Final Report

Creating an energy characteristics database for individual dwellings in the Netherlands

PBL Consultancy Team

Utrecht University

Utrecht University

Final report PBL consultancy group

Creating an energy characteristics database for individual dwellings in the Netherlands

Utrecht University
Energy Science
Consultancy project (GEO4-2519)

June 30, 2021

14 381 words

Supervisor:	dr. ir. Ioannis Lampropoulos	(UU)
Client representatives:	Folckert van der Molen, BSc	(PBL)
	drs. ing. Casper Tigchelaar	(TNO)

Willie van Schalkwyk – 2306861
Erik van Battum – 5715563
Thomas Brunekreef – 5525667
Rens Baardman – 7289340
Mark van der Boor – 6208541
Wessel Poorthuis – 5726034

Contents

List of Figures

List of Tables

1 | Executive summary

As stipulated in the Dutch climate agreement, to avoid an increase of 1.5 °C compared to pre-industrial times, the Netherlands has set out to reduce its greenhouse gas emissions by 49% of 1990 emissions by 2030. The built environment is responsible for more than half of the Netherlands' energy use and greenhouse gas emissions, which makes it a prime target for improvement (Anderson et al., 2015). One of the agencies charged with providing scientific studies for strategic policy analysis is PBL (*Planbureau voor de Leefomgeving*, Environmental Assessment Agency) and a current focus area for them is the residential sector. Part of their analysis consists of testing various policy scenarios by using predictive models. To do so, PBL requires the consolidated energy characteristics of all dwellings in the Netherlands. There are various datasets available, but they are incomplete, uncorrelated, and not at the required spatial scale. That is why we have been tasked with producing a consolidated dataset on an individual dwelling level that fully describes the energy characteristics of all Dutch dwellings.

Therefore, our research question is as follows: *What are the energy characteristics of specific dwellings in the Netherlands?* To answer this question, we set about defining what a dwelling is and what are its relevant energy characteristics. We identified the instruments available for reducing a dataset's spatial scale and to generate missing data. Furthermore, we developed a Python code that coupled with PostgreSQL can import, house, and retrieve large data sets. Finally, we developed algorithms in Python that apply the aforementioned instruments to ultimately produce the required data points.

A dwelling was defined as a building or part of a building that has a residential function, where one or more persons may live. Six dwelling energy characteristics were identified as relevant. Namely, space heating, water heating, space cooling, insulation, cooking and ventilation. However, only space heating, water heating, insulation and cooking were analysed. Dwelling energy labels are also important indicators in the enrgy use of a dwelling, so attention was given to them too. Lighting and appliances were not included due to their highly circumstantial presence. Human behaviour is also a factor needing to be considered, but this module was not fully developed and can be referenced in Appendix B.

To develop our consolidated output, we had to source, transform, reduce, and integrate data from multiple sources. Our primary data sources were the BAG, RVO, WoON, and CBS. Various methods were developed to predict the different energy characteristics. For space heating, water heating and cooking we reduced the spatial scale of available data through various averaging methods and extrapolated known data points through comparisons with national usage benchmarks. For energy labels and insulation characteristics we applied statistical instruments such as multiple linear regression to find correlations between different data points, allowing us to aggregate various datasets and thereby produce missing data points at our required spatial scale. To do this successfully, several assumptions had to be made. The extent to which we were able to predict or extrapolate information varied significantly from one dataset to the next. Because of this, we used descriptive statistics as well as confidence and prediction intervals to determine the accuracy of certain data points.

To finally combine all the different energy characteristics and their determination methods, the output data was generated through a data pipeline in Python. This program takes dwelling data from the BAG and, by passing it through different modules of comprised of algorithms, it generates data about the energy characteristics of these dwellings. Our space heating, water heating and cooking modules produces the predicted presence of a specific space heating, water heating and cooking device. The insulation module produces an estimated thermal resistance (R-value) for floors, roofs, windows, and the façade while the energy label module assigns each dwelling an EPI (*energieprestatieindex*, energy performance index) and an energy label class.

Due to a lack of comparative data, a comprehensive validation of the created dataset was not possible. A partial validation could be performed on the space heating, water heating and insulation modules based on a small sample of similar data found in an online map viewer. Overall, the space and water heating characteristics were found to be 67% and 66% accurate, respectively. Insulation R-value predictions were overall 64% accurate. More extensive validations can be performed if the energy label input data is made available by RVO.

Further research into the use of downscaling as a spatial reduction technique and its application to our research might also improve the accuracy of our heating modules. Furthermore, the space cooling, ventilation and human behaviour modules needs to be developed still.

2 | Introduction

2.1 Background of research

Major action on every level of our society is required to keep global warming beneath an increase of 1.5 °C compared to pre-industrial temperatures (Hoegh-Guldberg et al., 2018). In this light, the Dutch government has set its goal on reducing greenhouse gas emissions by 49% of 1990 emissions in 2030 (Rijksoverheid, 2019). Policy concerning the built environment, which is responsible for more than half of energy use and greenhouse gas emissions, is a major factor in the mitigation of climate change (Anderson et al., 2015).

The PBL (*Planbureau voor de Leefomgeving*, Environmental Assessment Agency) is the national environmental assessment agency in the Netherlands and provides scientific studies for strategic policy analysis on, inter alia, built environment energy use. These studies are conducted through the use of models like VESTA MAIS (van der Molen et al., 2021). To be calibrated and to produce results, these models need data. The quality of the assessment increases with the amount of data and the level of detail provided (Janssen & Heuberger, 1995).

As of now, data related to energy use in the built environment in the Netherlands can be found on multiple spatial scales. Data is provided at the level of the entire country, municipalities, postal codes and individual dwellings. These are datasets concerning the electricity and gas use of buildings and their energy labels and are provided by grid operators and governmental agencies. Every few years the WOON survey (*WoonOnderzoek Nederland*, Residential Research Netherlands) is conducted by CBS (*Centraal Bureau voor de Statistiek*, Central Bureau of Statistics). This survey provides detailed information about a particular subset of homes (CBS, 2018), like the insulation of different components, presence and type of heating installation and energy related behaviour.

However, currently there is no dataset available where all this information is aggregated and presented at the dwelling level. Either the data is presented at a higher spatial level, or if it is dwelling based, the dataset does not incorporate all dwellings in the Netherlands. Such a dataset could increase the accuracy of the predictions made by the models of PBL and serve as a basis for further additions of more datasets. Because the dataset will be open access, others, not affiliated with PBL, will also be able to use the data to do their research. This will increase the knowledge of how we use energy in the built environment and where and how improvements can be be made, giving us handholds in the transition to a more sustainable society.

