(2), 113–120. https://doi.org/10.1016/j.ufug.2010.01.001

Lu, B., Charlton, M., & Fotheringhama, A. S. (2011). Geographically Weighted Regression Using a Non-Euclidean Distance Metric with a Study on London House Price Data. Procedia Environmental Sciences, 7, 92–97. https://doi.org/10.1016/j.proenv.2011.07.017

Mahan, B. L., Polasky, S., & Adams, R. M. (2000). Valuing Urban Wetlands: A Property Price Approach. Land Economics, 76(1), 100. https://doi.org/10.2307/3147260

Meik, J. M., & Lawing, A. M. (2017). Considerations and Pitfalls in the Spatial Analysis of Water Quality Data and Its Association With Hydraulic Fracturing. Advances in Chemical Pollution, Environmental Management and Protection, 227–256. https://doi.org/10.1016/bs.apmp.2017.08.013

Miller, J., Franklin, J., & Aspinall, R. (2007). Incorporating spatial dependence in predictive vegetation models. Ecological Modelling, 202(3–4), 225–242. https://doi.org/10.1016/j.ecolmodel.2006.12.012

Mueller, J. M., & Loomis, J. B. (2008). Spatial Dependence in Hedonic Property Models: Do Different Corrections for Spatial Dependence Result in Economically Significant Differences in Estimated Implicit Prices? Journal of Agricultural and Resource Economics, 33(2), 212–231. http://www.jstor.org/stable/41220624

Myeong, S., Nowak, D. J., & Duggin, M. J. (2006). A temporal analysis of urban forest carbon storage using remote sensing. Remote Sensing of Environment, 101(2), 277–282. https://doi.org/10.1016/j.rse.2005.12.001

Nappi-Choulet, I., Maleyre, I., & Maury, T. -. P. (2007). A Hedonic Model of Office Prices in Paris and its Immediate Suburbs. Journal of Property Research, 24(3), 241–263. https://doi.org/10.1080/09599910701599290

Nilsson, P. (2008). Estimating economic values of meadows and grazings using hedonic housing modeling and GIS. JIBS, Economics, Finance and Stastistics, 22. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A423135&dswid=2331

Nur Syafigah, A.S., Abdul-Rahim, A.S., Mohd Johari, M.Y., & Tanaka, K. (2018). An economic valuation of urban green spaces in Kuala Lumpur City. An economic valuation of urban green spaces in Kuala Lumpur City, 2018, 469-490. https://www.researchgate.net/publication/324767044_An_economic_valuation_of_urban_green_spaces_in_Kuala_Lumpur_City

NVM. (2021). NVM-bestand Metadata [Dataset].

Openstreetmap. (2022). OpenStreetMap Wiki. Retrieved 15 March 2022, from https://wiki.openstreetmap.org/wiki/Main_Page

Palmquist, R. B. (2005). Chapter 16 Property Value Models. Handbook of Environmental Economics, 763–819. https://doi.org/10.1016/s1574-0099(05)02016-4

Panduro, T. E., & Veie, K. L. (2013). Classification and valuation of urban green spaces—A hedonic house price valuation. Landscape and Urban Planning, 120, 119–128. https://doi.org/10.1016/j.landurbplan.2013.08.009

Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Summer, M., MacQueen, D., Lemon, J., Lindgren, F., O'Brien, J., & O'Rourke, J. (2021). Package 'sp: Classes and Methods for Spatial Data'. CRAN. https://cran.r-project.org/web/packages/sp/index.html

Pietsch, M. (2012). GIS in Landscape Planning. Landscape planning, 55–84. https://books.google.nl/books?hl=nl&lr=&id=Du6dDwAAQBAJ&oi=fnd&pg=PA55&dq=Pietsch+M+(2012)+GIS+in+landscape+planning.+INTECH+Open+Access+Publisher&ots=FmMZxvjMWr&sig=L09UP6tCAIoQ4U63Qf7cWj2EjqQ&redir_esc=y#v=onepage&q&f=false

Plant, L., Rambaldi, A., & Sipe, N. (2017). Evaluating Revealed Preferences for Street Tree Cover Targets: A Business Case for Collaborative Investment in Leafier Streetscapes in Brisbane, Australia. *Ecological Economics*, *134*, 238–249. https://doi.org/10.1016/j.ecolecon.2016.12.026

Provincie Brabant. (2022). Brabantse verstedelijkingsstrategie. Retrieved from https://www.brabant.nl/onderwerpen/ruimtelijke-ontwikkeling/ruimtelijk-beleid/brabantse-verstedelijkingsstrategie#:~:text=Het%20stedelijk%20netwerk%20van%20Brabant,Brabant%20is%20sterk%20in%20beweging.&text=De%20steden%20groeien%20en%20nieuwe%20ontwikkelingen%20volgen%20elkaar%20snel%20op.

