Energy Efficiency Valuation

Estimating the increase in expected transaction price due to improved energy efficiency for houses in the Dutch housing market

by

M.L. Groot Beumer

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday September 27, 2022 at 16:00.

Student number:	4325370	
Thesis committee:	Dr. D. Kurowicka,	TU Delft, supervisor
	Dr. N. Parolya,	TU Delft

TUDelft

Abstract

Recent advancements in causal inference and machine learning research have brought forward methods to estimate effects of interventions from observational data [16, 36]. The augmented inverse probability weighted (AIPW) estimator is such a method, which can be used to obtain estimates of potential outcomes. Potential outcomes are defined as a hypothetical outcome pair $\left\{Y^{(1)}, Y^{(0)}\right\}$, of which only one outcome is observed in the data. Estimation of intervention effects boils down to effectively estimating these potential outcomes.

Using the AIPW estimator, we aim to evaluate the average effect of increasing the energy efficiency of houses in the Netherlands on their expected transaction price, defined as $\delta = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]$. Moreover, we investigate how this expected effect changes when we condition on a subset $X = x$, $\delta(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right]$.

Given that our assumptions hold, we find that on average, the estimated expected increase in transaction price is positive when improving the energy efficiency of a house. Improving an energy inefficient house to moderately energy efficient is expected to increase the transaction price by approximately €97.70 ± 20.31 per m^2, while the improvement from moderately energy efficient to energy efficient increases the expected transaction price by approximately €20.96 ± 11.56 per m^2. In general, older, smaller and more energy inefficient houses increase most in expected transaction price per m^2 when their energy efficiency is improved.

Preface

This MSc. thesis, with the title "Energy Efficiency Valuation - Estimating the expected increase in transaction price due to improved energy efficiency for houses in the Dutch housing market", marks the end of my time as a TU Delft Applied Mathematics student at the faculty of EEMCS.

In this thesis, I aim to provide insight in the expected increase in transaction price of a house when its energy efficiency is improved. Increasing awareness for climate change as a result of the Paris climate agreement has made this topic of research highly relevant. Recent advancements in the field of causal inference, with one of the pioneers being the Dutch-American recent Nobel prize winner economist Guido W. Imbens, have brought forward methods to estimate effects of interventions from observational data. These methods are investigated and used in this research.

I started this graduation project at Ortec Finance. Ortec Finance is a company that supports their clients, among which municipalities and housing corporations, in their investment decision making. For their clients, being able to evaluate the benefits of increasing energy efficiency is of large value.

I would like to thank Ortec Finance for the opportunity to pursue research in such an interesting and relevant topic. In particular, I would like to thank Dennis and Mark for their supervision during my time at Ortec Finance. Although I did not finish my graduation project at Ortec Finance, I would like to express my gratitude for the help and support to Dennis and Mark during a period I was having a difficult time mentally.

I am incredibly thankful for Academic Counsellor Paula, who has helped me get back on track during this difficult period in my project.

I wish to thank my daily supervisor Dorota Kurowicka for her extensive guidance during my challenging graduation project. Her willingness to dive deep into a new area of research other than her specialization in order to assist me is remarkable and greatly appreciated. The expertise, knowledge and enthusiasm from Dorota during our meetings has been a huge source of motivation for me during the entire project. Additionally I would like to thank Nestor Parolya for taking part in my MSc. thesis committee.

I would like to thank my parents, Herman and Manita, my brother Koert, my sister Karolijn and my girlfriend Amber for their love and support during my studies. Lastly I would like to thank my friends that have contributed to the incredible and unforgettable experiences during my time in Delft.

M.L. Groot Beumer
Delft, September 2022

Contents

List of Tables

List of Figures

1

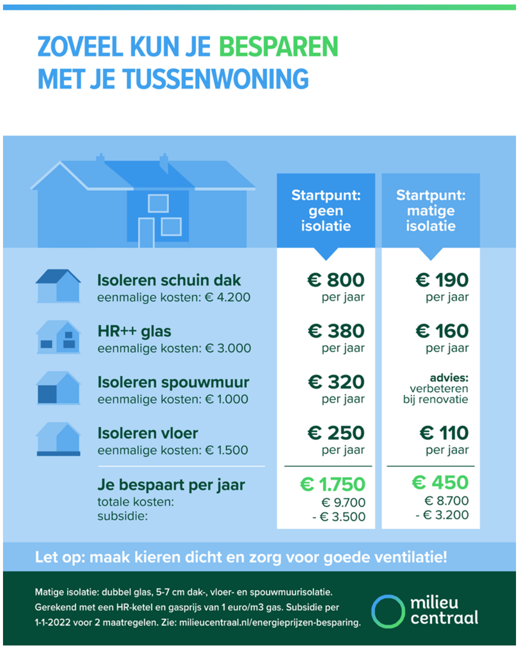
Introduction

1.1. Background

In the Netherlands, over 9% of carbon dioxide emissions in 2020 can be attributed to the burning of natural gas for the heating of houses [35]. The Paris climate agreement has set goals to limit the use of fossil fuels and reduce the carbon dioxide emissions. These goals impact almost all sectors, including the residential real estate sector. As a result of this agreement, homeowners are increasingly stimulated by their governments to increase the energy efficiency of their house.

In the Netherlands, the energy efficiency of a dwelling is indicated with an Energy Performance Coefficient label (EPC label). EPC labels range from A, a highly energy efficient house, to G, a highly energy inefficient house. Since 2015, EPC labels are mandatory when selling or renting out a house. EPC labels in the Netherlands are distributed by a range of certified commercial parties.

From 2015 until 2020, the period for which we have access to transactions, two different methods existed for determining the EPC label. Both these methods measure a number of characteristics of a dwelling and compare these against the NEN7120 standard, a standard for determining energy performance of a dwelling based on characteristics. These methods are the Energy Index (EI) method, and the Vereenvoudigd EPC label (VEL) method. Based on around 150 characteristics of a dwelling, the EI method estimates the Energy Performance Coefficient (EPC), which is a number on the interval $[0,5]$. Based on this EPC, the corresponding EPC label is handed out. The characteristics of a dwelling for this EI method are recorded by a certified expert.



ZOVEEL KUN JE BESPAREN
MET JE TUSSENWONING

	Startpunt: geen isolatie	Startpunt: matige isolatie
Isoleren schuin dak eenmalige kosten: € 4.200	€ 800 per jaar	€ 190 per jaar
HR++ glas eenmalige kosten: € 3.000	€ 380 per jaar	€ 160 per jaar
Isoleren spouwmuur eenmalige kosten: € 1.000	€ 320 per jaar	advies: verbeteren bij renovatie
Isoleren vloer eenmalige kosten: € 1.500	€ 250 per jaar	€ 110 per jaar
Je bespaart per jaar totale kosten: subsidie:	€ 1.750 € 9.700 - € 3.500	€ 450 € 8.700 - € 3.200

Let op: maak kieren dicht en zorg voor goede ventilatie!

Matige isolatie: dubbel glas, 5-7 cm dak-, vloer- en spouwmuurisolatie.
Gerekend met een HR-ketel en gasprijs van 1 euro/m3 gas. Subsidie per 1-1-2022 voor 2 maatregelen. Zie: milieucentraal.nl/energieprijzen-besparing.

milieu centraal

Figure 1.1: An overview of the average costs for improving the isolation of a terraced dwelling and the corresponding expected savings on the energy bills in 2022 in the Netherlands. Figure from Milieu Centraal [3].

The VEL method uses only a maximum of 10 of the most important characteristics, which are provided with proof by the home owner. The following 10 dwelling characteristics are measured.

- Construction age
- Dwelling type
- Wall insulation
- Roof & facade insulation
- Floor insulation
- Glazing
- Heating system (type and age of boiler)
- Ventilation
- Water heating system
- Renewable energy options (solar panels & solar water heater).

Based on these 10 characteristics, the energy efficiency of the dwelling is estimated by a certified party, based on the abovementioned NEN7120 norm. The VEL method is slightly less accurate, and therefore directly estimates an EPC label instead of an EPC score.

One of the main goals of the institution of EPC labels was to increase the transparency between the buyers and sellers of houses. The information the buyer and seller have about the dwelling is often unequal, as it is often difficult for a buyer to estimate the energy costs for a new house. As a result, a seller is often hesitant to invest in energy efficiency improving measures, as the return on investment may be hard to evaluate for a buyer. Policy makers often refer to this phenomenon as the energy efficiency gap [6]. EPC labels have been introduced in order to close this gap, allowing buyers to incorporate the energy efficiency of a dwelling in the price evaluation. This should encourage home owners to invest in the energy efficiency of their houses.

An investment in the energy efficiency of a house is twofold. On the one hand, there will be direct savings on the monthly energy bill. On the other hand, if energy efficiency is properly measured, the decrease of future bills should also be reflected in an increase of the expected transaction price of the house. Tools and approximations for estimating the savings on the energy bill when making a dwelling energy efficient are widely available, for example [1, 3]. An example is shown in Figure 1.1.

However, the fact that the expected transaction price may also increase when making a house more energy efficient is an underexposed subject. Perhaps because the estimation of this expected price premium is rather complex. The goal of this thesis is to estimate the expected price premium that is paid for a house in the Netherlands when its energy efficiency is improved, in comparison to the price of the house when its energy efficiency would not have been improved.

1.2. Problem Statement

In this Section we will summarize the research goals. The goal of this thesis is to estimate the expected increase in transaction price of a dwelling in the Netherlands when its energy efficiency is improved. To state the research question formally:

- What is the expected increase in transaction price of a dwelling in the Netherlands when its energy efficiency is improved?

This research question will be split into two parts. The reasons for doing so will become clear in Chapter 3. These research questions are:

1. What is the expected increase in transaction price of an energy inefficient dwelling in the Netherlands when its energy efficiency is improved to moderately energy efficient?

2. What is the expected increase in transaction price of a moderately energy efficient dwelling in the Netherlands when its energy efficiency is improved to energy efficient?

Furthermore, our interest lies not only in the overall expected increase in sale price of a dwelling. The factors impacting this price premium will be investigated as well. Do older houses benefit more from improving energy efficiency? Or do larger houses? Houses with what characteristics increase most in expected transaction price when improving the energy efficiency?

In order to answer these questions, the average treatment effect (ATE) and the conditional average treatment effect (CATE) (or: heterogeneous treatment effect) will be investigated. Treatment will be defined as improving the energy efficiency of a dwelling so that it falls in a higher category of energy efficiency. In Section 1.4, the mathematical definitions of the ATE and CATE, and the setting and methods used to estimate these will be introduced.

1.3. Relevance

In this Section, the literature regarding the valuation of energy efficiency of dwellings will be discussed. Afterwards, we will show how this thesis contributes to the existing literature and what is done differently in comparison with the current literature.

1.3.1. Literature summary

Research on the impact of a green status on the transaction price started as early as 1989, where Gilmer [24] found that the labels for energy efficient houses moderately shortens the time on the market, and Dinan and Miranowski [18] adopt a hedonic regression model to find that improving the energy efficiency of a house increased the expected selling price. A hedonic pricing model is a revealed preference model for estimating the willingness-to-pay for characteristics of a good. Hedonic models are most commonly estimated using regression analysis, and are very common in the use of real estate appraisal.

Positive effects of a house its energy performance on the sale conditions can also be seen at later times in the Netherlands. Brounen & Kok [11] research the capitalization of the energy efficiency label in the Dutch housing market, and the economic implications thereof. Brounen & Kok use a sample of 177,000 housing transactions in the period from January 2008 to August 2009. In the study, a logit model is used to estimate the adoption of energy label throughout the Dutch housing market. The study finds that energy labels had a negative sentiment at the time, which hindered capitalization of energy labels in the market. Moreover, using a Heckman two-step model on a subsample of 32,000 transactions, it finds that green labeled houses sell on average for 3.6% more than houses without an energy label. This premium can be partly related to the future energy savings due to the improved energy efficiency.

Building forth on Brounen & Kok, Aydin [5] implements both a hedonic regression model as well as an instrumented variable approach, to estimate the potential price premium paid for green labeled houses in the Netherlands. The study estimates the price premium separately per transaction year, uses a sample of transactions from 2008 to 2011, and limits its sample to single-family dwellings. Using an instrumental variable approach, the study finds that when the energy usage is halved as a result of energy efficiency improvements, the transaction price of the house increases by 11%. Moreover, as the study performed both ordinary least squares (OLS) regression, as well as an instrumental variable method, it concludes that OLS leads to downwards biased estimates of the market value of energy efficiency compared to the instrumental variable approach.

Chegut et al. [14] also base their study in the Netherlands, but only for the affordable housing sector (in Dutch: Sociale huur). The study looks at a sample of 17,835 homes which are sold in the period 2008-2013. By using a standard hedonic pricing model, the study finds that homes with a high energy efficiency sell for a 2.0% to 6.3% premium compared to otherwise similar dwellings. Moreover, it concludes that the price increase as a result of refurbishing and improving the energy label by 20% would more than pay for the retrofit in most instances.

In locations outside of the Netherlands, the effect of good energy efficiency of a dwelling on its transaction conditions is positive as well. Cajias, M., & Piazolo, D. (2013) [13] investigate the price premiums of energy efficient dwellings in Germany using a Hedonic approach. and find that 1% of energy conservation on average leads to a 0.45% increase in market price. A study by Eichholtz, P., Kok, N., & Quigley, J. M. [19] investigates American offices with a green rating by the top two rating companies in the US. The study clusters houses based on latitude and longitude, creating 893 clusters that contain nearby offices, where the average cluster contains 12 offices, each cluster containing at least one green rated office and a non-green rated office. Subsequently using a linear regression approach the study finds a price premium of 3% on the renting prices, and 16% on sale prices for green office buildings. Additionally, the study finds that the relative premium for green office buildings is systematically greater in the less expensive location clusters.

The study of Walls et al. [38] looks at three big regions in the US, namely Austin, Texas, Portland, Oregon and the Research Triangle, North Carolina. The study examines a sample of 170,000 transactions, of which a smaller subset is certified as green. The study investigates the capitalization of the Energy Star certification and local green certifications in the market price of dwellings. It uses several matching techniques to control for sample selection bias, and in order to match nearby houses with similar characteristics. Afterwards, the study uses OLS to estimate the difference in prices of certified homes to an appropriately matched set of non-certified homes. The study reports an increase in sale price of certified homes by 2%, 8% and 9%, for Austin, Portland and the Research Triangle, respectively.

The study of Fuerst et al. [21] looks at the impact of green certifications of commercial real estate in the United States. It performs an in-depth analysis of the price difference between certified and non-certified offices. The study finds three main drivers of price differences, namely additional occupier benefits, lower holding costs and a lower risk premium. Moreover, the study shows a detailed theoretical background of the use of the hedonic pricing model in the real estate sector. With this hedonic approach, the study controls for spatial coordinates of the properties, and for the submarket in which the house is located. The rent price premium is 5% and 4%, while the sale price premium is 25% and 26%, for LEED-certified and Energy Star-certified office buildings, respectively, according to the study.

Also, additional green attributes of a dwelling can contribute to an increase in the sales price. Dastrup et al. (2012) [17] perform a study on solar panels in the San Diego area. Using both a hedonic regression approach and a repeat sales index approach, the study finds that solar panels are capitalized in the market at a premium of roughly 3.5%. In the hedonic model the study controls for location by adopting a proxy for houses with the same zip code.

Some studies however do not find a positive effect of green attributes on the price of a house. Yoshida [39] finds a significant negative effect of energy efficient characteristics on the price of houses in Tokyo. Even though green houses in general sell for a price premium, when the model is corrected for quality and construction age, the green attributes negatively impact the transaction price. Yoshida thinks the main reason is the higher future maintenance cost of high quality energy efficient attributes of a house, that are taken into account by buyers.

1.3.2. Evaluation of existing literature

Before 2015, EPC labels were not mandatory in the Netherlands. As a result, studies that use transactions of dwellings before 2015 are impacted by sample selection bias. As concluded in [11], dwellings with and without an EPC label were vastly different from eachother in terms of characteristics, such as construction year and energy efficiency. As such, effects of having a certain label on the transaction price estimated for the sample of dwellings with an energy label may be very different for the group of labeled dwellings compared to the sample of all transacted dwellings. This phenomenon is commonly referred to as sample selection bias; the small sample that is studied is not representative for the sample as a whole. A study with more recent data where every sold dwelling is in possession of a label deals with problem.

Most of the current literature as discussed in 1.3.1 adopt a hedonic pricing model and subsequently estimate the model with OLS. There are a number of drawbacks with this approach. These drawbacks are the result of assumptions made for estimating a model with OLS, and are summarized below.

- Parametric models, such as OLS, assume a certain parametric form that relates the price of a dwelling to its characteristics. This parametric form is likely not correct. Non-parametric models can estimate the effect of interest without assuming a parametric form.
- Linear models, such as OLS, assume linear effects of variables on the price, which again is likely not correct.
- Using linear regression it is difficult to estimate variations in the effect of energy efficiency on the price as a function of other characteristics of dwellings. The variations in price as result of improvement of the energy efficiency as a function of other characteristics, is known as the heterogeneous treatment effect (HTE), and is our primary interest. A possibility to estimate the HTE is to add interaction terms, however, doing so again assumes a certain parametric form for the effect. Moreover, adding too many variables is known to decrease the general performance of OLS.
- Linear models such as OLS evaluate correlation, not the effect of improving EPC label on the price. As a result, it is not possible to draw valid conclusions regarding improving the energy label on the price of a dwelling using only OLS. The conclusions that are drawn only evaluate price differences between groups of dwellings with different energy efficiency when other characteristics are controlled for.

The drawbacks stated of OLS stated above will be substantiated in Chapter 2, where we also show how the method that we introduce in Section 1.4 can overcome these drawbacks.

Recent advancements in causal inference and machine learning literature, started by Chernozhukov [16], have made it possible to estimate treatment effects with non-parametric machine learning approaches, and performing valid inference. This makes it possible to directly evaluate the effects of a certain treatment, in our case this treatment is improving the energy efficiency of a house, on the expected transaction price of a house. Consequently, estimates produced by these methods can directly be used for decision-making and policy evaluation processes.

1.4. Problem Setting

In order to model the problem, we have access to a data set consisting of transaction prices, characteristics and treatment statuses of dwellings. How this data set was obtained and details regarding the variables used is being treated in Chapter 4. The available data on the transactions of dwellings are assumed to be realizations of independent and identically distributed (iid) random vectors $D_1,...,D_n$. A single observation D is distributed according to some unknown density function $p_D(d)$.

A dwelling can either be energy efficient or not, relative to some defined threshold, at the time of sale. Whether a dwelling is energy efficient or not at the time of sale will be referred to as the treatment. A dwelling that is energy efficient relative to the threshold is called treated, while a dwelling that is energy inefficient relative to the threshold is called untreated (or: control).

The possible treatments statuses that can be present in an specific house will be denoted by the random variable $T \in \{0,1\}$. Throughout this thesis, T will always be assumed to be a binary variable. For example, T may indicate if a house is energy efficient ($T = 1$) or energy inefficient ($T = 0$), relative to the defined standard.

The outcome variable will be denoted by $Y \in \mathbb{R}$, the transaction price per m^2. In this thesis, we consider a data set of sold houses, of which some of them were energy efficient ($T = 1$) and some of them were energy inefficient ($T = 0$), at the time of sale.

The data that is available will be assumed to be i.i.d. observations $D_i = (Y_i, T_i, X_i)$ for $i = 1,...,n$, where for the i^{th} dwelling the outcome is Y_i, the treatment received is T_i and its p-dimensional vector of covariates (or: controls/characteristics) will be $X_i \in \mathbb{R}^p$. We are interested in the effect of treatment assignment T on outcome Y.

We use uppercase letters such as Y, T or X when referring to a random variable. Observed values are written in lowercase; hence the ith observed value of X is written as x_i (where x_i is again a scalar or vector).

Recall that we would like to estimate the metric of interest, the expected increase in transaction price of a dwelling when its energy efficiency is improved. In order to do so, we introduce potential outcomes, first introduced by Neyman, Rubin and Rosenbaum [30], which are defined as the duo

$$\left\{Y^{(1)}, Y^{(0)}\right\}. \tag{1.1}$$

All dwellings are assumed to have both potential outcomes, however, only either of them is actually observed for every dwelling. The potential outcome $Y^{(1)}$ describes the outcome of interest, the transaction price, when a unit would be treated. Here, treatment indicates being energy efficiency at the time of sale. Similarly, $Y^{(0)}$ describes the transaction price of a unit when it would not be treated, e.g. energy inefficient at the time of sale. As a result, the setting could also be viewed as a missing data problem, where we have access to data $\tilde{D}_i = \left(Y_i^{(1)}, Y_i^{(0)}, T_i, X_i\right)$ for $i = 1,...,n$, where $Y^{(1)}$ is missing when $T = 0$ and $Y^{(0)}$ is missing when $T = 1$.

The potential outcomes of the transaction price of a house, $\left\{Y^{(1)}, Y^{(0)}\right\}$, are related to its characteristics X and its binary treatment status $T \in \{0,1\}$. The relation between the transaction price and the treatment status and characteristics will be referred to as the outcome model. The outcome model is described by (1.2).

The probability of a certain treatment status of a house is assumed to vary with its characteristics X, so that the probability of a dwelling having a certain treatment status is a function of X. Hence, the probability of treatment status equal to 1 is a function $p : \mathbb{R}^p \to [0,1]$. The relation between the probability of certain treatment status will be referred to as the propensity model, and is described by (1.3).

$$Y^{(t)} = \mu_t(X) + \epsilon \qquad \mathbb{E}[\epsilon \mid X] = 0 \tag{1.2}$$

$$\mathbb{P}(T = t \mid X) = p_t(X) \tag{1.3}$$

No further structural assumptions are forced upon functions μ_t and p_t.

Adopting this notation, the effects that we are interested in are the average treatment effect (ATE), defined as

$$\delta = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right], \tag{1.4}$$

and the heterogeneous treatment effect (HTE) (or: Conditional Average Treatment Effect (CATE)),

$$\delta(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right]. \tag{1.5}$$

In observed data, a house can only ever be either treated (e.g. energy efficient) or not at the time of sale, yet clearly not both. Consequently, assessment of the effect of a treatment often boils down to effectively estimating the counterfactual outcome, which is the outcome that is not observed for a certain unit.

A problem setting as described above is commonly referred to as the causal inference framework (or: potential outcomes framework) by Neyman [26] and Rubin [31]. Recent research has combined knowledge from semiparametric theory and machine learning with this area of causal inference. In particular, Chernozhukov et al. [16] analyzes the case of estimating $\delta(x)$ when it is constant or low-dimensional and linear, while allowing $\mu_t(X)$ to be high-dimensional. The study introduces valid methods for inference and construction of confidence intervals, while achieving $\sqrt{n}$-consistency and asymptotic normality for the estimation of $\delta(x)$ under mild regularity conditions. The estimator $\hat{\delta}_n$ of true function δ_0 is $\sqrt{n}$-consistent when

$$\left(\hat{\delta}_n - \delta_0\right) = O_p(n^{-\frac{1}{2}}), \tag{1.6}$$

where n is the sample size and $\alpha_n = O_p(n^d)$ denotes $\frac{\alpha_n}{n^d}$ is stochastically bounded. In other words, $\alpha_n = O_p(n^d)$ when for any $\epsilon > 0$ there exists a finite $M > 0$ and a finite $N > 0$ such that

$$\mathbb{P}\left(\left|\frac{\alpha_n}{n^d}\right| > M\right) < \epsilon \quad \forall n > N. \tag{1.7}$$

The estimator $\hat{\delta}_n$ is asymptotically normal if the difference with the true function δ_0 converges in distribution to a normal distribution as $n \to \infty$. That is,

$$\sqrt{n}\left(\hat{\delta}_n - \delta_0\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \tag{1.8}$$

for some finite σ^2. The studies of Oprescu [27] and Athey and Wager [4, 36] extend the results of Chernozhukov by estimating $\delta(x)$ with a random forest based estimator, allowing $\delta(x)$ to be estimated without specifying a parametric form. These studies show that under regularity conditions, and assumptions that will be discussed in Section 3.3, an asymptotically normal forest-based estimator $\hat{\delta}_n$ is available, allowing for the construction of valid confidence intervals.

Many of these methods for estimating treatment effects are implemented in the Python Library EconML [28]. EconML is open source software developed by the ALICE team of Microsoft Research. We will use this Python package, EconML, for the estimation of treatment effects.

The further Chapters in this thesis are structured as follows. In Chapter 2 an example will be introduced that shows different methods for estimating treatment effects in the framework mentioned above. In Chapter 3 the theory behind the estimators used in the example will be outlined, and how these estimators will be used on the real data set. In Chapter 4 it is described how the raw data was transformed into a cleaned up data set that can be used for analysis. In Chapter 5, the results are presented. Lastly, in Chapter 6 conclusions are drawn and the most important findings are discussed. Moreover, multiple directions for further research are considered.

2
Motivating Example

2.1. Problem setup

In short, the aim of my thesis is to investigate the effect of making a dwelling more energy efficient on its price. Furthermore, we would like to know how this effect changes for dwellings with different characteristics, for instance for dwellings in different locations, for dwellings with a different construction period or for dwellings of different size. Ultimately, it would also be of our interest to be able to pick dwellings with certain characteristics that benefit the most from improving their energy efficiency.

In order to gain insight into this problem, we have at our disposal a data set consisting of the transaction price of sold dwellings in the Netherlands in the period of 01-01-2015 until 01-01-2020. This data set is enriched with a large amount of characteristics concerning the individual dwellings, including the Energy Performance Coefficient label (EPC label or energy label). From this data we would like to infer the aforementioned effect for different dwellings. As we do not know much about the structure of the data, neither do we know the functional relations between the variables in our data set, and neither are we sure that we capture all important variables in our data set, a thorough analysis is necessary to be able to conclude anything about the relation between energy efficiency and dwelling price.

In order to start this thorough analysis, a simple simulated example will be presented, where problems that one has to deal with when estimating effects, as in our problem, are discussed. This approach has two main benefits.

1. The actual effect of increasing the energy efficiency of a single dwelling on its price will be known in this hypothetical problem. Consequently, it will be possible to evaluate the performance of the models to estimate this effect, in relation to the true effect. In our real problem, this true effect is not known and evaluation of the estimated effects is challenging.

2. An example with a limited amount of variables allows for visualizing the data easier, show the workings of the methods and helps building intuition for how the method will be used on the real problem.

The example will be extended step-by-step in order to substantiate and explain the methods used. This example will start in an overly simplified problem setting, and this setting will be extended so that the example will in the end closely resemble the real problem.

Throughout the example, we have a data set consisting of n i.i.d. units, $i = 1,...,n$. For every unit i in our sample, we have access to a response $Y_i \in \mathbb{R}$, which in our real problem corresponds to the transaction price of a house. Furthermore, we have a binary treatment indicator $T_i \in \{0,1\}$, which indicates whether unit i is treated or not, i.e. whether the dwelling i is energy efficient or not in our real problem. Lastly, we have a feature vector $X_i \in \mathbb{R}^p$, which consists of all the characteristics of unit i. Following the potential outcomes model of Neyman [26] and Rubin [31], potential outcomes for unit i are indicated with the notation $\left\{Y_i^{(0)}, Y_i^{(1)}\right\}$, where $Y_i^{(0)}$ denotes the response if unit i would not be treated, and $Y_i^{(1)}$ denotes the response if unit i would be treated. Note that only one of those outcomes is actually observed for unit i in our data set. The outcome that is not observed is defined as the counterfactual outcome.

