Using machine learning to study heterogeneity in the adoption of clean technologies in neighbourhoods

Kolen, A.M. (Sandra)

Award date:
2024

Using machine learning to study heterogeneity in the adoption of clean technologies in neighbourhoods

A.M. Kolen | 0815377

May 11, 2024

Prof. dr. T.A. Arentze
Dr. I.V. Ossokina
Ir. A.W.J. Borgers

7Z45M0 Graduation project Urban Systems & Real Estate (45 ECTS)
Eindhoven University of Technology
Master Architecture Building and Planning
Master track Urban Systems and Real Estate
This graduation thesis is publicly available
This project has been carried out with support from the MMIP 3 & 4 grant of the Dutch Ministry of Economic Affairs & Climate and the Dutch Ministry of the Interior & Kingdom Relations. The author is grateful to foundation Buurkracht for collaboration and providing access to the data.
This thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Integrity

Summary

The adoption of clean energy technologies in households is crucial in combating climate change. This energy transition implies large investments on the part of homeowners. One of the ways to stimulate and support homeowners in the transition are community-led initiatives to collectively purchases clean home technologies.

Neighbourhood dynamics play an important role in the adoption patterns of households and neighbourhoods, driven by factors like peer influence and socio-economic disparities. Understanding these dynamics is essential for effective policy implementation and targeted interventions. This thesis explores econometric and machine learning techniques to find out why some neighbourhoods are more successful in activating home owners to adopt than others, and how this is related to heterogeneity (i.e. differences) between residents and neighbourhoods. Therefore the research question of this thesis is: *Which econometric or machine learning technique is the preferred method to gain insight into the heterogeneity within and between various demographic groups in the adoption of clean energy technologies in neighbourhoods?* With this research question also 4 sub-questions are formulated: How can econometric and machine learning techniques be utilized to identify and explain the heterogeneity among various demographic groups in the adoption of clean energy technologies in neighbourhoods?, What are the comparative strengths and weaknesses of econometric and machine learning techniques in explaining heterogeneity among various demographic groups in the adoption of clean energy technologies in neighbourhoods?, and How do the predictive accuracy of econometric and machine learning techniques compare in the context of clean energy technology adoption, as assessed by evaluation metrics?

To answer the research question and sub-questions, first, a literature review was conducted. This review resulted in a conceptual model and a list of relevant variables that do influence energy use and the adoptions of clean technologies. These variables contain dwelling characteristics, household characteristics and neighbourhood characteristics that may cause heterogeneity in adoption probability. They are expected to have an effect on the decision to apply an energy efficient measure. How large the effect is is studied empirically. In order to build and test empirical models, data is required. For this thesis data from the foundation Buurkracht was used in combination with data on postal code level from Statistics Netherlands. Buurkracht is a foundation that supports collective purchases in various Dutch communities. The data from Buurkracht contains information about 74 thousand households across 82 communities. The data from Statistics Netherlands includes information on dwelling characteristics, gas use and electricity use.

Three techniques for explaining heterogeneity in adoption probability were compared: logistic regression, random forest and causal forest. The logistic regression is user-friendly, does not require much computational power and gives information on the direction and importance of the variables, however the method is prone to overfitting, a lot of assumptions should be met regarding the data and complex relationships are hard to capture. The random forest is more robust and less prone to overfitting and also provides an indication of the importance and direction of the effects of variables. The results are more difficult to interpret and more computational power is required. The logic from the model cannot be extracted, it is a black box model. The causal forest is especially focused on the treatment effect, which is defined as the presence of an initiator of the neighbourhood approach within 200 meters. This makes direct comparisons with other models challenging, since those models do not incorporate a similar treatment definition. The inclusion of the treatment enables us to investigate the heterogeneity around this one specific variable. The causal forest is more complex and requires more computational power. A Poisson regression was also used, a Poisson regression is specifically used for count variables and in this thesis the technique was used on a different level, community instead of household, hence this technique was not compared to the other techniques.

Empirical application of the models, unfortunately, produces inconclusive results: the models do not succeed to generate consistent results in terms of the impact of variables on adoption and fail to predict adoption probabilities well. The results from the logistic regression generally aligned the best with the intuitive expectations. The machine learning models, random and causal forest, succeed to predict non-adoption well,

but fail to predict the adoption. Digging into possible reasons of the poor model performance, the problem of an unbalanced dataset is identified. The data contains many household that do not adopt a technology and only very few households that do adopt a clean energy technology. A combination of an over and under sampling method is used to tackle the problem, however this does not lead to an improved performance. At the same time, the parametric logistic and Poisson regressions suggest that there exists a large unobserved community-specific variance that complicates prediction.

The predictive performance of the models has been compared using 5 evaluation metrics: accuracy, precision, recall, specificity and F1 score. The F1 score is the harmonic mean of precision and recall and useful in the case of imbalanced data. The predictive performance is tested based on both the original dataset and a dataset created with 20% households applying a measure and 80% not applying a measure. The independent variables that are included in the models are: gas use, electricity use, distance to initiator, living area, dwelling value and construction year. Recall is seen as the most important metric, since this is based on the correctly predicted positive outcomes. The causal forest performs the best on the recall metric, but for all models the recall score is very low. On the newly constructed data set (20% measure) the random forest performs the best on all evaluation metrics.

Unfortunately, this research did not identify a method that is preferred to gain insight into the heterogeneity in the adoption of clean energy technologies in neighbourhoods. The causal forest technique shows potential, but also its performance is far from desired. An important limitation is the data quality. Caveats of the data that were identified during the research are: class imbalance in the data regarding measure adoption, lack of household-specific data, a large unobserved community-specific variance, inability to measure community treatment effect since no data is available on households that were not involved in a community campaign and finally data on pre-existing measures is missing. Another notable missing element in current data and models is people's norms and values, which plays an important role in shaping their choices. Incorporating such factors into the models could improve their predictive accuracy and understanding of the decision to apply an energy efficient measure. It is recommended to replicate the study with better data and also other methods should be explored, since the field of machine learning and econometrics is much broader than the part that has been dealt with in this thesis.

Contents

1 Introduction

This chapter introduces the topic of the thesis. It begins by highlighting the importance of adopting clean technologies in households, followed by a focus on the neighbourhood related aspects of adoption. Subsequently, it delves into the heterogeneity between socio-economic groups concerning adoption probability and speed. Finally, the main research question and sub-questions are introduced, and an outline of the thesis is provided.

1.1 Importance of adoption of clean technologies in homes

Climate-change issues and greenhouse gas emissions are a prominent concern in the World. In the Netherlands the biggest challenges lie in the rising sea levels, dry springs and summers and the extreme summer rainfall KNMI (2021). In the Dutch climate act of June 28, 2019, (Rijksoverheid (2019)) it is stated that by 2030 CO_2 emissions should be 49% less than the levels of 1990 and by 2050 a 95% reduction of green house gas emissions should have taken place, compared to the levels of 1990. In 2050 the current government of the Netherlands wants the country to be climate neutral, only sustainable energy can be used. Greenhouse gas emissions in 2020 were 25.5 % below the level of 1990 (Ruyssenaars et al., 2022). This decrease in emission levels shows a trend in the right direction, but is not enough to reach the set goals.

By 2050 7 million dwellings and 1 million other buildings should no longer be connected to the natural gas supply. In addition, these buildings and dwellings must be better insulated, heated by sustainable sources and only clean electricity (such as solar power) can be used in these dwellings and buildings (EZK, 2019). These goals for the built environment are stated in the climate act. As a first step, 1.5 million existing homes will be made more sustainable until 2030 (Klimaatakkoord (nd)). This will be done neighbourhood by neighbourhood, the municipalities determine together with the residents and homeowners what the best solution is for their neighbourhood: heat pumps, fully electric, heat network or something else.

1.2 Neighbourhood adoption of clean technologies

Energielinq (2019), a hub for knowledge, experience and experts regarding energy renovations and new constructions, gives three main reasons why it has been decided to use a neighbourhood approach:

- buildings in one neighbourhood are often from the same construction period and hence the constructions have similar features making the sustainability opportunities comparable
- the surroundings are important in determining the opportunities, for example the proximity of a heat network creates the opportunity to extend the heat network to the neighbourhood
- the neighbourhood level is the level on which people know each other and it is a level on which people can communicate with each other about possibilities and preferences.

Next to these reasons, a neighbourhood approach can also result in scale benefits leading to economic benefits. Companies like Duurzame-VvE and De Energiebespaarders are promoting the adoption of clean technologies in dwellings by emphasising positive effects on comfort and health, stable and lower living costs, longer dwelling life span, a modern look, positive effects on the environment, a healthier living environment, a better future for the next generation, an increased dwelling value and an improved marketability.

Research has shown that in a number of ways peers/neighbours have positive effects on each other in the adoption of clean technologies. Ways in which neighbours influence each other are by word of mouth or by visibility of taken measures. In marketing word of mouth is considered free advertising triggered by customer experiences, where satisfied customers reflect this in their daily dialogues. The majority of people say they trust recommendations from friends and family above other forms of advertising (Freedman (2024)). Visibility is also an important concept. It is needed to increase the rate at which a brand and in

this case a measure is seen by the audience. By increasing the overall exposure more people are aware of the measures which triggers the adoption of the measures by more people. In most studies it is not clear which of these mechanisms is the cause of increasing adoption rates. Both mechanisms are at work when PV systems are implemented at a close distance. According to Bollinger and Gillingham (2012) both visibility of PV panels and word of mouth increase the adoption of PV panels. Graziano and Gillingham (2014) show that more adoptions of PV systems in the previous 6 months leads to an increase in the number of PV system adoptions in the next period within a certain distance of the systems. They find that in more dense neighbourhoods the adoption of PV systems is increasing. Also in neighbourhoods with a larger share of owner-occupied dwellings the adoption of PV systems is increasing. Bollinger et al. (2024) show that short and long campaigns are equally effective in promoting solar panels while they run, but longer campaigns generate more word of mouth leading to higher adoption rates. Bollinger et al. (2022) and Rode and Weber (2016) also show that the proximity of visible PV panels increases the adoption rate. The diffusion of PV panels depends on two types of costs: the costs of the actual systems and the cost of information to the consumers. Large amounts of information have to be collected related to the systems, future performance, maintenance requirements, quality and opportunities. To reduce uncertainties, interested parties take advantage of information from existing owners. These peer effects significantly decrease decision times (Rai and Robinson, 2013).

Community-based energy projects are increasing in popularity. These projects are defined as "organisations, initiated and managed by actors from civil society, that aim to educate or facilitate people on efficient energy use, enable the collective procurement of renewable energy or technologies or actually provide (i.e. generate, treat or distribute), energy derived from renewable resources for consumption by inhabitants, participants or members" (Boon and Dieperink, 2014). Kalkbrenner and Roosen (2016) report that community identity, trust, social norms and higher environmental concern are positively associated with the willingness to participate in community energy projects. They show that ownership of a renewable energy system and living in a suburban, or rural, rather than urban, community increases the likelihood of participation. Van der Schoor and Scholtens (2015) have found that for an community-based energy project to be successful the organizations need to entertain strong and continuous relations both on the local as well as on the global level. The level of activities is an important indicator of the effectiveness of a community-based approach. They see that on the smaller local scale clear local energy goals are missing, where on the scale of the municipality the active participation of citizens is lacking, but there is a more elaborate vision. They conclude that the community energy initiatives provide a useful approach for many citizens to engage in the transition to a sustainable energy future. However the further development of organisation structures and viable visions for local energy governance is necessary to achieve lasting results. Casteren et al. (2013) show that in the case of community-led retrofits the proximity to the residents who manage the community retrofit has a positive effect on other community member's willingness to retrofit. This positive effect arises through easy access to information about the technologies and by active campaigning. People living within 200 meter proximity to the managing residents have an up to 4 times higher probability to retrofit, compared to the average of their community.

1.3 Heterogeneity between socio-economic groups in adoption probability and adoption speed

Heterogeneity can be described as the presence of diversity or variations within a group, population, or system. Neighbourhoods and communities are important in the process of getting households to adopt clean technologies, but not all neighbourhoods are the same (heterogeneity between neighbourhoods). Neighbourhoods consist of households with a different willingness/ propensity to adopt clean technologies in their home(heterogeneity within neighbourhoods). People have a different willingness to pay for different positive aspects of clean technologies. Some people are focused on the positive aspects for health, some on the positive aspects on their spending, some feel like they have to preserve the planet for the future generations.

All these different reasons make that different people have a different adoption probability and adoption speed, illustrating the heterogeneity among households. It is important to understand which type of people are more likely to adopt through community projects in order to target the households which lead to the greatest successes.

One of the factors that has an influence is the socio-economic status. Important socio-economic factors are occupation, education, income and wealth. Brounen et al. (2012) showed that income has a big influence on energy use. De Groote et al. (2016) show that average income and the value of a house have a positive effect on the number of PV installations. Another thing they find and that is also shown in other research (Hammerle et al. (2023)) is that PV installations are more often installed on owner occupied dwellings than on rented dwellings.

The willingness to adopt clean technologies positively correlates with the willingness to behave environmental friendly. Different studies have examined the relationship between environmental concern and different socio-demographics. The results are unfortunately not unambigious. Mehmetoglu (2010) and Hersch and Viscusi (2006) find that younger people are more willing to behave environmentally friendly. However Bollinger and Gillingham (2012) find that people below 45 years or above 65 years old have a lower adoption rate for PV panels and Kwan (2012) finds that the share of PV panels decreases in age groups 25-34 and 55-64. A higher education level has a positive influence on pro-environmental behavior (Mehmetoglu, 2010) and (Casaló and Escario, 2018). Hunter et al. (2004) and Xiao and McCright (2015) find that women express greater concern for the environment. A research performed in Italy by Besagni and Borgarello (2018) showed that socio-economic characteristics have a higher explanatory power for electrical energy expenditures compared to dwelling characteristics and appliances. The building characteristics and variables are however important in determining the thermal energy expenditures. Since socio-economic characteristics have high explanatory power for electrical energy expenditures, this could also mean that they have a high influence on adoption probability or adoption speed.

1.4 Research question

It is important to find ways to improve the adoption of clean technologies in homes, at the current pace the set goals from the climate act will not be reached. Improving adoption rates can be done by using neighbourhood approaches. Neighbourhood adoption is receiving much attention from policy makers, but not all households and neighbourhoods are equally susceptible to particular approaches. Heterogeneity within and between neighbourhoods is very important in predicting the adoption probability of clean technologies for groups. Should heterogeneity not be considered then approaches with good results, high adoption rates, in one neighbourhood, could result in wrong predictions for other neighbourhoods. In this thesis the heterogeneity within and between neighbourhoods will be investigated. This will be done by using econometric techniques and machine learning techniques. The potential of both techniques will be investigated in order to find the target group that will lead to the best results in adoption of clean technologies when targeted. Therefore the research question of this thesis is:

Which econometric or machine learning technique is the preferred method to gain insight into the heterogeneity within and between various demographic groups in the adoption of clean energy technologies in neighbourhoods?

In order to answer the main research question the following sub-questions will be answered:

1. How can econometric and machine learning techniques be utilized to identify and explain the heterogeneity among various demographic groups in the adoption of clean energy technologies in neighbourhoods?

2. What are the comparative strengths and weaknesses of econometric and machine learning techniques

in explaining heterogeneity among various demographic groups in the adoption of clean energy technologies in neighbourhoods?

3. How do the predictive accuracy of econometric and machine learning techniques compare in the context of clean energy technology adoption, as assessed by evaluation metrics?

1.5 Academic and Practical Relevance

Academic Relevance

The neighbourhood approach is gaining popularity, however, its optimal utilization is currently hindered by insufficient research on its outcomes and effectiveness. This thesis aims to address this gap by developing predictive models to identify which independent variables, such as neighbourhood, household and dwelling characteristics, act as predictors of a successful neighbourhood approach. Through an exploration of various econometric and machine learning models, this thesis seeks to enhance our understanding of the determinants of a successful neighbourhood approach. This study represents an initial step towards investigating the outcomes and effectiveness of the neighbourhood approach. At the end, recommendations for improving both modeling techniques and data collection processes are provided to contribute to the advancement of research in this field.

Practical Relevance

Neighbourhood approaches have the potential to contribute significantly to increasing the sustainability of the dwelling stock. This research contributes to a better understanding of the effectiveness of neighbourhood approaches on specific neighbourhoods. A first step in identifying characteristics of dwellings, households or neighbourhoods that lead to an increased chance of the adoption of energy efficient measures after a neighbourhood approach is made. By getting a clearer image of relationships between the characteristics and the success of a neighbourhood approach, groups with potentially higher adoption rates can be targeted. This will lead to an increase in energy efficiency measures in homes, which should then lead to a decrease in the greenhouse gas emission in the Netherlands.

1.6 Outline of the thesis

This thesis is comprised of 10 chapters. Chapter 2 provides background of the research. Current policies to expedite the energy transition are discussed, who do they target and what are the effects. Various studies where econometric and machine learning methods were applied to predict and explain energy use and energy preservation in relation to heterogeneity of socio-economic groups are discussed. Chapter 3 presents the research design. In chapter 4 the theory of different regression methods is given. In chapter 5 machine learning methods are explained. Next in chapter 6 the data is discussed and descriptive statistics are given. Chapter 7 provides the results from the applied models. In chapter 8 the main issue that results in poor model performance is identified and further explained, methods to deal with the issue are given and the same models as before are applied to a newly created dataset, the results are given and discussed. In the next chapter the models are compared with each other based on five evaluation metrics. Finally, in the last chapter conclusions are drawn and recommendation for future research are made.

2 Background of the research

The importance of the adoption of clean technologies is not new, the adoption has been going on for a considerable amount of time and remains important. To encourage adoption, various policies have been introduced. This chapter will discuss a selection of these policies, starting with those targeting individuals and followed by those promoting group adoption. The chapter will explore the effects of these policies and their differential impacts across socio-economic groups. Additionally, previous research utilizing econometric and machine learning methods to study energy use and energy preservation in relation to heterogeneity in socio-economic groups will be examined.

2.1 Policies stimulating adoption

As a first step in order to meet the set goals for 2050 of the Dutch climate act, 1,5 million dwellings will be made more sustainable by 2030 (EZK (2022)). For specifically the group dwelling owners EZK (2022) states there are six policies of the dutch government that stimulate reaching these goals. These 6 policies are:

- Solar panel rebate scheme for feeding solar energy back into the grid. With this scheme home owners are allowed to deduct the energy that they supply back to the grid from their own energy use.
- Energy saving loan from "het warmtefonds" that can be used to make ones dwelling more sustainable against favourable conditions.
- A value added tax (VAT) refund on solar panels. The 21% VAT on purchase and installation is refunded.
- A programme called "Programma Aardgasvrije Wijken" has been set up in order to assist municipalities in transitioning neighbourhoods off the natural gas supply.
- The website verbeterjehuis.nl was built for more information on making dwellings more sustainable. On this website, you can check how well insulated your dwelling is and what sustainable improvements are possible.
- The investment subsidy for renewable energy and energy saving, in Dutch: Investeringssubsidie duurzame energie en energiebesparing (ISDE). This subsidy is meant for solar boilers, heat pumps, insulation measures and connection to a heat network. The percentage of subsidized costs rises when the number of measures rises, from 15 % for one measure to 30 % for two measures (Jorna (2024)).

Municipalities also have their own available subsidies and support for making homes in their municipality more sustainable. Naming all municipalities and their subsidies separately would get to extensive, but in order to give an idea some municipalities and their subisidies will be named. The municipality of Tilburg for example has subsidies for independent energy advice and for building green roofs, green facades and removing stones from gardens (Gemeente Tilburg (nd)). The municipality of Amsterdam has subsidies for making real estate natural gas-free (Gemeente Amsterdam (nd)). Den Haag has subsidies for roof-, floor- and facade insulation, making real estate natural gas-free, building green roofs and for cooking on electricity (Haag (nd)).

These policies and subsidies from the municipalities are however not clearly visible and the information is not easily accessible, hence making it for homeowner not easy to see where they can benefit from subsidies and financing, creating a threshold for individual homeowners to adopt clean technologies in their homes. This is one of the reasons why group buying can be beneficial.

Group buying adds value in the pre-, during- and post-acquisition stages according to Wang et al. (2013). In the pre-purchase phase information is shared. The participants collectively figure out how to best satisfy their needs. Together they learn more about the products and they generate preferences. During the negotiations this gives them a stronger position, due to collective expertise and knowledge. The role of

leaders is important in this process, they add benefit for both parties, on the one side better knowledge and experience and on the other hand better persuasion toward consumers. They mediate the dialogue between consumers and firms. Group strength is important and leads to better information and safeguarding of rights.

In the Netherlands a number of companies stimulating collective decision making have been formed. In Capelle aan den IJssel the ECC (Energie Collectief Capelle) has been formed. This collective consists of a group of volunteers who is committed to achieve energy savings, they assist the municipality in the recruitment of participants in neighbourhood actions and provide advice and guidance during the process. Other comparable companies are energiecoörperatie ONE, REL(Regionaal energieloket) and Buurkracht, all are focused on collectively making neighbourhoods more sustainable by organizing activities to inform homeowners collectively about the possibilities. Programma Aardgasvrije wijken is connecting municipalities, concerned parties and relevant stakeholders and offers support to accomplish the common goal of natural gas-free neighbourhoods.

2.2 What is known about the effects of policies

As mentioned in the previous section different policies have been implemented to stimulate adoption of clean technologies both by individuals and groups. Numerous studies have been conducted to study the effects of policies on different socio-economic groups. This field is very broad and not all effects of policies will be named here. To give some insight into possible effects of policies a number of studies will be discussed.