Ramn, F. (2021, July 8). OpenStreetMap Data in Layered GIS Format. OpenStreetMap. https://download.geofabrik.de/europe/netherlands.html

Rijkswaterstaat. (2021). NWB wegen - Wegvlakken historie 2020 (RWS) [Dataset]. Overheid.nl. https://data.overheid.nl/dataset/15503-nwb-wegen---wegvakken-historie-2020--rws-#panel-description

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82(1), 34–55. https://doi.org/10.1086/260169

Rovers, V., Bosch, P., Albers, R., & van Hove, B. (2015, March). Climate Proof Cities. Knowledge for Climate. https://www.researchgate.net/publication/274390866_Climate_Proof_Cities_-_Final_Report

Rozing, D. (2020). Hoeveel woningen telt Nederland? CBS. Retrieved from https://longreads.cbs.nl/nederland-in-cijfers-2020/hoeveel-woningen-telt-nederland/

Sander, H., Polasky, S., & Haight, R. G. (2010). The value of urban tree cover: A hedonic property price model in Ramsey and Dakota Counties, Minnesota, USA. Ecological Economics, 69(8), 1646–1656. https://doi.org/10.1016/j.ecolecon.2010.03.011

Sanesi, G., & Chiarello, F. (2006). Residents and urban green spaces: The case of Bari. Urban Forestry & Urban Greening, 4(3–4), 125–134. https://doi.org/10.1016/j.ufug.2005.12.001

Santamouris, M., Ding, L., Fiorito, F., Oldfield, P., Osmond, P., Paolini, R., Prasad, D., & Synnefa, A. (2017). Passive and active cooling for the outdoor built environment – Analysis and assessment of the cooling potential of mitigation technologies using performance data from 220 large scale projects. Solar Energy, 154, 14–33. https://doi.org/10.1016/j.solener.2016.12.006

Strand, J., & Vågnes, M. (2001). The relationship between property values and railroad proximity: a study based on hedonic prices and real estate brokers'appraisals. Transportation, 28(2), 137–156. https://doi.org/10.1023/a:1010396902050

Seo, K., Golub, A., & Kuby, M. (2014). Combined impacts of highways and light rail transit on residential property values: a spatial hedonic price model for Phoenix, Arizona. *Journal of Transport Geography*, *41*, 53–62. https://doi.org/10.1016/j.jtrangeo.2014.08.003

Suryowati, K., Bekti, R. D., & Faradila, A. (2018). A Comparison of Weights Matrices on Computation of Dengue Spatial Autocorrelation. IOP Conference Series: Materials Science and Engineering, 335, 012052. https://doi.org/10.1088/1757-899x/335/1/012052

Tajima, K. (2003). New Estimates of the Demand for Urban Green Space: Implications for Valuing the Environmental Benefits of Boston's Big Dig Project. Journal of Urban Affairs, 25(5), 641–655. https://doi.org/10.1111/j.1467-9906.2003.00006.x

Talca Universidad Chile, & Sarrias, M. (2020, October 7). Lecture 1: Introduction to Spatial Econometric [Slides]. https://www.msarrias.com/uploads/3/7/7/8/37783629/lecture1.pdf

Thiessen, A. H. (1911). PRECIPITATION AVERAGES FOR LARGE AREAS. Monthly Weather Review, 39(7), 1084–1089. https://journals.ametsoc.org/view/journals/mwre/39/7/1520-0493_1911_39_1082b_pafla_2_0_co_2.xml?tab_body=pdf

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography, 46, 234. https://doi.org/10.2307/143141

Torói, A. (2019). Spatial Econometrics. Lecture 2: Spatial Weight Matrix W. [Slides]. Https://Web.Sgh.Waw.Pl/. http://web.sgh.waw.pl/~atoroj/ekonometria_przestrzenna/2_W_EN.pdf

Tse, R. Y., & Love, P. E. (2000). Measuring residential property values in Hong Kong. Property Management, 18(5), 366–374. https://doi.org/10.1108/02637470010360669

United Nations. (2015). Sustainable Development Goals. Https://Www.Undp.Org/Content/Undp/En/Home/Sustainable-Development-Goals.Html.

World Health Organisation. (2017). Urban green space: a brief for action. Regional office for Europe. https://www.euro.who.int/__data/assets/pdf_file/0010/342289/Urban-Green-Spaces_EN_WHO_web3.pdf%3Fua=1

Zhang, Y., & Dong, R. (2018). Impacts of Street-Visible Greenery on Housing Prices: Evidence from a Hedonic Price Model and a Massive Street View Image Dataset in Beijing. *ISPRS International Journal of Geo-Information*, *7*(3), 104. https://doi.org/10.3390/ijgi7030104

Zhou, W., Huang, G., & Cadenasso, M. L. (2011). Does spatial configuration matter? Understanding the effects of land cover pattern on land surface temperature in urban landscapes. Landscape and Urban Planning, 102(1), 54–63. https://doi.org/10.1016/j.landurbplan.2011.03.009

Zygmunt, R., & Gluszak, M. (2015). Forest proximity impact on undeveloped land values: A spatial hedonic study. Forest Policy and Economics, 50, 82–89. https://doi.org/10.1016/j.forpol.2014.07.005