Using these definitions, the average effect of improving the energy efficiency can be described as

$$\delta = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]. \tag{2.1}$$

The conditional average treatment effect (CATE) (or: heterogeneous treatment effect), which will be denoted as $\delta(x)$, can be defined as

$$\delta(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right], \tag{2.2}$$

and corresponds to the expected effect of some treatment, on the response Y, for a unit with characteristics $X = x$. In this Chapter, the aim is to provide the reader some intuition of different methods for estimating the ATE and CATE, and the problems that are faced when pursuing this goal.

2.1.1. Problem 1 - Linear relations

Suppose we generate a tuple (Y_i, T_i, X_i) for $i = 1,\ldots,n$ where $n = 2000$, and in this case X is a one-dimensional characteristic. Suppose that the data is simulated from the following distributions:

$$
\begin{aligned}
&T \sim Bernoulli(p(X))\\
&X \sim Uniform(0,2)\\
&Y = \alpha + \beta X + \delta(X)T + \epsilon\\
&\epsilon \sim \mathcal{N}(0,1).
\end{aligned}
$$

The functions $p(x)$ and $\delta(x)$ will start out as constants, and T is independent of X, e.g. $T \perp\!\!\!\perp X$. However, throughout this example, $p(x)$, $\delta(x)$ and the functional form of Y will evolve into more complex forms, and T and X will be made dependent, so that the variables more closely resemble the true problem. Starting out, Y can be expressed as a linear function of X and T. For the sake of the example, these distributions and functional relations are assumed unknown. Our aim in this example would be to estimate the effect of T on Y, for which the true value is indicated by the function $\delta(x)$, from our data.

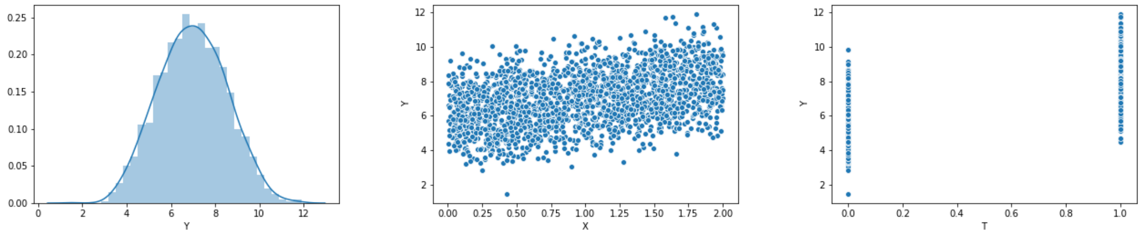
A well-known problem in estimating such effects in a causal inference setup, in contrast to when the aim is to develop a model for prediction, is that the ground truth for individual samples is unknown. In other words, for a single unit i with characteristics x_i, the true effect $\delta(x_i)$,

$$\delta(x_i) = \mathbb{E}\left[Y^{(1)} \mid X = x_i\right] - \mathbb{E}\left[Y^{(0)} \mid X = x_i\right] \tag{2.3}$$

is not observed from the data. The lack of an observed truth makes it difficult to assess the performance of models estimating the CATE. To illustrate this fact, consider the root mean squared error (RMSE), a metric often used to examine the difference between estimated values and true values,

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(\delta(x_i) - \hat{\delta}(x_i)\right)^2}, \tag{2.4}$$

where $\hat{\delta}(x)$ is the estimated function for $\delta(x)$. In causal inference problems, the true value of function $\delta(x)$ at point x is unknown, because for a single unit i, only either $y_i^{(1)}$ or $y_i^{(0)}$ is observed, but never both, as a single unit can not have multiple treatment statuses. In this simulated example, however, the true function is known, and so the RMSE can be calculated. In every step of the example, the RMSE is calculated using the known true function.



(a) Distribution plot of Y

(b) Scatter plot of X and Y

(c) Scatter plot of T and Y

Figure 2.1: Visualizations of the data distributions and relations between Y, T and X for problem 1.

In order to infer the effect of T on Y in this case, a possible approach would be to assume the functional relationship between Y and X and T is of a linear form. In other words, to assume the relation is $\hat{Y} = \hat{\alpha} + \hat{\beta}X + \hat{\delta}T$ and to find values $\hat{\alpha}, \hat{\beta}$ and $\hat{\delta}$ such that the difference between the sum of squared real values of Y and the estimated $\hat{Y}$ over $i = 1, \ldots, n$, i.e. $\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2$, is minimized. This is a basic problem in statistics and can be solved with ordinary least squares (OLS). As the relation between Y, X and T is linear, OLS will give consistent estimates for α, β and δ. Estimating coefficients for this example provides the following Table of estimates for the coefficients.

As can be seen from Table 2.1, OLS proves to be a good method for estimating the effect δ in this case.

2.1.2. Problem 2 - Misspecified Linear model

The problem from Subsection 2.1.1 is now extended, by simulating data in such a way that $\delta(x)$ and $p(x)$ are no longer constants. By varying $p(x)$, the probability that a unit is treated depends on its characteristics

	Real coefficient	Estimated coefficient	Standard deviation	Estimated 95% ci
α	5.00	4.96	0.049	$[4.87, 5.06]$
β	1.00	1.03	0.039	$[0.95, 1.11]$
δ	2.00	2.02	0.045	$[1.93, 2.10]$

Table 2.1: Coefficients estimated with OLS assuming a linear relation.

x. Furthermore, now $\delta(x)$ is a function of x, instead of estimating a constant $\hat{\delta}$, the aim is now to estimate a function $\hat{\delta}(x)$ that represents the effect of increasing the energy efficiency of a house on its price, as a function of the characteristics of a house. The fact that $\delta(x)$ is not a constant is in line with the real problem, where it might be the case that for large houses, an increase in energy efficiency has a larger effect on its price than for smaller houses.

$$p(x) = \frac{x}{2}$$
$$\delta(x) = 2 - x$$

In the following Section, it is illustrated what problems occur when the relations between Y and X and T are assumed linear and δ is assumed a constant, as was done in the previous Subsection. Suppose the data is estimated with a linear regression model similar to the model in Subsection 2.1.1, where δ is assumed constant, i.e. $Y = \alpha + \beta X + \delta T + \epsilon$. The coefficients that are estimated using linear regression are now as follows.

	Real coefficient/function	Estimated coefficient	Standard deviation	Estimated 95% ci
α	5.00	5.52	0.051	$[5.42, 5.62]$
β	1.00	0.47	0.040	$[0.39, 0.55]$
$\delta(x)$	2.00 - x	0.95	0.047	$[0.86, 1.04]$

In order to show the relation of the estimated δ to the true function $\delta(x)$, both are plotted below.

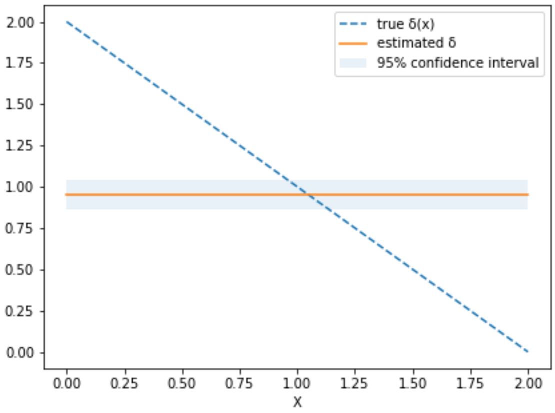


Figure 2.2: $\hat{\delta}$ estimated from misspecified linear model plotted against the true $\delta(x)$.

Obviously, assuming δ to be constant is not correct in this situation. However, if the true function $\delta(x)$ is not known, it is also not immediately clear what form $\delta(x)$ should be assumed to have. Certainly, it would be possible to add interaction terms such as $X \cdot T$ as predictors into the linear regression. Then, our model would be correctly specified,

$$Y = \alpha + \beta_1 X + \beta_2 TX + \delta T + \epsilon, \tag{2.5}$$

and $\delta(x)$ could be estimated by $\hat{\delta}(x) = \hat{\delta} + \hat{\beta}_2 x$. A problem with this approach, however, is that in the real problem, the functional form of $\delta(x)$ is not known. The relation between Y and X and T could be nonlinear and rely on many nonlinear interactions between X and T. In order to capture such relations between variables, a common approach is to add many cross-terms and nonlinear transformations of X as predictors. However, imagine a setting with 20 variables, where interaction terms (e.g. $X \cdot T$), logarithmic terms (e.g. $log(X)$), quadratic terms (e.g. X^2) and higher order terms are added as predictors. Adding many terms in this manner gives rise to other problems in linear regression:

1. Multicolinearity between X, X^2, X^3 etc. leads to inconsistent estimates.
2. If the amount of predictors grows larger than the amount of samples, the model will have very small statistical power.

Consequently, more advanced methods that can deal with nonlinearities and do not assume a specified form are required. In the next Subsections, it is illustrated how doubly robust estimators can be a convenient option to estimate $\delta(x)$, without the need to assume a functional form of the relation between Y and X and T. Doubly robust estimation is the method that will be used to estimate $\hat{\delta}(x)$ for our real problem. Subsequently, the theoretical basis of doubly robust estimation will be elaborated in Chapter 3. In short, doubly robust estimation consists of three separate modeling steps,

1. Estimate a classification model for the treatment indicator T, from characteristics X, in order to obtain estimates for so-called propensity scores, $\hat{p}_t(X) = \mathbb{P}[T = t \mid X]$.
2. Estimate regression models for response Y, from characteristics X and for treatment indicator $T = t$, in order to obtain $\hat{\mu}_t(X) = \mathbb{E}[Y \mid X, T = t]$.
3. Use estimated values for $\hat{p}(X)$ and $\hat{\mu}_t(X)$ in the doubly robust estimator to estimate $Y^{(1)}$ and $Y^{(0)}$, and regress $Y^{(1)} - Y^{(0)}$ on X to obtain estimates for $\delta(X)$.

These three steps are shortly elaborated below and implemented on the same example. The theoretical motivation for the doubly robust estimator is presented in Chapter 3.

Classification model

The classification model (or: propensity model) is used to estimate a probability for the treatment indicator T, as a function of characteristics X. Such a function can be estimated with several classification models, such as logistic regression, or even with non-parametric models like classification forests. In this step, the aim is to find the best performing classification model for the problem.

To find the best performing classification model, a grid search is applied to find the model with the hyperparameter settings that minimize the log loss for this step. Elaboration of the method used is presented in Chapter 3. The log loss is defined as

$$L(t,p)_{logloss} = -\frac{1}{n}\sum_{i=1}^{n}(t_i log(p_i) + (1-t_i) log(1-p_i)), \tag{2.6}$$

where p_i is the predicted probability of unit i belonging to class 1, and $t_i \in \{0,1\}$ is unit i its true class.

The motivation for minimizing log loss instead of another metric is that log loss penalizes the probability predictions rather than classification error. Further elaboration on this fact is given in Subsection 3.2.2.

Log loss is minimized with a combination of grid search and cross-validation, in order to ensure good generalization of the model. These methods are substantiated in Subsection 3.2.3. For the purpose of this example, two different models with different settings of hyperparameters are used to estimate a classification model as stated in Table 2.2.

Model	Hyperparameter settings
Logistic Regression l2 regularized	$regularize\_magnitude$: {0, 0.01, 0.1, 1.0, 10.0, 100.0}
Random Forest Classifier	$min\_samples\_leaf$: [10,50], max_depth: [$None$,3,5]

Table 2.2: Hyperparameter settings for grid search in the motivating example.

These models are fitted using 5-fold cross-validation subject to minimizing the log loss. The model that produces the smallest average log loss is saved. In this example, the random forest classifier with hyperparameters $max\_depth = 3$ and $min\_samples\_leaf = 10$ produces the smallest cross-validated log loss for the classification step.

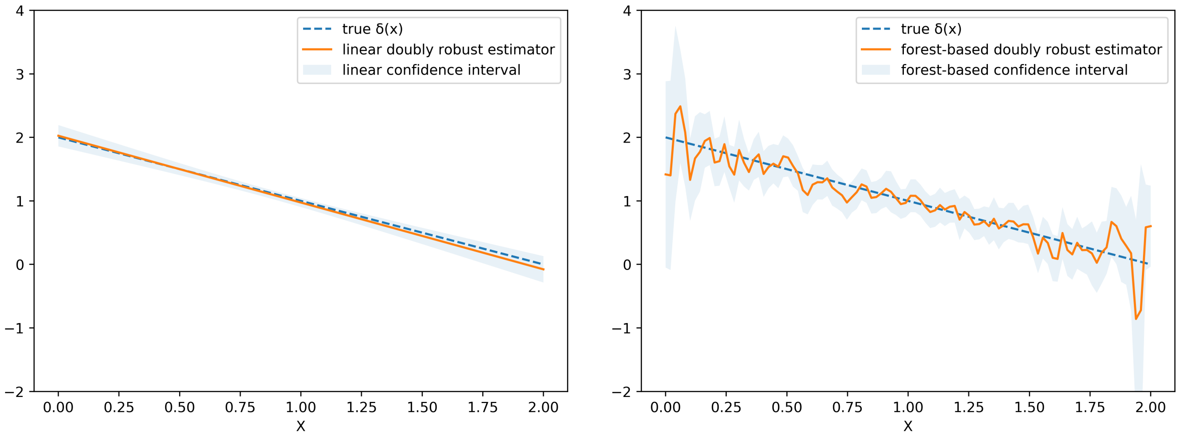
Regression model

Similar to the classification step, the aim is to find a regression model that estimates Y from X and T. This time, the loss function used is the Root Mean Squared Error (RMSE). Again, cross-validation in combination with grid search is used to find such a model and corresponding hyperparameters. In this example it is found that a random forest regression model with hyperparameters $max\_depth = 3$ and $min\_samples\_leaf = 50$ produced the smallest cross-validated RMSE for the regression step.

Final model

The final model is chosen based on what form $\delta(x)$ is expected to have. As this effect is unknown in our real example, it is not possible to optimise this model choice through the minimization of some error function. Using forest-based final models allows for non-specified functional forms of Y in relation to X. Wager and Athey [36] developed a non-parametric forest-based regression model for estimating treatment effects, which allows for statistical inference and produces valid confidence intervals.

If the function $\delta(x)$ is assumed to have a linear form, using a linear final stage model would be preferred. Naturally, it would be possible to assess the confidence of the estimations when using a linear final stage model. For the purpose of this example, both a linear and a forest-based doubly robust estimator are used to estimate the function $\delta(x)$. Both these methods will use the classification and regression models for the first two stages as mentioned above.



(a) Linear final model

(b) Forest-based final model

Figure 2.3: Doubly robust estimation of $\delta(x)$ with a parametric and non-parametric final model.

As can be seen from Figure 2.3, estimating δ assuming a linear form of $\delta(x)$ has good performance in this scenario. The confidence bounds for both the linear estimator as well as the forest-based estimator are relatively wide at both of the ends of the range of X. This is a result of the distribution of T. As T is simulated from a $Bernoulli(X/2)$ distribution, when X is close to 0 there will be a very low probability of T having a value of 1. Similarly, when X is close to 2, there will be a very low probability of T having a value of 0. As a result, not many data points exist with X close to 0, that have a treatment indicator $T = 1$, and vice versa when X is close to 2, which makes it hard for the model to estimate values of the function $\delta(x)$ at these values of x. These confidence bounds get narrower when more data points are available. This is also the intuition behind the positivity assumption discussed in Chapter 3, and is highly relevant for our real problem. The positivity assumption intuitively says that for all characteristics X, there is at least some probability that a unit with these characteristics has either treatment status. This assumption is reasonable, because when the probability of treatment is 0 at some $X = x$, the treatment effect at point x, which compares the treated and untreated state, does not have meaning, as the treated state can not occur. The function $\hat{\delta}(x)$ can not be estimated at a point x where there is a probability of 1 that a unit is either treated or untreated. For houses, this means that we might need to work with subsets of our complete dataset, if the probability of a very new house having a non-green label is close to 0.

The root mean squared error of both these models and of the misspecified linear regression model are summarized in the Table below, based on the prediction of δ for 100 data points ranging from $X = 0$ to $X = 2$.

Method	RMSE
Linear regression (misspecified)	0.638
Linear doubly robust estimation	0.021
Forest-based doubly robust estimation	0.121

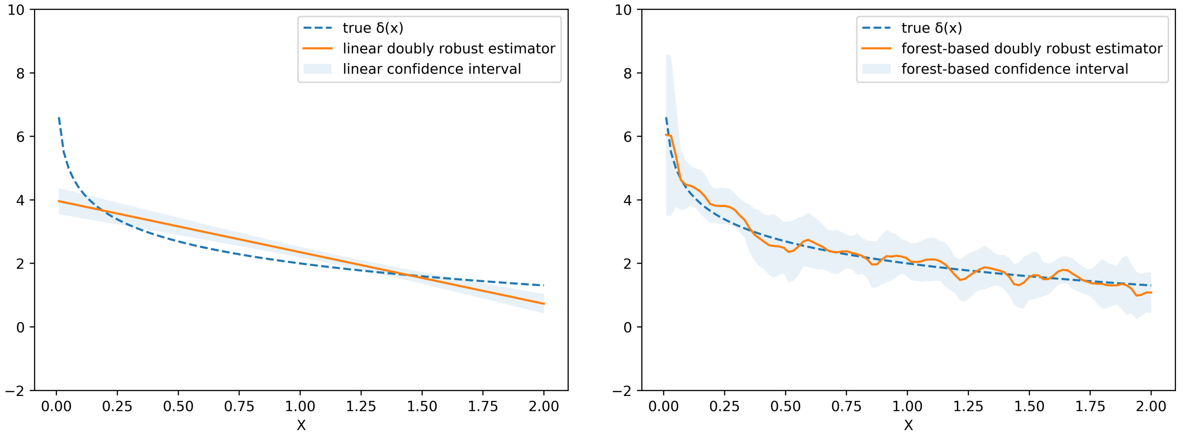
2.1.3. Problem 3 - Nonlinear effect

Now, the example is extended so that $\delta(x)$ has a nonlinear form,

$$\delta(x) = 2 - log(x).$$

Following a similar approach as in Subsection 2.1.2, a linear doubly robust estimation model and a forest-based doubly robust estimation model are used to estimate $\delta(x)$. For the first-stage classification and regression models, the same grid-search and hyperparameter optimization techniques are used. A random forest classifier with $max\_depth = 3$ and $min\_samples\_leaf = 10$ and a random forest regressor with $max\_depth = 3$ and $min\_samples\_leaf = 50$ perform best for the classification and the regression steps of the doubly robust estimator, respectively.

The results of the estimates for $\delta(x)$ of the final models are visualised in the Figure below.



(a) Linear final model

(b) Forest-based final model

Figure 2.4: Doubly robust estimation of $\delta(x)$ with a parametric and non-parametric final model.

The root mean squared error of these approaches are summarized in the Table below.

Method	RMSE
Linear regression (misspecified)	2.226
Linear doubly robust estimation	0.479
Forest-based doubly robust estimation	0.210

It can be concluded that when the relations between the variables increase in complexity and the functional form is unknown, forest-based doubly robust estimators can be a convenient method for estimating heterogeneous treatment effects. In the upcoming extensions of the example, the aim is to show how certain problems that exist in our real problem can be resolved.

2.1.4. Problem 4 - Unobserved variables

In the upcoming extension of the example, a sensitivity analysis is performed. The aim of this extension is to show how the results of our estimator may be biased when the unconfoundedness assumption does not hold. Unconfoundedness is an assumption made for our real problem, and roughly translates to the assumption that all variables that simultaneously strongly affect both the treatment assignment T, as well as the response variable Y, are observed and can be controlled for. A more thorough discussion on the unconfoundedness assumption for our real problem is presented in Chapter 3. Unconfoundedness is automatically violated when there are variables that strongly affect both the treatment assignment, T, as well as the response variable Y, which are not observed. Throughout this extension of the example from 2.1.3, there will be an unobserved variable U that affects either the response Y, the treatment assignment T, or both. This variable is correlated with variable X with correlation coefficient 0.7. The data is now generated as follows.

$$
\begin{aligned}
&T \sim Bernoulli(p(X,U))\\
&Y = \alpha + \beta X + \gamma U + \delta(X)T + \epsilon\\
&U \sim Uniform(0,2)
\end{aligned}
$$

In order to show how the unobserved variable U increased the error, γ and $p(x,u)$ will be varied so that U affects either the probability of treatment assignment, $\mathbb{P}(T = t \mid X)$, the outcome Y, or affects both, with different strengths.

In scenarios 1-3 from Table 2.3, the unobserved variable only affects the outcome Y, but not the probability of treatment, which will be kept fixed. In scenarios 4-6, the unobserved variable only affects the probability of treatment, but not the outcome variable. In scenarios 1-6, the estimates are expected to have larger confidence intervals than when there are no unobserved variables. However, the estimates should be unbiased estimates of the true effect.

In scenarios 7-9, the unobserved variable impacts both the outcome variable, as well as the probability of treatment. This is expected to produce biased estimates of the treatment effect.

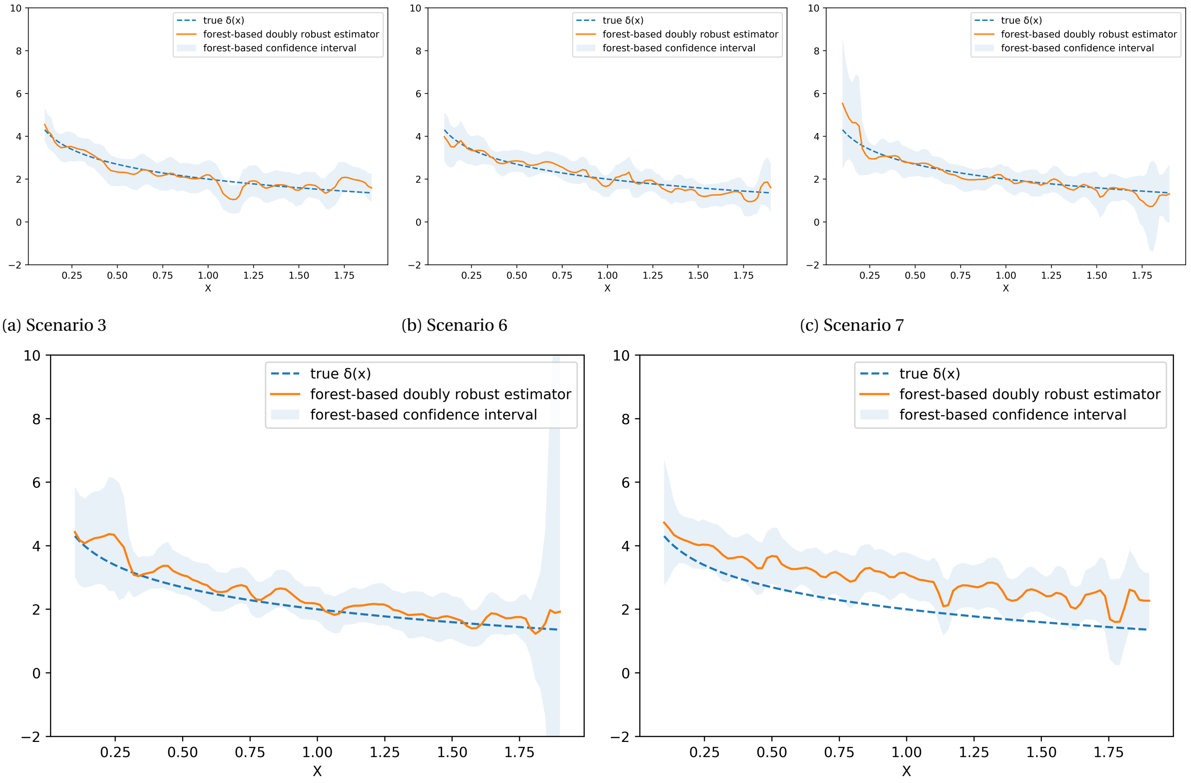
The results are summarized in Table 2.3.

Table 2.3: Estimates of the RMSE for the forest-based doubly robust estimates for different strengths of missing variables.

Case	Scenario	Variable values	RMSE
U only affects Y (fix $p(x,u) = 0.5$)	1	$\gamma = 0.5$	0.267
	2	$\gamma = 1$	0.275
	3	$\gamma = 2$	0.286
U only affects T (fix $\gamma = 0$)	4	$p(x,u) = 0.1u + 0.4x$	0.248
	5	$p(x,u) = 0.2u + 0.3x$	0.222
	6	$p(x,u) = 0.5u$	0.238
U affects both T and Y	7	$\gamma = 0.5$ & $p(x,u) = 0.1u + 0.4x$	0.305
	8	$\gamma = 1$ & $p(x,u) = 0.2u + 0.3x$	0.330
	9	$\gamma = 2$ & $p(x,u) = 0.5u$	0.825

As expected, the RMSE increases significantly when U has a strong impact on both the response Y, as well as the treatment assignment T. In all other cases, the RMSE remains fairly constant. In Figure 2.5, it can be seen that when the unconfoundedness assumption is violated, the estimated effect will be biased.

In this Chapter we presented several methods for estimating treatment effects with different forms. Furthermore, we have shown how violations of the assumptions, in particular the unconfoundedness assumption, can lead to biased estimates. The theoretical basis of these methods will be further elaborated in Chapter 3.



(a) Scenario 3

(b) Scenario 6

(c) Scenario 7

(d) Scenario 8

(e) Scenario 9

Figure 2.5: Forest-based estimation for different levels of unconfoundedness, scenarios from Table 2.3.

3
Theoretical Background

3.1. Estimating treatment effects

Throughout my thesis, a causal inference setting will be adopted. We denote $Y^{(t)}$ as the potential outcome that would have been observed if the sample would have energy efficiency $T = t$ at the time of sale. In this setting, we consider the following equations:

$$Y^{(t)} = \mu_t(X) + \epsilon, \qquad \mathbb{E}[\epsilon \mid X] = 0 \tag{3.1}$$

$$\mathbb{P}(T = t \mid X) = p_t(X). \tag{3.2}$$

No further assumptions are made on $\mu_t(X)$ and $p_t(X)$. Our goal is to estimate the average effect of improved energy efficiency on the transaction price of dwellings. When T is assumed to take binary values, this effect can be defined as

$$\delta = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]. \tag{3.3}$$

Here, δ will be referred to as the average treatment effect.

Additionally, we would like to infer whether and how the average treatment effect changes if we condition on a certain subsample $X = x$,

$$\delta(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right]. \tag{3.4}$$

In this equation, $\delta(x)$ will be referred to as the conditional average treatment effect. In this Section, possible methods for obtaining estimates of such effects are discussed.

