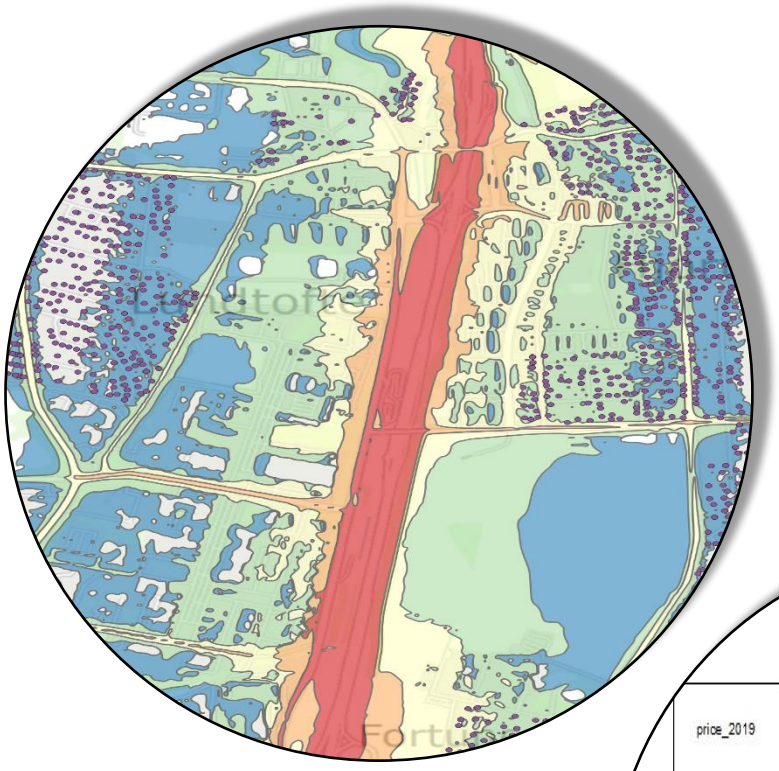
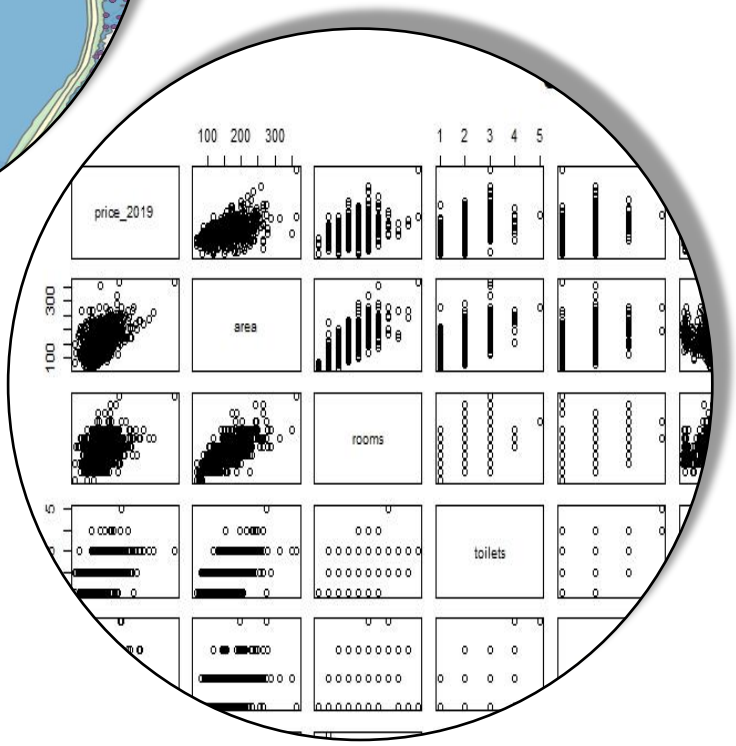
4 JUNE 2020





THE INFLUENCE OF TRAFFIC NOISE ON HOUSE PRICES

MASTER THESIS

KATRINE VALTERSDORF MØLLER
AALBORG UNIVERSITY COPENHAGEN
Geoinformatics 4th semester

AALBORG UNIVERSITET
STUDENTERRAPPORT

Title: The influence of traffic noise on house prices
Semester title: Master thesis

Education: Master's Programme (Cand.Tech.) in Surveying and Planning - Geoinformatics
Semester: 4th semester

Project Period:
February 2020 – June 2020

Group members:
Katrine Valtersdorf Møller

Supervisor:
Jamal Jokar Arsanjani

Number of pages:	71
Number of standard pages:	Approx. 48 (400 words per page)
Number of appendices:	8
Number of ZIP files:	1
Finished:	4 June 2020

Abstract

This project is about how the relationship between house prices and various variables that describe the house and its surroundings can be modeled. From this house price model, the traffic noise is examined to see if noise has any impact on the house price and if so how big is this impact on the house price.

The hedonic price method was used as a method to describe the house price. For the house price model data from BBR, SVUR, and noise data from the Danish Environmental Protection Agency were used. The variables that are used in the model describes the structural characteristics, the location of the house, the surroundings, and the environmental characteristics.
Through the project of modeling the best-fitted model that could describe the relationship between the house price and the variables, several different models have been tested. Models that have been tested are simple linear and logarithmic. OLS regression and Machine Learning have also been used to construct a model.

The best-fitted model was found when using Machine Learning using Linear Regression. This model had an accuracy of 73% which was significantly higher than the other models that have been tested. Based on this model, it is concluded that traffic noise does not have any significant effect on the house price compared to the effect of the other variables.

Resume

Dette projekt omhandler, hvordan man kan beskrive forholdet mellem huspriser og forskellige variabler, som beskriver huset og dets omgivelser. På baggrund af denne husprismodel undersøges trafikstøjs indvikling på husprisen og i hvor høj grad trafikstøj påvirker husprisen.

Til opstilling af en model, der kan beskrive husprisen, anvendes husprismetoden (hedonic regression). Til husprismodellen anvendes data fra BBR, SVUR og støj data fra Miljøstyrelsen. Der anvendes variabler, som beskriver husets strukturelle karakter, husets lokation og omgivelser og miljøet.
Gennem opstilling af den bedst mulige model til at beskrive forholdet med husprisen og variablerne, er flere forskellige modeller blevet testet gennem projekt. Modeller og metoder som er blevet testet, er simple lineære og logaritmiske modeller. OLS regression og Machine Learning er ligeledes blevet anvendt til at opstille en model.

Den bedste model blev fundet ved anvendes af Machine Learning ved brug af Lineær Regression. Denne model havde en nøjagtighed på 73%, som var væsentlig højere end de andre testede modeller. På baggrund af denne model er det konkluderet, at trafikstøj ikke har nogen væsentlig betydning for husprisen, sammenlignet med hvilke indflydelse nogen af de andre variable har.

Preface

This thesis is a result of an ending in my education in Geoinformatics at Aalborg University in Copenhagen. The thesis has been produced in the period from February 2020 to June 2020 and is done by Katrine Valtersdorf Møller.

The focus of the thesis was on how the relationship between house prices and different variables could be explored and from this see if traffic noise has an impact on the house prices that are placed along a highway. Noise is something that all are affected by, but it can be difficult to relate to. Through this project, models have been made to put a value on the traffic noise and see have big an influence the noise will have on a house price close to a road.

Through the work of the thesis data from The Building and Dwelling Register (BBR), State Sales and Valuation Register (SVUR) and the Danish Environmental Protection Agency were used. I would like to thank Lyngby-Taarbæk municipality for providing data to the project and thanks to my supervisor Jamal for guidance through the project.

For citations of the literature that are used in this report, the Harvard Method has been used. The citations are also used for figures, tables, and maps and if they do not have a reference it is my production.
The report does not contain all the results of the models and all tables. This can be found in the appendix report and zip files.

Terms of Reference

List of the abbreviations used in the report

- BBR = The Building and Dwelling Register
- CPR = Civil Registration System
- SVUR = State Sales and Valuation Register
- OLS = Ordinary Least Squares
- GWR = Geographically Weighted Regression
- Lden = The average noise level in a day
- Lnight = The noise level in the night
- dB = Decibel
- VIF = The Variance Inflation Factor

Table of contents

List of Figures

List of Tables

1 Introduction

In Denmark, there is a big focus on the infrastructure and more highways, railways, and more are planned. The government wants to have a good and effective infrastructure that connects Denmark. The infrastructure is a help for the citizens to easier go to work and visiting family and friends. A good infrastructure is also good for the companies and the customers and helps the growth and generate jobs. The government has plans for more investment in better infrastructure. Creating new highways, extend existing highways, and making new connections in the cities. (Statsministeret, 2020) This expansion of the infrastructure structure will create more traffic noise which is one of the major inconveniences of the infrastructure. This could have a negative impact on the surrounded areas.

Many people living in big cities or near the major roads and railways are affected by traffic noise. The properties close to the roads and people living in the cities are constantly affected by the noise. Some more than others. In 2011 one out of three European were bother by the noise in the day and one out of five were disturbed by the noise in the night. (Miljø- og Fødevareministeriet, 2020a)
In Denmark, many citizens are affected by traffic noise that exceeds the noise level for what is health-related correctly. In 2012 approximately 723,000 houses are affected by a noise level that is more than the Danish Environmental Protection Agency's threshold requirement. (Miljø- og Fødevareministeriet, 2020b)

The noise has some negative impact on people's health and according to the World Health Organization WHO it can give, among other things, headache, sleep issues, stress, and hypertension. (Miljø- og Fødevareministeriet, 2020b)

Besides the negative impact on people's health the traffic noise can have a negative impact on the valuation and the commercial value of the surrounding property. Marcel Theebe has stated in a report of how traffic noise affects the house price:

"*Properties located along a road with heavy traffic are likely to sell for less than comparable properties located elsewhere, while properties situated at very quiet locations might even sell at a premium compared to other properties."* (Theebe, 2004)

The biggest noise problem is the noise from the roads, where almost every third house is affected by noise that is over the threshold value. To avoid this the government is working on noise strategies to help reduce the noise from the roads. (Miljø- og Fødevareministeriet, 2020b)
This strategy focuses on helping the municipality in the work of reducing noise, as 9 out of 10 houses are placed along a municipal road. (Miljø- og Fødevareministeriet, 2020c)
The municipalities are working with the noise and trying to find solutions, so the noise affects the people living close to the road as little as possible. The big cities need to make noise plans that show which planes they have and what initiatives they take to reduce the noise.
The strategies from the Danish Environmental Protection Agency have shown that it can be socio-economic to reduce the road noise. By reducing the noise, the inconvenience of the noise gets lower

which is better for the health. Furthermore, the reduction of noise also has a positive effect on the housing market. By reducing the noise housing prices will rise. The single-family house will rise a bit more than 1 % pr. dB, while apartment will rise 0.5 % per. dB. (Miljø- og Fødevareministeriet, 2020c)

1.1 Problem statement and Research questions

The price of a house can be explained using several characteristics related to the house including house quality, house properties, environmental characteristics, and the surroundings. Some of these will have and positive influence on the house price and some will have a negative influence. Traffic noise is a negative effect when living in the cities or near a major road or railway. The valuation and the commercial value of the surrounding properties will be affected by traffic noise. Hence, this study aims at finding answers for the following research questions:

To what degree house prices can be modeled and which method can be used to explore the relationship between different variables and house prices?

How can machine learning be used to create the house price model compared to standard house price models?

Does traffic noise have an influence on property values? If so, how big an impact does traffic noise how on house prices?

2 Literature review

There have been many studies about which impact noise from railways, airports, and roads have on the house prices. In this section, some of these studies will be presented with their approach and results. These presented studies all have the same goal to see what impact the traffic noise has on the house prices, but they have a different approach for the study. The hedonic method has been used in different ways in the studies.

A study done by Blanco and Flindell (2011) looked at how road traffic noise affects the property prices in three different areas in different parts of England. The areas picked are both different in location, with differences in the market and different in the types of property. In these three areas, they are comparing hedonic price coefficients for road traffic noise.
One of the differences in this study compared to other studies is that they are using a lot of socio-economic variables in the hedonic equation. For the analyze they used the hedonic price method with the use of the standard regression equation. To find the best model they compared three different functional forms of the function, linear, log, and semi-log. Which also is the main function used in similar research. The semi-log and log model gave the best-fitted model.
This study gave some interesting results about how traffic noise has different effects on the different urban areas in England. The use of the many socio-economic variables in the hedonic equation that is related to the citizens living in the three different urban areas and are affected by the similar ranges of traffic noise showing a difference in the willingness to pay for lower noise for different people in different areas. The study suggests that some people will pay a higher price for a property that is in an area with higher traffic noise levels.
In London, where the property market is larger than in the two other areas, Birmingham, and Sutton Coldfield, shows an NSDI of 0.45 % per dB increase. But in Birmingham, there was no effect of the noise level and in Sutton Coldfield, the NSDI was positive, with a 5.8 % increase in the offer price with a 10 dB increase in the noise level. Of the variables that were tested in this study, the size of the property, that is measured at the floor area is the most important factor for the price. In London and Sutton Coldfield also the number of rooms was important. The differences in the population density did not have any effect which was a bit unexpected for them. (Blanco & Flindell, 2011)

As was seen in the study by Blanco and Flindell (2011) some people in some areas would pay a higher price in an area there was affected by a high noise level. A study by Rich and Nielsen (2004) from Denmark has revealed the opposite. Here it was shown that there was a clear indication that people would be willing to pay more to avoid areas with more noise. The focus of the study was to analyze the relationship between willingness-to-pay and annoyance. This was done by looking at the NSDI sensibility with different noise cut-off levels.
In the study, a noise assessment model for Copenhagen was created to see the cost of traffic noise on property values. In the study, they distinguish between the houses and apartments as they have different privileges and using a non-linear hedonic regression model. The final model was found with the use of the Box-Cox function, as it is flexible and contains a variety of functions.

One of the outputs of the study shows that the accessibility variables have a big effect on the estimated price. The results of the study showed an NSDI of 0.54 for houses and a bit lower for apartments with an NSDI of 0.47. (Rich & Nielsen, 2004)

A study done in the Netherlands by Theebe (2004) is not only focusing on the noise from cars but is analyzing the impact of traffic noise from cars, trains, and airplanes on the property prices. Like many other studies, the method that is used is the hedonic price method but Theebe has another approach to it. In many other studies, it is assumed that there is a linear relationship between noise and house prices, but this study focusing on estimating the nonlinear relationship.
For the estimating of the prices of noise, an index reflecting the noise is added to the variables that are explaining the property prices. They are also including the positive effects of the infrastructure by including accessibility variables.
To estimate the nonlinear relationship between noise and prices and to adjust for the positive eternality from accessibility variables, a spatial specification with noise dummies were added to their hedonic method.
The results from the study showed that the prices seem to be affected by traffic noise when the level exceeds 65 dB. The negative impact one the prices rise with the sound level and are affected by 5 to 6 percent with a maximum of 12 percent discount when they ignored the extremes. If they look at the quieter areas with a noise level between 40 and 65 dB the impact of noise is under 1 percent or not significant. (Theebe, 2004)

One report that many studies are refereeing to is the study by Wilhelmsson (2000) that is looking at the impact traffic noise has on the values of the single-family houses. As used in a lot of other studies he is also using the hedonic price method. His approach to this analysis is to test different hypotheses. One hypothesis is: *"…that during periods with rapid house price increases, the implicit price of traffic noise may be biased because observed house prices will be dynamic disequilibrium prices."* (Wilhelmsson, 2000). With a Chow test and by looking at recursive regressions he is testing the hypothesis.
This study differs from earlier and similar studies on how the noise variable is included in the hedonic equation. The attempt in the study was to separate the noise effect from the other effect that the roads can generate when estimation the marginal willingness to pay. In the analysis of finding the best model, they used the linear Box-Cox functional form, after a test of a simple linear function. For the Box-Cox the four functions were tested, linear, log-linear, semi-log, and inverse semi-log and found that log-linear was the best model.
The results from the study indicate that if the noise level increase with 1 % the price will be reduced with 0.2 % below 68 dBA and with 0.3% above 68 dBA. There will be a reduction of 0.3 % - 3 % of the property value per decibel. With the analysis, it gave an average noise discount of 0.6% of the property value per decibel, which will give a total discount of 30 % of the price for houses located in a noisy area compared to a similar house in a quiet area. (Wilhelmsson, 2000)

A study from Sweden done by Andersen, Swärdg, and Ögren (2015) is quite different from all other studies about noise impact on house prices. In the study, they are using multiple noise variables by

including both the equivalent level and the maximum level of noise in the estimation of the willingness to pay for traffic noise. The maximum level of noise is included to see which effect it has on the property prices. For the estimation, they are also using the hedonic method and both road noise and rail noise are estimated.
The analysis showed that there is a difference between road and rail. The maximum noise level had a negative influence on the property prices, but only for rails. For the road, the maximum noise did not affect the property prices as the equivalent noise.
Of other variables that were used in the study most of the property characteristics were important, having the accessibility and geographical variables as most important. Living away from a road to avoid the noise is better than the accessibility effect when living near a road.