Appendices

Appendix I: model selection approaches

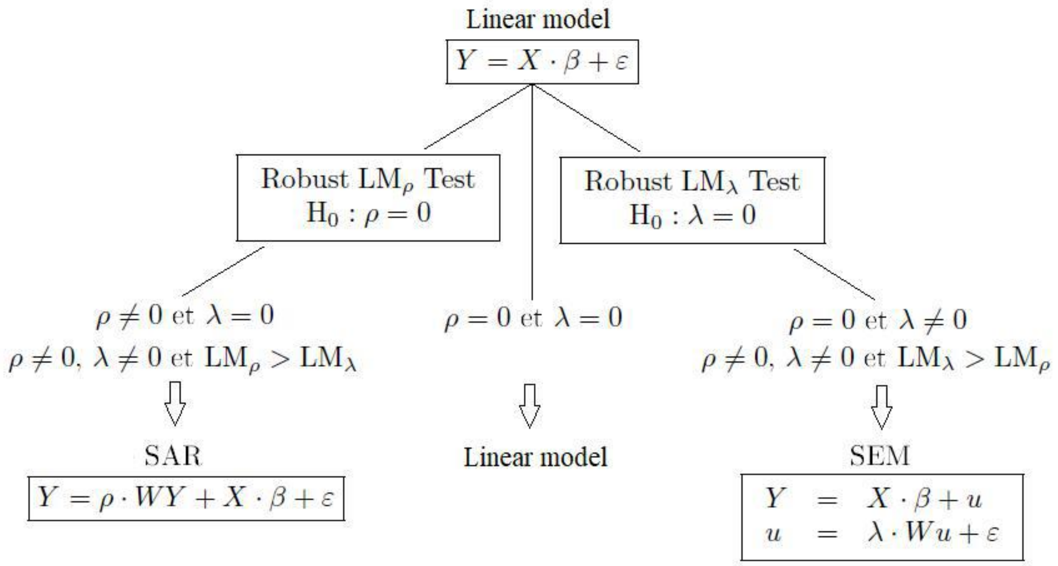


Figure a: the bottom-up approach (Florax et al. 2003)

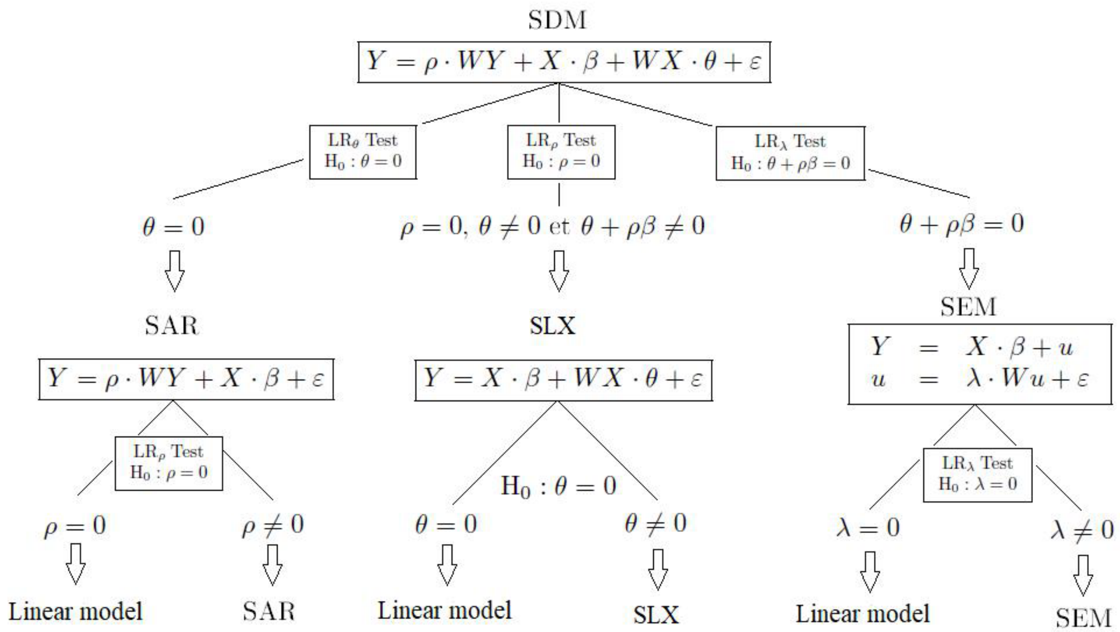


Figure b: the top-down approach (Lesage and Pace, 2009)