3.1.1. Randomized experiments

We consider the setting as described in Section 1.4 of the introduction Chapter. For completeness, we will shortly repeat the setting.

The available data on the transactions of dwellings are assumed to be realizations of independent and identically distributed (iid) random vectors $D_1, ..., D_n$. A single observation D is distributed according to some unknown density function $p_D(d)$.

A dwelling can either be energy efficient or not, relative to some threshold, at the time of sale. Whether a dwelling is energy efficient or not at the time of sale will be referred to as the treatment. A dwelling that is energy efficient relative to the threshold is called treated, while a dwelling that is energy inefficient relative to the threshold is called untreated (or: control).

The possible treatments statuses that can be present in an specific house will be denoted by the random variable $T \in \{0, 1\}$. Throughout this thesis, T will always be assumed to be a binary variable. For example, T may indicate if a house is energy efficient ($T = 1$) or energy inefficient ($T = 0$), relative to the defined standard.

The outcome variable will be denoted by $Y \in \mathbb{R}$, the transaction price per m^2. In this thesis, we consider a data set of sold houses, of which some of them were energy efficient ($T = 1$) and some of them were energy inefficient ($T = 0$), at the time of sale.

The data that is available will be assumed to be i.i.d. observations $D_i = (Y_i, T_i, X_i)$ for $i = 1, ..., n$, where for the i^{th} dwelling the outcome is Y_i, the treatment received is T_i and its p-dimensional vector of covariates (or: controls/characteristics) will be $X_i \in \mathbb{R}^p$. We are interested in the effect of treatment assignment T on outcome Y.

We could define population parameters $\mu_1 = \mathbb{E}[Y \mid T = 1]$ and $\mu_0 = \mathbb{E}[Y \mid T = 0]$ and $\Delta = \mu_1 - \mu_0$. Here, Δ would be the difference in mean outcome in the population of the treated and untreated units. Then we could estimate Δ simply by estimating the difference in mean outcomes for each group,

$$\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{1}{n_1} \sum_{i=1}^{n} Y_i T_i - \frac{1}{n_0} \sum_{i=1}^{n} Y_i (1 - T_i), \tag{3.5}$$

where $n_1 = \sum_{i=1}^{n} T_i$ and $n_0 = \sum_{i=1}^{n}(1 - T_i)$. Thus, $\hat{\mu}_1$ and $\hat{\mu}_0$ denote the mean outcome for the treated and untreated population, respectively.

Typically, Δ is not a good measure for examining the effect of a treatment on the outcome. If the treatment status of individuals is not random, the characteristics of treated and untreated units might be inherently different. For instance, dwellings that are energy efficient might be inherently younger, smaller or are in different locations than their energy inefficient counterparts. As a result, Δ would not only reflect the treatment effect of interest on the outcome, but also other effects on the outcome due to these differences.

Thus, another view on defining a treatment effect is necessary. In order to describe such a treatment effect better than we can with merely statistical association, we adopt the notion of potential outcomes, introduced by Neyman, Rubin and Rosenbaum [30, 31]. We assume that for every treatment $t \in T$ a potential outcome $Y^{(t)}$ exists. $Y^{(t)}$ denotes the outcome for a randomly selected unit, would treatment $T = t$, possibly in contrast to reality, have been present in that unit. As T is assumed to be binary, we define only two potential outcomes, $Y^{(1)}$ and $Y^{(0)}$. These are defined potential outcomes, because for a randomly selected unit, both potential outcomes are never simultaneously observed. However, using the definition of potential outcomes, we can define a treatment effect as $Y^{(1)} - Y^{(0)}$ for the randomly selected individual. Although this treatment effect can not be measured at an individual level, it may be possible under certain assumptions to estimate a treatment effect for the whole population,

$$\delta = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right], \tag{3.6}$$

where δ is referred to as the average treatment effect (ATE).

A randomized experiment is an experiment where treatment is given to a subset of individuals participating in the experiment in a random manner. Hence, in a randomized experiment, treatment assignment T is random and independent of characteristics X. Intuitively, one can reason that in this case, Δ is an unbiased estimate for the average treatment effect δ. That is, because units are assigned to treatment at random, the treated and untreated groups have similar distributions. Consequently, any difference in outcome can plausibly be contributed to the treatment.

In order to show this intuition formally, we need two assumptions. These assumptions are necessary to interpret the observed values, and indicate that the treatment assignment was indeed random.

Assumption 1. Consistency of potential outcomes

$$Y = TY^{(1)} + (1-T)Y^{(0)}. \tag{3.7}$$

That is, the observed outcome Y is equal to potential outcome $Y^{(1)}$ if the unit received treatment, and equal to $Y^{(0)}$ if not. This assumption intuitively implies that the observed outcome under treatment is equal to the potential outcome given treatment, and vice versa.

Additionally, treatment assignment has to be random, where random is defined as independent of the potential outcomes.

Assumption 2. Randomized treatment assignment

$$Y^{(t)} \perp\!\!\!\perp T \qquad \forall t \in T, \tag{3.8}$$

where $\perp\!\!\!\perp$ denotes independence.

In a randomized experiment, it is reasonable to assume that treatment assignment is independent of the potential outcomes. To make this more intuitive, one can view both potential outcomes as outcomes every unit possesses before treatment happened. It is only when a unit is either treated or not, that one of these outcomes will appear. Note that this independence does certainly not imply that $Y \perp\!\!\!\perp T$, which is the exact dependence of our interest.

When (3.7) and (3.8) hold, it can be easily shown that Δ is unbiased for the average treatment effect δ in (3.3),

$$\begin{aligned}
&\Delta = \mathbb{E}[Y \mid T=1] - \mathbb{E}[Y \mid T=0]\\
&\text{By (3.7)}\\
&\quad = \mathbb{E}\left[TY^{(1)} + (1-T)Y^{(0)} \mid T=1\right] - \mathbb{E}\left[TY^{(1)} + (1-T)Y^{(0)} \mid T=0\right]\\
&\quad = \mathbb{E}\left[Y^{(1)} \mid T=1\right] - \mathbb{E}\left[Y^{(0)} \mid T=0\right]\\
&\text{By (3.8)}\\
&\quad = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]\\
&\quad = \delta.
\end{aligned}$$

Consequently, as $\frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)$ is an unbiased estimator for Δ, it is also an unbiased estimator for the average treatment effect δ.

3.1.2. Observational data

The problem that we have to deal with, however, is not a randomized experiment, and so we can not assume (3.8). In observational data, like we have, treatment is not assigned at random, but by choice. As a result, the groups of treated and untreated units are likely inherently different. If this is indeed the case, Δ is no longer an unbiased estimate of the average treatment effect δ, because

$$\mathbb{E}\left[Y^{(1)} \mid T=1\right] \neq \mathbb{E}\left[Y^{(1)}\right].$$

In other words, for treated units, the potential outcome when treated is different than the potential outcome when treated for the sample as a whole. In order to be able to estimate treatment effects from observational data, a different assumption than (3.8) is necessary to progress. This assumption is the unconfoundedness assumption; conditional on covariates X, the potential outcomes are independent of treatment T. This assumption will be used in the derivations of the estimators constructed later in this Chapter.

Assumption 3. Unconfoundedness (or: conditional exchangeabililty) of treatment assignment.
Conditional on X, the potential outcomes $Y^{(t)}$ are independent of the treatment assignment T.

$$Y^{(t)} \perp\!\!\!\perp T \mid X, \qquad \forall t \in T. \tag{3.9}$$

This assumption is key to being able to estimate treatment effects from observational data. It says that the assignment of treatment is random, conditional on the set of observed characteristics X. For instance, imagine we have a group of dwellings with exactly the same values for all observed characteristics like construction age, amount of floors and total living area. Then within every such a group with the same characteristics, the probability of the dwelling being energy efficient is assumed to be constant. More generally, one can say that all characteristics of units that impact the treatment assignment as well as the outcome are captured in our set of characteristics X. If there are features that significantly impact treatment assignment T and outcome Y, that are not in our data, then (3.9) does not hold.

If (3.7) and (3.9) hold, the average treatment effect can be identified from the distribution of the observed data, (Y, T, X),

$$\begin{aligned}
\delta &= \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right] = \mathbb{E}\left[Y^{(1)}\right] - \mathbb{E}\left[Y^{(0)}\right] \\
&= \mathbb{E}_X\left[\mathbb{E}\left[Y^{(1)} \mid X\right]\right] - \mathbb{E}_X\left[\mathbb{E}\left[Y^{(0)} \mid X\right]\right] \qquad (3.10)\\
&= \mathbb{E}_X\left[\mathbb{E}\left[Y^{(1)} \mid T=1, X\right]\right] - \mathbb{E}_X\left[\mathbb{E}\left[Y^{(0)} \mid T=0, X\right]\right] \qquad (3.11)\\
&= \mathbb{E}_X\left[\mathbb{E}\left[Y \mid T=1, X\right]\right] - \mathbb{E}_X\left[\mathbb{E}\left[Y \mid T=0, X\right]\right] \qquad (3.12)\\
&= \mathbb{E}_X\left[\mathbb{E}\left[Y \mid T=1, X\right] - \mathbb{E}\left[Y \mid T=0, X\right]\right] \qquad (3.13)
\end{aligned}$$

where (3.10), (3.11) and (3.12) follow from the law of total expectation, the unconfoundedness assumption (3.9) and the consistency assumption (3.7), respectively. As a result, the average treatment effect can be estimated with the observations (Y, T, X). Note that some caution is required, as we have to make sure that the events that we condition on in (3.11), (3.12) and (3.13) are not events with probability zero. Hence, the assumption that this event indeed does not have zero probability is required. We will introduce and discuss the positivity assumption in more detail later in this Section, which will ensure that.

Conditional mean estimation

Using what we learned in the previous paragraph, an intuitive approach to estimate the average treatment effect δ would be to consider conditional mean models,

$$\mathbb{E}[Y|T=1, X] = \mu_1(X, \xi), \tag{3.14}$$

$$\mathbb{E}[Y|T=0, X] = \mu_0(X, \xi). \tag{3.15}$$

Here, ξ is a finite-dimensional parameter that describes the models. These conditional mean functions can be estimated by different models. In the motivating example in Chapter 2, the conditional mean functions were estimated with ordinary least squares in the first part, and later with random forest regression.

Estimates for these conditional mean functions, say $\mu(T=1, X, \hat{\xi}_n)$ and $\mu(T=0, X, \hat{\xi}_n)$ are then obtained. Under the assumption that these models are correct, then a consistent and unbiased estimator of the average treatment effect, δ, could be obtained by substituting the functions into (3.13), to get

$$\hat{\delta}_n^{REG} = \frac{1}{n} \sum_{i=1}^{n} \left(\mu_1(X_i, \hat{\xi}_n) - \mu_0(X_i, \hat{\xi}_n)\right). \tag{3.16}$$

Generally, such estimation works reasonably well as long as the regression models are good estimates of the true conditional mean functions. However, this method has a significant drawback.

When the set of covariates X is large or complex, it is common practice to use regression models that penalize complexity. This is called regularization, and is performed in order to refrain the variance from exploding when many variables are present. Regularization, in a sense, reduces the variance by introducing small bias. For prediction purposes, this is generally motivated by the fact that the overall error in the prediction, which is a combination of bias and variance, is reduced. However, this regularization leads to a biased treatment effect estimation. The study of Chernozhukov et al. [16] excellently shows how regularization leads to biased treatment effect estimates in an example of a partial linear model. Consequently, it is not possible to perform inference on the results and obtain valid confidence intervals.

Given this drawback, researchers have opted for methods that are aimed directly at estimating treatment effects directly rather than focusing on modeling the outcome with regression.

IPW estimation

Instead of estimating conditional mean functions, propensity scores can be used to overcome the aforementioned bias in this estimate. A propensity score is the probability of a unit having a certain treatment status $T = t$, conditional on covariates X, e.g. $\mathbb{P}(T = t \mid X)$. We will shorten the notation for propensity scores by using

$$p_t(X) = \mathbb{P}(T = t \mid X), \tag{3.17}$$

which will be referred to as the propensity score function.

Propensity scores can be estimated by a wide range of classification models, logistic regression being the most widely known. For now, we will assume that the propensity scores are known.

With the help of propensity scores, one can create unbiased estimates of the potential outcomes by reweighing the observed outcomes by their respective propensity scores. This approach is called inverse propensity weighting (IPW), and the intuition is the following. IPW aims to provide more weight to data points that are more relevant for estimating the causal effect. Each data point is given a weighting such that the data points for which treatment is unlikely, given the covariates, contribute more. For instance, an untreated unit with a very high propensity score is upweighted, because it does a good job representing a counterfactual to a treated unit. The aim is to make a model better able to simulate the outcome of a treated unit, had they not received treatment.

In order for this estimator to be well-defined, we need propensity scores to be strictly positive and have overlap for every possible treatment.

Assumption 4. Positive probability of treatments
For some $\epsilon > 0$ and for all $x \in \mathbb{R}^p$,

$$\epsilon < P(T = 1 \mid X = x) < 1 - \epsilon. \tag{3.18}$$

Note that since T is binary, (3.18) automatically implies that for some $\epsilon > 0$ and for all $x \in \mathbb{R}^p$, $\epsilon < P(T = 0 \mid X = x) < 1 - \epsilon$.

For the real problem setting, this assumption denotes that for all test points $X = x$, e.g. dwellings in the data set, the probability of every possible treatment status should be strictly bounded away from 0. Moreover, when this assumption holds, then for large enough n, there will be enough treated and untreated units near any test point x for local methods to work. Intuitively, this assumption is logical; it would not be reasonable to estimate the effect of increasing the energy efficiency on the transaction price for an energy inefficient dwelling with a very recent construction year, which by strict building codes can not possibly be energy inefficient. However, this implies that our data set needs to comply with the assumption above. How the data set is modified in order to comply with assumption (3.18) is explained in Chapter 4.

Theorem 1. Under assumptions (3.7), (3.9) and (3.18), we have that

$$\mathbb{E}\left[Y^{(t)}\right] = \mathbb{E}\left[\frac{Y \mathbb{1}_{T=t}}{\mathbb{P}(T = t \mid X)}\right]. \tag{3.19}$$

Proof. By the law of total expectation,

$$\begin{aligned}\mathbb{E}\left[Y^{(t)}\right] &= \mathbb{E}_X\left[\mathbb{E}\left[Y^{(t)} \mid X\right]\right]\\ &\text{which, by (3.9) and (3.7)}\\ &= \mathbb{E}_X\left[\mathbb{E}\left[Y^{(t)} \mid X, T=t\right]\right] = \mathbb{E}_X\left[\mathbb{E}\left[Y \mid X, T=t\right]\right]\\ &= \mathbb{E}_X\left[\frac{\mathbb{E}\left[Y\mathbb{1}_{T=t} \mid X\right]}{\mathbb{P}(T=t \mid X)}\right]\\ &= \mathbb{E}_X\left[\mathbb{E}\left[\frac{Y\mathbb{1}_{T=t}}{\mathbb{P}(T=t \mid X)} \mid X\right]\right]\\ &\text{which, by the law of total expectation in the other direction}\\ &= \mathbb{E}\left[\frac{Y\mathbb{1}_{T=t}}{\mathbb{P}(T=t \mid X)}\right]\end{aligned}$$

□

Theorem 1 is crucial, as it shows that counterfactual potential outcomes can be obtained from the distribution of the observed variables (Y, T, X).

Now consider the following setting, in which the goal still is to estimate (3.3). By weighting the observed outcomes by their respective propensity scores, an unbiased estimator for δ can be constructed. This estimator is called the IPW estimator,

$$\hat{\delta}_n^{IPW} := \frac{1}{n}\sum_{i=i}^{n}\Bigg(\underbrace{\frac{Y_i}{\mathbb{P}(T_i=1 \mid X_i)}\mathbb{1}_{T=1}}_{Y_i^{IPW,(1)}} - \underbrace{\frac{Y_i}{\mathbb{P}(T_i=0 \mid X_i)}\mathbb{1}_{T=0}}_{Y_i^{IPW,(0)}}\Bigg). \tag{3.20}$$

From Theorem 1, it immediately follows that the IPW estimator is an unbiased and consistent estimator for the average treatment effect δ under assumptions (3.7), (3.9) and (3.18). That is,

$$\mathbb{E}\left[\hat{\delta}_n^{IPW}\right] = \delta \tag{3.21}$$

and for all $\epsilon > 0$

$$\lim_{n\to\infty} P\left(|\hat{\delta}_n^{IPW} - \delta| > \epsilon\right) = 0 \tag{3.22}$$

Unbiasedeness as in (3.21) is an immediate consequence of Theorem 1. Consistency (3.22) immediately follows from the central limit theorem, because the IPW estimator is a sample average, the variance converges to 0.

In general, it will not be that $Y^{(t)} \sim Y \mid T = t$. However, under the conditions of Theorem 1 this becomes true after conditioning on X, so that $Y^{(t)} \sim Y \mid X, T = t$. Note that in reality, the propensity scores are unknown, and have to be consistently estimated. A large drawback of this method is that for some data points x, the propensity scores, $p_t(x)$, can be small if these samples are unlikely to be treated or untreated. As a result the variance of the estimator can grow very large in these instances. This will be the case when the amount of data points with a certain treatment status $T = t$ is very small in comparison to the group with the different treatment status for some subset of X.

3.1.3. AIPW estimator

The Augmented Inverse Propensity Weighted (AIPW) estimator (or: Doubly Robust estimator) [8, 23, 29] is an estimator for treatment effects when the treatment effect depends on high-dimensional data, or when it can not be satisfactorily estimated by parametric models. The method dates back to work from Robins [29], and is also excellently described in Tsiatis [34]. In this Section the motivation and construction of the Doubly Robust estimator will be substantiated.

The AIPW estimator combines a conditional mean model with the IPW estimator, e.g. the two estimators from (3.14) and (3.15), and the IPW estimator (3.20) in Subsection 3.1.2 into a single estimation technique. This reduces the drawbacks from the previous two approaches, as it reduces the model specification bias from the conditional mean models, and reduces the vulnerability to high variance in the propensity model. In particular, it fits a regression model, but then debiases the model, by applying the inverse propensity approach to the residuals of that model. The AIPW estimator is *Doubly Robust*. This means that the estimator is unbiased if either the conditional mean models as described by (3.14) and (3.15), or the propensity score model from (3.17) is correct. We will prove the doubly robust property later in this Section.

We define

$$\hat{\delta}^{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\underbrace{\mu_1(X_i) + \frac{Y_i - \mu_1(X_i)}{\mathbb{P}(T=1\mid X_i)}\mathbb{1}_{T_i=1}}_{Y_i^{AIPW,(1)}} - \underbrace{\left(\mu_0(X_i) + \frac{Y_i - \mu_0(X_i)}{\mathbb{P}(T=0\mid X_i)}\mathbb{1}_{T_i=0}\right)}_{Y_i^{AIPW,(0)}}\right\}. \tag{3.23}$$

The AIPW estimator again consists of a part that estimates either potential outcome $\{Y^{(0)}, Y^{(1)}\}$.

Additional to the unbiasedness of the AIPW estimator, another desired property can be deduced from equation (3.23). In contrast to the IPW estimator, when the propensity score is close to 0 or 1, the variance of the AIPW estimator will not blow up as much. That is, the terms $\frac{Y-\mu_1(X)}{p(X)}T$ and $\frac{Y-\mu_0(X)}{1-p(X)}(1-T)$ in (3.23) are centered around 0 when the conditional mean models are correct, and hence, cancel each other out on average. This can be seen by

$$\mathbb{E}\left[\frac{Y-\mu_1(X)}{\mathbb{P}(T=1\mid X)}\mathbb{1}_{T=1}\right] = \mathbb{E}\left[\frac{Y^{(1)}-\mu_1(X)}{\mathbb{P}(T=1\mid X)}\mathbb{1}_{T=1}\right] = 0 \qquad \text{and} \tag{3.24}$$

$$\mathbb{E}\left[\frac{Y-\mu_0(X)}{\mathbb{P}(T=1\mid X)}\mathbb{1}_{T=0}\right] = \mathbb{E}\left[\frac{Y^{(0)}-\mu_0(X)}{\mathbb{P}(T=0\mid X)}\mathbb{1}_{T=0}\right] = 0 \tag{3.25}$$

Besides the reduced variance in comparison to the IPW estimator, the AIPW estimator has another desirable statistical property, called double robustness, which will be proven in the following Theorem.

Theorem 2. Assume assumptions (3.7), (3.9) and (3.18) hold. Then the AIPW estimator as defined in (3.23) is a consistent estimator for the average treatment effect δ when either the conditional mean models, or the propensity score model is correct. In other words, when either the estimated conditional mean functions converge in probability to the true conditional mean functions, or the estimated propensity function converges in probability to the true propensity function, the AIPW estimator is consistent. This property is called doubly robustness.

In order to prove the doubly robustness property, we will closely follow Chapter 13 of Tsiatis [34]. We consider a model $\mathbb{P}(T=1\mid X) = p(X,\psi)$ for the propensity score of $T=1$, and models $\mathbb{E}[Y\mid T=0,X] = \mu_0(X,\xi)$ and $\mathbb{E}[Y\mid T=1,X] = \mu_1(X,\xi)$ for the regression models. In these models, ψ and ξ are finite dimensional estimated parameters that describe the propensity model and regression models, respectively. We use the convention that $\hat{\psi}_n \xrightarrow{p} \psi^*$ and $\hat{\xi}_n \xrightarrow{p} \xi^*$ to denote that under suitable regularity conditions, the estimators will converge, whether the model is correct or not. When the model is correct, $\psi^* = \psi_0$ and $\xi^* = \xi_0$ denote that the models converge to the truth. We wish to show the AIPW estimator is consistent when either $\psi^* = \psi_0$ or $\xi^* = \xi_0$.

Proof. Using the definitions of the models stated above, the AIPW estimator of (3.23) can be rewritten as

$$\begin{aligned}\hat{\delta}^{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\frac{T_iY_i}{p(X_i,\hat{\psi}_n)} - \frac{[T_i - p(X_i,\hat{\psi}_n)]\mu_1(X_i,\hat{\xi}_n)}{p(X_i,\hat{\psi}_n)} \\ &- \frac{(1-T_i)Y_i}{1-p(X_i,\hat{\psi}_n))} - \frac{[T_i - p(X_i,\hat{\psi}_n)]\mu_0(X_i,\hat{\xi}_n)}{1-p(X_i,\hat{\psi}_n)}\Bigg).\end{aligned}$$

Because the estimator is a sample average, $\hat{\delta}^{AIPW}$ converges in probability to

$$\mathbb{E}\left[\frac{TY}{p(X,\psi^*)} - \frac{[T-p(X,\psi^*)]\mu_1(X,\xi^*)}{p(X,\psi^*)} - \frac{(1-T)Y}{1-p(X,\psi^*))} - \frac{[T-p(X,\psi^*)]\mu_0(X,\xi^*)}{1-p(X,\psi^*)}\right]. \tag{3.26}$$

Using (3.7), note that

$$\frac{TY}{p(X,\psi^*)} = \frac{TY^{(1)}}{p(X,\psi^*)} = Y^{(1)} + \frac{[T-p(X,\psi^*)]Y^{(1)}}{p(X,\psi^*)}. \tag{3.27}$$

And similarly,

$$\frac{(1-T)Y}{1-p(X,\psi^*)} = Y^{(0)} + \frac{[T-p(X,\psi^*)]Y^{(0)}}{1-p(X,\psi^*)}. \tag{3.28}$$

Substituting (3.27) and (3.28) back into (3.26) to get

$$\mathbb{E}\left[Y^{(1)} - Y^{(0)}\right] \tag{3.29}$$

$$+\mathbb{E}\left[\frac{[T - p(X,\psi^*)][Y^{(1)} - \mu_1(X,\xi^*)]}{p(X,\psi^*)}\right] \tag{3.30}$$

$$+\mathbb{E}\left[\frac{[T - p(X,\psi^*)][Y^{(0)} - \mu_0(X,\xi^*)]}{1 - p(X,\psi^*)}\right]. \tag{3.31}$$

Now notice that (3.29) is the definition of the average treatment effect as defined in (3.3). So, in order to prove double robustness, it is sufficient to show that (3.30) and (3.31) equal 0 when either model is correct, e.g. $\psi^* = \psi_0$, or $\xi^* = \xi_0$. Let us first consider the case where the propensity model is correct, so that $\psi^* = \psi_0$. Using the law of total expectation we can write (3.30) as

$$\begin{aligned}
&\mathbb{E}_{(X,Y^{(1)})}\left[\mathbb{E}\left[\frac{[T - p(X,\psi_0)][Y^{(1)} - \mu_1(X,\xi^*)]}{p(X,\psi_0)} \mid X, Y^{(1)}\right]\right] \\
&= \mathbb{E}_{(X,Y^{(1)})}\left[\frac{[\mathbb{E}[T \mid X, Y^{(1)}] - p(X,\psi_0)][Y^{(1)} - \mu_1(X,\xi^*)]}{p(X,\psi_0)}\right] \\
&= \mathbb{E}_{(X,Y^{(1)})}\left[\frac{[p(X,\psi_0) - p(X,\psi_0)][Y^{(1)} - \mu_1(X,\xi^*)]}{p(X,\psi_0)}\right] = 0.
\end{aligned}$$

In the last equality, we used that $\mathbb{E}[T \mid X, Y^{(1)}] = p(X,\psi_0)$ by the unconfoundedness assumption (3.9). Analogously, (3.31) equals 0 when $\psi^* = \psi_0$.
Now we will look at the situation where $\xi^* = \xi_0$. In other words, the regression models are correctly specified, however, the propensity models may not be. Using a similar approach as above, we can rewrite (3.30) using the law of total expectation, as

$$\begin{aligned}
&\mathbb{E}_{(X,T)}\left[\mathbb{E}\left[\frac{[T - p(X,\psi^*)][Y^{(1)} - \mu_1(X,\xi_0)]}{p(X,\psi^*)} \mid X, T\right]\right] \\
&= \mathbb{E}_{(X,T)}\left[\frac{[T - p(X,\psi^*)][\mathbb{E}\left[Y^{(1)} \mid X, T\right] - \mu_1(X,\xi_0)]}{p(X,\psi^*)} \mid X, T\right] \\
&= \mathbb{E}_{(X,T)}\left[\frac{[T - p(X,\psi^*)][\mathbb{E}\left[Y^{(1)} \mid X, T = 1\right] - \mu_1(X,\xi_0)]}{p(X,\psi^*)} \mid X, T\right] \\
&= \mathbb{E}_{(X,T)}\left[\frac{[T - p(X,\psi^*)][\mu_1(X,\xi_0) - \mu_1(X,\xi_0)]}{p(X,\psi^*)} \mid X, T\right] = 0.
\end{aligned}$$

Where we used that

$$\mathbb{E}\left[Y^{(1)} \mid T, X\right] = \mathbb{E}\left[Y^{(1)} \mid T = 1, X\right] = \mathbb{E}[Y \mid T = 1, X] = \mu_1(X,\xi_0), \tag{3.32}$$

which follows from unconfoundedness (3.9), consistency (3.7) and the definition of $\mu(T = 1, X, \xi_0)$, respectively. Analogous calculations show that term (3.31) is equal to 0 when $\xi^* = \xi_0$. Hence, $\hat{\delta}^{AIPW}$ as introduced in (3.23) is consistent for the average treatment effect δ when either the regression model or the propensity model is correctly specified. This concludes the proof of the double robustness property. □

So far, in this Section, we focused only on the estimation of the average treatment effect, δ. Moreover, it was assumed that either the conditional mean models and propensity models could be correctly estimated with a parametric model. This setting will be extended in the next Sections.