2.2 Problem definition

Given the fact that no dwelling level dataset of energy related characteristics in the built environment exists, we can identify the aim of this research, which is to identify the main energy related characteristics of individual dwellings and to design a method to build a dataset with these characteristics for individual dwellings. This leads to our research question:

What are the energy characteristics of specific dwellings in the Netherlands?

In order to answer this question, we pose the following sub-questions:

1. *What are relevant energy characteristics?*
2. *What relevant datasets are available, and what is their spatial level, time period and completeness?*
3. *What are available instruments for reducing spatial scale in the datasets?*
4. *What are available instruments for generating missing data?*
5. *How can we aggregate the data into one large dataset?*

We define energy characteristics as those properties of a dwelling that influence the energy use of a dwelling providing its basic functions. These functions are space heating, water heating, space cooling,

insulation, cooking and ventilation. Examples of characteristics are the building type, presence of a heat pump and access to district heating. We will not look at energy use from appliances like washing machines, tumble dryers, and small electrical devices.

The term dwelling also needs some clarification. A dwelling is a building or part of a building that has a residential function, where one or more persons may live. A technical definition is given in Section 3.2.1, while the different possible types of dwellings are given in Section 3.4.

In Chapter 3, the theoretical background, the sub-questions posed above are answered. First, background information on the energy functions of a dwelling will be presented. In this section, the energy characteristics that are the subject of this research will be elaborated on. Afterwards, an overview of the relevant spatial scales is given and a discussion of the datasets used is provided. The two sections afterwards deal with the different building types present in the datasets and the energy labels that dwellings can have. Finally, a short introduction to the statistical instruments used to manipulate the data is provided.

In the methodology, Chapter 4, information on the data pipeline, which generates the results, is given. This pipeline comprises many different modules, which calculate the probability distributions for each energy characteristic that is researched. Information on how these modules work is given in the results section. The methodology also offers information about the statistical methods used to find correlations between and within datasets, which are used to link information.

In Chapter 5, the results of the research are presented. This comprises, for every module, a write up about the workings of the module, the data that comes in and the data that comes out. We choose to show the workings of the modules in the results section instead of in the methodology section, because we believe that the way the data is used is a more important result of this research than the dwelling level data the pipeline produces.

Finally, in Chapter 6 the validity and scope of the results is discussed. We will draw conclusions from these results and give recommendations for future research concerning the construction of this type of dataset.

3 | Theoretical background

In the Netherlands, the final energy use in dwellings for 2018 reported by CBS was 404 PJ, of which over 70% was provided by natural gas (2020). According to the RVO (*Rijksdienst voor Ondernemend Nederland*, Netherlands Enterprise Agency), applications such as heating and cooking account for 75% of the total final energy use in dwellings and generally overlap with the energy use provided by natural gas (2019). The final 25% of energy use in dwellings is attributed to energy use in lighting and other household electrical appliances (RVO, 2019).

An important difference between the energy use in primary applications and other energy use is that the energy use from primary applications is closely related to building specifications, whereas energy use in electrical appliances is often more dependent on the behaviour of the residents and hence more variable. In this research we have looked at the primary applications present in dwellings.

3.1 Energy characteristics

In order to create a database that describes the energy characteristics of specific dwellings in the Netherlands, it is first necessary to fully describe how energy is or can be used in a dwelling. We identified six categories that would allow us to classify different technologies under each energy consumption category (Nationaal Expertisecentrum Warmte, 2013).

1. Heating
2. Insulation
3. Ventilation
4. Cooking
5. Cooling
6. Human Behaviour

The classification of energy consumption as such can be expanded or narrowed down depending on the scope of the project. For this project we described three levels, the functional demand, the technology that fulfils the demand, and the energy consumption by the technology implemented. This is illustrated in Figure 3.1. The functional demand is something that we would expect to see in every dwelling, and the technologies and final energy carrier are outlined as options of what you could find in a household. The dataset we created consists of an index of which technologies are present in a household and what type of final energy carrier is required. In order to understand the prevalence of these technologies better, we will explore what the current state is of these technologies in the Netherlands.

Within heating, there are two aspects that must be considered, namely heating of space and heating of water. Space heating is mainly produced through the use of natural gas by boilers and stoves. There are a number of renewable alternatives, one of which is biomass burning. This comprises roughly 4% of total energy consumption for heat in the built environment and consists of the burning of wood (Segers et al., 2020). However, this last figure has an uncertainty of 30% (Segers et al., 2020). Furthermore, there is a growing consumption of electricity to produce heat through heat pumps and the pumps in boilers. Houses that use natural gas for heating, dedicate roughly 20% of that heat to heating tap water (Segers et al., 2020).

Another important aspect to consider when describing a dwelling's heating system is its access to district heating. Around 5.9% of dwellings in the Netherlands are connected to a district heating system (CBS, 2020). District heating systems are mostly located in densely populated areas, where they have the greatest potential to increase heating efficiency over individual heating systems (Segers et al., 2020). This increased efficiency could lead to lower overall energy use for heating of neighbourhoods (Werner, 2013).

Insulation is an important parameter as it affects how much energy is needed to maintain the interior temperature of a dwelling. CBS indicates that in 2018, 77% of building components' surfaces were

insulated in the Netherlands. This value shifts for private dwellings (80%), rental dwellings (73%) and private rental homes (61%) (CBS et al., 2020). This data was collected through the WoON survey, further described in section 3.3.2. The degree of insulation does not account for the thickness or the quality of insulation. This means that if a dwelling has a 100% degree of insulation, improvement could still be possible (CBS et al., 2020). Furthermore, they indicated that small houses had the same weight as large houses, meaning that this might skew the results.

Having a suitable ventilation system is important. However, it heavily competes with insulation in the sense that between 30-50% of heat produced in a space is lost through the ventilation system when no heat recovery mechanism is deployed. (Liddament & Orme, 1998). This means that having control over the air that enters and exits a dwelling plays an important role in lowering the heat demand and thus the energy consumption. Because of this, it is important to determine whether a dwelling has a mechanical ventilation system in place and whether heat recovery occurs, in order to better evaluate its effect on the heating system (Liddament & Orme, 1998).

Even though cooking is a fundamental part of domestic energy consumption, cooking only demands about 2% of the natural gas consumed in a dwelling (Segers et al., 2020). In the case of electric cooking, this represents about 4% of the electricity demand of a household (Topsector Energie, 2019). Even though cooking is one of the lowest energy consumers, data on gas and electric cooking was readily available and is therefore included in the results.