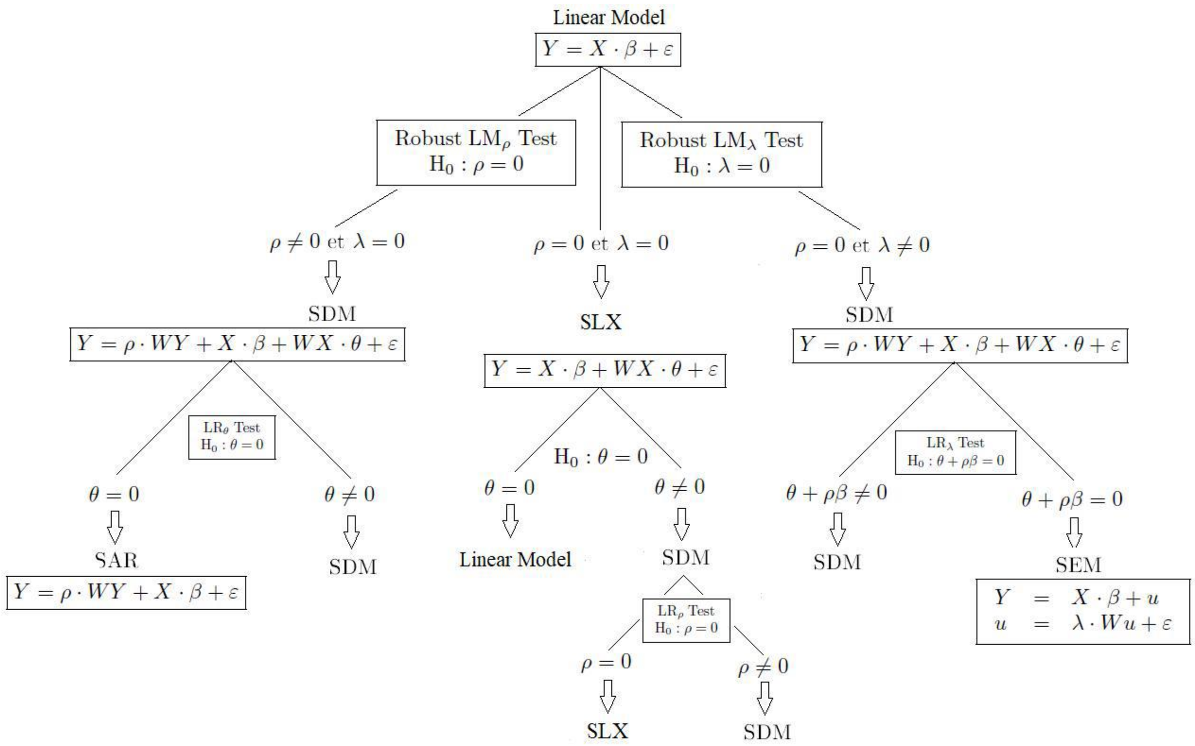


Figure c: the combined-approach (Elhorst, 2010)

Appendix II: tables with available variables

Variable	Description	Measurement level
Living area (m²)	Represent the adjusted floorspace of the dwelling	Continuous
Plot size (m²)	The size of the plot in m²	Continuous
Volume (m³)	Represents the m³ of the dwelling	Continuous
Construction year	Construction period of the dwelling	Categorical **0** = Unknown, **1** = 1500 – 1905 **2** = 1906 – 1930 **3** = 1931 – 1944 **4** = 1945 – 1959 **5** = 1960 – 1970 **6** = 1971 – 1980 **7** = 1981 – 1990 **8** = 1991 – 2000 **9** = ≥ 2001
Living room type	Classifies the type of living room	Categorical **0** = other **1** = L-room **2** = T- room **3** = Z or U – room **4** = through lounge **5** = room and suite
Number of rooms	The number of rooms	Discrete
Number of stories	The number of stories	Discrete
Number of balconies	The number of balconies	Discrete
Number of dormers	The number of dormers	Discrete
Number of rooftop terraces	The number of rooftop terraces	Discrete
Number of kitchens	The number of kitchens	Discrete
Number of pantries	The number of pantries	Discrete
Number of toilets	The number of toilets	Discrete
Number of bathrooms	The number of bathrooms	Discrete
Number of insulation	The number of insulation measures	Discrete
Heating type	Classifies the type of heating system	Categorical **0** = no heating **1** = gas or stove **2** = boiler, district, block or hot air heating **3** = air conditioning or solar collectors
Basement type	Classifies the type of basement	Categorical **0** = no basement **1** = provision basement **2** = boiler room **3 =** provision and boiler room
Parking space type	Classifies the type of parking space	Categorical **0** = no parking space **2** = parking space **3** = carport no garage **4** = garage no carport **6** = garage and carport **8** = large garage
Indoor parking space	Indicates the presence of an indoor parking space	Binary **0** = no **1** = yes
Indoor maintenance	Represent the indoor maintenance of the building	Categorical **1** = excellent **2** = excellent to very good **3** = very good **4** = good to sufficient **5** = sufficient **6** = sufficient to moderate **7** = moderate **8** = moderate to bad **9** = bad
Outdoor maintenance	Represents the outdoor maintenance of the building	Categorical **1** = excellent **2** = excellent to very good **3** = very good **4** = good to sufficient **5** = sufficient **6** = sufficient to moderate **7** = moderate **8** = moderate to bad **9** = bad
Garden orientation	Represents the orientation of the garden in relation to the sun	Categorical **-1** = No dwelling **0** = Unknown or no garden **1** = North **2** = North – east **3** = East **4** = South – east **5** = South **6** = South – west **7** = West **8** = North – west
City location	Indicates the location of the dwelling	Categorical **0** = outside urban areas **1** = unknown **2** = residential area **3** = city center
Road location	Indicates the location of the dwelling in relation to the roads	Categorical **0** = quiet road **1** = unknown **2** = busy road
Scenic location	Represent if the dwelling is located at a scenic location	Categorical **0** = Unknown **1** = At the edge of a forest **2** = by water **3** = next to a park **4** = unobstructed view
Elevator	Indicates the presence of an elevator	Binary **0** = no **1** = yes
Permanent stairs	Indicates the presence of permanent stairs	Binary **0** = no **1** = yes
Office room	Indicates the presence of an office room	Binary **0** = no **1** = yes
Attic	Indicates the presence of an attic	Binary **0** = no **1** = yes
Open portico	Indicates the presence of an open portico	Binary **0** = no **1** = yes
Loft	Indicates the presence of a loft	Binary **0** = no **1** = yes

Monumental status	Indicates if the object has a monumental status	Binary **0** = no **1** = yes
Ground lease	Indicates if there is a ground lease construction	Categorical **0** = no **1** = fixed **2** = variable

Table a: available structural variables (NVM, 2021, own edit)

