ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

"Evaluating drivers of the RET in Dutch municipalities: Predicting annual population-adjusted non-renewable energy consumption using consumer sentiment and characteristics, and RE investment data in the residential sector"

Name student: Kevin de Man

Student ID number: 485356km

Supervisor: Radek Karpienko

Second assessor: M. van de Velden

Date final version: 01/11/2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This study investigates factors influencing per capita non-renewable energy (PC NRE) consumption in the Netherlands' residential sector at the municipal level, providing insights into how individuals in households contribute to the RET. Consumer sentiment and characteristics, and renewable energy (RE) investments influenced by monetary policies were analysed using various models, including linear regression, random forest, gradient boosting, and extreme gradient boosting (XGB). Consumer sentiment was extracted from news publications associated with RE technologies through the Meltwater media monitoring platform, where sentiments were assigned to these articles The XGB model demonstrated superior predictive performance, revealing interesting insights. Among the 23 variables examined, 11 were identified as significant drivers of PC NRE consumption, including 2 sentiment-related and 3 investment-related variables. Furthermore, PC NRE consumption proved to be sensitive to sentiment, with publications with a positive sentiment resulting in lower consumption, while negative publicity increased consumption. Notably, municipalities with positive sentiment and high subsidized RE investment capacity showed reduced consumption, indicating the effectiveness of targeted monetary policies. Additionally, municipalities with the highest energy consumption exhibited high-income levels and property values, but their investments in renewable energy fell below the average. The study highlights the high potential these wealthy regions have for reducing PC NRE consumption by enhancing investments in RE technologies. To encourage their active participation in renewable energy initiatives, the study recommends exploring alternative, non-monetary strategies.

Table of contents

Introduction

Every year, the consequences of global climate change become more visible. Across the globe, occurrences of wildfires, floods, and droughts are increasing in both frequency and intensity. This is a result of extreme weather patterns driven by rising greenhouse gas (GHG) emissions from human activities. These events not only affect biodiversity but also carry significant economic implications (IPCC, 2021). The energy sector accounted for 73.2% of global GHG emissions in 2016 (Ritchie, 2020) with the built environment being one of the primary consumers of energy (Kabeyi & Olanrewaju, 2022). In particular, the residential segment, was responsible for 29.7% of global natural gas and 26.6% of total electricity consumption in 2019 (International Energy Agency, n.d. - a). Moreover, research by Niamir, Ivanova, Filatova, Voinov and Bressers (2020) indicates that households contribute to at least 70% of CO2 emissions, highlighting the importance of addressing energy consumption in the residential sector to combat GHG emissions. Consequently, this research focuses on energy consumption and its determinants in the crucial residential sector, aiming to comprehend the evolving transition to greener energy, known as the renewable energy transition (RET). While in other sectors energy consumption might be more commercially oriented, energy consumption within the residential sector can capture the behaviour of ordinary energy consumers. This study achieves this by modelling the effect of consumer sentiment, characteristics and renewable energy (RE) investment driven by monetary policies on per capita non-renewable energy (PC NRE) consumption.

The renewable energy transition (RET) is essential in achieving the climate objectives outlined in the 2050 Paris Agreement (United Nations Framework Convention on Climate Change, 2015) aimed at reducing GHG emissions. Regrettably, the global trajectory towards meeting these goals is inadequate, with certain countries exhibiting exceptionally high per capita emissions due to their significant fossil fuel production (Kharas, Fengler, Sheoraj, Vashold & Yankov, 2022). Acknowledging responsibility is crucial, yet achieving universal compliance in our complex world is challenging. Some may perceive that their actions have minimal impact on global emissions or prioritize other concerns. Additionally, this transition demands time and substantial financial resources, a level of flexibility not attainable for everyone. McKinsey (Pacthod et al., 2022) outlined a region-specific plan for the energy transition, making a clear distinction between the key stakeholders: individuals, governments and multilateral organisations, companies and financial institutions, able to shape the transition. This study centres on both individuals and the government as key stakeholders. Its objective is to understand the dynamics among the driving factors and provide actionable insights on how both individuals and governmental organisations can take responsibility for reducing PC NRE consumption.

While governments worldwide have implemented various market mechanisms, such as RE investment subsidies, and other monetary policies, the most effective strategy for accelerating the RET remains uncertain. Therefore, the following research question (RQ) emerges:

"What are the drivers of the Dutch energy transition in the residential sector and how can the government and individuals as stakeholders, play a role in this?"

The main RQ needs to be divided into more sub-questions (SQs) to be more specific in the research and cover as many relevant areas as possible. Therefore, the RQ is subdivided into the following questions:

Sub-question 1 - What drives the per capita non-renewable energy consumption?

Sub-question 2 - How do sentiment, RE investments and consumer characteristics affect the per capita non-renewable energy consumption?

Sub-question 3 - How do these factors interact with each other?

Sub-question 4 - What regions have the most potential for improvement?

To address the RQ and its SQs, the paper is structured into distinct sections. Initially, an extensive literature review is conducted within the field of RE to pinpoint the drivers of RET and energy consumption. This analysis occurs within the *Theoretical Framework* section. Once the drivers are identified, the *Data* section elaborates on the information used in the study, detailing the dataset containing these drivers. Subsequently, the *Methodology* section outlines the techniques and methods applied to the data, providing technical details. Using data science and machine learning techniques, factors that are crucial in driving or predicting PC NRE consumption in residential areas are explored. In the Results section, the analysis outcomes are presented without interpretations, whereas the Discussion section interprets and explains these results, addressing the study's SQs and main RQ. The study's final recommendations and limitations are addressed in the *Conclusion* section.

Theoretical framework

The renewable energy transition (RET)

The energy sector - counting direct and indirect emissions - remains the primary source of global greenhouse gas (GHG) emissions responsible for approximately two-thirds of all emissions. This can be attributed to fossil fuels, which accounted for 80% of the total global primary energy consumption in 2018 (Lu et al., 2020). Global energy demand exhibited an upward trend for almost every year for over half a century, with minor exceptions for the early 1980's, and the crisis periods in 2009 (financial crisis) and 2020 (COVID-19 pandemic). Although the growth in global energy demand is stagnating, it continues to grow at around 1 - 2% each year. This rise in energy demand is driven by population growth and the overall wealth increase of people (Ritchie, Roser & Rosado, 2022). Among all forms of energy, electricity stands as the primary energy source. Given its potential to decarbonize energy consumption and the electrification of various sectors worldwide, the electricity demand is expected to increase even further. Yet electricity is mainly generated by fossil fuels in the form of coal, natural gas and oil (61.3% in 2020) (Kabeyi & Olanrewaju, 2022). This rise in overall energy demand and more specific energy demand, such as electricity, shows the crucial role of transitioning from fossil fuel-based power to environmentally sustainable or renewable energy (RE) sources to tackle global GHG emissions effectively. Throughout history, there have been two major energy shifts: the shift from wood to coal and later from coal to fossil fuels. Currently, the world is actively transitioning from fossil fuels to renewable energy (RE) sources (Kabeyi & Olanrewaju, 2022), widely referred to as the green or renewable energy transition (RET).

Since the RET is a broad subject, there needs to be a measure that captures the essence of this term. However, monitoring and keeping track of the energy transition is another matter since it is difficult to quantify the energy transition. A substantial amount of research about enhancing the green energy transition, in general, creates a framework towards transitioning to cleaner energy (e.g. Kabeyi & Olanrewaju, 2022; Cantarero, 2020), or points out the necessity to monitor the energy transition in a context-specific way (e.g. Bisaga, et al., 2020; Abdullah, 2013). Some examples of quantifiable measures used in papers are the share of RE in final energy consumption, the number of green jobs and the energy use and emissions per capita (Cantarero, 2020). Some more specific examples are the number of green energy policies/initiatives (Bayulgen, 2020) or solar installations (Bennett, Baker, Johncox & Nateghi, 2020). Furthermore, using data science or machine learning models in combination with environmental-related issues is also common. Kannangara, Dua, Ahmadi and Bensebaa (2018) use decision trees and neural networks (NNs) to model the effect of socioeconomic and demographic factors on municipal solid waste generation and diversion in Canada at the municipal and regional level. They found that the NN generated the best predictive performance.

The RET can be split up into three more specific categories, following the structure of Kabeyi and Olanrewaju (2022). The RET typically consists of technological changes within energy demand or end-use sectors (energy saving), energy production (efficiency in energy generation) or fossil fuels substitutions (RE sources/ low carbon nuclear). Moreover, GHG emissions associated with energy are not solely derived from a single sector but rather encompass multiple industries. These include the conversion and delivery sectors of energy, as well as end-use sectors such as households and buildings.

The study will primarily concentrate on the end-use of the residential sector, known as one of the most energy-intensive sectors.

Effects of psychological or behavioural factors on the RET

Niamir, et al. (2020) studied how behavioural factors such as awareness, and personal and social norms, besides various sociodemographic and structural factors affect 3 energy-related actions in the residential sector. The researchers focus on a combination of bottom-up drivers affecting the transition to a lower carbon footprint. They study how behavioural, socioeconomic and structural factors affect energy-related actions among households in the Netherlands and Spain. This is assessed by using 3 measures: the probability of households' investing (in either insulation, solar panels or energy-efficient appliances), energy conservation and switching of energy providers. They observed that a higher level of personal norm about environmental issues significantly leads to a higher probability of investing in green energy solutions, conserving energy and a household switching to a green energy supplier. This personal norm was in turn strongly correlated with knowledge and awareness. According to their literature study, social and personal norms also seemed to positively influence other energy-related actions beyond the scope of their research, such as recycling or fuel conservation.

Another form of energy-related action is green energy adoption. Wall, Khalid, Urbański and Kot (2021) surveyed five crucial cities in Thailand to investigate the impact of behavioural factors on consumers' intention to adopt renewable energies in 5 vital cities of Thailand. Respondents were asked to provide their responses using a Likert scale. Their findings indicate that the behavioural variables perception of self-effectiveness, environmental concern, RE awareness, and beliefs about RE benefits have a significant and positive effect on the willingness to adopt. Furthermore, they propose that these beliefs are shaped by consumers' knowledge about RE, which can, in turn, be influenced by their awareness. In turn, this awareness can be increased through publicity or word of mouth. Yang, Zhang and Zhao (2016) arrived at comparable conclusions when investigating the influence of psychological and socio-demographic factors in shaping both residents' direct and indirect energy-saving behaviours. Their research centres around energy behaviour, with actions such as turning off the light, without the quantifiable effect. Their findings indicate that individuals with a strong sense of environmental responsibility and curtailment attitudes are more likely to participate in energy-saving actions, both directly or indirectly. To clarify, direct energy use encompasses immediate energy consumption from sources like gas, electricity, and water, while indirect energy use is integrated with the products and services individuals consume. Interestingly, despite the greater influence of psychological and socio-demographic factors on indirect energy-saving behaviours, people tended to engage more frequently in direct energy-saving actions. Moreover, the paper of Arkesteijn and Oerlemans (2005) combines different types of variables to study green energy adoption. They conducted a study on the early adoption of green energy among Dutch households, employing a logistic regression to classify households as either adopters (yes) or non-adopters (no). A combination of variables that incorporated theoretical insights from various disciplines in adoption studies increased the statistical significance of the model. This involves a model with a combination of variables related to the technical system (e.g. such as level of trust in green energy suppliers that they provide green energy or the perception of ease of switching/use), individuals (e.g. attitude towards the environment or knowledge about RE sources) or variables derived from economic theories (e.g. disposable income or willingness to pay). Due to greater significance, the combination was the most powerful model to predict green energy adoption. They find that individuals who perceive a sense of responsibility towards the environment and display a higher willingness to pay are more likely to adopt green energy. Additionally, higher prior understanding of respondents and historically shown environmentally friendly behaviour had a higher chance of adopting.

The previous paragraphs demonstrated how psychological factors impacted energy-related actions. The direct effect of psychological factors on energy consumption is studied by Abrahamse and Steg (2011). They consider psychological factors besides socio-demographic factors or building attributes when measuring household-level energy consumption and willingness to conserve energy. They found that attitude towards energy conservation played a significant negative role in energy usage and had a positive effect on the intention to reduce household energy use. Furthermore, regarding intentions to reduce energy use or energy conservation, Chen, Xu and Day (2017) find that in low-income households thermal energy conservation was positively affected by behavioural factors such as personal attitude. Additionally, Frederiks, Stenner and Hobman (2015b) study psychological factors along with sociodemographic factors on residential energy consumption and conservation in a comprehensive literature review. Their research revealed that normative social influence, such as observing energy-related practices of peers or neighbours and experiencing social pressure from family and friends to conserve energy, had a significant impact. Individuals tend to follow and compare with the behaviour of the people around them, aligning with group or societal guidelines and behavioural expectations of what is considered normal. A disparity remains between what people express and the eventual energy-related outcome in terms of energy consumption and conservation. This observation is consistent with another publication by the same authors (Frederiks et al., 2015a), in which they apply behavioural economics to comprehend behavioural biases in the complex energy-related decision-making and behaviour of consumers in household energy use. They state that a difference exists between individuals' underlying knowledge, values, beliefs, attitudes or intentions and their actual energy consumption or conservation. They emphasize the necessity to gain insights into these phenomena and narrow the gap between an individual’s core values, and energy-related consumer behaviour. By doing so, policymakers' public interventions can achieve greater effectiveness, and promote renewable and sustainable energy utilization among energy consumers.

Effects of governmental policy interventions on the RET

Different government-driven market mechanisms or policy interventions are essential to address private energy consumption. Effective policies can help in bridging the gap between what people express and the eventual energy outcome as came forward at the end of the previous section. Frederiks et al. (2015b) emphasize the need for energy-saving initiatives to align with people's values, beliefs, and attitudes, translating them into concrete energy consumption, conservation or pro-environmental behaviours. This alignment ensures optimal return on investment and cost-effectiveness. Yang, Zhang and Zhao (2016) suggest that policy interventions, accomplished through strengthening publicity and educational initiatives, to elevate environmental awareness, effectively contribute to reduced household energy consumption. Moreover, Niamir et al. (2020) suggested a strategic approach involving targeted policies, along with widespread social advertising and educational initiatives, to enhance knowledge, awareness, individual norms or sentiment with RE usage among the broader public. This approach could complement and enhance the efficiency of other incentivizing mechanisms, such as subsidies, a form of monetary policy.

Monetary policies are another way government policy can positively affect private energy consumption by encouraging investments. Lu, Khan, Alvarez-Alvarado, Zhang, Huang, and Imran (2020) assert that addressing concerns related to risk and trust in RE adoption typically requires policies regulating reward and penalty values for products or services, e.g., with a feed-in-tariff (FiT) or carbon tax. Nicolini and Tavoni (2017) investigated how monetary incentives promote RE in the five largest European countries. Their study revealed a positive relationship between subsidies and both incentivized energy production and invested installed capacity in the short and long-term. They also observed that the amount and average tariffs of these incentives were positively related to RE production. Qadir, Al-Motairi, Tahir and Al-Fagih (2021) explore specific approaches aimed at funding the transition towards renewable energy for governments, corporations, and individuals in the general public. As a result of insufficient awareness regarding the advantages of RE and misinterpretations of the associated installation and operational expenses, individuals and households tend to avoid investments in RE. Monetary policy design by the government plays a crucial role in the RET since financing investments in RE has been a major issue. This design can be categorized into a "demand-pull" approach involving incentives like subsidies, and a "supply-push" strategy to enhance the business environment, such as providing support for research and development (R&D). In certain countries, the continued subsidization of fossil fuels is hindering the advancement of the RET, highlighting the necessity to enhance monetary policy designs and allocate greater resources to energy-related subsidies.

More direct effects of monetary policy on the RET are examined by Ouyang & Lin (2014), studying the impacts of renewable energy subsidies on China's economy, emissions and the stability of energy distribution and consumption. Their study revealed that an increase in subsidies resulted in a decrease in energy consumption per unit of GDP, a decrease in CO2 emissions and other favourable macroeconomic outcomes. Moreover, increasing subsidies for renewable energy seems to help China with the problem of imbalanced energy distribution and consumption. Furthermore, Xue, Gong, Zhao, Ji and Xu (2019) conducted a study to examine the impact of government subsidies aimed at promoting energy-efficient products for manufacturers. They assessed various factors, under which the level of energy conservation and energy saving of products. The study revealed that government subsidies had a significant positive effect on the enhancement of energy-saving products. Additionally, the research found a positive correlation between government subsidies and energy conservation.