The energy demand used for cooling is expected to increase in the future due to multiple factors such as the urban heat island effect, thermal comfort, population growth and ageing (Topsector Energie, 2019). Apart from these factors it is common for new buildings to be better insulated but to also have relatively large windows. Due to the reduced size of newly built homes, they easily overheat in summer (Topsector Energie, 2019). The Dutch government has taken on the *TO-juli* indicator, which accounts for the demand for cooling in the summer in new buildings and should prevent buildings from overheating in future (Topsector Energie, 2019). The adoption of this indicator is an indication that the government anticipates an increase in the demand for cooling technologies. Considering the increased adoption of heat pumps for space and water heating and the heat pump's ability to both heat and cool a room, information on heat pump installations derived in the heating module can also be rolled over to indicate cooling characteristics.

Finally, household behaviour is an important factor to consider when evaluating the cost-effectiveness of energy saving measures (Tigchelaar et al., 2011). It has been found that predictions solely based on building characteristics are likely to overpredict energy consumption in inefficient buildings and underpredict in efficient buildings (Majcen et al., 2016). As much as 49% of residential heating consumption can be attributed to human behaviour (van den Brom, 2020). Therefore, including behavioural indicators when analysing energy characteristics could improve energy consumption estimates.

It is important to note that even though lighting and appliances contribute approximately 25% of the total energy use in a dwelling, we did not include it in our analysis as data on the current and possible future presence of these appliances are too circumstantial (RVO, 2019). Theorising a strategy that can accurately predict the current presence of an appliance can be rendered invalid within months, considering the fast-paced technological developments.

Due to time constraints and the lack of sufficient data it was decided to prioritise space heating, water heating, insulation and cooking. Specific attention also had to be paid to developing energy labels as they are a critical input to space and water heating. This resulted in the descoping of cooling, ventilation and human behaviour. However, some comments are made on the correlations between occupant demographics and energy consumption in Addendum B which could proove useful for further research.

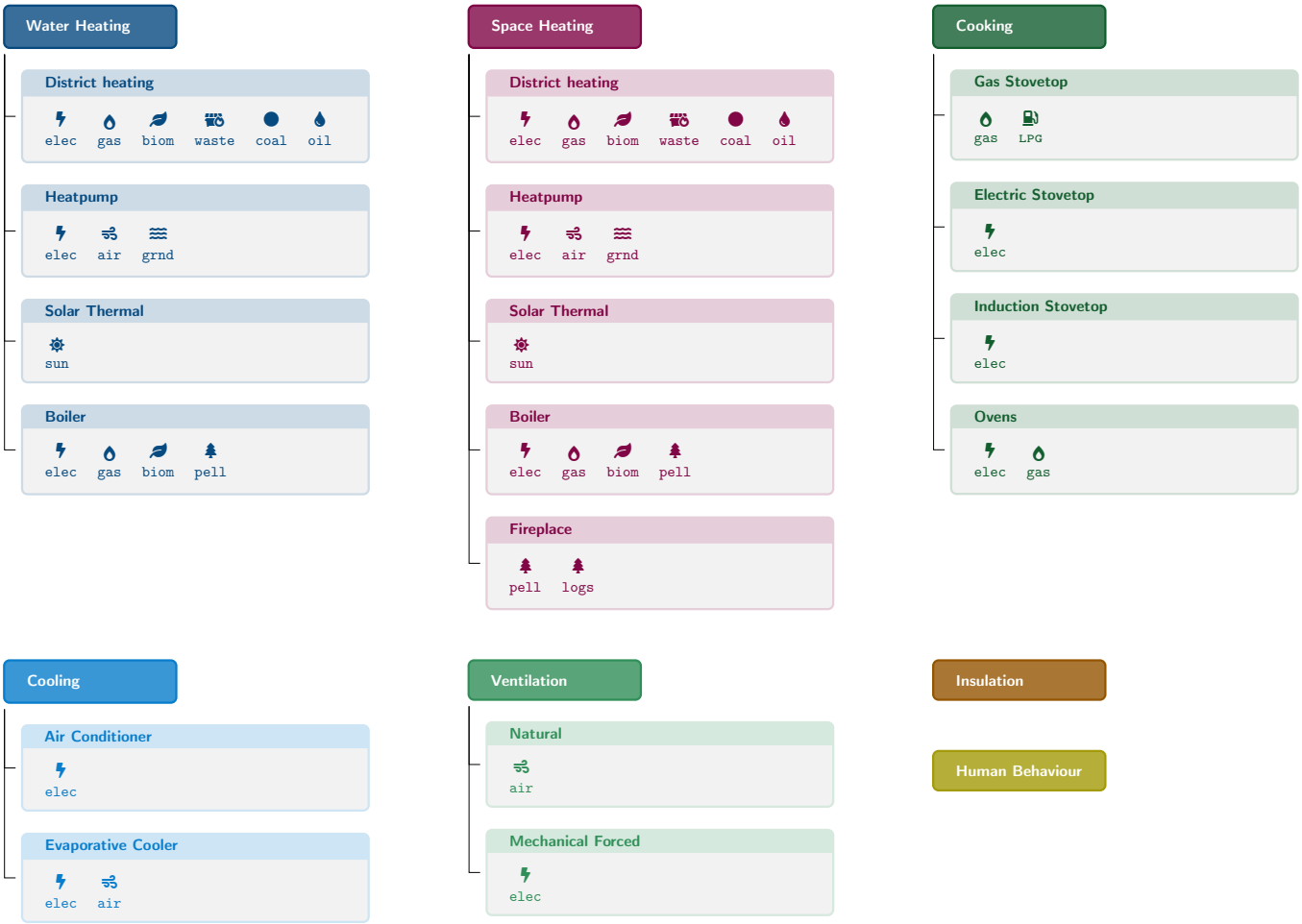


Figure 3.1: A taxonomy of the relevant energy characteristics. Note the three layers: function – appliance – final energy carrier. Abbreviations:

`elec`	electricity
`biom`	biomass
`waste`	municipal waste
`grnd`	ground
`pell`	wood pellets
`logs`	chopped wooden logs

3.2 Spatial levels

3.2.1 Definition of a dwelling

A dwelling is (a place in) a building where one or more persons can live. In this research, the definition of a dwelling is based on the definition of a 'verblijfsobject' as defined in the BAG (*Basisadministratie Adressen en Gebouwen*, Base Registry of Adresses and Buildings) records, which is:

> *The smallest unit of use, located within one or more buildings, fit for residential, commercial or recreational purposes, which has its own lockable entry from the public road, a yard or shared traffic space and which can be the subject of legal acts under property law and is functionally independent.*
>
> Translated from Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2018).