Variable	Description	Measurement level(s)
Community health center	Building in which one or more practitioners work together	Network distance to nearest health center (km) Average number within 1, 3 or 5 kilometres calculated by road
Hospital exc. outdoor clinic	In a hospital, patients can be admitted for more than 24 hours and major surgeries can be performed. An outdoor clinic is a location of a hospital where non-bedridden patient are treated or monitored.	Network distance to nearest hospital exc. outdoor clinic (km) Average number within 5, 10 or 20 kilometres calculated by road
Hospital inc. outdoor clinic	In a hospital, patients can be admitted for more than 24 hours and major surgeries can be performed. An outdoor clinic is a location of a hospital where non-bedridden patient are treated or monitored.	Network distance to nearest hospital inc. outdoor clinic (km) Average number within 5, 10 or 20 kilometres calculated by road
General practitioner clinic	A place where general practitioners provide evening, night and weekend services	Network distance to nearest clinic (km) Average number within 1, 3 or 5 kilometres calculated by road
Large supermarket	Store with several types of daily groceries and minimum surface of 150 m²	Network distance to nearest supermarket (km) Average number within 1, 3 or 5 kilometres calculated by road
Other stores for daily groceries	Examples of other stores are bakery, toko, coffee and tea store, chocolate store, cheese store, tabaco store, delicacies store, butcher, poulterer, liquor store or mini supermarket.	Network distance to nearest 'other store' (km) Average number within 1, 3 or 5 kilometres calculated by road
Department store	A large shopping store with a variety of goods in different departments	Network distance to nearest department store (km) Average number within 1, 3 or 5 kilometres calculated by road
Daycare center	Place where children aged 0 to 4 years are cared for one or more half-days a week throughout the year.	Network distance to nearest daycare center (km) Average number within 1, 3 or 5 kilometres calculated by road
After school care	Place where children from the primary school are cared for before and after schooltime and/or during the holidays	Network distance to nearest after school care (km) Average number within 1, 3 or 5 kilometres calculated by road
Primary school	In addition to regular primary schools, primary education also includes schools for children of people without a permanent place of residence, the so-called mobile schools and mooring schools for sailing pre-school children. Special primary education and special schools are not included.	Network distance to nearest primary school (km) Average number within 1, 3 or 5 kilometres calculated by road
High school	Schools where student can follow government-funded full-time secondary education. This includes VMBO, HAVO and VWO education.	Network distance to nearest high school (km) Average number within 1, 3 or 5 kilometres calculated by road
VMBO school	High school with VMBO education	Network distance to nearest high school with VMBO education (km) Average number within 1, 3 or 5 kilometres calculated by road
HAVO/VWO school	High school with HAVO/VWO education	Network distance to nearest high school with HAVO/VWO education (km) Average number within 1, 3 or 5 kilometres calculated by road
Main road	Access to a national or provincial road. As a starting point for the driveways, the National Road Database from the Ministry of Infrastructure and Environment, was used	Network distance to nearest entrance main road (km)
Central train station	Access to a central train station	Network distance to nearest central train station (km)

Transfer train station	Access to a transfer train station	Network distance to nearest transfer train station (km)
Pharmacy	Medicine store or dispensing general practitioner	Network distance to nearest pharmacy (km)
Pub	Establishment like pubs, coffeehouses, coffeeshops, discotheques, sex/nightclub or party centers	Network distance to nearest pub (km) Average number within 1, 3 or 5 kilometres calculated by road
Cafeteria	Establishment like cafeteria, fast food joint, grillroom, kebab joint, lunchroom pancake store or ice cream vendor	Network distance to nearest cafeteria (km) Average number within 1, 3 or 5 kilometres calculated by road
Restaurant	Venue where people can go to eat or order food to be delivered at home	Network distance to nearest restaurant (km) Average number within 1, 3 or 5 kilometres calculated by road
Hotel	Accommodation with sleeping places in predominantly single and double rooms against booking per night	Network distance to nearest hotel (km) Average number within 5, 10 or 20 kilometres calculated by road
Theater	Establishments like concert hall, pop podium, neighbourhood center or theater. Locations with a performance arts as a secondary activity are not taken into account.	Network distance to nearest theatre (km) Average number within 5, 10 or 20 kilometres calculated by road
Pop podium	Building or outdoor area were performance art, festivals or music performances are given. Locations with a performance art as a secondary activity are not taken into account.	Network distance to nearest pop podium (km)
Library	Public institution were books, films or music can be borrowed. Libraries where taken from the G!DS database.	Network distance to nearest library (km)
Swimming pool	Public swimming pool	Network distance to nearest swimming pool (km)
Tanning salon	Establishment where customers can expose themselves to ultraviolet light to darken their skin	Network distance to nearest tanning salon (km)
Museum	Public institution were object of historical, scientific, artistic or cultural interest are exhibited	Network distance to nearest museum (km) Average number within 5, 10 or 20 kilometres calculated by road
Amusement park	Establishment like an amusement park, indoor play hall or zoo	Network distance to nearest amusement park (km) Average number within 5, 10 or 20 kilometres calculated by road
Ice rink	Ice rink, open during winter season	Network distance to nearest ice rink (km)
Cinema	Movie theater for viewing films for entertainment	Network distance to nearest cinema (km)
Sauna	Building designed to experience dry or wet heat sessions	Network distance to nearest sauna (km)
Fire department station	Fire department station	Network distance to nearest Fire department (km)

Table b: available locational variables (CBS, 2021, own edit)