De Groote et al. (2016) show that local subsidies for PV have a positive and statistically significant effect on the number of PV installations. Their results suggest that subsidization can be very effective in promoting PV. They find that richer households disproportionately benefited from the subsidies in Flanders. Larger households are more likely to invest in PV installations because their savings will be bigger since they consume more energy.

Sloot et al. (2019) have researched involvement in community energy initiatives. Their research showed that even though people rate financial and environmental motives as important for involvement in community energy initiatives, environmental and communal motives are mostly related to different indicators of initiative involvement.

Jans et al. (2023) show that the environment and the community are important reasons for participating in local energy initiatives. They also suggest that involvement in local energy initiatives motivates people to act in a sustainable way.

Awareness is an important factor for the choice for more energy efficient appliances (Mills and Schleich, 2010). Mills and Schleich (2010) did research into the awareness of energy labels for household appliances and the choice of class-A energy-efficient appliances. For this they used data from a large survey in Germany, more than 20,000 German households participated in this survey. By framing the energy performance of an appliance in financial savings the adoption of energy efficient appliances increases. They also find that household characteristics have little impact on the purchase of energy efficient appliances. When more households in the neighbourhood are aware of energy labels, the awareness of an individual household rises. Renting a residence instead of owning a residence leads to an increased probability of class-A appliances. They suggest that provision of economic information on the likely economic benefits of energy efficient appliances can further influence purchase decisions. Access to information through personal computers is also likely to influence consumer purchase decisions.

The introduction of mandatory energy efficiency certificates for household appliances has a significant negative impact on residential energy use (Aydin and Brounen, 2019), meaning the residential energy use decreases. They looked into the effect of policies on the residential energy consumption. Other factors that have a significant effect on the residential energy use according to their research are: income and heating

degree days. A higher allowable U-value of the dwelling materials and more heating degree days lead to a higher non-electricity energy consumption.

With the improvement of energy efficiency in dwellings the rebound effect comes into play. This effect happens when energy efficiency improvements lead to a decrease in costs and hence in an adjustment of behavior which then leads to an increase in energy consumption. This effect is important since it undoes the energy savings that are accomplished by certain energy efficiency improvements. Aydin et al. (2017) investigate this rebound effect in residential heating. They find that this rebound effect varies by wealth, income and energy use level of the household. These characteristics are thus important to incorporate in a model to predict the energy savings of certain energy efficiency measures. In their research they have used instrumental variables and fixed-effects approaches.

2.3 Models for heterogeneity in socio-economic groups

Econometric methods as well as machine learning methods are used in recent studies about energy use and energy preservation in relation to heterogeneity of socio-economic groups. In this section we will look at different studies that have used these methods to study heterogeneity. First we will start with the studies that have used econometric models on real data.

2.3.1 Econometric applications

Zou and Mishra (2020) use a multinominal logit model to predict the patterns of electric appliance usage in rural China. They provide evidence for the leading role of income and education in the selection of energy-efficient appliances. Income is the most important factor in the decision to buy energy-efficient appliances. Other factors that play a role in the purchase of energy-efficient appliances are a larger household size and larger dwelling areas.

Souza (2018) finds that compared to renters, homeowners are significantly more likely to have energy-efficient appliances. Linear probability models are used. He investigates heterogeneity in tenancy modes and utility payment regimes, heterogeneity in tenancy duration and heterogeneity in energy prices. Bivariate probit models are estimated to correct for potential biases due to correlations of variables with the error term. The effect of owner-occupation is attenuated when housing attributes are included and when the landlord pays the utility bills. Higher energy prices lead to an increase in the likelihood of efficient appliance adoption and the investments in rented homes are likely to occur at later periods of tenancy.

McCoy and Kotsch (2021) explore the impacts of energy efficiency measures. The baseline of the used model is a fixed effects panel specification controlling for unobserved time-invariant household characteristics which might affect energy consumption. With the baseline specification they perform a series of Event-Study analyses. To explore heterogeneity a series of quantile regression models are estimated. They use statistical matching on the data in order to obtain consistent and unbiased estimators. They find that cavity wall insulation and heating system replacement result in an energy saving of about 10 percent and loft insulation in a 3 percent reduction. Households in more deprived areas observe lower energy savings.

Filippidou et al. (2017) perform a trend analysis to determine the energy renovation pace in the Dutch non-profit housing sector. Although a number of energy improvements have been realized, they only resulted in small changes of the energy efficiency of the dwellings. A linear extrapolation of a used energy performance measure shows that the renovation pace should increase in order to reach the goals.

Brounen et al. (2013) measure the awareness of households toward energy use and whether they are able to make a trade-off between long-term savings and the investment in energy efficiency measures. They do this by a using a survey. They also measure the willingness to conserve energy. A logit model is used. Awareness of energy use is to some extent determined by demographics, the age of the respondent is the most important.

Rational decision-making is determined primarily by education. Energy literacy and awareness are unrelated to conservation behavior and actual energy consumption. Older people with an higher income choose higher comfort levels.

Brounen et al. (2012) document a strong relationship between the age of a dwelling and the energy resource consumption. Older buildings use more gas and newer buildings more electricity. Families with children consume more gas than single households or couples, but not per capita. Elderly households consume more gas. Households with children consume more electricity. The gas and electricity consumption also differs with income. These results are found by regressions on the available data.

2.3.2 Machine learning methods

Different machine learning methods can be used to estimate heterogeneity along a number of covariates. Knaus et al (2019) have looked into this, they find in different research that the following machine learning methods are used: regression trees, random forests, the least absolute shrinkage and selection operator (Lasso), support vector machines, boosting, neural nets and Bayesian machine learning.

O'Neill and Weeks (2018) examine the heterogeneity of demand response following the introduction of time-of-use electricity pricing. They find that households that are younger, more educated and that consume more electricity respond more to a new pricing scheme. Part of their research is the comparison between past consumption information and survey information in producing estimates using a causal forest. They find that a causal forest favours past consumption information to describe heterogeneity.

Murakami et al. (2022) use both a difference-in-difference regression to calculate the average treatment effect and the machine learning technique causal forest to estimate heterogeneous treatment effects at the household level. They do this by using the GRF algorithm from the R package grf. A sample splitting approach is used to examine whether the predicted treatment effects reflect true treatment effect heterogeneity. They want to shed light on the cruciality of heterogeneity in nonmonetary interventions. The results show that heterogeneity has a more crucial impact on policy effectiveness for nonmonetary interventions than for monetary interventions. The results also show that sophisticated targeting will increase the impact of interventions both monetary and nonmonetary.

Knittel and Stolper (2021) and Knittel and Stolper (2019) also start with a difference-in-difference regression to calculate average treatment effects. After this they use a causal forest to predict the heterogeneous treatment effects. The six characteristics that are most used in their forest are: baseline consumption, home value, home square footage, building year, income and respondent's age. From these six, baseline consumption and home value are the most frequently used to grow the forest.

2.4 Conclusions

The adoption of clean energy technologies stands as an important challenge nowadays. As shown in this chapter, various policies have been introduced to stimulate adoption, targeting both individuals and groups. While these policies strive for a sustainable energy transition, their effectiveness varies across socio-economic groups. Understanding the effects of policies on adoption behavior is essential for developing fair and inclusive energy transition strategies. By using insights from econometric and machine learning analyses, policymakers can design targeted interventions that address the diverse needs and preferences of different socio-economic groups.

As shown a number of different econometric and machine learning models can be used to find relations and explain or predict differences between different socio-economic groups. Important factors that follow from the discussed articles are: income, education level, households size, heating degree days, energy consumption,

home value, dwelling area, dwelling age, energy prices, income, age and homeownership. These variables are thus important to incorporate in research for heterogeneity in energy adoption.

3 Research Design

In this chapter the research design from the thesis will be described. In order to answer the main and sub research questions first a conceptual model will be created. Next, the foundation that provided the data that will be used is introduced. Finally a short overview of the next steps in this research will be given.

3.1 Conceptual model

In order to answer the research questions, models will have to be constructed. The models will be based on the conceptual model, this model and the included variables follow from the literature research. The conceptual model can be found in figure 1. The characteristics that are included in the research are categorized in three groups: dwelling characteristics, household characteristics and neighbourhood characteristics. Other variables that also influence the decision are the energy prices and heating degree days. The variables included in the conceptual model are found in previous research to have a relation with energy use and/or the adoption of energy efficient measures. In this research the effect of the neighbourhood approach on the relationship between these variables and the decision to apply energy efficient measures/ adopt clean energy technologies is tested. A neighbourhood approach involves a series of organized activities within a neighbourhood aimed at promoting the adoption of energy efficient measures in dwellings. As can be seen in the model, the neighbourhood treatment is expected to have an effect on the relationship between the dwelling, household and neighbourhood characteristics and the decision to apply energy efficient measures. The variable neighbourhood treatment refers to whether the neighbourhood has received a neighbourhood approach from a company that focuses on implementing energy efficient measures. Success is indicated by more people applying energy efficient measures to their home. A higher conversion rate, where conversion rate is the number of households that apply an energy efficient measure divided by the total number of households, is an indicator for a successful neighbourhood approach. Not all variables from the conceptual model will in the end be included in the models, since not on all variables information is available. Information on energy prices and heating degree days is not available as well as information on some characteristics.

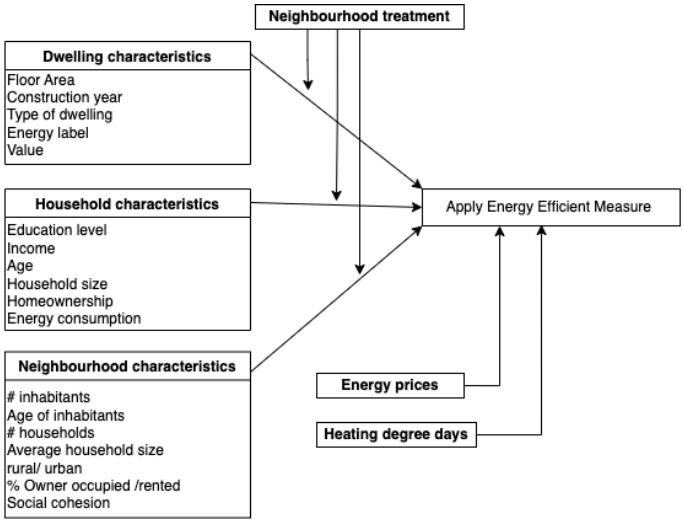


Figure 1: Conceptual model

This conceptual model will be the starting point from which the empirical models used to answer the main and sub research questions will be built. The first sub-question focuses on how econometric and machine learning techniques can identify and explain the relations between the different variables and the decision to apply an energy efficient measure. Hence, the focus lies on the arrows between the characteristics on the left side and the dependent variable: "apply energy efficient measure". By using different models this

relationship will be explored. The second sub-question will be answered in the methodology chapters; in these chapters the strengths and weaknesses of the different techniques will be given. After exploring the techniques, they will be used to predict the decision to apply energy efficient measures based on given variables. So, using the different characteristics it will be predicted, using the different models, whether a household/homeowner will apply an energy efficient measure or not, which allows to compare the models on their predictive accuracy(sub-question 3). All households/homeowners that are included in the data used for this thesis have received a neighbourhood treatment. These neighbourhood treatments include activities like meetings, leaflets, posters and other activities to inform the homeowners about the possibilities. The relation between the characteristics/variables and the decision to apply an energy efficient measure has thus been influenced by the neighbourhood treatment as depicted in the conceptual model. However, as no observations without neighbourhood treatment are available, the influence of the neighbourhood treatment cannot be measured from the data. For this research an alternative treatment will be used. This treatment is defined as an initiator of the neighbourhood approach living within 200 meters of the household. This alternative treatment has been chosen since a big contribution to the neighbourhood approach is delivered by the initiators, being within reach of these initiators can be seen as an indicator of a more intense neighbourhood treatment.

A company that uses the neighbourhood approach/treatment to stimulate the adoption of the use of clean energy technologies in homes is Buurkracht. Buurkracht applies this treatment/approach on communities, which are selected parts of neighbourhoods. Buurkracht will be the case study of this thesis. More information on Buurkracht will be given in section 3.3.

3.2 Econometric and Machine Learning Methods

The methods that will be used in this thesis are logistic regression, Poisson regression, random forest and causal forest. At first, these methods will be used to identify and explain the relationship between different characteristics, neighbourhood, dwelling and household, and the decision to adopt/apply an energy efficient measure. Second, the methods will be used to predict the decision to adopt energy efficient measures based on given characteristics. The logistic regression is chosen because of the binary dependent variable (measure/no measure), in this logistic regression fixed community effects will be included. The Poisson regression will be used at the community level where the total number of households that applies a measure is counted and thus we are dealing with a count variable, for which a Poisson regression is especially suitable. The random forest will be used because it is a classification algorithm that can handle large datasets, is less susceptible to overfitting than other machine learning techniques and it can handle non linear relationships in the data. Finally, the causal forest will be used. This algorithm is a valuable tool for estimating treatment effects in observational studies which makes it well-suited for causal inference.

The results from the different methods will be compared. The effects of the different characteristics, dwelling, household and neighbourhood, will be compared based on their impact and the ranking of their importance according to the models. The predictive power of the models will be compared based on five evaluation metrics: accuracy, precision, recall, specificity and F1 score. The F1 score is the harmonic mean of precision and recall and useful in the case of imbalanced data.

3.3 Buurkracht

For this study data from Buurkracht will be used. Buurkracht is a non-profit making independent foundation that helps people to get a grip on their energy consumption. Their main focus is on helping neighbourhoods with energy-saving measures. They believe in the power of neighbours, by letting neighbours work together they can make the neighbourhood nicer, greener, better, safer, more sustainable and cosier. They provide tips, tricks and tools to support the neighbourhoods in their process to a better neighbourhood.

The basic idea of the Buurkracht approach can be summarised in three steps.

1. *A good idea:* In this first step one of the neighbours comes up with an idea and takes the initiative to do something with the idea.

2. *Mobilize neighbours:* To implement the idea in the neighbourhood, more neighbours have to be found that are interested in implementing the idea, this can be done in a number of ways, one of which is the Buurkracht app.

3. *Execute the idea together:* After the idea is made more concrete the idea can be implemented together.

Based on their website (Buurkracht (nd)) and their data (see figure 3), the process in a community can be presented as in figure 2. The initiators in this process are the driving force behind the neighbourhood approach. They decide which households are included in the project, thus setting the physical boundaries of the communities for their neigbourhood approach. They organise activities and provide information to the neighbours in this community. These initiators do not necessarily have to adopt an energy efficient/saving measure/technology themselves. In figure 3 one can see that after the first measure has been taken, there are still a number of activities, and then more measures follow. The goal of Buurkracht in this community is to get as many households as possible to take a measure in order to make their homes more energy efficient.

Figure 2: Buurkracht in a neighbourhood

In the data from Buurkracht information is included on the households that were part of a community that has received a neighbourhood approach which was led by initiators who were supported by Buurkracht. In the data is included which neighbours were the initiators of the neighbourhood approach, it is also included who have implemented a measure during the campaign, these measures were mostly solar panels and insulation, but also high-efficiency boiler, direct current ventilation, insulating glass, hybrid heat pump, heat pump and solar boiler were included.

3.4 Conclusions

This thesis adopts a multifaceted approach to understanding the factors influencing household decisions to adopt clean technology measures, such as solar panels, insulation or heat pumps. By employing machine learning and econometric techniques, this study aims to find the variables related to dwellings, neighbourhoods and households that contribute to the adoption or non-adoption of these measures. Subsequently, the developed models are utilized for predictive purposes to identify households likely to adopt such measures. Finally, an evaluation of these models is conducted using metrics that assess predictive accuracy and model performance.

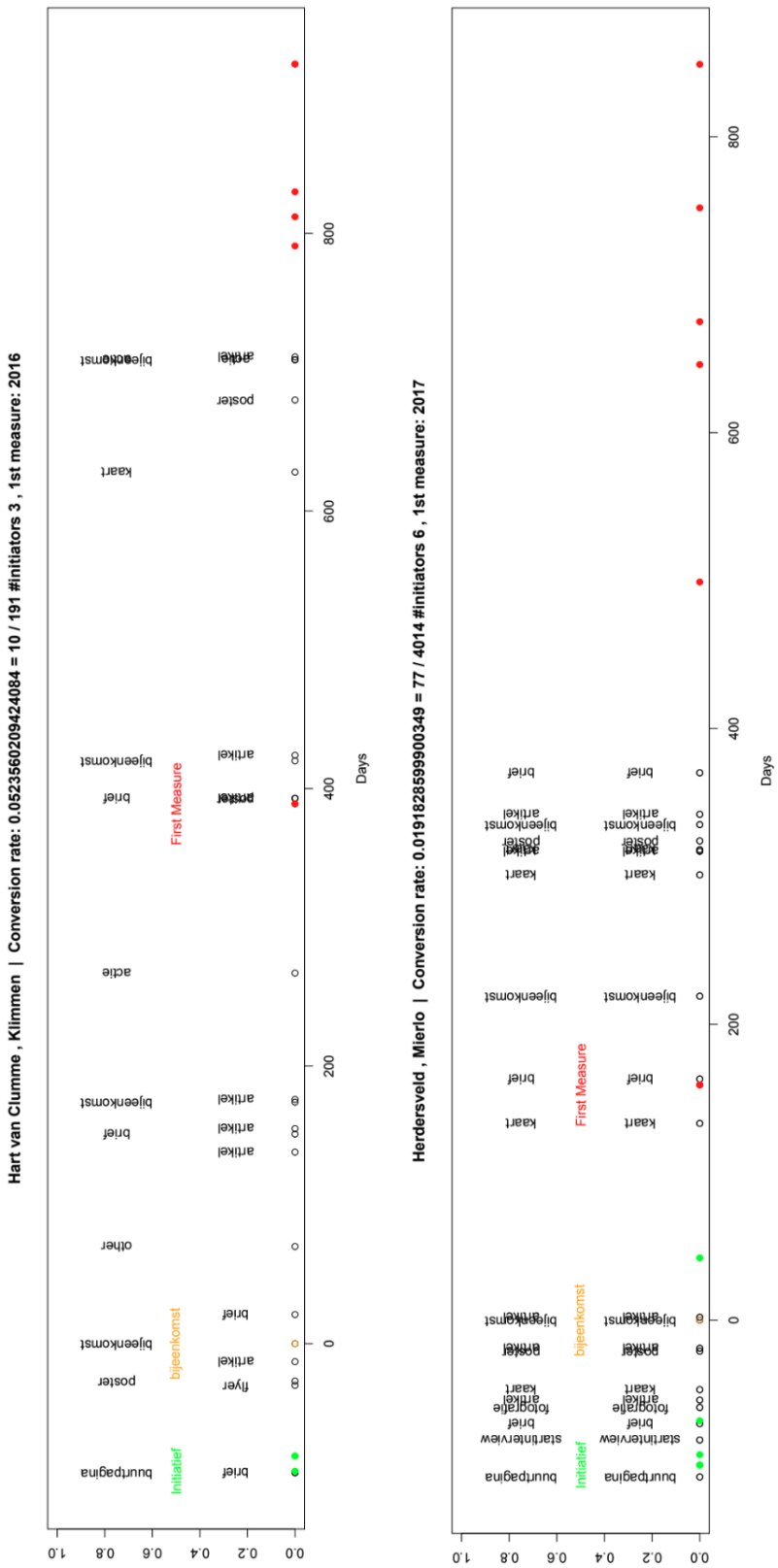


Figure 3: Timeline in neighbourhoods

4 Methodology - Regressions

In this chapter the methodology regarding regressions will be described. First the logistic regression is described. The logistic function is given, methods to assess the logistic model fit are discussed, it is explained how the results can be interpreted, the assumptions that are made for the use of logistic regressions are given and the advantages and disadvantages of the logistic regression method are listed. Next the poisson regression is described. The general form of a poisson regression is given, ways to assess model fit are discussed, it is explained how to interpret the results, the assumption are given and finally advantage en disadvantages are listed.

4.1 Logistic Regression

We are interested in whether a household applies an energy efficient measure in their house or not. That is why we will consider regression models where the dependent variable is binary, measure applied(1) or no measure applied(0). The most well know regression method for a categorical dependent variable is logistic regression and for specifically a binary dependent variable a binary logistic regression can be used.

In a logistic regression a logit transformation is applied on the odds. The odds is the probability of success (p_i) divided by the probability of failure ($1 - p_i$). This logit transformation of the odds, also called log odds will be the dependent variable and the x_i's will be the independent variables. A logistic function is represented by the following formula:

$$ln(\frac{p_i}{(1-p_i)}) = \beta_0 + \beta_1 * x_{i1} + ...\beta_n * x_{in} \tag{1}$$

The β's in a logistic regression model are estimated via maximum likelihood estimation (Penman (2022)).

In the logistic regression fixed effects can be included. These fixed effects are used to control for unobserved heterogeneity across different groups in the data. For our data this can be used to capture the unobserved neighbourhood characteristics that are the same for each individual in a selected community from Buurkracht. By including these fixed neighbourhood effects the estimates of the effects of the explanatory variables are not biased by omitted variables at the community level or unobserved heterogeneity between communities. A logistic function with fixed effects is represented by the following formula:

$$ln(\frac{p_i}{(1-p_i)}) = \beta_0 + \beta_1 * x_{i1} + ...\beta_n * x_{in} + \gamma_1 * D_1 + ...\gamma_m * D_m \tag{2}$$

The dummy variables $D_{1...m}$ are dummy variables for the different communities in which individuals are located. The coefficients $\gamma_{1...m}$ reflect the fixed effect of these communities on the dependent variable. In turn these fixed effects can be regressed against the neighbourhood characteristics of the communities.