As per our client's instruction, we will however, only look at dwellings that are used for residential purposes. This definition is used as a standard throughout different datasets used in our research.

3.2.2 Higher spatial levels

The dwelling is the lowest spatial level that we consider in our research. Most data is available on higher spatial levels. An overview of the spatial levels and how they relate to dwellings is given in Table 3.1.

Table 3.1: Different spatial levels, sorted on number of entities on that level in the Netherlands. Note that there is no guaranteed hierarchy: a building might have multiple PC6 codes, a PC4 code might lie in different districts, etc. Empty values are unknown. All values in the example refer to the same dwelling. Surface area of the country is from the CBS *Kerncijfers* and might not match other definitions. When a range is given, it is the 95% range (2.5% percentile – 97.5% percentile) of entries. The area for dwellings and buildings is usable area from the BAG. Unless stated differently, the codes in the example are CBS-codes.

Unit	Dutch/definition	Area	Nº in NL	Nº dwellings	Example
country		33 671 km^2	1	7 951 730	the Netherlands
municipality	*gemeente*	11–335 km^2	355	1111–33 518	Utrecht (municipality) (`GM0344`)
district	*wijk*	35–4597 ha	3177	30–10 995	Wijk 05 Oost (`WK034406`)
PC4	four digits of the postal code		4053	18–6739	3584
neighbourhood	*buurt*	7–1683 ha	13 808	4–2748	Utrecht Science Park (`BU03440533`)
PC6	full postal code		437 900	2–51	3584 SB
building	*pand*	70–465 m^2	5 487 293	1–4	building 'Johanna' (BAG ID: `0344100000138505`)
address		—	7 951 730	1	Bisschopssteeg 1 (BAG ID *nummeraanduiding*: `0344200000170569`)
dwelling	*verblijfsobject* with residential function	37–204 m^2	7 951 730	1	BAG ID: `0344010000167627`

3.3 Data sources

To create our database we will use data from several established sources. Several databases exist that contain energy related information on all or a significant number of dwellings. Available data can be classified into three types: Type 1 data are indicators for which we have an entry for every dwelling in the Netherlands, the dataset is, so to say, complete. This is the same spatial scale as what our output data will be. There is Type 2 data, which is data for which we have an entry for specific dwellings but not for all dwellings. Finally, we have Type 3 data, this is data for which we have information for all of the Netherlands, but only at a higher than dwelling level spatial scale. For the final two types we will need a way to use this incomplete or higher level data to generate entries at individual dwelling level. An overview of these data types can be found in table 3.2

Table 3.2: Classification of different types of data

	Type 1	**Type 2**	**Type 3**
Completeness	Complete	Incomplete	Complete
Spatial scale	Specific	Specific	Higher level

The datasets used in this research can be found in Table 3.3. Tables 3.4, 3.5, and 3.6 provide an overview of the relevant data labels as scattered across the different data sources. Information on spatial levels can be found in Table 3.1. Furthermore, a complete scope means that all the dwellings in the Netherlands are encompassed in the dataset. Note that the names of the datasets are clickable hyperlinks which lead to the web page where the data can be found. In Table A.1 in Appendix A the datasets that were found in the proposal phase, but were not used in the research, are shown. Reasons for not using that data were that data was not easily fed into our program, that the spatial scale was too great, or that there was simply no use for the data. However, we feel that it is important to keep a record of all datasets that were reviewed, for future reference by other researchers.

3.3.1 BAG

The *Basisregistratie Adressen en Gebouwen* (BAG, Basis Registration Addresses and Buildings) is a database that indexes all buildings in the Netherlands, and is thus Type 1 data. It lists for every building in the Netherlands what kind of building it is, what its current function is, its floor space, and the year that it was built. It also describes the exact location, orientation and shape (of the footprint) of all buildings. The 3D BAG is a database that is created by using the previously describes dataset (2D BAG if you like) as a filter for the actual height map of the Netherlands (*Actueel Hoogtebestand*) so that only buildings are projected in three dimensions whereas the rest of the map remains two dimensional. The BAG is compiled by the *Kadaster* (Dutch Land Register) and updated monthly. It is based on data kept and delivered by municipalities.

3.3.2 WoON

The WOON survey is a survey done by the CBS to assess the quality of the Dutch housing stock. It covers roughly 40 000 houses. It contains a sub-survey of 4 500 dwellings called the energy module, that specifically assesses the energetic quality of the specified dwellings. Hereafter when we mention the WOON survey we will mean specifically the energy module of the WOON survey. This Type 2 dataset will be useful in creating a predictive model to deal with lacking actual data.

The WOON survey has a great deal of highly detailed information. It consists of two parts, a survey filled in by residents themselves, and a technological assessment done by an expert. The survey part mostly considers attitudes towards renovation and energy efficiency. It also tries to map energy using behaviour. The assessment part is a very thorough report on the technical state of the dwelling (energetically speaking). It lists the present systems for heating and cooling, renovation history (if applicable), and a detailed overview of the insulation of each building component. Additionally, there are some derived

Table 3.3: Relevant data sources and their characteristics.

Name	Description	Organisation	Spatial level	Completeness	Year
BAG	Dwelling Registration	Kadaster	Individual dwelling	Complete	2021
Kerncijfers per postcode	Dutch demographics	CBS	PC6	Complete	2017
WOON survey energy module	Built environment energy use survey and technical evaluation	CBS	Individual dwelling	4500 homes	2018
EP-Online	Energy labels	RVO	Individual dwelling	3.8 million homes	2021
Energielevering aan woningen en bedrijven naar postcode	Gas and electricity use of dwellings (and companies)	CBS	PC6	Complete	2019
CBS83878 Aardgaslevering vanuit het openbare net	Gas use distribution classified by energy label, building year, building type and floor space	CBS	Dwelling	Complete	2019
CBS83882 Elektriciteitslevering vanuit het openbare net	Electricity use distribution classified by building type number of inhabitants	CBS	Dwelling	Complete	2019
CBS84948 Woningen; hoofdverwarmingsinstallaties, regio	Heating and cooking installation types	CBS	Neighbourhood	Complete	2019
Warmtenetten	Number of households per neighbourhood connected to district heating	RVO	Neighbourhood	Complete	2021