Furthermore, well-targeted policies can make some indirect effects more tangible. Yang et al. (2016) highlight that policy interventions leading to reduced energy consumption are most effective when tailored to specific behaviours and demographics. Yang & Zhao (2015) study the RET in China by measuring the energy efficient and renewable energy equipment (EERE) purchases and base this on sociodemographic factors. Family income and better familiarization or awareness with subsidy incentives positively affect the transformation of a green attitude to a purchasing intention of the EERE equipment. Poruschi and Ambrey (2019) conducted a study in Australia analyzing data from postal codes over 14 years. Their research centred on the impact of feed-in tariffs (FiT), a form of subsidy, along with income levels and city density on the adoption of solar panels in the built environment. Their findings revealed several key insights. Firstly, they found a positive relation between solar installations and income. However, they noted that the benefits of these installations seem to be concentrated in higher-income areas, potentially increasing energy inequality if poor policy design neglects income differences. Additionally, FiT subsidies positively influenced the adoption of solar PV panels. Moreover, the study found that denser urban areas exhibited fewer short and long-term solar PV installations, possibly due to reduced light and space availability. Interestingly, the implementation of FiTs appeared to more than offset this effect, emphasizing the significance of built environment subsidies compared to the relatively modest impact of city density on solar panel adoption. Poruschi and Ambrey drew multiple implications from their results. Poorly designed government policies may lead to unnecessary subsidy expenditure. Therefore, the implementation of fair and well-considered subsidies is crucial, drawing lessons from past FiT policies while accounting for local conditions and potential socio-demographic interactions or interactions with other variables. Subsidies could prove beneficial for supporting multi-household investments in apartments, simplifying ownership structures and addressing split incentives. The researchers recommend further investigation into policy outcomes for specific socio-demographic groups, such as disadvantaged categories (renters, low-income individuals, and apartment dwellers). This information will be used specifically to address variables that can capture subsidies or loans to stimulate RE investments.

Effects of consumer characteristics on the RET

Shifting from a consumer and policy-oriented perspective to examining consumers' and their residence's characteristics, Niamir et al. (2020) model and quantify 3 energy-related actions as a measure of the RET as mentioned in section 2.2 based on socioeconomic and building characteristics, such as the size or type of the house. The economic comfort of residents was found to have a positive and significant effect on a household's willingness to invest in green energy solutions. Owners of a house instead of house renters were found to be significantly more willing to invest in green energy solutions. The age of the house was also significant, having a positive effect on the likelihood of investment in isolation but a negative effect on the likelihood of installing solar panels. To households' energy conservation, the economic comfort of the resident (lower economic comfort) and the energy label (energy efficiency label) of the residence seemed to have a positive effect. Switching to a green energy provider is positively influenced by the age of the residence and negatively influenced by income and energy rating. Overall, economic comfort/income seemed to be the most important sociodemographic driver over all three measures while several structural attributes were contributing most to the three energy-related actions. Lastly, to capture the influence of institutional, cultural, and climatic factors, the researchers used a country dummy variable as a proxy. They find that the effects of the 3 measures differed between the countries emphasizing the significance of considering spatial and geographical information when analysing energy-related behaviour. Bennett, Baker, Johncox and Nateghi (2020) use the number of solar installations in households as a measure of green energy adoption at the zip code level in the US. They explore how social, economic, environmental (solar radiation) and other factors influence the number of solar installations per capita. The findings reveal that factors such as average electricity consumption, income and solar radiation have a positive effect on the population-adjusted adoption rate. People exhibiting higher electricity consumption seemed to be more keen to invest in solar installations. They use multiple machine learning algorithms to capture the relationship between the variables but find that the extreme gradient boosting (XGB) algorithm outperforms the other algorithms followed by random forest (RF).

Extensive research has been conducted in the field of energy consumption, proposing various frameworks to model energy consumption within the residential sector. One of the fundamental papers that identify the key differences among various models concerning energy consumption in the residential sector, used in multiple papers after publication, was written by Swan and Ugursal (2009). Their study conducted an extensive examination of the current body of literature. They make the distinction between bottom-up and top-down approaches, both relying on different types of information or calculations still functioning as the fundamental structure in energy consumption studies. Top-down methods utilize aggregated historical data, often dependent on macroeconomic factors (e.g. GDP or employment rates), energy price or the general climate. The bottom-up approach concentrates on the energy consumption of individual end-uses, single houses or a group of houses. Any model that uses data inputs at a lower-level hierarchy than the sector as a whole is considered to be part of the bottom-up approach. Detailed variables such as building characteristics can be incorporated into this approach, however, this approach also has the advantage of being able to use macroeconomic, energy prices, income data and other regional or national indicators from a sample of houses consequently integrating the strengths of the top-down approach. Furthermore, they distinguish between the engineering method (EM) which primarily emphasizes the physical aspect of a building to assess energy performance at the building level and the statistical method (SM) which relies more on historical data to model energy usage through influencing variables. Similarly, in their comprehensive literature review regarding the prediction of building energy consumption, Zhao & Magoulès (2012) categorize EM and SM alongside machine learning methods capable of capturing nonlinear relationships, such as neural networks (NNs), support vector machines (SVMs) or decision trees.

This fundamental structure was also adopted by Wiesmann, Azevedo, Ferrão and Fernández (2011) who modelled the per capita electricity consumption at the municipality (top-down) and the individual household level (bottom-up) using an ordinary least squares (OLS) regression. The models at both scales contain socioeconomic variables and building attribute variables such as income, (average) persons per household and building age. Additionally, at the household level, variables such as dwelling type, occupancy type and urbanization level are considered. However, data availability is limited on the aggregated level. To address this, the researchers use proxies of different levels of urbanisation, dwelling type and dwelling floor area. Climate and regional effects were also accounted for at the municipality level. The findings from both scales align with each other, showing similar significance and signs for the independent variables, including for 2 out of the 3 proxies. Nevertheless, the municipality scale model exhibited a superior goodness of fit (R-squared). The study reveals that an increase in income or a decrease in household members resulted in higher per capita electricity consumption. Moreover, per capita consumption of single-family homes is greater than that of multi-family homes or apartments. At last, urban households consumed more electricity per capita than rural households.

This aligns with the research of Wiedenhofer, Lenzen and Steinberger (2013) who study the direct and indirect per capita energy consumption for the average households in Australia. Their findings revealed that regions with greater urbanization and wealth exhibited the highest per capita consumption (indirect and total). Moreover, suburban areas are more reliant on energy-intensive car transportation, compared to urban areas where public transport is more prevalent. However, due to the multitude of options for consuming goods and services, urban areas still exhibited higher consumption rates. Similarly, the study by Abrahamse and Steg (2011) delved into the impact of sociodemographic factors on household-level energy consumption and individuals' intentions to reduce their energy usage. They found that variables such as income, household size and age (of the household member that filled in the questionnaire) most significantly increased energy usage. Moreover, Frederiks et al. (2015b) conducted a comprehensive literature review concerning the sociodemographic and building characteristics impact on energy consumption and conservation within the residential sector. The key findings of the researchers indicate that among the most influential sociodemographic factors, there exists a positive relationship between household income and both energy consumption and energy-efficient investments. Similarly, larger buildings or detached dwellings (single-family homes) are positively associated with energy consumption. Additionally, homeownership, in contrast to renting, is positively linked to both energy consumption and capital investments. On the contrary, family size exhibits a negative impact on per capita energy consumption. Nonetheless, the effects are not straightforward and should be nuanced due to the complex nature of the variables. The effects appear to be domain-specific and depend on the context with potential interactions among variables. An example is the income variable. While higher income could potentially positively influence energy-conserving behaviour through the financial flexibility to make increased investments in energy efficiency, this influence is offset by the observation also put forth by Goldstein, Gounaridis and Newell (2020) in terms of CO2 emissions. According to these researchers, households in the United States with a higher income exhibit 25% higher per capita CO2 emission compared to those with lower income, primarily caused by the ownership of larger properties. Lastly, Frederiks et al. (2015b) propose the development of customer profiles capable of encompassing a blend of sociodemographic and psychological variables.

Gassar, Yun and Kim (2019) employ various machine learning models to predict the gas and electricity consumption in residential areas of London based on a combination of sociodemographic factors, building characteristics, and economic variables at the district level. The models they use encompass multiple regression, a multilayer neural network (MNN), gradient boosting (GB) and an RF. Their analysis reveals that residential electricity and gas consumption are positively influenced by household income, population density, and median house prices. Additionally, various building characteristics, including the average number of rooms per house, the number of buildings, available household spaces, and land area, all exhibit a significant positive relationship with energy consumption. In addition to identifying the most influential variables, they observe that the MNN model outperforms the other models in terms of predictive accuracy.

Summary and expected relationships

In the preceding literature review, various variables indicative of the energy transition are identified and their determinants discussed, with a focus on the residential sector.

Understanding these determinants and their relationships with the energy transition is crucial when addressing the RQ. Various determinants, such as psychological variables, subsidy-related variables, sociodemographic aspects and building characteristics play a pivotal role in the RET. Stakeholders have the potential to exert influence on certain determinants, providing valuable insights for this research. Primarily, psychological factors impact individuals' energy-related decisions, encouraging actions such as investing in energy-efficient products (Niamir et al, 2020), embracing energy conservation practices or adopting green energy (Wall et al, 2021; Arkesteijn & Oerlemans, 2005). In turn, this can reduce (non-renewable) energy consumption. Government policies can promote more knowledge and awareness about the RET (Niamir et al., 2020; Yang et al., 2016), which impacts these psychological factors (Niamir et al., 2020; Wall et al., 2021). Additionally, government-implemented monetary policies, such as subsidies related to the RET, directly decrease energy consumption (Ouyang & Lin, 2014) and indirectly influence consumption through energy-related actions like solar PV adoption (Poruschi & Ambrey, 2019) and the use of energy-saving products (Xue et al., 2019). While sociodemographic aspects and building characteristics are more challenging for stakeholders to influence directly, specific combinations of these factors could be strategically targeted, as suggested by Poruschi and Ambrey (2019) and Yang, Zhang and Zhao (2016). Research by Arkesteijn and Oerlemans (2005) has demonstrated that combining determinants from different disciplines enhances the statistical significance of models. Consequently, this research aims to investigate a combination of all the determinant types outlined in the theoretical framework. Table 1 provides a clear overview of the expected relationships between the different types of determinants and a metric representing the RET and subsequently what the expected outcome of these determinants would be on energy consumption.

Table 1

Expected relationships of psychological, policy indicators and sociodemographic and building characteristics on the corresponding RET-related variables and the eventual expected relationship with energy consumption. Relationship expressed with either a positive (+) or negative (-) sign.

Variable	**Expected relationship with the corresponding variable in literature**	**Resulting relationship with energy consumption**	**Variable type**
Attitude (sentiment)	Energy consumption: - (Abrahamse & Steg, 2011); intention to conserve energy: + (Abrahamse & Steg, 2011 and Chen, et al. 2017)	A better attitude towards RE will decrease energy consumption	Psychological
Normative social influence	Energy conservation: + (Frederiks et al., 2015b)	More normative social influence will decrease energy consumption	
Personal norms	All 3 energy-related actions: + and strongly correlated with knowledge and awareness (Niamir et al., 2020)	A higher level of personal norm will decrease energy consumption	
Awareness (number of publications)	"There is a knowledge-action gap", so no significant impact on energy-related behaviour (Frederiks et al., 2015b); the purchasing intention of EERE equipment: + (Yang & Zhao, 2015)	More awareness regarding RE will decrease energy consumption	
Renewable energy subsidy	Energy consumption & CO2 emissions: - (Ouyang & Lin, 2014), enhancement of energy-saving products: + (Xue et al., 2019), invested installed capacity: + (Nicolini & Tavoni, 2017)	Higher energy subsidies will decrease energy consumption	Monetary policy
Income/wealth	Population-adjusted solar installation adoption: + (Bennett et al., 2020); likelihood of switching energy provider: - (Niamir et al., 2020); per capita electricity consumption: + (Wiesmann et al. 2011); energy consumption: + (Wiedenhofer et al. , 2013); energy consumption: + (Abrahamse & Steg, 2011); energy consumption & energy efficient investment: + (Frederiks, Stenner & Hobman, 2015b); purchasing intention of EERE equipment: + (Yang & Zhao, 2015)	A higher income will increase consumption	Socio-demographic

Residential arrangement (homeownership/ renting or house/apartment)	for homeowners -> likelihood of the willingness to invest in solar panels and insulation: + (Niamir, et al., 2020); for single-family homes -> per capita electricity consumption: + (Wiesmann et al. 2011); energy consumption & energy investments in energy-efficient equipment: + (Frederiks et al., 2015b)	More homeowners and single-family homes will increase energy consumption	
Urbanisation/household density	per capita electricity consumption: + (Wiesmann et al. 2011) & energy consumption: + (Wiedenhofer et al., 2013)	Higher density will increase energy consumption	
Number of household members	Per capita electricity consumption: - (Wiesmann et al., 2011); Energy consumption: - (Frederiks et al., 2015b)	Higher average household size will decrease (per capita) energy consumption	
Age of a building	Likelihood of household investment in insulation: +; Likelihood of household investment in solar panels: – ; switching energy provider: + (Niamir et al., 2020)	Older buildings will increase energy consumption	Building characteristics
Size of the building	Likelihood of the willingness to invest in PV and insulation: + (Niamir, et al. , 2020); per capita electricity consumption: + (Wiesmann et al. 2011); energy consumption: + (Abrahamse & Steg, 2011); energy consumption: + (Frederiks et al., 2015b)	Bigger homes will increase energy consumption	
Energy label	Better energy labels resulted in more conservation (Niamir et al., 2020)	Better energy labels will decrease energy consumption	
House price	Electricity and gas consumption: + (Gassar, Yun & Kim, 2019)	Higher house prices will increase energy consumption	

Data

Introduction

Several types of variables were identified in the theoretical framework that could potentially influence the energy transition, with a clear distinction between them. The objective is to analyse these variables to assess their impact on the energy transition, particularly energy consumption, which is the variable of interest.

Data is collected for the residential sector in the Netherlands (NL) - aggregated on the municipality level. A municipality refers to a geographic area within the Netherlands that operates within the second level of governance. This dataset contains energy consumption data, besides independent variables containing information about news publications (psychological variables), RE investments made possible through monetary policy, sociodemographic and building characteristics. This results in a total number of features of 23. Furthermore, the dataset is aggregated at the municipal level, covering 342 municipalities for each year over a span of 6 years, from January 1, 2015, to December 31, 2020, resulting in a total of 2052 data points.

In the upcoming section, there will be an explanation regarding the essential data and required to address the research question and sub-questions. This section will detail the data cleaning and processing procedures as well as the methods for accessing this data.

Data collection, description and pre-processing

Psychological data – Meltwater

Meltwater software

To obtain psychological data for the municipalities in NL, the commercial media monitoring platform *Meltwater* is used. Meltwater uses natural language processing algorithms to analyse media data about a certain subject, person, event, etc. and determine the sentiment of each publication, categorizing them as positive, neutral, or negative (Whitney W., n.d. - a). By using a combination of user-defined keywords, the Meltwater software retrieves all publications that correspond with that set of keywords sourcing from over 270,000 online news, 15 social media and other media sources, globally (*Media Monitoring & Analysis | Meltwater*, n.d.).

For every publication, information is given about, for instance, the source, reach, sentiment, date and region of all of these publications. The reach metric quantifies the potential viewership of a source and is derived from the monthly count of unique page visitors for the specific source, which is made available by *Meltwater*'s partnership with **Similarweb** (Whitney W., n.d. - b). Reach consists of the desktop reach and mobile reach referring to the monthly unique page visitors that used either the desktop or mobile. Furthermore, the region from where the news is published is provided by the source of the publication and therefore the data can be analysed on the municipality level.

This research is not focused on the specific publications, but rather focuses on the accumulated number of publications and their corresponding sentiment and the potential reach linked to a set of RE-related keywords. This information about the publication is relevant since the number of publications of a certain sentiment reflects public attitudes towards RE. Moreover, news publications not only convey a certain sentiment but also contribute to raising overall awareness and knowledge about a certain subject. Therefore, both the total number of publications and their potential reach will serve as proxies for overall knowledge and awareness or potential knowledge and awareness that is gained by the news publications regarding RE.

Determination of keywords

Before proceeding with the import of the relevant data, first, a set of keywords related to the energy transition is established. The keywords are derived from the study of Zhang, Abbas and Iqbal (2022) that investigate the perceptions of GHG emissions and renewable energy sources. This investigation involves text analysis of Twitter publications (Tweets) based on certain keywords, identifying word collocations, and comparing the findings with the Google search interest regarding the same words. The researchers used the keywords “greenhouse gas”, “GHG”, “renewable energy”, “coal”, “natural gas”, “solar energy”, “wind energy”, “biomass”, “hydro energy”, “geothermal energy”, “tidal energy” for the Twitter and Google search. These keywords are complemented by terms found on the website of the International Energy Agency (n.d. - b) related to technologies within the RET encompassing areas such as hydropower/hydroelectricity, bioenergy, heating, heat pumps, energy efficiency and ocean power/tidal power). This combination of keywords will be referred to as the initial set of keywords.

Data pre-processing

Before incorporating the set of keywords into the primary analysis of this research, preliminary text analysis is conducted to identify related words through high occurrences and collocations/co-occurrences. The process is explained in the following steps:

1. **Initial Keyword Selection**: The Dutch initial set of keywords is utilized, for the period 1, January - 31, December 2021, along with specific instructions for the Meltwater software to find publications with one of those words in the title of the publication ensuring more significant results. The specific code used in the Meltwater software in combination with these keywords with an example of how this looks are shown in the Appendix (spread over Appendix C).

2. **Data Extraction**: Meltwater extracts all publications meeting these criteria resulting in a list of publications alongside information about these publications. The data is then imported and processed using the R programming language.

3. **Identification of Relevant Terms**: High occurrences of other relevant words and collocations and co-occurrences between words from the initial keyword set and other relevant terms are

investigated. The *udpipe* R package can categorize each word (e.g., verb, adjective, noun). A selection of relevant nouns (e.g., wind energy, heat pump) and adjectives (e.g., green, renewable) is made.

4. **Occurrence plot**: An occurrence plot featuring the 30 most frequent words is created.