4.1.1 Model fit

A test that is often used to assess the model fit is the Hosmer-Lemeshow test. The Hosmer-Lemeshow test is a goodness of fit test. It tests whether the observed number of events matches the expected number of events in population subgroups. In this test the data is first ordered based on the predicted probabilities from the logistic regression model and grouped in a predefined number of groups(g), the number of groups is set larger than the number of covariates and in many cases 10 groups is used, this is also the default in statistical software. The test statistic is then calculated with the following formula:

$$G^2_{HL} = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{(E_j(1 - \frac{E_j}{n_j})} \sim \chi^2_{g-2} \tag{3}$$

where n_j is the number of observations in the j^{th} group, O_j is the number of observed events in the j^{th} group, E_j the number of expected events in the j^{th} group. If this test is performed in statistical software

the values that will be returned are the Hosmer-Lemeshow chi-squared and a p-value. Small p-values(0.05) implicate a poor model fit (Hosmer et al. (2013)).

McFadden's pseudo R-squared (Bartlett (2014)) is also a measure for goodness of fit. It is used to assess how well the model fits the observed data compared to a model with no predictors (only an intercept). McFadden's R-squared is defined as:

$$R^2_{McFadden} = 1 - \frac{Log(l_c)}{Log(l_{null})} \tag{4}$$

L_c is the log-likelihood value for the fitted model and L_{null} is the log-likelihood value for the null model, which is a model with only an intercept and no covariates. The statistic represents the proportion of variance in the dependent variable that is explained by the independent variables. A higher value indicates a better fit of the model to the data.

The Hosmer-Lemeshow test and McFadden's pseudo R-squared are not the only goodness of fit tests for a binary logistic regression, other ways to check for the goodness of fit include: Chi-square goodness of fit tests and deviance, classification tables, ROC curves, logistic regression R^2 or model validation via an outside data set or by splitting the data set (Joseph (nd)).

The outcome of a logistic regression is a probability between 0 and 1. The β's from a logistic regression are often interpreted by using the odds ratio. This ratio is found by applying exp() on the β's. A change in x_i by one unit changes the odds by a factor $\exp(\beta_i)$.

4.1.2 Assumptions

Linear regressions make a lot of assumptions regarding the data, logistics regressions do not make the same assumptions. For a logistic regression a linear relationship between dependent and independent variables is not necessary. The error terms do not need to be normally distributed and homoscedasticity is not required. For binary logistic regressions other assumptions arise (Zach (2020)). In order to use a binary logistic regression the following assumptions are made:

1. The dependent variable is dichotomous (binary)
2. The observations should be independent of each other. Multiple repeated observations from one individual cannot be used and the observations cannot be related to each other in any way.
3. Linearity between the independent variables and the log odds is required. This can be checked by inspecting scatterplots of the variables against the log odds.
4. A large sample size is needed. A guideline that can be used is that at least 10 cases with the least frequent outcome are needed for each independent variable in the model. For example is you have 4 explanatory variables and the expected probability of the least frequent outcome is 0.20, then the sample size should be at least (10*4)/0.20 = 200.
5. There can be no extreme outliers in the data. Cook's distance can be used to detect the extreme outliers, these outliers could be removed, replaced with the mean or median or kept and reported on. Other ways to remove outliers are: 1. Using interquartile ranges or 2. Using standard deviations. Standard deviations are best used in removing outliers from the data in the case of normally distributed data. Interquartile ranges are better to use in the case of skewed data.
6. No high correlations between the explanatory variables are allowed (multicollinearity). If two or more explanatory variables are highly correlated they do not provide unique information to the regression model. This can lead to problems with fitting and interpreting the model. A correlation matrix among the predictors can be used to assess the multicollinearity. Another way to check for multicollinearity

is by using the variance inflation factor (VIF). A VIF greater than 5 suggest multicollinearity with poorly estimated coefficients and questionable p-values.

4.1.3 Advantages and disadvantages

For every method there are advantages and disadvantages (AIML.com (2023), Jain (2018)). Below some of the advantages and disadvantages related to a logistic regression are named. These pros and cons are not relative to linear regression but instead are in the context of a variety of potential methods.

Advantages:

- Quick and easy to implement, no high computation power required.
- The β's give information about the importance and the direction of the variables.
- Especially efficient when the data set has features that are linearly separable, meaning positive and negative outcomes can be effectively separated by a linear boundary.
- Performs well on a low dimensional data set, the regression will be less sensitive to overfitting on a low dimensional data set.

Disadvantages:

- The interpretation is more difficult than for a linear regression. The interpretation of the weights is multiplicative and not additive.
- Logistic regression can be prone to overfitting, especially when a large number of predictor variables is used in the model. Overfitting means that the generalizability of the model beyond the data that is used to fit the model reduces. By adding more independent variables the amount of variance explained in the log odds increases. However, adding more and more variables to the model will lead to overfitting.
- The previously mentioned six assumptions should be met.
- It is difficult to capture complex relationships, because of the linear nature of the method, methods like decision trees, random forests or neural networks are better at capturing complex relationships.

4.2 Poisson regression

Another variable that we are interested in is the number of measures taken per neighbourhood. This variable is a count variable. A regression that is specifically useful for count variables is the Poisson regression. In this section the Poisson regression will be explained.

4.2.1 Model

For large means the normal distribution is a good approximation for the poisson distribution. Poisson regression is especially useful in the case of a small response variable, which is the case in the Buurkracht data. The number of measures taken per community is small. Sometimes it is suggested that the mean count should not be larger than 10 (Lund and Lund (2018)).

The general form of a poission regression model is:

$$log(\lambda_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_n x_{in} \tag{5}$$

Here, λ_i is the response variable, in our case the number of measures taken in a community, α is the intercept, β_n's are numeric coefficients and the x_{in}'s are the values of the n-th explanatory variables for the i-th observation. The explanatory variables can be a mixture of numeric or categorical variables. The coefficients are estimated using maximum likelihood estimation (Jabeen (2019)).

4.2.2 Model fit

In order to assess the goodness of fit of the model, pearson chi-squared and deviance test statistics can be used. The pearson chi-squared test statistic is:

$$\chi^2 = \sum_{j=1}^{c} \frac{(o_j - e_j)^2}{e_j} \tag{6}$$

Where o are the observed values, e are the expected values and c is the number of classes. The χ^2 follows a chi-squared distribution with degrees of freedom equal to the number of classes minus the number of parameters minus 1. A large statistic and hence a small p-value suggests that the model does not fit the observed data (Vul and Wenhao (nd)).

Residual deviance values can also be used to assess model fit. If the residual deviance is lower than the expected chi-square value for the corresponding degrees of freedom, than this indicates a good model fit. If the value is higher than the expected chi-square value, than this indicates a lack of fit.

All parameter estimates are on the log scale and need to be transformed for interpretation. The coefficients can be interpreted as follows: for a one unit change in the explanatory variable, the log of the count is expected to change by the respective regression coefficient, given the other explanatory variables stay the same. So the count is changed by a factor $e^{regressioncoefficient}$. Statistical significance is indicated by small p-values.

4.2.3 Assumptions

Poisson regressions can be used if the data meets 5 assumptions:

1. The response variable is a count variable, described by a poisson distribution.
2. There should be one or more independent variables and these variables could be measured on a continuous, ordinal or nominal scale.
3. The observations must be independent of one another
4. The mean of the poisson random variable must be equal to its variance. This is one of the properties from a poisson distribution. If this assumption is satisfied this is called equidispersion. Often this is however not the case in the data, in that case overdispersion can be assessed using the pearson dispersion statistic. It can be dealt with by using a dispersion parameter for small differences or a negative binomial regression model for large differences.
5. The relation between the log of the mean rate, $\log(\lambda)$, and the explanatory variable is linear.

4.2.4 Advantages and disadvantages

As for all methods there are advantages and disadvantages of using a Poisson regression. Some of these advantages and disadvantages are listed below.
Advantages:

- It is a simple model compared to other models that can be used for count data.

Disadvantages:

- The assumption that the mean is equal to the variance may not always be true, this will lead to incorrect standard errors
- A poisson regression does not perform well in the case of many zero counts.

4.3 Conclusions

In this section both the logistic regression method and the Poisson regression methods were discussed. Pros and cons were presented for each method, and methods for assessing the model fit were explored. One of the primary advantages of both methods is their relative simplicity compared to other alternatives. However, this simplicity also presents a drawback, as these models may not be complex enough to capture all potential relationships.

5 Methodology - Machine learning

This chapter provides an overview of the machine learning methodology that will be employed in this research. We begin with an introduction into machine learning, followed by the discussion of multiple methods. The initial method explored is decision trees, next the theory of random forest is given and finally the causal forest is discussed.

5.1 Machine learning

In machine learning two types of techniques are distinguished: supervised learning and unsupervised learning. In supervised learning a model is trained on known input and output data in order to predict future outputs. In unsupervised learning hidden patterns and structures are found in the input data. In this research the focus will be on supervised machine learning techniques since known data on the output that we are trying to predict is available. In supervised learning two techniques are used to develop machine learning models, classification and regression techniques. Classification is about predicting discrete outcomes and with regression techniques a continuous quantity can be predicted. A classification problem where the outcome has two options like "yes" or "no: is called a binary classification problem. The accuracy of a classification problem is given by the number of correct predictions divided by the total number of predictions. In some cases classification techniques can be used for a regression problem and the other way around. This can for example be done by converting a ratio variable to an ordinal variable thus constructing classes. Examples of machine learning classification algorithms are: k-nearest neighbours, support vector machines, naive bayes, decision tree and random forest (Sarker (2021)).

5.2 Decision tree

Decision trees can be produced using different algorithms. One algorithm that is often used is the Classification and Regression Trees (CART) algorithm by Breiman et al. (1984). It is based on binary splitting of the attributes. CART consists of three steps. 1. growing a large initial tree. 2. using a pruning algorithm to prune the tree and 3. using a validation method for determining the best tree size. The growing of the tree is done by a greedy search. This means that the best possible solution for each split is chosen, this does not necessarily lead to the best overall result. In CART the best possible solution at a split is based on the Gini Index, this index is comparable with the variance. The pruning algorithm is based on weakest link cutting. A node is pruned away if the resulting change in the cost-complexity criterion will be less than α times the change in tree complexity. α is the tuning parameter that indicates the allowed trade-off between the complexity and the accuracy. The best tree size is found using cross validation. The α that minimizes the cross-validated sum of squares is chosen. Other well known algorithms used to build decision trees are: C4.5, CHAID and QUEST (Song and Lu (2015)).

Overfitting is a well known limitation of decision trees. The problem of overfitting is making the tree so excessive that the tree perfectly predicts in the case of the training data, but in the case of the test data the results are way of. This happens because the leaves are made too small/specific. The tree is not made for general input, but made specific for the training data. Overfitting should be prevented by pre-pruning or post-pruning. Pre-pruning is done for example by setting a maximum depth or a minimum leaf size. For post-pruning, algorithms are used that remove non-significant branches, some algorithm that can be used are: reduced error pruning, pessimistic error pruning, minimum error pruning and error based pruning. These methods work from the bottom of the tree upwards and consider for each node, that one level down only has leaf nodes, whether it can be pruned (Song and Lu (2015)).

High variance is a problem of trees. A small change in the data will lead to a very different tree. The hierarchical nature of the tree makes that an error at one of the top splits is propagated to all the splits below. One way to reduce the variance is by bagging. Bagging is also called bootstrap aggregation. Several subsets of the data are created with replacement and with each subset a decision tree is created, at the end

the average of these decision trees is taken to create the final decision tree (Breiman (1996)).

5.3 Random forest

A random forest is a machine learning technique that combines bagging and decision trees. Random forests were first proposed by Breiman (2001). The technique adds an extra layer of randomness to bagging. In a standard decision tree, each node is split by considering all variables and choosing the variable that gives the best split. In a random forest not all features are considered for each split, only a random subset of all features is considered for each split. For each node a new random subset of features is chosen. This creates more variation and hence leads to a lower correlation between trees. In a random forest the majority vote or average is taken from different trees. This process is depicted in figure 4. The predictions from the different trees should have a low correlation with each other. This means that the individual decisions from each tree in the forest should be as independent as possible from one another. Each tree learns a unique pattern of the data and combining them enables the model to better generalize the overall patterns in the data. In our case majority vote will be used, since we want to predict a class variable, "measure taken" or "measure not taken".

A random forest prevents overfitting in two ways, by selecting random subsets of the entire data set for training the different trees and by randomly selecting a number of features to choose from at each split.

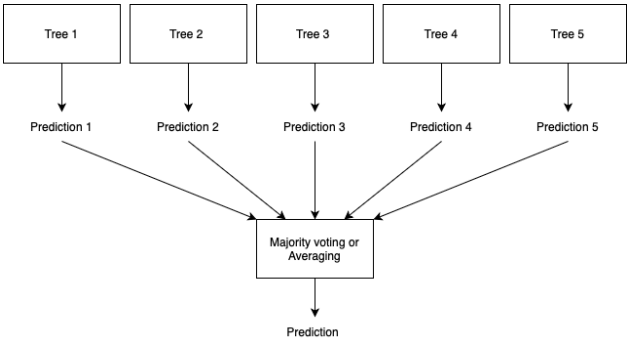


Figure 4: Forest

The random forest algorithm consists of 3 steps:

1. For the number of trees that will be used bootstrap samples have to be drawn from the original data. For each set around one-third of the instances is left out.
2. From each of the bootstrap samples a tree has to be grown, this tree is unpruned and at each node the best split among a random sample of variables is made.
3. Aggregating the different trees that were constructed at step 2 leads to the required predictions.

Breiman uses out-of-bag data to estimate the error rate of a random forest. This out-of-bag data is the data that is not in the bootstrap sample. The out-of-bag data is used to predict the outcome by using the tree that is grown with the bootstrap sample. These out-of-bag predictions are compared to their real outcome and the differences are aggregated for the different trees and give the error rate. This removes the need for a test set.

5.3.1 Interpretation

A random forest cannot be interpreted based on parameters such as regression models can. The random forest can be seen as black box model. It can make predictions but the inner workings are not easily interpretable or explainable. The randomForest package in R (Liaw and Wiener, 2022) produces a measure of

variable importance. This measure is calculated by looking at how much the prediction error increases when out-of-bag data for the specific variable is permuted while the data of the other variables is not changed. If permuting the out-of-bag data for a specific variable leads to a substantial increase in prediction error, it indicated that the variable is important for the model's accuracy (Breiman (2001)). The mean decrease accuracy value given by the package represents how much removing the variable reduces the accuracy of the model. A higher value thus means that the variable is more important for making an accurate prediction. Other measures about the accuracy of the model that the package produces are the out-of-bag estimate rate and a confusion matrix. The out-of-bag estimate rate gives an estimation of the error rate of the model. It indicates how well the model estimates on the data that is not used in the model. The confusion matrix indicates the number of correct predictions per class and number of wrong predictions. It also gives a percentages of correct predicted and wrong predicted outcomes per class.

A graphical tool to visualize the relationship between the features of a random forest model and the predicted classes is a partial dependence plot (PDP). This partial dependence plot can be created in R using the package "pdp" (Greenwell (2017)). In a PDP, the effect of a single feature on the model's predictions is visualized. It demonstrates how variations in a single feature influence the model's predictions while keeping all other features fixed at their average or representative value.

5.3.2 Advantages and disadvantages

Below the advantages and disadvantages of a random forest are given in relation to the use of single decision trees.
Advantages:

- A random forest is more robust and less prone to overfitting. It prevents overfitting in two ways. First, it selects random subsets for training the different trees and at the end the majority vote is taken, this lowers prediction error and overall variance. Second, at each split of the different trees not all features are considered, this creates more variation between the trees and thus lowers the correlation between the trees.
- More robust to outliers, because it aggregates multiple decision trees.
- Random forest provides an indication of variable importance. This is valuable for understanding which features have the biggest impact on predictions.

Disadvantages:

- The results from a random forest are more difficult to interpret than the results from one single decision tree.
- Constructing a large random forest requires more time and computational power.
- The way a random forest is constructed makes it impossible to extract detailed decision rules or logic that is used for the predictions.

5.4 Causal forest

A causal forest is a nonparametric method for heterogeneous treatment effect estimation. It is especially useful when the number of covariates increases and outperforms other nonparametric methods like nearest-neighbour matching and kernel estimations in that case. The method is developed and described by Stefan Wager and Susan Athey in their 2015 paper (Wager and Athey (2015)).

The causal forest method is an extension of the random forest algorithm, adapted to address questions related to causal relationships between variables, especially in the context of observational data and experiments. With a causal forest, questions such as: "What is the causal effect of a certain treatment or intervention on

the outcome of interest?" can be answered. Causal forests capture heterogeneity in treatment effect. They can estimate how the effect of a treatment varies across different subpopulations or units.

5.4.1 Honest tree

An important part of constructing a causal forest is the use of "honest trees". A honest tree is a tree where, for each training sample only the response is used to estimate the within-leaf treatment effects or to decide where to place the splits, but not both. Information is used for either selecting the model structure or for estimation given a model structure. There are two ways to construct honest trees. The first method to construct a honest tree is by splitting the training sample in two parts such that one part can be used for selecting the model structure and the other part can be used to estimate the leaf-wise responses. By separating these tasks, the model ensures that the splits are chosen without any knowledge of the treatment effects, preventing bias in the estimation process. The second procedure is by ignoring the outcome data when placing splits and train a classification tree for the treatment assignments. By excluding the outcome data from the split placement process, the model ensures that the tree structure is determined independently of any potential biases introduced by the outcome variable. Both methods ensure that the tree structure is determined without any influence from the outcome variable.

5.4.2 Causal forest construction

A causal forest can be constructed using the generalized random forest (grf) R package. At each split the heterogeneity in average treatment effects is maximized. This treatment effect is calculated by taking the difference in outcome between the treatment and control conditions within one leaf. The causality arises by making the split where it will produce the biggest difference in treatment effect across leaves, but still give an accurate estimate of the treatment effect (Green and White (2023)). By using honest trees asymptotic normality is present and confidence intervals can be constructed (Wager and Athey (2015)).

5.4.3 Interpretation

Like in the case of a random forest, we are dealing with a black box model. A black box model makes it difficult to make interpretations about the influence of variables and the complete function of the model. Like in a random forest we can retrieve variable importance from the R package. This gives an indication of the importance of each feature/variable in the model. Features with a higher importance are more influential in determining the heterogeneous treatment effects.

Statistical significance of the estimated treatment effects has to be assessed. It can be tested by using a t-test, where the null hypothesis is that the treatment effect is zero and the alternative hypothesis is that the treatment effect is not zero.

A subgroup analysis can be performed using the causal forest. These subgroups have to be chosen, they can be based on expectations and hypotheses that were formed by the literature. In this thesis they will be chosen by regression analysis. By looking at the treatment effects and their statistical significance for the different subgroups it can be determined whether there is a relevant difference in the impact of the treatment on the different subgroups.

The causal forest estimates a treatment effect, which distinguishes its interpretation from that of logistic regression and random forest models. However, the causal forest is still capable of predicting the adoption of an energy-efficient measure based on this treatment effect. By utilizing data from a training set, the causal forest identifies treatment effects. Subsequently, implementing a test set which includes treatment assignments and other variables allows for predictions regarding the adoption of an energy-efficient measure. In essence, the causal forest model captures the causal relationship between the treatment and the outcome,

providing valuable insights into the effectiveness of the treatment.

5.4.4 Advantages and disadvantages

Below the advantages and disadvantages of using a causal forest instead of a random forest are given.

Advantages:

- A causal forest is well-suited for identifying and quantifying the impact of specific variables or treatments on the outcome.
- A causal forest enables inference for identifying heterogeneous treatment effects.

Disadvantages:

- Causal forests are more complex to implement and understand, because of additional steps like estimating treatment effects and propensity scores.
- For validity causal forest may require more extensive data collection and cleaning.
- More time and computational power is required for large datasets.

5.5 Conclusions

We have discussed decision trees, random forests and causal forests. The random forest consists of multiple decision trees and other than that it takes more computational power and interpreting the results is more difficult it is a big improvement compared to a single decision tree. The causal forest is an extension on the random forest, making implementation and interpretation more difficult, but the main advantage for our research is that it can be used to find heterogeneous treatment effects.

6 Data and Descriptive Statistics

In this chapter the data that will be used for this research will be discussed. First the data will be described, next the data will be cleaned and finally descriptive statistics will be given.

6.1 Data description

For this research the research area will be the Netherlands. The data that will be used is the data that Buurkracht has collected during their previous projects in communities, this data will be combined with openly available data from Statistics Netherlands that gives information on people/households by postal code. For privacy reasons not all information is available in this data set, postal codes with 0 till 4 observations or postal codes that have classified information have missing values.

Before cleaning, the data retrieved from Buurkracht contains 100 communities and 112365 observations. These observations are not all unique households, some household appear as multiple data entries because they have purchased multiple clean energy technologies/measures. Only one data entry from this households will be kept for our analysis. In addition, the data also includes information on buildings that do not have the function "dwelling". Since we are only interested in dwellings, all other data will be removed. We are left with 100 communities and 108817 observations.