In Section 3.2 it is shown how the conditional mean models and propensity models are estimated. In Section 3.3, the theory from previous Sections is combined to show how conditional average treatment effects can be estimated. Moreover, we will present the theoretical basis for valid inference using estimated mean and propensity score functions.

3.2. Estimating mean and propensity score functions

In the previous Section, the assumption was made that the conditional mean functions $\{\mu_1(X), \mu_0(X)\}$, as well as the propensity score function $p_t(X)$ were known, or at least one of either could be correctly estimated with parametric models. This, however, is barely ever a valid assumption, and neither is that the case in our problem. Nevertheless, it is possible to estimate the functions $\{\mu_1(X), \mu_0(X)\}$ and $p_t(X)$ by some arbitrary machine learning methods, which, under some conditions, provide valid inference. This will be elaborated in Section 3.3.

In the following Subsections, methods for estimating the conditional mean functions, (3.14) and (3.15), as well as the propensity score function (3.17) will be elaborated.

3.2.1. Random Forests

In order to estimate the conditional mean models of (3.14) and (3.15), random forest [10] regressor will be used. For estimating the propensity scores model (3.17), a random forest classifier will be used. In this Subsection these methods will be elaborated.

A random forest is a machine learning method that can be used for classification and regression tasks. A random forest is an extension of a decision tree model. Unlike a decision tree model, which models a problem with a single decision tree, random forests make use of many trees, using a technique called bagging [9]. Bagging creates an ensemble model, in this case a random forest, by averaging the predictions of many different instances of the same underlying model, in this case a decision tree. Doing so improves the stability and accuracy, decreases variance and helps reduce overfitting [9].

Before the random forest is introduced, the theory and construction of a decision tree model will be explained. When the workings of a decision tree model are clear, the random forest is a relatively simple extension.

3.2.2. Decision Tree Learner

In this Subsection, the theory behind a decision tree learner is elaborated. Decision tree learners can be used for either regression tasks and classification tasks. We will start by focusing on using decision trees for a regression problem, and then make the translation to a classification problem afterwards.

In a regression setting with a one-dimensional outcome variable $Y \in \mathbb{R}$, the goal is to estimate a function $f : X \to Y$ from features in a p-dimensional feature vector, $X \in \mathbb{R}^p$. A decision tree aims to learn this function by constructing a decision tree with the use of the data.

We will start by showing a simple example of a grown decision tree, and afterwards explain the concept of creating such a decision tree.

Example: Decision Tree for regression Imagine that we would like to predict house prices $Y \in \mathbb{R}$ from two features, $Area \in [50, 200]$ and $Urbanity \in \{1, 2, 3, 4, 5\}$. Here, *Area* is the living area of a dwelling, and *Urbanity* is a categorical variable that represents the degree of urbanity of the location of the dwelling. We generate $n = 500$ data points following the distributions as stated below,

$$\begin{aligned} Y &= 100000 + 2000 * Area + 40000 * Urbanity + \epsilon \\ Area &\sim U(50, 200) \\ Urbanity &\sim U\{1, 5\} \\ \epsilon &\sim \mathcal{N}(0, 50000). \end{aligned}$$

Here $U(a, b)$ and $U\{a, b\}$ denote the uniform and discrete uniform distributions, respectively, and $\mathcal{N}(\mu, \sigma)$ denotes the normal distribution. A decision tree is trained on the generated data points, using the methods explained in the previous paragraphs. We take *Area* and *Urbanity* as features, and Y as the outcome variable. A maximum depth of 2 is assigned to the tree model, making the splits in the decision tree stop when a depth of 2 is reached. The full grown tree is visualized in Figure 3.1. The decision process for an individual data point is shown in the dotted trajectory. For a new point, at every depth of the tree the decision rule is followed until a terminal node, e.g. an 'end' node, is reached. At the terminal node, the prediction is equal to the mean outcome value in that terminal node.

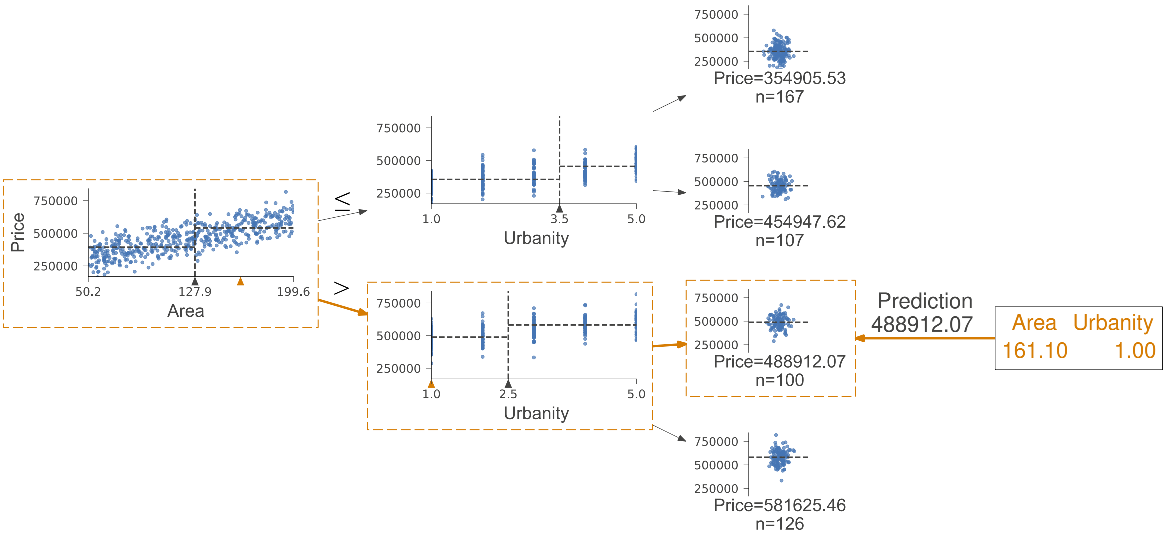


Figure 3.1: A decision tree prediction visualization. The individual scatter plots show the distribution of a feature versus the outcome variable of the remaining data points. A new data point with values 161.1 and 1.0 for the *Area* and *DegreeofUrbanity*, respectively, is predicted to have a Y value of 488,912.07. At each decision node, the splitting criterion are followed until a termination node is reached, as can be seen by the dotted trajectory. Figure created with open source python library graphviz [20].

Splitting criterion

Imagine we have $i = 1,...,n$ samples of training data, where the i^{th} sample consists of a p-dimensional feature vector and outcome pair, (X_i, Y_i). The problem of creating a decision tree that can map a feature vector X to an estimated outcome value $\hat{Y}$, boils down to finding the optimal features and corresponding feature values to split the data on in each node of the tree.

In particular, we try to find the feature and corresponding feature value so that when we split the data according to this feature and value, the sum of within-group variances has decreased the most. This is done in a brute-force manner, which is shown in pseudocode in Algorithm 1. The metric that is optimized in every split generally depends on whether the decision tree is used for regression or for classification.

Algorithm 1 Find the optimal feature and feature value for the next split in a decision tree

```
function FindBestSplit(X, Y)
    define $MSE(Y) = \sqrt{\sum_{i=1}^{n}(Y_i - mean(Y))^2}$        ▷ Define a metric to optimize
    $bestfeature \leftarrow None$        ▷ Initialize the goal variables
    $bestvalue \leftarrow None$
    $smallest\_MSE \leftarrow \infty$
    for feature in X do        ▷ Try all features one at a time
        $sortedvalues \leftarrow Sort(feature.values)$        ▷ Sort the values of the feature
        for value in sortedvalues do        ▷ Try all values in feature X
            $left = Y$[where $feature.values \leq value$]        ▷ Split units based on current feature and value
            $right = Y$[where $feature.values > value$]
            $total\_MSE = MSE(left) + MSE(right)$        ▷ Calculate total MSE of this split
            if $total\_MSE < smallest\_MSE$ then        ▷ If total MSE is the smallest so far, save variables
                $bestfeature \leftarrow feature$
                $bestvalue \leftarrow value$
                $smallest\_MSE \leftarrow total\_MSE$
            end if
        end for
    end for
    return bestfeature, bestvalue
end function
```

Stopping criterion

For building the decision tree model, we recursively find the best split for all levels of the decision tree, starting at the root of the tree with all the data. However, the last concept that is necessary to ensure the decision tree will grow properly, is a criterion when the splitting stops. If the splitting is not stopped in time, the tree will keep growing and overfit on the training data, making the model less accurate for predictions on unseen data. The most commonly used stopping criterion for a decision tree is to assign a maximum depth. Doing so makes the tree stop growing when the maximum depth has been reached. Another stopping criterion that is often used is the minimum samples that need to be left in a certain node in order to continue splitting.

Prediction

Decision trees predict the outcome values by simply taking the mean of all outcome values in the terminal node that the feature values of a data point fall in to. When we want to predict the outcome value for point x, we simply check which terminal leaf R_m the point x falls into. Then we predict a constant value c_m, which is the mean of training data Y that are in leaf R_m. In mathematical terms,

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{1}\{x \in R_m\} \tag{3.33}$$

$$\text{where } c_m = \frac{1}{n_m}\sum_{i}^{n} Y_{i,train}\mathbb{1}\left\{X_{i,train} \in R_m\right\} \text{ and } n_m = \sum_{i=1}^{n} \mathbb{1}\left\{X_{i,train} \in R_m\right\}. \tag{3.34}$$

Decision Tree for classification

A decision tree for a classification task works in a very similar manner as a decision tree for a regression task. A classification task, however, models a discrete outcome. In our problem, the classification task is to model a binary outcome, e.g. $T \in \{0,1\}$, and estimate the probability of both binary outcomes.

Instead of creating the decision tree by splitting based on the minimum mean squared error, the splitting criterion used is log loss. For a setting with a binary outcome, log loss per data point is the negative log-likelihood of the classifier, given the true label. Hence, in mathematical terms, the total negative log loss is defined as

$$L(t,p)_{logloss} = -\frac{1}{n}\sum_{i=1}^{n}(t_i log(p_i) + (1-t_i)log(1-p_i)), \tag{3.35}$$

where p_i is the predicted probability of unit i belonging to class 1, and $t_i \in \{0,1\}$ is unit i its true class. For the classification step, we set the splitting criterion as minimizing the negative log loss, as this optimizes the predicted probabilities, rather than the predicted classes. The predicted probabilities will be used in the doubly robust estimator, so optimizing prediction for this metric is in line with our goal.

Random Forest

A random forest is a bootstrap aggregated (bagged) ensemble of single decision trees instances. Bootstrapping is the random sampling with replacement of the training sample, with the goal of obtaining $b = 1,...,B$ samples of training data. On all these training samples, a different instance of a decision tree T_b is trained. Subsequently, these B individual decision trees are combined together, which is the aggregation step. After these steps, prediction on unseen data is performed by a majority vote or averaging predictions from individual decision trees, for classification and regression, respectively.

Bootstrapping and aggregating in this manner creates a more robust model. The idea of bootstrapping is to average many noisy but approximately unbiased models, in order to reduce variance. Decision trees are great candidates for bagging, as they can capture complex structures in the data and have relatively low bias. Since the variance is relatively large in a decision tree, averaging can greatly improve the model.

Bagged trees are the average of B decision trees, each with expectation m and variance σ^2. Bagged decision trees can be seen as random variables that are identically distributed, but not necessarily independent, because the data used to train each decision tree comes from the same sample. If B random variables are identically distributed with positive pairwise correlation ρ, the variance of the average is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{3.36}$$

Proof. To see this, assume that we have variables X_i, for $i = 1, ..., B$, where $\mathbb{E}[X_i] = m$ and $Var(X_i) = \sigma^2$, where for $i \neq j$, X_i and X_j are correlated with correlation coefficient ρ. We are interested in

$$Var\left(\frac{1}{B}\sum_{i=1}^{B} X_i\right).$$

Recall that

$$Var(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2. \tag{3.37}$$

Hence,

$$Var\left(\frac{1}{B}\sum_{i=1}^{B} X_i\right) = \frac{1}{B^2} Var\left(\sum_{i=1}^{B} X_i\right) = \frac{1}{B^2}\left(\mathbb{E}\left[\left(\sum_{i=1}^{B} X_i\right)^2\right] - \mathbb{E}\left[\sum_{i=1}^{B} X_i\right]^2\right). \tag{3.38}$$

For the second expectation of the right-hand side of (3.38), note that simply

$$\mathbb{E}\left[\sum_{i=1}^{B} X_i\right]^2 = (Bm)^2 = B^2 m^2.$$

For the first expectation of the right-hand side of (3.38), recall that

$$\left(\sum_{i=1}^{B} X_i\right)^2 = \sum_{i,j=1}^{B} X_i X_j.$$

And thus we have that

$$\mathbb{E}\left[\left(\sum_{i=1}^{B} X_i\right)^2\right] = \sum_{i,j=1}^{B} \mathbb{E}[X_i X_j]. \tag{3.39}$$

For the definition of the correlation coefficient, we have that

$$\frac{\mathbb{E}[(X_i - m)(X_j - m)]}{\sigma^2} = \rho \quad \text{for i} \neq \text{j and } \rho > 0.$$

With straightforward calculations we get

$$\mathbb{E}[X_i X_j] = \rho\sigma^2 + m^2.$$

For $i = j$, from (3.37) we have that $\mathbb{E}[X_i X_i] = \sigma^2 + m^2$. Hence

$$\mathbb{E}[X_i X_j] = \begin{cases} \sigma^2 + m^2, & \text{if } i = j \\ \rho\sigma^2 + m^2, & \text{if } i \neq j \end{cases}.$$

Hence, the right-hand side of (3.39) can be expanded to

$$\begin{aligned} \sum_{i,j=1}^{B} \mathbb{E}[X_i X_j] &= B\mathbb{E}[X_i^2] + (B^2 - B)\mathbb{E}[X_i X_j] \\ &= B(\sigma^2 + m^2) + (B^2 - B)(\rho\sigma^2 + m^2) \\ &= B\sigma^2 + B^2\rho\sigma^2 + B^2 m^2 - B\rho\sigma^2. \end{aligned}$$

Then, we can substitute the findings in (3.38), to find that

$$\begin{aligned} Var\left(\frac{1}{B}\sum_{i=1}^{B} X_i\right) &= \frac{1}{B^2}\left(B\sigma^2 + B^2\rho\sigma^2 + B^2 m^2 - B\rho\sigma^2 - B^2 m^2\right) \\ &= \frac{\sigma^2}{B} + \rho\sigma^2 - \frac{\rho\sigma^2}{B} \\ &= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \end{aligned}$$

□

Consequently, if a sufficient amount of trees B is used, the second term in (3.36) goes to 0. In order to additionally reduce the pairwise correlation between trees, ρ, random forests use a different random subset of features for the splitting steps in each of its decision trees. The intuition for this is the following. Even when building many different decision trees with sampled data, the same features will often be important in every decision tree. As a result, the splits on top of the decision tree are generally done with the same features in each tree. This is a result of the fact that those features can generally describe the outcome very well, and so are prime targets for the splitting steps. This, however, makes the individual trees more correlated. When the bootstrapped trees are highly correlated, their ensemble will be less powerful as proven above. Hence, a random subset of features will be used for each splitting step, in order to reduce correlation. How large this random subset should be is determined by hyperparameter optimization, explained in Subsection 3.2.3. Thus, while decision trees are normally highly vulnerable to noise in the training data overfitting, a random forest generally performs better.

3.2.3. Hyperparameter optimization

As illustrated in the previous Subsections, there are some settings within random forests that can be tweaked. These settings are often referred to as hyperparameters. For the random forest, we will be tweaking three different hyperparameters. The method for selecting the optimal hyperparameters are grid search in combination with K-fold cross-validation. This can be seen as a brute-force search of the hyperparameters that provide the best score on some defined metric on unseen data. This approach will be elaborated below.

In order to define the hyperparameters for our models, K-fold cross-validation will be used. K-fold cross-validation divides the available data into K folds. Then it trains a model on $K-1$ folds, and afterwards evaluates the trained model on the holdout fold. In this way, K instances of the model with the same hyperparameters will be trained on separate $K-1$ folds of the data. The score will be the average score of these models on the holdout fold. This whole process is repeated for every combination of hyperparameters, and is described in steps below.

1. Split the available training data with size n into K folds of size n/k.
2. Train the model on $K-1$ folds.
3. Predict the values of the holdout fold.
4. Evaluate the model by comparing the predicted values of the holdout fold with their true values by a chosen metric.
5. Repeat steps 2-4 with a different holdout fold until every fold has been the holdout fold.
6. Calculate the average score of the chosen metric over all holdout folds.
7. Repeat steps 2-6 for every combination of hyperparameter values.

When all the steps are completed, the model with the best average score is picked. We will use 5-fold cross-validation for optimizing the hyperparameters. The hyperparameters that are optimized in this manner are the the maximum depth of the underlying trees, the minimum samples in the leaf when splitting stops, and the amount of features that are used for splitting.

3.2.4. Feature importance evaluation

Evaluation of feature importance allows us to learn which features are important for the model. Permutation importance estimates the variable importance by the mean increase in relative error over cross-validated predictions on the test set, when one of the values in a variable are randomly shuffled in the test set. As a result, for the test set, this variable has no predictive power. The variable that makes the error increase the most when randomly shuffled, can be ranked as most important for the model. The steps done for calculating permutation importance for a random forest regressor, with the mean squared error as evaluation metric, are summarized below.

1. Train the random forest on train data.
2. Measure the mean squared error on test data, D_{test}, say MSE_{base} as reference.

3. For each repetition k in $1,...,K$: Randomly shuffle values from feature j to generate a corrupted version of the test set, $\tilde{D}_{j,k}$. Then measure the mean squared error on the corrupted test set, $MSE_{k,j}$.

4. Importance of feature j, $i_j = \frac{1}{K}\sum_{k=1}^{K} MSE_{k,j} - MSE_{base}$.

5. Repeat steps 3-4 for all features $j = 1,...,J$.

For a random forest classifier, the steps are similar. The difference is that instead of the mean squared error, the log loss as defined in (3.35) is measured in steps 2 and 3. In step 4, the importance is of feature j is calculated as $i_j = \frac{1}{K}\sum_{k=1}^{K} L_{k,j} - L_{base}$. Here, L_k, j is the kth repetition of the measured log loss on the corrupted test set as defined similarly for the MSE in step 3. L_{base} is the reference log loss on the test set.

The benefit of estimating permutation importance is that it directly measures the importance of every single feature by evaluating the predictive performance. A drawback is that permutation importance is computationally heavy, and thus often performed by only removing features one-by-one. Hence, standard permutation importance does not take the relations between features into account. As a result, correlated features heavily impact the permutation importance. For instance, if two important features are heavily correlated, removing either does not impact model performance much, and hence the individual feature will not be evaluated as important. However, these features are in fact both important for the model, because when they would both be permuted, model performance would get much worse. There are options for dealing with this drawback, however, take into account that feature importance generally only provides a rough indication of the importance of features. It is difficult to give a clear interpretation of the absolute values of the estimated feature importance.

It is important to note that feature importance does not necessarily explain the predictive value of a feature, but rather the importance of the feature for the model. Feature importance may be very different depending on the goal of the model. When modeling the price of a dwelling, the important features could be very different than when predicting the probability of treatment. Hence, feature importance will be evaluated for all models that are used in the AIPW estimator.

3.3. AIPW estimator and evaluation

In recent literature, many advancements have been made in proving statistical properties of estimators that use machine learning estimates of functions $\{\mu_1(X), \mu_0(X)\}$ and $p_t(X)$ as plug-ins to doubly robust estimators [4, 15, 16, 27, 36]. In particular, Chernozhukov et al. [16] analyzes the case of estimating $\delta(x)$ as defined in (3.4) when it is constant or low-dimensional and linear, while allowing the mean and propensity models to be high-dimensional. The study introduces valid methods for statistical inference and construction of confidence intervals. In particular, the study constructs a method it calls "Double Machine Learning", and shows for a point $x \in X$, point estimates can be made that concentrate in an $N^{-\frac{1}{2}}$-neighborhood of the true parameter values and are approximately unbiased and normally distributed. This theory will be used to construct the linear doubly robust estimator that will be elaborated in Subsection 3.3.1.

Wager & Athey [4] and Oprescu [27] develop similar results as Chernozhukov. However, instead of a linear parametric form of $\delta(x)$, they focus on a random forest based estimator for $\delta(x)$. In a recent study, Athey [4] slightly modified the original random forest by applying honesty and subsampling, explained in detail later this Section, which allows to make valid statistical inference using random forests. In particular, Athey shows that predictions by the modified random forest are asymptotically Gaussian and unbiased. Specifically, for a test point x,

$$\frac{(\hat{\delta}(x) - \delta(x))}{\sqrt{Var(\hat{\delta}(x))}} \rightarrow \mathcal{N}(0,1), \tag{3.40}$$

under conditions that will be discussed in Subsection 3.3.2.

In this Section we will elaborate how the estimates from Section 3.2 are used in the AIPW estimator. Moreover, we present how we construct estimates for the average and conditional average treatment effect.

3.3.1. Linear Doubly Robust Estimator

Recall that the AIPW estimator for the average treatment effect is defined as

$$\hat{\delta}^{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\underbrace{\mu_1(X_i) + \frac{Y_i - \mu_1(X_i)}{\mathbb{P}(T=1 \mid X_i)}) \cdot \mathbb{1}_{T_i=1}}_{Y_i^{AIPW,(1)}} - \underbrace{\left(\mu_0(X_i) + \frac{Y_i - \mu_0(X_i)}{\mathbb{P}(T=0 \mid X_i)} \cdot \mathbb{1}_{T_i=0}\right)}_{Y_i^{AIPW,(0)}}\right\}.$$

Using the theory presented in Section 3.2, we can estimate models for

$$\begin{aligned} \mathbb{E}[Y \mid T=t, X] &= \mu_t(X) \qquad \text{and} \\ \mathbb{P}(T=t \mid X) &= p_t(X) \end{aligned}$$

in the first stage, to obtain $\hat{\mu}_t(X)$ and $\hat{p}_t(X)$. For constructing these models, a random split of half the data is used.

For the other half of data, estimates $Y_i^{AIPW,(1)}$ and $Y_i^{AIPW,(0)}$ for $i = 1, ..., N$ are created using the estimated functions from the first stage, where

$$Y^{AIPW,(t)} = \hat{\mu}_t(X) + \frac{Y - \hat{\mu}_t(X)}{\hat{p}_t(X)} \cdot \mathbb{1}_{T=t}.$$

The linear doubly robust estimator assumes a linear parametric conditional average treatment effect. In other words, the assumption is made that the treatment effect is a linear function of the variables X. The linear doubly robust estimator performs an unregularized linear regression on the differences between $Y_i^{AIPW,(1)}$ and $Y_i^{AIPW,(0)}$, e.g. the linear doubly robust estimator regresses $Y^{AIPW,(1)} - Y^{AIPW,(0)}$ on X. Mathematically, the setting is as follows,

$$\begin{aligned} & Y^{AIPW,(1)} - Y^{AIPW,(0)} = \delta(X) + \epsilon, \\ \text{where} \quad & \delta(X) = X\beta, \\ \text{and} \quad & \epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right). \end{aligned} \tag{3.41}$$

This model is estimated with OLS and hence valid confidence intervals can be constructed via asymptotic normality arguments.

3.3.2. Forest Doubly Robust Estimator

The forest doubly robust estimator is similar to the random forests regression as explained previously in Section 3.2. Only now, the target variable is again $Y^{AIPW,(1)} - Y^{AIPW,(0)}$.

However, in order to perform valid statistical inference, two main alterations have to be done for the random forest as explained in Section 3.2. First of all, instead of training the individual decision trees on bootstrapped training data, the individual trees have to be trained on subsampled training data. Subsampled training data are data points that are selected from our data set without replacement, instead of with replacement.

The second alteration is that the individual decision trees have to be honest, as defined by Athey & Wager[4]. A decision tree is honest if a different, disjoint sample is used for finding optimal splits with Algorithm 1 from Section 3.2 and for the final leaf-based response calculation.

Taking these alterations into account, the approach for constructing a decision tree is now as follows.

1. Draw a random subsample of size s from the training data set without replacement. Divide the random subsample into two disjoint sets of size $|\mathbb{I}| = \lfloor s/2 \rfloor$ and $|\mathbb{J}| = \lceil s/2 \rceil$.
2. Construct a decision tree. Use set $\mathbb{I}$ in Algorithm 1 to find optimal splits, where splits are made that maximize the variance of $Y_i^{AIPW,(1)} - Y_i^{AIPW,(0)}$ for $i \in \mathbb{I}$.
3. Calculate leaf-based averages of $Y_j^{AIPW,(1)} - Y_j^{AIPW,(0)}$ where $j \in \mathbb{J}$ for all leafs in the constructed decision tree.

Remark In the second step, we split so that the variance of $Y_i^{AIPW,(1)} - Y_i^{AIPW,(0)}$ is maximized. The motivation for this splitting criterion is the following. In regression trees we generally minimize the mean squared error of predictions. Regression trees compute predictions $\hat{Y}(X_i)$ by averaging the training outcomes in a leaf and hence we can verify that

$$\sum_{i \in \mathbb{I}} (\hat{Y}(X_i) - Y_i)^2 = \sum_{i \in \mathbb{I}} Y_i^2 - \sum_{i \in \mathbb{I}} \hat{Y}(X_i)^2. \tag{3.42}$$

Consequently, the split that minimizes the mean squared error is equivalent to the split that maximizes the variance of $\hat{Y}(X_i)$ for $i \in \mathbb{I}$. Note that for a decision tree, $\sum_{i \in \mathbb{I}} \hat{Y}(X_i) = \sum_{i \in \mathbb{I}} Y_i$, and as a result, maximizing the variance is equivalent to maximizing the sum of $\hat{Y}(X_i)^2$. Concluding, by maximizing the variance of $Y_i^{AIPW,(1)} - Y_i^{AIPW,(0)}$ in step 2 we hope to minimize the mean squared error of the treatment effect.

When the random forest is built of trees that are honest and use subsampled data points, the predictions made by the random forest are asymptotically Gaussian and unbiased. For a test point x,

$$\frac{(\hat{\delta}(x) - \delta(x))}{\sqrt{Var(\hat{\delta}(x))}} \to \mathcal{N}(0,1).$$

Moreover, this asymptotic variance can be estimated with the infitesimal jackknife for random forests developed by Wager, Hastie and Efron [37].

4

Data analysis

4.1. Data Preprocessing

In this Section the steps made to transform the raw data sets into data sets that can be used for analysis are presented. The requirements for the final data sets are the following.