5. **Refinement of Occurrence Plot**: Words in the occurrence plot that are not directly relevant to the energy transition (e.g., research, year, application) are removed once.

6. **Final occurrence plot**: Step 4 is repeated, creating the final occurrence plot for this iteration. Additionally, word co-occurrences and collocations are examined.

7. **Incorporation of new terms**: New words identified through high occurrences and collocations are added to the initial keyword list.

8. **Iteration**: The iteration of all steps is complete and will be repeated once (2 iterations), combining the new set of keywords with the initial set to uncover additional relevant keywords.

The *udpipe* package analyzes a single language at a time. Further analysis is conducted exclusively on the Dutch translation of keywords due to the consistent inclusion of numerous irrelevant terms in the English translation after multiple iterations. The occurrence plots from the first and second iteration (after step 6) and tables detailing the co-occurrences of specific words and frequently appearing phrases (comprising three words) after the second iteration are provided in the Appendix.

After the first iteration, key terms such as 'green,' 'windmill,' 'sustainable,' and 'CO2-emission' were identified. Specifically, adjectives like 'green' or 'sustainable' were paired with 'energy' in the successive search to ensure relevance to the energy transition. Additionally, the term 'subsidy' prominently surfaced in the first search. This suggests that renewable energy publications could enhance awareness and understanding of renewable energy subsidies.

Following the second iteration, the key terms extracted from the occurrence plot, and the co-occurrence and collocation tables include 'energy source' and 'fossil'. Additionally, the term 'investment' surfaced, suggesting that these publications might be enhancing awareness regarding RE investments and potential opportunities in the field. Too broad terms such as energy consumption or generation were omitted from consideration.

Ultimately, the new relevant terms (relative to the initial set of keywords) obtained through this text analysis are 'windmill' and 'CO2-emission'. Furthermore, 'sustainable' and 'green' are relevant when paired with the terms 'energy' or 'energy source.' These words are integrated with the initial set to form the final set of keywords. This final set is then used to conduct the ultimate search in Meltwater, utilizing both the Dutch and English translations with the Netherlands specified as the region. This search aims to retrieve the cumulative number of publications and their reach, serving as essential psychological data for this research.

Data Import

With the final set of keywords, the psychological data required for this research is extracted. Before the data import, a sanity check is conducted of the publications retrieved by Meltwater to ensure that the publications are relevant to the keywords. Subsequently, the data is imported, covering the period from January 1, 2015, to December 31, 2020. After importing the Meltwater data, it was observed that a part of the publications did not have a region specified. Consequently, additional checks were performed based on the source of the publications to identify their respective regions. Specifically, if the source name contained the name of a municipality, manual matching was carried out to assign the correct region. Moreover, additional regional sources, often encompassing multiple provinces, were designated to the municipalities they covered. The publications that were not regional or municipality-specific are seen as national sources. Since these sources have the same effect on all municipalities, they were excluded from the analysis. Following the data-cleaning process, the publication dataset consisted of 345,514 separate publications. These publications were aggregated, summing the number of publications with positive, neutral, or negative sentiment for each region and year. This aggregation was done to maintain consistency with the dimensions of the main dataset. To conclude, the five independent psychological variables that result from this data import process: are the number of publications categorized as having a negative, neutral or positive sentiment, the total number of publications and the potential reach of these publications.

Klimaatmonitor

The RE investment, socio-demographic and building characteristic data for the residential sector in NL, along with energy consumption data, are publicly obtainable through *Klimaatmonitor* (Klimaatmonitor, n.d.). Klimaatmonitor is a website managed by a branch of Rijksoverheid, the Dutch government's central authority responsible for all national-level legal duties. It is set up to monitor the decentralized climate transition. The primary objective of the website is to enable decentralized governments to track their policy objectives, evaluate their policies and make necessary adjustments. Klimaatmonitor does not collect its data but integrates information from various reputable Dutch data sources, including the Central Agency for Statistics (CBS) and the Netherlands Enterprise Agency (RVO) (Regionale Klimaatmonitor, 2023). Certain variables are derived by combining datasets. For a detailed explanation of how specific variables are constructed and which assumptions were made, refer to Klimaatmonitor's methodology documentation.

Energy consumption

The in-depth analysis conducted on energy consumption, presented in the theoretical framework, demonstrates that it serves as a robust metric capable of capturing the real energy outcomes resulting from collective actions, irrespective of whether the energy sources are sustainable or fossil-fuel-based. Through the examination of this metric, patterns in energy consumption behaviour can be discerned, revealing insights into the relationship with the RET, such as consumers' ability or willingness to invest in energy conservation. This study primarily examines per capita energy consumption in residential areas, aiming to create tangibility around energy-related behaviours within this sector.

It is crucial not only to examine overall energy consumption but also to differentiate between its sources. Total energy consumption (TEC) can arise from either RE consumption or non-renewable energy (NRE) consumption. Ideally, our energy demand would be fully green, sourced entirely from renewable sources completely substituting energy from non-renewable sources to mitigate carbon dioxide emissions. Most previous research focuses on total energy consumption. However, only non-renewable energy seems relevant since this is the emitting part of the sector. Consequently, the most significant potential for improvement is within this emitting segment of the energy sector, which will be the primary focus of this research.

Since Klimaatmonitor only publishes TEC and RE consumption, the dependent variable is calculated as follows: TEC - RE Consumption = NRE Consumption (expressed in gigajoule - GJ). This is divided by the number of residents in a municipality to obtain the final dependent variable: the per capita NRE consumption (PC NRE consumption). It is important to note that a decrease or lower PC NRE consumption is seen as beneficial for the RET.

This approach enables capturing the potential substitution effect of RE, stemming from variables like income discussed in the theoretical framework. Higher-income households may exhibit higher total energy consumption but, on the other hand, tend to make more environmentally-friendly investments in energy-conserving or green energy-producing technologies. The RE generated through these investments in energy-producing technologies can serve as a substitution for NRE consumption.

To calculate TEC and RE consumption, different elements of the data that Klimaatmonitor publishes are used. Klimaatmonitor categorizes TEC into electricity, natural gas, and district heating, which are combined to derive TEC. Additionally, Klimaatmonitor has only published the proportion of RE of the overall TEC across all sectors. It's crucial to emphasize that Klimaatmonitor assumes that the renewable energy consumed remains within the individual municipalities, aligning it with the renewable energy produced within those regions. Additionally, this research makes a specific assumption that the proportion of renewable energy across all sectors remains consistent within the residential sector as well. Consequently, the residential RE consumption is calculated as follows: %RE consumption * TEC = RE consumption.

RE investment, sociodemographic and building characteristic data

To capture the effect or progress of monetary policy or incentives on the RET, the dataset includes data on investments done regarding RE, covering the cumulative count and capacity of realized projects that have received an SDE (Stimulation of Sustainable Energy Production & Climate Transition) subsidy. This subsidy was introduced by the Dutch government in 2008 to reduce GHG emissions in the Netherlands. It was designed to incentivize companies and non-profit organizations to actively engage in the large-scale generation of RE or the reduction of carbon emissions. Notably, this stands as the most generously funded subsidy aimed at addressing carbon emissions in the Netherlands. Over time, the scheme evolved to accommodate emerging technologies, starting with SDE+ in 2011 and the most recent version, SDE++ introduced in 2020. All versions are accumulated in the dataset. The SDE(++) subsidy program encompasses various green categories spanning renewable electricity, gas, heat, and low-carbon heat and production (Rijksdienst voor Ondernemend Nederland, 2012, 2020). Another two variables were included to capture another type of monetary incentive, by retrieving data on the quantity and magnitude of loans distributed for sustainability and energy conservation purposes. Furthermore, two variables were added containing information about the percentage or capacity of houses with registered solar panels. In the Netherlands, individuals can receive subsidies for installing solar panels on their homes. Yet, no available data indicates whether the registered solar panels were directly stimulated through subsidies or other monetary policies. Despite this, these variables are considered RE investments and are categorized within the same group of variables. This results in six independent variables containing information about RE investment, often facilitated through monetary policies or incentives. This data may also indicate the entrepreneurial mindset within a municipality concerning RE.

Within the socio-demographic category, additional socio-demographic variables are investigated beyond those explicitly stated in the theoretical framework to examine the possibility of additional relationships. This means that within the socio-demographic category, variables such as average household income, average household size, percentage of rented and owner-occupied homes, housing density (urbanisation) as well and the value of a house (WOZ-value) are included. Moreover, the dataset incorporates data related to building characteristics, including the age of the residence expressed as a percentage of buildings constructed before or after the year 2000 and the percentage of residences with the highest energy label (ranging from A to A++++). In summary, the dataset comprises nine sociodemographic variables and three building attribute variables.

Data tests and modifications

Upon merging all the datasets, the resulting dataset consisted of 23 independent variables alongside the dependent variable, the year and municipality columns with 2052 observations. The majority of the variables within the dataset are continuous, except for the municipality and year columns. The variables in the complete dataset, before undergoing data cleaning, are provided in Table 2.

Table 2

Information on the variables included in the analysis

Variable type	Variable	Description	Additional information
Dependent variable			
Energy consumption	*y*	Per capita energy consumption from non-renewable sources in the residential sector expressed in Gigajoules (GJ)	(Total energy consumption in residential sector – renewable energy consumption in residential sector)/ population
Independent variables			
Psychological	Sentim_pos	The number of news publications with positive sentiment concerning the energy transition	Captures positive awareness (/knowledge) about renewable energy
	Sentim_neut	The number of news publications with neutral sentiment concerning the energy transition	Captures awareness (/knowledge) about renewable energy
	Sentim_neg	The number of news publications with negative sentiment concerning the energy transition	Captures negative awareness (/knowledge) about renewable energy
	Tot_pub	The number of total news publications	Captures awareness (/knowledge) about renewable energy
	Tot_reach	The number of unique page visitors of the source of the publications	Popularity of the source of the publications (reliability)
Investment related variables	No. of subsidies: #_subs	The number of SDE subsidised (granted) projects	Captures how many green energy investments are made available through SDE subsidies (monetary incentives)
	Capacity of subsidies: Cap_subs	The capacity of SDE subsidised (granted) projects (MW)	Captures how large green energy investments are made available through SDE subsidy (monetary incentive). Could also capture the

			availability/supply of renewable energy in a region.
	No_loans_100kres	Number of sustainable/energy-conserving loans per 100,000 residents	Loans (#) to make sustainable or energy-conserving investments (monetary incentive)
	Loan_amount_100kres	The amount of sustainable/energy-conserving loan per 100,000 residents in euros	Loans (€) to make sustainable or energy-conserving investments (monetary incentive)
	P_solar	Percentage of properties with registered solar panels [%]	
	Cap_solar	Capacity of registered solar panels per house (watt-peak)	
Socio-demographic	Tot_res	Total residents	
	Income	Average household income in €	
	HH_size	Average household size	
	Total_HH	Total households	
	Density	Housing density – number of homes per hectare	Captures the degree of urbanisation of an area
	P_homeowner	Percentage of home owner-occupied properties	
	P_rental	Percentage of properties rented	
	WOZ_value	Average house value [1000euro]	
	Land_area	Area of land in hectares	
Building - characteristics	Homes_<2000	Percentage of homes built earlier than 2000	
	Homes_>2000	Percentage of homes built from 2000 onwards	
	P_A+	Percentage of labelled properties with the highest valid energy label (A – A++++)	

The dataset underwent several tests and modifications before training. Initially, missing values in variables, such as the percentage of renewable energy (22) and total electricity consumption of homes (14), which were necessary for calculating the dependent variable, were addressed using a technique developed by Wright and Ziegler (2015) and executed through the ranger (short for RANdom forest GENerator) package in R. This method involved imputing missing values by predicting them with an RF model, a method elaborated upon in the Methods section, where other variables served as covariables. Alternative methods like using municipality mean or medians were deemed unsuitable due to their lack of consideration for time effects in this panel-type data, focusing solely on municipalities as groups and showing unlogical values. There was a minor concern about the internal correlations of the imputed values, which involved two variables in calculating the dependent variable, with the final analysis outcome since an RF will also be used in the main analysis. However, to mitigate this concern, the integration of the dataset with psychological variables was postponed until the missing values were resolved in combination with the other variables. Considering the psychological data the missing values are assigned a value of 0, indicating the absence of publications for that specific municipality in that particular year.

Additionally, the relationships among variables were assessed using the variance inflation factor (VIF). High VIF values within the psychological variables, and relatively lower yet notable values were observed for the percentage of properties with registered solar panels (*P_solar*) and capacity of registered solar panels per house (*Cap_solar*) as well as the percentage of homeowners (*P_homeowner*) and rental homes (*P_rental*). These findings indicate substantial correlations among these features. However, only mild efforts were made to address multicollinearity, considering that tree-based ensemble algorithms, which will be discussed in the *Methods* section, can effectively handle collinearity. Income and house value (*WOZ_value*) exhibit a correlation (0.86) but were identified in the literature as separate variables influencing energy consumption and thus are retained. Conversely, certain other features were logically removed from the model. Features like *P_homeowner* and *P_rental* exhibited a nearly perfect negative correlation (-0.99) with each other; therefore, one of the two was omitted. The total number of publications (*Tot_pub*) displayed the highest correlation with another psychological variable, especially with the number of publications with a neutral sentiment (*Sentim_neut*; 0.99*),* while adding no logical value to the model. Consequently, it will also be excluded from the analysis. Moreover, the variable *Total_HH*, representing total households, will be removed from the dataset. Its exclusion is essential because it could impact the dependent variable, which is derived from the municipality's resident count. As the number of residents increases, total households naturally rise, potentially exerting excessive influence on predicting the dependent variable. Removing these variables enhances the interpretability of other features in the dataset.

Methods

The following section provides an overview of the methods utilized to handle the data and to address the research question. Various machine-learning models were explored in the literature review to capture complex relationships in the energy transition. These models include simple linear regressions, SVM, (extreme) gradient boosting, RF, and neural network algorithms. Besides the Municipality (categorical) and Year (factor) columns, the research data comprises numerical variables, containing information on different municipalities over time, indicating panel data characteristics. The problem is framed as a regression task since the dependent variable, energy consumption, is a continuous variable.

Support vector machines are primarily used in tasks involving high-dimensional classification tasks and are not ideal for managing panel data. On the other hand, neural networks, more specifically, Long Short-Term Memory networks can handle panel data well but require larger datasets and more sequential data. Tree-based ensemble methods such as RF, GBM and XGB provide greater flexibility in handling diverse data types and distributions and eliminate the need for extensive datasets. Apart from their sensitivity to outliers, these methods excel in addressing non-linear problems, multicollinearity among features and managing panel data complexities. Additionally, they are well-suited for regression tasks and can effectively handle both categorical and numerical data. Therefore, to predict energy consumption aggregated on the municipality level per year and distinguish key performance drivers, the RF, GBM and XGB algorithms will be employed in this research. A simple linear regression model (LM) is used as a benchmark for the machine learning models' predictive performance and enhances the interpretability of the data. These methods will be discussed in the following sections.

Linear regression (LM)

Linear regression, a fundamental statistical technique, models a linear relationship between a dependent variable (or outcome) and one or more independent variables (or predictors) by employing coefficients that include intercepts and slopes. These coefficients, crucial for predicting the dependent variable, are estimated during the training process on the data. The primary objective of an LM is to discover the most accurate linear relationship that forecasts the values of the dependent variable based on the provided independent variables. This relationship is expressed through the general linear equation (Equation 1):

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + \varepsilon \qquad (1)$$

Where:

- y is the dependent variable.
- $x_1, x_2, \ldots, x_n$ are the independent variables.
- b_0 is the intercept (the value of y when all independent variables are 0).

- $b_1, b_2, \ldots, b_n$ are the coefficients (indicating the change in y for a unit change in each x).
- ε represents the error term, accounting for the difference between the predicted and actual values.

The regression model calculates the coefficients ($b_0, b_1, \ldots, b_n$) that minimize the sum of squared differences between the predicted and actual values. Once the model is trained, it can be used to make predictions on new data.

Random forest (RF)

A random forest (RF), introduced by Breiman (2001), is an ensemble machine-learning method employed for classification, regression and various other tasks. An ensemble algorithm is a method that aggregates multiple models, combining *weak learners* to create a robust predictive model. These ensembles enhance predictive accuracy and minimize overfitting. A weak learner, such as a decision tree, is a model that predicts better than random guessing but does not have high predictive performance. Another ensemble method closely related to the RF algorithm is *bagging*, which is short for bootstrap aggregating. Bagging is an algorithm that takes random samples with replacement from the data and for each sample, a separate weak learner is constructed in the same way using bootstrapping. These weak learners are then combined to create the final output of the RF which is determined through averaging (for regression) or majority vote (for classification) among the weak learners.

RFs are a specific form of bagging, specifically composed of decision trees and differ from bagging in their approach. While both methods use random samples, RFs employ a distinct splitting rule. Instead of using all features for each sample, RFs only use a random subset of features to split trees. This crucial difference ensures that the decision trees within a random forest are uncorrelated, setting them apart from traditional bagging where decision trees are correlated.