Layer	Class	Description
Land use	Forest	Forest or woodland
Land use	Park	Park
Land use	Residential	Residential area
Land use	Industrial	Industrial area
Land use	Cemetery	Cemetery or graveyard
Land use	Allotments	Area with small private gardens
Land use	Meadow	Meadow, possibly used for grazing cattle
Land use	Commercial	Commercial area
Land use	Nature reserve	Nature reserve
Land use	Recreation ground	An open green space for general recreation ground
Land use	Retail	Area mainly used by shops
Land use	Military	Military land use, usually no access for civilians
Land use	Orchard	Area used for growing fruit-bearing trees
Land use	Vineyard	Area used for growing grapes
Land use	Scrub	Area where scrub grows
Land use	Grass	Area where grass grows
Land use	Heath	Heath areas
Land use	National Park	National park
Land use	Farmland	Agricultural land (areas where crops are grown)
Land use	Farmyard	Area of land with farm building and trees around them
Water	Water	Unspecified bodies of water. Typically lakes, but can also be large rivers, harbours, etc.
Water	Reservoir	Artificial lakes, typically above a dam
Water	River	Polygons for larger rivers
Water	Dock	Dock (to repair ships)
Water	Glacier	Glaciers
Water	Wetland	Wetland
Roads	Motorway	Motorway/freeway
Roads	Trunk	Important roads, typically divided
Roads	Primary	Primary roads, typically national
Roads	Secondary	Secondary roads, typically regional
Roads	Tertiary	Tertiary roads, typically local
Roads	Unclassified	Smaller local roads
Roads	Residential	Roads in residential areas
Roads	Living street	Streets where pedestrians have priority
Roads	Pedestrian	Pedestrian only streets
Roads	Service	Service roads for access to buildings, parking lots etc
Roads	Track	For agricultural use, in forests, etc. Often gravel roads
Roads	Bridleway	Paths for horse riding
Roads	Cycleway	Paths for cycling
Roads	Footway	Footpaths
Roads	Unspecified paths	Unspecified paths
Roads	Steps	Flights of steps on footpaths
Roads	Unknown	Unknown type of road or path

Table c: description of the variables in the OSM data set (OpenStreetmap 2021, own edit)

UGS class	**Almere**				**Eindhoven**			
	Mean	St.dev	Min	Max	Mean	St.dev	Min	Max
Allotments	1	0.64	0.00	3.21	1.47	0.75	0.07	4.08
	1.32	0.77	0.02	3.66	1.93	0.82	0.05	5.06
Beaches	2.71	1.98	0.5	8.03	3.26	1.23	0.58	6.9
	3.26	2.21	0.07	9.10	3.92	1.59	0.57	7.87
Cemetery	4.53	3.02	0.06	10.56	1.12	0.7	0.03	3.73
	4.73	2.85	0.10	11.21	1.45	0.87	0.01	4.22
Farmland	1.94	1.04	0.07	4.52	1.21	0.56	0.05	3.15
	2.24	1.10	0.04	4.75	1.58	0.65	0.03	3.27
Orchard	1.18	0.51	0.15	3.16	1.55	0.74	0.02	3.83
	1.56	0.61	0.05	3.30	1.84	0.77	0.05	3.53
OGS	1.91	0.92	0.11	4.47	1.6	0.89	0.01	4.0
	2.36	1.01	0.09	5.38	1.9	0.92	0.02	5.88
Meadow	1.5	0.81	0.06	4.91	1.55	0.74	0.06	4.16
	2.06	0.99	0.01	5.94	2.1	0.84	0.08	5.71
Park	0.55	0.38	0.00	2.88	0.5	0.29	0.01	1.80
	0.74	0.48	0.02	3.42	0.71	0.40	0.00	3.65
Scrub	2.64	1.32	0.00	5.38	0.9	0.49	0.02	2.7
	2.36	1.01	0.09	6.1	1.18	0.59	0.02	4.01

**The top numbers show the Euclidean distance in kilometres while the figures below show the Network distance in kilometres.*