So far the data contains 44 variables, not all variables can be used for our models. Some variables are just id's other variables are only applicable for a very limited number of households/dwellings. The variables that are of interest to us can be found in table 1. In the end not all variables listed in this table and in table 2 are included in the models, but all variables were considered during the model-building process.

Table 1: Variables in Buurkracht data

Code	**Description**	**Measurement level**
pc	Postal code	
name_buurt	Community name	Nominal
vbo_opp	Net floor area dwelling	Ratio
pand_bouwjaar	Construction year	Ratio
mtr_type	Type of measure taken	Nominal
maatregel	Took a measure(1) or not(0)	Dichotomous
num_xcoord	x-coordinates of the dwelling	
num_ycoord	y-coordinates of the dwelling	
mtr	Measure within 2 years of the campaign start in the community	Dichotomous
pltf	Platform within 2 year of the campaign start in the community	Dichotomous
init	Initiator of the campaign in the community	Dichotomous
init_mtr	Initiator that takes a measure	Dichotomous
mtr_iso	Insulation measure taken	Dichotomous
mtr_zonne	Solar panel measure taken	Dichotomous

As one can see, number of households per community, number of measures per community, number of initiators per community are not included in the original dataset. These numbers are calculated and added to the data. Another variable that is added to the dataset is the distance to an initiator. The location from each household is known and the location from the initiators is known, hence the distance from each household to the nearest initiator can be determined. This variable will be added as a ratio variable and as a categorical variable (0-50 meters, 50-100 meters, 100-200 meters, 200-300 meters, 300-400 meters and 400+ meters). Based on this variable also a dummy is made for being within 200 meters of an initiator, this variable will be used as the treatment in the causal forest model.

Also missing from the Buurkracht data is information on socio-demographics from the households that have participated in the Buurkracht approach. Since this information is also not freely available due to privacy issues, another way to find information on socio-demographics had to be found. This is why data from Statistics Netherlands will be used.

Statistics Netherlands has information available on socio-demographics in the form of statistics per postal code. In these statistics information is available from the people/households living in a postal code. The relevant variables from this data set are given in table 2.

Table 2: Variables in CBS postal code statistics

Code	**Description**	**Measurement level**
PC6	Postal code of 4 digits and 2 letters	
INWONER	Number of inhabitants	Ratio
INW_014	Number of inhabitants from 0 to 14 years	Ratio
INW_1524	Number of inhabitants from 15 to 24 years	Ratio
INW_2544	Number of inhabitants from 25 to 44 years	Ratio
INW_4564	Number of inhabitants from 45 to 64 years	Ratio
INW_65PL	Number of inhabitants older than 65 years	Ratio
AANTAL_HH	Number of households	Ratio
TOTHH_EENP	Number of households of one person	Ratio
TOTHH_MPZK	Number of households of more persons without children	Ratio
HH_EENOUD	Number of single parent households	Ratio
HH_TWEEOUD	Number of two parent households	Ratio
GEM_HH_GR	Average households size	Ratio
WONVOOR45	Number of dwellings constructed before 1945	Ratio
WON_4564	Number of dwellings constructed between 1945 and 1965	Ratio
WON_6574	Number of dwellings constructed between 1965 and 1974	Ratio
WON_7584	Number of dwellings constructed between 1975 and 1984	Ratio
WON_8594	Number of dwellings constructed between 1985 and 1994	Ratio
WON_9504	Number of dwellings constructed between 1995 and 2004	Ratio
WON_0514	Number of dwellings constructed between 2005 and 2014	Ratio
WON_1524	Number of dwellings constructed after 2014	Ratio
P_KOOPWON	Percentage of owner occupied dwellings	Ratio
G_GAS_WON	Average gas consumption	Ratio
G_ELEK_WON	Average electricity use	Ratio
P_LINK_HH	Percentage of households belonging to 40% lowest incomes in NL	Ratio
P_HINK_HH	Percentage of households belonging to 20% highest incomes in NL	Ratio
WOZWONING	Average WOZ woning in PC6	Ratio
P_NL_ACHTIG	Percentage of people with both parents born in the Netherlands	Ratio
M_INKHH	Median income of the households	Ratio

These statistics can be merged with the data from Buurkracht because in both data sets the postal code is known. This leaves us with 108755 observations. Next it is time to detect possible problems like outliers in the data and remove or restore them. This will be done in the next section.

6.2 Data cleaning

In this section, we will detect and potentially remove any outliers present in the data. We will describe each step taken in detail. At the conclusion of the section, we will summarize all the steps in table 5, along with the remaining number of observations and measures in the dataset.

The variables vbo_opp, pand_bouwjaar, WOZWONING, G_ELEK_WON, G_GAS_WON, P_NL_ACHTIG, GEM_HH_GR and M_INKHH are all variables that will be included in the models. The models cannot be fitted if one of these variables contains a negative number, because these numbers should all be positive. This leaves us with 94218 observations in 99 communities. All communities that are left are plotted on a map from the Netherlands, this leads to figure 5.

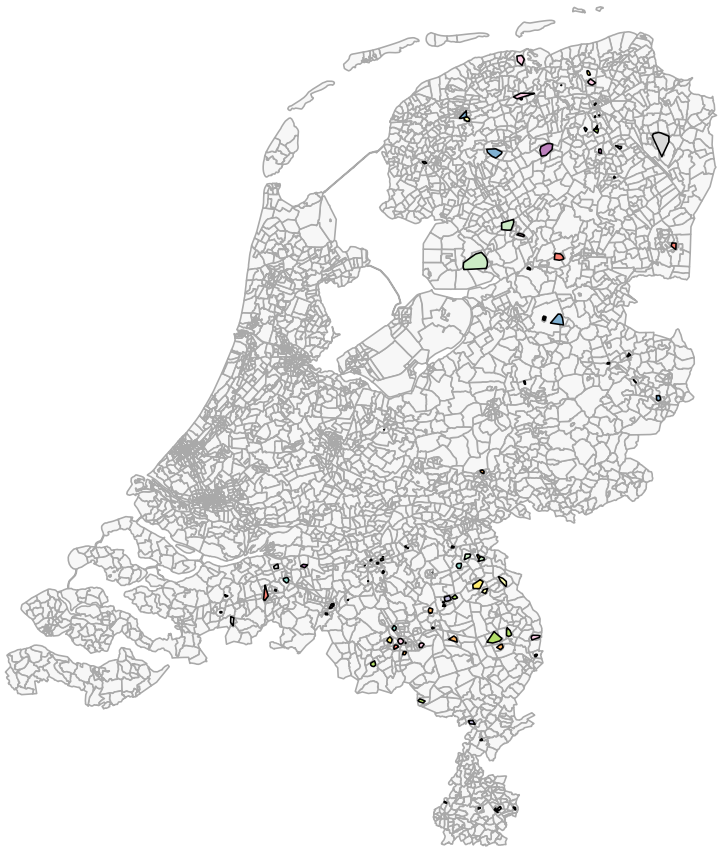


Figure 5: Communities

In all of these communities a number of initiators is present. In table 3 it is depicted in how many communities 0 till 7 initiators are present. 7 is the most initiators present in a community in the data. As one can see there are 14 communities without initiators, this is either a fault in the data collection, or the initiators

have been removed in a previous step. Since a community without initiators does not give reliable results for our analysis, these communities will be removed, leaving us with 84 communities and 84661 observations.

Table 3: Number of initiators

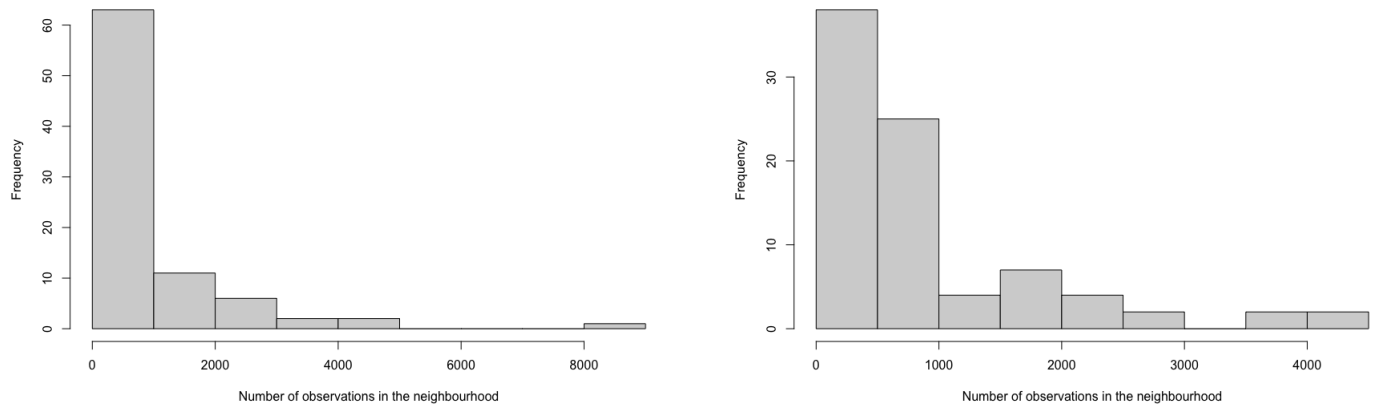
Number of initiators	0	1	2	3	4	5	6	7
Number of communities	14	19	28	18	13	3	3	1

Table 4 tells us that there are some outliers in the data from the living area of the dwellings. A dwelling from 10 m^2 is not realistic and the same goes for a dwelling with a living area of 7877 m^2. It is decided to remove all observations with a living area bigger than 400 m^2 and smaller than 30 m^2.

Table 4: Living area dwellings in m^2

	Mean	St. dev	Min	Max
Before removal	131.3	71.1	10.0	7877.0
After removal	129.7	50.5	30.0	400.0

Figure 6 shows two histograms. The before removal histogram shows that the community with more than 8000 observations is an extreme outlier in our data. This community will hence be removed, resulting in the histogram at the right side. We are left with 75217 observations and 2119 measures taken.

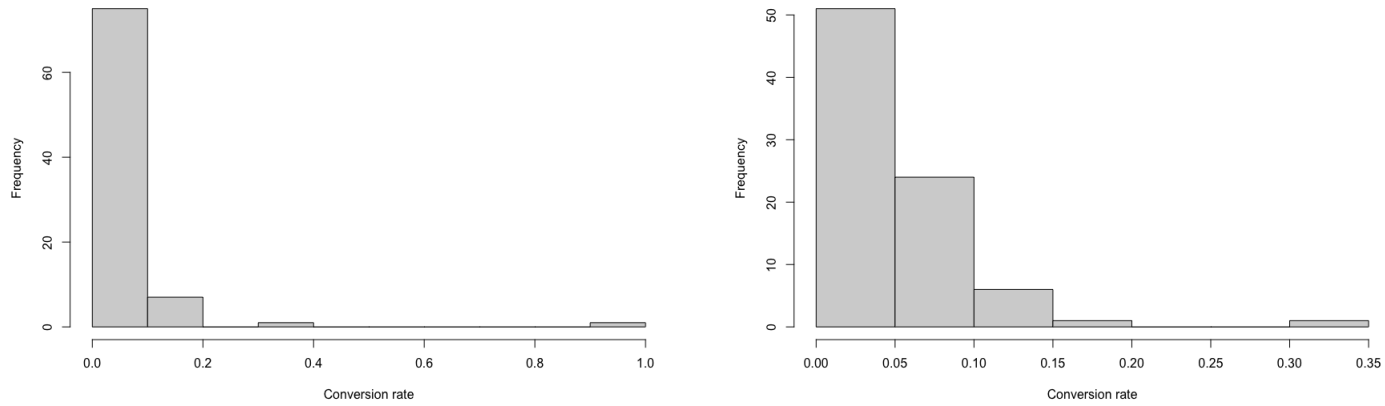


Before removal

After removal

Figure 6: Number of observations per community

Next, let us look at the conversion rate from the communities. The conversion rate is defined as ratio of the number of households that have implemented an energy-efficient measure to the total number of households in the community. In figure 7 one can see that there is one extreme outlier. This outlier will be removed, which results in the histogram on the right and a final number of observations of 75134 in which 2039 measures are taken.



Before removal

After removal

Figure 7: Conversion rate per community

Table 5: Data cleaning steps

Step	**Description**	**Number of observations**	**Number of measures**
1	Buurkracht data	112365	3459
2	Function = dwelling	109341	3393
3	Remove duplicates	108817	2869
4	Merge CBS data	108755	2869
5	Remove cases with negative numbers	94218	2526
6	Remove initiators = 0	84661	2195
7	Remove vbo_opp <30 and >400	87392	2186
8	Remove n_buurt >5000	75217	2119
9	Remove conversie >0.5	75134	2039
10	Merge with event data	74750	2030
11	Measure within 3 years of initiative	74750	1851

The final two steps to complete the dataset are as follows:

- The data is merged with information from Buurkracht regarding the activities organized in the communities during the campaign/neighbourhood approach.
- Only measures that were implemented within three years of the first initiative in the neighborhood are retained as observations of a measure. This time frame ensures that the adoption of the measure can be reasonably attributed to the neighborhood approach or campaign.

6.3 Descriptive Statistics

In this section descriptive statistics will be given for variables of interest. These descriptives can be given on two levels: at the households level, so based on the 74750 observations, or at the community level, in that case the 74750 observations will be grouped based on the community name and the data will be based on the 84 communities that are left after cleaning the data.

6.3.1 Household level

In table 6 statistics are given for some relevant variables. As one can see some new variables are mentioned that were not explained before.

- ln_g_elek_won is the natural log transformation of G_ELEK_WON, the average electricity use in the pc6.
- ln_g_gas_won is the natural log transformation of G_GAS_WON, the average gas consumption in the pc6.
- ln_vbo_opp is the natural log transformation of vbo_opp, the net floor area of the dwelling.
- ln_wozwoning is the natural log transformation of WOZWONING, the average WOZ woning in the pc6.
- m_ink_high has value 1 if the average income in the pc6 is above 60000 euro and 0 otherwise.
- bouwjaar_n1992 has value 1 if the construction year is after 1992 and 0 otherwise. The descriptives show that 24.8% from the dwellings in the data set is constructed after 1992.
- bouwjaar_v1975 has value 1 if the construction year is before 1975 and 0 otherwise. The descriptives show that 45.3% from the dwellings in the data set is constructed before 1975.
- bouwjaar_n1975v1992 has value 1 if the construction year is between 1975 and 1992 and 0 otherwiste. The descriptives show that 29.9% from the dwellings in the data set is constructed in this period.

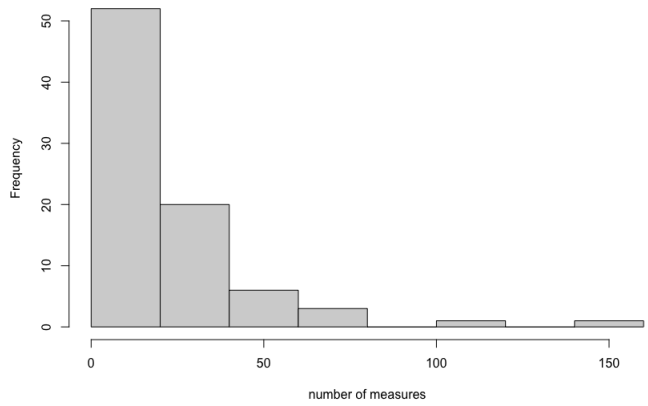
Table 6: Descriptives - Household level

Statistic	N	Mean	St. Dev.	Min	Max
vbo_opp	74.750	132.069	50.148	30.000	400.000
pand_bouwjaar	74.750	1973.755	28.607	1.500	2.020
ln_g_elek_won	74.750	7.975	0.276	6.856	8.963
ln_g_gas_won	74.750	6.925	1.559	0.000	8.331
ln_vbo_opp	74.750	4.817	0.365	3.401	5.991
ln_wozwoning	74.750	5.362	0.385	4.078	6.918
m_ink_high	74.750	0.266	0.442	0.000	1.000
bouwjaar_n1992	74.750	0.248	0.432	0.000	1.000
bouwjaar_v1975	74.750	0.453	0.498	0.000	1.000
bouwjaar_n1975v1992	74.750	0.299	0.458	0.000	1.000

6.3.2 Community level

The variables from the combined data sets can be used to find statistics on the communities. This is done by grouping the different households together based on the name from the community. This results in numbers/statistics per community, the most important variables per community are mentioned below:

- **Number of measures taken per community**
  The number of measures taken variable will be used as a dependent variable, to see if and how this number can be predicted based on other neighbourhood characteristics. Measures that are included in the data are: High-Efficiency Boiler, Roof Insulation, Direct Current Ventilation, Insulating Glass, Hybrid Heat Pump, Cavity Wall Insulation, Floor Insulation, Heat Pump, Solar Boiler and Solar Panels. The number of measures taken per community is actually the number of households per community that take a measure. Households could take multiple measures, but this is not included.



Mean	Min	Max
24.57	1.00	149.00

Figure 8: Number of measures taken per community

- **Number of activities per community**
  During the neighbourhood approach/treatment Buurkracht organises several activities. These activities include organized events, written materials such as articles and letters, dwelling scans as well as the dissemination of information through visual aids like posters. In the appendix in table 25 a complete list of all activities included in the data can be found. The number of activities is a variable that will be used for the analysis. This variable will be used as an independent variable in the poisson regression where the dependent variable will be the number of measures taken per community. In order to find these numbers another data set from Buurkracht is used, in this data set all activities are listed per community. The number of activities is counted per community and merged with the already cleaned data set. Activities are only included if they take place within 3 years after the start date of the Buurkracht campaign in that community. The frequencies from the number of activities are depicted in figure 9.

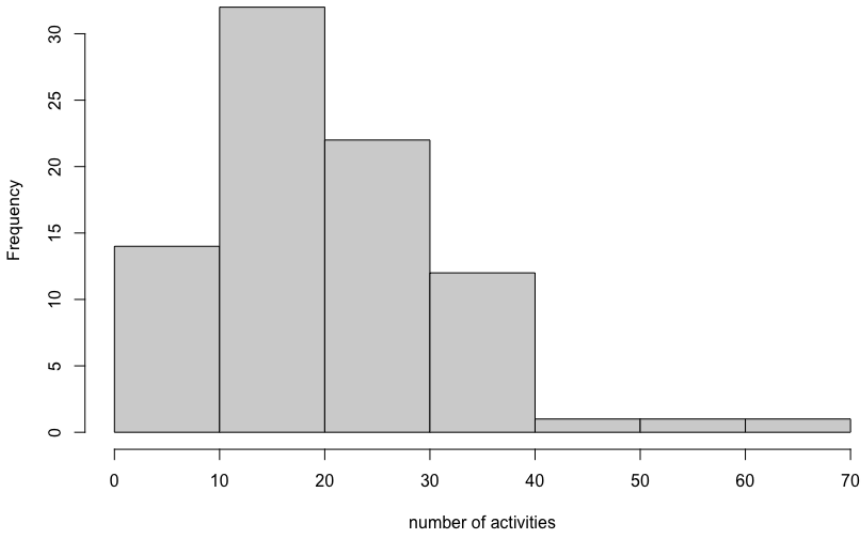
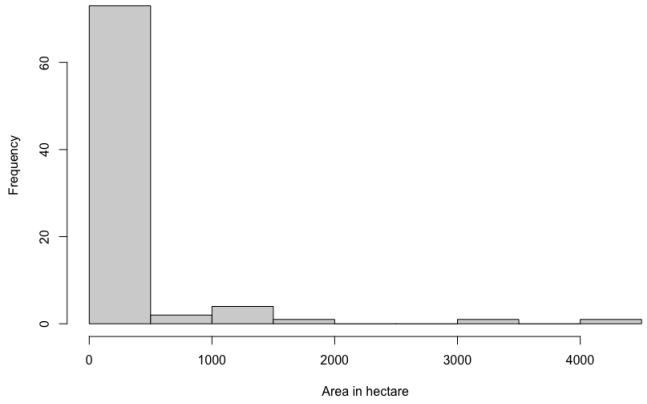


Figure 9: Number of activities from Buurkracht in a community

- **Area**
  The Area of the communites differs a lot as can be seen in figure 10. By taking the logarithm of the area, the distribution of the data improves, as can been seen in figure 10.



Mean	Min	Max
303.79	1.59	4100.23

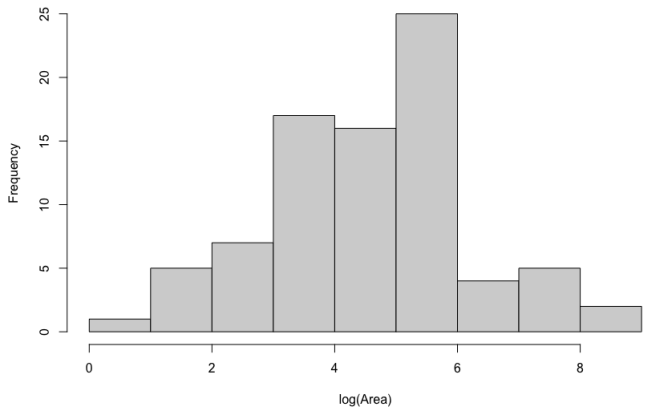


Figure 10: Area in hectare of the communities

6.4 Conclusions

After thoroughly looking at the data and deciding which data to remove, the dataset is now ready for regression analysis and the application of machine learning techniques. In the next chapters this will be done based on the two data sets that are left after the cleaning process: the complete set of households observations and the set that is grouped per community with neighbourhood characteristics.