In RFs, hyperparameters are predefined settings related to the specifications of the decision trees or bootstrapping type used to create the RF model. Examples of these hyperparameters include the number of decision trees, the number of randomly selected features considered for the best split, or the minimum number of observations in a node for a split in a decision tree to occur. Defining combinations of hyperparameters, known as “tuning”, is essential to optimize the model's performance. The tuning process can be accomplished by trial and error or through a grid-search using cross-validation, to find the optimal set of hyperparameters resulting in the most accurate predictions. The hyperparameters that are tuned for the RF in this research are the minimum number of observations in a node for a split to occur (*min.obs.size*) and the number of variables randomly sampled as candidates at each split (*mtry*). RF was employed throughout energy-related research, as demonstrated in studies found in the literature research like Bennett et al. (2020), where it predicted the number of solar installations, and in the work of Gassar et al. (2019), where it was used to forecast gas and electricity consumption.

Gradient boosting (GBM)

Similarly, the boosting algorithm is a tree-based ensemble method. It is a supervised machine learning algorithm designed for classification and regression tasks, primarily focused on reducing bias. In contrast to bagging, boosting does not independently select random data samples. Instead, boosting creates the subsets and trees sequentially, incorporating errors from prior trees. This sequential approach contrasts with bagging's parallel and independent sampling. Moreover, in boosting, weak learners are assigned varying weights, unlike bagging, where all weak learners have equal weight. Both techniques reduce variance, but only boosting reduces bias, making it susceptible to overfitting. The output relies on weighted averages for regression or weighted voting for classification (Freund, Schapire & Abe, 1999).

One prominent boosting algorithm is gradient boosting (Friedman, 2001). Gradient boosting was found to be a useful algorithm in energy-related literature to predict, e.g. gas and electricity consumption (Gassar, Yun & Kim, 2019). In regression tasks, a gradient boosting machine (GBM) minimizes a *loss function*, a function that measures the difference between predicted and actual values, also known as the *loss* or the *cost*. GBM uses the mean squared error (MSE) as its loss function, defined as $MSE = \sum_{i=1}^{D}(x_i - y_i)^2$, where x_i the actual value of the i^{th} observation, y_i the predicted value for the i^{th} observation. and optimizes it using the gradient descent algorithm, hence the name *gradient* boosting. The process begins with an initial guess (typically the average of the target variable in the first iteration), which is subtracted from the actual values, creating residuals. These residuals are then utilized to construct decision trees with a predetermined depth. The trees predict residuals for each observation, and the leaves or terminal nodes of the trees represent the averages of the residuals within those leaves.

To minimize the loss function, the gradient descent algorithm sequentially adjusts the fitted trees. It achieves this by reducing the residuals of the trees along the gradient of the loss function. Incremental steps are taken guided by the first-order derivative of the loss function, often referred to as the pseudo residual. However, a potential challenge in this process is overfitting, which results in low bias but high variance. Consequently, precautions are necessary to mitigate these issues.

To prevent overfitting, a learning rate is employed with a value between 0 and 1. It determines the step size in each iteration, gradually moving towards the loss function's minimum where the slope is lowest. Smaller step size enhances accuracy but extends computation time. Multiplying the tree, or more specifically the residuals in that tree, by the learning rate scales the contribution of a new decision tree, facilitating gradual minimization of MSE. The gradient descent process, involving regression, residual addition, and tree construction, continues until predictions can no longer improve, reaching the minimum of the loss function.

Comparable to an RF, several hyperparameters require tuning before training with gradient boosting. Tree-specific parameters like tree depth and minimum observations in terminal nodes of the trees can be set, reducing overfitting. Additionally, boosting hyperparameters, including the number of trees and the aforementioned learning rate to scale the trees, can be specified. The hyperparameters that are tuned for the GBM in this research are the tree depth, learning rate (*shrinkage*) and the minimum number of observations in a node for a split to occur (*n.minobsinnode*).

Extreme gradient boosting (XGB)

Extreme Gradient Boosting (XGB) is an advanced version of GB that operates on similar principles but is notably faster. Like GB, XGB fits regression trees to residuals with sequential tree building relying on residuals from prior trees, refining predictions iteratively until residuals are minimized or a predetermined tree limit is reached. However, the XGB is trained using regression trees that are unique to the XGB algorithm.

As discussed in the previous section, the first-order derivative of the loss function is used in the gradient descent to find the direction of the step to which the tree parameters must be adjusted to minimize the error. In contrast to GB, XGB calculates errors using the second-order derivative of the loss function, enhancing accuracy when adjusting parameters. Additionally, XGB integrates a regularization parameter (lambda) that prevents the training data from overfitting. Without regularization, the output or leaf equals the average of residuals in that leaf, the same as with normal GBM. This parameter reduces sensitivity to individual observations, controlling the weights of certain leaves. Another hyperparameter in XGB that avoids overfitting and scales the output of a tree is a learning rate called eta, similar to GBM's learning rate (Chen, et al., 2015). The XGB hyperparameters tuned in this study include the number of rounds (maximum iterations or trees fitted), tree depth, learning rate (eta), and minimum loss reduction (gamma).

XGBoost (XGB) holds significance in this study as identified in energy literature research, where it has been utilized to forecast metrics such as the count of solar installations (Bennett et al., 2020).

Furthermore, the XGB requires that the data comprises predominantly numerical variables, except for variables denoting municipality and year, no preliminary data adjustments are necessary.

Interpretability methods

The methods discussed in the previous sections, random forest, GBM and XGB, are expected to outperform the simple linear regression in predictive accuracy. While these models can handle nonlinear relationships, their interpretability is limited. To address this challenge, interpretability methods have been developed that generate a model that is interpretable by itself. Interpretability methods are either model-specific, tailored for specific machine learning models, or model-agnostic, applicable to any complex model. They can be categorized into global and local interpretability methods. Global methods provide insights into how models make decisions overall, while local methods focus on specific decisions within the model concerning particular observations. A prime example of a global interpretability technique is the Variable Importance (VI) plot. This method assesses the increase in the model's prediction error by permuting the values of a predictor while keeping others constant. The greater the increase in the model's error due to this permutation, the more important the feature is considered. Plotting these feature importance values results in a highly interpretable outcome, with the most significant variables appearing at the plot's top.

While the previous plot provides overall insights, it can not reveal the direction of a specific feature's impact. To gain more clarity on the relationship direction, another global interpretability technique can be employed: the Partial Dependence Plot (PDP) introduced by Greenwell (2017). This plot illustrates the extent to which the target variable changes when a particular feature varies, considering the values of all other input features. Moreover, it is also possible to visualize how two distinct features influence the target variable in a multi-predictor PDP.

Model evaluation

To measure the predictive performance of the machine learning models the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE) will be measured. The formulas of the performance measures are shown in Equation 2 and Equation 3, respectively.

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \qquad (2)$$

$$\text{MAPE}(y,\hat{y}) = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (3)$$

N denotes the number of observations, y_i the actual value of the i^{th} observation, $\hat{y}_i^2$ the predicted value for the i^{th} observation. The model with the best predictive performance will be employed for further analysis and to address both the main research question and subquestions. This model most effectively captures the key features influencing energy consumption, which will be analysed with the interpretability methods.

Analysis & Results

Feature selection

Since a simple linear regression model (LM) is used as a benchmark for the other machine learning models, assumptions for a linear regression must be met. Given the absence of normality and the presence of skewness in the independent variables, they undergo a log transformation, while the dependent variable remains unchanged. Afterwards, the dataset underwent an 80%/20% split, resulting in a training set with 1641 observations and a test set with 411 observations. To enhance interpretability and computational efficiency for the other machine learning models, a recursive feature elimination (RFE) technique was employed. RFE is a method that iteratively eliminates less important features from the model, utilizing the performances obtained during model fitting, which in this instance involved an RF. This approach selects the most relevant features, making the model more understandable. The feature selection process revealed that the most accurate RMSE is attained with 11 variables out of the total independent variables of 23, as shown in Figure 1.

Figure 1

Plot showing RMSE is the lowest for 11 variables

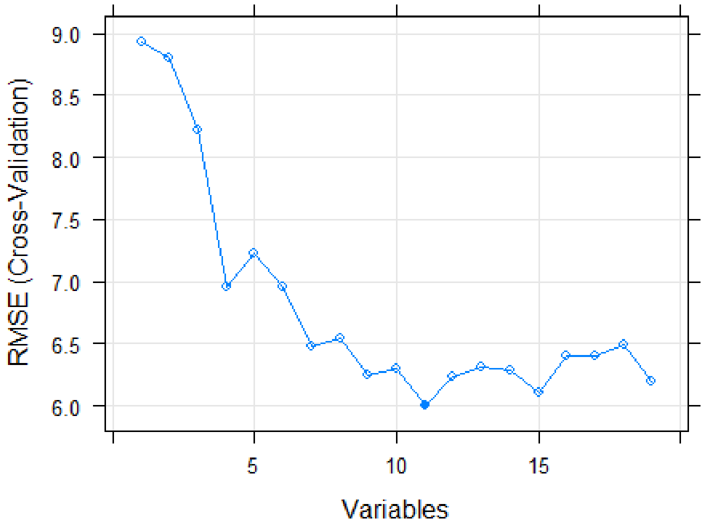


Table 3 displays summary statistics for the dataset before log-transformations for the final 11 variables selected through RFE, that will be used to perform further analysis. These variables encompass a combination of all different types of determinants distinguished in the theoretical framework demonstrating a various range of relationships. Out of the 11 features examined, 2 provide information about the psychological state in a municipality, 3 are related to investments of which 1 is monetary policy-driven, 5 pertain to sociodemographic factors and 1 provides information about building characteristics.

Table 3

Summary statistics of the 11 most important features together with the dependent variable PC NRE (y)

Statistic	N	Mean	St. Dev.	Min	Max
Cap_subs	2,052	21.975	63.551	0	893
Homes2000	2,052	0.150	0.056	0.030	0.440
P_solar	2,052	0.109	0.067	0.000	0.360
Cap_solar	2,052	381.992	254.743	0	1,609
Density	2,052	393.329	479.307	14.070	3,183.650
Sentim_neg	2,052	9.587	10.198	0	59
Land_area	2,052	98.475	89.186	7.000	523.000
Sentim_pos	2,052	11.001	9.202	0	77

HH_Size	2,052	2.276	0.183	1.700	3.300
WOZ_value	2,052	241.331	65.478	120.000	735.000
Income	2,052	59,931.740	8,905.472	38,000.000	110,900.000
y	2,052	23.551	9.958	-1.561	382.358

An intriguing observation concerning the dependent variable is the presence of negative values, indicating municipalities where total energy consumption is lower than the renewable energy they consume. However, these negative values stem from an assumption made by Klimaatmonitor, outlined in the Data section, that energy produced in a specific municipality remains confined within its borders, equalizing energy consumption with production. In reality, municipalities might utilize the surplus of RE produced by others.

The municipality Zeewolde stands out as the only municipality with negative values, aligning with Marijnissen and Straver's (2020) publication about Zeewolde which was the first energy-neutral municipality in the Netherlands. Other variables that require attention are *Income* or *WOZ_value*, practically impossible to be 0. Fortunately, this is not the case and does not require any adjustments.

Model comparison

The results of the linear regression are shown in Table 4. An interaction effect is added, which was expected from the literature to have a combined effect with each other.

Table 4

Linear regression results for the relationship between the top 11 variables including an interaction effect and the dependent variable PC NRE consumption

	Dependent variable:
Variable	**PC NRE consumption**
Cap_subs	0.233
	(0.387)
New_homes	-14.644***
	(5.595)
P_solar	94.945***
	(7.565)
Cap_solar	-11.123***
	(0.531)

Density	-4.955***
	(0.349)
Sentim_neg	0.492
	(0.335)
Land_area	-1.903***
	(0.407)
Sentim_pos	0.450
	(0.460)
HH_Size	-30.962***
	(5.913)
WOZ_value	5.073**
	(1.991)
Income	-2.818
	(3.850)
Cap_subs:Sentim_pos	-0.318**
	(0.150)
Constant	153.131***
	(29.981)
Observations	1,641
R2	0.366
Adjusted R2	0.361
Residual Std. Error	8.732 (df = 1628)
F Statistic	78.179*** (df = 12; 1628)

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

Except for *Income*, number of positive or negative publications and the capacity of SDE subsidised investments, each variable exerts a significant influence on NRE consumption to a varying degree. The *WOZ_value* is significant at the 5% level, while other significant variables exhibit even greater significance with lower p-values. The model achieves an R-squared value of 0.36, indicating it explains 36% of the variance. Since the features underwent log transformation, the model follows a linear log format, necessitating a specific interpretation of coefficients. For instance, one of the most significant variables, *Cap_solar*, decreases NRE consumption by 0.011 GJ/capita for a 1% increase in density. On the other hand, a 1% growth in the percentage of homes with registered solar panels increases PC NRE consumption by 0.95 GJ. Furthermore, the most interesting finding is that the interaction effect between *Sentim_pos* and *Cap_subs* is significant (at the 5% level).

For hyperparameter tuning the performance measures MAPE and RMSE are used. Through a grid search, the optimal combination of these parameters was determined to select the final model. Firstly, for the RF model with 500 trees, the best hyperparameters were a minimum node size of 12 observations and sampling 8 variables randomly at each split. Predicting values on the test set with this RF model resulted in RMSE and MAPE of 2.28 and 5.44% respectively. Similarly, for the GBM model, the ideal configuration included a learning rate of 0.3, tree depth of 7, and minimum node size of 5 observations. Predictions on the test set with this GBM-fitted model resulted in RMSE and MAPE of 2.61 and 5.76% respectively. Finally, the optimal hyperparameter combination for XGBoost (XGB) was several rounds of 400, a tree depth of 7 and a learning rate of 0.05. The RMSE and MAPE that resulted from this XGB-fitted model were 1.49 and 4.59%. A comparison of the predictive performances of the linear regression, RF, GBM and XGB that are fitted with the optimal hyperparameters is shown in Table 5.

Table 5

Predictive performances of the LM, RF, GBM and XGB expressed in RMSE and MAPE

Model	**RMSE**	**MAPE**
Linear Regression	4.538	0.139
Random Forest	2.285	0.054
GBM	2.607	0.058
XGBoost	1.495	0.046

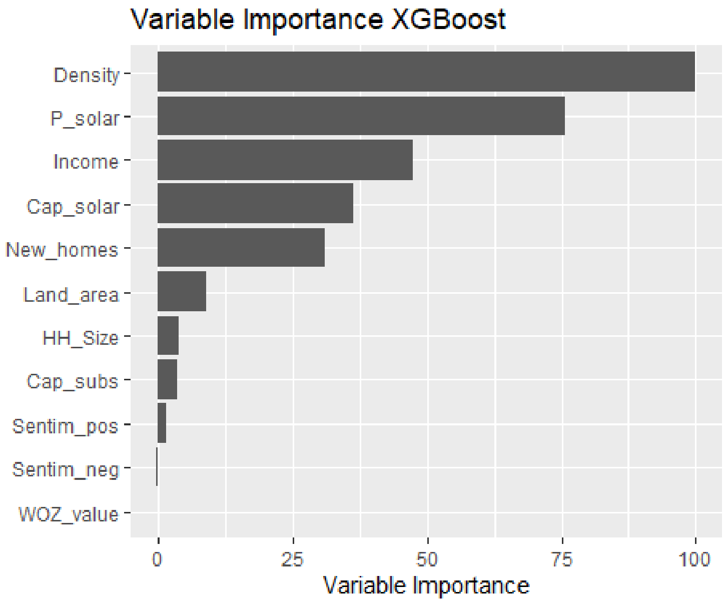
The XGB model resulted in the lowest RMSE and MAPE when compared to the other models. It seemed to most accurately capture the combination of variables to predict energy consumption. Therefore, the XGB model will be used to analyse further with the interpretability methods and draw conclusions, which are applied in the following sections.

Variable importance (VI)

Initially, a method for global interpretation is employed to analyze the XGB model. This involves examining the variable importance, which ranks the variables from most important at the top to least important at the bottom, as demonstrated in Figure 2. The population density, percentage of homes with a registered solar panel, and income were the most important drivers in predicting NRE consumption. In contrast, house value and the number of positive or negative publications had comparatively minimal impact. For interpretative purposes, the analysis focused on the three most crucial variables (*Density, P_solar, Income)*, along with variables stakeholders can influence, including psychological factors (*Sentim_pos* and *Sentim_neg*), the variable representing monetary policy-driven investment (*Cap_subs*) and the other investment variable (*Cap_solar).*

Figure 2

Variable importance (VI) plot

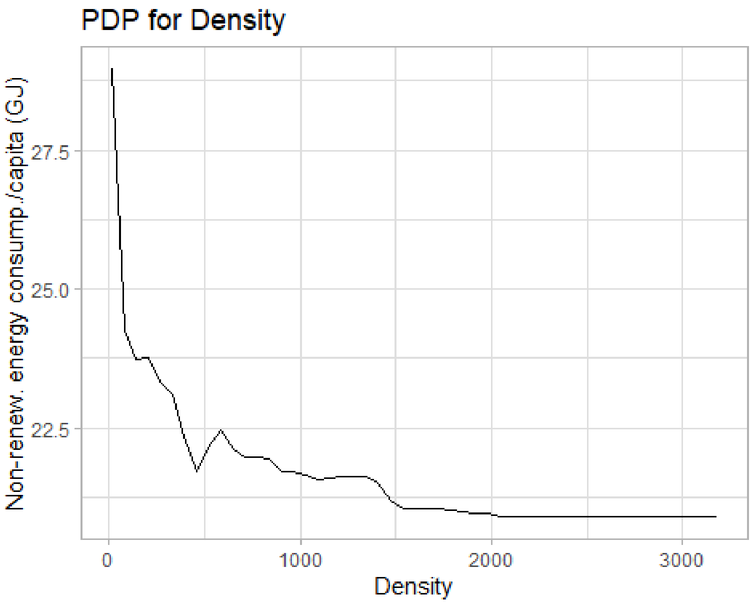


Partial dependence plots (PDPs)

To interpret how individual features affect the energy consumption PDPs are generated. The PDPs provided can be directly interpreted, as changes in independent variables, which were initially logarithmic, have been reversed. The y-axis represents the target values and the x-axis represents the values of the corresponding feature. To prevent misunderstandings, a negative relationship with PC NRE consumption is considered advantageous, leading to a greener municipality. Whether an effect is positive or negative only indicates its direction, not the perception concerning RET. Regarding the PDP of the most crucial variable, Density, consumption sharply drops from 0 to around 300 homes/hectare. For densities above 300, the decline continues but at a gentler rate until approximately 1500 homes/hectare. Beyond 1500, consumption stabilizes. To conclude, higher density results in lower PC NRE consumption, with the PDP seen in Figure 3.