As seen from these studies there are many ways on how to analyze what impact traffic noise has on property prices. One thing that is similar in all the presented studies is the use of the hedonic price model, but the use of it is different and different variables are being used depending on the focus of the study. Variables that are important to include in the study are the accessibility variables that are mentioned in some of the studies as the most important. Also, how the noise variable is being used in the study is different, some is using the noise as a dummy variable and some are using the noise as it is, while others include other variables to describe the noise.

3 Theory

In this chapter, the theory behind the hedonic price model, that is used for creating a house price model is presented. Furthermore, there is a section about noise in Denmark, how the noise is measured and calculated, and which kind of noise the noise data from the Danish Environmental Protection Agency is.

3.1 Hedonic price model

Noise is not something that has a value and that is being sold on the market. Sometimes it can be necessary to put a value on the noise. Different valuation methods can be used to put a price on the noise. The valuation methods can be either a direct or indirect method. For this project, the indirect method is relevant as the good that needs to be valued has a connection to a good that is on the market. One of the most used methods is the hedonic price method. This method is relevant for this project as we want to see which impact traffic noise has on house prices. (Miljøstyrelsen, 2003)

One of the first to develop the hedonic price method was Rosen (1974) and have since then been used in many different economic projects and is later been developed by Palmquist and Freeman. (Miljøstyrelsen, 2003)
The hedonic prices are defined by Rosen as: *"… the implicit prices of attributes that are revealed to economic agents from observed prices of differentiated products and the specific amounts of characteristics associated with them*." (Owusu-Ansah, 2013)

With Rosen's definition, it is assumed that housing is a heterogeneous good that can be viewed as several different characteristics related to the location, the house structure, and the environmental surroundings. Each house will have its unique characteristics and no houses are the same. All these characteristics that the house can have are not traded explicitly and do not have a direct price on the market. (Owusu-Ansah, 2013)

The hedonic price method is therefore used when non-market variables like road, noise, air quality need to be valued. In general, the method assumes that the price is reflecting a combination of the effect from different variables and characteristics of the house like the number of rooms, size, location, and environmental variables like traffic noise, etc. If a house mainly has positive characteristics it is assumed that the price will be higher than for a house that mainly has negative characteristics. (Blanco & Flindell, 2011)
To find the value of the non-market variables a statistical analysis can be made. The price of the house will be regressed on all the selected variables that will affect the house price. This can be described by the following function:

$$P = f(S_i; N_i; Q_i) + \epsilon$$

(Blanco & Flindell, 2011) and (Rich & Nielsen, 2004)

Here is P a vector of the house price, S is a vector of physical variables, N is a vector location variable, Q is a vector of environmental variables and ε represents the error-term. F represents the functional function in the hedonic model. (Blanco & Flindell, 2011) (Rich & Nielsen, 2004)

The hedonic price coefficient for the sound levels variables can be described as $\frac{\Delta P}{\Delta Q_i}$ which is the marginal implicit price of noise. (Blanco & Flindell, 2011) It is the marginal willingness to pay for a given variable here the traffic noise. (Rich & Nielsen, 2004)

The idea with the hedonic regression is that it will show the relationship between a dependent variable for instance the house price and an independent variable like the number of rooms or one of the house's other characteristics. (Owusu-Ansah, 2013)

The functional function, f, of the hedonic price method is not given by the literature but needs to be described from the use of the data and the selected variables. These variables can either be of a continuous form which is a variable that can have any value such as the size or age or the variable can be a dummy-variable which is a variable that only can two values like 0 or 1 for having a view or not. Usually, there would either be a linear or logarithmic correlation between a variable and the house price. (Miljøstyrelsen, 2003)
Many previous studies have rejected the linear function of the hedonic equation. The most used function in the previous studies is the Box-Cox function, as it contains the relevant functions like linear, logarithmic, etc. and through a test can decide which functional function that fits best to the data. (Rich & Nielsen, 2004) (Miljøstyrelsen, 2003)

There are different ways of modeling the hedonic price model. One is the parametric approach where the regression curve in the hedonic model will have some pre-specified functional form. These will be described as a finite set of parameters that will be the coefficients of the independent variables. An example of a function can be the multiple parametric hedonic function which can be as following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_i X_i + u$$

(Owusu-Ansah, 2013)

Here is Y the dependent variable and X_1 to X_i are the independent variables. The β are unknown parameters that need to be estimated through the equation. U is the error, that is assumed to be normally distributed with a mean of 0 and having a constant variance σ^2. (Owusu-Ansah, 2013)

One of the advantages of the parametric models is that estimating and interpreting the coefficients is simple. The estimation will be precise, and the interpretation of the model will be easy if the assumptions for the model is correct. But these assumptions that this approach has like the linearity between the dependent and independent variables have been criticized as it can make the model inconsistent and give a misleading relationship between the variables. (Owusu-Ansah, 2013)

The most common parametric models are:

- Log-linear Ordinary Least Squares (OLS) Model
- Box-Cox OLS Model
- Weighted Least Square (WLS) Models

(Owusu-Ansah, 2013)

3.2 Noise

The next sections will focus on noise and how the Danish Environmental Protection Agency is working with noise. Also, what the noise limits are in Denmark and how the noise is being calculated, to see which kind of noise that should be included in the house price model.

3.2.1 Noise limits

To help the public authorities in the valuation of noise pollution there are different noise limits for the different types of noise and different areas. The noise limits are instructive, and the limits are what the Danish Environmental Protection Agency thinks is environmental and health acceptable. They are used in the strategic noise work where the number of houses that are exposed to noise is calculated and is used when the municipality is working with noise plans.
The noise limits that the Danish Environmental Protection Agency has given is a so-called A-weighted equivalent corrected noise level. This equivalent noise level is the noise average, which is measured over a longer period. If the noise has some high tones, then 5 dB is added to the equivalent noise level to have the right noise impact. (Miljø- og Fødevareministeriet, 2020e)

These are the recommended limits for noise in housing areas:

- Road noise Lden is 58 dB
- Train noise Lden is 64 dB
- Airplane noise Lden 55 dB

(Miljø- og Fødevareministeriet, 2020d)

The noise is measured as Lden which is the daily weighted average of the noise. The noise in the evening is added 5 dB and the noise in the night is added 10 dB before the average value of the noise is calculated. (Miljø- og Fødevareministeriet, 2020g)

3.2.2 Noise map

The Danish Environmental Protection Agency has created a noise map (Figure 1) that gives a view of the noise impact from the major road, railways, and the noise in the big urban areas (Copenhagen areas, Aarhus, Odense, and Aalborg). (Miljø- og Fødevareministret, 2020)

The biggest cities are required to measure the noise. The cities are the 14 municipalities in Copenhagen areas: Copenhagen, Frederiksberg, Tårnby, Hvidovre, Brøndby, Vallensbæk, Rødovre,

Glostrup, Alberslund, Ballerup, Herlev, Gladsaxe, Gentofte and Lyngby-Taarbæk. Furthermore Aarhus, Odense and Aalborg need to measure and map the noise.
The noise that needs to be measured is noise from all major roads, major railways, and the noisiest companies. The mapping and measuring of the noise are done by the responsible authorities. The municipalities measured the noise from the municipal roads and the other roads are measured by the Ministry of Transport. (Miljø- og Fødevareministeriet, 2020d)

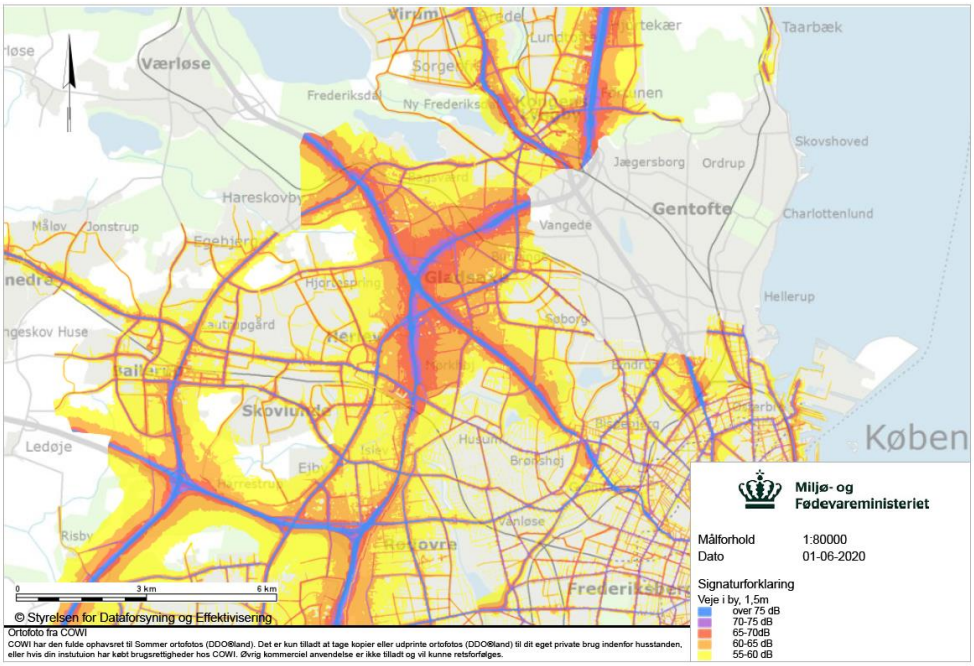


Figure 1: Noise map. Source: (Miljøstyrelsen, 2020)

The noise is mapped for each type of noise and is mapped in 4 different noise groups, showing the daily noise and the noise in the night, which is measured at different heights:

- Daily average (Lden) of noise at a height of 1.5 meters
- Daily average (Lden) of noise at a height of 4 meters
- Noise in the night (Lnight) at a height of 1.5 meters
- Noise in the night (Lnight) at a height of 4 meters

(Miljø- og Fødevareministeriet, 2020d)

3.2.3 Noise calculation

To calculate the noise a Nordic method of calculation, Nord2000, is used. This method is more precise than the other method that was used before. Nord2000 can calculate the noise "moving/spreading" under different weather conditions. This allows calculating the average noise level of a year. The average noise level of a year is calculated by calculating the noise in 9 different weather conditions. This is added up having different weights depending on how often the different weather conditions appear. Depending on the area that the noise calculations are done in, the calculations time can be long. In these situations, only 4 weather classes are used and in the big cities, only 1 weather class can be used. (Miljø- og fødevareministeret, 2020h)

4 Data

In this chapter, the data that have been used in this project will be presented. Data have mainly been used from The Building and Dwelling Register (BBR) and the Environmental Protection Agency. This data has been used in the main analysis. Also, data from Kortforsyningen have been used, which have been used in the data preparation.

4.1 BBR - The Building and Dwelling Register

I Denmark there is data registering about the buildings and dwellings. The Building and Dwelling Register (BBR) have a lot of different datasets that are categorized into 3 different levels depending on the detail of the building or dwelling. The 3 different levels (Figure 2) of the data are property, building, and dwelling. (Bygnings- og Boligregistret, 2009)

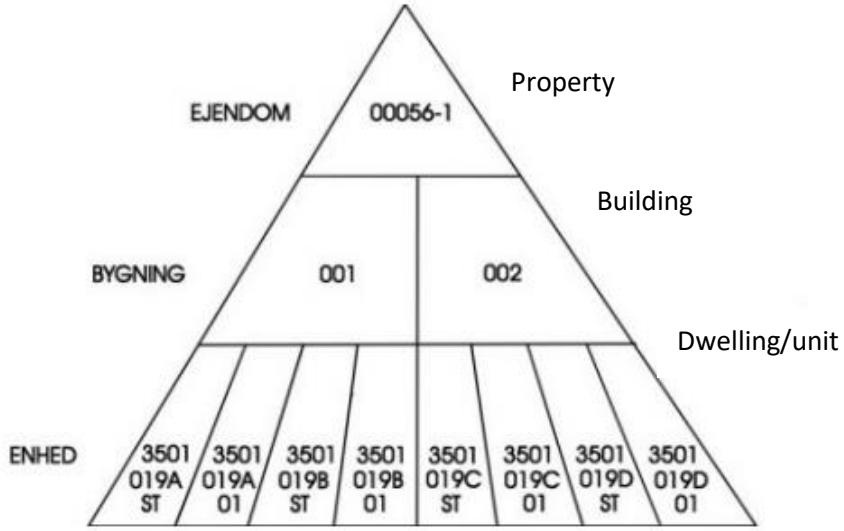


Figure 2: The levels in BBR. Source: (Bygnings- og Boligregistret, 2009)

At the property level, the data is about all the buildings on the property and covers the hole establish property. These data are more general and are information that is the same for all the buildings or dwelling on the property. (Bygnings- og Boligregistret, 2007a) Data on the property level contains information about the number of buildings, number of apartments, building area, house number, etc. (Bygnings- og Boligregistret, 2020)

At the building level, the data are more specific about each building on the property. (Bygnings- og Boligregistret, 2007b) The data on the building level contains information about the age of the building, the material used in the building, area of the building, area of carport/garage/annex, etc. (Bygnings- og Boligregistret, 2020)

At the dwelling level, the data are for each unit, which is defined as the dwelling that has its own entrance and address. (Bygnings- og Boligregistret, 2007c) A dwelling could be a single-family house with an address, or it could be an apartment, which will be more dwellings in a building. Data on the dwelling level is about the specific unit and contains information about the total size, number of rooms, toilets, bathrooms, kitchens, house number, type of house, etc. (Bygnings- og Boligregistret, 2020)

In this project, the data from the building level and the dwelling level are being used. The data that is used from the building level (BBR_bygning) is the building and renovation year and the size of the garage, carport, and annex.
The data from the dwelling level (BBR_enhed) contains all kinds of buildings and houses. As the analysis of the project is to see what impact noise has on houses the information about the use of the house is used in the project. Furthermore, information about rooms, bathrooms size, etc. will be used. This will be described in more detail in section 5.3 about the variables for the analysis.

4.2 SVUR – State Sales and Valuation Register

For the project, the sales price of the properties is being used. The data about the sales prices of the different properties are found in State Sales and Valuation Register (SVUR). This register contains information about the property value and land value. Furthermore, it also contains sales prices of all properties. The information in SVUR comes from Valuation Register (VUR), while the sales price comes from the land charges register. (Ejendomsinfo, 2020) The valuation of the properties in SVUR are based on the information that is in BBR about property age, size, installations, etc. (Boligejer, 2020) In the project, the data set, sales price, from SVUR, is being used, which contains information about the sales price and the date for the sale of the property.

4.3 Noise data

As it is described in the theory in section 3.2.2 about noise map, the Danish Environmental Protection Agency has a map of the noise from the major roads and the noise from the roads in the biggest cities.
The measured noise in the map is shown in different colors depending on how high the noise is. The colors go from blue, which is the highest noise level, Lden over 75 dB, to yellow which is the lowest measured noise, Lden between 55-59 dB. Noise lower than 55 dB is not measured. The noise is mapped as the weight averaged of the day (Lden) and as the noise is at night (Lnight). (Miljø- og Fødevareministeriet, 2020d)

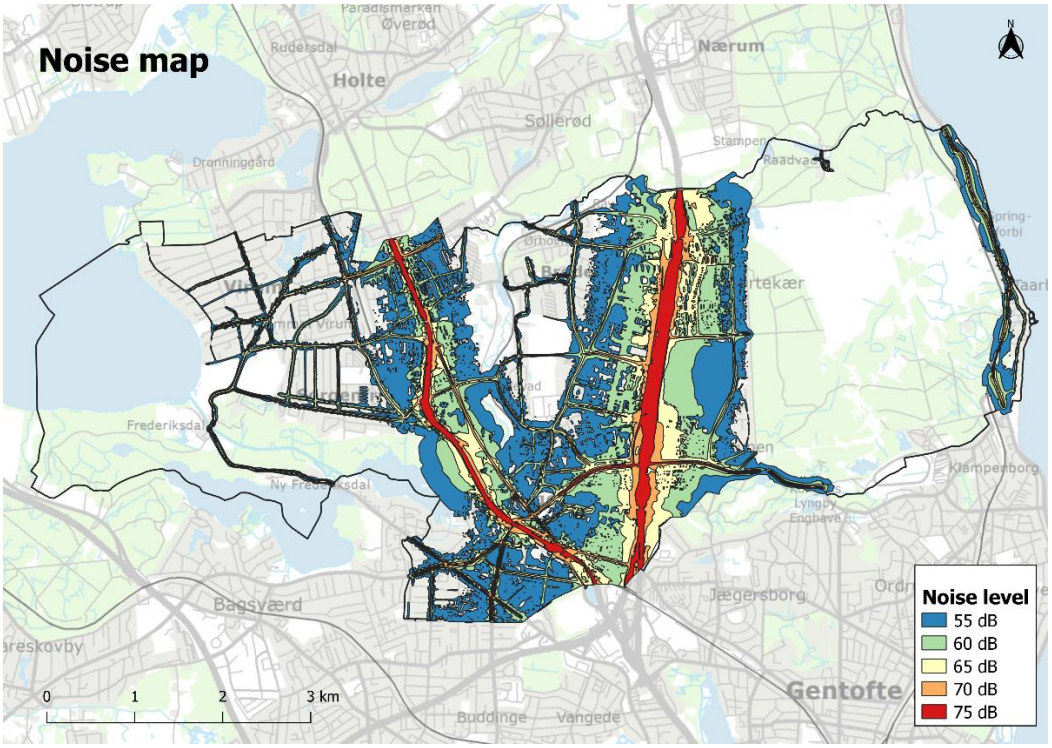


Figure 3: Overview of the noise in Lyngby-Taarbæk municipality.
Data from the Environmental Protection Agency, Kortforsyningen, and GeoDanmark.