7 Modelling results

In this chapter the modeling results will be discussed. The four methods, logistic regression, Poisson regression, random forest and causal forest, of which the theory is discussed in chapter 4 and chapter 5 are applied to the dataset. First, a logistic regression and a Poisson regression are performed. Next, the random forest model has been applied and finally the causal forest model has been applied. After the four methods have been applied the different results are compared and conclusions are drawn.

7.1 logistic regression

Regressions are conducted at two different levels: the household level and the community level. To identify any potential issues stemming from correlations between variables in the data, the correlations among the variables will be examined first. Subsequently, the results from both models will be discussed.

7.1.1 Correlation

Correlation coefficients above 0.7 and below -0.7 would give a reason to expect problems with the correlations between variables for the regression analysis (Moore et al. (2013)). As one can see in table 7 this is not the case for the variables in the constructed dataset. The correlation between dwelling value and electricity use is almost at this boundary of 0.7 and also the correlation between value of a dwelling and the dwelling area is relatively high. This should be kept in mind, but generally no problems are expected regarding the correlation between variables.

Table 7: Pearson correlation

	Dwelling Value	Electricity	Gas	Year<1975	Year>1992	Distance to initiator
Living Area	0.610	0.588	0.081	−0.040	0.117	0.002
Dwelling Value	x	0.698	−0.025	−0.119	0.284	0.012
Electricity		x	0.086	0.000	0.075	−0.011
Gas			x	0.243	−0.421	−0.105
Year<1975				x	−0.523	−0.031
Year>1992					x	0.079
Distance to initiator						x

In table 8 the variance inflation factor is given for the different variables. The VIF is a measure used to detect multicollinearity among predictor variables in regression analysis. A rule of thumb from Cohen et al. (2003) is that VIF values exceeding 10 may indicate potential multicollinearity issues. Since all VIF values are below 10 multicollinearity will probably not be an issue. The gas dummy is left out of this analysis since this dummy is based on the variable gas, thus between those variables the correlation will definitely be very high.

Table 8: VIF

Distance to initiator	Living Area	Dwelling Value	Electricity	Gas	Year>1992	Year<1975
1.033	1.731	2.368	2.119	1.323	1.577	1.236

From both the pearson correlation and the variance inflation factor it can be concluded that correlation between the variables should not be a problem in this research.

7.1.2 Household level

The logistic regression at the household level is performed on the variables in the dataset. A stepwise method in both directions is used to select the relevant variables. A stepwise method is an automated procedure for selecting variables to include or exclude from a regression model. For this model the stepwise procedure from the MASS package in R is used (R Documentation (2022)). The procedure adds and removes variables based on significance and non-significance of the variables. This leads to the following formula:

$$\begin{aligned}\ln(\frac{P(\text{Measure}=1)}{P(\text{Measure}=0)}) = \beta_0 + \beta_1 \ln(\text{Gas}) + \beta_2 \ln(\text{Electricity}) + \beta_3 \text{Gas dummy} + \beta_4 \text{Distance to initiator 50-100} \\ + \beta_5 \text{Distance to initiator 100-200} + \beta_6 \text{Distance to initiator 200-300} + \beta_7 \text{Distance to} \\ \text{initiator 300-400} + \beta_8 \text{Distance to initiator 400+} + \beta_9 \ln(\text{Living area}) \\ + \beta_{10} \ln(\text{Value dwelling}) + \beta_{11} \text{Construction year after 1992+} \\ \beta_{12} \text{Construction year before 1975} + \gamma_{1\ldots m} \text{Community fixed effect} \end{aligned} \tag{7}$$

The results from the logistic regression analysis are summarized in table 9. The table includes the coefficients associated with the predictor variables, while the constant term (intercept) and fixed effects are excluded from the table. These are omitted as they provide information about the baseline or reference category and may not offer meaningful standalone insights. Including these additional parameters would result in a more extensive table of several pages.

Table 9: Logistic regression - Households level

Variable	Coefficient (Standard Error)
ln(Gas)	-0.519 $(0.155)^{***}$
ln(Electricity)	-0.635 $(0.175)^{***}$
Gas dummy	3.830 $(1.122)^{***}$
Dist initiator 50_100	-0.882 $(0.089)^{***}$
Dist initiator 100_200	-1.127 $(0.081)^{***}$
Dist initiator 200_300	-1.476 $(0.093)^{***}$
Dist initiator 300_400	-1.703 $(0.107)^{***}$
Dist initiator 400+	-2.348 $(0.092)^{***}$
ln(Living Area)	1.137 $(0.103)^{***}$
ln(Dwelling Value)	0.999 $(0.160)^{***}$
Construction year after 1992	-1.037 $(0.106)^{***}$
Construction year before 1975	-0.349 $(0.068)^{***}$
Community fixed effects (82)	Yes
Num. obs.	74,750
McFadden's R-squared	0.128 (df=94)
Hosmer-Lemeshow	24.677 (df=8), p-value =0.002
Note:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

As can be seen in table 9 all coefficients are significant. This is the case because a stepwise method was used, which excludes variables that do not improve the model. In the conceptual model, we initially considered more variables, as shown in figure 1. Some variables were excluded from the regression analysis due to lack of available information in the dataset, while others were excluded by the stepwise method. Both McFadden's R-squared and the Hosmer-Lemeshow test statistic indicate a disappointing model fit. Both gas and electricity use have a negative effect on the purchase of a measure if the usage increases. However, the gas dummy variable has a positive effect, indicating that households using gas are more likely to purchase a measure. An increase in the distance to the initiator has a progressively negative effect on the likelihood of purchasing a measure, as indicated by the decreasing coefficients for the distance to initiator categories. A larger living area makes the purchase of a measure more likely. A higher value of the dwelling gives

a bigger chance on the purchase of a measure and both a construction year before 1975 and after 1992 have a negative effect on the purchase, which means that a construction year between 1975 and 1992 leads to a increased chance of purchasing a measure. In the regression community fixed effects were taken into account. The coefficients obtained for each community in this regression will be utilized for the regression at the community level.

7.1.3 Community level

At the community level a regression will be conducted with the community fixed effects as the dependent variable. These community fixed effects were obtained from the logistic regression performed in the previous section, the results of which are presented in table 9. By examining the impact of various neighbourhood variables on the community fixed effect, we can determine the effect of these variables on the likelihood of a measure being taken. The community fixed effect indicates the influence of community characteristics on the likelihood of a measure being taken. In table 10 the results from different regressions can be found. M1 represents the first model and is based on the following formula:

$$\begin{aligned}\text{community fixed effect} = \beta_0 &+ \beta_1 \text{Number of households in the community} + \beta_2 \text{Number of initiators per capita} \\ &+ \beta_3 \text{Area of the community} + \beta_4 \text{Number of activities in the community} \end{aligned} \tag{8}$$

Stepwise M1 is based on M1 and the stepwise method in both directions is used to select the relevant variables. For M2 the formula slightly changed. The number of activities is replaced by 3 new categories for the 3 different types of activities. These three categories were constructed in the following way. In category 1 are all activities that are meetings/gatherings. Category 2 are all flyers, folders, stickers, posters and so on, all communication/promotion on paper. Category 3 are all other activities for example questionnaires, scans or giveaways. In category 1 486 activities are present in the dataset, in category 2 2381 activities are in the dataset and 598 activities of category 3 are present in the dataset. In Stepwise M2 the stepwise method is applied to M2 to find the relevant variables. The formula corresponding to M2 is:

$$\begin{aligned}\text{community fixed effect} = \beta_0 &+ \beta_1 \text{Number of households in the community} + \beta_2 \text{Number of initiators per capita} \\ &+ \beta_3 \text{Area of the community} + \beta_4 \text{Number of activities in the community of category 1} \\ &+ \beta_5 \text{Number of activities in the community of category 2} \\ &+ \beta_6 \text{Number of activities in the community of category 3} \end{aligned} \tag{9}$$

In table 10 it can be seen that the neighbourhood variables that are present in the used dataset are not able to predict the community fixed effects accurately. None of the coefficients obtained from the M1 regression are significant at the 10% level. Applying the stepwise method on M1 results in leaving only the number of households in the community in the model and the coefficient is not significant at the 10% level. Dividing the activities in three categories gives a bit more information. In this case the number of households is significant as well as the activities in category 2. The effect of the activities in category 2 is negative and more households in the community has a negative effect on the fixed effect of the community and thus on the chance that measures will be applied. In the appendix the list of activities per category can be found and an extra regression at the community level on the conversion rate has been performed, not leading to interesting results. In the appendix more test that were performed but do not lead to relevant results can be found.

7.1.4 Community level - Poisson regression

Another test performed at the community level is the Poisson regression. The dependent variable used is the number of measures taken in the community. The independent variables used in the regression can be found in table 11.
As one can see, the total area of the community, the average dwelling value and the average gas use are not significant in predicting the number of households that do apply a measure. The constant has by far the

Table 10: Regressions on community fixed effects

	M1	Stepwise M1	M2	Stepwise M2
# Households	−0.0002 (0.0002)	−0.003 (0.0002)	−0.0003 (0.0002)	−0.0003 (0.0002)*
Initiators per capita	16.719 (28.442)		17.721 (28.157)	
Area of community	0.0004 (0.0003)		0.0004 (0.0003)	
# Activities	−0.023 (0.018)			
# Activities Cat. 1			0.222 (0.015)	0.238 (0.148)
# Activities Cat. 2			−0.083 (0.035)**	−0.074 (0.033)**
# Activities Cat. 3			0.087 (0.127)	
Observations	82	82	82	82
R^2	0.065	0.028	0.110	0.086
Adjusted R^2	0.016	0.016	0.039	0.051
Residual Std. Error	1.706 (df = 77)	1.707 (df = 80)	1.687 (df = 75)	1.676 (df = 78)
F Statistic	1.338 (df = 4; 77)	2.279 (df = 1; 80)	1.546 (df = 6; 75)	2.457* (df = 3; 78)

Notes: *p<0.1; **p<0.05; ***p<0.01, Standard errors are in parentheses

largest value hence, the number of households that do take a measure is largely explained by factors other than the included variables. All other coefficients are very small and the effects of the variables are thus not significant in relation to the the unexplained variables included in the constant. Also, the large Pearson chi-squared statistic with a p-value of 0 suggests that the model does not fit the observed data.

7.2 Random forest

The target variable in the random forest is the variable "Measure". This variable can have the values "Measure" or "No Measure". "Measure" is seen as class 1 and "No Measure" as class 0. The variables that are included to build the random forest are: ln(Gas), ln(Electricty), Gas dummy, ln(Distance to initiator), ln(Area), ln(Value) and Construction year categories. The forest is constructed using 500 trees. The Out-of-Bag error rate from the constructed random forest is 2.41 %. This implies that the forest can predict the outcome of "Measure" fairly accurate. However when we look at the confusion matrix in table 12 we see that this happens because "No measure" is predicted correct in 99.88% of the cases and "Measure" is only predicted correct in 6.95% of the cases. Due to class imbalance in the data, meaning that only a small portion of the dataset has a positive outcome (adopts a measure), while a large portion of the data has a negative outcome (no measure), the overall results from the random forest may appear accurate, but in reality, they are not accurate at all. In fact, a model that would always predict "no measure" in all cases, would be correct in 97.5% of cases, because of this class imbalance.
In table 13 the result of including more trees in the forest is presented. The addition of more trees could theoretically result in a more stable model with more robust predictions, thus increasing generalizability. In the table per 50 extra trees in the forest the predictions are given. As can be seen the error rates do not decrease by including more trees in the forest.

In table 14 the variable importance of the included variables in the forest is given. This variable importance is measured by the mean decrease in accuracy. The mean decrease in accuracy is calculated by comparing the accuracy of the model before and after randomly shuffling the values of a specific variable. The importance of the variables signifies the impact of the variables on the predictive accuracy of the model. A higher value indicates a more significant contribution to the model's accuracy. This means that in this random forest the distance to the initiator is the most important variable and the gas dummy the least important variable.

Table 11: Poisson Regression - Community level

Variable	Coefficient (Standard Error)
Number of households	0.0004 (0.00003)***
Total area of the community	−0.000 (0.00004)
Number of activities	0.009 (0.002)***
Average living area	0.009 (0.002)***
Average construction year	0.010 (0.002)***
Average dwelling value	−0.001 (0.001)
Average gas use	0.00005 (0.0001)
Average electricity use	−0.0003 (0.0001)***
Constant	−18.232 (4.360)***
Observations	82
Log Likelihood	−473.915
Pearson chi-squared	588.42, $p - value = 0$
Null Deviance	1147.11
Residual Deviance	566.89
Degrees of Freedom (df)	73

Note: *p<0.1; **p<0.05; ***p<0.01

Table 12: Confusion Matrix and Class Error - Random forest

	0	1	Class Error
Actual 0	72818	90	0.0012
Actual 1	1714	128	0.9305

Partial dependence plots

In table 15 partial dependence plots are given for the variables that were used to create the random forest. In a partial dependence plot the relationship between a variable and the predicted outcome of a machine learning model is shown. All other variables are held constant. These plots are helpful to understand the marginal effect of a single variable on the outcome. The plots are especially useful in explaining the output from black box models like random forests (Greenwell (2017)). It is crucial to exercise caution when interpreting the results from a partial dependence plot. One underlying assumption is that all other variables are held constant, which may not always align with real-world scenarios.
Analysis of the plot reveals the following insights:

- The likelihood of implementing a measure is the same for different amounts of gas usage until a certain limit and after this limit an increased gas usage leads to a decreased likelihood of implementing a measure.
- The relationship between electricity usage and the likelihood of applying a measure is not clear.
- An increased distance to the initiator initially increases the likelihood of a measure. However, beyond a certain threshold, the increase diminishes.
- Smaller living areas of a dwelling have a bigger positive effect on the likelihood of implementing a measure.
- The more expensive a dwelling becomes the smaller the likelihood of implementing a measure, with an unclear relation in the medium priced segment.
- Construction years before 1975 have the biggest effect on the likelihood of implementing a measures, whereas buildings constructed between 1975 and 1992 exhibit the smallest effect.

Table 13: Out-of-Bag Error Rates for Different Numbers of Trees - Random forest

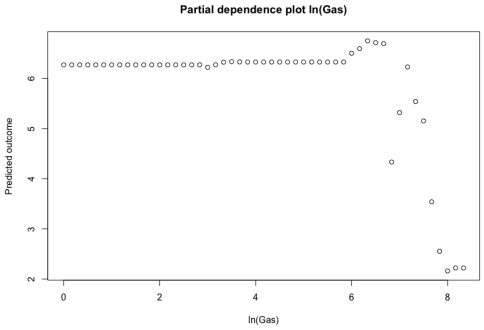
number of trees	OOB	No Measure error rate	Measure error rate
50	2.46%	0.17%	92.89%
100	2.43%	0.15%	92.62%
150	2.42%	0.14%	92.67%
200	2.42%	0.14%	92.83%
250	2.42%	0.13%	93.11%
300	2.41%	0.13%	92.94%
350	2.41%	0.12%	92.94%
400	2.41%	0.12%	92.89%
450	2.41%	0.13%	92.94%
500	2.41%	0.12%	93.05%

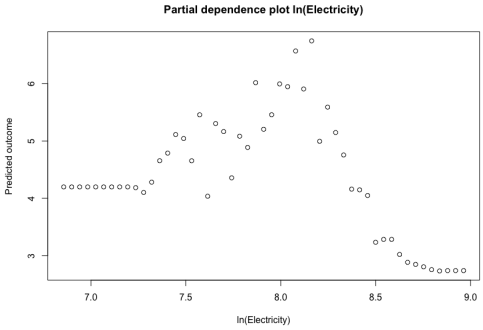
Table 14: Variable Importance - Mean Decrease in Accuracy - Random forest

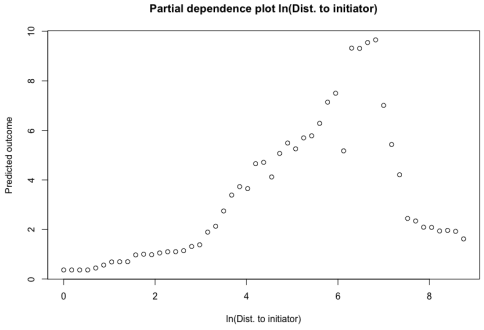
Variable	Mean Decrease Accuracy
ln(Dist. to initiator)	79.55
ln(Living area)	72.80
ln(Dwelling value)	60.67
ln(Electricity)	59.77
ln(Gas)	54.11
Construction year	49.51
Gas dummy	5.41

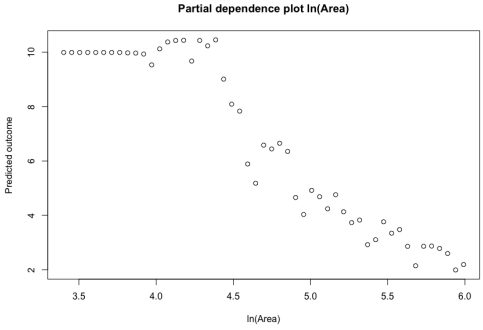
- Gas usage, as indicated by the gas dummy variable, positively influences the likelihood of implementing a measure.

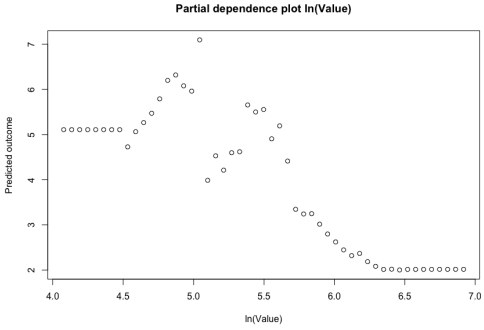
Table 15: Partial dependence plots - Random forest

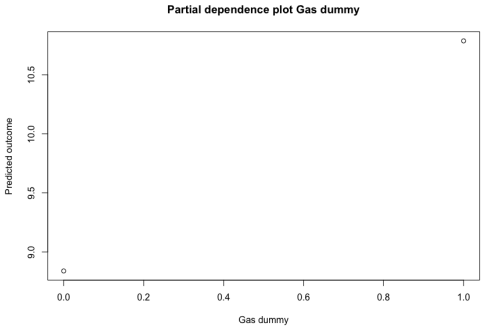












7.3 Causal forest

The causal forest is a nonparametric method for heterogeneous treatment effect estimation (Wager and Athey (2015)). Ideally this method would be used to estimate the effect of the treatment, previously referred to as the neighbourhood approach or the Buurkracht campaign. All data in our dataset however has received this treatment. Therefore, in order to use the causal forest method another treatment has to be selected. A big contribution to the campaign is delivered by the initiators of the campaign. In the previous models the distance to the initiator was an important variable. The most important in the random forest even. It could be argued that households closer to the initiator receive more treatment from the campaign than households further away from the initiators. This is why it has been decided to use the distance to the initiator as the treatment in the causal forest. More specifically, the treatment used is: distance to an initiator being 200 meters or less. This distance of 200 meters is chosen because of figure 11. From the figure, which illustrates the likelihood of implementing a measure plotted against the distance to the nearest initiator, it becomes clear that beyond a distance of 200 meters, the likelihood of implementation does not change much.

The causal forest was created by dividing the dataset in a training set and a test set. The training set contained 70% of the data and the test set the other 30%. This division in two groups with a 70/30 ratio for training and validation purposes is often used in machine learning because it often gives the best results like in the research from Nguyen et al. (2021). This distribution of the data over the sets is made randomly. The forest is created with 5000 trees. The variable importance from the variables used to build the causal forest can be found in table 16. The variable importance is also depicted in figure 12. A confusion matrix cannot be made as easily for a causal forest as for a random forest. The predictions from the causal forest are given as a likelihood for measure adoption rather than as direct predicted values. Which means that a threshold should be chosen above which the likelihood will lead to a prediction of one in order to compare the predictions to the actual decision that has been made to adopt a measure or not. In order to test the predictive power of the causal forest this will be done in chapter 9.