Table 3.4: Data label matrix part 1

Description	BAG	WOON	Energy Label	Heating Systems	Energy Use (Postal Code)	Energy Use (Neighbourhood)	Benchmark Gas Use	Benchmark Electricity Use	District Heating Network	Demographics	Gas and DH Connections	Solar Panels	Kleinverbruiksdata
BAG ID	Yes	-	Yes	-	-	-	-	-	-	-	-	-	-
Geographic location	Yes	-	-	-	-	-	-	-	-	-	-	-	-
Postal code	PC6	-	PC6	-	PC6	-	-	-	-	PC4	-	-	PC6
House number	Yes	-	Yes	-	-	-	-	-	-	-	-	-	-
Floor space	Yes	Yes	Yes	-	-	-	Yes	Yes	-	-	-	-	-
Building year	Yes	Yes	-	-	-	-	Yes	Yes	-	Yes	-	-	-
Demolition year	Yes	-	-	-	-	-	-	-	-	-	-	-	-
Number of floors	Yes	Indirect	-	-	-	-	-	-	-	-	-	-	-
Building per dwelling	Yes	-	-	-	-	-	-	-	Yes	-	-	-	-
Dwelling type	Yes	Yes	Yes	-	-	Yes	Yes	Yes	-	-	-	-	-
Building ID	Yes	-	-	-	-	-	-	-	-	-	-	-	-
Neighbourhood ID	Yes	-	-	Yes	-	Yes	-	-	Yes	-	-	-	-
Neighbourhood name	-	-	-	-	-	-	-	-	Yes	-	Yes	-	-

Table 3.5: Data label matrix part 2

Description	BAG	WOON	Energy Label	Heating Systems	Energy Use (Postal Code)	Energy Use (Neighbourhood)	Benchmark Gas Use	Benchmark Electricity Use	District Heating Network	Demographics	Gas and DH Connections	Solar Panels	Kleinverbruiksdata
Municipality ID	Yes	-	-	-	-	-	-	-	Yes	-	-	Yes	-
Municipality name	-	-	-	-	-	Yes	-	-	Yes	-	-	-	Yes
Energy index	-	-	Yes	-	-	-	-	-	-	-	-	-	-
Energy label	-	-	Yes	-	-	-	-	-	-	-	-	-	-
Net heat demand	-	-	Yes	-	-	-	-	-	-	-	-	-	-
Heating system	-	-	-	Yes	-	-	-	-	-	-	-	-	-
Avg natural gas use per dwelling	-	Yes	-	-	Yes	Yes	-	-	-	Yes	-	-	Yes
Avg electricity use per dwelling	-	Yes	-	-	Yes	Yes	-	-	-	Yes	-	-	Yes
National benchmark gas use	-	-	-	-	-	-	Yes	Yes	-	-	-	-	-
National benchmark electricity use	-	-	-	-	-	-	-	-	-	-	-	-	-
Number of dwellings	-	-		-	-	-	-	-	Yes	Yes	-	-	-

Table 3.6: Data label matrix part 3

Description	BAG	WOON	Energy Label	Heating Systems	Energy Use (Postal Code)	Energy Use (Neighbourhood)	Benchmark Gas Use	Benchmark Electricity Use	District Heating Network	Demographics	Gas and DH Connections	Solar Panels	Kleinverbruiksdata
Percentage district heating users	-	-	-	-	-	Yes	-	-	Yes	-	-	-	-
Number of gas grid connections	-	-	-	-	-	-	-	-	-	-	Yes	-	Yes
Number of district heating connections	-	-	-	Yes	-	-	-	-	-	-	Yes	-	-
Number of electrical connections	-	-	-	-	-	-	-	-	-	-	-	-	Yes
Number of residents	-	-	-	-	-	-	-	-	-	Yes	-	-	-
Age groups	-	-	-	-	-	-	-	-	-	Yes	-	-	-
Family types	-	-	-	-	-	-	-	-	-	Yes	-	-	-
Distance to amenities	-	-	-	-	-	-	-	-	-	Yes	-	-	-
Number of solar panels installed	-	-	-	-	-	-	-	-	-	-	-	Yes	-
Capacity solar panels (kW)	-	-	-	-	-	-	-	-	-	-	-	Yes	-
Electricity grid feed-in	-	-	-	-	-	-	-	-	-	-	-	-	-

variables that compile previous information and calculate compounded variables. For example: there is a derived variable for roof insulation that takes the insulation and surface of all individual roof segments and generates a single number.

3.3.3 CBS84948 heating and cooking system installations

The CBS84948 data set provides a conditional relationship between space heating, water heating and cooking applications based on dwelling level energy label input data and gas and electricity usages. This raw data is aggregated by CBS to a neighbourhood level to protect the privacy of individuals.

CBS classifies this aggregated data into eight usable categories, from which the presence of a heating or cooking application can be deduced. Table 3.7 provides an overview of the eight categories and their respective allocations. A ninth category is also provided for, but this category represents the dwellings for which there is no energy label, where no district heating has been identified based on customer files, no gas connection has been found, and no heat pump subsidy has been linked. We therefore do not apply this ninth category in our analysis. The data set ultimately provides the percentage of dwellings which fall into each category in a particular neighbourhood. As explained later on, these percentages are used to determine the base probabilities for the presence of space heating, water heating and cooking applications.

3.4 Dwelling types

In order to link our input data to BAG ID's it is useful to understand the dwelling types that are specified in the different datasets. The first distinction in dwelling category is made in 'eengezinswoningen' or single household dwellings and 'meergezinswoningen' or multiple household dwellings. Within these categories further specifications on dwelling type can be made. A single household is defined as a dwelling which does not share a building with other dwellings. However, the dwelling can be connected horizontally to other dwellings. Examples for single household dwellings are corner houses, terraced houses and freestanding houses. Table 3.8 describes the different dwelling types for single household dwellings. Multiple household dwellings are defined as a dwelling which shares a building with other dwellings. In multiple household dwellings distinctions can be made between top, bottom cornered an terraced apartments for example. Table 3.9 describes the different dwelling types for multiple household dwellings. Figure 3.2 shows the different categories for dwelling type with an illustration.

Table 3.8: Overview of the different dwelling types of single household dwellings.

Building type in Dutch	Translation	Description
Rijtjeshuis	Terraced house	A dwelling which is connected to at least two other single household dwellings.
Hoekwoning	Corner house	A dwelling which is connected to exactly 1 terraced house. A corner house lies at the end of a row of terraced houses.
Twee onder 1 kap	Two under 1 roof	A dwelling which is connected to exactly one other dwelling which is not a terraced house.
Vrijstaand	Free standing	A dwelling which has no connecting walls to another building.