The data from the Danish Environmental Protection Agency have different datasets, some with noise from all the major roads in Denmark and some datasets only with noise from the roads in the biggest cities. For this project, the dataset with the noise from all roads measured at 1.5 meters will be used (Figure 3). The noise data is measured in 5 class 55-60 dB, 60-65 dB, 65-70 dB, 70-75 dB and above 75 dB. The noise level that is showed on the map and is the value in the attribute table of the dataset, is the lowest noise in the noise group. For 55-60 dB the noise level is set to 55 dB, for 60-65 dB the noise level is set to be 60 dB, etc. This can give some issues as the noise in the different noise areas can be higher in some places than is registered in the dataset.

5 Method

In this chapter, the method that is used in the project will be described. Furthermore, some of the choices that are made in the project are going to be presented. The first part of this chapter will give an overview of the process of the project, which will show the connection between the different parts of the project. Here follows a description of the variables and how the work has been with the model.

5.1 Process of the project

Figure 4 gives an overview of the process of the project. The colors in the diagram indicate the different processes in the project. The blue color is the beginning and ending of the project and is a more general part of a project. The orange color is the process of finding a good model that can describe the connection between the house price and the different variables which are described more detailed in chapter 7. The green color is the result of the models which is presented in chapter 8.

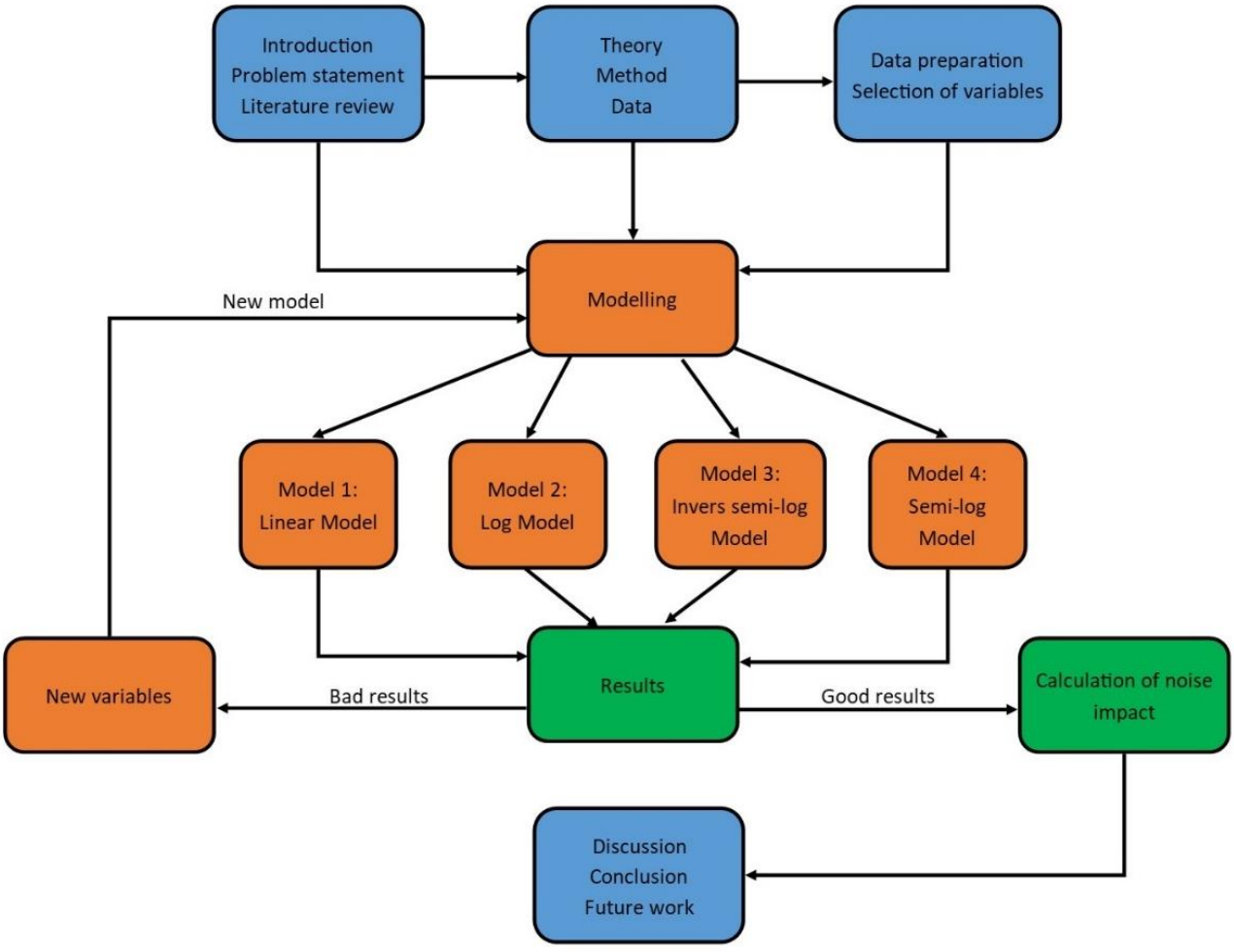


Figure 4: Process of the project.

The first part of the project includes the introduction, problem statement, theory, data preparation, etc. which leads up the main part of the project the modeling. All the first parts (blue colors) of the project contribute to the modeling part.
In the modeling part, the hedonic price model is used, and different models are being tested to find a model that best could describe the relationship between house prices and the variables. The result of the models are being evaluated and if the models do not give any good results new variables are being added to the model. Then the model is being tested again. As the results are better it is possible to conclude on how big an influence the noise has on house prices. From the results of the model the discussion, conclusion, and future work follows.

5.2 Study area

The study area (Figure 5) for the project is Lyngby-Taarbæk commune as I have access to data here and I also made an internship in this area. For the modeling part, a smaller area in the municipality has been chosen. There is a major highway that is going through the municipality and an area around the highway is chosen to be the study area.

Figure 5: Study area with noise zones and houses.
Data from the Environmental Protection Agency, Kortforsyningen, and BBR.

This is chosen as it is the biggest area where the noise is measured and the variation in the noise levels that the properties are affected by is highest here. Also, the area along the highway is chosen as it is assumed that the traffic noise is more constant here than on the smaller road in the city.
As the project only focuses on the single-family houses the areas outside the main cities are more optimal for the project.

5.3 Variables

For the use of the hedonic model several different variables need to be identified, which can describe the house price. It is important to find the variables that have an influence on the house price to find the model that describes the house price best. Too many variables can also be bad for the model as there would be a risk of multicollinearity, which can cause a bad model. (Miljøstyrelsen, 2003)

In this section, the variables that will be used in the hedonic price model will be described. First, the house price will be presented followed by the variables that are describing the house and the surroundings. These variables can be divided into 3 different groups: Structural variables, Environmental variables, and Accessibility variables. The variables come from the data set described in section 4, BBR, SVUR, and the Environmental Protection Agency.

5.3.1 House price

The SVUR data gives information about the sales prices from 1992 to 2020. The prices from different years cannot be compared directly as there is a development in the housing market and prices can go up and down. This means that the same house or similar houses can be sold at two very different prices depending on the year the house is sold.
To avoid this and make the comparison of the house prices easier, the price will be changed to be the price as if they all were sold in 2019. This is done by using the price index from Statistic Denmark. (Danmarks Statistik, 2020a)
The price index that is used to calculate the 2019 price is the price index that covers a year, going from 2006 to 2019. Each year has a price index. The price index for the year is calculated as an unweighted average of the quarter index. This can give some minor differences. (Danmarks Statistik, 2019) The price index that is used in the calculation can be seen in Table 1.
To calculate the 2019 price the following equation is used including the new and old price index:

$$\frac{price \cdot new\ index}{old\ index} = new\ price$$

(Danmarks Statistik, 2020b)

SVUR data contains different types of prices. The price that is being used is the “Købesum beløb” which is the purchase price that was agreed upon the sale of the property.

5.3.2 Structural variables:

The hedonic price model contains different characteristics of the house. For this project, the main variables that are selected to make the model are variables that are describing the house.
The following selected variables are describing the structure of the house: size, number of rooms, number of toilets, number of bathrooms, age, and renovation year. Later variables as lot size, garage area, carport area, and annex area were included.

5.3.3 Environmental variables:

The focus for the project is to see which impact noise has on the house price, the only environmental variables are going to be the noise. The noise data is collected from the Environmental Protection Agency. As described in the theory about the noise data the noise is measured in different ways.
For this project, Lden at 1.5 meters is used, which is the noise on average of the whole day at 1.5 meters. The noise is categorized into 5 bands in a 5 dB interval going from 55 to 75 dB. The noise that is used in the project is not only the noise that is measured from the highway but is also the noise from the secondary roads.

5.3.4 Accessibility variables:

To gets some more variables included in the hedonic price model accessibility variables got add through the project. The new variables focus on the surroundings and location. Being close to green areas, school and the city can have a positive influence and something that buyers are looking for when buying a house.
A straight line from the house and to different locations was measured in QGIS. This accessibility variables are describing the distance to the train station, Lyngby centrum, business and industrial areas, the school, the coast, the forest, and lakes.

5.4 Hedonic price model

Through the analysis, the goal is to find if noise has any impact on the house price and how big an impact it has. For the analysis of the noise impact of the house price the hedonic price method as presented in theory 3.1 is being used. There are different ways to use the hedonic method and how the best-fitted model is found as described in the literature and theory.

For the hedonic model, a variety of different variables have been selected. Through the project, more and more variables have been added to the list of variables. The variables are categorized into three different groups, structural, environmental, and accessibility variables, which are the independent variables. The house price variables will be the dependent variable in the hedonic function.

The connection and relationship between the independent variables and the house price are not known. The goal is to find a function that describes this relationship best as possible. It is an iterative process to find the best model and the variables need to be tested in different functions. The process of finding the best-fitted model in the project is shown in Figure 6. In this process, different methods were tested to find what method best could explain the relationship between the variables and the house price.

The start of the process was to test with a few variables from the structural and environmental groups in simple functions like linear, log-linear, semi-log, and inverse semi-log. Next, the same functions are tested again but with more variables to see the effect of having more variables.
The method is to find the variables that are describing the model best, by looking at the results and looking at the relationship to the dependent variables.
To find the best model with the experience from the first part of the analysis new methods were tested. Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR) in ArcMap were tested to see if these methods better could describe the relationship between the variables and the house price. The last part of the analysis was to test different machine learning models to see if they could outperform the more standard models that have been tested.

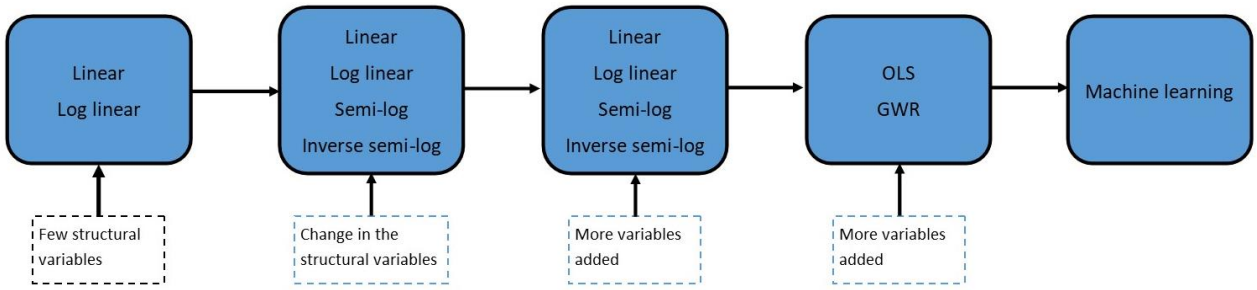


Figure 6: Process of the modeling

This analysis and the work with the models are described in more detailing in chapter 7, where the analysis is presented.

6 Data preparation

This chapter contains the preparation of the data from BBR and from SVUR for a selection of which houses to include in the analysis and preparing of the final dataset for the analysis. The final dataset will contain sold houses in the study area with all the selected variables from section 5.3.

For the data preparation, the tool pgAdmin and QGIS have being used. Furthermore, PostGIS has been used to create a database for the project in QGIS. The data that are used are the BBR data from the property and dwelling level and data from SVUR about sales prices.

For the project, only buildings that are used as a dwelling will be used. The use of the buildings is collected in BBR (BBR_enhed) under ("enh_anvend_kode_t"), which is being sorted and only the relevant use is being selected. The buildings with the following use are selected:

- Linked house (Rækkehus and dobbelthus)
- Single-Family house (Fritliggende enfamiliehus)
- Farmhouse (Stuehus til landbrugsejendom)

Not all the data from the dataset (BBR_enhed) is relevant for further analysis. From the dataset the following attributes are selected:

- Budiling id (Bygning_id)
- Property number (ejd_nr)
- Use of the unit (enh_anvend_kode_t)
- Housing type (boligtype_kode_t)
- Size of the unit (enh_arl_saml)
- Size of dwelling (bebo_arl)
- Number of rooms (vaerelse_ant)
- Number of toilets (antvandskyltoilletter)
- Number of bathrooms (antbadevaerelser)
- Rental type (enh_udlej2_kode_t)
- Coordinates (koornord, kooroest)

Furthermore, only houses that are used by the owner ("Benyttet af ejeren") are selected to avoid having rental properties in the dataset.

```
CREATE EXTENSION postgis;
-- Selecting relevant type of residence (the use of the house) from 'BBR enhed' --
-- and creating af new table 'bolig' --
SELECT bygning_id, ejd_nr, enh_anvend_kode_t, boligtype_kode_t, enh_arl_saml,
bebo_arl, vaerelse_ant, antvandskyltoilleter, antbadevaerelser, enh_udlej2_kode_t,
koornord, kooroest
INTO noise_price.boliger
FROM noise_price.bbr_enhed
WHERE (enh_anvend_kode_t = '(UDFASES) Række-, kæde- eller dobbelthus (lodret adskillelse mellem enhederne).'
OR enh_anvend_kode_t ='Dobbelthus'
OR enh_anvend_kode_t ='Fritliggende enfamilieshus (parcelhus).'
OR enh_anvend_kode_t ='Række- og kædehus'
OR enh_anvend_kode_t ='Stuehus til landbrugsejendom')
and enh_udlej2_kode_t = 'Benyttet af ejeren';
```

The new data table (boliger) is then joined with the data from BBR from the property level (BBR_bygning) to get the age of the building and if the building has been renovated. From the table BBR_bygning, the attribute age (opførelsesår) and renovations year (ombygningsår) are selected. Furthermore, this data set contains information about the garage, carport, and annex size which is also selected. The join is done on the property number and building number, to get the right age of building as the property can contain more buildings with different ages.

```
--- Join the housing (bbr_enhed) and the buildings (bbr_bygning) on ejd_nr
--- Selecting age, garage, carport and annex from BBR bygning
SELECT boliger.*, bbr_bygning.opfoerelse_aar, bbr_bygning.ombyg_aar,
bbr_bygning.garage_indb_arl, bbr_bygning.carport_indb_arl, bbr_bygning.udhus_indb_arl
INTO noise_price.boliger30
FROM noise_price.boliger
JOIN noise_price.bbr_bygning
ON (boliger.ejd_nr = bbr_bygning.ejd_nr and boliger.bygning_id = bbr_bygning.bygning_id);
```

The SVUR data (svur_grund) contains the lot size for each of the properties. The data table is joined with the SVUR data to get the lot size that is connected to each of the houses.

```
---Join boliger30 with lot size from SVUR (svur_grund)
SELECT boliger30.*, svur_grund.grund_pris_kode_t, svur_grund.grund_arl_spec
INTO noise_price.boliger31
FROM noise_price.boliger30
JOIN noise_price.svur_grund
ON boliger30.ejd_nr = svur_grund.ejd_nr
WHERE  grund_pris_kode_t = 'Kvadratmeterpris'
OR grund_pris_kode_t = 'Kvadratmeterpris ved standardberegning' ;
```

The new table (boliger31) now contains all dwellings, except apartments, with the selected structural variables described in section 5.3.2. To find which of the houses that have been sold and the sales price of houses the table is joined with the SVUR data (SVUR_salgspris). From the SVUR data, the attributes about the date of the sale and the sale price are selected. The property can be sold in different ways and only the property sales that are a free sale (code = 1) is selected as this sale is the only one where the sales price is representative. (DinGeo, 2020)

```
34  --- Join boliger31 with the house sales price from SVUR (svur_salgspris)
35  SELECT boliger31.*, price.omregnings_dato, price.koebesum_beloeb,
36  price.overdragelses_kode
37  INTO noise_price.boliger_sold32
38  FROM noise_price.boliger31
39  JOIN noise_price.price
40  ON boliger31.ejd_nr = price.ejd_nr
41  WHERE overdragelses_kode = '1';
```

To visualize the data in QGIS and to select the data in the study area a geometry column is added to the data table from the column koorost and koornord.

```
43  -- Adding a geometry column to the table --
44  -- The geometry column is created by the east and west coordinates in the table --
45  ALTER TABLE noise_price.boliger_sold32
46  ADD COLUMN geom geometry(Point,25832);
47
48  UPDATE noise_price.boliger_sold32
49  SET geom = st_PointFromText('POINT('||kooroest||' '||koornord||')', 25832);
```

A study area along the highway as mentioned in section 5.2 is created in QGIS. To find the houses that are in the study area and the properties from the table (boliger_sold32) is selected.

```
51  -- Selectiong the housing around the highway from the study area
52  SELECT boliger_sold32.*
53  INTO noise_price.house_sold33
54  FROM noise_price.boliger_sold32
55  INNER JOIN noise_price.omraade
56  ON st_within(noise_price.boliger_sold32.geom, noise_price.omraade.geom);
```

Before starting to create a house price model the new dataset has to be clean up and the variables should be adjusted, and the sale prices should be changed so the different sales prices can be compared. As was mentioned in section 5.3.1 the prices need to be changed to be the price if the house were sold in 2019. A price index from Statistic Denmark going from 2006 to 2019 is used (Table 1). Before the recalculating, the dataset is reduced to only have sales that are between 2006 and 2019. To calculate the new price the following equation is used:

$$\frac{price \cdot new\ index}{old\ index} = new\ price$$

(Danmarks Statistik, 2020b)

With the following the price index for 2006-2019:

Year	2006	2007	2008	2009	2010	2011	2012
Price index	100	98.3	89.2	74.6	80	78.7	76.2
Year	2013	2014	2015	2016	2017	2018	2019
Price index	80.3	84.5	91.4	96.7	101.6	106.6	110.2

Table 1: Price index. Source: (Danmarks Statistik, 2020c)

The properties where the sale price is set to be 0 kr. are deleted from the data set. To make the dataset more understandable the names of the variables are changed, and all dates are changed to only being the year.