1. The impact of location on the price of a house is limited.
2. Only numerical or categorical variables are available in the data set.
3. The amount of features is limited, however, have good predictive power of the price or the energy efficiency.
4. Enough transactions are available for further analysis.

4.1.1. Data Sources

Starting out, we have available a raw data set of transactions of dwellings, containing $n = 3,964,318$ samples of transactions and $p = 40$ predictors, consisting of characteristics of the dwellings and information regarding the transaction. This sample is obtained from three sources; the Kadaster (a Dutch land registry agency), Funda (a house marketplace) and the Basisadministratie Adressen en Gebouwen (Dutch property registration). The total data set contains over 99% of all publicly sold houses in the Netherlands in the period of January 1993 until March 2020.

In 2015, a new standard for estimating the Energy Performance Coefficient of dwellings and providing dwellings with Energy Performance Coefficient labels (EPC labels) was instituted in the Netherlands, called NEN7120. This method is vastly different from the standards used to measure Energy Performance before 2015, as the NEN7120 no longer requires the judgment of an expert to examine the energy performance of a dwelling, merely the EPC label will be estimated from important dwelling characteristics provided by the house owner. Moreover, the requirements for labels have been made more strict, meaning that dwellings which had a certain EPC label provided by older methods, might obtain a worse label by the NEN7120 standard. Lastly, from 2015 onwards, a dwelling is obliged to have an EPC label before it is allowed to be sold, while before 2015 EPC labels were not required. As a result of these policies, it is difficult to compare transactions from dwellings sold before 2015 with transactions of dwellings sold after 2015, which used a different method for measuring Energy Performance. To deal with this difficulty, all transactions from before 2015 are removed from the data set. Methods used for EPC label classification before 2015 that period are vastly different from methods used after 2015. Moreover, the optionality of EPC labels before 2015 can lead to sample selection bias, meaning it might be the case that only owners of green houses opted to obtain an EPC label before selling the house. Removing transactions before 2015 leaves a set of $n = 909,945$ transactions in the period of January 2015 until March 2020.

Only 31.4% of transactions in this raw data set contain an EPC label. To handle this problem, another available data source from the Rijksdienst voor Ondernemend Nederland (RVO) is used. This data source contains all registrations of EPC labels for dwellings in the Netherlands. This data contains among other things the EPC label, EPC label registration date, method used for determining the EPC label and the address. From this data source, the registrations that did not comply with the NEN7120 standard are removed. Afterwards, the transactions data set and EPC label data set are merged on address, creating a new variable EPC_label, which contains the latest registered EPC label before the sale date for the corresponding address. All transactions that do not have an EPC label registered by the NEN7120 standard are removed from our sample. Now the data set contains $n = 802,874$ transactions of dwellings that have an EPC label conform the NEN7120 standard at the date of sale.

Throughout the next Subsections, the steps to modify the raw data sets in order to obtain one cleaned up data set that can be used for analysis are addressed. The problems one has to deal with to obtain such a data set are presented, and it is shown how these problems are handled one by one.

4.1.2. EPC labels

In the Netherlands, the energy efficiency of dwellings is communicated to the market through the use of EPC labels. There are a total of 11 different EPC labels ranging from A++++ to G, where A++++ denotes a highly energy efficient dwelling, and G denotes a highly energy inefficient dwelling. Transactions of dwellings with EPC labels ranging from A++++ to A+ are extremely rare, and do not occur at all in our data set. Hence, only the 7 EPC labels ranging from A to G are considered.

The impact of an increased EPC labels on the price of a dwelling is expected to be very small in relation to other, more important variables. As a result, it will be extremely difficult to estimate the effect of increasing the energy efficiency by 1 EPC label, as this effect is expected to be very small. Hence, the of merging transactions with comparable labels into the same groups is considered.

There are two main reasons for merging individual EPC labels into groups. Firstly, when merging groups of EPC labels together, the jump from one energy efficiency group to another becomes larger, making the corresponding expected premium on the price larger as well. This will make the energy efficiency relatively more important for modeling the price of a dwelling compared to other variables. As a result, the effect of increasing the energy efficiency on the price will be stronger, hence easier to estimate.

Another reason for merging groups of dwellings with similar labels is to have more data available. With a limited amount of data, the confidence bounds of the estimated effect of an increased EPC label on the price will be wider. The confidence bounds might become so wide that there are no meaningful conclusions to draw. This problem is amplified when the expected jump of impact between two labels is small.

A disadvantage of this approach is the fact that it will no longer be possible to estimate the effect of improving energy efficiency within an EPC label group. Moreover, the definition of the improved price due to an increased EPC label becomes wider. A jump from label C to B will be treated in the same manner as a jump from label D to A. This will cause the estimated results to lose a level of detail, however, it does make results more general.

Considering the advantages outweigh the disadvantages, the choice is made to group together transactions with different, but comparable EPC labels. The next step is to determine which specific EPC labels these groups should consist of.

Table 4.1: Relevant descriptive statistics per EPC label. The values are the shown as mean ± standard deviation

Labels	Label A	Label B	Label C	Label D	Label E	Label F	Label G
Sample size (N)	92,938	83,046	162,330	69,321	50,630	47,283	45,682
Construction Year	2002 ± 30	1991 ± 13	1975 ± 18	1957 ± 26	1946 ± 29	1933 ± 35	1925 ± 33
Dwelling Area (m^2)	143.8 ± 44.4	135.3 ± 40.5	124.3 ± 34.9	126.6 ± 43.9	117.2 ± 43.1	123.5 ± 52.2	128.8 ± 61.6
Price per m^2 (€ / m^2)	2,386 ± 714	2,260 ± 667	2,105 ± 695	2,212 ± 881	2,374 ± 997	2,488 ± 1,100	2,353 ± 1,151

From Table 4.1, one can see that there exist some large differences in characteristics between transactions of dwellings with different labels. By far the most number of transactions are from dwellings with EPC label C. This group also has the lowest mean price per m^2. Moreover, the construction year of a dwelling is highly correlated with the energy efficiency. The sample of dwellings that are energy inefficient tend to be more heterogeneous throughout their characteristics, which can be seen from the somewhat higher standard deviations in the construction year, area and price. This heterogeneity was also present in most other variables, which are not present in this Table.

The grouping of EPC labels has two main goals. First, the EPC label groups should be similar in size after merging. Second, the EPC label groups should have small within-group differences, and large differences from other groups.

The group of transactions with EPC labels C are very different from both the most energy efficient dwellings, as well as the most energy inefficient dwellings. Considering this is by far the largest sample, adding these transactions to either of the extreme groups will make the estimated effect of increasing energy efficiency on the price biased. To conclude, a middle group is needed containing dwellings that are neither energy efficient nor energy inefficient.

Dwellings with EPC labels C or D are put into this middle group, called 'Moderately Energy Efficient'. Transactions of dwellings with EPC labels A and B are put together in group 'Energy Efficient, and transactions of dwellings with EPC labels E, F and G are put together in group 'Energy Inefficient'.

4.1.3. Dwelling Types

In our data set, there are 5 different main types of dwellings present. Detached dwellings, semi-detached dwellings, corner houses, terraced dwellings and apartments. Within some of these groups, subgroups exist. For instance, apartments have subgroups 'ground floor apartments' and 'penthouses'. The dwelling type is of critical importance to both the value and the EPC label of a dwelling. Not only that, dwelling types vastly differ from one another. To see this, some summary statistics of transaction prices are shown in Table 4.2.

Table 4.2: Descriptive statistics of relevant variables per dwelling type. *Area* is the total floor area of the dwelling. *DoU* is the mean Degree of Urbanity, a score ranging from 5 (highly rural) to 1 (highly urban). *isCity* is the percentage of transactions within cities with over 50,000 residents. The last 3 columns display the percentage-wise distribution of EPC labels per dwelling type.

	Sample size	Price (€ / m^2)	Area (m^2)	DoU	isCity (%)	A+B (%)	C+D (%)	E+F+G (%)
Detached	106,643	2,524 ± 1,027	165 ± 67	3.99 ± 1.11	14.2%	31.5	31.9	36.5
Semi detached	91,614	2,314 ± 869	130 ± 38	3.46 ± 1.21	18.1%	30.7	36.7	32.6
Corner house	99,663	2,211 ± 912	121 ± 33	2.82 ± 1.18	29.8%	28.9	47.1	24.0
Terraced house	253,611	2,166 ± 805	117 ± 30	2.62 ± 1.15	36.5%	33.7	46.2	20.1
Apartment	198,191	2,806 ± 1,461	84 ± 32	1.70 ± 0.96	62.7%	30.2	42.4	27.5

A closer look at the group of apartments shows difficult problems with this group. Apartments are vastly different from all the other groups in terms of the area, *isCity* and *DegreeOfUrbanity*. Another problem is that because apartments are often located in cities, the variance of the impact of location on the price is much larger for this group than for all other groups. Taking these reasons into account, it is decided to split off the transactions of apartments.

Corner houses and terraced houses have very similar characteristics and location attributes. Moreover, within these types of dwellings, the price variation is relatively low. Lastly, these types of dwelling transactions together still consist of almost half the total sample of transactions. Hence, these types of dwellings are chosen to be part of the main data set, consisting of $353,274$ dwelling transactions.

4.1.4. Location

So far, the data set contains data corresponding to the transaction, the characteristics of the dwelling and data related to the EPC label. However, one of the most important drivers for the price of a dwelling is its location. A house in a favorable location can potentially be valued many times the amount of a house with similar characteristics in a less favorable location. Many choices can be made regarding the modeling of the location of a dwelling, however, the methods that have been considered for my thesis can roughly be categorized in the following four types.

1. Add fixed effect terms for each distinct geographical area, such as neighborhoods, municipalities and provinces.

2. Classify all dwellings into a cluster based on dwelling density, neighborhood characteristics and longitude and latitude data, in order to obtain clusters of dwellings with a comparable location.

3. Fix the scope of research to a particular cluster of homogeneous locations in order to limit price variations due to location.

4. Describe the location of every house by a large set of location variables, such as neighborhood income levels and crime rates, and add these variables as input in a model.

It is important to know the advantages and drawbacks of each of these possible methods. One could argue that the most elegant solution would be to represent the location of a house by a large set of location related variables that best describe it. By doing so it is possible to capture effects of similar locations, and learn what attributes make a certain location favorable. To achieve this goal, data from the Centraal Bureau voor de Statistiek (CBS) can be added to every sample. The CBS grants open access to many socio-economic characteristics, such as average income, crime rate and house density on different levels of location, for instance on neighborhood level, municipality level or province level. All these variables can be added as input in a model, however, this approach has some limitations. As explained before, describing the location is one of the most important factors in modeling house prices. Unfortunately, the available data per location in our case is not able to describe the location well enough. A possible explanation is that the price of dwellings in a certain neighborhood can not be fully captured by the attributes that define the location. Either because certain important variables that describe location are missing in our data set, or because some attributes, such as the hype or popularity of a location, simply can not easily be captured in data.

The other approaches of modeling location that are considered could potentially overcome this drawback, because the other approaches indirectly compare the transaction prices of dwellings in a similar location. By finding a subset of data that is approximately location invariant, the need for explicit modeling of the price for certain locations is no longer required. Therefore, the choice is made to fix the scope of research to a particular cluster of transactions that come from a similar location, in order to limit price variations due to location, which increases the relative impact of EPC label on the price and makes the sample easier to examine due to the decreased sample size. Moreover, this approach also makes the sample more homogeneous, as the characteristics of dwellings in similar locations are more alike. A large disadvantage of this approach is that it is very difficult to find transactions in locations so that the location has insignificant impact on the price. Similar dwellings in the same street can already differ in price due to for instance relative location to the city center, let alone dwellings in different cities or provinces. In Subsection 4.1.5, the approach is discussed.

4.1.5. Finding a location invariant subset of transactions

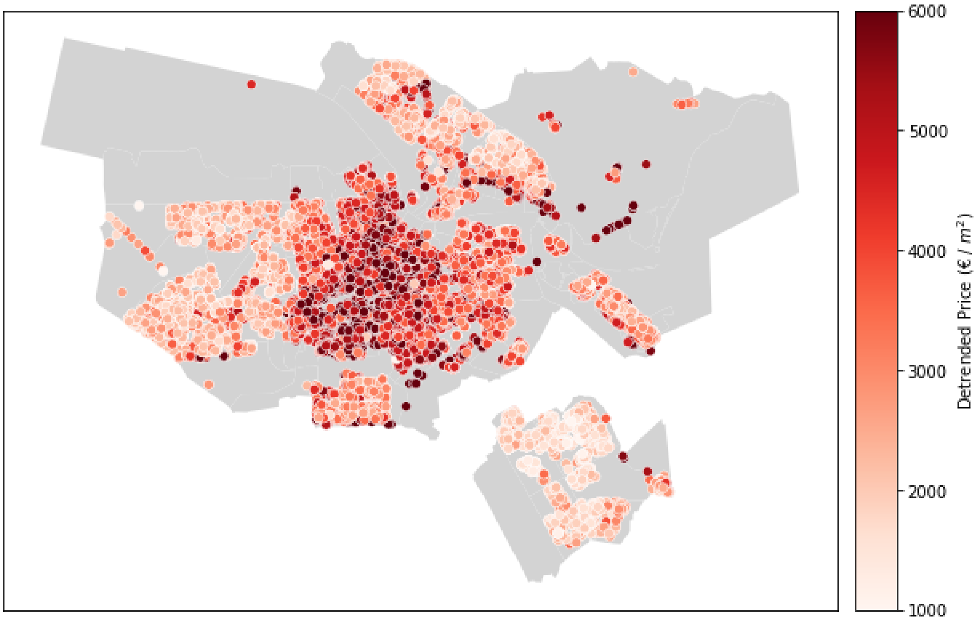


Figure 4.1: Scatter plot of house transactions in Amsterdam. The scatterplot shows that prices of dwellings in Amsterdam heavily vary by location.

Finding a subset of transactions wherein location has minimal impact on the transaction price is extremely difficult. On the one hand the aim is to find a certain region so that the difference of the impact of location is minimal, and on the other hand as many transactions as possible are required for proper analysis. On top of that, not all regions are suited for analysis. In some regions, there simply is too much variety available regarding the impact of location on the transaction prices of dwellings. When there is much variety in the impact that location has on the price of a dwelling, this impact of location on a dwellings price is often enormously difficult to capture within data. This variety often depends on data that is either unknown or hard to measure. This can for instance be variables regarding whether a dwelling has a very favorable location in relation to other nearby dwellings, or a sudden hype that increases the popularity and prices of dwellings in certain neighborhoods. Such large varieties often occur mainly in the largest and most dense cities of a country. Figure 4.1 shows that the prices of dwellings in Amsterdam heavily vary depending on the location. Note that this does not immediately imply that location itself directly impacts the prices of dwellings. It might also be the case that most dwellings in the centre of Amsterdam all have some rare, highly demanded characteristics that other dwellings do not have.

For every transaction in our data set, the specific location is described by the variables Neighborhood, Municipality, COROP-area, longitude/latitude and address. As most of these variables are categorical and not ordinal, the usual approach is to one-hot encode these categorical variables. Encoding neighborhoods in this manner alone provides the data set with over 3000 distinct variables, encoded as $neighborhood_A$, $neighborhood_B$ and so on. Not only is the amount of variables created in this way far too large for analysis, it also removes the spatial correlation between close neighborhoods. Hence a more intelligent approach has to be used in order to properly encode location.

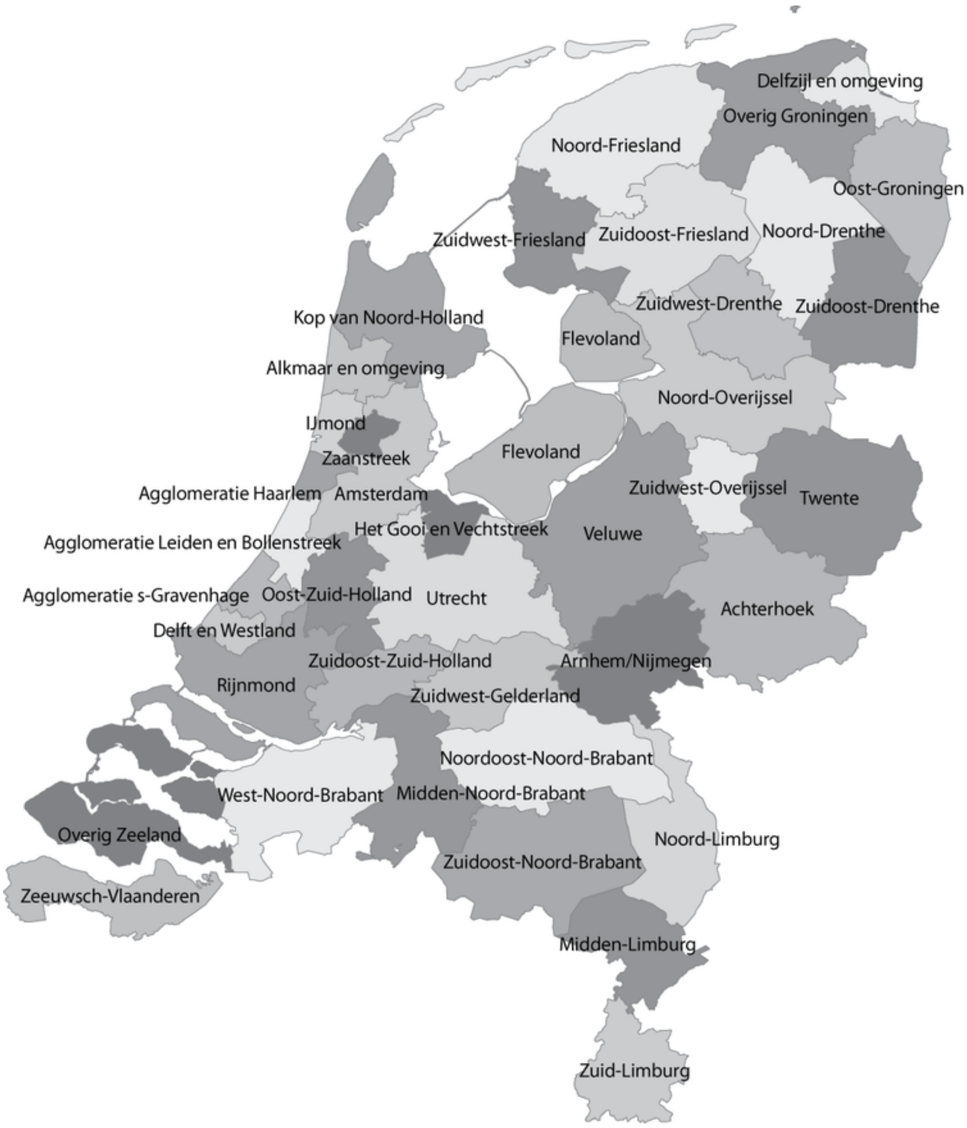


Figure 4.2: Map of the 40 COROP areas in the Netherlands. Grey scales are used to visualise the COROP area borders, but have no meaning regarding data or statistics. Source: CBS [2]

.

The Netherlands can be divided into 40 regional areas, called COROP areas. COROP areas are part of the Dutch Nomenclature of Territorial Units for Statistics or NUTS (French: Nomenclature des Unités Territoriales Statistiques). NUTS is a standard issued by the European Union for referencing subdivisions of countries. COROP areas are the Dutch level 3 NUTS regions, the smallest level of NUTS regions. COROP areas consist of one or more municipalities, however, they are bounded by province borders. Figure 4.2 shows a map of all COROP areas in the Netherlands. COROP areas often consist of a core, such as a business center or city, along with its corresponding catchment area. The Dutch CBS collects many regional data and statistics about these COROP regions in order to analyse regional differences. This makes them suitable for examining which regions have similar impact on the price of a dwelling.

In order to reduce the dimensions of the data set, while properly encoding the information from the location variables, K-means clustering is used. K-means clustering partitions the *N* COROP areas into *K* different clusters. The final goal is to be able to select a cluster of COROP areas with similar transaction and location attributes. Within this cluster, the the variance of the impact of location on the price of a dwelling should be reduced.

K-means clustering sequentially performs the followings steps:

1. Randomly assign K initial COROP areas as centroids.
2. Assign each COROP area to the closest centroid to obtain K clusters.
3. Calculate the new centroids (mean) of the clusters.
4. Repeat steps 2 and 3 until the total sum of distances from all points to the centers has converged.

In order to perform K-means clustering, there are two main choices that have to be made; we have to define what close is, secondly, how many clusters there should be. The goal is for COROP areas to be close when the impact of the location on the price is similar in these areas. With this goal in mind, relevant statistics from different COROP areas are collected. The statistics gathered per COROP area are the amount of dwelling transactions normalized by population size, the median degree of urbanity of municipalities within the COROP area, the mean, median and standard deviation of the transaction price per COROP area, and the $1st$ and $3rd$ quantile of the transaction price per COROP area. Afterwards, these statistics are standardized, so that they have similar magnitude. The notion of close will be the fact whether or not these statistics are close to each other for different COROP areas in Euclidean space.

In order to determine an appropriate amount of clusters, we perform K-means clustering for a several number of clusters. As there is no set way in determining the amount of clusters, some experiments are performed. The quality of the clustering can be derived from the sum of total distance from all cluster points to the center of the cluster the point belongs to. Naturally, as K is increased, the total distance decreases. In Figure 4.3, the sum of total distance from the points to their corresponding cluster centers is plotted against the number of clusters K. The elbow method aims to find the amount of K clusters, so that when more clusters are added, the sum of total distances from the individual points to the cluster centers no longer significantly decreases.

With the elbow method the aim is to find this point from inspecting this plot. Finding the elbow corresponds to the point where the slope of the total sum of distances flattens out, which from Figure 4.3 can be seen as approximately $K = 4$ clusters. As the choice of initial centers of K-mean impacts the clustering outcome, K-means is not deterministic. As such, 5 separate simulations are performed. For each separate simulation, the sum of distortions did not change much as can be seen in Figure 4.3. $K = 4$ was the optimal number of clusters in each of the simulations.

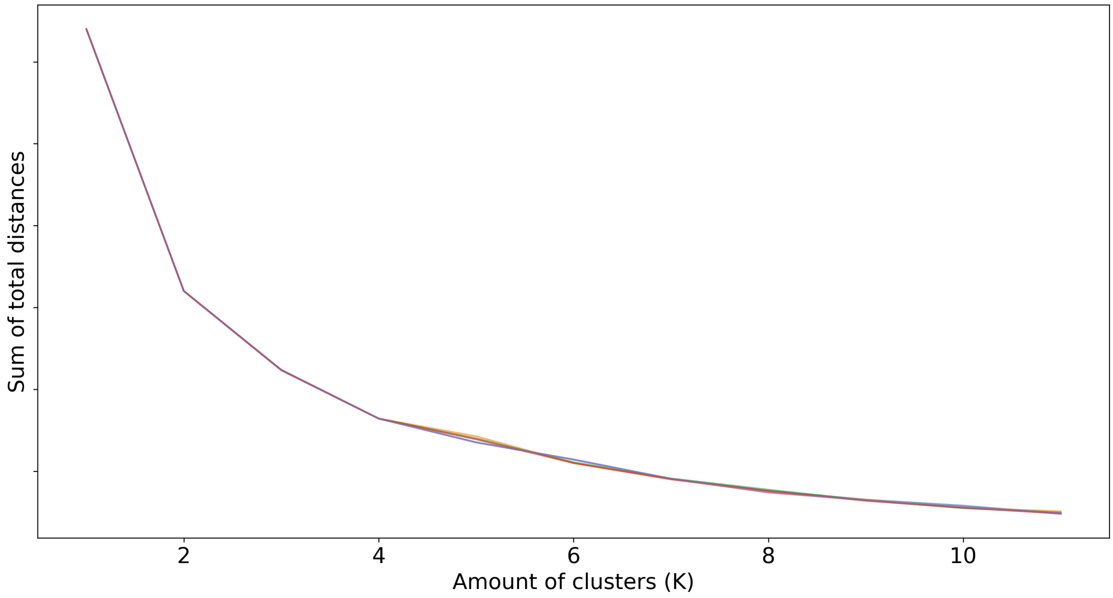
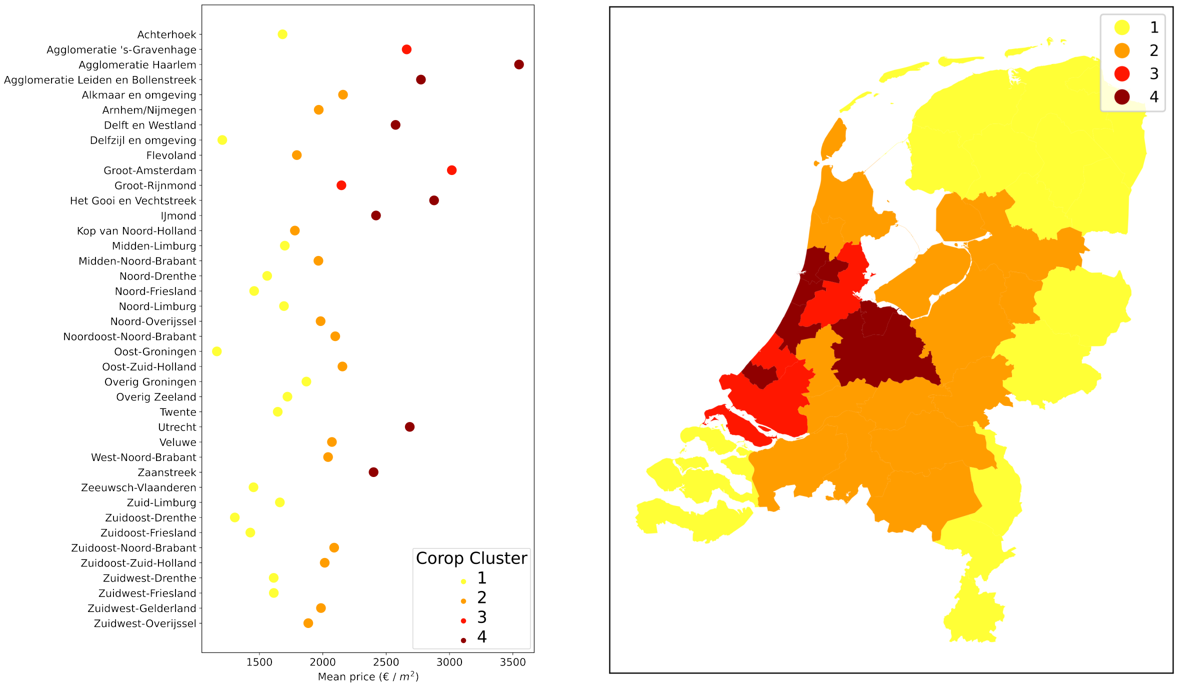


Figure 4.3: Elbow plot for finding the ideal amount of clusters in K-means clustering. The elbow is at $K = 4$.