Table 3.7: Overview of data set CBS84948 providing a breakdown of the conditional relationships between electricity and gas use, space heating, water heating and cooking applications

Category	Conditions	Space Heating					Water Heating					Cooking	
		Individual Gas Boiler	**Block Heating Boiler**	**District Heating**	**Electric Boiler**	**Heat Pump**	**Individual Gas Boiler**	**Block Heating Boiler**	**District Heating**	**Electric Boiler**	**Heat Pump**	**Gas**	**Electric**
1	DH* with high gas use	backup	backup	yes	no	possible	backup	backup	yes	no	possible	possible	possible
2	DH with low gas use	no	no	yes	no	possible	yes	no	possible	no	possible	yes	possible
3	DH with no gas use	no	no	yes	no	no	no	no	possible	yes	possible	no	yes
4	EH* with high gas use	no	no	no	no	Hybrid	yes	no	no	no	no	yes	no
5	EH with low gas use	no	no	no	yes (or)	yes (or)	yes	no	no	no	no	yes	no
6	EH with no gas use	no	no	no	yes (or)	yes (or)	no	no	no	yes (or)	yes (or)	no	yes
7	Individual gas boiler	yes	no	no	no	possible	yes	no	no	no	possible	Yes	no
8	Block heating	no	yes	no	no	possible	no	yes	no	no	no	Yes	no

*DH = District Heating, EH = Electric Heating

Table 3.9: Overview of the different dwelling types of multiple household dwellings. Within the category of multiple household dwellings a distinction can be made between dwellings with 1 floor level or multiple floor levels. However, the subcategories described below do not differ between these dwelling types.

Horizontal type	Vertical type	Translation	Description
Hoek	Vloer	Corner floor	A ground floor apartment with at least three walls not connected to another apartment or building
	Midden	Corner middle	A corner apartment which has at least one heated floor above and below.
	Dak	Corner roof	A corner apartment which has no heated floors directly above.
	Dak vloer	Corner roof floor	A corner apartment with no heated floors directly above or below
Tussen	Vloer	Terraced floor	A ground floor apartment with less than three walls not connected to another apartment or building
	Midden	Terraced middle	A terraced apartment which has at least one heated floor above and below.
	Dak	Terraced roof	A terraced apartment which has no heated floors directly above.
	Dak vloer	Terraced roof floor	A terraced apartment with no heated floors directly above or below

3.5 Energy labels

The energy label is a general indication of the energy efficiency of a building. It is expressed as a letter ranging from G to A^{+++++} where G indicates an inefficient building and A^{+++++} indicates an efficient building. Currently approximately half of the dwellings (4 364 061 (RVO, June 2021) out of 7 892 928 dwellings (BAG, June 2021) in the Netherlands have a registered energy label.

As of the 1st of January 2021, energy labels have to be calculated using the NTA-8800 method ("Veelgestelde vragen Energielabelverplichting woningen", n.d.). Previously multiple methods were available to calculate the so called 'energieprestatieindex' (energy performance index, EPI) of a dwelling. The value that this index took would then directly translate to an energy label according to a nationally decided key. In the Netherlands this became the NTA-8800. The NTA-8800 is different from other methods in that it is more extensive as it considers the energy performance indicator in kilowatt hours per square meter per year (kWh/(m^2·yr)) which requires a detailed survey of the dwelling (*Besluit energieprestatie gebouwen*, 2006) ("Veelgestelde vragen Energielabelverplichting woningen", n.d.). In time, as energy labels are updated, this will mean that all dwellings will be evaluated according to this method. For now, however, not all dwellings have been evaluated using the same methods. While the differences in outcome between the varying methods are minor, it is expected that 90% of dwellings will remain in the same tier regardless of the calculation method and if there would be a difference, the change would not be greater than 1 tier (a D label will not become a B label) ("Veelgestelde vragen Energielabelverplichting woningen", n.d.). However, because of this updating of the labels, it is necessary to use caution when using energy labels.