The dataset is loaded in QGIS to see what noise level the properties are affected by and which properties that are not affected by significant noise being under 55 dB. In QGIS the properties table is joined with the noise data and the column with the noise level is added to the property table.

7 Finding the best model

In this chapter, the modeling of the hedonic price function will be described. Different models will be studied to find the model that best describes the relationship between the house price and its characteristics. The work of finding the best model is an iterative process which is giving different kinds of models and results. The most relevant part of the process will be presented and discussed. Before the modeling part is presented, a discussion of the variables and the expectation of the influence on the house price is described. Followed by a short study of the final data set looking for outliers.

7.1 Variables

All the variables that are being used through the analysis are presented in the method section, where the variables are split into three main groups of different variables that are describing the house structure, environment, and the surroundings.
Before going into the modeling part, it is good to set some expectations of how the relationship between the variables and the house price will be. If the variable is expected to have a positive or a negative impact on the house price.
In the list Table 2 below is the different variables listed with their expected sign:

Variables	Expected sign
Area	+
Rooms	+
Toilets	+
Bath	+
Age	-
Garage	+
Carport	+
Annex	+
Lot size	+
Noise	-
Distance to city	+
Distance to industry/business area	-
Distance to the coast	+
Distance to school	+
Distance to forest	+
Distance to lake	+
Distance to train station	-

Table 2: Variables with the expected sign.

It is mainly expected that the variables have a positive impact on house prices. It is expected that variables that are describing the house would have a positive impact on the house price, except the age of the house where it is expected to be negative as an older house would be less worth than a new house. It is expected that the noise is having a negative impact on the house price.
For variables describing the distance to different places in the area is more difficult to say what influence they might have on the house price. Being close to the city and school is expected to be positive but that also means that you are close to bigger roads and therefore near traffic noise which is negative. Being close to the coast, forest and lakes are also expected to be positive, but this also often means that you are further away from the main cities, which in some cases can be negative. The distance to industry/business area and train stations is expected to have a negative impact on the house price as being close to these also means more noise from trains and the industrial areas.

7.2 Outliers

Before starting to create a hedonic price model the final dataset will be studied. In the following boxplot is studied to see if there are any extreme outliers. The test of outliers is only done on the structural variables, as there can be some errors in these datasets from BBR and SVUR. Only the boxplot where there are outliers to find will be presented.

The boxplot of the area (Figure 7) is showing a few outliers that show that some of the houses are over 250 sqm. It is considered that these few houses are not extreme outliers and it is realistic that some of the houses are over 250 sqm. For the number of rooms (Figure 7), there is one extreme outlier which indicates that a house should have 95 rooms. This does not seem realistic and it seems to be an error in the dataset. The house with the 95 rooms will be removed from the dataset. The rest of the data seems right as the highest number of rooms is 12.

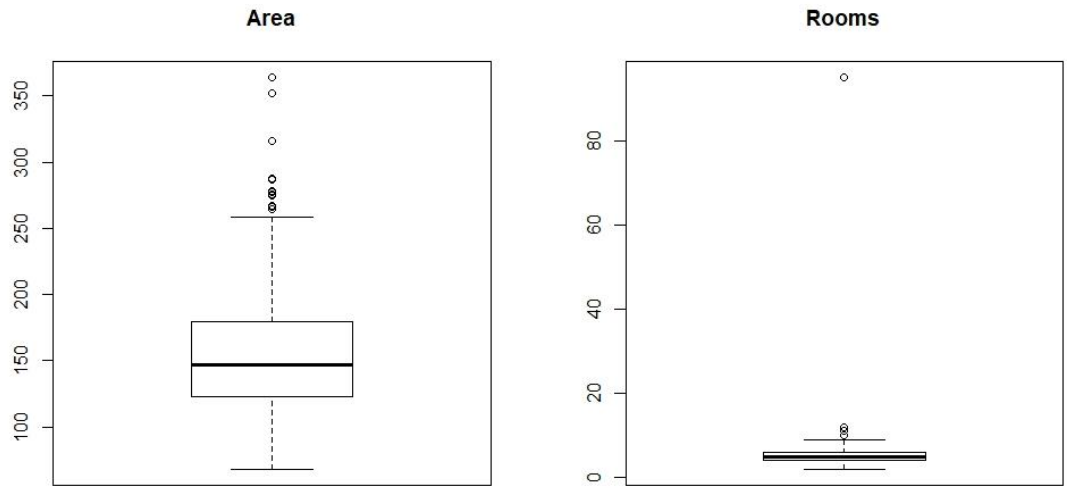


Figure 7: Boxplot of area and rooms.

The boxplots of the toilets and bathrooms (Figure 8) do not show any extreme outliers. There a few houses that have 4 and 5 toilets and some houses have 4 bathrooms. This does not seem unrealistic as some of the houses are between 250-350 sqm.

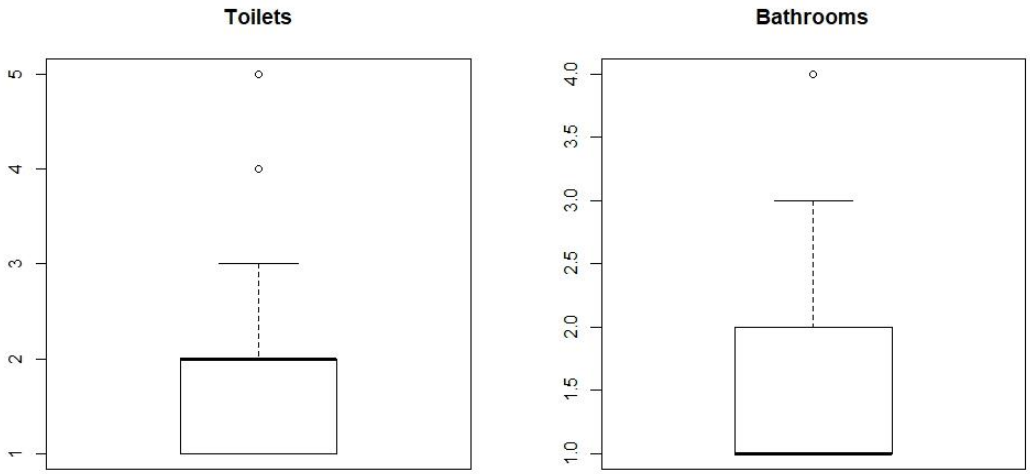


Figure 8: Boxplot of toilets and bathrooms.

The boxplot of the square meter price (Figure 9) shows a huge variation in the price. The sqm price is going from around 100 kr. per sqm to almost 70.000 kr. per sqm. As the study area is just a small part of the municipality and as the houses are placed around the highway it would be considered that the prices are lower here in the rest of the municipality. All houses with a sqm price under 5000 kr. are being removed.

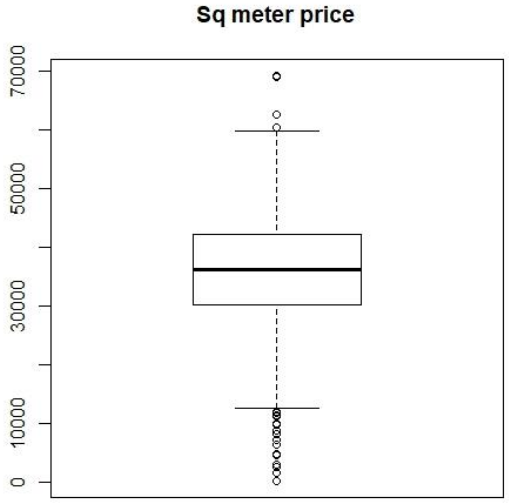


Figure 9: Boxplot of square meter price.

7.3 First model

From the theory of the hedonic price model the standard equation contains the price as the dependent variable and a variety of other variables describing the structure of the house and surrounding being the independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_y X_y + u$$

For the first part of the modeling 7 of the listed variables were used to set up a simple linear model. The independent variables were area, age, bathroom, rooms, toilers, rebuild, and noise. The first linear model that is tested is:

$$Price = \beta_0 + \beta_1 \cdot area + \beta_2 \cdot age + \beta_3 \cdot bathroom + \beta_4 \cdot room + \beta_5 \cdot toilets + \beta_6 \cdot rebuild + \beta_7 \cdot noise + e$$

We wish to estimate the coefficients, β. The exact relationship between the different variables and the sales price is not known. Most of the variables used in the model are continuous variables but rebuild is used as a dummy variable, indicating if the house is rebuilt or not.

The linear model is set up in R, which gave the following results (Table 3):

```
Call:
lm(formula = X2019_pris ~ enh_arl + vaer_ant + anttoil + antbadev +
    isovalue + opfoerelse + rebuild, data = house)

Residuals:
     Min       1Q   Median       3Q      Max
-5502038  -692334     6294   766117  6979092

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17624498    3242227   5.436 6.70e-08 ***
enh_arl        16218       1558  10.407  < 2e-16 ***
vaer_ant       71348      43257   1.649 0.099347 .
anttoil       314753      89591   3.513 0.000461 ***
antbadev      145692      94652   1.539 0.124034
isovalue       -3136       1587  -1.976 0.048363 *
opfoerelse     -8045       1668  -4.823 1.61e-06 ***
rebuild      -205550      89713  -2.291 0.022139 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1350000 on 1109 degrees of freedom
Multiple R-squared:  0.3079, Adjusted R-squared:  0.3036
F-statistic: 70.49 on 7 and 1109 DF,  p-value: < 2.2e-16
```

Table 3: Summary of the linear model.

The estimated coefficient for the 7 variables have the expected relationship with the price, except the rebuild variable that has a negative influence on the price, but you should expect a positive

influence. There is a need to look more into the model to see how good this model is and if it can be used to tell what impact traffic noise have on the house price.

The Standard Error is a measure of how much the average value the coefficient estimates differ from the actual average value of the variable. The Std. Errors should be a low number in relation to the estimated coefficients. (Rego, 2015)
The Std. Errors are very high, and these should be as close to zero as possible. The standard errors of the coefficient are high for most of the variables, indicating that the model is not good, as the variables can vary a lot from the estimated coefficient.

The t-value needs to be far away from 0 to reject the null hypothesis and it will indicate that there is a relationship between the house price and the independent variable. (Rego, 2015) For the variable area, age, and toilets the t-value is high and acceptable, while the rest of the variables having a t-value close to 0.
The t-value is also being used to calculate the p-value, Pr(>t). The p-value is the probability of examining a value that is equal to or larger than t. If the p-value is small this is an indication that the relationship between the dependent variable and the independent variables is not due to chance. (Rego, 2015) The p-value is for some of the variables good as they are under 0.05, but some of them are also very high indicating that the variables are not statistically significant.

The residual standard error of the model is a measure of how good the linear regression fits. (Rego, 2015) The residual standard error is very high and shows that the observed sale price is far away from the predicted sales price. Also, the intercept is very high.
Another way to see how good the model fits the data is the Multiple R-squared and Adjusted R-squared. These will have a value between 0 and 1. If it is near 0 the model does not explain the variance in the variables good, and if it is near 1 the model explains the variance in the variables well. The adjusted R-squared is the one to prefer as the multiple R-squared increases when more variables are added, while the adjusted R-squared will adjust for the number of variables in the model. (Rego, 2015)
From the output of the summary of the linear model, the adjusted R-squared of 0.30 tells that around 30 % of the variation in the sale price of the house can be explained by this linear model. This is not very high, and this model is not precise. The reason it is that low could be the number of variables and including more variables could help.

The last thing to check is the F-statistic, which indicates if there is a relationship between the dependent variable and independent variables. Further away from 1 the better the relationship it is, but it also depends on the number of data points and variables. When the data set is big and having an F-statistic that is slightly larger than 1, means that it is already enough to reject the null hypothesis (H0: There is no relationship). When the data set is small a large F-statistic is needed to ensure that there is a relationship between the dependent variable and the independent variables. (Rego, 2015)

The F-statistic is far away from 1 but looking at the number of data points which is large it seems that the model does not describe the relationship between the house price and the variables that well.

In general, this model does not fit the data very well and the model needs to be improved to get a better answer to what influence traffic noise might have on the house prices. One way of improving this model is trying to transform the dependent variable (the house price), which is also know from the previous studies using the hedonic price model. Another way to improve the model could be including more variables or making transformations of the independent variables.

The next step in creating a better-fitted model is to log transform the sales price of the house, to see if this has any effect on the model. The result of this model can be seen in Table 4.

```
Call:
lm(formula = lprice2 ~ enh_arl + vaer_ant + anttoil + antbadev +
    isovalue + opfoerelse + rebuild, data = house)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6570 -0.1157  0.0330  0.1761  0.6868

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.8922038  0.6598188  25.601  < 2e-16 ***
enh_arl      0.0025394  0.0003171   8.007 2.94e-15 ***
vaer_ant     0.0166791  0.0088031   1.895   0.0584 .
anttoil      0.0576803  0.0182325   3.164   0.0016 **
antbadev     0.0273437  0.0192625   1.420   0.1560
isovalue    -0.0004169  0.0003230  -1.291   0.1970
opfoerelse  -0.0010465  0.0003395  -3.083   0.0021 **
rebuild     -0.0262629  0.0182572  -1.438   0.1506
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2747 on 1109 degrees of freedom
Multiple R-squared:  0.2374, Adjusted R-squared:  0.2326
F-statistic: 49.32 on 7 and 1109 DF,  p-value: < 2.2e-16
```

Table 4: Summary of the semi-log model.

With the same independent variables, but with a log transformation of the sale price R-squared gets lower and the model only explains around 23 % of the variation in the sales price. Also, the p-value of the variables gets worse.

Before going further with the modeling, we start to look at the variables that have been used in the two first models. The age and the rebuild are two variables that need to be considered, as the age is the year that the house was build and not the exact age of the house that is used. Rebuild is used as a dummy variable which is giving a wrong impact on the house price.
The age variable is being changed to indicate how old the house is and not what year it is built. The rebuild variables are removed as it is positive if a house is rebuilt which is good for old houses but

for new houses, the rebuild is having a negative influence as it is not rebuilt but a new house would be more worth than an old renovated house.

With the change in the variables the same two models, linear and semi-long models, are tested with the variables, area, rooms, toilets, bathrooms, age, and noise. Furthermore, are two other models tested, log model and inverse semi-log model as this is similar models that are tested in other studies. In the log model, both the dependent and the independent variables are being log-transformed. In the inverse semi-log model, all the independent variables are being transformed and the dependent variable, the house price, is in linear form.

The test of these four models gives similar results with the linear model and the inverse semi-log model as the best with an R-squared 0.30 and 0.31. The results of the 4 models can be seen in appendix D.