Clustering the COROP areas using K-means with $K = 4$ provides each COROP area with a cluster number. Figure 4.4 shows clustering the COROP areas based on similarities in transaction data and density does quite a good job in separating COROP regions. The transaction price of the dwellings within clusters have much lower variance than the total sample. It is therefore plausible that the variance of the impact of location on the transaction price is reduced, which the aim was. The next step is to select a cluster of COROP areas for further analysis.



(a) Scatterplot of clustered COROP areas and mean price per m^2

(b) Clusters of COROP areas located on the map of the Netherlands.

Figure 4.4: Estimated clusters of COROP areas in the Netherlands.

Even within these clusters of COROP areas, the impact of varying location on the price of a dwelling is still present. In order to be able to capture the effect of location as well as possible, two additional variables are created. The aim is to be able to capture as much information about the impact of location on the price as possible, with the least as possible amount of variables. To reach this goal, the variables *DegreeOfUrbanity* and *isCity* are created. *DegreeOfUrbanity* is a variable that is a measure for the amount of dwellings per km^2 in a neighborhood. The *DegreeOfUrbanity* is a score ranging from 1 to 5, denoting < 500, $500 - 1000$, $1000 - 1500$, $1500 - 2500$ and > 2500 dwellings per km^2 in a neighborhood, respectively. The *isCity* variable describes whether or not the dwelling can be seen as located within a city. Although the Netherlands has no official notion of when a place is called a city, an often used lower limit for a municipality to be called a city is 50.000 residents, which will also be used throughout this thesis. These two variables had most predictive power as location attribute compared to all variables that have been experimented with. The aim now is to select a COROP area cluster that can be used for later analyses.

It is a difficult task to determine which cluster of COROP areas should be used in later analyses, as there are multiple contradicting constraints in selecting suitable COROP areas. First of all, a COROP area, where as many samples as possible are available in the data set, has the preference. A large set of transactions makes our model able to fit better, which benefits the uncertainty in the final evaluation of the effect of increasing the energy efficiency on the transaction price of a dwelling. Moreover, a large sample is also more likely to have sufficient overlap between the characteristics of dwellings with different EPC labels. Lastly, having a large sample makes the conclusions of the research more generally applicable for dwellings in the Netherlands.

Together with this constraint of requiring as many transactions as possible within cluster of COROP areas, there should not be many very large cities in the COROP area. The transaction price of a dwelling in a very large city can not be sufficiently described by the location variables that are used. When a very large city like Amsterdam or Utrecht is examined, even within this city the impact of different locations on the transaction price will be enormous.

Lastly, there should be at least some overlap between the characteristics of dwellings in the different EPC label groups. The intuition for this is that for every dwelling, there should be at least a small chance of it having a certain EPC label X, based on its characteristics. If this would not be the case, the effect of improving a dwelling its EPC label to X can not be estimated, because there can not exist similar dwellings with a different

EPC label. This assumption was thoroughly discussed in Chapter 3.

Taking these reasons into account, one would like to select a cluster of COROP areas with as many transactions as possible, however, where the impact of the location on the price within municipalities of more than 50.000 residents, and within municipalities with less than 50.000 residents is minimal. All transactions from this certain cluster of COROP areas are then selected, and are supplemented with the data regarding the neighborhood *DegreeOfUrbanity* and the boolean variable *isCity*, which denotes whether the dwelling is part of a municipality with more than 50.000 residents or not.

The cluster of COROP areas that is selected is COROP cluster 2 in Figure 4.4. This COROP cluster has the largest amount of transactions of any cluster. Moreover, this cluster also contains no municipalities with more than 250.000 residents, with Eindhoven and Groningen having the most residents in 2015. This limits the difficulties of having to deal with transaction prices in very large cities, which are mainly captured in clusters 3 and 4. After selecting all transactions from cluster 2, this leaves $153,551$ transactions for further analysis.

4.1.6. Time trend

As the selected cluster of transactions consists of transactions at fixed points in time, one has to deal with the changes of the price of a house over time. In general, assets such as dwellings are expected to grow in value over time at an exponential rate. When Y_{t_1} and Y_{t_0} denote the prices per m^2 of dwellings at time t_1 and t_0, respectively, where $t_1 > t_0$, the expected exponential growth is described by

$$\mathbb{E}\left[\frac{Y_{t_1}}{Y_{t_0}}\right] = exp\left(\alpha(t_1 - t_0)\right) \quad \text{for some constant } \alpha. \tag{4.1}$$

A common approach is to remove the trend in time series data, in order to better analyse the underlying structures of the data. This approach is called detrending. An exponential curve is estimated from the data, and all transaction prices will be scaled, so that every transaction price displays an estimated price as if the dwelling had been sold on 01-01-2015. In Figure 4.5, the positive trend for price per m^2 over time is illustrated for transactions in our data set. This trend is estimated to be approximately 0.53% per month, which comes down to around 6.55% per year.

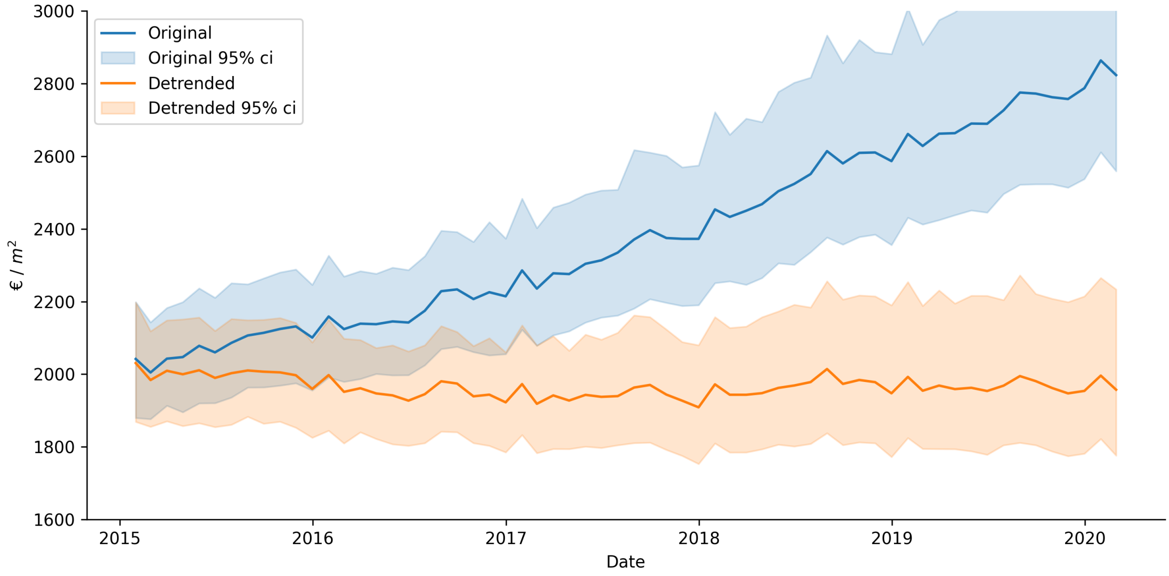


Figure 4.5: The original and detrended price (€ / m^2) over time for transactions, aggregated by month, excluding apartments.

After detrending, the monthly-aggregated transactions are approximately time-stationary, as shown in the same Figure. The choice to detrend the data in this manner is still not perfect. This can be seen from the plots in Figure 4.6. If the detrended data is partitioned by certain characteristics, such as in this case whether on not the dwelling is located in a city, there is still a slight trend present over time. This can be seen in Figure 4.6, where it can be seen that the trend of dwellings in the city is positive, while the trend of cities outside cities is negative after detrending. This is the case because the price trend of dwellings within a city grew faster than the average trend, and vice versa for dwellings outside cities.

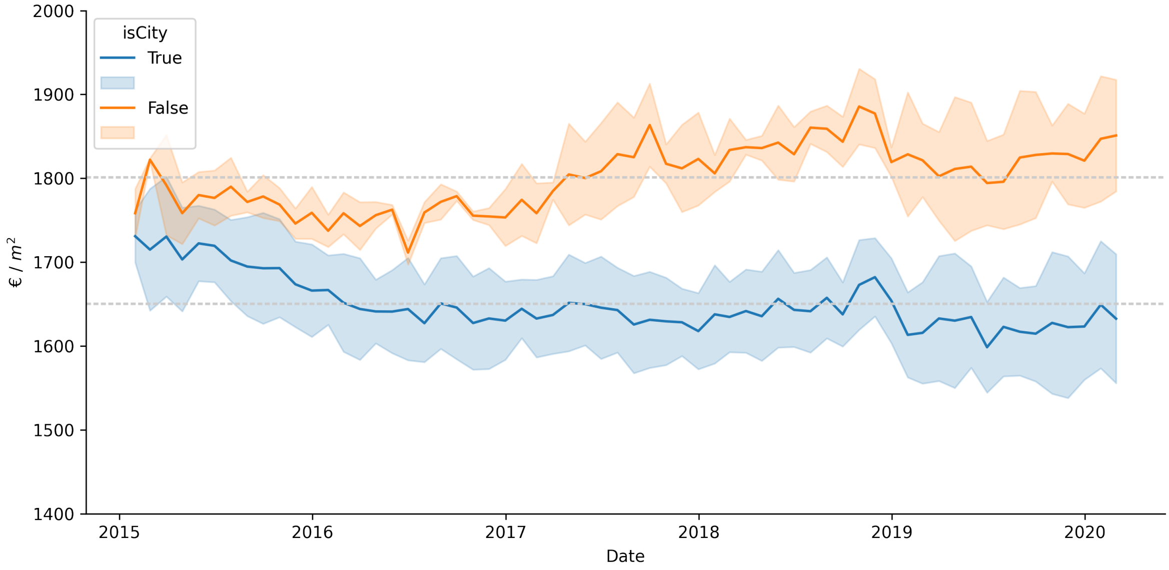


Figure 4.6: Detrended price per m^2 for dwellings within a city, and outside of cities. The plot indicates that the trend is not fully removed when subtypes of dwellings are examined.

However, there is no easy solution for this problem. Moreover, this will probably not be a large problem in our setting, as our goal is not prediction of dwelling prices, but rather trying to estimate a certain driver of this price.

4.1.7. Other problems in the data

In this Subsection we discuss which important features are either missing or are difficult to capture in data. Three main groups of variables are missing, which impact the errors in modeling. These three groups of variables are related to the circumstances of the sale, and the location and the quality of the dwelling, and will be elaborated below.

Circumstances of the sale Some variables related to the circumstances of the sale might heavily impact the price. Examples are whether the dwelling is sold to friends or family, or whether or not renters are living in the dwelling. Basically, such circumstances create outliers in our data that are difficult to detect without having access to variables related to them.

Location Most of the impact of location on the price is attempted to reduce in Subsection 4.1.4. However, some features related to the location of a dwelling are either unobserved or difficult to be captured by data. Examples of such variables are for instance the uniqueness compared to close by dwellings, whether or not the ground beneath the dwelling is owned or whether or not there are big plans for renovating the neighborhood of the dwelling. All these variables might heavily impact the price of the dwelling.

Quality Lastly, variables related to the maintenance condition and quality of the dwelling are very important regarding both the price and the energy efficiency. Unfortunately, variables directly measuring the maintenance condition or quality are not available in our data set. In future research, language processing or image recognition might be used for scanning advertisements, in order to obtain variables that better resemble the quality of different aspects of a dwelling.

The general consequence of missing variables will be increased errors and larger variance in the obtained estimates. In particular, problems arise when any of the missing variables has large impact on both the energy efficiency of a dwelling, as well as on the price. If this is the case, the results of the estimates of improving energy efficiency on the price might be biased. The larger the impact of this missing variable is on both outcomes, the more biased the estimate will be. In Chapter 2, tests are done to show this phenomenon and show how large this bias gets with different strengths of missing features.

From the groups of variables listed above, features belonging to the group of quality might give problems. I suspect that overall quality of a dwelling is at least somewhat positively correlated with EPC label, as renovating and insulating a dwelling are tasks that tend to be done simultaneously by home owners. Due to this fact, the impact of EPC label on the price could be overestimated, because part of the effect may be attributed to the improved quality aspect of a dwelling, rather than its energy efficiency. This bias will be hard to show in our results. Consequently, conclusions should be drawn with caution.

4.1.8. Feature engineering, missing data & outliers

Feature engineering resembles the creation of meaningful variables that can be used for modeling from the raw data. In Subsection 4.1.5, some meaningful features were already created for location, namely variables *isCity* and *DegreeOfUrbanity*. A similar approach has to be used in order to create variables for the characteristics of a dwelling.

Starting out with the amount of rooms and bathrooms. The raw data for these variables still needs to be altered in order to be used. In Table 4.3, the 5 first entries in both these columns are shown. This combination of text and numerical inputs can not be handled by models.

Table 4.3: First 5 entries in raw data set

AmtOfRooms	**AmtOfBathrooms**
5 kamers (4 slaapkamers)	1 badkamer en 2 aparte toiletten
4 kamers (3 slaapkamers)	1
NaN	1 badkamer
5	NaN
5 kamers (3 slaapkamers)	2 badkamers en 1 apart toilet

Table 4.4: First 5 entries after extracting numbers

AmtOfRooms	**AmtOfBathrooms**
5	1
4	1
NaN	1
5	NaN
5	2

The entries in *AmtOfRooms* consist of either a string 'X kamers (Y slaapkamers)', or a plain number. The aim is to extract the amount of rooms (Dutch: kamers), and drop the information on the amount of bedrooms (Dutch: slaapkamers). The amount of bedrooms is not expected to add much information, because basically any room can be classified as a bedroom. In both columns, the string entries are replaced by the number before the 'kamers' and 'badkamers', respectively, which is extracted from the string. The NaN values will be handled later on in this Subsection. The result is show in Table 4.4.

Next, the variable *DwellingVolume* is divided by the variable *LivingArea* in order to obtain the *AvgHeight*. This is done so that this variable becomes independent from the living area.

Finally, there are some variables related to the type of isolation of a dwelling. For instance, the type of glazing and whether the roof and floor are properly isolated in a dwelling. These variables directly impact the EPC label of a dwelling. These variables will be dropped from our data set, as they are part of the total effect of the EPC labels on the price, which is the goal to estimate. In conclusion, the assumption is made that improving isolation itself only affects the price of a dwelling through the improved EPC label.

Table 4.5: Variables and the percentage of missing entries.

Variable	**Missing entries (%)**
LotArea	0.16
hasGarage	1.28
hasGardenShed	2.71
hasMonumentalStatus	2.91
hasBasement	9.14
AmtOfBathrooms	28.09
AmtOfFloors	8.68
AmtOfRooms	5.34

Now that the raw data columns are turned into useful features, we are left with 13 variables. 10 of these variables contain entries with missing values. Missing data in the crucial variables, which are *EnergyEfficiency*, *SalePricePerM2*, *TransactionDate*, *FloorArea* and the proxy variables for location were already removed in earlier stages. Table 4.5 shows the variables and the corresponding percentage of missing data entries. Those entries will be handled as follows.

Variable *LotArea* is a numerical variable. This variable is highly correlated with the *FloorArea* of a dwelling. As such, a linear regression model is fitted to estimate *LotArea* from *FloorArea*. Afterwards, missing values in *LotArea* are imputed by predicting it from the *FloorArea*.

Variables *hasGarage*, *hasMonumentalStatus* and *hasBasement* are all binary variables, containing either 1 or 0 values as data entries, corresponding to the dwelling having the characteristic or not, respectively. All these variables have a relatively small fraction of 1 values. Moreover, intuitively it feels more likely that missing values in these columns denotes that said variables are not available in the dwelling. Hence, the missing values in these columns are imputed with the value 0.

Variables *AmtOfBathrooms*, *AmtOfFloors* and *AmtOfRooms* all contain ordinal values. Similar to *LotArea*, these variables are all correlated to the *FloorArea*. Hence, missing entries for these variables are imputed by fitting a linear regression model to estimate the missing variables from *FloorArea*. Afterwards, the estimated values are rounded to their closest integer values.

Outliers in all variables are removed as follows. First, any variables with negative values that are not expected are removed from the data set. Afterwards, 34 transactions of dwellings with a detrended *SalePricePerM*^{2} above €10,000 and under €500 are removed, as those are assumed to be either erroneous data points, or not representative for analysis. Similarly, 218 data points of dwellings with *LivingArea* over 1000m^2 are removed. Afterwards, 295 data points of dwellings with *LotArea* over 1000m^2 are removed, for similar reasons. Lastly, 16 transactions of dwellings with *AvgHeight* over 10m^2 are removed.

4.1.9. Data set splitting

In Chapter 1, we noted that the treatment will always be a binary variable. The treatment, energy efficiency, currently has three categories of values. In order to obtain adequate data sets that can be used for analysis, the data set is split into two separate instances.

The first data set contains all transactions of dwellings with moderate and good energy efficiency. The binary variable *EnergyEfficiency* will assigned the value *TRUE* if the dwelling is energy efficient, and *False* if moderately energy efficient. This data set will be referred to as the energy efficient data set.

The second data set contains all transactions of dwellings with bad and moderate energy efficiency. The binary variable *EnergyEfficiency* will assigned the value *TRUE* if the dwelling is moderately energy efficient, and *False* if energy efficient. This data set will be referred to as the energy inefficient data set. Note that both data sets overlap on all transactions of moderately energy efficient dwellings.

Building codes regarding energy efficiency have become more strict over the last decades in the Netherlands. As a result, dwellings with a construction year later than 1997 are always in possession of EPC label A or B. Similarly, dwellings with a construction year later than 1974 are always in possession of EPC label D or better. This will give problems in our later analysis, which will be elaborated further in Chapter 3. Intuitively, one might understand that it is impossible to estimate the effect of improving energy efficiency from moderately energy efficient to energy efficient on the transaction price of a dwelling that can not possibly be moderately energy efficient due to building codes. Hence, we remove the sample of dwellings with a construction year after 1997 from the data set containing energy efficient dwellings. Similarly, the sample of dwellings with a construction year after 1974 are removed from the energy inefficient data set.

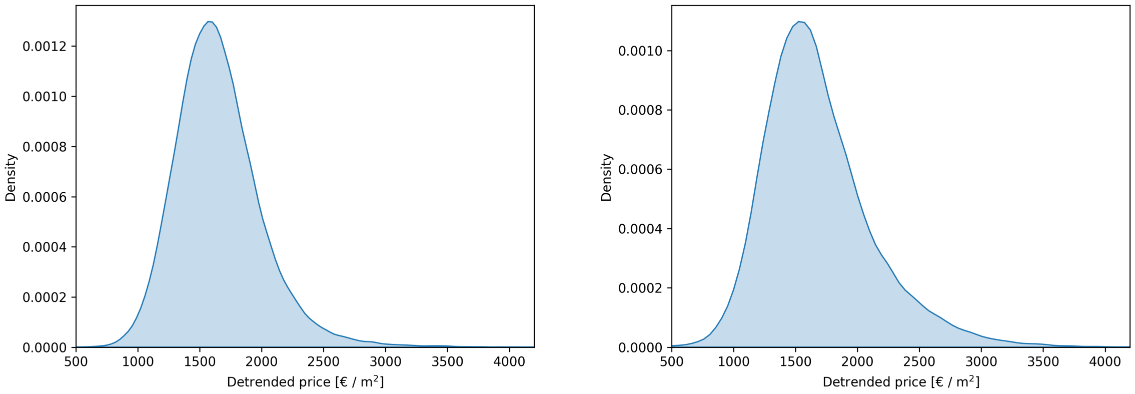
This marks the completion of operations in order to obtain our final data sets. The energy efficient data set contains $96,267$ transactions of terraced or corner dwellings, all located in cluster 2 as defined in Subsection 4.1.5. The energy efficient data set contains $60,154$ transactions of terraced or corner dwellings located in cluster 2. Both data sets contain 15 describing features, among which the variables of interest, *EnergyEfficiency* and its impact on the detrended *SalePricePerM*^{2}.

4.2. Analysis on final data sets

In this Section the aim is to get a feel for the final data set that will be used for analysis. In Appendix A, descriptive statistics of all variables in the final data set are shown, among which the mean, median and standard deviation of all variables present in the data sets. In this Section, some specific, interesting visualizations are presented.

4.2.1. Transaction price distribution

In Figure 4.7, the distribution of the detrended price of dwellings is shown. For readers that are familiar with the housing market, these prices might seem low. These low prices are the result of the selection of a location cluster with relatively low prices, the exclusion of apartments and detached dwellings, and detrending all prices to the price in 2015. The distribution of the detrended price is slightly skewed for both data sets. The prices of dwellings in the energy inefficient data set have slighly more variation than the prices of dwellings in the energy efficient data set, as can be observed from the Tables in Appendix A.

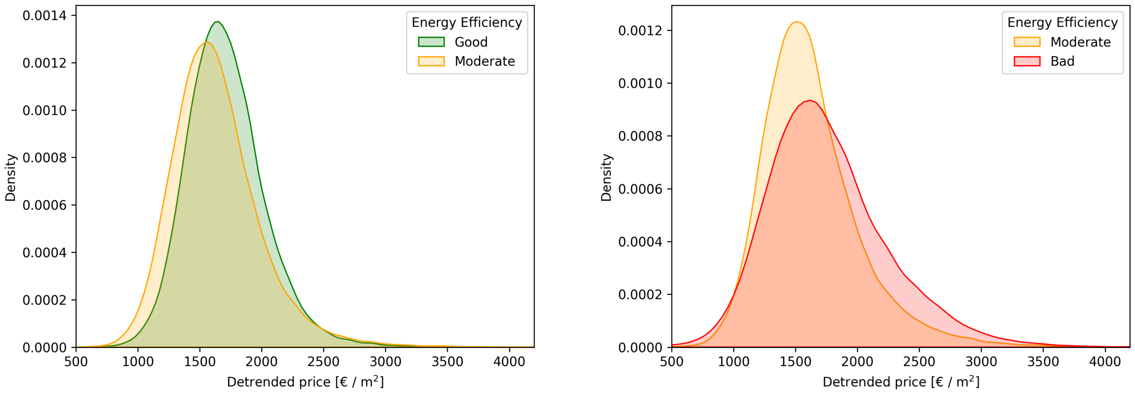


(a) Energy efficient data set

(b) Energy inefficient data set

Figure 4.7: Distribution of the detrended price per m^2 of dwellings in the final data sets.

In Figure 4.8, the distributions of the detrended prices of dwellings with different levels of energy efficiency is plotted. The distributions for these groups are significantly different from eachother. Dwellings with moderate energy efficiency have a smaller mean transaction price in comparison with the group of energy efficient dwellings. The sample of dwellings with bad energy efficiency has relatively fat tails, meaning that the price variations in this group are larger. This might mean that prices are more difficult to model accurately for this group.



(a) Energy efficient data set

(b) Energy inefficient data set

Figure 4.8: Distribution of the detrended price per m^2 per energy efficienciy level for dwellings in the final data set.

4.2.2. Correlations

In Figure 4.9, pearson correlation coefficients between all variables are displayed when the energy efficient and energy inefficient data sets are merged. Interesting to see is that *ConstructionYear* and *FloorArea* are positively correlated with *EnergyEfficiency*. *DegreeOfUrbanity* is negatively correlated with energy efficiency, meaning dwellings in urban areas are generally less energy efficient than dwellings in rural areas.

It is interesting to note that the variable most correlated with the *SalePricePerM*^{2} is *AvgHeight*. A possible explanation for this is that higher ceilings are most commonly found in dwellings of higher quality, and are perceived as aesthetic.

The amount of rooms and the floor area are negatively correlated with the *SalePricePerM*^{2}. This might seem strange, however, this probably can be explained by the fact that larger sized dwellings are often found in more rural areas, which are generally less expensive. Smaller houses on the other hand, are often found in more urban areas, which are generally more expensive.

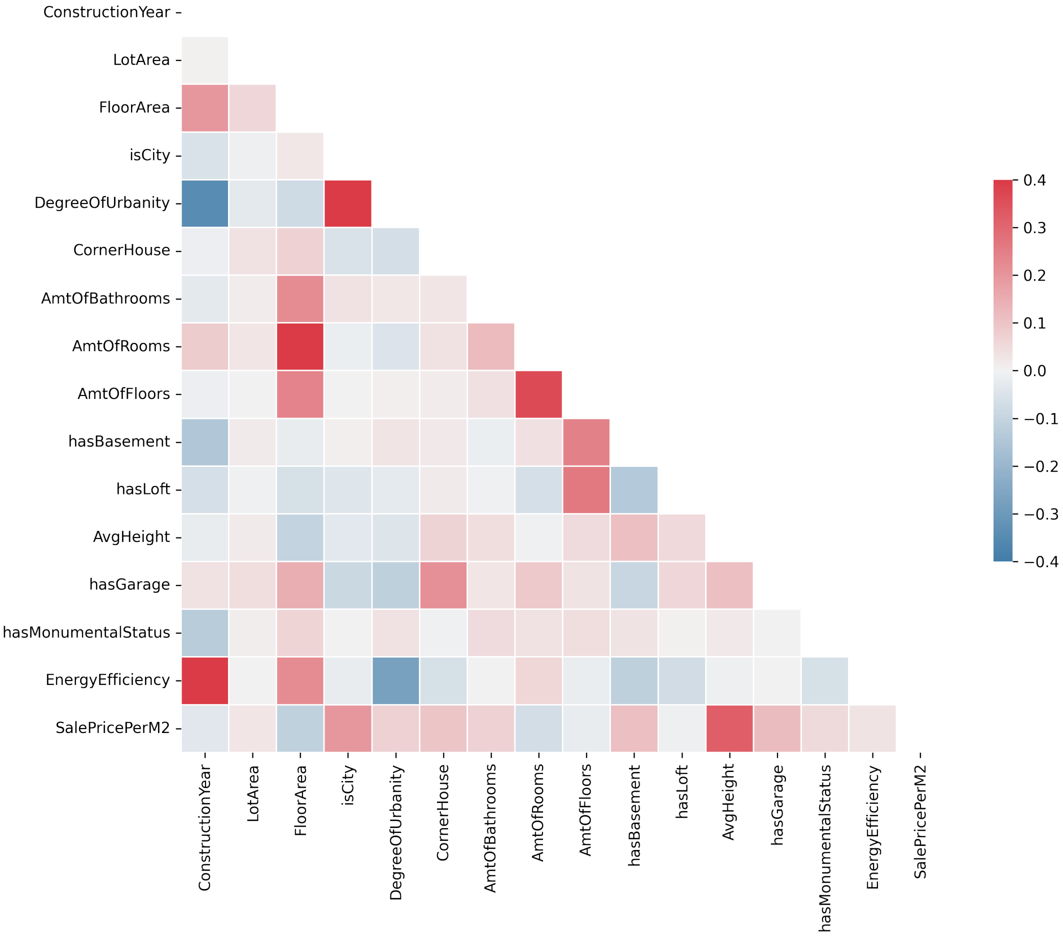


Figure 4.9: Heatmap of Pearson correlation coefficients between all variables.

In this Chapter, we have presented how the data from the raw data sources was modified in order to obtain data sets that can be used for analysis. Important choices were made regarding the modeling of energy efficiency, house types and the location of houses. Analysis is performed on the final data sets, and the results are presented in Chapter 5.