Figure 3

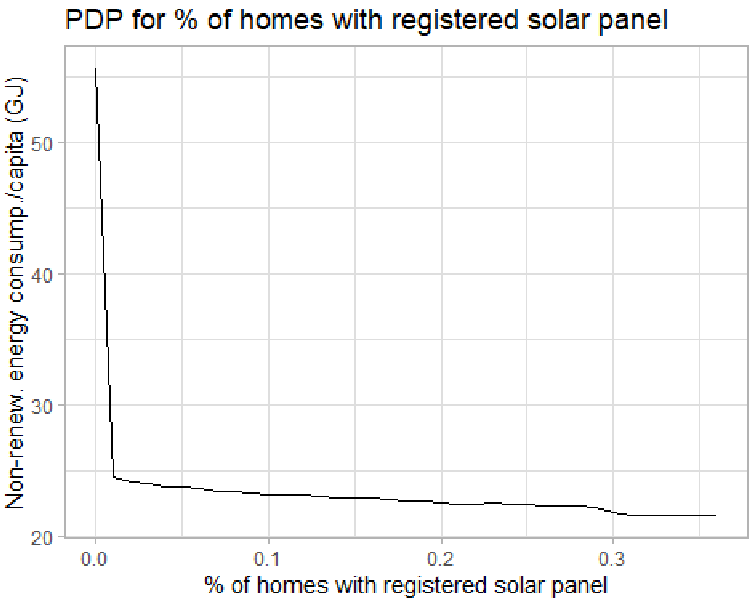
PDP for Density



The second most influential variable, the percentage of homes with registered solar panels, demonstrates a clear pattern. PC NRE consumption decreases sharply from 0 to 0.02 (2%), with a more gradual decline afterwards. At around 30% of homes with solar panels, consumption remains stable. In summary, an increase in the percentage of homes with solar panels correlates with decreased PC NRE consumption in a municipality. The PDP is shown in Figure 4.

Figure 4

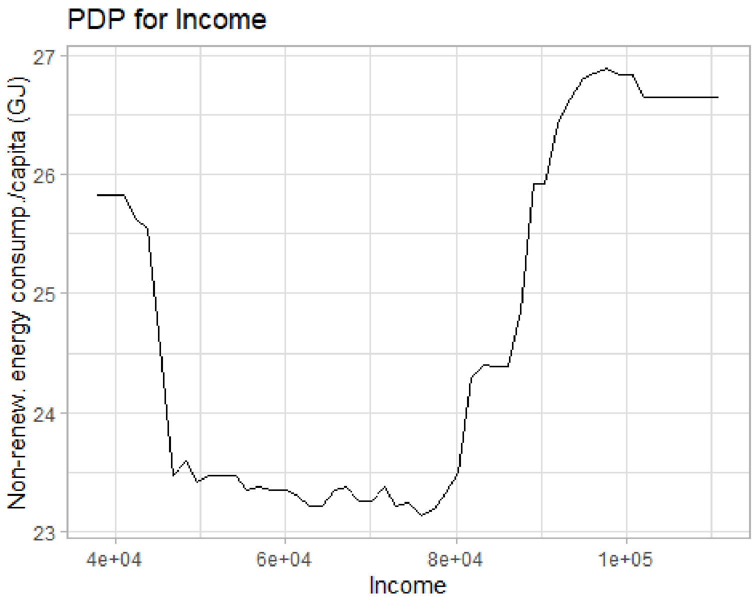
PDP for the percentage of homes with registered solar panels



According to the PDP shown in Figure 5 of the income variable, which ranks as the third most significant factor, a sharp decline in consumption occurs after €45,000, followed by a gradual decrease up to an income of approximately €75,000. Municipalities with higher average incomes experience significantly higher PC NRE consumptions, levelling off after an income of around €95,000. This pattern highlights that municipalities with a moderate average household income, ranging from €45,000 to €75,000, demonstrate relatively lower PC NRE consumption.

Figure 5

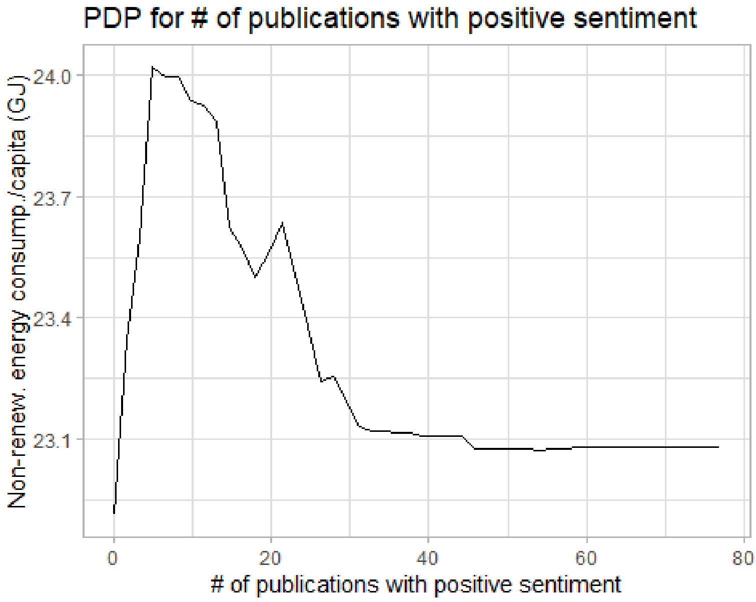
PDP for Income



Additionally, the PDP will be utilized to analyze psychological variables. Specifically, examining the PDP related to the number of positive publications in Figure 6 reveals a short initial rise in consumption corresponding to an increase in publications. This initial rise might be influenced by municipalities with minimal or no publications due to the absence of local publishers, which are not considered outliers. Improving the data granularity of psychological variables could enhance accuracy. However, after approximately 7 publications, consumption kept on steadily declining, reaching a stable point after 40 publications. Overall, a higher number of publications with a positive sentiment has a negative relationship with PC NRE consumption.

Figure 6

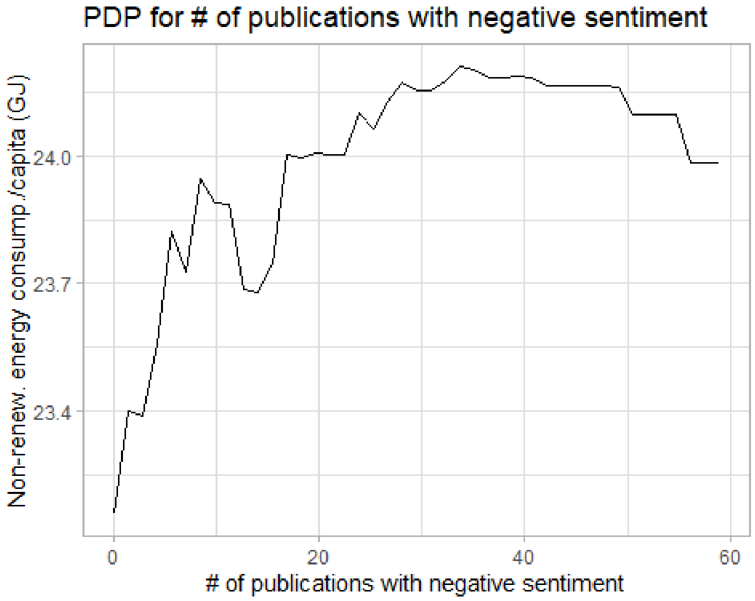
PDP of the number of publications with a negative sentiment



Regarding the PDP of the number of publications with a negative sentiment in Figure 7, a positive relationship is observed. In general a higher number of negative publications of up to around 30 leads to a higher PC NRE consumption. However, for higher publication numbers, the value remains stable on PC NRE consumption.

Figure 7

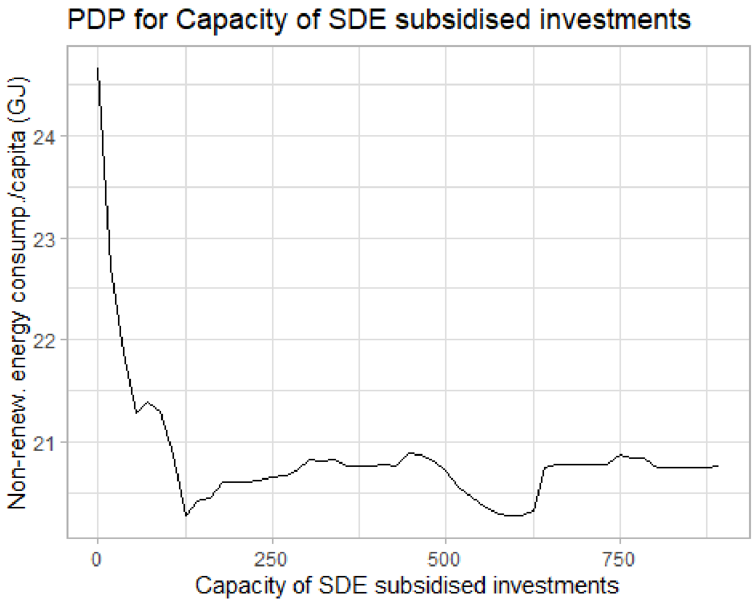
PDP of the number of publications with a negative sentiment



Analyzing the PDP Figure 8 of SDE-subsidized investments capacity, a decrease in per capita NRE consumption is observed from 0 MW up to 125. From that point onwards, although there are fluctuations, the consumption remains relatively consistent. In conclusion, higher capacities of SDE-subsidized investments lead to lower per capita NRE consumption.

Figure 8

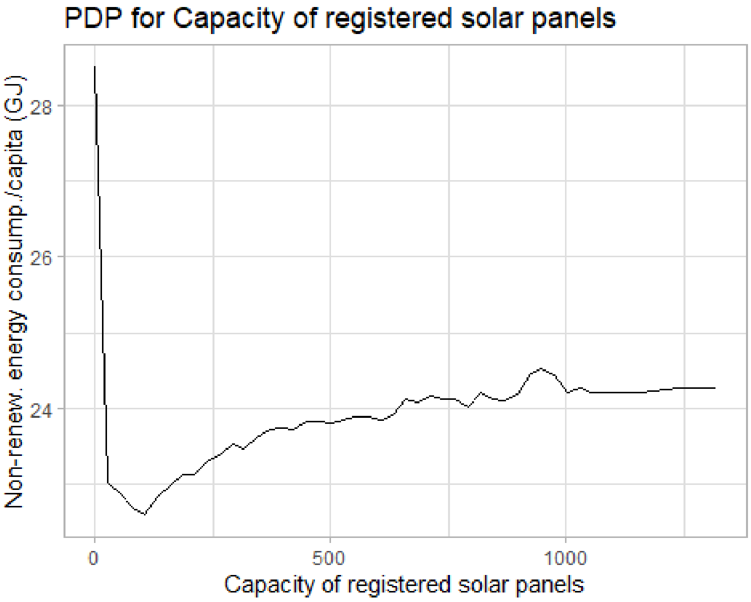
PDP for the capacity of SDE subsidised investments



The PDP of the capacity of registered solar panels on homes exhibits a sharp decline in PC NRE consumption from 0 capacity up to approximately 50 Watt-peak as seen in Figure 9. Beyond 100 Watt-peak, there is a constant slight increase in consumption. In summary, municipalities with no registered solar panel capacity experience the highest PC NRE consumption compared to municipalities with higher capacities. However, a capacity exceeding 50 Watt-peak does not significantly reduce PC NRE consumption any further and even increases slightly.

Figure 9

PDP for the capacity of homes with registered solar panels

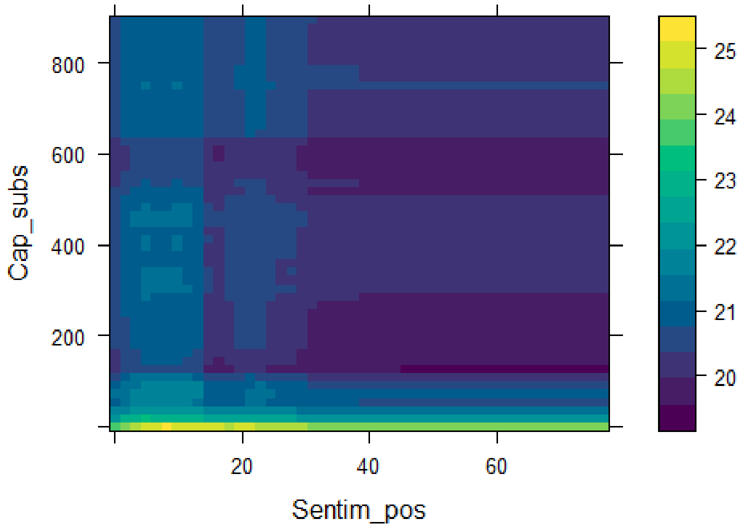


Multi-predictor PDP

The XGB considers interaction effects, however, it is not yet clear how it does this. To visualise how features have a combined relationship or an interaction effect multi-predictor PDPs are generated. These plots feature a layered bar on the right, using colours to represent various values of the target variable. Yellow shades indicate higher, while blue shades indicate lower NRE consumption PC, corresponding to specific features. Initially, the multi-predictor Partial Dependence Plot (PDP) in Figure 10 combines *Sentim_pos* and *Cap_subs* to explore potential interaction effects between psychological and policy-related variables, as outlined in the theoretical framework. The analysis reveals that municipalities with higher levels of both *Sentim_pos* and *Cap_subs* correspond to lower PC NRE consumption compared to municipalities with lower levels of these variables. Notably, when the number of publications exceeds approximately 16 or when the capacity surpasses 100 MW, PC NRE consumption reaches lower levels.

Figure 10

Multi-predictor PDP for Cap_subs and Sentim_pos



The theoretical framework suggests a second possible interaction effect: policies are more impactful when directed at specific socio-demographic groups. Consequently, a multi-predictor PDP is conducted on the two socio-demographic variables that are in the top 5 of most important variables, *Density* and *Income*, in combination with *Cap_subs*. The multi-predictor PDP has shown that PC NRE consumption is primarily impacted by income, reaching its minimum in municipalities with an average household income ranging from €50,000 to €80,000. Interestingly, there is no interaction effect between income and the capacity of SDE-subsidized projects. This is evident as consumption levels remain consistent regardless of low or high capacities. Similarly, no interesting interaction effect was found when combining *Cap_subs* with *Density*. For the other multi-predictor PDPs refer to the appendix (A.3 and A.4).

Removal of outliers

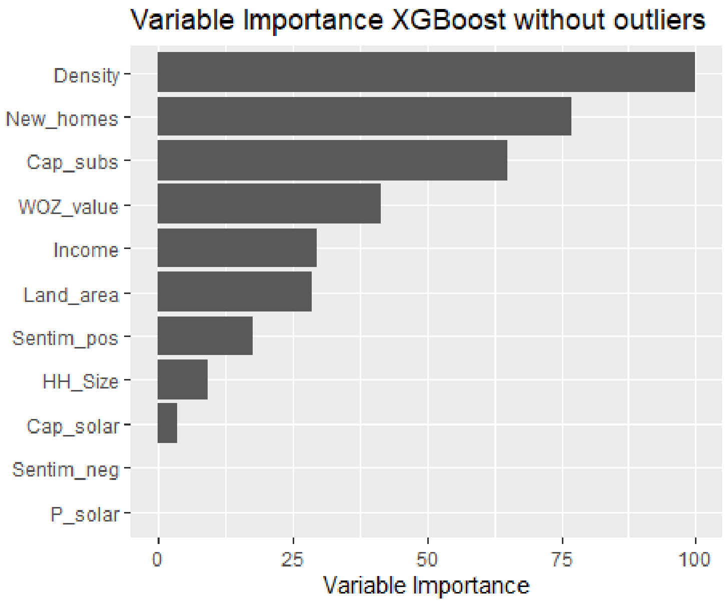
The dependent variable displayed extreme outliers extending far beyond the upper and lower bounds of the quantiles. These outliers, although extreme, are considered natural and represent actual observations from municipalities in the Netherlands. Known as 'true outliers,' they require careful handling. However, due to the sensitivity of boosting algorithms, under which XGB, to outliers, the analysis was repeated after removing outliers from the dependent variable. A comparison is made to evaluate whether the inclusion or exclusion of these extreme values affects the predictive power or interpretability of the features related to energy consumption using a more flexible definition for top and lower bounds: either 0.95 quantile (+) OR 0.05 quantile (-) 2 times the Interquartile Range (IQR).