7.4 More variables

The next step to try to improve the model is to include some more variables to the model. The new variables that are added are variables that are describing the surroundings and accessibility. The nearest distance to the city Lyngby, the coast, the school, the forest, the lake, and the train station is measured and added to the hedonic model. The linear hedonic function is now:

$$Price = \beta_0 + \beta_1 area + \beta_2 rooms + \beta_3 toilets + \beta_4 bath + \beta_5 noise + \beta_6 age + \beta_7 city + \beta_8 coast + \beta_9 school + \beta_{10} forest + \beta_{11} lake + \beta_{12} train + e$$

The same four models, linear, log, semi-log, and inverse semi-log is tested, to see if more variables added to the model will improve the model.

For all the four models the adjusted R-squared gets higher having the linear and inverse semi-log model with the highest adjusted R-squared. The results of all four models can be seen in appendix E.

The result of the linear model (Table 5) shows a better model with an R2 at 0.35, indicating that adding more variables to the model will improve the model. The sign of the estimated coefficients is for most of them what was expected. The variable for forest and lakes have a negative sign which was expected to be positive, while the train station and age have a positive sign but were expected to have a negative influence on the house price. One thing that seems very wrong for this model is that the Intercept is negative, meaning that the mean price of the house when the variables are 0 are minus 1358682 kr.
The t- and p-value of the model is not that good for many of the variables, indicating that they do not fit well in the model. This could mean that some of the relationships between the house price and the independent variables are not linear and it could maybe help to transform the variables, which is done in the next model.

Linear Model:

```
> print(linearMod2)

Call:
lm(formula = price_2019 ~ area + rooms + toilets + bathrooms +
    noise + age + city + coast + school + forest + lakes + trainstation,
    data = house2)

Coefficients:
 (Intercept)          area         rooms       toilets     bathrooms
   -1358682         13507         54126        294067        144210
       noise           age          city         coast        school
       -4251          8689        362732        335410        215422
      forest         lakes  trainstation
       -7604       -548448        815220

> summary(linearMod2)

Call:
lm(formula = price_2019 ~ area + rooms + toilets + bathrooms +
    noise + age + city + coast + school + forest + lakes + trainstation,
    data = house2)

Residuals:
     Min       1Q   Median       3Q      Max
-5088309  -681608     -614   703312  6751653

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1358682    1745242  -0.779 0.436437
area            13507       1528   8.838  < 2e-16 ***
rooms           54126      41799   1.295 0.195617
toilets        294067      87607   3.357 0.000816 ***
bathrooms      144210      92102   1.566 0.117690
noise           -4251       1861  -2.285 0.022521 *
age              8689       1556   5.585 2.94e-08 ***
city           362732     177680   2.041 0.041440 *
coast          335410     243380   1.378 0.168442
school         215422     138737   1.553 0.120773
forest          -7604      87381  -0.087 0.930666
lakes         -548448     140156  -3.913 9.67e-05 ***
trainstation   815220     209343   3.894 0.000104 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1302000 on 1104 degrees of freedom
Multiple R-squared:  0.3592, Adjusted R-squared:  0.3522
F-statistic: 51.57 on 12 and 1104 DF,  p-value: < 2.2e-16
```

Table 5: Summary of the linear model.

In the inverse semi-log model (Table 6), all the independent variables have been log-transformed except the noise variable. This is giving a better model but having around the same R-squared as the linear model. This model also has several problems indicating that this model is not the best-fitted model that can be created. As the linear model, the p-value of the variables is high for half of the variables indicating that they do not fit in the model and there is no relationship between them and the house price.

Inverse Semi-Log model:

```
> print(Invers2)

Call:
lm(formula = price_2019 ~ larea + lrooms + ltoilets + lbath +
    noise + lage + lcity + lcoast + lschool + lforest + llake +
    ltrain, data = house2)

Coefficients:
(Intercept)        larea       lrooms     ltoilets        lbath        noise         lage
   -6462901      2250955       294216       410377       255181        -7560       434095
      lcity       lcoast      lschool      lforest        llake       ltrain
     -74544     -1200783       112106        80829      -287041       123586

> summary(Invers2)

Call:
lm(formula = price_2019 ~ larea + lrooms + ltoilets + lbath +
    noise + lage + lcity + lcoast + lschool + lforest + llake +
    ltrain, data = house2)

Residuals:
     Min       1Q   Median       3Q      Max
-5207054  -697711    21409   703360  8117491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6462901    1397624  -4.624 4.20e-06 ***
larea        2250955     240282   9.368  < 2e-16 ***
lrooms        294216     218520   1.346  0.17845
ltoilets      410377     149990   2.736  0.00632 **
lbath         255182     142595   1.790  0.07380 .
noise          -7560       1826  -4.141 3.72e-05 ***
lage          434095      58222   7.456 1.80e-13 ***
lcity         -74544     206775  -0.361  0.71854
lcoast      -1200783     584735  -2.054  0.04025 *
lschool       112106     100473   1.116  0.26476
lforest        80829      52472   1.540  0.12374
llake        -287041      62933  -4.561 5.66e-06 ***
ltrain        123586     108448   1.140  0.25470
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1300000 on 1104 degrees of freedom
Multiple R-squared:  0.3612, Adjusted R-squared:  0.3542
F-statistic: 52.01 on 12 and 1104 DF,  p-value: < 2.2e-16
```

Table 6: Summary of the inverse semi-log model.

7.5 OLS and GWR

As a result of the previous models that are tested the linear model, seems to be the one that is the best, and adding more variables to the model is having a positive impact on the model performance. The previously tested model can still get better. Two new methods to create a better model and to

get more knowledge about the data have been tested. The two regression methods that are tested are Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR) in ArcGIS.

The GWR is often a good try to get a regression model better as it makes regression for smaller areas and not for the whole area at once. (Esri, 2013a) The first model to test using GWR has the house price as the dependent variables and the same variables from the linear model above as the independent variables. The model could not be created as there either were severe global or severe local multicollinearity.
To fix this problem and checking for global multicollinearity an OLS model can be created. The variables that are having large Variance Inflation Factor (VIF), above 7.5 are redundant and should be removed one by one if more than one.

Two models using OLS were tested to check for the errors from the GWR model. One model with all the variables was made to see which of the variables that had a VIF over 7.5 and a model where the variables having a VIF over 7.5 were removed from the model. Furthermore, other models were also tested. For each model, a spatial autocorrelation on the residuals was made to see if the residuals were clustered or random. For all the models the residuals were clustered which could indicate that some key variables are missing from the model.

Before going further into the analysis and the results of the models, the dataset got updated with some extra variables as it seems that the model gets better with more variables. The variables that got add were lot size, area of the garage, carport, and annex, and the distance to business/industry area was added.

With the updated dataset a model using OLS was made with the house price as the dependent variable and the following independent variables: area, number of rooms, number of toilets, number of bathrooms, age, size of the garage, size of the carport, size of the annex, lot size, noise, distance to city (Lyngby), distance to industry/business area, distance to coast, distance to school, distance to forest, distance to lake and distance to train station.

The new function of the hedonic price model is:

$$price = \beta_0 + \beta_1 \cdot area + \beta_2 \cdot rooms + \beta_3 \cdot toilets + \beta_4 \cdot bathrooms + \beta_5 \cdot age + \beta_6 \cdot garage + \beta_7 \cdot carport + \beta_8 \cdot annex + \beta_9 \cdot lot\ size + \beta_{10} \cdot noise + \beta_{11} \cdot city + \beta_{12} \cdot industry + \beta_{13} \cdot coast + \beta_{14} \cdot school + \beta_{15} \cdot forest + \beta_{16} \cdot lake + \beta_{17} \cdot train + \epsilon$$

When creating an OLS model in ArcGIS it can generate an output report that contains the summary of the OLS results and additional graphics that can help understand the model. (Esri, 2013c) Looking at the overall results in the OLS report of model 1 (see appendix F) indicates that the model is not good and that there are several errors in the model that needs to be checked and change to get a better model. In the following the 6 checks, an OLS model should pass to be a good model and which problems there can be with the model is covered.

7.5.1 The checks of the output report: (the six checks)

The following are the six things that need to be checked in the model to see how good the model is.

1. Are the independent variables helping the model?
2. Are the relationships between the dependent variable and independent variables as expected?
3. Are some of the independent variables redundant?
4. Is the model biased?
5. Does the model contain all the key variables?
6. How good is the model explain the dependent variable?

(Esri, 2013b)

In the following section, each of these six checks is covered to see how good the model is and which problems there might be with the model. In Table 7 is the summary of model 1:

Summary of OLS Results - Model Variables

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	5630156,4848	887405,37862	6,344515	0,000000*	1002412,9698	5,616604	0,000000*	--------
AREA	10595,557279	1635,241350	6,479507	0,000000*	1849,058157	5,730246	0,000000*	3,384172
ROOMS	42803,050399	41300,373227	1,036384	0,300242	43284,274246	0,988882	0,322927	2,246329
TOILETS	309596,51926	88061,235753	3,515696	0,000471*	85250,442891	3,631612	0,000308*	2,108009
BATH	118984,66266	91449,622332	1,301095	0,193508	95410,978144	1,247075	0,212642	1,899883
AGE	7803,710980	1593,805638	4,896275	0,000002*	1893,376105	4,121585	0,000047*	1,360604
GARAGE	-7000,148072	6205,851928	-1,127991	0,259567	4507,336215	-1,553057	0,120714	1,170767
CARPORT	9552,991451	7005,159900	1,363708	0,172952	11778,754401	0,811036	0,417509	1,190560
UDHUS	-26323,43925	13044,623318	-2,017953	0,043831*	15099,042623	-1,743385	0,081555	1,167306
GRUND_ARL_	659,720599	168,860818	3,906890	0,000108*	245,496096	2,687296	0,007310*	2,091815
NOISE	-5404,455201	1948,364443	-2,773842	0,005635*	2012,437109	-2,685528	0,007348*	1,759199
CITY	1123645,6800	282221,08432	3,981438	0,000081*	323092,84541	3,477780	0,000541*	41,080882
ERHVERV	116351,14387	188184,84986	0,618281	0,536521	189431,65441	0,614212	0,539206	1,959092
KYST	-798340,8606	131379,18423	-6,076616	0,000000*	145523,56443	-5,485990	0,000000*	10,614279
SKOLE	275306,69665	156820,96859	1,755548	0,079452	159627,41759	1,724683	0,084876	1,819458
SKOV	-114875,1267	88442,724905	-1,298865	0,194272	88620,328765	-1,296262	0,195167	2,058843
SOER	250756,13288	195406,92871	1,283251	0,199683	220699,63435	1,136187	0,256124	2,491021
STOG_ST	-1502274,447	305832,18826	-4,912087	0,000002*	358879,47976	-4,186014	0,000036*	61,819477

Table 7: Summary of OLS results of model 1

First check

The first thing to check is statistically significant, by checking the probability (p-value) of the variables and looking at the estimated coefficient.

If the estimated coefficient is very small, near zero, these variables are not helping the model. (Esri, 2013b) For this model, all the variables are far away from zero showing that they are helping the model.

Looking at the t-value of the variables some of these are not that high and some of them are close to 0 indicating that there is no relationship between the variable and the house price. We want the t-value to be as far away from 0, as the null hypothesis can be regretted and say there is a relationship between the variable and the house price. The t-value is being used to calculate the p-value.
The p-value can tell something about the statistically significant. If the probabilities have an asterisk with them, they are statistically significant, which is important for the model. (Esri, 2013b) Looking at the p-value for this model around half of the variables have an asterisk next to them and acceptable p-value (probability). Some of the variables have a high p-value that could indicate that it is not useful in the model and there is a need to work more with the variables.

Also, in the first check, we can look for nonstationary by looking at the Koenker Statistic. In the OLS report, the overall model significance is measured using Joint F-statistic and Joint Wald Statistic. The Joint F-statistic can only be trusted if the Koenker (BP) Statistic is not statistically significant. If the result from the Koenker (BP) Statistic is significant the Joint Wald Statistic should be used to see if the overall model is significant and the robust probabilities should be used to check the coefficients of the variables are significant or not. (Esri, 2013c) The results of these tests can be seen in Figure 10.

Joint F-Statistic [e]:	40,037759	Prob(>F), (17,1093) degrees of freedom:	0,000000*
Joint Wald Statistic [e]:	436,799684	Prob(>chi-squared), (17) degrees of freedom:	0,000000*
Koenker (BP) Statistic [f]:	164,639008	Prob(>chi-squared), (17) degrees of freedom:	0,000000*

Figure 10: The three statistics test, Joint F-Statistic, Joint Wald Statistic, and Koenker Statistic.

For this model, the three presented Statistic test is all having a value of 0 and an asterisk which shows the overall is significant and as Koenker (BP) Statistic is significant the relationships that are modeled are not consistent due to either non-stationarity or heteroskedasticity.
If there are problems with non-stationarity it means that the relationship that is tried to be model is changing across the study area and if the relationships vary in relation to the size of the variables that are tried to be predicted there is a problem with heteroskedasticity. (Esri, 2013c)
As the Koenker test is statistically significant the robust probabilities are the ones that can be trusted and to see if the variable is helping the model or not. Only eight variables are statistically significant and there is a need to look further into the variables.

Second check
The next check is to look at the sign of the coefficients of the variables to see if it is as expected. The OLS generates the estimate of the variable's coefficients. For these variables we have some expectation of what relationship it would have with the house price and if they would have a negative or a positive influence on the house price. In Table 2 the variables are listed with the

expected sign. Some of the variable's coefficients have a different sign that was expected like age is positive which was expected to be negative as the old a house would have a negative impact of the house price.

Third check

The third check is to see if the variables are redundant by looking at the VIF. The VIF is measuring the redundancy of the variables. These value needs to be under 7.5 and if it is over the variables should be removed. (Esri, 2013b)

Three variables, train, coast, and city is having a VIF over 7.5. As the VIF is over 7.5 indicates that these variables are telling the same side of the story and therefore need to be removed one by one.

Fourth check

The next thing to check is to see if the model is biased. If the model is biased the distribution of residuals is unbalanced. A statistically significant Jarque-Bera (Figure 11) diagnostic indicates your model is biased and shows if the residuals are normally distributed or not. If the p-value of this test is small, it means that the residuals are not normally distributed, and the model can be biased. (Esri, 2013b)

Jarque-Bera Statistic [g]:	274,916667	Prob(>chi-squared), (2) degrees of freedom:	0,000000*

Figure 11: Jarque-Bera Statistic

For this model, it is zero showing that the residuals are not normally distributed. Checking the residuals with spatial autocorrelation, also shows that the residuals are clustered, meaning that the model is missing a key variable (see the fifth check) and therefore the model is biased.

With the Jarque-Bera test results showing that the model does have a linear relationship, but as the model is biased it can be a result of a nonlinear relationship or there might be some outliers that have a big influence on the model. This can be checked by looking at histograms and scatterplots and try to transform some of the variables to create a more linear relationship.

The scatterplot (Figure 12) of some of the variables (see appendix H for the rest of the variables) shows that some of the variables have a linear relationship with the house price, but also that some of them do not have a linear relationship.

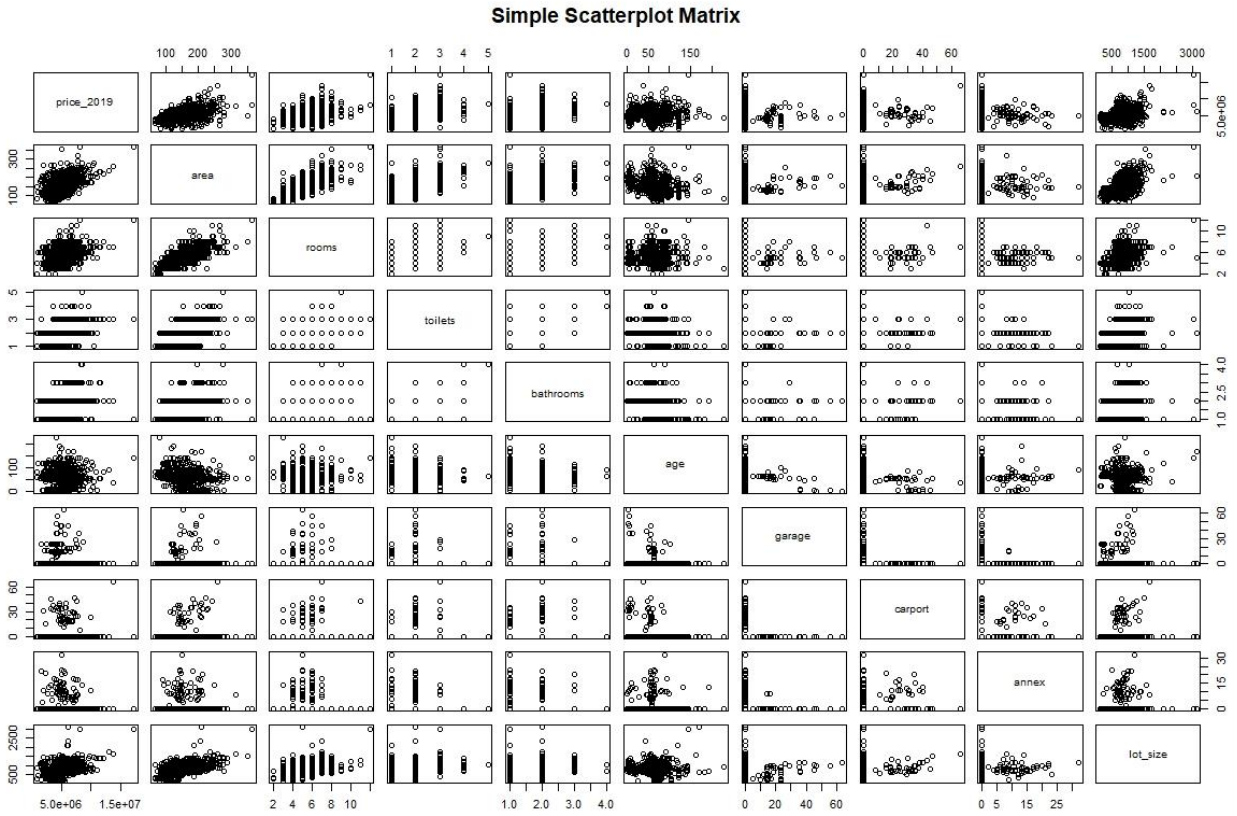


Figure 12: Scatterplot Matrix for some of the variables.