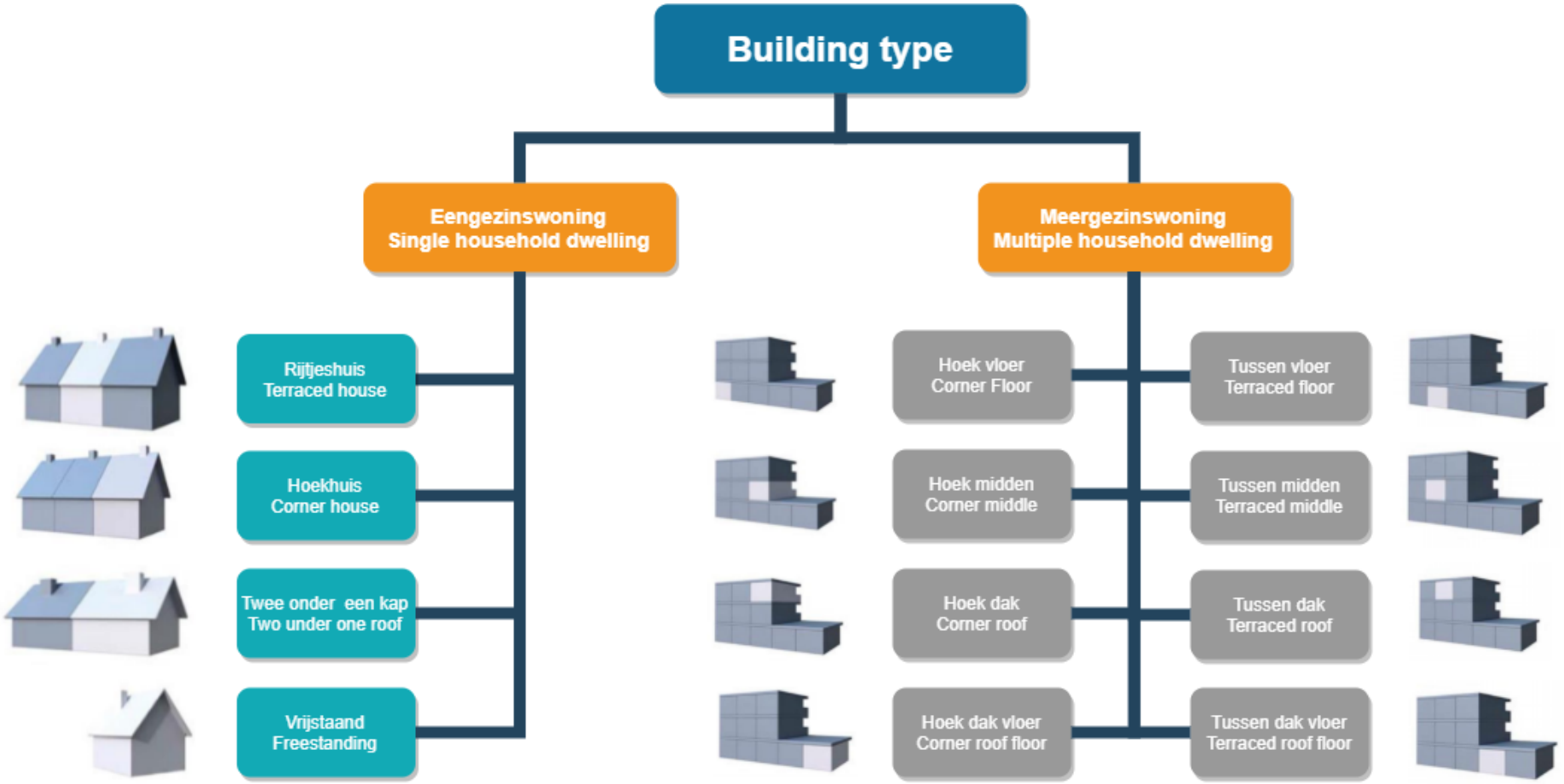


Figure 3.2: Diagram of dwelling types. Building illustrations adapted from Handleiding bij opnamelijst WoON energie (CBS, 2018)

The registration of energy labels since 1850 is shown in Figure 3.3. Since the year 2000, the majority of dwellings that have had an energy label registered have been assigned energy label A. Since 2008 it has been mandatory to get an energy label at the sale, renting or delivery of a building (*Besluit energieprestatie gebouwen*, 2006). However, energy label evaluation is not exclusively done in these situations. A homeowner can get an energy label assigned to their dwelling at any point in time. In Figure 3.4, the number of registered energy labels per calculation method is represented throughout the years. We can see that over time, many different calculation types are used and that the implementation of the detailed evaluation through the NTA-8800 method only started in 2020.In Table **??**, the number of energy labels that were determined per calculation type is reported. Here we can see that a significant number of energy labels were determined through the *Nader Voorschrift* and *Rekenmethodiek Definitief Energielabel* methods.

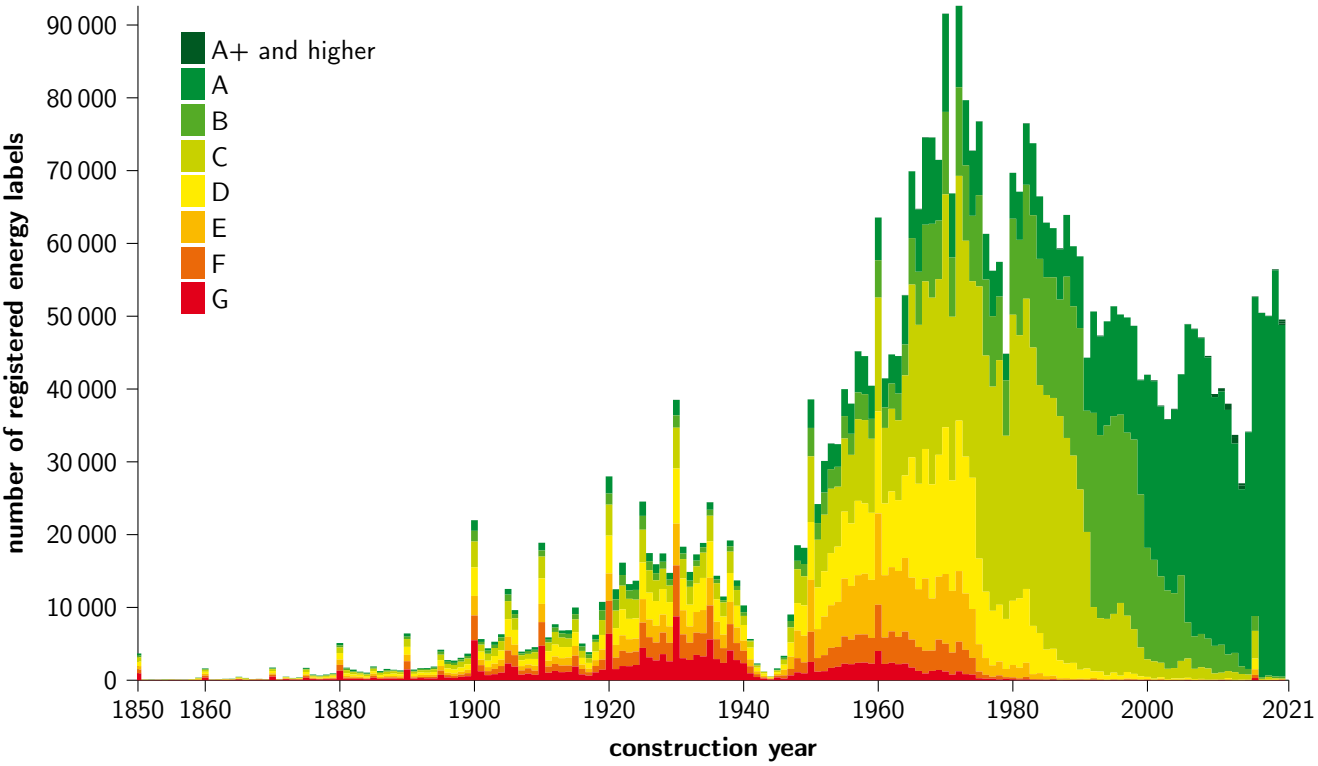


Figure 3.3: Registered energy labels per construction year, derived from the EP-Online database Rijksdienst voor Ondernemend Nederland, 2021 (own analysis). The peaks at every decade indicate that building years in the BAG are given as the estimated nearest decade when the precise building year is unknown. Note the large dip in the amount of buildings built around the World War II period, and the smaller dip after the financial crisis of 2008. Only very few labels have been registered yet for buildings built in the year 2021 (Rijksdienst voor Ondernemend Nederland, 2021).