Everything beyond these top and lower bounds is removed, resulting in a removal of 18 observations for the training set and 2 observations for the test set.

After removing these outliers, the XGB model was reevaluated. Firstly, the RMSE and MAPE of the XGB model, without outliers, improved to 1.3 and 4.00%, respectively. Secondly, the interpretability models were analysed. From the VI plot can be seen that the importance of the variables was more evenly distributed. Also, the top five variables changed: *Cap_subs* joined the list of the top 3 most important variables, alongside the previously most important variables *Density*, *New_homes*, and *Income*. *Sentim_pos* also became a more crucial feature. The VI plot after removing outliers is shown in Figure 11.

Figure 11

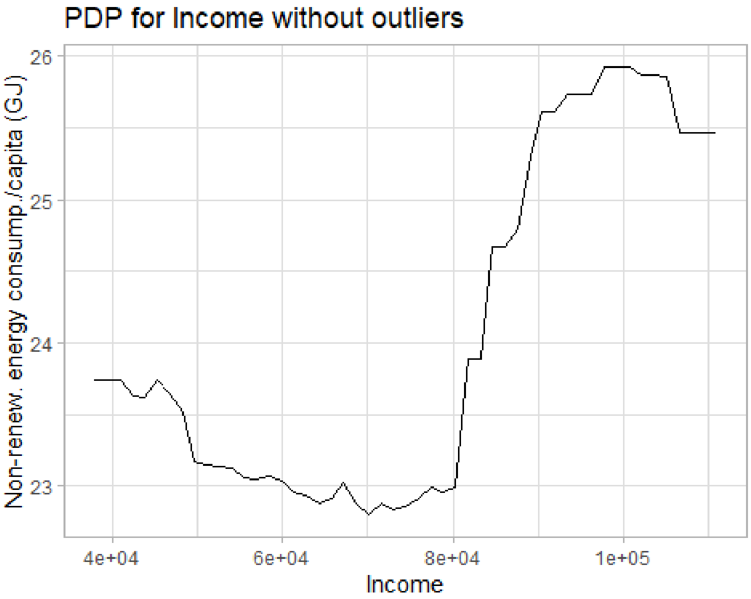
VI Plot after removal of the outliers



While most interpretations, based on the PDPs were consistent with the outcomes before outlier removal, the interpretations of key variables, namely *Income* and *Cap_Solar*, shifted somewhat after outlier removal. While moderate incomes consistently had low consumption, the difference between low and high incomes was more emphasised. Higher average incomes were now associated with considerably higher consumption compared to lower and middle incomes. The updated Partial Dependence Plots (PDPs) highlighted that municipalities with higher average household incomes, rather than lower incomes, exhibited elevated PC NRE consumption, and is shown in Figure 12.

Figure 12

PDP for Income after outlier removal



Just like the PDP of *Cap_solar* before outliers were removed, the PDP after outlier removal exhibited an initial dip. However, rather than stabilizing or showing just a slight increase, PC NRE consumption sharply rose for higher *Cap_solar* values. This suggests that a greater capacity of registered solar panels on homes significantly amplifies PC NRE consumption. Additionally, the PDP for the newly crucial variable, *WOZ_value*, followed a comparable path to the PDP of the *Income* variable. It indicated a rise in PC NRE consumption for municipalities with higher average housing values. In summary, when outliers were removed, interpretations remained consistent, enhancing clarity without altering the overall understanding.

Additional analysis

Top and bottom 10 analysis

An alternate analysis was undertaken, involving the extraction of the highest and lowest 10 PC NRE consumption values along with the top 11 features from the dataset, both before and after removing outliers. To ensure a diverse representation and prevent concentration within a single municipality over 6 years, the top and bottom 10 values were chosen. Additionally, the averages of these chosen top and bottom 10 features were computed to enhance comparability. These averages were then set side by side with the feature means of the remaining dataset to identify any distinct patterns. The outcome of this analysis is presented in Table 6, both before and after removing outliers with the corresponding features.

Table 6

Feature mean values for the top and bottom 10 consuming municipalities in comparison with the remaining data. Comparison is made before (3 left columns) and after removing outliers (3 right columns)

Variable	Top_10	Bottom_10	Remaining_data	Top_10_noout	Bottom_10_noout	Remaining_data_noout
Sentim_pos	8.2	15.9	10.991	11.9	15.1	10.958
Sentim_neg	5.8	16.6	9.571	12.2	16.5	9.545
Cap_subs	1.2	139.8	21.498	0.1	100.8	21.626
P_solar	0	0.187	0.110	0.065	0.163	0.110
Cap_solar	0	682.6	382.392	239.4	587.4	383.727
Homes_>2000	0.083	0.257	0.150	0.127	0.174	0.150
Density	138.755	46.211	396.290	243.157	134.426	397.259
Land_area	32.58	251.38	98.047	24.12	225.86	98.264
HH_Size	2.05	2.34	2.276	2.18	2.33	2.277
WOZ_value	308.1	249.1	240.964	474.8	221.7	240.475
Income	58650	62280	59926.493	85760	59590	59882.214
y	111.09	5.45198	23.210	35.60	13.301	23.145

When examining the averages of various features within the top 10 values before outlier removal (three left columns of Table 6), all features, except *WOZ_value*, displayed a lower average compared to the means of the remaining data. Oppositely, among the bottom 10 values, all features, except *Density*, exhibited a higher average in comparison to the remaining data. In both the top 10 and bottom 10 cases, *Density* had a consistently lower average, and *WOZ_value* had a consistently higher average compared to the remaining data. Despite the distinct impact of other features on either the top or bottom consumption values, when focusing solely on the means, *Density* and *WOZ_value* did not demonstrate a difference in their influence on either the top or bottom consumption values.

However, the scenario shifted after removing outliers (three right columns of Table 6). In the averages of the top 10 values' features, a more diverse pattern emerged. *Sentim_pos*, *Sentim_neg, WOZ_value*, and *Income* displayed higher averages compared to the remaining data, deviating from the predominantly lower averages observed earlier. On the other hand, among the bottom 10 values, Density, *WOZ_value*, and Income showed lower averages, while the rest exhibited higher averages compared to the remaining data. Density had a similar ambiguous impact on the top and bottom consumption values even after outliers were removed. However, a noticeable distinction emerged between the top and bottom values in terms of how *WOZ_value* influenced these variables. Interestingly, when outliers were removed, there was no clear differentiation in how *Sentim_pos* and *Sentim_neg* influenced the top and bottom consumption values.

To gain a deeper understanding of the highest and lowest values, the names of the municipalities associated with the top and bottom 10 PC NRE consumption values are examined. Interestingly, Ameland, Schiermonnikoog and Vlieland were initially among the top 10 municipalities with the highest PC NRE consumption before outlier removal. However, these municipalities belong to the West Frisian Islands (Waddeneilanden) in the Netherlands, situated off the Dutch mainland with very few residents (3761 in Ameland, 947 in Schiermonnikoog and 1155 in Vlieland for 2020), leading to higher per capita consumption. To maintain focus on the Dutch mainland, the island municipalities will be excluded from the analysis, achieved through outlier removal. Consequently, Blaricum, Laren, Rozendaal, and Wassenaar (excluding Schiermonnikoog) emerged as the top PC NRE-consuming municipalities after the removal of outliers.

Concerning the bottom 10 consumption municipalities, Zeewolde had the lowest PC NRE consumption. However, as detailed in the *Data* section, Zeewolde was the first energy-neutral municipality in the Netherlands and thus is not considered an outlier. Other municipalities that were among the bottom 10 PC NRE consumptions were Dronten, Duiven, Schouwen-Duiveland and Veere. Since a specific focus on the bottom 10 consuming municipalities was not necessary, conclusions about the top and bottom consuming municipalities are drawn based on the data after outlier removal, effectively excluding most of the island municipalities from this analysis.

XGB with all variables

Using XGBoost with all variables, including outliers, resulted in a slight increase in both RMSE and MAPE of 1.35 and 4.56%, respectively, indicating overfitting. Removing outliers improved predictive performance significantly, achieving the best results among all models with reduced RMSE and MAPE of 1.17 and 3.87%, respectively. However, incorporating all variables substantially compromised the model's interpretability. A balance between interpretability and predictive performance is crucial for this study. Given the importance of interpretability in addressing the research question, this model will not undergo further analysis.

Discussion - Interpretation of results

In the following section, the findings will be interpreted to address the RQ. The research question is addressed in a detailed manner, segmented into a distinct answer to each one of the four SQs. Among the LM, RF, GBM & and XGB models, XGB demonstrated the best predictive performance. Consequently, all interpretations, conclusions, and recommendations will stem from the XGB model. It is essential to clarify that when referring to negative or positive relationships, the interpretation is solely based on the direction of the effect on PC NRE consumption. A negative relationship, indicating a decrease or lower PC NRE consumption, is considered beneficial for the RET in a municipality. This research aims to identify the optimal combination of drivers that exhibit the strongest negative relationship with PC NRE consumption.

Sub-question 1: Drivers of PC NRE consumption

From the theoretical framework, it became evident that psychological state, monetary policy, socio-demographic factors, and building characteristics were variables influencing the RET in residential areas. Although the dataset encompassed all these variables, this research revealed that they affected PC NRE consumption to varying extents. To address SQ 1, the feature selection process was utilized. Out of the initial 23 determinants, only 11 remained as pivotal drivers in predicting PC NRE consumption. These variables included 2 psychological factors (out of 5): the number of publications with either positive or negative sentiment; 1 related to SDE subsidy-driven investment (out of 4): the capacity of SDE subsidized investments; 5 socio-demographic factors (out of 9): Income, Density (urbanization), Housing value, Land area, and average household size; and 3 building characteristics (out of 5): both the percentage and capacity of homes with registered solar panels and homes built after the year 2000. These variables emerged as the most influential factors driving PC NRE consumption and, consequently, the RET. The variables about the SDE subsidised investment and the percentage or capacity of homes with registered solar panels show how important RE investments are in driving RE consumption. When removing outliers, the order of the most important variables seems to be influenced.

Sub-question 2: In-depth analysis of the drivers of PC NRE consumption

To further explore these factors and answer SQ 2, the impact of these variables on PC NRE consumption is examined using PDPs and LIME plots. Nonetheless, not all variables hold relevance for addressing the RQ. The attention will be directed toward the top five crucial drivers and variables that can be influenced by individuals and/or the government, specifically emphasizing the psychological and investment-related variables. For psychological variables, the number of publications with a negative sentiment was found to have a positive impact on PC NRE consumption. Conversely, publications with a positive sentiment, except for a very low number, had a negative effect. These variables might reflect the municipal attitudes towards RE, general awareness or knowledge about RE, or the overall positivity of its residents. The impact of awareness and knowledge spread through RE-related publications appears to be influenced by the sentiment expressed. In conclusion, residents in municipalities with numerous publications expressing negative sentiment exhibit higher PC NRE consumptions, in contrast to areas with fewer such publications. Lower PC NRE consumptions are observed in areas predominantly featuring positive publications.

For both the percentage and capacity of registered solar panels, similar patterns emerge in their effect on PC NRE consumption. Initially, there is a sharp decline in consumption from 0 values up to 2% or 50 Watt-peak, respectively. However, beyond this threshold, the decrease levels off or even show a slight increase. Surprisingly, a higher capacity or percentage of homes with registered solar panels does not necessarily imply that the RE generated offsets NRE, contrary to what one might expect from an increase in RE investments. It appears people might tend to consume more energy overall, possibly because they perceive their energy consumption as renewable. In municipalities where fewer than 2% of homes have registered solar panels, investing in solar panels offers the highest marginal benefit for reducing PC NRE. Numerous subsidies, primarily administered by housing corporations, encourage solar panel investments. Consequently, the conclusions of this analysis might also carry implications for monetary policy. Unfortunately, there is a lack of data specifying this information and no conclusions can be made regarding the effect of monetary policy on this, however, could carry implications for individual investors.

Furthermore, increased Density or urbanization levels and Cap_subs overall appear to correlate with reduced PC NRE consumption. Denser population living conditions may lead to lower energy usage per person, possibly due to enhanced energy reuse. Furthermore, although it seems to fluctuate a lot, a higher capacity of SDE subsidised investments led to a lower PC NRE consumption.

Moreover, the study revealed that municipalities with higher average income levels exhibited the highest PC NRE consumption, while those with moderate average household incomes demonstrated the lowest PC NRE consumption. This finding is interesting as existing literature had primarily indicated a purely positive correlation between income and population-adjusted consumption. One possible explanation could be that municipalities with lower average household incomes might lack the financial means to adopt sustainable practices or may prioritize other needs. Conversely, literature suggested that higher-income households consume more energy due to their larger homes. Consistent with this notion, assuming a house's value correlates with its size, the research confirmed that the average house value positively influenced PC NRE consumption.

Sub-question 3: Interaction effects

SQ 3 is considered by using the multi-predictor PDPs and the outcome of the linear model to analyse how the determinants jointly influence PC NRE consumption. As brought forward in the theoretical framework, it is crucial for political interventions to bridge the gap between public sentiments and actual energy outcomes (Frederiks et al., 2015a) or target specific behaviours and psychological factors (Yang et al., 2016) to enhance policy effectiveness. Additionally, increasing knowledge and awareness can augment the efficiency of incentivizing measures like subsidies (Niamir et al., 2020). In alignment with these argumentations, this study combined the variables Sentim_pos and Cap_subs. The multi-predictor PDP analysis revealed that higher values for both variables corresponded to lower energy consumption in a municipality. Moreover, the interaction effect between these variables, determined through linear regression, was negative and statistically significant at the 5% level. This implies that investments driven by SDE subsidy to reduce PC NRE consumption are indeed more effective when coupled with positive sentiment and awareness regarding RE.

Moreover, studies suggested that policies yield greater effectiveness when tailored to specific socio-demographic groups (Yang et al., 2016). Hence, the SDE subsidy-driven variable, Cap_subs, was combined with key socio-demographic variables: Density and Income, identified as the top 5 most crucial factors. Contrary to expectations from existing literature, the multi-predictor PDP demonstrated no interesting interaction effect with Cap_subs. The consumption was primarily influenced by socio-demographic factors, regardless of the capacity of SDE-subsidized investments.

Furthermore, in line with the PDP findings related to psychological factors, the multi-predictor PDP indicated that combining publications could not be generalized regardless of their sentiment to generate overall awareness and knowledge about RE. A lower number of publications with negative sentiment and a higher number of positive sentiment publications led to reduced PC NRE consumption.

Sub-question 4: High-potential municipalities

Previous SQs focused on providing a broad data overview. Aligning with literature suggestions that advocate tailoring policies to specific socio-demographic groups or behaviours (Yang et al., 2016), the focus shifted to individual municipalities, analyzing how distinct features specifically influence them. This analysis concentrates on municipalities with the top and bottom 10 PC NRE consumptions. The goal is to pinpoint municipalities with substantial improvement potential, comparing them with the best-performing ones to grasp a combination of their unique characteristics, addressing SQ4. Because the outlying values included non-mainland Dutch municipalities, the conclusions of the following analysis predominantly centre on data after outlier removal.

Most interestingly, municipalities in the top 10 PC NRE consumption had higher average household incomes and house values while showing lower average capacities for SDE subsidized investments and percentages and capacities of homes with registered solar panels compared to the mean values of the remaining data. Examples of such municipalities are Blaricum, Laren, Rozendaal, and Wassenaar. On the contrary, the bottom 10 consuming municipalities, such as Dronten, Duiven, Schouwen-Duiveland, and Veere, displayed excellent performance in terms of PC NRE consumption, with opposite trends in income, property values, and investment percentage and capacities. These findings contradict current literature suggesting that wealthier households, with the help of their financial flexibility, would outbalance their higher overall energy consumption through increased investments in RE (Frederiks et al., 2015b).

Several factors could explain the unexpectedly low investment rates in these affluent municipalities. Firstly, when looking at the SDE subsidy-driven investment variable, these results imply that economic factors might play a minimal role in these wealthy municipalities. Monetary incentives like subsidies may not be as effective in promoting RE investments. It's intriguing to note that higher energy costs per person might have a lesser impact in these regions. These wealthy municipalities possess financial flexibility; paradoxically, this might be precisely why subsidies fail to incentivize them. Additionally, there remains a possibility that residents in these municipalities might lack awareness of the benefits associated with these investments.

Given the high importance of the investment-related variables in driving PC NRE consumption, as brought forward in the feature selection, the most potential lies in increasing investment rates in these wealthy municipalities. Relating to the solar panel investment variables, as brought forward in the PDP analysis of the percentage and capacity of homes with registered solar panels revealed that the highest marginal benefits are derived within a certain range, typically from 0% or Watt-peak up to a specific threshold (2% or 50 Watt-peak, respectively). Given that the mean values of these municipalities fall within this range, these areas possess large potential for improvement by investing more in solar panels, thus effectively decreasing PC NRE consumption.