Table d: Descriptive statistics UGS distances in kilometres

Appendix III: overview explanatory variables and VIF scores

Variable	Description	Measure	Expected impact
Dependent variable			
Transaction price (log)	Price x €1000	Number	
Structural characteristics			
Living area (log)	Log of living area in m^2	Number	+
Plot size (log)	Log of plot size in m^2	Number	+
Volume (log)	Log of volume in m^3	Number	+
Building year 1960 – 1980	Building year (0/1)	Binary	-
Building year 1981 – 1990	Building year (0/1)	Binary	+
Building year 1991 – 2000	Building year (0/1)	Binary	+
Building year 2001 ≥	Building year (0/1)	Binary	+
Rooms	Number of rooms	Number	+
Stories	Number of stories	Number	+
Isolation measures	Number of isolation measures	Number	+
Kitchen	Dwelling has a kitchen (0/1)	Binary	+
Bathroom	Dwelling has a bathroom (0/1)	Binary	+
Basement	Dwelling has a basement (0/1)	Binary	+
Dormer	Dwelling has a dormer (0/1)	Binary	+
Rooftop terrace	Dwelling has a rooftop terrace (0/1)	Binary	+
Pantry	Dwelling has a Pantry (0/1)	Binary	+
Balcony	Dwelling has a Balcony (0/1)	Binary	+
Heating system	Dwelling has a heating system (0/1)	Binary	+
Parking space	Dwelling has a parking space (0/1)	Binary	+
Garage	Dwelling has a garage (0/1)	Binary	+
Garden	Dwelling has a garden not located at a South direction (0/1)	Binary	+
Garden South	Dwelling has a garden located at a South direction (0/1)	Binary	+
Monument	Dwelling has a monumental status (0/1)	Binary	+
Ground lease	Dwelling has a ground lease construction (0/1)	Binary	+
Locational characteristics			
Supermarket	Distance to nearest supermarket in kilometres	Number	-
Department store	Distance to nearest department store in kilometres	Number	-
Daily groceries	Number of stores for daily groceries within 3 km from the dwelling	Number	+
Primary schools	Number of primary schools within 1 km from the dwelling	Number	+
High schools	Number of high schools within 1 km from the dwelling	Number	+
Daycare centre	Number of daycare centres within 3 km from the dwelling	Number	+
Afterschool care	Number of afterschoolcare centres within 3 km from the dwelling	Number	+
Hospital	Number of hospitals within 20 km from the dwelling	Number	+
General practitioner	Number of general practitioners within 5 km	Number	+
Pharmacy	Distance to nearest pharmacy in kilometres	Number	-
Central train station	Distance to nearest central train station in kilometres	Number	-
Train station	Distance to nearest train station in kilometres	Number	-
Highway	Distance to highway in kilometres	Number	-
Cafeteria	Number of cafeterias within 3 km	Number	+
Pub	Number of pubs within 3 km	Number	+
Restaurant	Number of restaurants within 3 km	Number	+
Theatre	Number of theatres within 20 km	Number	+
Library	Distance to nearest library in kilometres	Number	-
Cinema	Distance to nearest cinema in kilometres	Number	-
Swimming pool	Distance to nearest swimming pool in kilometres	Number	-
Tanning salon	Distance to nearest tanning salon in kilometres	Number	-
Sauna	Distance to nearest sauna in kilometres	Number	-
UGS characteristics			
Allotments	Network distance to nearest allotments in kilometres	Number	-
Beach	Network distance to nearest beach in kilometres	Number	-
Cemetery	Network distance to nearest cemetery in kilometres	Number	+/-
Farmland	Euclidean distance to nearest farmland in kilometres	Number	+
Orchard	Network distance to nearest orchard in kilometres	Number	-
Open green space	Network distance to nearest OGS in kilometres	Number	-
Meadow	Network distance to nearest meadow in kilometres	Number	-
Park	Network distance to nearest park in kilometres	Number	-
Scrub	Network distance to nearest scrub in kilometres	Number	+/-
Grass	Land covered with grass within the 6-digit zip code from the dwelling	%	+

Trees	Land covered with trees within the 6-digit zip code from the dwelling	%	+
Water	Land covered with water within the 6-digit zip code from the dwelling	%	+
Health	Health of the vegetation within the 6-digit zip code from the dwelling based on NIR-radiations	%	+
Control variables			
Terraced	Building type (0/1)	Binary	-
Semi-detached	Building type (0/1)	Binary	+
Corner	Building type (0/1)	Binary	+
Duplex	Building type (0/1)	Binary	+
Detached	Building type (0/1)	Binary	+
2015	Transaction year (0/1)	Binary	+
2016	Transaction year (0/1)	Binary	+
2017	Transaction year (0/1)	Binary	+
2018	Transaction year (0/1)	Binary	+

Table A: overview explanatory variables

Variable	VIF
Living area (log)	3,2
Plot size (log)	5,9
Building year 1960 - 1980	2,5
Building year 1981 – 1990	3,1
Building year 1991 – 2000	3,8
Building year 2001 ≥	4,2
Rooms	3,2
Stories	4,8
Isolation measures	1,9
Kitchen	2,1
Bathroom	1,6
Dormer	1,2
Rooftop terrace	1,1
Pantry	1,7
Balcony	1,4
Heating system	2,4
Parking space	1,3
Garage	1,6
Garden	3,2
Garden South	3,1
Basement	1,0
Monument	1,0
Ground lease	1,8
Supermarket	1,7
Department store	1,6
Primary schools	1,6
High schools	4,2
Pharmacy	1,8
Central train station	4,0
Highway	2,5
Swimming pool	2,1
Allotments	1,8
Beach	3,2
Cemetery	4,6
Farmland	1,9
Orchard	1,7
Open green space	1,4
Meadow	1,5
Park	1,3
Scrub	2,6
Grass	1,4
Water	1,1
Health	1,5
Semi-detached	1,2
Corner	1,2
Duplex	1,5
Detached	1,5
2015	1,8
2016	1,8
2017	1,8
2018	1,8