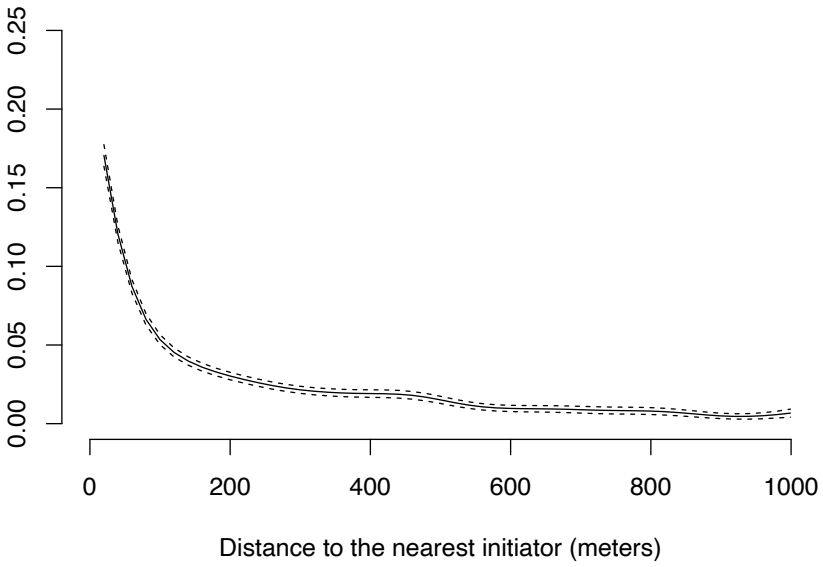


Figure 11: The chance of applying a measure against the distance to the initiator

Table 16: Variable importance - Causal forest

Variable	Mean Decrease Accuracy
ln(Living area)	3.84×10^{-1}
ln(Dwelling value)	3.27×10^{-1}
ln(Electricity)	1.21×10^{-1}
ln(Gas)	1.19×10^{-1}
Construction year 1975-1992	3.58×10^{-2}
Construction year before 1975	7.28×10^{-3}
Construction year after 1992	6.41×10^{-3}
Gas dummy	1.72×10^{-5}

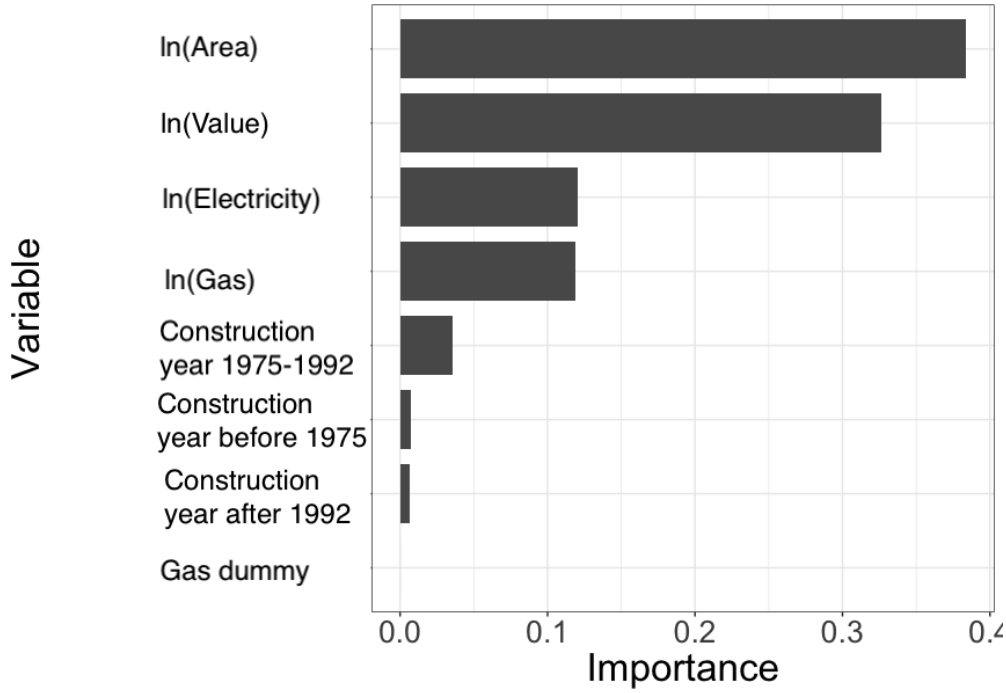


Figure 12: Variable importance - Causal forest

To test whether the treatment effect is equal to zero a one sample t-test is used, this gives the following results as depicted in table 17. The p-value is close to zero and hence there is strong evidence against the null hypothesis, which says that the estimated treatment effect is equal to zero. The confidence interval does not include zero and the sample mean estimate is a positive number, hence it can be concluded that there is strong evidence that suggests that the true mean of the estimated treatment effects is significantly different from zero.

Table 17: Results of One Sample t-test on treatment effect

Test	One Sample t-test
Data	estimated causal effects
t-value	221
Df	52324
P-value	$< 2.2 \times 10^{-16}$
Alternative hypothesis	true mean is not equal to 0
95% Confidence Interval	(0.03669095, 0.03734757)
Sample mean estimate	0.03701926

Next the predicted treatment effects are used to select the people that are targeted from the test set. In the first regression (t1) only the people with the 50% highest treatments effects are targeted. In the second regression (t2) 50% of the test set is randomly targeted. In table 18 the results from both regressions are given. As can be seen in the first regression the effect of the treatment (an initiator within 200 meters) is bigger, 0.051 for t1 and 0.033 for t2, for the people that have an higher predicted treatment effect which is predicted using the causal forest. This suggest that the causal forest can be used to select the people that should be targeted with a treatment.

Table 18: Linear Regression Results

Model	Intercept	Treatment	Std. Error	P-value
t1	0.01903441	0.05110892	0.002017569	4.71×10^{-21}
t2	0.01666876	0.03251157	0.001775494	7.28×10^{-21}

In table 19 the mean values of the different variables are given for different quintiles. These quintiles are made by sorting the data based on the predicted treatment effect and then taking the 20% with the lowest predicted treatment effect and so on. The higher the predicted treatment effect the higher the gas use. For the lower predicted treatment effects we can see a strong increase in electricity use with an increase in treatment effect, but for the higher predicted treatment effects this relations is not seen. Dwelling area is higher for the higher treatment effects and the same goes for dwelling value. For the construction years, no clear relationship is visible.

Table 19: Mean variable values in 5 quintiles

Variable	All	0-20	20-40	40-60	60-80	80-100
ln(Gas)	6.9347	6.8055	6.7213	6.9010	7.0489	7.1966
ln(Electricity)	7.9785	7.7155	7.9663	8.0485	8.0936	8.0684
Gas dummy	0.9551	0.9694	0.9329	0.9420	0.9541	0.9773
ln(Living area)	4.8157	4.5217	4.6992	4.8522	4.9901	5.0153
ln(Dwelling value)	5.3623	5.0064	5.2855	5.4532	5.5408	5.5253
Construction year after 1992	0.2490	0.2119	0.2794	0.2974	0.2589	0.1973
Construction year before 1975	0.4532	0.4884	0.4571	0.4582	0.4649	0.3975
Construction year 1975-1992	0.2978	0.2997	0.2635	0.2444	0.2763	0.4053

7.4 Comparison

In this section the results from the previous applied methods on the data will be compared to each other. This will be done by looking at the relevant variables separately. It is important to note that all households in the data have been subjected to a neighbourhood approach, implying that the effects of the variables on the decision are also influenced by this neighbourhood campaign from Buurkracht. Therefore, the relationship between each variable and the decision to apply a measure is not solely based on the variable itself but is

rather conditioned on having received the neighborhood approach.

Gas and Gas dummy
From all three models it follows that gas use in itself is an indicator for an increased chance of applying a measure, but the models do not agree on what an increase in gas use does for this chance. The logistic regression suggest more gas use has a decreasing effect on the chance, where the random forest shows that this is only the effect after a certain threshold and before that threshold the effect of gas use on the chance is constant. These results are not as one would expect beforehand. One would expect that with an increased gas use the demand for energy efficient measures would increase, since this would in the end result in saving money. The observation that the use of gas correlates with an increased likelihood of applying a measure aligns with initial expectations. It is reasonable to anticipate that households already disconnected from the gas may have undertaken additional energy efficiency measures compared to those that still use gas. It is essential to acknowledge that the dataset does not contain pre-existing measures implemented by households. Consequently, this limitation may have influenced the analysis by not accounting for the cumulative effect of previously adopted energy-saving practices. By not incorporating these existing measures, this study may have underestimated the true impact of gas usage and other variables on the adoption of energy-saving measures.

Electricity
The logistic regression model indicates that increased electricity usage has a negative impact on the likelihood of implementing energy-saving measures. However, both the random forest and causal forest models fail to find an interpretable relationship between electricity usage and the adoption of energy-saving measures. This raises questions about the true nature of the relationship between electricity consumption and the energy-saving measures. The negative effect from electricity use on the adoption of energy efficient measures is not as one would expect. With more use one would expect that the urge to apply a measure would increase. This deviation from expectations may be due to the nature of the electricity consumption data, which, like gas consumption data, are aggregated at the postal code level. As such, these statistics represent averages across multiple households but are treated as if they relate to individual households. Unfortunately, this aggregation was necessary because household-level electricity consumption data were not available. It also remains unclear what electricity use is included in the data, no information is available on whether the yield from the solar panels has been deducted or not. One possible explanation for the negative effect of increased electricity use on the adoption of energy-efficient measures could be that households have already installed a heat pump, replacing gas heating. This change leads to higher electricity consumption and may reduce the likelihood of these households implementing additional energy-saving measures during the neighbourhood approach.

Distance to initiator
Distance to an initiator is in the logistic regression model a negative predictor for the chance of applying a measure, the further away the smaller the chance. This finding is in line with intuitive expectations, suggesting that households located farther away from initiators are less inclined to take energy efficient measures. Interestingly, distance to an initiator is indicated as the most important variable in the random forest, but the direction of the effect is opposite of that from the logistic regression. This discrepancy underscores the complexity of the relationship between distance and measure adoption. In the causal forest we have used an initiator within 200 meters as the treatment, where the treatment is shown to be significant. This aligns with the logistic regression model results, that greater distances correspond to diminished probabilities of adoption. A possible problem in the random forest could be that fixed community effects are not included, which is the case in the logistic regression. The lack of these community effects in the model could result in less accurate predictions, particularly for individual cases heavily influenced by these community effects.

Living Area
For the living area variable the models are not in line with each other, the logistic regression and the causal forest give a positive relationship, but the random forest gives a negative relationship. One would expect a positive relationship in the sense that more living area means more space to heat, a higher electricity use, more loss of energy via poor insulation and a higher property value hence more money and thus bigger savings in the case of improvements. The results from the logistic regression and causal forest thus make the most sense.

Dwelling value
The dwelling value has a positive relation with the chance of applying a measure according to the causal forest and the logistic regression, for the random forest the relation is not clear. The relationship in the logistic regression and the causal forest is in the direction one would expect the relationship to be. A higher value would suggest more money and thus more money available to invest.

Construction year
The effect of the construction years differs per model and is not considered important in the causal and random forest. For the logistic regression the construction years between 1975 and 1992 are considered to have the highest positive effect on applying a measure. This could be explained by more regulations for older buildings due to its historical value, hence a higher threshold to implement energy efficient measures, or these buildings might have already been improved. A reason why the newer buildings lead to a lower chance of applying a measure could be that these dwellings do not require the measures, since they are already better insulated and already of the gas grid for example.

7.5 Conclusions

The logistic regression provides insights into the factors influencing the likelihood of implementing a measure at both household and community levels and generally aligns the best with intuitive expectations. The Poisson regression does not provide insights into the number of measures taken in a community, since to much information is included in the constant.The random forest suffers from class imbalance issues, leading to inaccurate predictions for the minority class. The causal forest analysis confirms the significant impact of the treatment (distance to initiator) on the likelihood of implementing a measure, providing valuable insights into causal relationships. In the next chapter, we will further discuss the problems with class imbalance that became evident when using the random and causal forest. Finally, in chapter 9, the predictive use of the models will be compared based on five evaluation metrics.

8 Class imbalance

One major issue that followed from the modelling results is that not applying a measure is often predicted correctly in contrast to the prediction of applying a measure, this is often predicted incorrectly. In this chapter it is discussed why this happens and how this could be solved.

8.1 Introduction

As can be seen from the data, the number of households that implement a measure is far less than the households that do not implement a measure. This class imbalance leads to problems in the prediction of the classes using the models, especially for the use of the random forest model. The random forest predicts "No measure" accurately but the prediction of "Measure" is way off. This is mainly due to the class imbalance. A solution to this problem of class imbalance has to be found.

Abd Elrahman and Abraham (2013) have written a review about class imbalance problems. They summarize various solutions given by different researchers. There is not one general approach that works best for all class imbalance problems. For the problem of this research the focus will be on sampling methods. In sampling methods there is the possibility of over sampling and under sampling. In oversampling, extra copies of the minority class, in our case "Measure", are created. In under sampling, data entries from the majority class, in our case "No Measure", are removed.

8.2 Over sampling

The simplest methods of over sampling is the random replication of positive examples (García et al. (2012)). This makes it more likely that overfitting will occur. Chawla et al. (2002) have proposed a method that is called SMOTE, Synthetic Minority Over-Sampling TEchnique. This method is an over sampling technique that creates new instances of the minority class by interpolating between existing positive examples that are in close proximity. This approach aims to address the overfitting issue that is associated with random replication. The SMOTE algorithm relies on the k nearest neighbour algorithm. These nearest neighbours are used to create the synthetic samples by interpolating between the existing sample and the randomly selected neighbours. SMOTE can only be used for binary class problems, which aligns with our challenge. SMOTE is however not suitable for continuous independent variables. This means that we need to consider another method. In R there is a package ROSE that uses a function called ovun.sample. This function is based on SMOTE in combination with under sampling based on Edited Nearest Neighbors. This method removes synthetic examples whose labels differ from the majority of their k-nearest neighbours.

8.3 Under sampling

In under sampling the number of instances of the majority class is reduced. Under sampling may lead to the loss of useful information since data is removed. The most straightforward method of under sampling is randomly removing instances from the majority class. The data set will become more balanced, but valuable information might be lost. A more sophisticated technique for under sampling is Tomek Links. This algorithm eliminates boundary instances (Devi et al. (2017)). Essentially, it removes instances of the majority class if they are located too close to an instance of the minority class or if they are to similar to the instances of the minority class. This is done by pairing instances from both classes that are similar, a Tomek-link pair. Under sampling is especially useful in the case of a large dataset, as over sampling would require too much computational power in that case. The main limitation of under sampling is the loss of potentially crucial information. Several under sampling techniques have been proposed in the literature, but a detailed exploration goes beyond the scope of this research.

8.4 Conclusions

In this chapter, the effect of class imbalance in the dataset on modeling results was addressed. The imbalance is caused by a considerable difference between households implementing measures and those that do not. Consequently, modeling, especially with the random forest algorithm, struggles to accurately predict the minority class.

Various methods to address class imbalance were explored, with a focus on over-sampling and under-sampling techniques. Over-sampling methods aim to generate synthetic instances of the minority class and under-sampling methods selectively reduce instances of the majority class to balance the dataset.

A combination of an over and under sampling method has been applied to the dataset. The logistic regression, random forest model and causal forest model were re-run on the balanced dataset. While logistic regression results showed minor differences, the random forest model exhibited significant improvements in prediction accuracy for both classes, but the effects of the different variables found by the partial dependence plots were not convincing. The results from the causal forest do not show improvement. The variable importance was the same for the most important variables, only in the construction year categories the results changed. For the impact of the different variables on the decision to apply an energy efficient measure, making the dataset more balanced did not improve the results, for model prediction a more balanced dataset leads to improved prediction results. The results from the application of the method on the data and next on the models can be found in the appendix. In the next chapter model predictions on a more balanced dataset will be compared to model predictions on the original dataset.

9 Predictive accuracy of the models

In this chapter the predictive accuracy of the previously determined models, logistic regression, random forest and causal forest, will be tested. This will be done with the original dataset and with a new dataset. The new test dataset is made by deleting part of the previously used dataset. In the previous dataset problems occurred with class imbalance. There were far more households that did not apply any energy efficient measure than there were households that did apply energy efficient measures to their dwellings. This resulted in skewed results and especially in the results from the machine learning models it was shown that applying a measure was predicted incorrectly in many cases, even though not applying a measure was predicted correctly in many cases.

The new dataset is created by removing observations/households that do not apply a measure. The observations are selected randomly. It has been chosen to randomly select 4 times as many households that do not apply a measure as households that do apply a measure. This means that all observations are kept that do apply a measure(1842) and 7368 households are chosen randomly that do not apply a measure. This approach was chosen after several other less random options were explored. The other options that were considered are:

- Keeping all entries within a certain distance of an initiator. This option showed not to be feasible since the percentage of measures taken always stayed too small to not experience problems from class imbalance. With lowering the distance to the initiator also the number of measures taken declined and below 200 meters there would be a problem with the causal forest model since this is based on the treatment of being within 200 meters of an initiator, hence this model could not be used if all observations would be within these 200 meters.
- The second option that was considered, was keeping all observations within a certain distance of households that had taken a measure. This however did not lead to a good percentage of measures taken if this distance was above 50 meters, which would mean that only entire streets would be kept in which a measure was taken. Since data is used on PC6 level this would lead to too much of the same data.

In order to test the models the test data is divided in two sets: a training set and a test set. The train set consists of 70% of the data and the test set contains the other 30%. These 70 and 30 percent of the households are selected at random. The three models, logistic, random forest and causal forest, will be tested based on 5 evaluation metrics. These evaluation metrics and how they are calculated are listed below.

- **Accuracy**
  Overall correctness of the model
  (true positive + true negative)/ all cases

- **Precision**
  Correctly predicted positive cases of all predicted positive cases
  true positive/ (true positive + false positive)

- **Recall**
  Correctly predicted positive cases of all actual positive cases
  true positive/ (true positive + false negative)

- **Specificity**
  Correctly predicted negative cases of all actual negative cases
  true negative / (true negative + false positive)

- **F1 score**
  Measure of model accuracy. It is often used in machine learning and is especially useful in the case of class imbalance.
  2* (precision * recall)/(precision + recall)

9.1 Descriptive statistics from the dataset

Creating the dataset as explained before leaves us with 9210 observations from which 1842 households (20%) apply an energy efficient measure. Some descriptives from the data are given in table 20. Compared to the descriptives from the original dataset in table 6 the means are approximately the same, the main difference is the percentage of households that do apply a measure. In this new dataset the percentage is 20% where in the original dataset this percentage is 2.5%.

Table 20: Descriptives - new dataset - household level

Statistic	N	Mean	St. Dev.	Min	Max
vbo_opp	9,210	134.032	48.798	31	397
pand_bouwjaar	9,210	1,974.054	27.441	1,644	2,019
n_Categorie1	9,210	2.797	1.628	0	11
n_Categorie2	9,210	13.109	6.844	1	43
n_Categorie3	9,210	3.030	1.757	0	8
n_activities	9,210	18.936	8.899	2	60
mtr	9,210	0.200	0.400	0	1
dist2init	9,210	440.052	407.156	0	6,275
ln_g_elek_won	9,210	7.982	0.265	6.856	8.963
ln_g_gas_won	9,210	6.916	1.593	0.000	8.331
ln_vbo_opp	9,210	4.836	0.352	3.434	5.984
ln_wozwoning	9,210	5.372	0.371	4.094	6.918
bouwjaar_n1992	9,210	0.233	0.423	0	1
bouwjaar_v1975	9,210	0.451	0.498	0	1
bouwjaar_n1975v1992	9,210	0.315	0.465	0	1
gas_dummy	9,210	0.951	0.215	0	1

9.2 Logistic regression

In order to test the logistic regression model first the logistic regression model is determined with the training dataset, consisting of 70% of the 9210 observations. After this predictions are made on the test dataset, the other 30% of the 9210 observations. With these predictions the evaluation metrics can be determined, since the predicted outcomes can be compared to the actual outcomes. The predicted outcomes are not binary, but they lie between 0 and 1, hence they will be converted to a binary outcome by setting a threshold at 0.5 and everything below will be set to 0 and everything above will be converted to 1. The outcomes from the performance evaluation metrics can be found in table 21.

Table 21: Performance evaluation metrics for the logistic regression model

Metric	Value test data	Value original data
Accuracy	0.803	0.976
Precision	0.568	0.000
Recall	0.239	0.000
Specificity	0.952	1.000
F1 score	0.336	n.a.

As one can see the differences between the evaluation metrics for the test data and the original data are very big. This has to do with the class imbalance of the original data. The old model predicted the zeros very good but the ones very bad. This is the reason why the precision and recall are 0 since they are a measure of correctly predicted positive cases. The accuracy and specificity also include the correctly predicted negative outcomes, hence these numbers are high. The F1 score cannot be predicted since the precision and recall are zero.

9.3 Random forest

The random forest is created based on the training set. After this the random forest model is tested with the test set. This leads to the evaluation metrics and their numbers in table 22.

Table 22: Performance evaluation metrics for the random forest model

Metric	Value test data	Value original data
Accuracy	0.831	0.976
Precision	0.699	0.530
Recall	0.338	0.066
Specificity	0.962	0.999
F1 score	0.455	0.117

For the random forest the number of the original dataset seems better, except for the precision and recall and hence the F1 score. This suggests that the positive cases are predicted worse for the original data than for the test data. Again this has to do with the class imbalance. In the test data the recall has the lowest value, this means that the correctly predicted positive cases of all actual positive cases is still pretty low, but much better than for the original data. This makes sense since the test data is still unbalanced 20% positive and 80% negative.

9.4 Causal forest

The causal forest is build using the training set. To construct the forest, this set is again split into a 70-30 ratio, resulting in three subsets: one for building the forest (training set), one for assigning numbers to the forest (validation set) and one for prediction purposes (test set).

For the causal forest the performance evaluation is a bit more difficult. Where in the previous two models the predictions could be converted to binary predictions by setting a threshold at 0.5, this is not possible for the causal forest. In the causal forest all predictions lie below the threshold of 0.5 hence this would mean that the model would always predict no measure. The threshold should be set at a different number. Since we have created the dataset in such a way that 20% of the households will apply a measure, the threshold is set in such a way that the 20% households with the highest prediction in the causal forest will get the prediction measure taken and the other 80% will get the prediction no measure taken. This results in the performance evaluation metrics that can be found in table 23. In this table also the values for the original data are given. In this case also the threshold is set at 20%.

Table 23: Performance evaluation metrics for the causal forest model

Metric	Value test data	Value original data
Accuracy	0.600	0.782
Precision	0.024	0.011
Recall	0.022	0.096
Specificity	0.753	0.798
F1 score	0.023	0.020

The causal forest model does not necessarily perform better on the test data than on the original data, this is different than for the other models. This could have to do with a lot of factors, one of which is the threshold.

9.5 Comparison of the models

In table 24 the results from the different models on the test data are put together in one overview. As one can see the random forest has the best score on all parts and the logistic regression outperforms the causal forest on all parts. With the original data the random forest also seemed to perform the best. The F1 score was the highest, but much lower than for the test data.