5

Results

5.1. Mean model estimation

In this Section, the Random Forest models used for the estimation of $\mu_1(X) = \mathbb{E}[Y \mid X, T = t]$ and $\mu_0(X) = \mathbb{E}[Y \mid X, T = t]$ are presented. This Section will discuss the following.

- Which hyperparameter values lead to a minimized mean squared error in the validation data set?
- Performance evaluation and comparison to benchmarks
- Variable importance evaluation

As explained in Chapter 4, the total data set is split based on the energy efficiency of dwellings, and two separate models are fit on these separate data sets throughout the Chapter. The first data set contains transactions of only energy efficient and moderately energy efficient dwellings, and the second data set contains transactions of only energy inefficient and moderately energy efficient dwellings. We will refer these data sets as the "energy efficient data set" and the "energy inefficient" data set, respectively, throughout this whole Chapter. Note that transactions of moderately energy efficient dwellings are present in both data sets.

Unless explicitly stated otherwise, evaluation of performance of models is always evaluated on unseen data (or: test data).

5.1.1. Hyperparameter evaluation

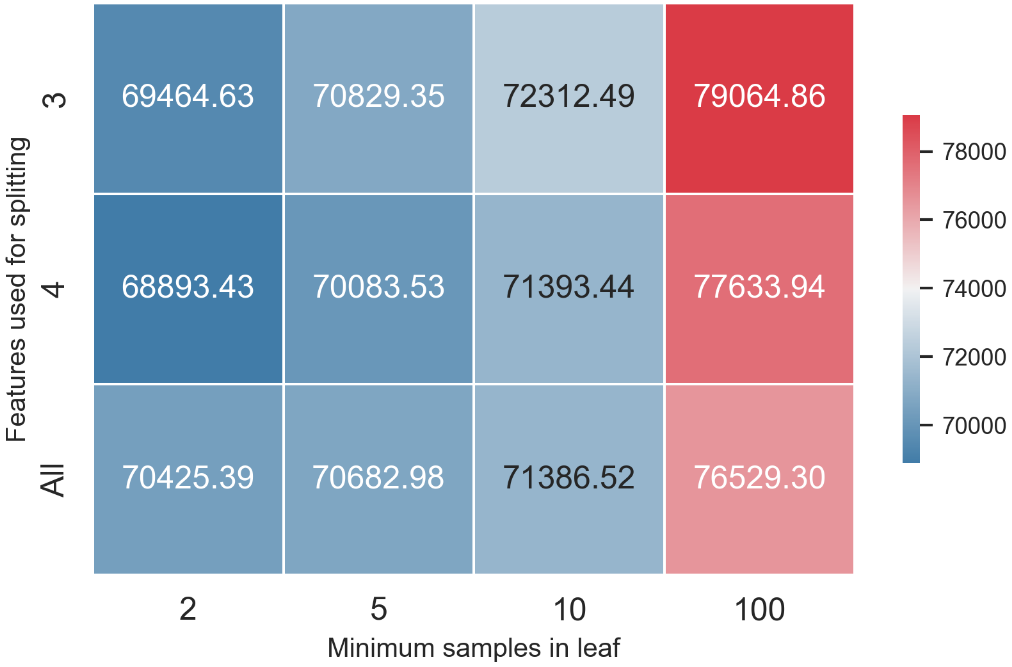


Figure 5.1: Average Mean Squared Error over the 5 folds of test sets, for each combination of hyperparameters for the Energy Efficient data set.

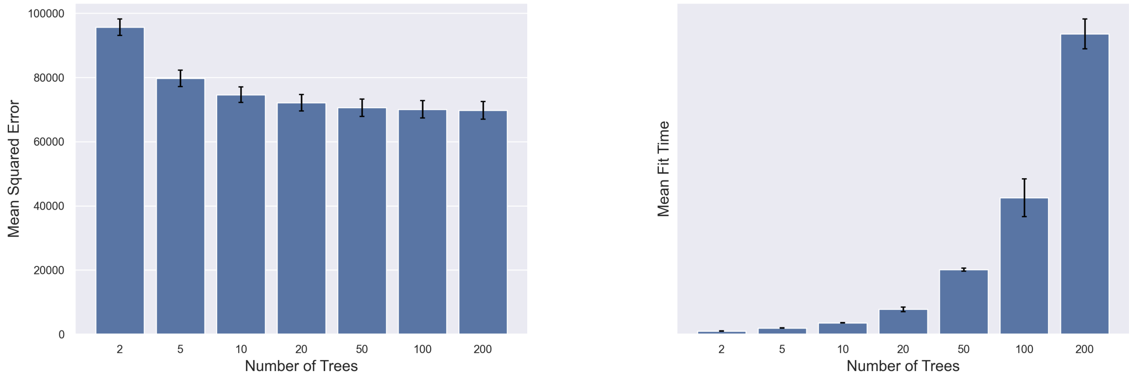
In Figure 5.1, the mean squared error is shown for different combinations of hyperparameters. In the Figure, the hyperparameter *maximumdepth* was left out, because in all hyperparameter combinations, no restrictions on the maximum depth led to the lowest mean squared error. The hyperparameter combinations that achieve the lowest mean squared error for both data sets are summarized in Table 5.1.

	Energy Efficient Data Set	Energy Inefficient Data Set
Features used for each split	4	4
Minimum samples in leaf	2	2
Maximum depth	None	None

Table 5.1: Hyperparameter settings that produce the lowest errors for the Random Forest regression models for the energy efficient data set and the energy inefficient data set. For both data sets, the hyperparameters that achieve the lowest mean squared error are equal.

It is interesting that the best model is obtained when barely any restriction is forced on the complexity of individual decision trees. That is, there is no maximum depth, nor a significant minimum leaf sample size assigned. As a result, the individual trees are grown very deep, and hence are very prone to overfitting. However, combining these deep unrestricted trees in a random forest has the best performance on unseen data in comparison to other hyperparameter settings, as is explained by the theory in Subsection 3.2.2.

The computation time grows almost linearly with the amount of individual trees used, as can be seen from Figure 5.2. Moreover, the mean squared error does not improve much when adding more than 100 estimators. Consequently, the amount of trees that will be used is $n_{trees} = 100$.



(a) MSE

(b) Fit time

Figure 5.2: cross-validated mean squared error and mean fit time versus the number of trees used. Mean squared error is estimated on a held-out test data set.

5.1.2. Model evaluation

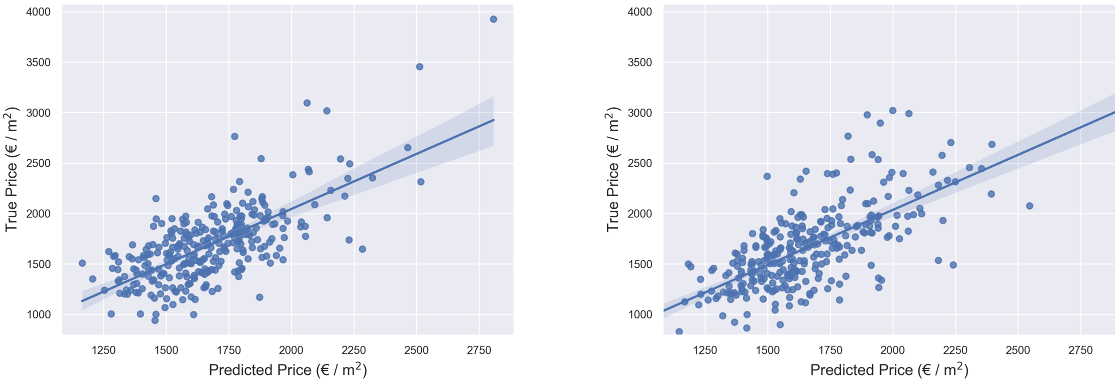
The conditional mean models are evaluated with cross-validation. The Random Forest model with the best hyperparameter settings as discussed in the previous Subsection was fit on the training data. Afterwards, the predictions made on the test data are compared with the true values of the test data in terms of mean squared error and R^2. These are compared to benchmarks, the default settings for Random Forest regression, and linear regression. The results are shown in Table 5.2.

	The energy efficient data set			Energy inefficient data set		
	RF (best)	RF (default)	OLS	RF (best)	RF (best)	OLS
Mean Squared Error (train)	70±2	72±2	104±5	119±4	121±3	157±6
Mean Squared Error (test)	73±2	75±2	133±5	129±4	131±5	158±5
R^2 (train)	0.44 ± 0.01	0.43 ± 0.01	0.23 ± 0.01	0.41 ± 0.01	0.40 ± 0.01	0.22 ± 0.01
R^2 (test)	0.42 ± 0.01	0.41 ± 0.01	0.23 ± 0.01	0.38 ± 0.02	0.37 ± 0.02	0.22 ± 0.02

Table 5.2: Performance of different conditional mean models. Mean squared errors are times 10^3, but are truncated for readability. RF (best) indicates the random forest with the hyperparameters as in Table 5.1. RF (default) indicates a random forest with default settings. OLS indicates a linear regression solved with ordinary least squares.

As can be seen from Table 5.2, the mean squared error and R^2 score are significantly better for the random forest model with custom hyperparameters in comparison with the benchmarks. The mean squared errors are generally slightly higher for the energy inefficient data set, which might be the result of the fact that energy inefficient dwellings are more heterogeneous than energy efficient dwellings. A slight difference in performance is noticeable between the test set metrics and the traning set metrics. This is most likely the result of slight overfitting on the training data set, and is expected.

In Figure 5.3, for $n = 300$ data points, the price predictions made by the random forest with hyperparameters as presented in Subsection 5.1.1 are plotted versus the true observed prices.



(a) Energy efficient data set

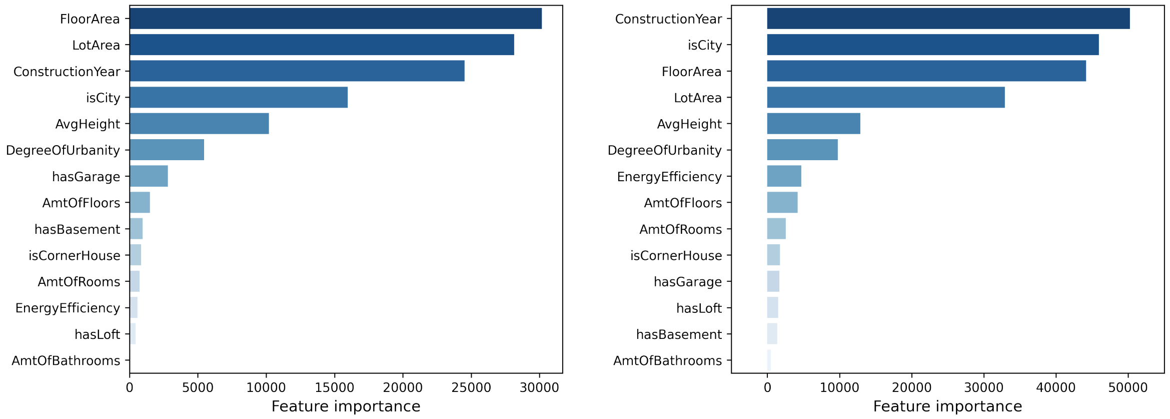
(b) Energy inefficient data set

Figure 5.3: Predictions from the Random Forest model with hyperparameters as in Table 5.1 versus true observed values. $n = 300$ random data points are plotted. A linear regression line is plotted through the data points.

The prediction errors are still quite significant. This can be largely attributed to the fact that transaction prices of dwellings are generally very noisy. Many variables that might be important are not captured in our feature set, or can not be properly captured in data at all, as explained in Chapter 4.

5.1.3. Feature importance evaluation

To estimate which features were most important for these predictions in the random forest model, the permutation importance as explained in Subsection 3.2.4 is used. This leads to the feature importance ranking as presented in Figure 5.4. For the models for both the energy efficient data set and the energy inefficient data set, the most important features are the same. It is interesting to note that energy efficiency is not important in either model. The performance of this model does not suffer much from permuting the energy efficiency variable. This can possibly be explained by two reasons. The first being that energy efficiency is heavily correlated with construction year, one of the most important variables. If the construction year is known, knowing energy efficiency barely adds more predictive power. The second reason is that indeed energy efficiency is likely barely of any importance for prediction, relative to most other variables in the model.



(a) Energy efficient data set

(b) Energy inefficient data set

Figure 5.4: Feature importance ranking for features in the conditional mean models. Feature importance values should be interpreted as a rough ranking of the importance of features. The x-axis shows the increase in mean squared error when values in this feature are randomly shuffled between samples, and hence have no predictive value.

5.2. Propensity model estimation

For the propensity score model, we use the same data that we used for the mean model. Thus, similar to the conditional mean models, two propensity score models are estimated. One model for the sales of neutral and energy efficient dwellings, called the energy efficient data set, and one for the sales of neutral and energy inefficient dwellings, called the energy inefficient data set. A similar approach is used to perform hyperparameter selection, model evaluation and feature importance evaluation as in the previous Section.

5.2.1. Hyperparameter evaluation

	Energy efficient data set	Energy inefficient data set
Features used for each split	3	3
Minimum samples in leaf	5	5
Maximum depth	None	None

Table 5.3: Hyperparameter settings that produce the lowest log loss for the Random Forest classification models for the energy efficient data set and the energy inefficient data set.

The minimum samples in a leaf value is slightly higher for the propensity score models in comparison with the mean models from previous Subsection.

5.2.2. Model evaluation

While log loss is the main metric of interest, we also compare the accuracy of the Random Forest models with hyperparameters as in 5.3 with their benchmarks, a default Random Forest and a logistic regression. These results are summarized in Table 5.4

	Energy efficient data set			Energy inefficient data set		
	RF (best)	RF (default)	LR	RF (best)	RF (default)	LR
Mean Log Loss (train)	0.35 ± 0.00	0.38 ± 0.00	0.54 ± 0.00	0.45 ± 0.00	0.48 ± 0.01	0.62 ± 0.00
Mean Log Loss (test)	0.36 ± 0.01	0.38 ± 0.01	0.55 ± 0.00	0.45 ± 0.01	0.49 ± 0.02	0.63 ± 0.01
Accuracy (train)	0.88 ± 0.00	0.88 ± 0.00	0.76 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.67 ± 0.00
Accuracy (test)	0.87 ± 0.00	0.87 ± 0.00	0.76 ± 0.00	0.79 ± 0.00	0.78 ± 0.00	0.66 ± 0.01

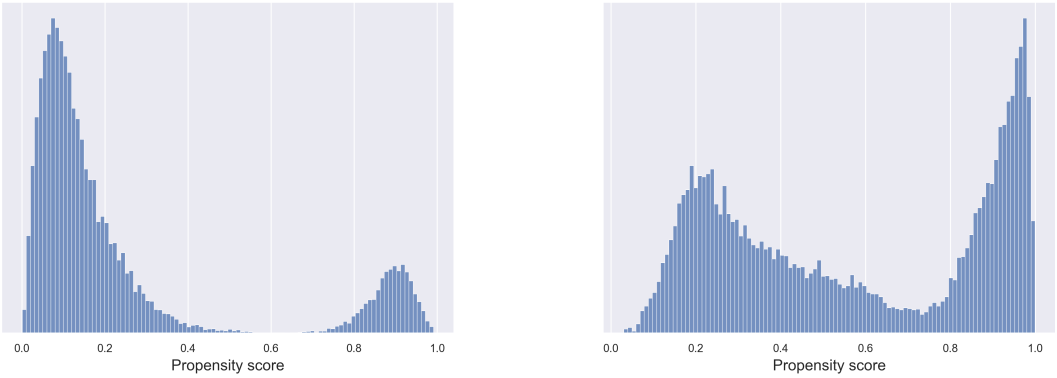
Table 5.4: A comparison of mean log loss and accuracy of the random forest model with custom hyperparameters. RF (best), RF (default) and LR indicate a random forest with hyperparameters as in 5.3, a random forest with default hyperparameters and a logistic regression model, respectively.

The random forest model with the best hyperparameters performs significantly better in terms of mean log loss in comparison with its benchmarks. The performance of all the models on the energy inefficient data set is slightly worse than the performance on the energy efficient data set. Hence, it is more difficult for the models to distinguish between energy inefficient and moderately energy efficient dwellings, in comparison with moderately energy efficient and energy efficient dwellings.

The differences between scores on the training data set and on the test data set are in line with the expectations. Performance on the training data sets is slightly better, as the models have slightly overfit on the training data.

5.2.3. Propensity score evaluation

In this Section, the estimated propensity scores are evaluated. As described in Chapter 3, the propensity scores have to be strictly between 0 and 1 for the AIPW estimator. A histogram of the estimated propensity scores for the energy efficient data set and the energy inefficient data set, with their respective models with hyperparameters as in 5.4, are presented in Figure 5.5.



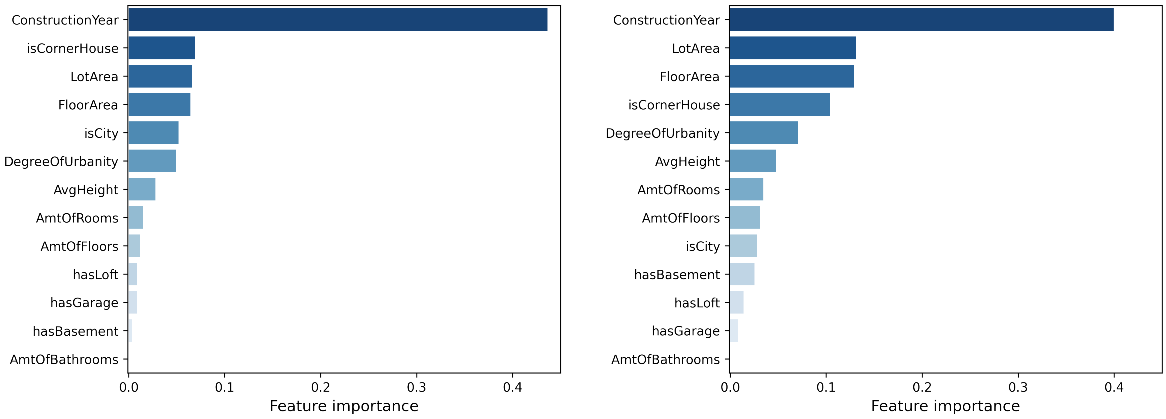
(a) Energy efficient data set

(b) Energy inefficient data set

Figure 5.5: Histogram of the estimated propensity scores.

When propensity scores are strictly between 0 and 1, then for a dwelling with any vector of characteristics $x \in X$, the estimated probability of it having any treatment status is positive. The minimum and maximum estimated propensity scores are 0.006 and 0.991 for the energy efficient data set, and 0.030 and 0.996 for the energy inefficient data set. This implies that for the energy efficient data set, the dwelling which is estimated to be the least likely to have a green EPC label has a 0.6% estimated probability of having a green label, which is over 0% as required. The maximum estimated probability is 99.1% of having a green EPC label. Hence, the estimated propensity scores are bounded between 0 and 1. Similarly, for the energy inefficient data set the least likely dwelling to have a moderate EPC label has a probability of 3.0% for this event, and the most likely dwelling to have a moderate EPC label has a 99.6% probability of having a moderate EPC label. Hence, the estimated propensity scores for the energy inefficient data set are also between 0 and 1. This ensures overlap, which is required for the AIPW estimator to function, as explained in Chapter 3. These estimated probabilities imply that the assumption that true propensity scores are strictly between 0 and 1 is plausible for all dwellings in both data sets.

5.2.4. Feature importance evaluation



(a) Energy efficient data set

(b) Energy inefficient data set

Figure 5.6: Feature importance for the propensity score models. Construction year is clearly the most important variable for classifying a dwelling as energy efficient or not. The x-axis shows the increase of log loss when the sample values of this feature are randomly permuted.

Unsurprisingly, for both data sets, construction year is by far the most important feature for modeling the probability of a dwelling being energy efficient. In the energy efficient data set, other variables are barely of any importance. In the energy inefficient data set, the total floor area and whether the dwelling is at the corner of a street are somewhat important for the model to estimate the probability of the dwelling having a certain treatment status.

5.3. Treatment effects

In this Section, the treatment effects for the energy efficient data set and the energy inefficient data set will be discussed. That is, the average effect of improving the energy efficiency of a dwelling on its price. Later on, in Subsection 5.3.2, the variation in treatment effect conditional on the characteristics of a dwelling, the heterogeneous treatment effect, will be evaluated.

5.3.1. Average Treatment Effects

The average treatment effect is always estimated on a population level on held-out data from the energy efficient data set and the energy inefficient data set. Results are presented for the linear doubly robust estimator, and for the forest doubly robust estimator, for both data sets, in Table 5.5.

	Energy efficient data set		Energy inefficient data set	
	Linear DR	Forest based DR	Linear DR	Forest based AIDRPW
Average treatment effect	20.96	20.98	96.03	97.70
Standard error	3.84	11.56	4.23	20.31
95% interval	$[13.45, 28.50]$	$[-1.71, 43.62]$	$[89.40, 105.99]$	$[56.22, 135.84]$
P-value	0.00	0.07	0.00	0.00

Table 5.5: Average treatment effects estimated by a linear AIPW estimator and a forest-based AIPW estimator. The p-values relate to the null hypothesis of no average effect of treatment.

As can be seen from Table 5.5, the average treatment effects as estimated by the forest DR estimator and linear DR estimator are similar in magnitude and sign. The forest-based estimator does not assume a parametric form. Thus, the variance of the average treatment effect estimate is computed differently and more conservatively, which results in a standard error of significantly larger magnitude. As a result, the forest-based average treatment estimate for the energy efficient data set is not significantly different from 0, based on its p-value of 0.07. All other estimates of the average treatment effect are significantly different from 0 on the 5% level.

For the energy efficient data set, the average estimated effect of increasing the energy efficiency from moderately energy efficient to energy efficient on the price is 20.96 and 20.98 for the linear and forest-based DR estimator, respectively. That is, if the energy efficiency of a dwelling is improved from moderate to good, the estimated transaction price per m^2 increases by approximately €21, on average.

For the energy inefficient data set, the average estimated effect for improving energy inefficient dwellings to be moderately energy efficiency is 96.03 and 97.70 for the linear and forest-based DR estimator, respectively. Hence, the estimated price increase per m^2 when improving an energy inefficient dwelling to be moderately energy efficient is approximately €97 per m^2 on average.

5.3.2. Treatment effect heterogeneity

In this Subsection, the treatment effect is evaluated for different values of characteristics X. Again the results of both a linear DR estimator and forest-based DR estimator are presented. The coefficient estimates of the linear DR estimator are presented in Table 5.6.

	Energy efficient data set			Energy inefficient data set		
	Point estimate	Standard error	P-value	Point estimate	SD	P-value
Construction year	-0.91	0.45	0.04	-1.15	0.24	0.00
Degree of urbanity	-9.29	3.20	0.00	-0.42	3.89	0.91
Total floor area	0.13	0.18	0.46	-0.28	0.19	0.13
Intercept	1828.54	838.18	0.04	2370.58	472.14	0.00

Table 5.6: Coefficients of the linear parametric conditional average treatment effect model.

From Table 5.6, the estimated treatment effect functions are given by

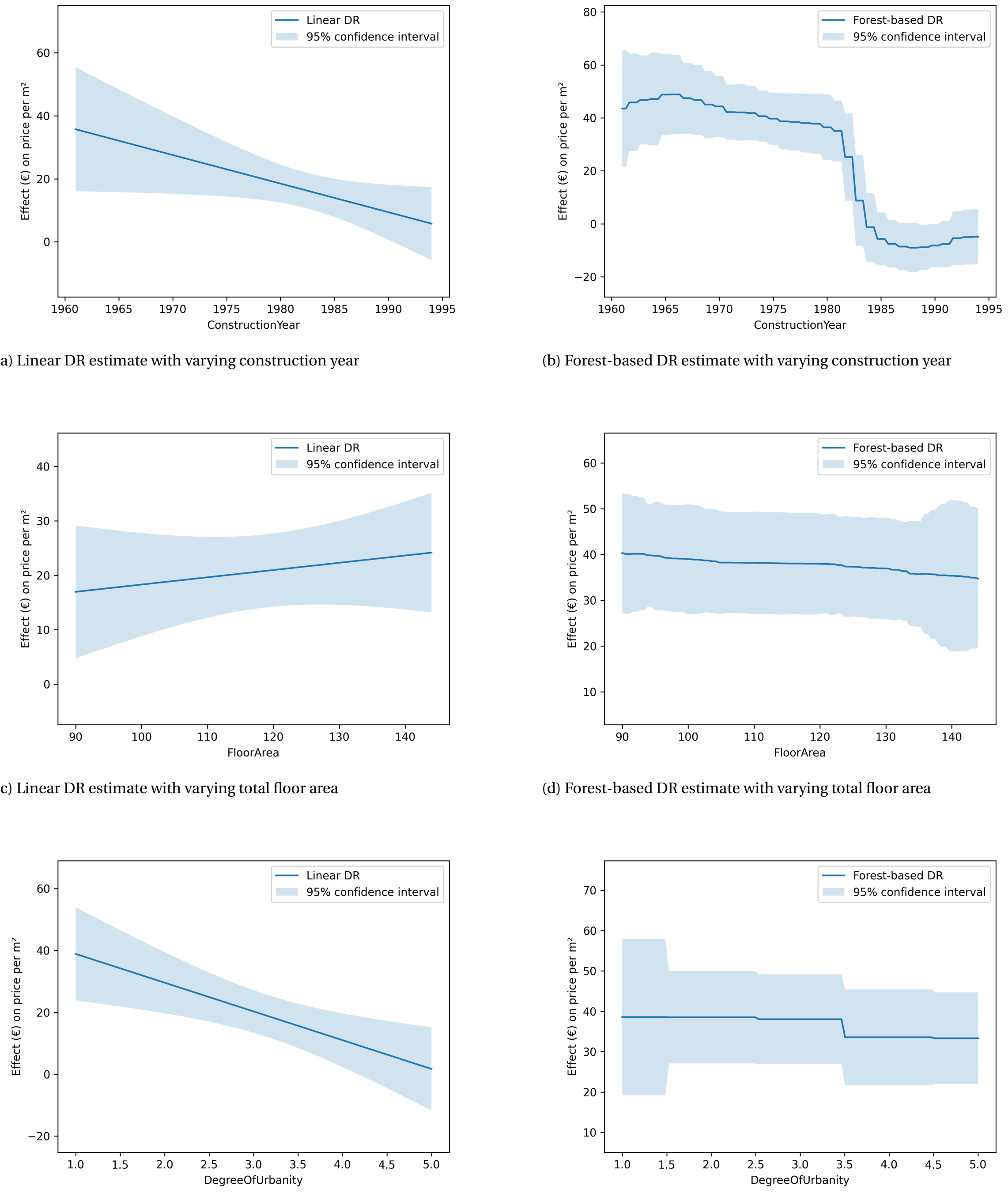
$$\delta_{EE}(x) = 1828.54 - 0.91 \cdot \text{Construction year} - 9.29 \cdot \text{Degree of urbanity} - 0.13 \cdot \text{Total floor area}, \tag{5.1}$$

$$\delta_{EI}(x) = 2370.58 - 1.15 \cdot \text{Construction year} - 0.42 \cdot \text{Degree of urbanity} - 0.28 \cdot \text{Total floor area}, \tag{5.2}$$

where $\delta_{EE}(x)$ and $\delta_{EI}(x)$ denote the treatment effect functions for the energy efficient data set and the energy inefficient data set, respectively. The linear DR model assumes that the treatment effect is linear, which is likely not the case. As a result, the magnitude of the coefficients should not be interpreted as the absolute truth, but rather provide an intuition of the impact of the corresponding variables on the treatment effect.