Fifth check

The fifth check is to see if all the key variables are found. To see if some variables might be missing from the model is to look for statistically significant spatial autocorrelation of the model's residuals. If the residuals are clustered it indicates problems with spatially autocorrelated residuals. The Spatial Autocorrelation tool is used to see if the model's residuals are clustered or not. (Esri, 2013b) The result of the spatial autocorrelation (Figure 13) shows that the model has clustered residuals, meaning that some key variables are missing from the model.

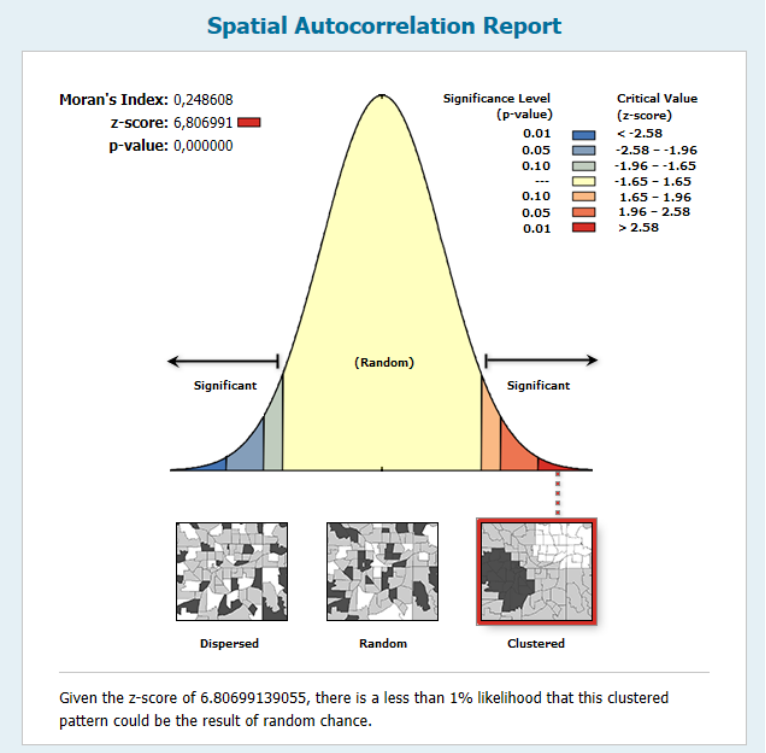


Figure 13: Spatial Autocorrelation report of model 1.

Sixth check

The last check of the model is to see how well the model is explaining the dependent variable. The multiple R-squared and Adjusted R-squared indicates how good the model performance is. (Esri, 2013b) For this model, the results are: Multiple R-Squared is 0,38 and Adjusted R-Squared 0,37. Meaning that the model can explain around 37 % of the variation in house price.

7.5.2 The solution to the problems

The first model shows that there are several different problems with the model, which needs to be fixed to see if this will give a better-fitted model.

From the six checks of the models the following problems for the model were found:

- Not all variables are statistically significant
- Koenker test is statistically significant, which means that the relationship that is being modeled is not consistent because of non-stationarity or heteroscedasticity.
- Some of the estimated coefficients have a different sign than expected
- Some variables are redundant
- The model is biased
- The residuals are clustered

To get a better model these problems need to be studied to see if they can be fixed and to see if it have any effect on the model or if there is a bigger problem connected to the model. Different tests will be done to find a solution to the listed problem of the model.

Redundant variables

The first problem to look at is the redundant variables. The variables that have a VIF higher than 7.5, is train, coast, and city. The train station has the highest with a VIF above 61.
These variables might be telling the same part of the story and needs to be removed one by one to see what effect it has on the VIF for the other variables.
The train variable is the first one to be removed as it has the highest VIF and being close to a train station often also dictates being close to the city.
By removing the train station from the model gives the model (OLS3) a bit lower Adjusted R2 at 0.36 (see appendix F for OLS report), but none of the variables have a VIF that is higher than 7.5. The coast variable is still high, near 7.5, and could be considered removed in the next step.

The coast variable is removed from the model to see what effect it has on the model (OLS4 – see appendix F for result). By removing the coast variable, it gives an acceptable VIF for all the variables. The variables that are left in the model are the variables that are used in the next models. The spatial autocorrelation of the residuals (Figure 14) still shows that these are clustered and removing these two variables did not have an overall effect on the model.

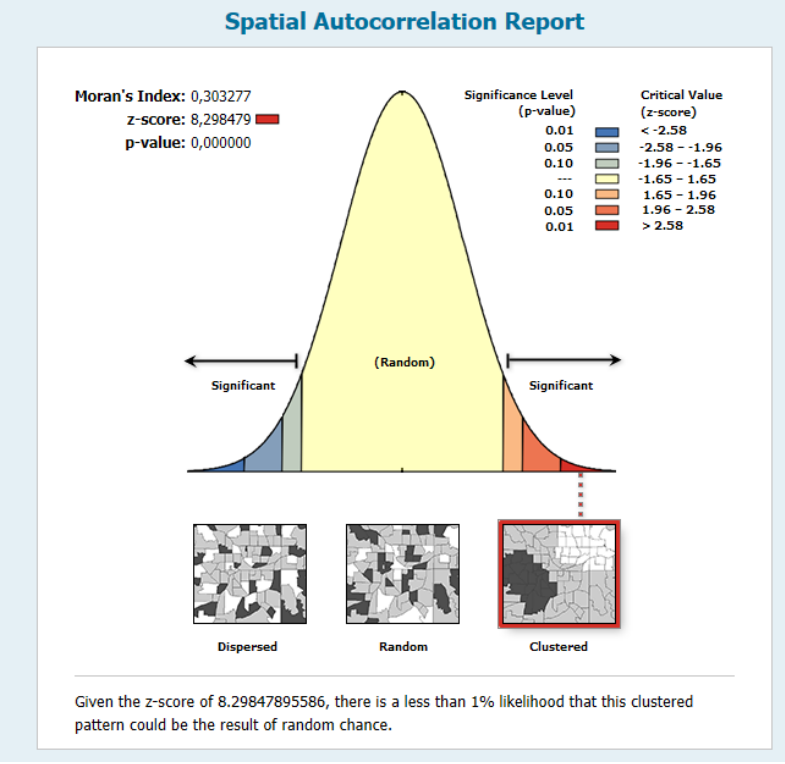


Figure 14: Spatial Autocorrelation report of model 4.

Model bias

The results also indicated that the model is biased. This can maybe be solved by looking at the distributions of the variables to see if they are skewed. If they are some of the variables have a skewed distribution a transformation of these variables could give a better model and eliminate the bias. The transformation of some of the variables would also help nonlinear relationships and that

eliminates the bias of the model. Outliers can also give a bias model and the variables should be checked to find outliers that could affect the model.
To solve this problem there is a need to study the variables for nonlinear relationships, outliers, and see if any of the variables should be transformed.

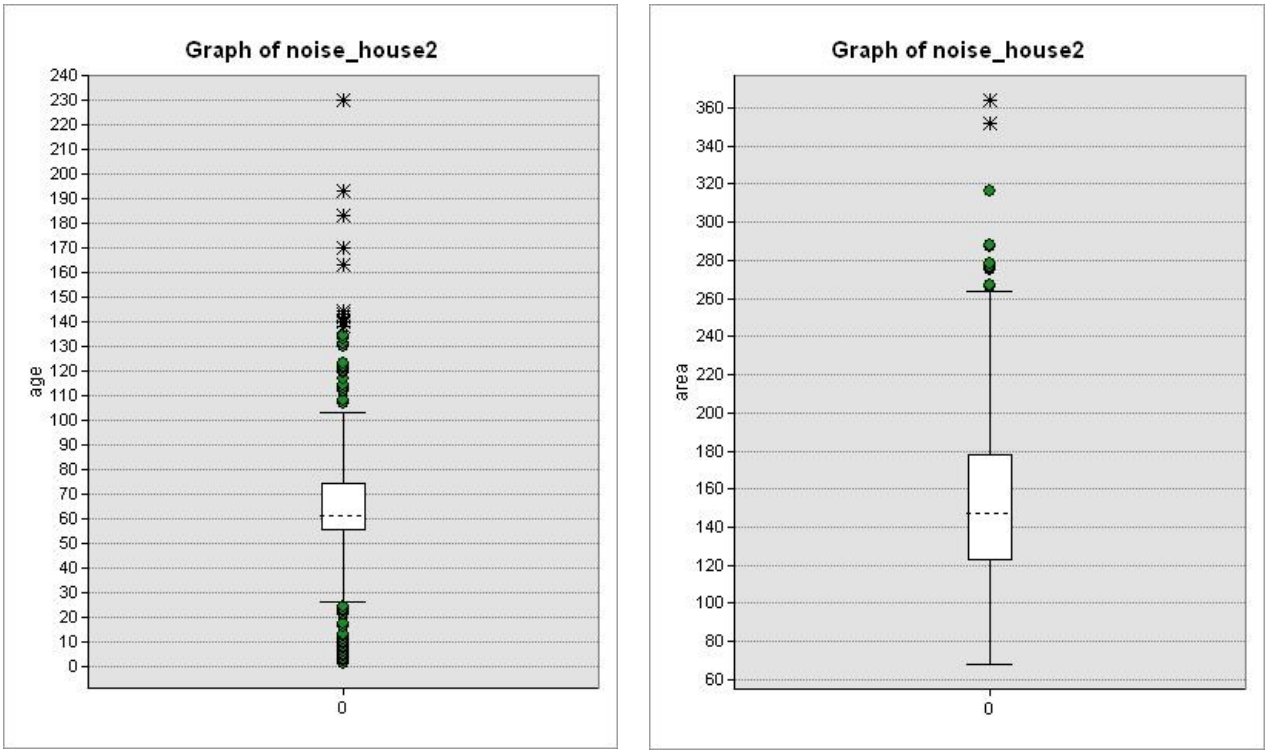


Figure 15: Boxplot of age of area to find outliers.

Boxplots (Figure 15) was used to find if there were any extreme outliers. There were only a few outliers for the variables area, age, and lot size. These outliers were removed from the dataset to see the effect of the outliers and if it would help the model to get better. A new model (OLS5) were tested. The removal of the outliers did not have any effect on the output of the model (see appendix F).

Histograms were used to see the distribution of the variables and to check if a transformation of the variables could help if the distribution were skewed. For the variables area, rooms, and forest the skewness of the distribution got better with a log transformation. In Figure 16 the histogram of the area with none transformation and with log transformation can be seen. See Appendix G for the histogram of rooms and forest.

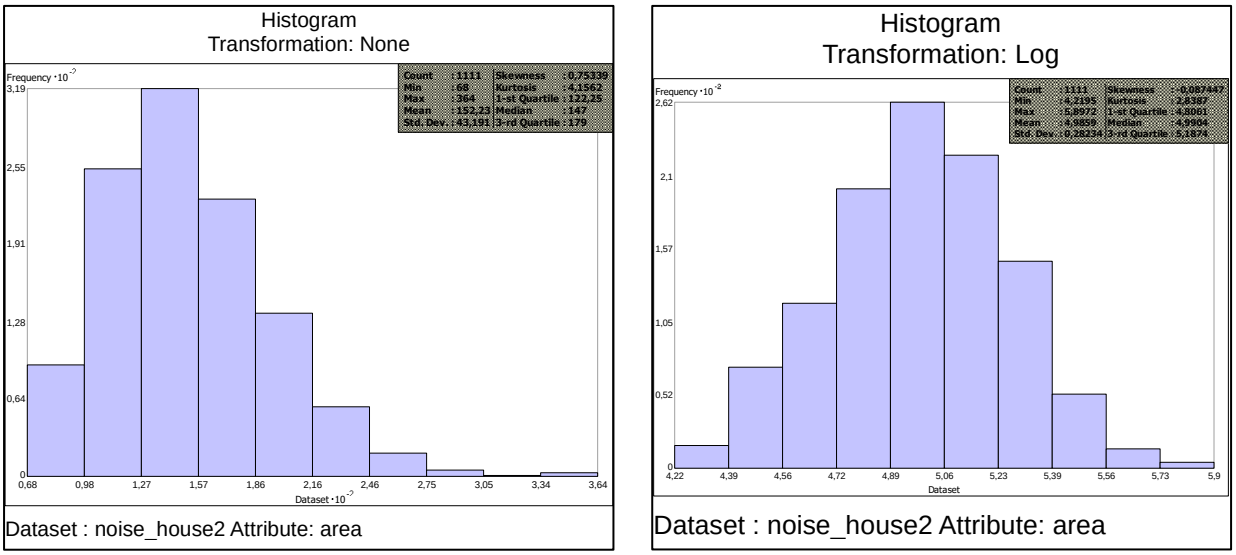


Figure 16: Histograms of area with none and log transformation.

A new model (OLS 10) was created with the transformed data and with the outliers as these did not have any effect on the model. The results of the model with the transformed variable got a bit better than the model before with R2 at 0.36 and more of the variables were statistically significant. This could indicate that not all of the variables have a linear relationship with the house price and needs to be transformed to better fit in the model. See appendix F for the output report of the model. The residuals of this model are still clustered (Figure 17), meaning that there is still some problem.

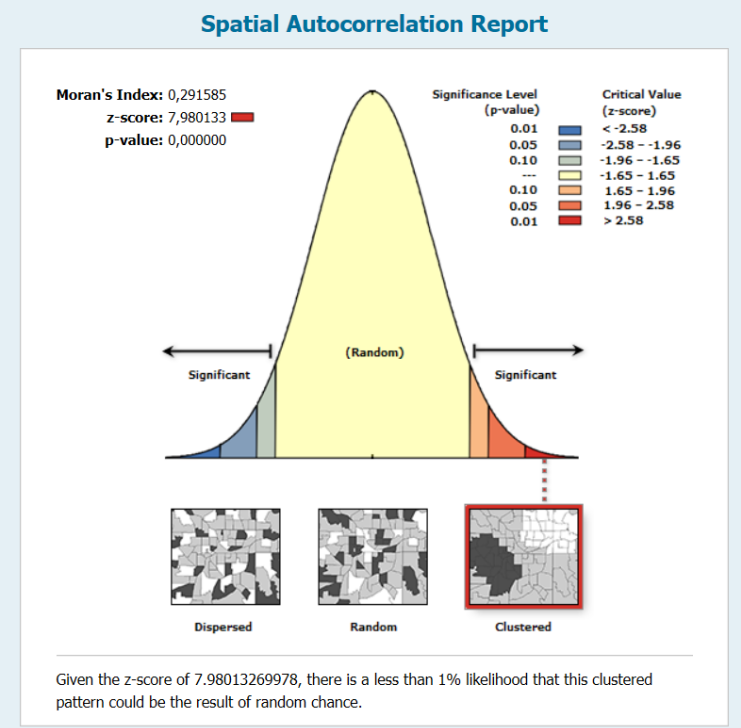


Figure 17: Spatial Autocorrelation report of model 10.

Variables not statistically significant

As these solutions to some of the problems could not help to get a better model there is a need to look more at the variables as many of them are not statistically significant. Try to test which of the variables that have an impact on the model and which of them could be removed. This could help to get a better-fitted model. From the model above with the transformed variables, the variables bathrooms, garage, carport, business area, rooms, and forest are not statistically significant.
The variables, bathrooms and rooms, are some of the variables that do not help the model as there are some variables, toilets and area, there tell the same thing about the house. In the next model (OLS 11) bathrooms and rooms are removed and see what effect this might have on the model. Removing some different variables that were not statistically significant did not have any impact on the model result. See appendix F for the result of the model 11.

Clustered residuals

One of the biggest problems with the model is that the residuals are clustered, and it seems that one or more key variables are missing in the model. A way to try to avoid the clustered residuals is to divide the area into smaller subsections and try to find an OLS model for each of these areas.
The areas were divided into areas covering the different cities in the study area. For the area of the city Kgs. Lyngby which is a small area gives a better result, where Jarque-Bera Statistic (Figure 18) is not statistically significant, and the model is not biased.