Table 24: Performance evaluation metrics

	Logistic	Random forest	Causal forest
Accuracy	0.803	0.831	0.600
Precision	0.568	0.699	0.024
Recall	0.239	0.338	0.022
Specificity	0.952	0.962	0.753
F1 score	0.336	0.455	0.023

9.6 Conclusions

In this chapter the models have been evaluated based on five evaluation metrics and based on two different datasets: the original dataset and a constructed dataset. The five metrics are: accuracy, precision, recall, specificity and the F1 score. On all metrics the random forest gets the highest scores for the test dataset, which means that for a more balanced dataset the random forest model predicts the outcome the best. This is for a dataset that is balanced 20 percent positive outcomes and 80 percent negative outcomes. The logistic regression outperforms the causal forest in this case. A direct comparison between the models is difficult, since the causal forest model requires a treatment variable, which the other models do not. Also a balanced dataset is not a real-world scenario, it is highly unlikely that 20 percent of the households will apply a measure after a campaign by Buurkracht, because this has not yet happened so far. In selecting a good model the priority should be on correctly predicting positive cases. Recall should be seen as the most crucial criterion. For the original dataset, the causal forest scores the highest on recall. This suggests that the causal forest model is more effective in identifying positive cases, which is essential for predicting the adoption of energy efficient measures. However, one should note that the recall metric is very low for all models.

10 Conclusions and Future Research

In this final chapter, the findings of this thesis are summarized. The thesis is finished by providing suggestions for further research.

10.1 Conclusions

The aim of this thesis was to investigate which econometric or machine learning techniques is the preferred method to get insight into heterogeneity within and between various demographic groups in the adoption of clean energy technologies in neighbourhoods. In order to answer the main research question three sub-questions were formulated. The research was done in two parts. First it has been investigated, using the obtained data, which factors do influence the choice on whether or not to take energy efficient measures, next the models have been tested on their predictive power.

In this thesis data from the foundation Buurkracht has been used. First the Buurkracht data has been cleaned in such a way that it can be used for the desired methods. Next the data has been combined with the data from Statistics Netherlands. This complete data set has been used for a logistic regression model with fixed effects, an ordinary regression on the estimated fixed effects from the logistic regression, a Poisson regression model, a random forest model and a causal forest model. The Poisson regression model did not lead to insights as the constant in the model explained most of the variation in the data and the other coefficients were very small in comparison to this constant. From the other three models conclusions can be drawn regarding the effect of different variables on the decision to apply a measure or not. The variables that were investigated are: gas use, electricity use, distance to initiator, living area, dwelling value and construction year. The predictions from the logistic regression models regarding the variables were most in line with the expectations. This was not the case for the gas use and the electricity use.

Let us start with answering the first sub-question, *How can econometric and machine learning techniques be utilized to identify and explain the heterogeneity among various demographic groups in the adoption of clean energy technologies in neighbourhoods?* Econometric and machine learning techniques are not required to identify heterogeneity. That heterogeneity among different demographic groups in the adoption of clean energy technologies exists becomes clear from the data, otherwise the results from a neighbourhood approach would be the same in each community and for each household, this is clearly not the case. Econometric and machine learning techniques can be utilized to explain the heterogeneity among various demographic groups in the adoption of clean energy technologies in neighbourhoods. With the logistic regression, random forest and causal forest, effects of different variables, related to households or neighbourhoods, on the decision to apply a measure can be found. These effects explain the heterogeneity between the households and neighbourhoods in their decision to adopt clean energy technologies. However, in this research the directions of these effects do not comply with each other, indicating the need for further research.

Subsequently, the results from the models were compared to the results from the same models on a dataset that has been made more balanced using a method that combines over and under sampling. This did not lead to better results for the logistic regression model. The random forest model became more accurate and the variable importance slightly shifted, but still the distance to the initiator was the most important and the gas dummy the least important variable. Also the partial dependence plots do not provide new enlightening insights. For both the random forest and the causal forest a shift is seen in the importance of the newer dwellings in predicting a positive outcome.

The second sub-question, *What are the comparative strengths and weaknesses of econometric and machine learning techniques in explaining heterogeneity among various demographic groups in the adoption of clean energy technologies in neighbourhoods?*, has been addressed in the methodology part of the thesis. In summary, the logistic regression is user-friendly, does not require much computational power and gives information on the direction and importance of the variables, however the method is prone to overfitting, a lot of

assumptions should be met regarding the data and complex relationships are hard to capture. The random forest is more robust and less prone to overfitting and also provides an indication of the importance and direction of the effects of variables. The results are more difficult to interpret and more computational power is required. The logic from the model cannot be extracted, it is a black box model. The causal forest is especially focused on the treatment effect and hence difficult to compare to the other models that have not incorporated a treatment effect. The causal forest is more complex and requires more computational power.

The third sub-question, *How do the predictive accuracy of econometric and machine learning techniques compare in the context of clean energy technology adoption, as assessed by evaluation metrics?*, has been answered in chapter 9. The performance of the different constructed models has been compared using a test dataset. The models are compared on 5 different evaluations metrics: Accuracy, Precision, Recall, Specificity and F1 score. The random forest model outperformed the other models on all metrics for the test dataset. It is highly unlikely that a dataset will be balanced. The positive outcomes (implementing measures) should be seen as most important, because of the goal to implement as many measures as possible. Leading to the conclusion that the causal forest performs best in predicting, as this model performs best on recall for the original dataset. However, it should be noted that the prediction accuracy of the models is low on the original unbalanced dataset, especially regarding the positive outcomes.

Unfortunately, the research did not identify a preferred method to gain insight into the heterogeneity within and between various demographic groups in the adoption of clean energy technologies in neighbourhoods. An important limitation in identifying this method is the data quality. During the research several caveats of the data were identified. The data consists of many more homeowners that did not adopt a measure than homeowners that did adopt a measure (class imbalance). No household specific data on energy use, gas and electricity, was available. The values that were used for gas and electricity use were derived from the data on postal code level. Other households specific characteristics that were listed in the conceptual model were not available. Another notable missing element in current data and models is people's norms and values, which plays an important role in shaping their choices. Incorporating these factors into models could improve their predictive accuracy and understanding of the decision to apply an energy efficient measure. The parametric regressions identified a large unobserved community-specific variance (contained in the community fixed effects), this variance determines largely the outcome and complicates predictions. To use the causal forest in the way it was intended, by using as treatment the neighbourhood approach of Buurkracht, it is necessary to get data on households that did not receive the neighbourhood approach from Buurkracht. Also data on measures that were already taken before the campaign from Buurkracht started in the community is missing, which could explain part of the variation currently observed in the data.

10.2 Future Research

The research field on econometric and machine learning methods is broad, in this thesis only a small part of all possible methods were applied. Unfortunately, the results did not provide sufficient insights for Buurkracht to develop a more effective targeting strategy aimed at increasing conversion rates. Recommendations for further research are:

- Buurkracht could try to include more household specific data in their dataset. In this research CBS postal code data was used for household variables, but it would be more beneficial to include household specific data. In the conceptual model many variables are included, which in the end could not be used in the models, as the information was not available.
- In order to use the causal forest more effectively it would be advantageous to have a treatment and a control group, where the treatment group is the group that has taken part in a Buurkracht campaign. In this research all households were part of a Buurkracht campaign and thus it remains unclear whether the measures taken are a result of the campaign or from the rising awareness and popularity of energy efficient measures in society.

- As there are many methods that were not explored in the thesis, maybe some enlightening insides could follow from applying other methods on the data, the recommendations would be to focus on methods that are very good at handling unbalanced datasets.
- Norms and values play a significant role in shaping human behavior and decision-making processes. By integrating norms and values into predictive models, there is potential to enhance both predictive accuracy and understanding of decision-making. Future research should explore methods to effectively incorporate norms and values into models, thereby advancing the ability to predict and interpret human behavior more accurately.

Bibliography

Abd Elrahman, S. M. and Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1(2013): 332–340.

AIML.com (2023). What are advantages and disadvantages of logistic regression? https://aiml.com/what-are-advantages-and-disadvantages-of-logistic-regression/. Accessed: August 24, 2023.

Aydin, E. and Brounen, D. (2019). The impact of policy on residential energy consumption. *Energy*, 169: 115–129.

Aydin, E., Kok, N., and Brounen, D. (2017). Energy efficiency and household behavior: the rebound effect in the residential sector. *The RAND Journal of Economics*, 48(3): 749–782.

Bartlett, J. (2014). R-squared in logistic regression. https://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/. Accessed: May 5, 2024.

Besagni, G. and Borgarello, M. (2018). The determinants of residential energy expenditure in italy. *Energy*, 165: 369–386.

Bollinger, B. and Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, 31(6): 900–912.

Bollinger, B., Gillingham, K., Kirkpatrick, A. J., and Sexton, S. (2022). Visibility and peer influence in durable good adoption. *Marketing Science*, 41(3): 453–476.

Bollinger, B., Gillingham, K., Lamp, S., and Tsvetanov, T. (2024). Promotional campaign duration and word-of-mouth in solar panel adoption. *Marketing Science*.

Boon, F. P. and Dieperink, C. (2014). Local civil society based renewable energy organisations in the netherlands: Exploring the factors that stimulate their emergence and development. *Energy Policy*, 69: 297–307.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2): 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1): 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*.

Brounen, D., Kok, N., and Quigley, J. M. (2012). Residential energy use and conservation: Economics and demographics. *European Economic Review*, 56(5): 931–945. Green Building, the Economy, and Public Policy.

Brounen, D., Kok, N., and Quigley, J. M. (2013). Energy literacy, awareness, and conservation behavior of residential households. *Energy Economics*, 38: 42–50.

Buurkracht (n.d.). Buurkracht. https://www.buurkracht.nl/. Accessed: April 16, 2024.

Casaló, L. V. and Escario, J.-J. (2018). Heterogeneity in the association between environmental attitudes and pro-environmental behavior: A multilevel regression approach. *Journal of Cleaner Production*, 175: 155–163.

Casteren, T. v., Ossokina, I. V., and Arentze, T. A. (2013). Do you listen to your neighbour? block leader effects in community-led energy retrofits.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Mahwah, NJ.

De Groote, O., Pepermans, G., and Verboven, F. (2016). Heterogeneity in the adoption of photovoltaic systems in flanders. *Energy Economics*, 59: 45–57.

Devi, D., Purkayastha, B., et al. (2017). Redundancy-driven modified tomek-link based undersampling: A solution to class imbalance. *Pattern Recognition Letters*, 93: 3–12.

Energielinq (2019). Wat is de wijktransitie? https://energielinq.nl/wijktransitie/wat-is-de-wijktransitie/.

EZK (2019). Klimaatakkoord hoofdstuk gebouwde omgeving. https://www.klimaatakkoord.nl/gebouwde-omgeving/documenten/publicaties/2019/06/28/klimaatakkoord-hoofdstuk-gebouwde-omgeving.

EZK (2022). Gebouwde omgeving.

Filippidou, F., Nieboer, N., and Visscher, H. (2017). Are we moving fast enough? the energy renovation rate of the dutch non-profit housing using the national energy labelling database. *Energy Policy*, 109: 488–498.

Freedman, M. (2024). Why word of mouth trumps traditional advertising? https://www.businessnewsdaily.com/2353-consumer-ad-trust.html. Accessed: February 28, 2024.

García, V., Sánchez, J. S., and Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1): 13–21.

Gemeente Amsterdam (n.d.). Subsidie duurzame amsterdamse gebouwen – techniekneutraal aardgasvrij. https://www.amsterdam.nl/subsidies/subsidieregelingen/subsidie-techniekneutraal-aardgasvrij/. Accessed: May 15, 2023.

Gemeente Tilburg (n.d.). Duurzaam tilburg. https://www.duurzamertilburg.nl/. Accessed: May 15, 2023.

Graziano, M. and Gillingham, K. (2014). Spatial patterns of solar photovoltaic system adoption: The influence of neighbors and the built environment ‡. *Journal of Economic Geography*, 15(4): 815–839.

Green, J. and White, M. H., I. (2023). *Machine Learning for Experiments in the Social Sciences*. Elements Series in Experimental Political Science. Cambridge University Press.

Greenwell, B. M. (2017). pdp: An r package for constructing partial dependence plots. *R J.*, 9(1): 421.

Haag, G. D. (n.d.). Subsidies en leningen voor duurzame maatregelen. https://duurzamestad.denhaag.nl/woning/subsidies/. Accessed: May 15, 2023.

Hammerle, M., White, L. V., and Sturmberg, B. (2023). Solar for renters: Investigating investor perspectives of barriers and policies. *Energy Policy*, 174: 113417.

Hersch, J. and Viscusi, W. K. (2006). The generational divide in support for environmental policies: European evidence. *Climatic Change*, 77(1-2):121–136.

Hosmer, D. W., J., Lemeshow, S. A., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley, Hoboken, NJ, 3rd edition.

Hunter, L. M., Hatch, A., and Johnson, A. (2004). Cross-national gender variation in environmental behaviors*. *Social Science Quarterly*, 85(3): 677–694.

Jabeen, H. (2019). Poisson regression in r. https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/. Accessed: January 15, 2024.

Jain, A. (2018). Advantages and disadvantages of logistic regression in machine learning. https://medium.com/@akshayjain_757396/advantages-and-disadvantages-of-logistic-regression-in-machine-learning-a6a247e42b20. Accessed: August 24, 2023.

Jans, L., Djoera, E., and Sloot, D. (2023). Kunnen lokale energie-initiatieven mensen motiveren voor een duurzame energietransitie? *Real Estate Research Quarterly*, 22(4):110–118.

Jorna, J. (2024). Isde-subsidie voor het verduurzamen van je woning. https://www.hier.nu/je-huis-aardgasvrij/isde-subsidie-voor-het-verduurzamen-van-je-woning. Accessed: April 1, 2024.

Joseph, L. (n.d). Logistic regression diagnostics. https://www.medicine.mcgill.ca/epidemiology/joseph/courses/epib-621/logfit.pdf. Accessed: January 15, 2024.

Kalkbrenner, B. J. and Roosen, J. (2016). Citizens' willingness to participate in local renewable energy projects: The role of community and trust in germany. *Energy Research Social Science*, 13: 60–70. Energy Transitions in Europe: Emerging Challenges, Innovative Approaches, and Possible Solutions.

Klimaatakkoord (n.d.). Klimaatakkoord: Gebouwde omgeving. https://www.klimaatakkoord.nl/gebouwde-omgeving. Accessed: March 31, 2024.

Knittel, C. R. and Stolper, S. (2019). Using machine learning to target treatment: The case of household energy use. Working Paper 26531, National Bureau of Economic Research.

Knittel, C. R. and Stolper, S. (2021). Machine learning about treatment effect heterogeneity: The case of household energy use. *AEA Papers and Proceedings*, 111: 440–44.

KNMI (2021). Knmi klimaatsignaal'21. https://www.knmi.nl/kennis-en-datacentrum/achtergrond/knmi-klimaatsignaal-21.

Kwan, C. L. (2012). Influence of local environmental, social, economic and political variables on the spatial distribution of residential solar pv arrays across the united states. *Energy Policy*, 47: 332–344.

Liaw, A. and Wiener, M. (2022). randomforest: Breiman and cutler's random forests for classification and regression. R package version 4.7-1.1.

Lund, M. and Lund, A. (2018). Poisson regression using spss statistics. https://statistics.laerd.com/spss-tutorials/poisson-regression-using-spss-statistics.php. Accessed: June 5, 2023.

McCoy, D. and Kotsch, R. A. (2021). Quantifying the distributional impact of energy efficiency measures. *The Energy Journal*, 42(01).

Mehmetoglu, M. (2010). Factors influencing the willingness to behave environmentally friendly at home and holiday settings. *Scandinavian Journal of Hospitality and Tourism*, 10: 430–447.

Mills, B. and Schleich, J. (2010). What's driving energy efficient appliance label awareness and purchase propensity? *Energy Policy*, 38(2): 814–825.

Moore, D. S., Notz, W. I., and Flinger, M. A. (2013). *The Basic Practice of Statistics*, chapter 4. W. H. Freeman and Company, New York, NY, 6th edition.

Murakami, K., Shimada, H., Ushifusa, Y., and Ida, T. (2022). Heterogeneous treatment effects of nudge and rebate: causal machine learning in a field experiment on electricity conservation. *International Economic Review*, 63(4): 1779–1803.

Nguyen, Q. H., Ly, H.-B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., and Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021: 1–15.

O'Neill, E. and Weeks, M. (2018). Causal tree estimation of heterogeneous household response to time-of-use electricity pricing schemes. Cambridge Working Papers in Economics 1865, Faculty of Economics, University of Cambridge.

Penman, P. (2022). Binary logistic regression: An introduction. https://www.datascienceinstitute.net/blog/binary-logistic-regression-an-introduction. Accessed: September 2, 2023.

R Documentation (2022). stepAIC: Choose a model by aic in a stepwise algorithm. https://www.rdocumentation.org/packages/MASS/versions/7.3-60.0.1/topics/stepAIC. Accessed: April 1, 2024.

Rai, V. and Robinson, S. A. (2013). Effective information channels for reducing costs of environmentally-friendly technologies: evidence from residential pv markets. *Environmental Research Letters*, 8(1): 14 – 44.

Rijksoverheid (2019). Klimaatakkoord. https://www.klimaatakkoord.nl/documenten/publicaties/2019/06/28/national-climate-agreement-the-netherlands. Accessed: May 5, 2024.

Rode, J. and Weber, A. (2016). Does localized imitation drive technology adoption? a case study on rooftop photovoltaic systems in germany. *Journal of Environmental Economics and Management*, 78: 38–48.

Ruyssenaars, P., van der Net, L., Coenen, P., Rienstra, J., Zijlema, P., Arets, E., Baas, K., Dröge, R., Geilenkirchen, G., 't Hoen, M., Honig, E., van Huet, B., van Huis, E., Koch, W., te Molder, R., Montfoort, J., and van der Zee, T. (2022). Greenhouse gas emissions in the netherlands 1990–2020. http://hdl.handle.net/10029/625728.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2.

Sloot, D., Jans, L., and Steg, L. (2019). In it for the money, the environment, or the community? motives for being involved in community energy initiatives. *Global Environmental Change*, 57:101936.

Song, Y.-Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2): 130–135.

Souza, M. (2018). Why are rented dwellings less energy-efficient? evidence from a representative sample of the u.s. housing stock. *Energy Policy*, 118: 149–159.

Van der Schoor, T. and Scholtens, B. (2015). Power to the people: Local community initiatives and the transition to sustainable energy. *Renewable and Sustainable Energy Reviews*, 43: 666–675.

Vul, E. and Wenhao, Q. (n.d.). Chi-squared test. https://vulstats.ucsd.edu/chi-squared.html. Accessed: January 20, 2024.

Wager, S. and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242.

Wang, J. J., Zhao, X., and Li, J. J. (2013). Group buying: A strategic form of consumer collective. *Journal of Retailing*, 89(3): 338–351.

Xiao, C. and McCright, A. M. (2015). Gender differences in environmental concern: Revisiting the institutional trust hypothesis in the usa. *Environment and Behavior*, 47(1): 17–37.

Zach (2020). Assumptions of logistic regression. https://www.statology.org/assumptions-of-logistic-regression/. Accessed: September 2, 2023.

Zou, B. and Mishra, A. K. (2020). Appliance usage and choice of energy-efficient appliances: Evidence from rural chinese households. *Energy Policy*, 146: 111800.

Appendix

Extra analysis on Activities

In order to check wheter an effect from different activities in the community has an effect on taking a measure, the activities were divided in 3 categories: category 1 consist of of all kinds of meeting, category 2 consists of various forms of written communication and category 3 consists of all other activities. The different activities in the categories can be found in table 25

Table 25: Categories in the provided data

Category 1	**Category 2**	**Category 3**
Bijeenkomst	Artikel	Other
Coöperatievergadering	Brief	Actie
Buurtbijeenkomst	Buurtpagina	Overig
	Uitnodigingskaart	Startinterview
	Poster	Vragenlijst
	Poster Ledlampactie	Fotografie
	Posters LPB	Algemeen materiaal
	Persbericht	Giveaways kaart
	Verslag	Energiefiets
	Sticker	Belonen
	Introductie + Uitnodigingsbrief	Duurzame dag
	Flyer	Digitaal
	Kaart	
	Communicatieoverzicht	
	Buurtpagina online	

In total there are 486 activities is category 1, 2381 activities in category 2 and 598 activities in category 3 in the dataset.

A linear stepwise regression model in R with conversion as dependent variable and all possible neighbourhood variables(17) from the dataset including the number of activities in category 1, 2 and 3 as independent variables results in the results from table 26. The only significant variables appear to be the number of households in a community, the number of activities in category 3 and average dwelling value, but all effects are extremely small.

Table 26: regression results from the neighbourhood variables on the conversion rate

Coefficient	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	1.368×10^{-2}	1.725×10^{-2}	0.793	0.42990
n_buurt	-1.715×10^{-5}	3.673×10^{-6}	-4.668	1.24×10^{-5} ***
n_Categorie3	5.728×10^{-3}	2.024×10^{-3}	2.831	0.00591 **
gem_woz	1.167×10^{-4}	6.073×10^{-5}	1.922	0.05825 .

In table 27 the results from a stepwise approach for a linear regression of conversion on the neighbourhood characteristics are shown. This stepwise approach leaves in the end only the number of households, the community area and the average dwelling area in the community as variables in the model.