Alternative incentives might be necessary to stimulate RE investments and consequently, a lowered PC NRE consumption in these areas, which brings the focus on the psychological variables. Unfortunately, no substantial insights emerged from the psychological variables, as there were no differences in the mean values of these features between municipalities with the top high and low consumptions when compared to the mean of the remaining data. Both the top and bottom 10 municipalities had a higher number of both positive and negative publications on average compared to the rest of the data. Both sentiment-driven and monetary policy-related features do not appear to have a beneficial impact on PC NRE consumption when specifically comparing the top and bottom 10 PC NRE-consuming municipalities.

In conclusion, municipalities with the highest potential for improvement in terms of energy consumption tend to have higher average household incomes and housing values while also exhibiting below-average investments in solar panels or SDE-incentivized investments. A lot of improvement in

PC NRE consumption can be reached by increasing these below-average RE investment rates. Two reasons could explain why these municipalities have low RE investment rates resulting in higher PC NRE consumption. Initially, the SDE subsidy, operating as a monetary policy incentive, proves ineffective in regions with significant financial flexibility, where the incentive to invest is relatively too marginal. Secondly, there is a lack of awareness regarding the advantages of RE investments linked to monetary incentives. Additionally, psychological variables in this model did not exhibit notable differences in consumption patterns between these wealthy top-consuming regions and bottom-consuming municipalities. To reduce PC NRE consumption in these high-consuming areas, enhancing awareness about the benefits of RE investments is crucial. Additionally, exploring alternative individual core values or implementing different monetary policy measures becomes essential.

Conclusion

Recommendations

Upon arriving at the final answer to the research question and deriving actionable insights from the results, several recommendations can be made for both individuals and the government as stakeholders in the Netherlands concerning Dutch PC NRE consumption in the residential sector.

Regarding the number of publications within a municipality, it was evident that solely positive publications had a beneficial impact on RE consumption, leading to a reduction in PC NRE consumption. On the contrary, negative publications could increase PC NRE consumption, while publications with a neutral sentiment did not substantially affect PC NRE consumption when compared to other variables. Hence, it is crucial to promote positive publicity about RE to decrease PC NRE consumption and consequently contribute to the RET. Individuals aiming to contribute should create positive publicity surrounding RE through word-of-mouth, social media or other platforms and refrain from negative comments about RE. This recommendation also extends to the Dutch government which can address this issue on a larger scale through awareness campaigns focusing on the positive aspects of various RE technologies. The aim is to cultivate a positive perception not only of widely accepted technologies but also of controversial ones such as nuclear power or biomass.

This effect becomes even more pronounced when combined with regions characterized by substantial SDE subsidy-driven investments. The quantity of positive publications appears to significantly amplify the negative relationship of SDE-subsidized investment capacity on PC NRE consumption. Conversely, SDE-subsidized investment capacity magnifies the negative relationship of positive publications on PC NRE consumption. Positive publicity proves to be most impactful in regions with a higher prevalence of incentive-driven entrepreneurship in the field of renewable energy, or vice versa. Consequently, a specific recommendation emerges for the government. To decrease PC NRE consumption and contribute to the RET, the Dutch government should focus their positive awareness campaigns on RE technologies in regions with the highest levels of incentive-driven entrepreneurship in the RE sector. Alternatively, the government could monitor media sources to ensure positive sentiment of RE within a region and stimulate RE entrepreneurship through campaigns highlighting the possible subsidies available to entrepreneurs.

Furthermore, the findings indicate that municipalities with higher average incomes and property values, coupled with limited investments in RE technologies, hold significant potential for improvement. These areas could greatly benefit from increased investments in RE technologies, especially given their financial flexibility. A clear recommendation would be for individuals in municipalities with low investment rates in RE and higher financial flexibility to play an active role in the RET. They could invest more in RE technologies or conserve more energy, contributing to the reduction of PC NRE consumption. However, the situation might be more complex, and there could be underlying reasons behind these low investment rates. Firstly, there might be a lack of awareness regarding the benefits of RE investments in these municipalities. Hence, a recommendation is made for the government to enhance awareness about subsidies through targeted campaigns, specifically focusing on high-income municipalities. These campaigns should highlight the advantages available to entrepreneurs, aiming to stimulate RE investments. Another recommendation emerges from the consideration that the SDE subsidy, a form of monetary policy, might not be effective for wealthier municipalities. The government should employ alternative incentivizing mechanisms, tailored specifically for wealthy regions, to boost RE investments.

Limitations & future research

This study encountered limitations, that restricted interpretations and data explorations. Firstly, the lack of information regarding whether solar panel investments were directly influenced by subsidies hindered drawing conclusions related to monetary policies. While this variable indicated household-level RE investments, it was more robust for individual conclusions compared to the SDE subsidy-driven investment variable, which was associated with companies or non-profit organizations instead of regular households. Another limitation was following the methodology and assumptions of Klimaatmonitor. Specifically, assuming equal RE consumption and production within a municipality occasionally led to peculiar outcomes, such as negative values in the dependent variable's minimum value.

A challenge in this study, to be tackled in future research, was the availability of publication data for evaluating psychological variables using the selected keywords. Primarily, the research relied on a Meltwater subscription utilized by an employee at Meltwater, limiting the accessible data. Consequently, ideas generated later in the project couldn't be incorporated into the research. Additionally, numerous municipalities lacked publications, complicating data analysis. Although the chosen keywords were derived from specific sources, a more precise subset of keywords capturing terms relevant to the energy transition could be explored in future studies. Concepts like energy storage, smart grids, energy neutrality, and biofuels, not clearly defined in existing sources, could enrich the psychological variables data, enhancing granularity. Additionally, exploring alternative forms of psychological indicators in combination with data science methods could provide valuable insights.

Another interesting road for future research is to explore developing countries in the Global South, as they hold the most potential for improvement in the context of RE consumption also highlighted by Cantarero (2020). Data in this research was confined to the Dutch region but to make the model more general and understand the factors influencing RET in various regions or specific types of regions, incorporating data from other countries, as suggested by Niamir et al (2020), could be valuable.

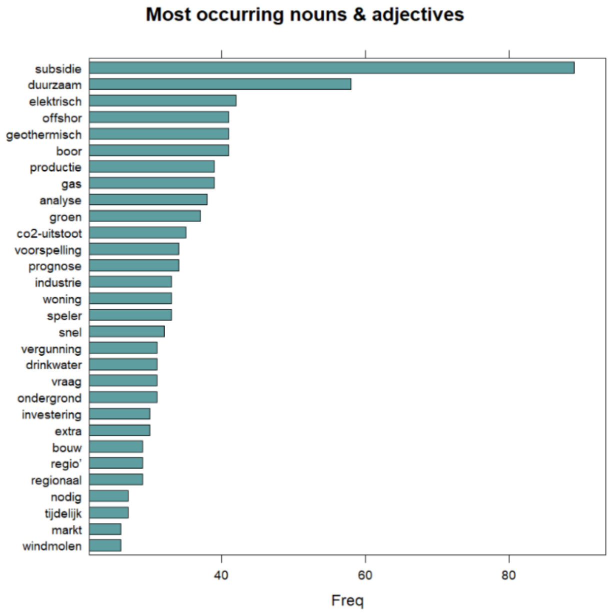
Utilizing a path less commonly taken for policy recommendations to address the RET problem such as data science, illustrates that every individual can play a role in the RET, even if they consider the field as irrelevant. Collaborative efforts can make the transition from fossil fuels to RE more achievable.

Appendix

Appendix A - Graphs

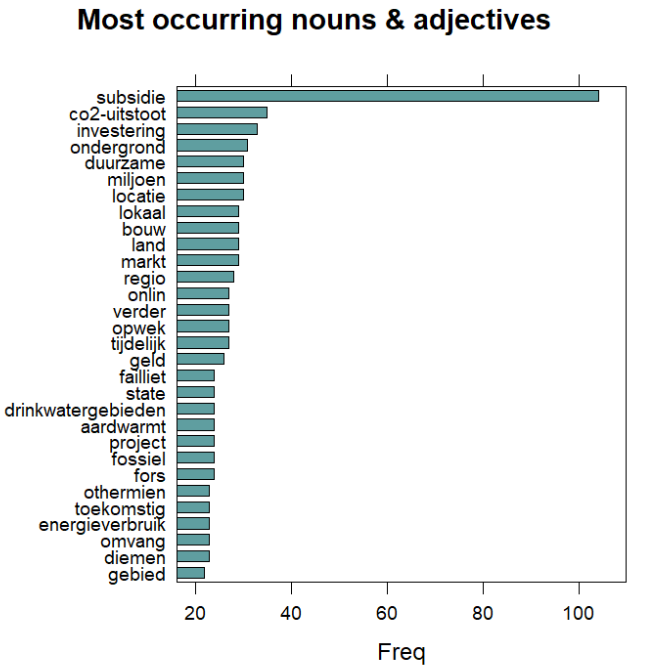
Appendix A. 1

The 30 most occurring nouns and adjectives after the first iteration



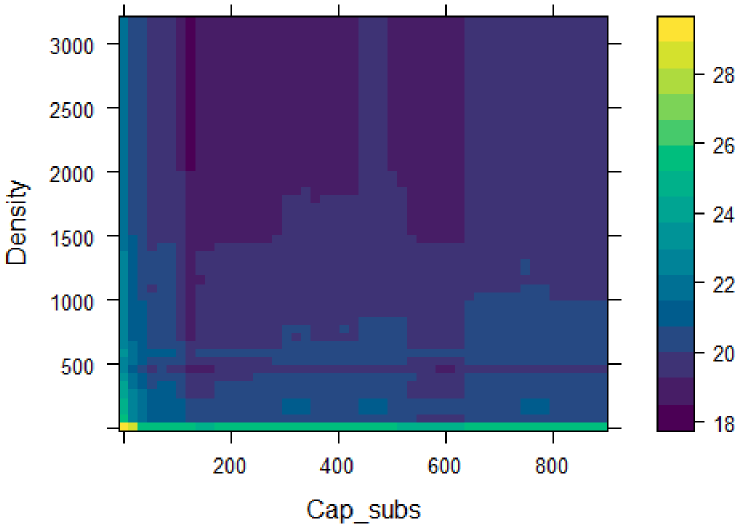
Appendix A. 2

The 30 most occurring nouns and adjectives after the first iteration



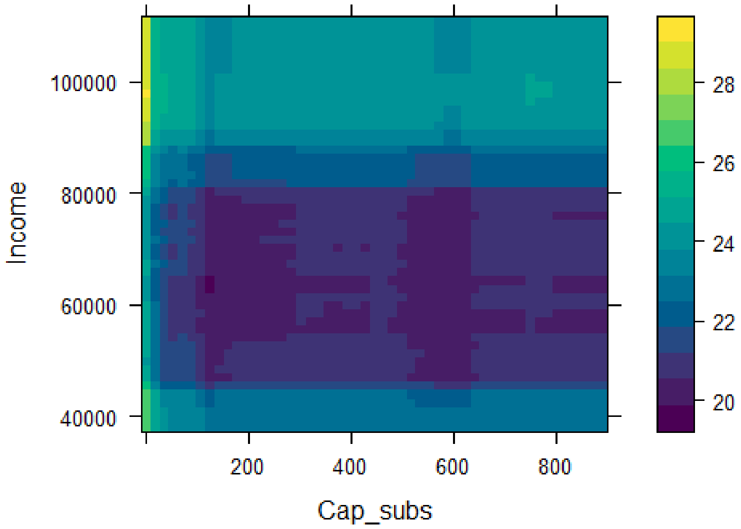
Appendix A. 3

Multi-predictor PDP of Density and Cap_subs



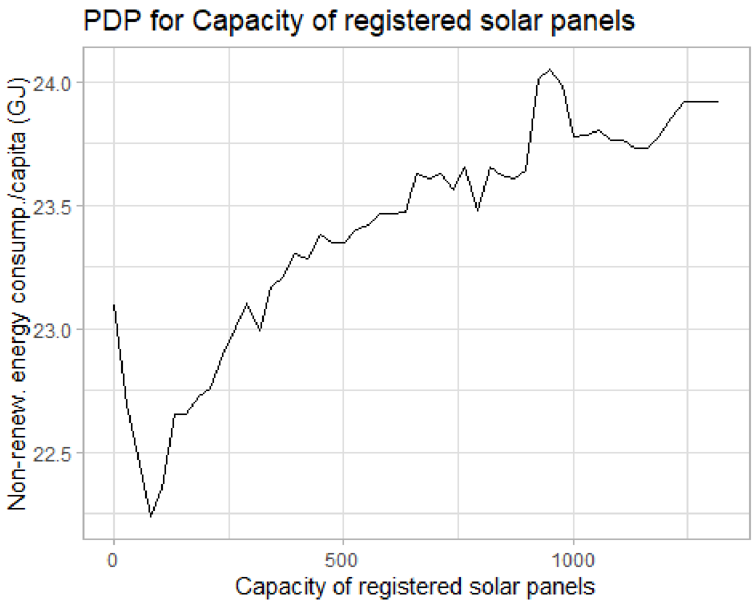
Appendix A. 4

Multi-predictor PDP of Income and Cap_subs



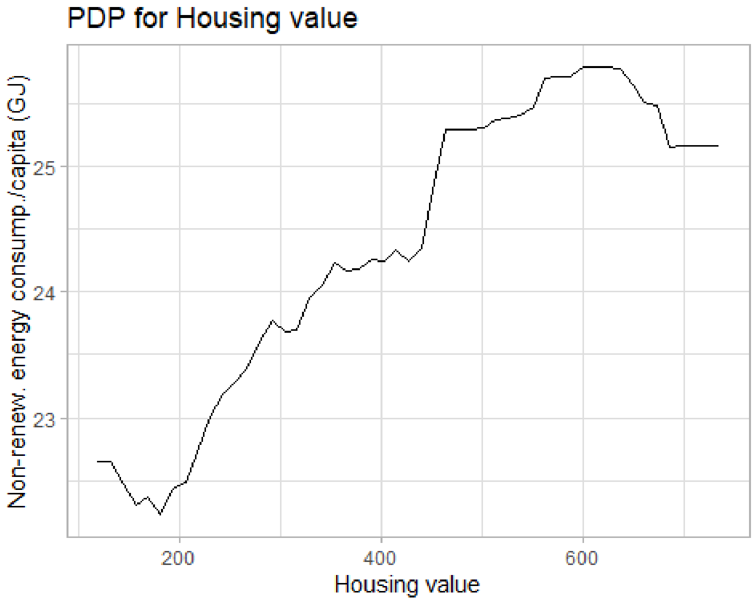
Appendix A. 5

PDP for Cap_solar after removing outliers



Appendix A. 6

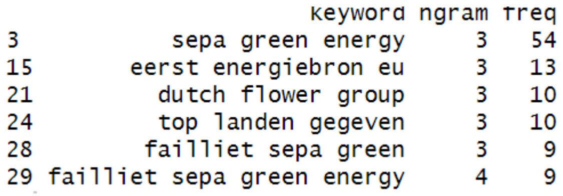
PDP for WOZ_value after removing outliers



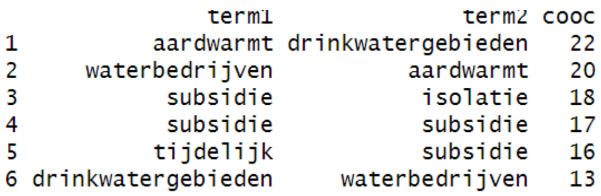
Appendix B - Tables

Appendix B.1

Most occurring co-ocurrences and phrases



```
                     keyword ngram freq
3             sepa green energy     3   54
15        eerst energiebron eu      3   13
21           dutch flower group     3   10
24           top landen gegeven     3   10
28          failliet sepa green     3    9
29 failliet sepa green energy     4    9
```



```
               term1             term2 cooc
1          aardwarmt drinkwatergebieden   22
2     waterbedrijven          aardwarmt   20
3           subsidie           isolatie   18
4           subsidie           subsidie   17
5          tijdelijk           subsidie   16
6 drinkwatergebieden     waterbedrijven   13
```

Appendix C – Meltwater keyword search

Appendix C.1

Boolean code

ONLY NL but dutch and english as language with news as source

- **Search 1.1.1 - RET "positive (solutions to the problem)" keyword News search:** (sentiment: "positive" OR sentiment: "neutral") AND (title:"renewable energy" OR title:"hernieuwbare energie" OR title:"hernieuwbare energietransitie" OR title:"solar energy" OR title:"zonne-energie" OR title:"wind energy" OR title:windenerg* OR title:biomass* OR title:bioenerg* OR title:"hydro energy" OR title:hydroelec* OR title:"hydropower" OR title:waterkracht OR title:"geothermal energy" OR title:aardwarmte* OR title:geotherm* OR title:"tidal energy" OR title:"tidal power" OR title:getijdenenergie* OR title:"ocean power" OR title:"golf energie" OR title:"heat pump" OR title:"warmtepomp" OR title:"energy efficiency" OR title:"energie efficientie")
  - Language Dutch (same as removing all the english words)
  - New words based on text analysis: **Groen -> groene energie, duurzaam -> duurzame energie, CO2-uitstoot, windmolen,**
  - **Refresh the search -> search 1.1.2**