Jarque-Bera Statistic [g]:	3,690022	Prob(>chi-squared), (2) degrees of freedom:	0,158024

Figure 18: Jarque-Bera Statistic.

By dividing the area into smaller subsections the residuals become random for some of the areas (Figure 19), but the accuracy of the model is still not high with a Multiple R-Squared at 0.40 which is higher than other model but have a lower Adjusted R-Squared at 0.33.

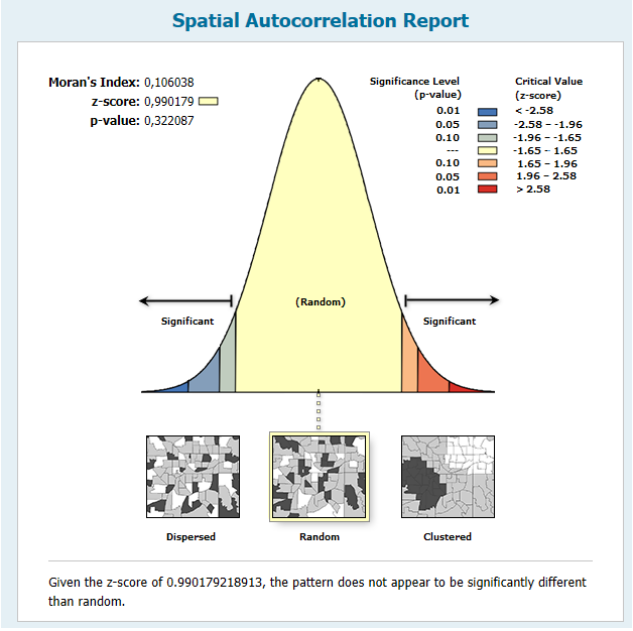


Figure 19: Spatial Autocorrelation report of subarea Kgs. Lyngby.

Key variables missing

Another big problem with all the models that have been tested is that some key variables are missing from the model. With the examine the model residuals through a Hot spot analysis (Figure 20), problems with clustered over- and underpredictions were found for all the models. There were found clusters of over- and under prediction in more places around the study area indicating that some key variable is missing (Esri, 2013c), which maybe could be a variable that was describing the different sub-areas in the study area.

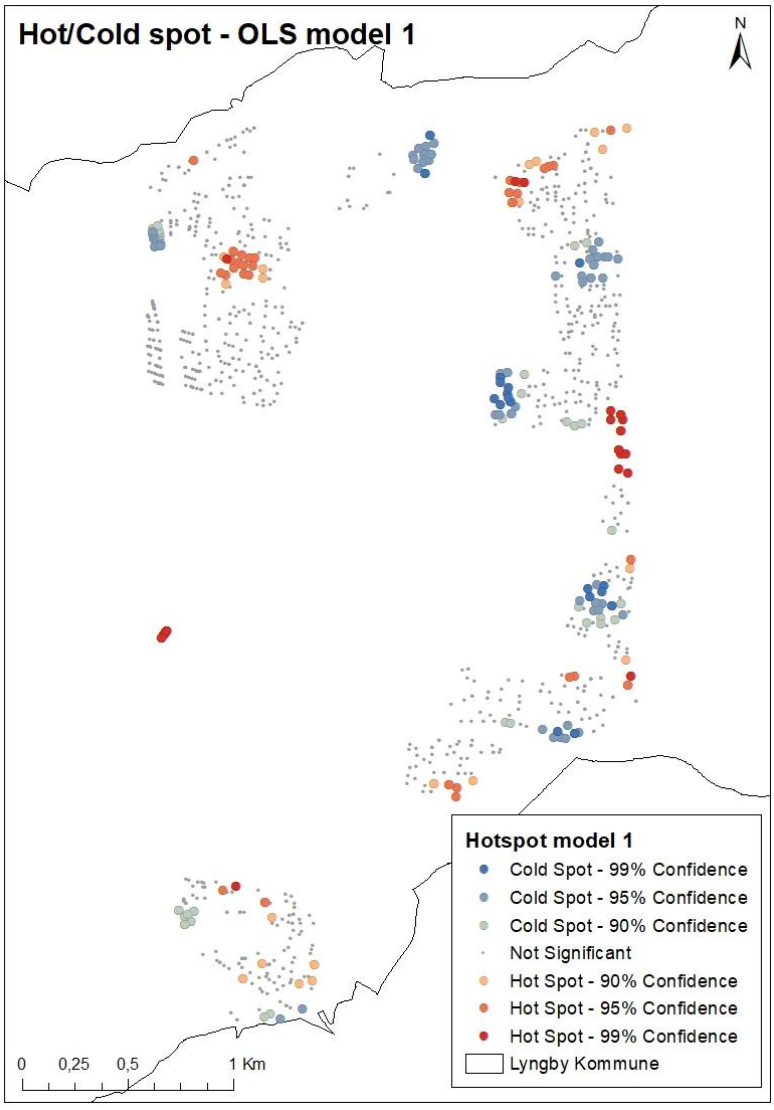


Figure 20: Hot/cold spot of the residuals of model 1.

Non-stationarity or heteroscedasticity

In all the models that were tested the Koenker test shows statistically significance, having the model being non-stationarity or heteroscedasticity. A way to deal with non-stationarity is to try and run the model using GWR. (Esri, 2013c) Trying to run the best model that was found with OLS gave errors when running the model with GWR. The model could not run because of problems with either

global or local multicollinearity, meaning that some of the variables have a large VIF, but these are already removed from the model.
In the output report of the models, a graph shows if there is some problem with heteroskedasticity. The graph (Figure 21) shows the relationship between the residuals and the predicted value. This scatterplot should look random and have a little structure. This scatterplot is clustered and as the shape of the graph is a bit cone shape it indicates that the model is performing differently depending on the size of the estimated value. (Esri, 2013c)

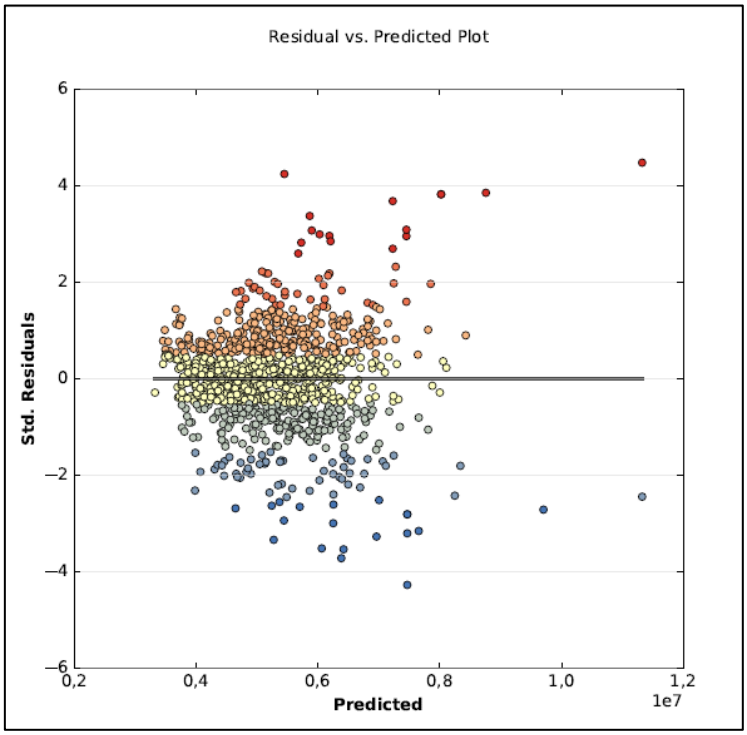


Figure 21: Graph showing the relationship between the residuals and the predicted value.

As it is difficult to find a better-fitted model with the use of OLS another method is tried. Machine Learning is tested to see if this method can find a better-fitted model.

7.6 Machine Learning

As it is complex to create a house price model as many factors play a role in the model and many things that have an influence on the house price in general.
The variables that are used for the house price model do not seem to explain the house price well enough. Some problems with the data could be that they are not precise enough or there is some misleading in the data set from BBR. As it is different places in the municipality that is being studied some of these areas can be more popular than others and some houses are more popular, and they will be sold faster and to a higher price than others. These are some of the things that make the house price model complex to model.

Machine Learning is a good tool to use as of the complex house prices. Machine Learning could make a better-fitted model. For this, the program Weka is used where different classifiers are being tested. Random Forest and Linear Regression are the two models that had the highest R2 with RF as the highest.
These are the two models that will be used in further work and to see if the models can get better by tuning the models by changing some of the model's parameters. In the Weka, four different test options can be changed. Cross-validation and Percentage split are the two that are used to tune the models. Different tests where cross-validation with different numbers of folds was tested and different percentages of the Percentage split were tested. With these tests, the model performance got improved a lot for some of the models. From the different test the following 4 models had the best results:

- Linear regression with cross-validation 10 folds with an R2 at 0.5879
- Linear regression with a percentage split at 80 % with an R2 at 0.7331
- Random Forest with cross-validation 10 folds with an R2 at 0.6111
- Random Forest with a percentage split at 70 % with an R2 at 0.6690

Another step to try to create a better-fitted model was to test these four models on four subsets of the project area. The four subareas that the areas got divided into were the four city areas in the study area, Fortunen, Hjortekær, Kgs. Lyngby and Lundtofte. The test on the four areas did not improve the model. The results of the four models and the four areas can be seen in Table 8.

	RF 10 folds	RF 70 %	LR 10 folds	LR 80 %
Fortunen	0.3496	0.3319	0.3864	0.3563
Hjortekær	0.6170	0.4434	0.5224	0.5047
Kgs. Lyngby	0.5103	0.5459	0.4883	0.4679
Lundtofte	0.5366	0.5695	0.5020	0.5066

Table 8: The results of the RF and LR model on the four subareas.

The best-fitted model that was found was the Linear regression model with an R2 at 0.73. The output of the linear regression is the following equation:

$$Price = 11584.24 \cdot area + 331191.09 \cdot toilets + 109221.76 \cdot bathrooms + 7846.49 \cdot age + (-19381.20) \cdot annex + 658.25 \cdot lot\ size + (-5606.92) \cdot noise + 1104219.45 \cdot city + (-945759.87) \cdot coast + (-129858.18) * local\ train + 307288.15 \cdot school + (-109154.59) \cdot forest + 274611.81 \cdot lakes + (-155766.74) \cdot train\ station + 6654060.51$$

Only 13 of the 19 variables were used in the model. In the next chapter, the model will be described in more detail and what this model can tell about how big an impact noise is having or not.

8 The house price model

In this chapter, the results of the analysis are presented. Through the project, many different methods have been used to find the best-fitted model that could describe the house price and which impact noise might have on house prices.
The best model that was found and some of the problems will be described. A further discussion of the model and the issues that have been can be found in the next chapter.

In the work of finding the best-fitted model different methods and models have been tested and more variables were adding though the work with the models. In Figure 22 the different model and the improvement in is showed:

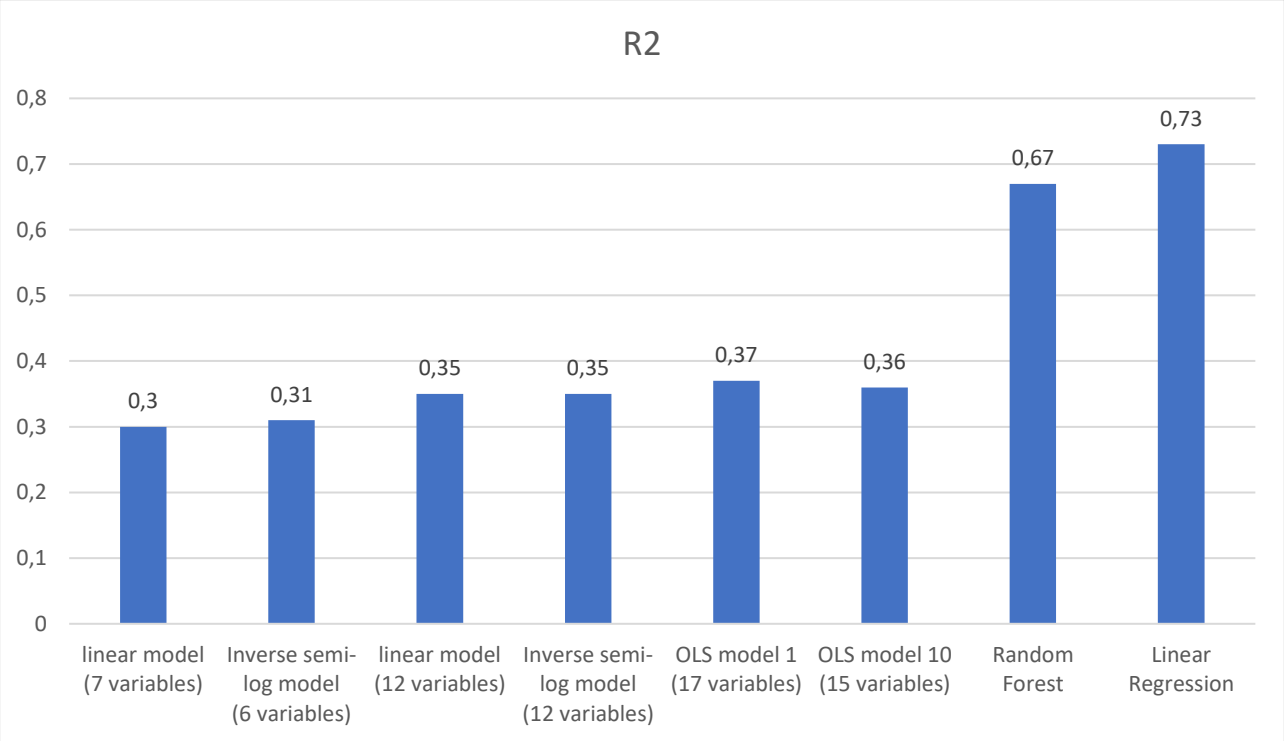


Figure 22: The improvement of the model through the project.

The best improvement of the model where when more variables got added and when the method got changed to machine learning,

The best-fitted model was found with the use of Machine learning using Linear Regression as the classifier. This model had the highest R-squared at 0.73 which is much better than all of the other models that have been tested. This model is not perfect as we wish to have a higher R-squared around 0.90. A lot of other factors is playing a role in the house price model which has an impact on house prices but can be difficult to include in the model.

From the output of the Linear Regression model follows this equation, with the chosen variables and the estimated coefficients:

$$Price = 11584.24 \cdot area + 331191.09 \cdot toilets + 109221.76 \cdot bathrooms + 7846.49 \cdot age + (-19381.20) \cdot annex + 658.25 \cdot lot\ size + (-5606.92) \cdot noise + 1104219.45 \cdot city + (-945759.87) \cdot coast + (-129858.18) \cdot local\ train + 307288.15 \cdot school + (-109154.59) \cdot forest + 274611.81 \cdot lakes + (-155766.74) \cdot train\ station + 6654060.51$$

The data set that was used to create this equation content 19 different variables and only 13 of the variables were used to create the model. Compared to similar studies the number of variables in this model is relatively low and different variables are used. From the literature review of other studies, it was shown that some of the most important variables to have in the model were the accessibility variables which are also the variables that are used in the final model.

The main goal of the project was to see which method there best could describe the relationship between several variables and discover the correlation between them and house prices and from this to see which role noise played and how big an impact noise would have on house prices.
The estimated coefficient of the noise variable (-5606.92) is relatively low compared to some of the other variable's coefficient. From the result of the model, it seems that noise does not play a significant role in the house price. But the results of the noise impact can be difficult to trust as the performance of the model is not perfect.
Compared to similar studies they have shown that noise in some places can have a positive impact as seen in studies of Blanco and Flindell (2011), which also showed that the noise had a negative or non-effect on different areas. The general results of the different studies show a decrease in the house price between 0.3 % to 1.3 % per dB, with much of them under 1% decrease in house price per dB. From all these studies the main results are that noise, in general, has a negative impact on house prices, but the impact is relatively low and many different factors play a role in the modeling of the house price giving all these different results. (Miljøstyrelsen, 2003)

9 Discussion

In this chapter a discussion of the work that has been done will follow. The results of the analysis of finding the best model are discussed and which issues there have been and what can have caused these problems.

9.1 Over/under prediction and clustered residuals

One of the main problems with the models was that the hot and cold spots were clustered indicating that the model made over- and underprediction in the study area.
It can be difficult to tell what the real reason for this problem could be, but it might indicate that there is some irregularity in the house prices. There would be a need to look more into the house prices in these hot and cold areas to find if there would be any connection to why there would be an over- or underprediction of the price.
House prices are a complex thing to model and many factors play a role when a house is sold. It is not only the structural and surroundings that have an impact on the house price. The neighborhood can also be an important factor in the variation of the house price. Some areas of a city will be more popular than others and some areas will have a bad rumor. The social profile of an area is an important factor to have in a model because this will also attract different groups of people.
Some houses will also be more popular than others and would be sold faster and properly sold to a higher price than the house nearby but is not an attractive house for a buyer. All these things can play a role in the house price and it can be difficult to see which of the houses that are more popular and can be sold to a higher price.