Table 27: Linear regression conversion on neighbourhood characteristics

Coefficient	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	1.964×10^{-2}	3.070×10^{-2}	0.640	0.524309
n_buurt	-1.582×10^{-5}	3.998×10^{-6}	-3.958	0.000166 ***
area	-1.003×10^{-5}	5.874×10^{-6}	-1.708	0.091578 .
gem_opp	2.922×10^{-4}	2.091×10^{-4}	1.397	0.166284

Gas and electricity use

In figure 7 the gas use is plotted against the electricity use and also for the ln of both. This shows a sort of linear increasing relationship between the two. Which is not what one would expect, since a decrease in gas usage would normally mean that more electrical appliances are used for example for cooking and heating. This could explain why the results in the results chapter are not as expected.

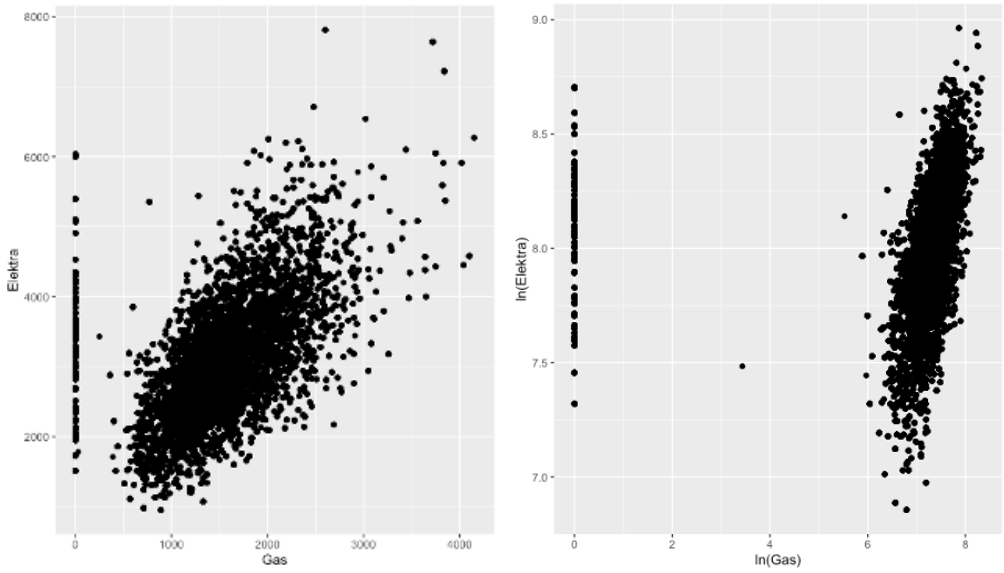


Figure 13: Gas against Electricity use

Fixed effects

The fixed effects found by the logistic regression with fixed effects can be ordered. The three most positive fixed effects are found for the communities: Kortland, Akker-Warande and Klein Brabant. The three most negative fixed effects are found for the communities: Brouwersstraat, Slotjes-West and Lindeplein. The communities are depicted in table 14. No remarkable patterns can be detected from these pictures, expect for the number of households in the communities. It also followed from a linear regression performed before that this is indeed a significant variable in predicting the fixed effects. Also the location of the communities in the Netherlands was checked, but this also did not show anything which could contribute to better predicting fixed effects.

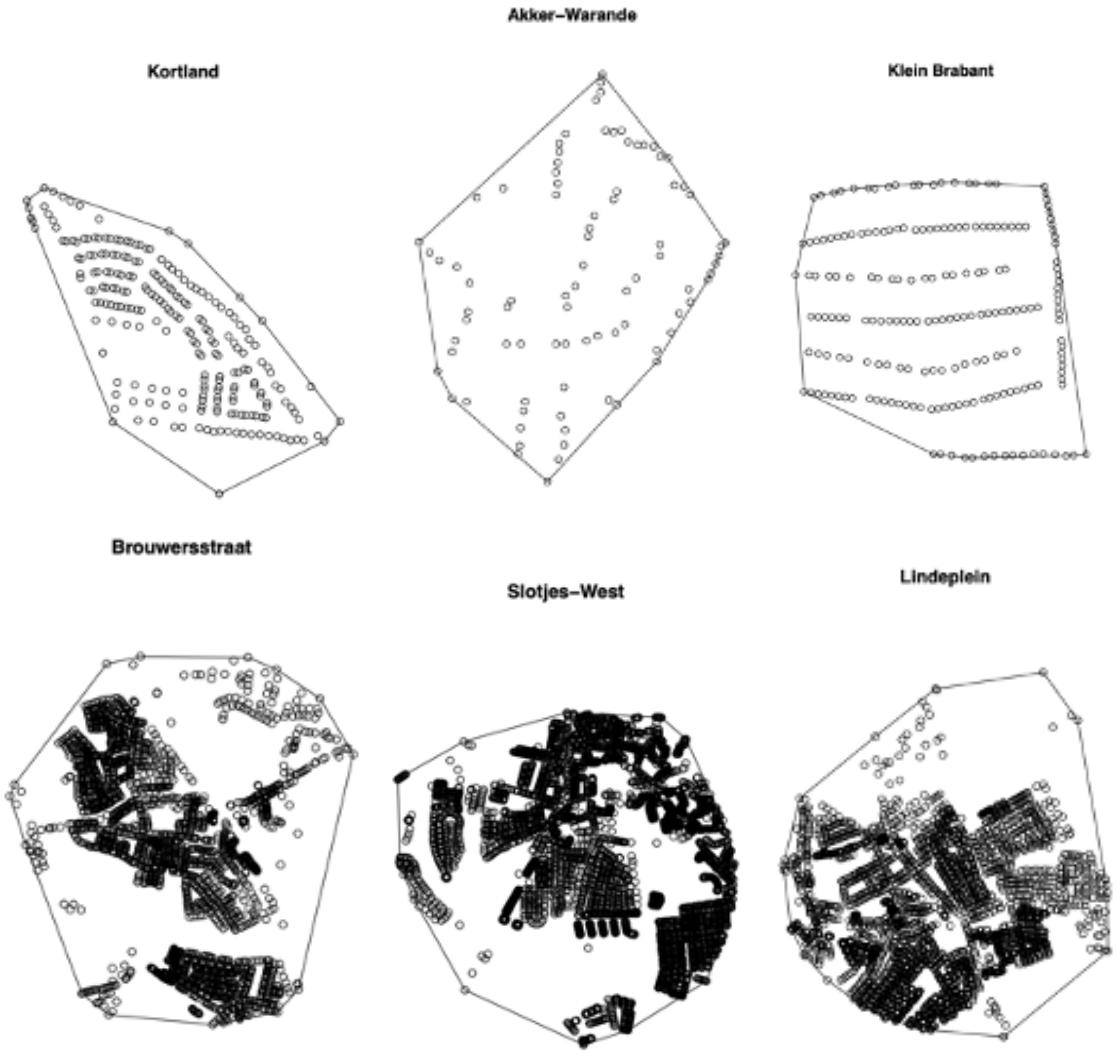


Figure 14: Communities with most positive and most negative fixed effects

Balanced modelling results

The modelling results will be improved by correcting for the class imbalance. As can be seen for example from the model performance from the random forest in chapter 7 class imbalance causes problems in the model predictions of the positive outcomes. That is the main reason why a compensation had to be found for the class imbalance. In this section an over and under sampling method is applied to the dataset and the modeling methods are applied again to the adjusted dataset.

Logistic regression

In order to perform the logistic regression again on a dataset without a class imbalance, a new dataset is created in R. This dataset is created by a combination of over and under sampling with the function ovun.sample from the ROSE package. The new dataset contains 80000 households, 40355 of these households do not apply a measure and 39645 of these households do apply a measure. The results from the same logistic regression as before, but now applied to the new dataset can be found in table 28.

Table 28: Logistic regression

	Model 1	Model 1 (old data)
ln(Gas)	-0.501 (0.056)***	-0.519 (0.155)***
ln(Electricity)	-0.693 (0.060)***	-0.635 (0.175)***
Gas dummy	3.789 (0.399)***	3.830 (1.122)***
Dist initiator 50–100	-1.085 (0.041)***	-0.882 (0.089)***
Dist initiator 100–200	-1.306 (0.038)***	-1.127 (0.081)***
Dist initiator 200–300	-1.709 (0.040)***	-1.476 (0.093)***
Dist initiator 300–400	-1.839 (0.042)***	-1.703 (0.107)***
Dist initiator 400+	-2.476 (0.038)***	-2.348 (0.092)***
ln(Living area)	1.375 (0.037)***	1.137 (0.103)***
ln(Dwelling value)	1.148 (0.056)***	0.999 (0.160)***
Construction year after 1992	-1.078 (0.034)***	-1.037 (0.106)***
Construction year before 1975	-0.308 (0.024)***	-0.349 (0.068)***
PC4 fixed effects (82)	Yes	
Num. obs.	80,000	

Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors are in parentheses.

The results only differ very little from the results of the previous logistic regression, hence over and under sampling the data with this method does not provide new insight into the effects that are researched using the logistic regression.

Random forest

The same balanced dataset that was created for the logistic regression before is used to create a random forest model. The out-of-Bag error rate from the random forest that is created is 1.93%. Which suggest the model is very accurate. The confusion matrix, which is depicted in table 29 shows us that for this data set indeed the predictions are accurate for both classes, an error of 3.8% for the predication of "No measure" and a prediction error of 0.0003% for the prediction of "Measure".
Let us look at table 30 where at every step 50 trees are added to the forest. As one can see the error rate decreases slightly by adding an extra 50 trees after the first 50 trees, but after this the positive effects of adding more trees to the random forest can be considered negligible.
In table 31 the variable importance of the included variables in the random forest is given. The variable importance is based on the mean decrease in accuracy. A higher value indicates a more significant contribution

Table 29: Confusion Matrix and Class Error - Random forest

	0	1	Class Error
Actual 0	38816	1539	0.0038
Actual 1	1	39644	0.00003

Table 30: Out-of-Bag Error Rates for Different Number of Trees - Random forest

Number of Trees	Out-of-Bag Error	No Measure Error Rate	Measure Error Rate
50	2.12%	4.17%	0.03%
100	1.99%	3.94%	0.01%
150	1.98%	3.92%	0.01%
200	1.97%	3.89%	0.02%
250	1.98%	3.91%	0.01%
300	1.98%	3.93%	0.01%
350	1.94%	3.85%	0.01%
400	1.92%	3.81%	0.00%
450	1.93%	3.82%	0.00%
500	1.93%	3.81%	0.00%

to the model's accuracy. For this new random forest the ranking of importance has changed in comparison to the random forest applied on the old dataset. But the most important and least important variables are the same. Distance to initiator is the most important and the gas dummy the least important.

Table 31: Variable Importance - Mean Decrease in Accuracy - Random forest

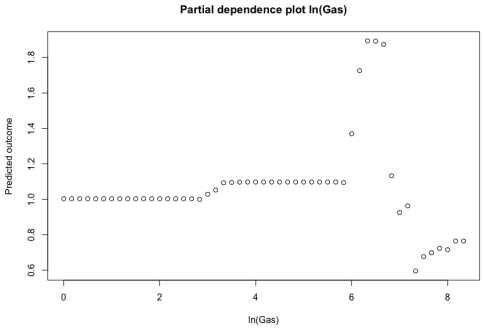
Variable	Mean Decrease Accuracy
ln(Dist. to initiator)	206.9098
ln(Dwelling value)	193.7144
ln(Gas)	176.5150
ln(Living area)	176.2378
ln(Electricity)	161.6372
Construction year	141.0464
Gas dummy	33.3562

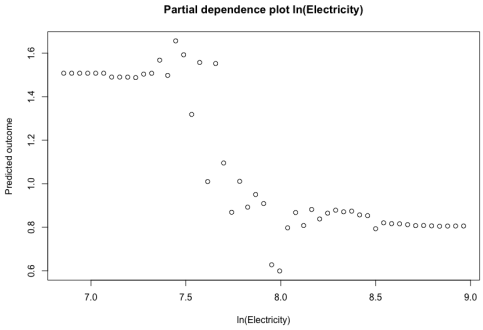
In table 32 the new partial dependence plots can be found. Comparing them to the the partial dependence plots from the previous random forest gives the following insights:

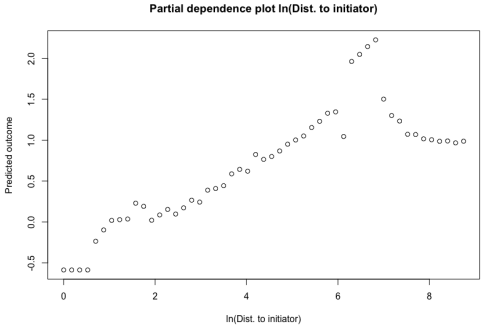
- Gas use still has a continuous effect and after a certain threshold there is an increase of the effect of gas use on the likelihood of applying a measure and then a decrease like in the previous forest. Looking at the scale from both pdp's the increase is comparable, but the decrease in the previous forest was larger.
- In the pdp from electricity for the new forest there is a pattern visible. For low energy use the likelihood of applying a measure is greater than for high energy use.
- The distance to the initiator behaves comparable to the previous forest. A larger distance to the initiator results in a larger likelihood of applying a measure.
- Area also behaves in a same manner as previously, an increase in area results in a smaller likelihood of applying a measure.
- The lower the value of the dwelling the larger the likelihood of applying a measure. This is the same as in the previous random forest

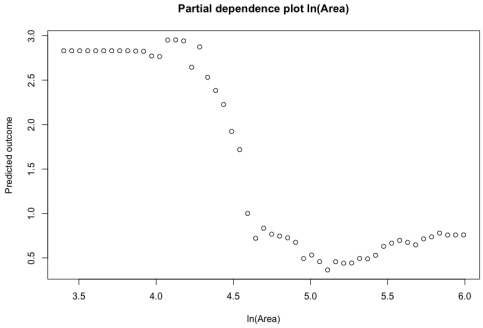
- In the construction year a shift is visible relative to the previous forest. The newer the dwelling the more likely it will be that a measure will be applied.
- The gas dummy has the opposite effect as from before. A connection to the gas results in a decreased likelihood of applying a measure.

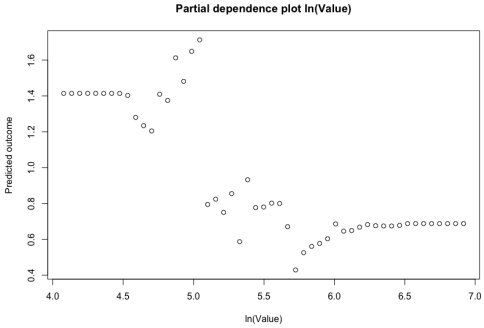
Table 32: Partial dependence plots - Random forest

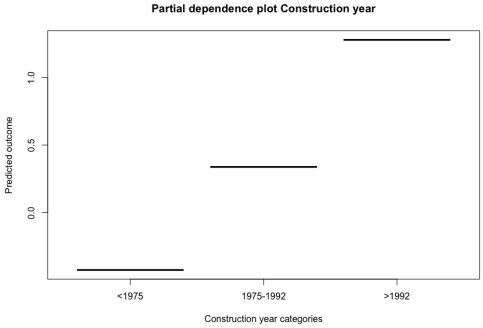


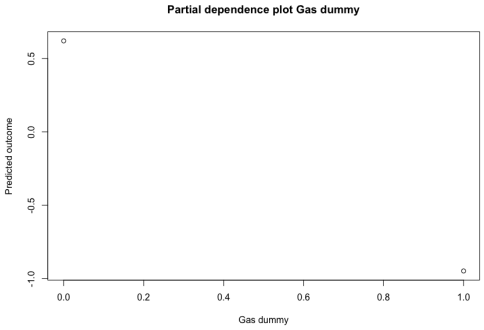












Causal forest

The same causal forest algorithm is used as before, but now the newly created balance dataset is used. The causal forest was created by dividing the dataset in a training set and a test set. The training set contained 70% of the data and the test set the other 30%. This distribution of the data over the sets is made randomly. The forest is created with 5000 trees. The variable importance from the variables in the forest is given in table 33. This importance in also depicted in figure 15.

The variable importance for the most important variable is comparable to the previous results. In the construction year categories a small shift in importance can be seen. In this new causal forest, the newer dwellings become more important in estimating the treatment effects and thus the likelihood of applying a measure when treated. This is in line with this change that was seen this the new data set in the random forest result. There the newer dwellings also became more important.

Table 33: Variable importance - Causal forest

Variable	Mean Decrease Accuracy
ln(Living area)	0.529
ln(Dwelling value)	0.189
ln(Electricity)	0.142
ln(Gas)	0.119
Construction year after 1992	0.0112
Construction year before 1975	0.00490
Construction year 1975-1992	0.00394
Gas dummy	0.000106

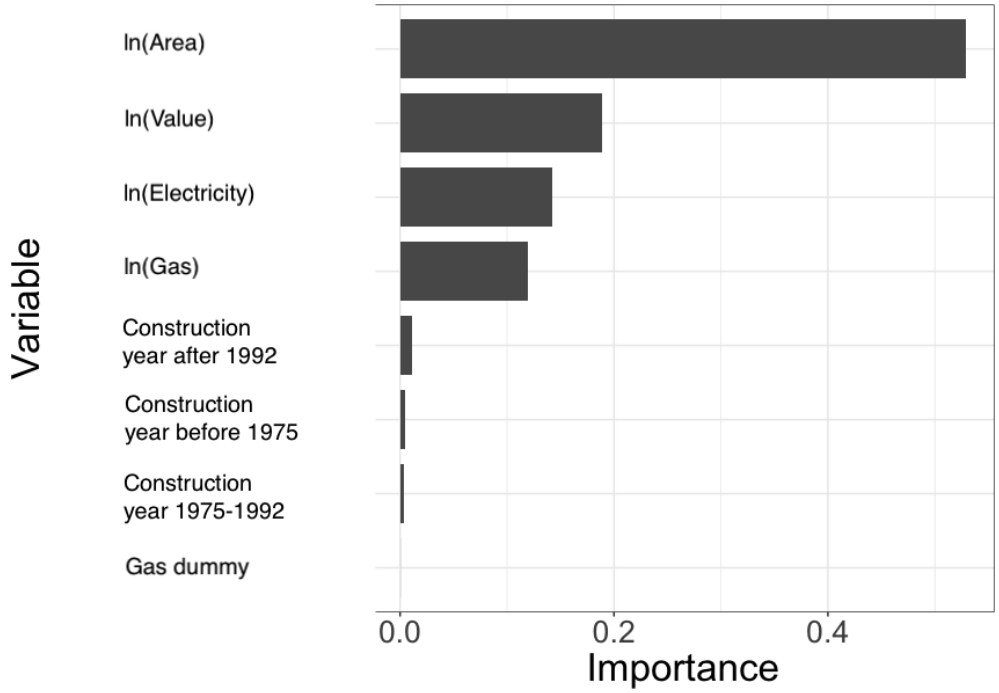


Figure 15: Variable importance - Causal forest

Next a test is performed to check whether the predicted treatment effect differs from 0. This is tested with

a one sample t-test. The results are depicted in table 34. The p-value from this test is very low, almost 0, indicating that there is strong evidence to reject the null hypothesis. The null hypothesis in this case is that the mean of the treatment effects is 0. The t-value and the confidence interval in which zero is not included support this rejection of the null hypothesis.

Table 34: One Sample t-test Results on treatment effect

Test	One Sample t-test
Data	Estimated causal effects
t-value	130.82
Df	55999
P-value	$< 2.2 \times 10^{-16}$
Alternative hypothesis	True mean is not equal to 0
95% Confidence Interval	(0.06148799, 0.06335857)
Sample mean estimate	0.06242328

The data can be divided in different groups based on the expected treatment effect. Quintile 1 is the 20% of the households with the smallest predicted treatment effects and quintile 5 is the 20% of the households with the largest predicted treatment effects. By performing linear regressions on the data from these quintiles it is aimed to predict the outcome from Measure. As one can see in table 35 the effect of an initiator being within 200 meters of the household on the variable measure increases for the higher quintiles, in the table this is seen because the slope increases. Which means that the higher the predicted treatment effect from the causal forest the higher the chance of a measure being taken by the household that has an initiator within 200 meters of them.

Table 35: Regression Coefficients for Different Quintiles

Quintile	Intercept	Slope	t-value	P-value
1	0.6237	-0.3460	76.8936	0.0000
2	0.2607	0.0168	34.4300	0.2309
3	0.3195	0.3050	37.3742	0.0000
4	0.3285	0.4747	37.5697	0.0000
5	0.2831	0.6321	32.8811	0.0000

In table 36 the mean values of the different variables are given for the different quintiles. The gas use is increasing with the predicted treatment effect. The electricity use does not show a clear pattern, neither does the gas dummy, the area and the value. The dwellings with a construction year after 1992 are mostly present in the quintiles with the smaller expected treatment effects.

Table 36: Mean variable values of the quintiles

Variable	All	0-20%	20-40%	40-60%	60-80%	80-100%
ln(Gas)	6.993	6.692	6.776	7.079	7.280	7.138
ln(Electricity)	8.004	7.959	7.888	8.033	8.062	8.080
Gas dummy	0.957	0.924	0.947	0.966	0.987	0.963
ln(Living area)	4.891	4.845	4.661	4.886	4.995	5.071
ln(Dwelling value)	5.416	5.359	5.229	5.428	5.504	5.560
Construction year after 1992	0.220	0.267	0.238	0.227	0.198	0.172
Construction year before 1975	0.445	0.403	0.448	0.460	0.472	0.441
Construction year 1975-1992	0.335	0.330	0.314	0.313	0.330	0.387