Add new phrases or keywords from the text analysis (co-occurrences) using R

- **Search 1.1.2 - RET "positive (solutions to the problem)" keyword search:** (sentiment: "positive" OR sentiment: "neutral") AND (title:"renewable energy" OR title:"hernieuwbare energie" OR title:"hernieuwbare energietransitie" OR title:"solar energy" OR title:"zonne-energie" OR title:"wind energy" OR title:windenerg* OR title:biomass* OR title:bioenerg* OR title:"hydro energy" OR title:hydroelec* OR title:"hydropower" OR title:waterkracht OR title:"geothermal energy" OR title:aardwarmte* OR title:geotherm* OR title:"tidal energy" OR title:"tidal power" OR title:getijdenenergie* OR title:"ocean power" OR title:"golf energie" OR title:"heat pump" OR title:"warmtepomp" OR title:"energy efficiency" OR title:"energie efficientie" OR title:"groene energie" OR title:"green energy" OR title:"duurzame energie" OR title:"sustainable energy” OR title:windmolen* OR title:”windmill” OR title:”duurzame energiebron” OR title:”hernieuwbare energiebron” OR title:”groene energiebron” OR title:”duurzame energiebron”)

- After search 2 we can see that subsidy is one of the highest (#4) occurring keywords -> raises awareness of the available subsidies. Kept that keyword in and removed the rest of the keywords that are not relevant to the energy transition (in the top 30), are already part of one of the previous keywords or do not form a co occurrence with other words that creates a relevant phrase
- Only new keyword is “energiebron” -> “duurzame energiebron” / “hernieuwbare energiebron”. These are added in the final search in dark blue shown above. & CO2-uitstoot -> this can be used for the “negative sentiment”

- **Search 1.1.3 - Final DUTCH RET search (10,4k)**: (sentiment: "positive" OR sentiment: "neutral") AND (title:“renewable energy” OR title:“hernieuwbare energie” OR title:“hernieuwbare energietransitie” OR title:”solar energy” OR title:”zonne-energie” OR title:“wind energy” OR title:windenerg* OR title:biomass* OR title:bioenerg* OR title:”hydro energy” OR title:hydroelec* OR title:”hydropower” OR title:waterkracht OR title:“geothermal energy” OR title:aardwarmte* OR title:geotherm* OR title:“tidal energy” OR title:”tidal power” OR title:getijdenenergie* OR title:”ocean power” OR title:”golf energie” OR title:”heat pump” OR title:”warmtepomp” OR title:”energy efficiency” OR title:”energie efficientie” OR title:”groene energie” OR title:”green energy” OR title:”duurzame energie” OR title:”sustainable energy” OR title:windmolen* OR title:”windmill” OR title:”duurzame energiebron” OR title:”hernieuwbare energiebron” OR title:”groene energiebron” OR title:”duurzame energiebron”)
  - Deze zoekterm maar dan zonder titel levert voor 2022 de volgende nieuwe zoektermen op gekeken naar de top 50 most frequent words in de titels: windpark, zonnepark, waterstof, windturbine, kernenergie
  - title:hydrogen* OR title:waterstof*
- **Search 1.1.4 - Final DUTCH RET search with negative words (13.5k results):** title:coal OR title:steenkool* OR title:”natural gas” OR title:aardgas* OR title:“fossil fuel” OR title:“fossiele brandstof” OR title:”greenhouse gas” OR title:GHG OR title:broeikasgas* OR title:“renewable energy” OR title:“hernieuwbare energie” OR title:“hernieuwbare energietransitie” OR title:”solar energy” OR title:”zonne-energie” OR title:“wind energy” OR title:windenerg* OR title:biomass* OR title:bioenerg* OR title:”hydro energy” OR title:hydroelec* OR title:”hydropower” OR title:waterkracht OR title:“geothermal energy” OR title:aardwarmte* OR title:geotherm* OR title:“tidal energy” OR title:”tidal power” OR title:getijdenenergie* OR title:”ocean power” OR title:”golf energie” OR title:”heat pump” OR title:”warmtepomp” OR title:”energy efficiency” OR title:”energie

efficientie” OR title:”groene energie” OR title:”green energy” OR title:”duurzame energie” OR title:”sustainable energy” OR title:windmolen* OR title:”windmill” OR title:”duurzame energiebron” OR title:”hernieuwbare energiebron” OR title:”groene energiebron” OR title:”duurzame energiebron”

Appendix C.2

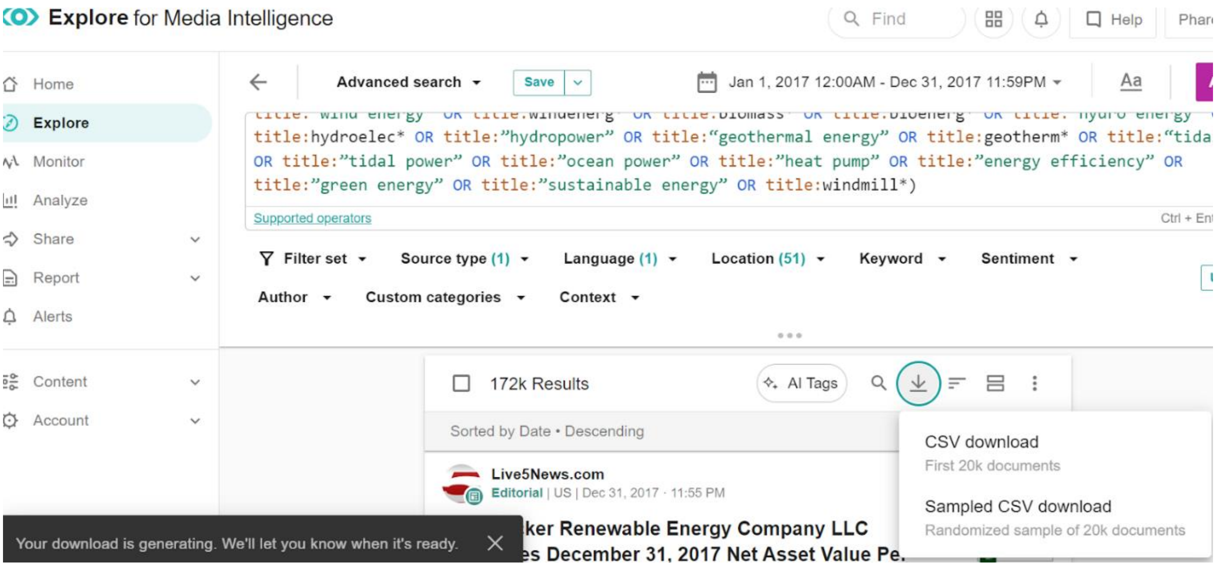
Ellaboration of the searches

Search 1.2:

- Zhang, Y., Abbas, M., & Iqbal, W. (2022). Perceptions of GHG emissions and renewable energy sources in Europe, Australia and the USA. *Environmental Science and Pollution Research*, *29*(4), 5971-5987. : Use mentioned rate in twitter and search interest in google search. The keywords they used are #greenhouse gas”, “#GHG” and “#renewable energy”, “#coal”, “#natural gas”, “#solar energy”, “#wind energy”, “#biomass”, “#hydro energy”, “#geothermal energy”, “#tidal energy. They use word colocation analysis. They compare this with google trends for each source of energy: **Tidal energy, hydro energy, biomass energy, geothermal energy, solar energy, wind energy**, coal, natural gas. Also investigate **word collocation.**
- New terms based on link below: hydropower/hydroelectricity, bioenergy, heating, heat pumps (warmte pomp), “energy efficiency” (“energie efficientie”), ocean power/tidal power/ energy -> The terms are based on technologies surrounding the renewable energy transition as are given in https://www.iea.org/fuels-and-technologies/renewables

Appendix C.3

Example of the Meltwater process



```
title:"wind energy" OR title:windenerg* OR title:biomass* OR title:bioenerg* OR title:"hydro energy"
title:hydroelec* OR title:”hydropower” OR title:“geothermal energy” OR title:geotherm* OR title:“tida
OR title:”tidal power” OR title:”ocean power” OR title:”heat pump” OR title:”energy efficiency” OR
title:”green energy” OR title:”sustainable energy” OR title:windmill*)
```

Jan 1, 2017 12:00AM - Dec 31, 2017 11:59PM

Source type (1) Language (1) Location (51)

172k Results

Sorted by Date • Descending

Live5News.com
Editorial | US | Dec 31, 2017 · 11:55 PM

ker Renewable Energy Company LLC
es December 31, 2017 Net Asset Value Pe.

CSV download
First 20k documents

Sampled CSV download
Randomized sample of 20k documents

Your download is generating. We'll let you know when it's ready.

Reference list

Abdullah, K. (2014). National renewable energy policy and action plan: highlights and updates. *Applied Mechanics and Materials*, *465*, 275-279.

Abrahamse, W., & Steg, L. (2011). Factors related to household energy use and intention to reduce it: The role of psychological and socio-demographic variables. *Human ecology review*, 30-40.

Arkesteijn, K., & Oerlemans, L. (2005). The early adoption of green power by Dutch households: An empirical exploration of factors influencing the early adoption of green electricity for domestic purposes. Energy Policy, 33(2), 183-196.

Bayulgen, O. (2020). Localizing the energy transition: Town-level political and socio-economic drivers of clean energy in the United States. *Energy Research & Social Science*, *62*, 101376

Bennett, J., Baker, A., Johncox, E., & Nateghi, R. (2020). Characterizing the key predictors of renewable energy penetration for sustainable and resilient communities. *Journal of Management in Engineering*, *36*(4), 04020016.

Bisaga, I., Parikh, P., Tomei, J., & To, L. (2020). Mapping synergies and trade-offs between energy and the sustainable development goals: A case study of off-grid solar energy in Rwanda. Energy Policy.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Cantarero, M. M. V. (2020). Of renewable energy, energy democracy, and sustainable development: A roadmap to accelerate the energy transition in developing countries. *Energy Research & Social Science*, *70*, 101716.

Carley, S., & Konisky, D. M. (2020). The justice and equity implications of the clean energy transition. *Nature Energy*, *5*(8), 569-577.

Chen, C. F., Xu, X., & Day, J. K. (2017). Thermal comfort or money saving? Exploring intentions to conserve energy among low-income households in the United States. *Energy Research & Social Science*, *26*, 61-71.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.

Climate Change 2021: The Physical Science Basis. (2021). IPCC. https://www.ipcc.ch/report/ar6/wg1/

Doel en geschiedenis. Regionale Klimaatmonitor. (2023). Retrieved August 29, 2023, from https://klimaatmonitor.databank.nl/content/doel-geschiedenis

Energy system. International Energy Agency. (n.d. - b). Retrieved September 08, 2023, from https://www.iea.org/energy-system

Final consumption. International Energy Agency. (n.d. - a). Retrieved September 08, 2023, from https://www.iea.org/reports/key-world-energy-statistics-2021/final-consumption

Frederiks, E. R., Stenner, K., & Hobman, E. V. (2015a). Household energy use: Applying behavioural economics to understand consumer decision-making and behaviour. *Renewable and Sustainable Energy Reviews*, *41*, 1385-1394.

Frederiks, E. R., Stenner, K., & Hobman, E. V. (2015b). The socio-demographic and psychological predictors of residential energy consumption: A comprehensive review. *Energies*, *8*(1), 573-609.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gassar, A. A. A., Yun, G. Y., & Kim, S. (2019). Data-driven approach to prediction of residential energy consumption at urban scales in London. *Energy*, *187*, 115973. (53 citations)

Goldstein, B., Gounaridis, D., & Newell, J. P. (2020). The carbon footprint of household energy use in the United States. *Proceedings of the National Academy of Sciences*, *117*(32), 19122-19130.

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

Hawkins, D., Hong, S. M., Raslan, R., Mumovic, D., & Hanna, S. (2012). Determinants of energy use in UK higher education buildings using statistical and artificial neural network methods. *International Journal of Sustainable Built Environment*, *1*(1), 50-63.

Kabeyi, M. J. B., & Olanrewaju, O. A. (2022). Sustainable energy transition for renewable and low carbon grid electricity generation and supply. *Frontiers in Energy Research*, *9*, 1032.

Kannangara, M., Dua, R., Ahmadi, L., & Bensebaa, F. (2018). Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning appoaches. *Waste management*, *74*, 3-15.

Kharas, H., Fengler, W., Sheoraj, R., Vashold, L., & Yankov, T. (2022, November 29). *Tracking emissions by country and sector | Brookings*. Brookings. https://www.brookings.edu/articles/tracking-emissions-by-country-and-sector/

Klimaatmonitor. (n.d.). https://klimaatmonitor.databank.nl/jive

Lu, Y., Khan, Z. A., Alvarez-Alvarado, M. S., Zhang, Y., Huang, Z., & Imran, M. (2020). A critical review of sustainable energy policies for the promotion of renewable energy sources. *Sustainability*, *12*(12), 5078.

Marijnissen, H., & Straver, F. (2020, March 3). *Als enige gemeente in Nederland lukt het Zeewolde om energieneutraal te zijn*. Trouw. https://www.trouw.nl/binnenland/als-enige-gemeente-in-nederland-lukt-het-zeewolde-om-energieneutraal-te-zijn~bd7b1f97/?referer=https%3A%2F%2Fwww.google.com%2F

Media Monitoring & Analysis | Meltwater. (n.d.). Meltwater. https://www.meltwater.com/en/products/media-monitoring

Niamir, L., Ivanova, O., Filatova, T., Voinov, A., & Bressers, H. (2020). Demand-side solutions for climate mitigation: Bottom-up drivers of household energy behavior change in the Netherlands and Spain. *Energy Research & Social Science*, *62*, 101356.

Nicolini, M., & Tavoni, M. (2017). Are renewable energy subsidies effective? Evidence from Europe. *Renewable and Sustainable Energy Reviews*, *74*, 412-423.

Ouyang, X., & Lin, B. (2014). Impacts of increasing renewable energy subsidies and phasing out fossil fuel subsidies in China. *Renewable and sustainable energy reviews*, *37*, 933-942.

Pacthod, D., Pinner, D., Polymeneas, E., Samandari, H., Tai, H., Bolano, A., Lodesani, F., & Pratt, M. P. (2022). The energy transition: A region-by-region agenda for near-term action. In *McKinsey & Company*. https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/the-energy-transition-a-region-by-region-agenda-for-near-term-action

Poruschi, L., & Ambrey, C. L. (2019). Energy justice, the built environment, and solar photovoltaic (PV) energy transitions in urban Australia: A dynamic panel data analysis. *Energy Research & Social Science*, *48*, 22-32.

Qadir, S. A., Al-Motairi, H., Tahir, F., & Al-Fagih, L. (2021). Incentives and strategies for financing the renewable energy transition: A review. *Energy Reports*, *7*, 3590-3606.

Rijksdienst voor Ondernemend Nederland (2012, August 2). *Feiten en cijfers SDE(+)(+)*. RVO.nl. Retrieved August 29, 2023, from https://www.rvo.nl/subsidies-financiering/sde/aanvragen/feiten-en-cijfers

Rijksdienst voor Ondernemend Nederland (2020, May 15). *SDE++: Oriënteren*. RVO.nl. Retrieved August 29, 2023, from https://www.rvo.nl/subsidies-financiering/sde/orienteren

Ritchie, H. (2020). Sector by sector: where do global greenhouse gas emissions come from?. *Our World in data*.

Ritchie, H., Roser, M., & Rosado, P. (2022) - "*Energy*". OurWorldInData.org. Retrieved September 13, 2023, from: https://ourworldindata.org/energy

Royston, P. (1995). Remark AS R94: A remark on algorithm AS 181: The W-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *44*(4), 547-551.

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and sustainable energy reviews*, *13*(8), 1819-1835.

United Nations Framework Convention on Climate Change. (2015). Paris Agreement. https://unfccc.int/sites/default/files/english_paris_agreement.pdf

Wall, W. P., Khalid, B., Urbański, M., & Kot, M. (2021). Factors influencing consumer's adoption of renewable energy. *Energies*, *14*(17), 5420.

Whitney W. (n.d. - a). *How is sentiment assigned?.* Meltwater Help Center. Retrieved August 30, 2023, from **https://help.meltwater.com/en/articles/4064558-how-is-sentiment-assigned**

Whitney W. (n.d. - b). *How is article reach measured?*. Meltwater Help Center. Retrieved August 30, 2023, from https://help.meltwater.com/en/articles/4064552-how-is-article-reach-measured

Wiedenhofer, D., Lenzen, M., & Steinberger, J. K. (2013). Energy requirements of consumption: Urban form, climatic and socio-economic factors, rebounds and their policy implications. *Energy policy*, *63*, 696-707.

Wiesmann, D., Azevedo, I. L., Ferrão, P., & Fernández, J. E. (2011). Residential electricity consumption in Portugal: Findings from top-down and bottom-up models. *Energy policy*, *39*(5), 2772-2779.

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Xue, J., Gong, R., Zhao, L., Ji, X., & Xu, Y. (2019). A green supply-chain decision model for energy-saving products that accounts for government subsidies. *Sustainability*, *11*(8), 2209.

Yang, S., Zhang, Y., & Zhao, D. (2016). Who exhibits more energy-saving behavior in direct and indirect ways in China? The role of psychological factors and socio-demographics. *Energy Policy*, *93*, 196-205.

Yang, S., & Zhao, D. (2015). Do subsidies work better in low-income than in high-income families? Survey on domestic energy-efficient and renewable energy equipment purchase in China. *Journal of Cleaner Production*, *108*, 841-851.

Zhang, Y., Abbas, M., & Iqbal, W. (2022). Perceptions of GHG emissions and renewable energy sources in Europe, Australia and the USA. *Environmental Science and Pollution Research*, *29*(4), 5971-5987.

Zhao, H. X., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586-3592.