All these things can be related to the problem with the clustered residuals which could indicate missing some key variables but also that there would be some outliers in the house price data. These outliers would not be in terms of being noisy but instead of that, some houses will be more attractive and unique. They will be oversold or be sold quicker than other similar houses.
The houses can maybe be similar but one of the houses could have some issues which could scare buyers away and to get it sold it needs to be sold for a cheaper price. These types of outliers give issues with modeling a better model, than the best model that was found.

9.2 More variables

The issues of finding a better model depended on the data set and which variables were included in the model. The variables mainly focused on the structure of the house and its surrounding, which is not a complete enough data set to cover all the factors that have an impact on the house price. Many other factors play a role in the modeling of a house price which is not included. As seen in other studies some also included socio-economic variables that give an indication of the area and which kind of people there live there. The area environment and popularity also attract different kinds of people and what they are looking for in a house. The human part of a house sale related to

the seller and buyer of the house is important to think about but is something that is difficult to include in the model.

Another issue with finding a better-fitted model is if the data that is used can be trusted. There can be some errors in the BBR data and things that are not registered in the system. Some houses can contain illegal annexes, extra rooms, etc. which are not in the data, but is playing a major role in the house price. As it is the owner of the house and the municipality that needs to check if the BBR is correct the data from BBR can contain many errors. If changes of the house are not reported to the municipality the BBR is not complete enough giving that some houses maybe contain more rooms than is registered or the house seems better than it is. All this gives issues in not be able to find a better model.

9.3 Machine learning

In the work of finding the best model different methods were tested. Four different models, linear, log-linear, semi-log, and inverse semi-log were tested, which were the models with the lowest R2. The next methods that were used were OLS and GWR in ArcGIS which improve the model a bit but the model was not good enough. As the house is complexed to model machine learning was tried. This resulted in a lot better model with an R2 at 0.73.
Machine learning for this project was the best but compared to other studies this method is quite different in the approach of creating a price model. From other studies, the semi-log and log model was the one that often was the best.
As mentioned before the data and the variables that are in the model are not complete enough to cover all aspects of the house price the machine learning seems to be the one that could create the best possible model with this data. But is this model a good or a bad model? The linear regression model from machine learning has an R2 at 0.73, which indicates that there still is something to make better. This model cannot be trusted fully but is giving some indication of the variable connection and relationship with the house price and how big of a role the noise is playing.

9.4 Noise variable

The noise variable is an important part of this analysis and is the variable that is focused on to find how big an impact the noise might have on the house price. But how is the noise variable best included in the model? As seen in other literature it is quite different how the noise is used as a variable.
Theebe (2004) is using a spatial specification with noise dummies as he also includes the positive effects of the infrastructure by including accessibility variables. The study by Andersen (2015) is not only using the equivalent level of noise but also having the maximum level of noise to see the effect on the house price. A third way noise is being included in the house price model is a study by the Danish Environmental Protection Agency (2003) where it is not the noise that is used but the distance to the road and if the house is in the front row to the road that indicates the noise.

In this project, the noise has been included as the daily average of the noise in the area with an interval of 5 dB starting at 55 dB. For the houses that were not placed in an area with measured noise, the noise for this was set to be 0 dB to have the biggest difference between noisy and quiet areas. This cut-off point for these areas does not seem right as there would always be some kind of noise. A test of these cut-off points for the lowest noise could be changed to different levels under 55 dB to see if this would give a change in the noise impact on the house price.
As mentioned, the noise variable can be used and expressed in different ways. An example of this could be the distance to the city, as being closer to the city also will give more traffic noise because of more people and cars. Having both the distance to the city and the noise as variables could cause some problems as these variables would be correlated. If removing the city variable, the noise could both explain the noise but also the distance to the city. From the study of Rich and Nielsen (2004), they conclude that it seems that: *"…the ability to describe accessibility has a greater impact on the description of traffic noise, e.g. to obtain a reliable estimate of noise effects also accessibility must be properly described."* (Rich & Nielsen, 2004)
The use of the different variables and what the different variables should describe plays an important role in the result of the model and which relationships there are between the variables and the house price.

10 Conclusion

In this chapter, the conclusion of the 3 research questions is presented which will answer the project problem statement. The research questions will be included in the text as they will be answered.

To what degree house prices can be modeled and which method can be used to explore the relationship between different variables and house prices?
The focus of the project was to find a model that could explain the relationship between different variables and house prices and see if traffic noise has any negative effect on house prices. The house price can be explained by several characteristics that are related to the house like house quality, house structure, environmental characteristics, and surroundings. All these things will have an impact on the house price some will have a positive impact, and some will have a negative impact.
To explain the relationship between the variables and the house price the hedonic price method was used to create a house price model. This model describes the house price trough different variables that are connected to the house structure, location, environmental characteristics, etc. were used.
The best-fitted model that was found in the project was a model with an accuracy of 73%. The method that was used to find this model from the data and the selected variables where machine learning with Linear Regression as the classifier. Linear Regression was the method that best could explain the relationship between the different variables and the house prices.
To get this model different methods have been tested through the project. Different linear models got tested but these models all had a low performance with an R2 around 0.25-0.35. Also, OLS and GWR were tested.

How can machine learning be used to create the house price model compared to standard house price models?
Compared to the other more standard models that were tested in the project machine learning difference in the approach and method of finding the best house price model. Machine learning seemed to be the best method because of the complexity of modeling house prices which the other standard models could not handle and cover. Because of the machine learning iteration process and splitting the data into training and test data, the complexity of house prices could better be explained through this method.

Does traffic noise have an influence on property values? if so, how big an impact does traffic noise how on the house's prices?
With a model with 73% accuracy, this model can be trusted and used for checking if noise how any influence on the house prices. The model shows that the traffic noise does not have any significant impact on the house prices compared to some of the other variables that were included in the model. That the noise does not have any significant influence on the house price is what some other studies also have found. As mentioned in the result section 8 different other studies show different results of the noise impact on house prices. Most of these studies show that noise has a negative impact, but some also showed that in some areas the noise had a positive impact on the house price.

11 Future work

In this section different aspect of how this analysis can be developed and what other factors that could be included in the house price model to improve the model will be discussed. Furthermore, how this type of analysis and models can be used in other content like in the municipality will be discussed.

11.1 House price model on apartments

This analysis only focused on single-family houses and terraced houses, which mainly are houses that are placed outside the bigger cities. These types of houses often also attract some groups of people. To get another perspective on the house price model and see if the output of the noise impact could get different. This analysis could be done in a more central city center and focus on the apartments as this would be a different group of people and these might have a different meaning on the noise.

The noise in the city will be more constant because of more people living closer and more cars. The people living in the city might be a different group of people than those who are living outside the city in more quiet areas. These people would be more acceptable for the noise and might be willing to pay more for noise to be able to live in the city.
For the people living outside the city and have a garden to there house the noise would have a more negative impact on them than the people that are living in the city.
The results of this type of analysis could maybe show that the noise impact on the house price would be positive because of all the positive effects it can be to live in the city like close to work, recreational areas, shopping, etc.
These kinds of aspects of how people are seeing noise and what they can accept also play a major role in how the output of these house price model will be and what kind of influence noise would have on a house price.

11.2 Real estate

The house price model had different variables, but these variables did not cover all the aspects of the house price. The house price is a difficult thing to handle and there is a need to have some knowledge about house prices is created.
A way to give a better insight into how the house price is constructed could be to involve a real estate agent in terms of how they value a house, and which values it is based on. By including this in the analysis could help to clarify the importance of the different variables or which variables that might be necessary to include in the model.
This will also give an idea of how much house prices are changing over time and which areas there is more popular. Furthermore, it could give some information about what buyers are seeing as positive things of a house and the surroundings compared to what buyers see as negative things there can cause a discount in the house price.

11.3 Work in municipality

For this kind of analysis how can this be used in another way like in the planning in the municipality and what can they use it for. A way this type of analysis can play a role in the municipality is when planning new roads or railways that might need to go through properties that are established and have been there for a long time. With the use of this type of analysis, it could look at what consequence both economic but also health-related it will have on the properties and the owners. When establishing a road or a railway noise is not the only negative consequence that comes from the traffic also air pollution and the change in nature and surroundings.
To get a better picture of what consequence a new road will have on house prices more variable describing the effect of a new road needs to be included. Another way is to create a house price model an run it on the areas before a road is established and run the same model a few years after the road was established to see what difference that has been in the house price and the effect of being close to a road.

The work with traffic and noise in the municipality can also be seen in relation to the Sustainable Development Goals as there is a focus on the infrastructure and health. One of the goals (goal 3) is to ensure a healthy life for all, where one of the sub-goals is that the number of deaths and diseases caused by among other things air pollution should be reduced. When planning new roads this should be in focus that the major road with high traffic should not be near cities to avoid air pollution is too high in the cities. Another goal (goal 11) is the focus on making cities safe, resilient, and sustainable. The traffic and infrastructure also play a big role here. Here is there also a focus on air pollution and that cities should be safe and finding a good sustainable solution for urban planning. (Udenrigsministeriet, 2017)

12 Bibliography

Andersson, H., Swärdh, J.-E. & Ögren, M., 2015. *Traffic noise effects on property prices: Hedonic estimates based on multiple noise indicatores,* Stockholm: Centre for Transport Studies.

Blanco, J. C. & Flindell, I. H., 2011. Property Prices in Urban Areas Affected by Road Traffic Noise. *Applied Acoustics,* Volume 72(Issue 4), pp. 133-141.

Boligejer, 2020. *SVUR - Statens Salgs- og Vurderingsregister.* [Online]
Available at: https://boligejer.dk/statens-salgs-og-vurderingsregister
[Accessed April 2020].

Bygnings- og Boligregistret, 2007a. *2.2.1 Ejendomsniveauet.* [Online]
Available at: https://instruks.bbr.dk/ejendomsniveauet/0/30
[Accessed April 2020].

Bygnings- og Boligregistret, 2007b. *2.2.2 Bygningsniveauet.* [Online]
Available at: https://instruks.bbr.dk/bygningsniveauet/0/30
[Accessed April 2020].

Bygnings- og Boligregistret, 2007c. *2.2.3 Niveau for bolig- eller erhvervsenheder.* [Online]
Available at: https://instruks.bbr.dk/niveau_boli_erhverv/0/30
[Accessed April 2020].

Bygnings- og Boligregistret, 2009. *2.2. Niveauopdeling i BBR-systemet.* [Online]
Available at: https://instruks.bbr.dk/niveauopdeling/0/30
[Accessed April 2020].

Bygnings- og Boligregistret, 2020. *3.2.1 Ejendomsniveau.* [Online]
Available at: https://instruks.bbr.dk/ejendomsniveau/0/30
[Accessed April 2020].

Bygnings- og Boligregistret, 2020. *3.2.2 Bygningsniveau.* [Online]
Available at: https://instruks.bbr.dk/bygningsniveau/0/30
[Accessed April 2020].

Bygnings- og Boligregistret, 2020. *3.2.3 Bolig-/erhvervsenheden.* [Online]
Available at: https://instruks.bbr.dk/boligerhvervsenheden/0/30
[Accessed April 2020].

Danmarks Statistik, 2019. *Statistikdokumentation for Ejendomssalg 2019.* [Online]
Available at: https://www.dst.dk/da/Statistik/dokumentation/statistikdokumentation/ejendomssalg
[Accessed April 2020].

Danmarks Statistik, 2020a. *Ejendomssalg.* [Online]
Available at: https://www.dst.dk/da/Statistik/emner/priser-og-forbrug/ejendomme/ejendomssalg
[Accessed April 2020].

Danmarks Statistik, 2020b. *Forbrugerprisindeks.* [Online]
Available at: https://www.dst.dk/da/Statistik/emner/priser-og-forbrug/forbrugerpriser/forbrugerprisindeks
[Accessed April 2020].

Danmarks Statistik, 2020c. *Statistikbanken.* [Online]
Available at: https://www.statistikbanken.dk/statbank5a/SelectVarVal/saveselections.asp
[Accessed June 2020].

DinGeo, 2020. *Salgspriser.* [Online]
Available at: https://www.dingeo.dk/data/salgspriser/
[Accessed April 2020].

Ejendomsinfo, 2020. *SVUR - Statens Salgs- og Vurderingsregiste.* [Online]
Available at: https://ejendomsinfo.dk/ois-svur
[Accessed April 2020].

Esri, 2013a. *Geographically Weighted Regression (GWR) (Spatial Statistics).* [Online]
Available at:
resources.arcgis.com/en/help/main/10.1/index.html#/Geographically_Weighted_Regression_GWR/005p00000021000000/
[Accessed June 2020].

Esri, 2013b. *What they don't tell you about regression analysis.* [Online]
Available at:
resources.arcgis.com/en/help/main/10.1/index.html#/What_they_don_t_tell_you_about_regression_analysis/005p00000053000000/
[Accessed May 2020].

Esri, 2013c. *Interpreting OLS results.* [Online]
Available at:
http://resources.arcgis.com/en/help/main/10.1/index.html#/Interpreting_OLS_results/005p00000030000000/
[Accessed May 2020].

Miljø- og fødevareministeret, 2020h. *Nord2000-beregningsmetoden.* [Online]
Available at: https://mst.dk/luft-stoej/stoej/trafikstoej/nord2000-beregningsmetoden/
[Accessed May 2020].

Miljø- og Fødevareministeriet, 2020a. *Trafikstøj og sundhed.* [Online]
Available at: https://mst.dk/luft-stoej/stoej/trafikstoej/trafikstoej-og-sundhed/
[Accessed March 2020].

Miljø- og Fødevareministeriet, 2020b. *Trafikstøj.* [Online]
Available at: https://mst.dk/luft-stoej/stoej/trafikstoej/
[Accessed March 2020].

Miljø- og Fødevareministeriet, 2020c. *Vejstøjsstategien.* [Online]
Available at: https://mst.dk/luft-stoej/stoej/trafikstoej/vejstoejsstrategi/
[Accessed Mach 2020].

Miljø- og Fødevareministeriet, 2020d. *Støj-Danmarkskortet.* [Online]
Available at: https://mst.dk/luft-stoej/stoej/saerligt-for-borgere-om-stoej/hvad-er-stoej/kortlaegning-af-stoej/stoej-danmarkskortet/
[Accessed April 2020].

Miljø- og Fødevareministeriet, 2020e. *Støjgrænser og begreber om støj.* [Online]
Available at: https://mst.dk/luft-stoej/stoej/saerligt-for-borgere-om-stoej/hvad-er-stoej/stoejgraenser-og-begreber-om-stoej/
[Accessed May 2020].

Miljø- og Fødevareministeriet, 2020g. *Støj fra vejtrafik.* [Online]
Available at: https://mst.dk/luft-stoej/stoej/saerligt-for-borgere-om-stoej/er-du-generet-af-stoej/vejtrafik/
[Accessed May 2020].

Miljø- og Fødevareministret, 2020. *Kortlægning af støj.* [Online]
Available at: https://mst.dk/luft-stoej/stoej/saerligt-for-borgere-om-stoej/hvad-er-stoej/kortlaegning-af-stoej/
[Accessed May 2020h].

Miljøstyrelsen, 2003. *Hvad koster støj? - værdisætning af vejstøj ved brug af husprismetoden.* [Online]
Available at: https://www2.mst.dk/udgiv/publikationer/2003/87-7972-568-6/pdf/87-7972-569-4.pdf
[Accessed February 2020].

Miljøstyrelsen, 2020. *Støjkortlægning - Miljøstyrelsen.* [Online]
Available at: http://miljoegis.mim.dk/spatialmap?&profile=noise
[Accessed June 2020].

Owusu-Ansah, A., 2013. A REVIEW OF HEDONIC PRICING MODELS IN HOUSING RESEARCH. *A Compendium of International Real Estate and Construction,* Volume Volume 1, pp. 17-38.

Rego, F., 2015. *Quick Guide: Interpreting Simple Linear Model Output in R.* [Online]
Available at: https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R
[Accessed May 2020].

Rich, J. H. & Nielsen, O. A., 2004. ASSESSMENT OF TRAFFIC NOISE IMPACTS. *International Journal of Environmental Studies,* Volume 61(Issue 1), pp. 19-29.

Rosen, S., 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy,* Volume 82(1), pp. 34-55.

Statsministeret, 2020. *Transport og infrastruktur.* [Online]
Available at: https://www.regeringen.dk/regeringens-politik-a-å/transport-og-infrastruktur/
[Accessed March 2020].

Theebe, M. A. J., 2004. Planes, Trains, and Automobiles: The Impact of Traffic Noise on House Prices. *The Journal of Real Estate Finance and Economics,* Volume 28, pp. 209-234.

Udenrigsministeriet, 2017. *Handlingsplan for FN's 17 verdensmål.* [Online]
Available at: https://www.regeringen.dk/publikationer-og-aftaletekster/handlingsplan-for-fns-verdensmaal/
[Accessed June 2020].

Wilhelmsson, M., 2000. The Impact of Traffic Noise on the Values of Single-family Houses. *Journal of Environmental Planning and Management ,* Volume 43(Issue 6), pp. 799-815.