Table B: VIF-scores OLS variables

Variable	Description	Measure	Expected impact
Dependent variable			
Transaction price (log)	Price x €1000	Number	
Structural characteristics			
Living area (log)	Log of living area in m²	Number	+
Plot size (log)	Log of plot size in m²	Number	+
Building year 1960 - 1980	Building year (0/1)	Binary	-
Building year 1981 – 1990	Building year (0/1)	Binary	+
Building year 1991 – 2000	Building year (0/1)	Binary	+
Building year 2001 ≥	Building year (0/1)	Binary	+
Rooms	Number of rooms	Number	+
Isolation measures	Number of isolation measures	Number	+
Kitchen	Dwelling has a kitchen (0/1)	Binary	+
Dormer	Dwelling has a dormer (0/1)	Binary	+
Rooftop terrace	Dwelling has a rooftop terrace (0/1)	Binary	+
Pantry	Dwelling has a pantry (0/1)	Binary	+
Balcony	Dwelling has a balcony (0/1)	Binary	+
Basement	Dwelling has a basement (0/1)	Binary	+
Heating system	Dwelling has a heating system (0/1)	Binary	+
Parking space	Dwelling has a parking space (0/1)	Binary	+
Garage	Dwelling has a garage (0/1)	Binary	+
Garden South	Dwelling has a garden located at a South direction (0/1)	Binary	+
Monument	Dwelling has a monumental status (0/1)	Binary	+
Ground lease	Dwelling has a ground lease construction (0/1)	Binary	+
Locational characteristics			
Supermarket	Distance to nearest supermarket in metres	Number	-
Department store	Distance to nearest department store in metres	Number	-
Primary schools	Number of primary schools within 1 kilometre from the dwelling	Number	+
Pharmacy	Distance to nearest pharmacy in metres	Number	-
Central train station	Distance to nearest central train station in metres	Number	-
Highway	Distance to highway in metres	Number	-
Pub	Number of pubs within 3 kilometres	Number	+
Cinema	Distance to nearest cinema in metres	Number	-
UGS characteristics			
Allotments	Network distance to nearest allotments in kilometres	Number	-
Beach	Network distance to nearest beach in kilometres	Number	-
Cemetery	Network distance to nearest cemetery in kilometres	Number	+/-
Farmland	Euclidean distance to nearest farmland in kilometres	Number	+
Orchard	Network distance to nearest orchard in kilometres	Number	-
Open green space	Network distance to nearest OGS in kilometres	Number	-
Meadow	Network distance to nearest meadow in kilometres	Number	-
Scrub	Network distance to nearest scrub in kilometres	Number	+/-
Grass	Land covered with grass within the 6-digit zip code from the dwelling	%	+
Trees	Land covered with trees within the 6-digit zip code from the dwelling	%	+
Water	Land covered with water within the 6-digit zip code from the dwelling	%	+
Health	Health of the vegetation within the 6-digit zip code from the dwelling based on NIR-radiations	%	+
Control variables			
Semi-detached	Building type (0/1)	Binary	+
Corner	Building type (0/1)	Binary	+
Duplex	Building type (0/1)	Binary	+
Detached	Building type (0/1)	Binary	+
2015	Transaction year (0/1)	Binary	+
2016	Transaction year (0/1)	Binary	+
2017	Transaction year (0/1)	Binary	+
2018	Transaction year (0/1)	Binary	+

Table C: included variables in the final OLS model

Appendix IV: R script used for model estimations

```
Thesis script
#Author: Roelof Lammes
#Date: 09-03-2022

#installing (spatial) packages
install.packages("spdep")
install.packages("sp")
install.packages("rgdal")
install.packages("spatialreg")

#load libraries
library(rgdal)
library(spdep)
library(spatialreg)

#reading in dataset as a shapefile
spat.data=readOGR(dsn=".", layer="finaldatasetfull")
summary(spat.data)

#reading in spatial weights
w1=read.gwt2nb("maxmin.gwt", region.id=spat.data$ID)
listw.wts=nb2listw(w1) #Minmax weight

#(spatial) regression models

#step 1: define the regression equations
reg.eq1=transact_2~m2_log10+perceel_lo+nkamers+nbalkon+ndakkap+ndakterras+nkeuken+nbijke
uk+parkingspa+garage+tuinzuid+isol+verw+erfpacht_t+kelder+monument+p6_grass+p6_water+p6_t
rees+nd_a_km+nd_b_km+nd_c_km+ed_f_km+nd_m_km+nd_o_km+nd_r_km+nd_s_km+D_superma
r+D_departme+AV1_CAFE+D_highway+D_centrals+D_Cinema+Elementery+D_pharmacy+bwper1+b
wper2+bwper3+bwper4+year2015+year2016
+year2017+year2018+Corner+Duplex+Semi_detac+Detached
options(scipen=7)

#step2: Run OLS model
ols=lm(reg.eq1,data=spat.data)
summary(ols)
lm.morantest(ols,listw.wts)
lm.LMtests(ols,listw.wts, test="all")

#step3: Run SLX model
slx=lmSLX(reg.eq1,data=spat.data,listw.wts)
summary(slx)

#step4: Run SAR model
sar=lagsarlm(reg.eq1,data=spat.data,listw.wts)
summary(sar)

#step5: Run SEM model
```

```
sem=errorsarlm(reg.eq1,data=spat.data,listw.wts)
summary(sem)

#step6: Run SDEM
sdem=errorsarlm(reg.eq1,data=spat.data,listw.wts, etype="emixed")
summary(sdem)

#step7: Run SDM
sdm=lagsarlm(reg.eq1,data=spat.data,listw.wts, type="mixed")
summary(sdm)

#step 8: "Impacts command" is used to calculate the direct, indirect and total marginal effects for SDEM model
impacts(sdem,listw=listw.wts)
summary(impacts(sdem,listw=listw.wts,R=500),zstats=TRUE) #for p values

#Calculating R2 for SDEM model fit
1-(sdem$SSE/(var(spat.data$transact_2)*(length(spat.data$transact_2)-1)))

#Running a spatial Breusch-Pagan test for heteroskedasticity
bptest.sarlm(sdem,studentize=TRUE)

#step 9: "Impacts SDM model"
impacts(sdm,listw=listw.wts)
summary(impacts(sdm,listw=listw.wts,R=500),zstats=TRUE) #for p values

#step 10: Likelihood ratio (LR) Tests: Test Model Restrictions:

#LR Test 1: SDEM restricted to SEM?
LR.Sarlm(sem,sdem)

#LR Test 2:SDEM restricted to SLX?
LR.Sarlm(slx,sdem)

#LR Test 3:SLX restricted to OLS?
LR.Sarlm(slx,sdem)

#LR Test 4: SEM restricted to OLS?
LR.Sarlm(sem,ols)

#LR Test 5: SDM restricted to SAR?
LR.Sarlm(sar,sdm)

#LR Test 6: SDM restricted to SLX?
LR.Sarlm(slx,sdm)
```