The total floor area does not have an impact on the treatment effect that is significantly different from 0 on the 5% level in both data sets. This indicates that the effect of increasing energy efficiency of a dwelling on its price is not significantly affected by the dwelling its total floor area. The construction year of a dwelling is estimated to have a negative impact on the treatment effect for both data sets, implying that newer dwellings on average do not increase in price as much as older dwellings when the energy efficiency is improved. The degree of urbanity is estimated only to be significant for the energy efficient data set, and has a negative sign. Hence, indicating that for dwellings in neighborhoods with a higher density of dwellings, the effect of increasing the energy efficiency of a dwelling from moderate to good is smaller than for dwellings in lower house density neighborhoods.

In order to evaluate the forest-based DR estimator, which is non-parametric, a dwelling with median values for all characteristics X of the corresponding data sets are assumed. The effect of changes in a single variable value on the treatment effect is evaluated. These effects are presented in the plots below.



(a) Linear DR estimate with varying construction year

(b) Forest-based DR estimate with varying construction year

(c) Linear DR estimate with varying total floor area

(d) Forest-based DR estimate with varying total floor area

(e) Linear DR estimate with varying degree of urbanity.

(f) Forest-based DR estimate with varying total floor area

Figure 5.7: Conditional average treatment effect estimation for three variables in the energy efficient data set. The y-axis indicates the average expected price increase in € per m^2 when the energy efficiency of a dwelling is increased from moderate (EPC label C or D) to good (EPC label A or B). The left side plots indicate the estimates made by the linear DR estimator, and the right side plots indicate estimates made by forest-based DR estimator. The variables that are varied are, from top to bottom, the construction year, the total floor area and the degree of urbanity. All other variables values are fixed at the median value.

The linear and forest-based DR estimator generally seem to agree on the trend and magnitude of the treatment effect. For all scenarios, the conditional average treatment effect is estimated to be positive, which is in line with the expectation.

The treatment effect estimates are fairly constant with respect to the total floor area for both estimators. For a dwelling with median values for all its characteristics, but with a varying total floor area, the estimated treatment effects do not change much.

While the forest-based estimated treatment effect is fairly constant for a dwelling with median values for every X and a varying degree of urbanity, the linear DR estimator estimates a downward trend. This downward trend indicates smaller treatment effects when the density of dwellings in a neighborhood is larger.

Both the linear and forest-based estimators identify that the older a dwelling is, the larger the effect of increasing the energy efficiency on the price. The forest-based DR estimator estimates a drastic change in treatment effect for dwellings with a construction year between 1980 and 1985. The forest-based estimates imply that the transaction price of dwellings with a construction year after 1985 do not benefit from increased energy efficiency at all. This is counter-intuitive, and consequently these estimates have been investigated further.

In order to explain these estimates, the following three hypotheses have been constructed and tested. The first two hypotheses refer to the fact that the plot shows estimates on a dwelling with median characteristic, while varying only one characteristic. It could be the case that median characteristics of the whole sample are not representative for median values of a smaller subsample.

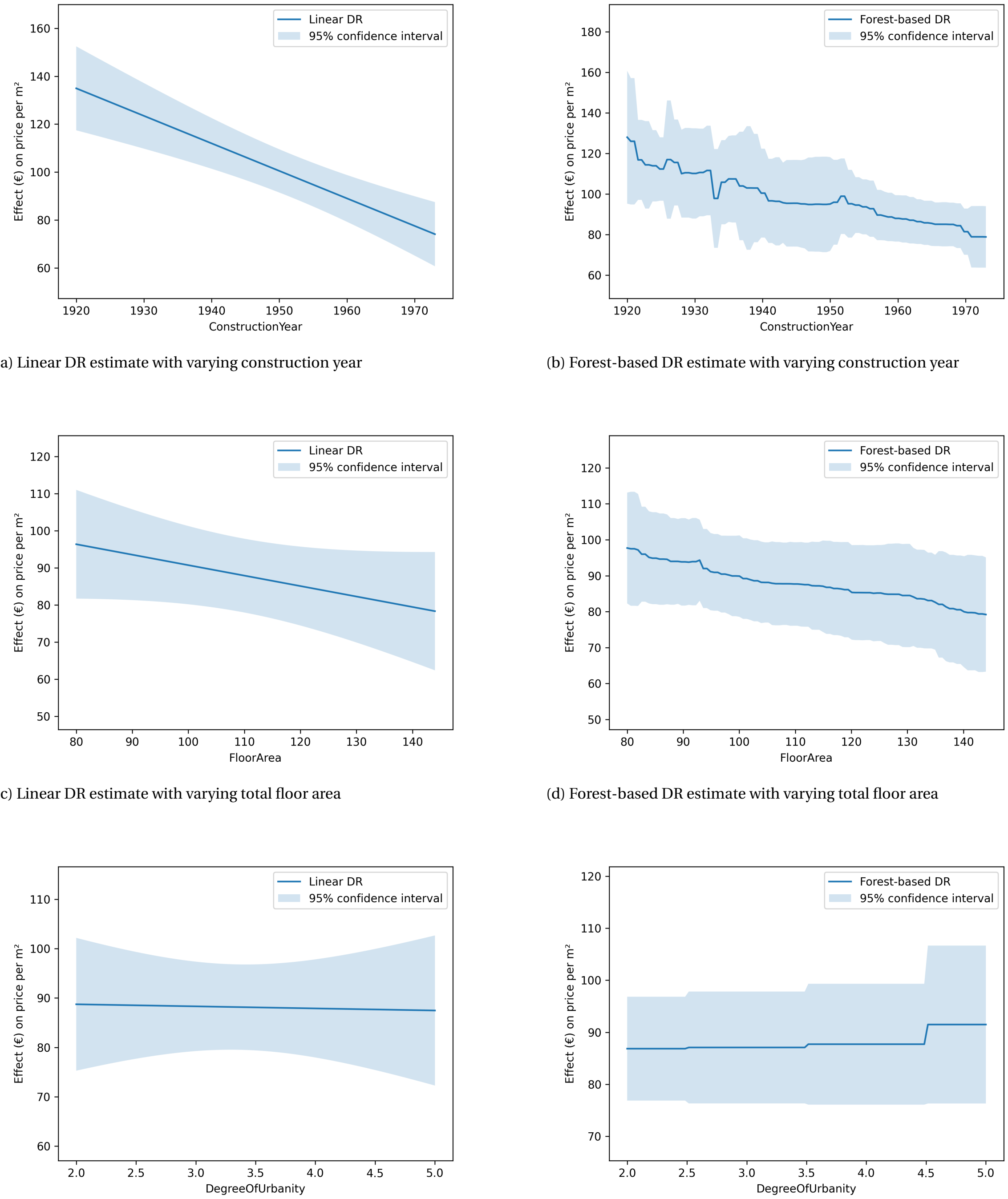
1. Median values of the characteristics of the whole sample are not representative for the median values of characteristics of newer houses.

2. Median values of the characteristics of the whole sample are not representative for the median values of characteristics of older houses.

After investigation, the median characteristics of the sample of dwellings with a construction year after 1984 are very similar to the median characteristics of the sample as a whole. Similarly, dwellings with a construction year before 1981 are very similar to the median characteristics of the entire sample. Hence, it is not the case that the median characteristics of the whole sample are not representative for the smaller subsets. Hypotheses 1 and 2 are rejected.

A possible explanation could be that for dwellings with a construction year after 1984, the probability of some level of energy efficiency is very difficult to model. The only variable that has significant predictive power for such dwellings is the construction year. As a result, due to the randomness of the individual decision trees used, the probability of a certain level of energy efficiency for these dwellings are essentially random, conditional on the construction year. When in this subsample the estimated probabilities of a treatment status is random, an obvious consequence is that the effect is zero. It is difficult to provide a good solution for this problem, other than adding additional variables to the data that have predictive power for the energy efficiency status of a dwelling.

A similar analysis is performed for the energy inefficient data set, and is presented in Figure 5.8.



(a) Linear DR estimate with varying construction year

(b) Forest-based DR estimate with varying construction year

(c) Linear DR estimate with varying total floor area

(d) Forest-based DR estimate with varying total floor area

(e) Linear DR estimate with varying degree of urbanity.

(f) Forest-based DR estimate with varying total floor area

Figure 5.8: Conditional average treatment effect estimation for three variables in the energy inefficient data set. The y-axis indicates the average expected price increase in €per m^2 when the energy efficiency of a dwelling is increased from bad (EPC label E, F or G) to moderate (EPC label C or D). The left side plots indicate the estimates made by the linear DR estimator, and the right side plots indicate estimates made by forest-based DR estimator. The variables that are varied are, from top to bottom, the construction year, the total floor area and the degree of urbanity. All other variables values are fixed at the median value.

Generally, both estimators again agree on the trend of the impact of the evaluated variables on the effect of improving the energy efficiency. The estimated effects on the price per m^2 are significantly larger, however, for the energy inefficient data set. This implies that increasing the energy efficiency of a dwelling from bad to moderate has a significantly larger estimated expected effect on the price than increasing the energy efficiency from moderate to good.

Similar to the energy efficient data set evaluated in Figure 5.7, the estimated effect of improving energy efficiency on the price is much smaller for newer dwellings. This trend seems to be fairly linear. For dwellings with a larger total floor area, the estimated effect of increasing the energy efficiency from bad to moderate on the price per m^2 decreases as the total floor area increases. The degree of urbanity of a dwelling does not impact this estimated effect much.

5.3.3. For which houses is the effect the largest?

In this Subsection, we will evaluate what type of dwellings increase in expected transaction price the most when the energy efficiency is improved. In order to do so, the top 1% of dwellings are examined on the three variables for which the treatment heterogeneity was estimated. The characteristics of dwellings within this top 1% are compared against the average characteristics of all dwellings in order to learn the differences. The results are presented in Table 5.7.

	Energy efficient data set		Energy inefficient data set	
	Top 1%	Sample	Top 1%	Sample
Construction Year	1960.5 ± 6.2	1976.6 ± 16.5	1908.4 ± 17.1	1952.0 ± 23.4
Degree of Urbanity	2.84 ± 1.38	3.16 ± 1.04	3.83 ± 1.42	3.51 ± 1.19
Total Floor Area	75.3 ± 11.8	117.2 ± 22.9	91.0 ± 74.0	112.7 ± 30.7
Treatment effect	59.6 ± 5.5	21.0 ± 22.2	219.1 ± 12.3	96.0 ± 29.9

Table 5.7: The characteristics of the top 1% of dwellings that benefit most from improving the energy efficiency versus the entire sample.

As can be seen, the dwellings that benefit most from improving the energy efficiency on the price per m^2 are generally smaller in total floor area. Moreover, such dwellings are also significantly older on average. The degree of urbanity does not seem to make much difference.

In further studies it may be beneficial to evaluate optimal policies for improving energy efficiency, when costs are also taken into account.

6
Conclusion, discussion and further research

6.1. Conclusion

In this thesis, the effect of increasing the energy efficiency of a house in The Netherlands on its expected transaction price is investigated. On an example, we presented several methods that can be used to estimate the price premium paid for more energy efficient dwellings.

To estimate the effects of improving the energy efficiency on the expected transaction price, we adopted the potential outcomes framework of Neyman [26] and Rubin [31]. In this framework, it is assumed that all transacted dwellings in the data set have two potential outcomes, $\{Y^{(1)}, Y^{(0)}\}$, which indicate the hypothetical transaction price when the dwelling has been sold with and without being energy efficient, of which only one status is observed in the data.

The expected difference between these outcomes, $\delta = \mathbb{E}[Y^{(1)} - Y^{(0)}]$, is referred to as the average treatment effect (ATE) of improving the energy efficiency on the transactions price. Additionally, the expected increase in transaction price conditional on a certain subsample $X = x$, denoted as $\delta(x) = \mathbb{E}[Y^{(1)} - Y^{(0)} \mid X = x]$, is referred to as the conditional average treatment effect (CATE).

Estimation of these effects of increasing the energy efficiency on the expected transaction price boils down to effectively estimating the potential outcome that is not observed for the individual unit. The Augmented Inverse Probability Weighted (AIPW) estimator aims to do this.

In the first stage, we estimate models for both potential outcomes, $Y^{(t)} = \mu_t(X) + \epsilon$ and for the propensity scores, denoted as $\mathbb{P}(T = t \mid X) = p_t(X)$. These models are both estimated with random forests and combined in the AIPW estimator. Recent research [4, 16, 27, 36] has combined semiparametric theory of the AIPW estimator with machine learning methods such as random forests to allow for valid statistical inference, even when the first-stage models are estimated with machine learning methods.

Using this research, two final models were implemented to estimate the conditional average treatment effects; a model assuming the relation between the average effect of increasing energy efficiency on the transaction price is a linear function of the characteristics of a house, and a model based on the random forest [36] that makes no parametric assumptions on the effect.

Under the assumptions discussed in Chapter 3, these final models provided the following results. The estimated average effect of increasing the energy efficiency of a dwelling on its transaction price is positive. Improving an energy inefficient house to moderately energy efficient seems to have a larger effect on the transaction price per m^2 than improving an already moderately energy efficiency to energy efficient. The estimated average effect for improving a dwelling from energy inefficient to moderately energy efficient is €97.70 ± 20.31 per m^2. The effect of improving moderately energy efficient dwellings to energy efficient is smaller, namely €20.96 ± 11.56 per m^2.

These estimated effects are significantly larger for older than for newer dwellings. Whether a dwelling is located in a neighborhood with a high density of dwellings does not significantly affect this estimated effect on the price. For moderately energy efficient dwellings, the total floor area has no significant impact on the effect of improving energy efficiency on the price per m^2.

For energy inefficient dwellings, a larger total floor area negatively impacts the effect of improving the energy efficiency on the price per m^2. This indicates that for a house that is twice the size, the total increase in absolute transaction price (not per m^2) when improving the energy efficiency is smaller than two times the increase in expected transaction price of the smaller house. This may be a consequence of the fact that expected savings on the energy bill due to improving energy efficiency often do not increase at a linear rate with the size of a house; as the energy bill consists of fixed costs, and some forms of energy consumption do not necessarily scale with house size, such as energy consumption for showering and cooking.

The top 1% of dwellings that increase most in expected transaction price when the energy efficiency is improved are generally older and smaller in comparison with the sample as a whole. Improving the energy efficiency of this top 1% is on average about 2 to 3 times as effective in comparison with the average dwelling in the sample when only the price increase of a house is considered. Knowing this can help policy makers improve incentives for making houses more energy efficient, and can stimulate homeowners to improve energy efficiency of their homes.

6.2. Discussion and directions for further research

The estimated effect of improving the energy efficiency by the AIPW estimator is quite significant. However, one should take into account the assumptions discussed in Chapter 3 when evaluating the results, and in particular the unconfoundedness assumption. The unconfoundedness assumption (3.9) that is made in this thesis is a rather strict assumption, which additionally is difficult to test. The unconfoundedness assumption implies that there are no unobserved variables that have significant impact on the energy efficiency as well as on the price of a dwelling. One could argue, however, that homeowners often simultaneously improve the energy efficiency of a dwelling, as well as improve its quality by renovating. As a result, the estimated effect of improving energy efficiency on the price may partly be contributed to the improved quality of a dwelling due to renovations. This seems even more likely for older dwellings, where our estimated effect was in general larger. Consequently, one should be careful to draw strong conclusions without explicitly taking this assumption into account.

Including more variables that impact the transaction price and the probability of a house being energy efficient may help motivating conformance with the unconfoundedness assumption. Additionally, the estimates produced by the AIPW estimator would get more accurate if more meaningful variables are added. In order to get access to more important variables, it would be possible to use Natural Language Processing (NLP) techniques on house advertisement websites in order to infer useful variables regarding the dwelling from the advertisement text. Such useful variables could for instance be the approximate quality of the dwelling, or whether it has recently been renovated. Additionally, pictures of dwellings could be analyzed by image recognition techniques to gain more accurate estimates of such variables.

Alternatives to the unconfoundedness assumption, however, are not easily attainable. One could investigate repeated transactions of the same house, that has improved its energy efficiency in between the transactions. In theory, in doing so one could immediately estimate a price difference on a house that has had most of its characteristics fixed over time. Such an approach has been carried out by some researchers in the past. For instance, Bruegge et al.[12] investigated the willingness to pay for an 'Energy Star' (the American equivalent of a green EPC label) in Florida using repeated sales of houses. A similar study was performed by Fuerst et al. [22] in England. However, for the data available to us this approach suffers from similar problems as mentioned above; as quality is unobserved, improvements in energy efficiency may be biased upwards due to renovations that happened simultaneously. Additionally, this method initiates a new problem of estimating the price trend of different locations over time, which is a vastly complex task. Moreover, the sample of houses that has had multiple sales is extremely small in comparison to all sold houses, which leads to much higher variance in the estimates. Lastly, one could argue that this small sample of houses with repeated transactions over a small time frame are not representative for the housing market as a whole.

Problems in a causal inference settings arise, because it is never possible to observe the change in transaction price for an individual house from observed data, because a house is either energy efficient or not at the time of sale. Consequently, evaluating and validating the estimated effects is extremely challenging. Methods for evaluating estimated treatment effects are of broad and current interest in causal inference research. The importance is stressed by the study of Schuler et al. [32], which compares different treatment effect evaluation metrics with the known ground truth in a simulation study. Methods for verification of causal inference assumptions and causal graph discovery are quickly evolving research areas [25, 33]. These research areas could help verifying the assumptions that are made in such a setting.

In order to avoid having to deal with the causal inference assumptions made in this thesis at all, one could research the preference of people participating in the housing market with survey-like instruments. A substantial benefit is that in a survey we can create hypothetical situations. Therefore, for a single house, respondents can estimate both an energy efficient kind, as well as an energy inefficient kind, that are otherwise completely equal. If we relate this setting to the causal inference setting in our problem, in a hypothetical situation people can directly estimate both potential outcomes, $Y^{(1)}$ and $Y^{(0)}$ for the exact same house, because in a hypothetical situation those outcomes can exist simultaneously for the same house. These preferences can then again be modeled in several different ways. Such an approach was performed by Banfi et al.[7] to estimate the willingness to pay for individual energy-saving measures in residential buildings in Switzerland using a fixed-effects logit model.

In further research, it would be beneficial to additionally model the cost-side of the investment to improve energy efficiency in a similar manner. Additionally, a discounted cash flow (DCF) method could be used to value the expected savings on the energy bills for different dwellings. When the total costs, and the total revenues consisting of savings on the energy bill and the increase in expected transaction price are modeled, it would be possible to estimate optimal policies for improving energy efficiency. This information could be used by the government to nudge homeowners with detailed investment opportunities. Moreover, for homeowners the investment to improve energy efficiency of their house can be evaluated with less effort.

A

Descriptive Statistics

	mean	std	min	25%	50%	75%	max
AmtOfBathrooms	1.04	0.19	1.0	1.00	1.00	1.00	5.00
AmtOfFloors	2.91	0.46	1.0	3.00	3.00	3.00	9.00
AmtOfRooms	4.99	0.91	1.0	5.00	5.00	5.00	45.00
AvgHeight	3.33	0.41	0.0	3.06	3.30	3.51	9.99
ConstructionYear	1976.67	16.37	1800.0	1971.00	1978.00	1988.00	1997.00
DegreeOfUrbanity	3.16	1.04	1.0	2.00	3.00	4.00	5.00
EnergyEfficiency	0.24	0.43	0.0	0.00	0.00	0.00	1.00
FloorArea	117.01	23.23	27.0	103.00	115.00	129.00	635.00
LotArea	193.67	489.54	17.0	140.00	165.00	219.00	999.00
SalePricePerM2	1653.79	355.78	146.0	1419.31	1617.73	1840.48	6691.04
hasBasement	0.11	0.31	0.0	0.00	0.00	0.00	1.00
hasGarage	0.20	0.40	0.0	0.00	0.00	0.00	1.00
hasLoft	0.26	0.44	0.0	0.00	0.00	1.00	1.00
isCity	0.33	0.47	0.0	0.00	0.00	1.00	1.00
isCornerHouse	0.28	0.45	0.0	0.00	0.00	1.00	1.00

Table A.1: Descriptive statistics of all variables in the energy efficient data set. EnergyEfficiency is a binary variable, where the mean denotes the fraction of energy efficient dwellings in the data set.

	mean	std	min	25%	50%	75%	max
AmtOfBathrooms	1.04	0.21	1.0	1.00	1.00	1.00	4.00
AmtOfFloors	2.93	0.53	1.0	3.00	3.00	3.00	9.00
AmtOfRooms	4.91	1.08	1.0	4.00	5.00	5.00	39.00
AvgHeight	3.37	0.46	0.0	3.08	3.34	3.57	9.98
ConstructionYear	1952.01	23.14	1800.0	1936.00	1961.00	1970.00	1974.00
DegreeOfUrbanity	3.52	1.19	1.0	3.00	4.00	4.00	5.00
EnergyEfficiency	0.57	0.50	0.0	0.00	1.00	1.00	1.00
FloorArea	112.19	29.93	21.0	92.00	110.00	126.00	649.00
LotArea	192.73	483.75	18.0	134.00	161.00	211.00	999.00
SalePricePerM2	1688.40	450.46	146.0	1387.64	1618.47	1909.28	9352.71
hasBasement	0.17	0.37	0.0	0.00	0.00	0.00	1.00
hasGarage	0.17	0.38	0.0	0.00	0.00	0.00	1.00
hasLoft	0.30	0.46	0.0	0.00	0.00	1.00	1.00
isCity	0.36	0.48	0.0	0.00	0.00	1.00	1.00
isCornerHouse	0.29	0.46	0.0	0.00	0.00	1.00	1.00

Table A.2: Table of descriptive statistics of all variables in the energy inefficient data set. EnergyEfficiency is a binary variable, where the mean denotes the fraction of moderately energy efficient dwellings in the data set.

Bibliography

[1] Woning verduurzamen - energiebesparende maatregelen. https://www.berekenhet.nl/nieuws/woning-verduurzamen-energiebesparende-maatregelen.html. Accessed: 2022-07-27.

[2] URL https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84929NED/table?dl=343E. Accessed: 2022-07-27.

[3] Milieu centraal - praktisch over duurzaam. https://www.milieucentraal.nl/. Accessed: 2022-07-27.

[4] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. The Annals of Statistics, 47 (2):1148 – 1178, 2019. doi: 10.1214/18-AOS1709. URL https://doi.org/10.1214/18-AOS1709.

[5] Erdal Aydin, Dirk Brounen, and Nils Kok. The capitalization of energy efficiency: Evidence from the housing market. Journal of Urban Economics, 117:103243, 2020.

[6] Sandra Backlund, Patrik Thollander, Jenny Palm, and Mikael Ottosson. Extending the energy efficiency gap. Energy Policy, 51:392–396, 2012.

[7] Silvia Banfi, Mehdi Farsi, Massimo Filippini, and Martin Jakob. Willingness to pay for energy-saving measures in residential buildings. Energy Economics, 30(2):503–516, 2008. ISSN 0140-9883. doi: https://doi.org/10.1016/j.eneco.2006.06.001. URL https://www.sciencedirect.com/science/article/pii/S0140988306000764.

[8] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973, 2005.

[9] Leo Breiman. Bagging predictors. Machine learning, 24(2):123–140, 1996.

[10] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[11] Dirk Brounen and Nils Kok. On the economics of energy labels in the housing market. Journal of Environmental Economics and Management, 62(2):166–179, 2011.

[12] Chris Bruegge, Carmen Carrión-Flores, and Jaren C Pope. Does the housing market value energy efficient homes? evidence from the energy star program. Regional Science and Urban Economics, 57:63–76, 2016.

[13] Marcelo Cajias and Daniel Piazolo. Green performs better: energy efficiency and financial return on buildings. Journal of Corporate Real Estate, 2013.

[14] Andrea Chegut, Piet Eichholtz, and Rogier Holtermans. Energy efficiency and economic value in affordable housing. Energy Policy, 97:39–49, 2016.

[15] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

[16] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

[17] Samuel R Dastrup, Joshua Graff Zivin, Dora L Costa, and Matthew E Kahn. Understanding the solar home price premium: Electricity generation and “green” social status. European Economic Review, 56 (5):961–973, 2012.

[18] Terry M Dinan and John A Miranowski. Estimating the implicit price of energy efficiency improvements in the residential housing market: A hedonic approach. Journal of Urban Economics, 25(1):52–67, 1989.

[19] Piet Eichholtz, Nils Kok, and John M Quigley. Doing well by doing good? green office buildings. American Economic Review, 100(5):2492–2509, 2010.

[20] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. Graphviz—open source graph drawing tools. In International Symposium on Graph Drawing, pages 483–484. Springer, 2001.

[21] Franz Fuerst and Patrick McAllister. Green noise or green value? measuring the effects of environmental certification on office values. Real estate economics, 39(1):45–69, 2011.

[22] Franz Fuerst, Patrick McAllister, Anupam Nanda, and Peter Wyatt. Does energy efficiency matter to home-buyers? an investigation of epc ratings and transaction prices in england. Energy Economics, 48: 145–156, 2015.

[23] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. American journal of epidemiology, 173(7):761–767, 2011.

[24] Robert W Gilmer. Energy labels and economic search: an example from the residential real estate market. Energy Economics, 11(3):213–218, 1989.

[25] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. Frontiers in genetics, 10:524, 2019.

[26] Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Roczniki Nauk Rolniczych, 10(1):1–51, 1923.

[27] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In International Conference on Machine Learning, pages 4932–4941. PMLR, 2019.

[28] Microsoft Research. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML, 2022. Version 0.13.1.

[29] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427):846–866, 1994.

[30] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

[31] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688, 1974.

[32] Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. arXiv preprint arXiv:1804.05146, 2018.

[33] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. arXiv preprint arXiv:2011.04216, 2020.

[34] Anastasios A Tsiatis. Semiparametric theory and missing data. 2006.

[35] Centraal Bureau voor de Statistiek. Welke sectoren stoten broeikasgassen uit? https://www.cbs.nl/nl-nl/dossier/dossier-broeikasgassen/welke-sectoren-stoten-broeikasgassen-uit-. Accessed: 2022-07-27.

[36] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

[37] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. The Journal of Machine Learning Research, 15(1):1625–1651, 2014.

[38] Margaret Walls, Todd Gerarden, Karen Palmer, and Xian Fang Bak. Is energy efficiency capitalized into home prices? evidence from three us cities. Journal of Environmental Economics and Management, 82: 104–124, 2017.

[39] Jiro Yoshida and Ayako Sugiura. Which “greenness” is valued? evidence from green condominiums in tokyo. 2010.