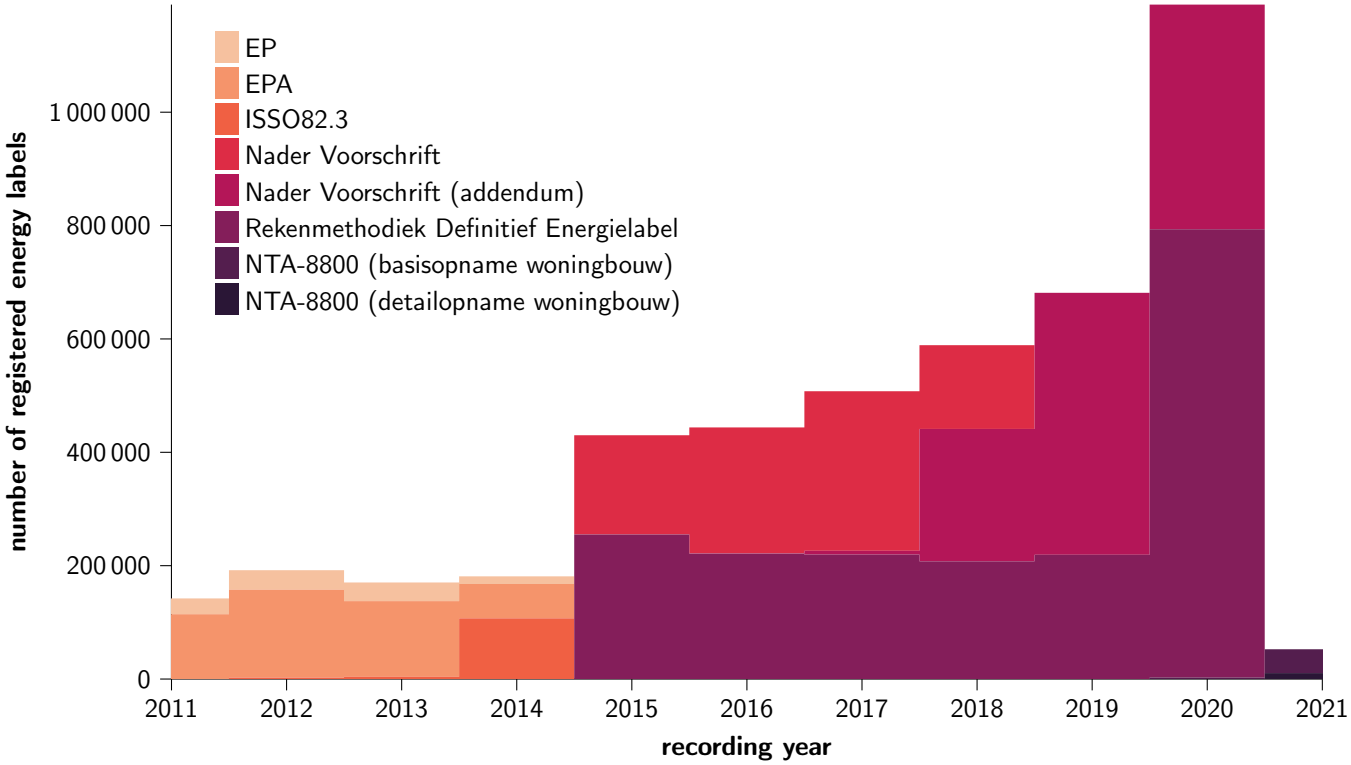


Figure 3.4: Number of registered energy labels per calculation type throughout the years (Rijksdienst voor Ondernemend Nederland, 2021).

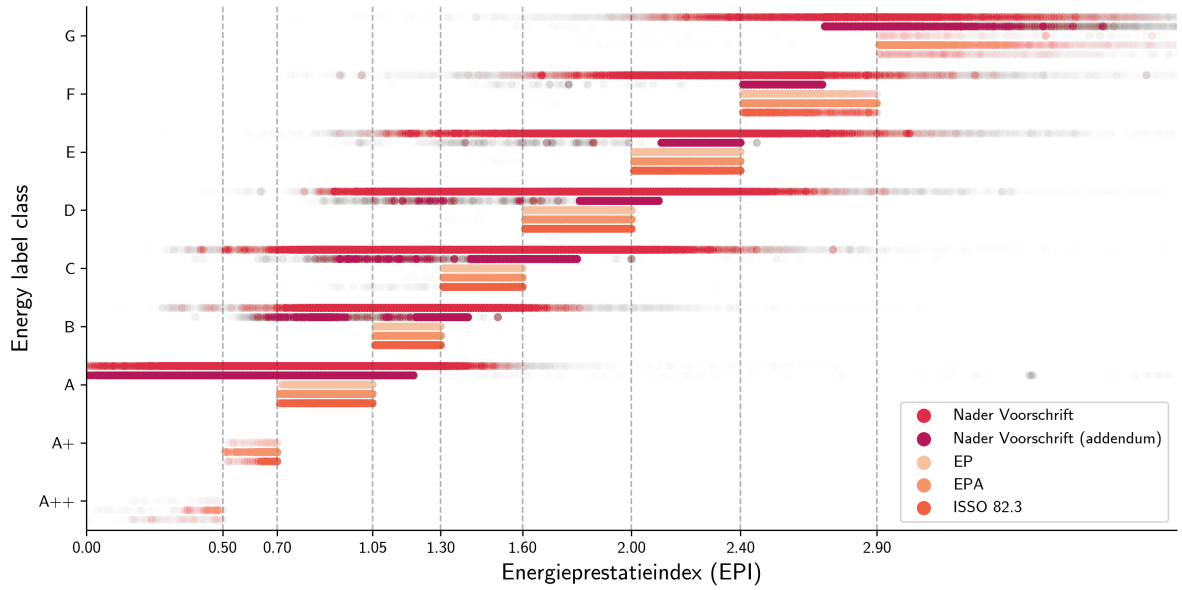


Figure 3.5: Value of Energieprestatieindex per calculation type and energy label (Rijksdienst voor Ondernemend Nederland, 2021).

In Figure 3.5, the determined EPI through various calculation methods is plotted together with the resulting energy label that the dwelling received. Something to note here, is that in the case of the *Nader Voorschrift* and *Nader Voorschrift (addendum)* is that there is an overlap in the EPI of dwellings from different energy label tiers. This is a problem as it would indicate that there is not a clear method to determine what EPI belongs to what energy label class, as well as houses that received a C classification being able to receive an A classification and vice versa. This is however not the case for EPA, EP and

ISSO 82.3 where the boundaries are well separated. In Table **??** we established the visible ranges for each energy label according to the data from EPA, EP and ISSO 82.3.

Table 3.10: The boundaries and averages of the *Energieprestatieindex* (EPI) for the different energy labels, considering only those calculated with EP, EPA or ISSO 82.3 version 3.0, October 2011. Furthermore, the new boundaries that have to be used starting from the 1st of January of 2021.

Energy Label Class	**EPI bounds according to the data** (kWh/m^2)	**EPI average according to the data** (kWh/m^2)	**Nº of labels in the data**	**EP-2 range** (from 1st of January 2021) (kWh/m^2)
A++++	-	-	-	≤0.00
A+++	-	-	-	0.01-0.50
A++	(∞, 0.5]	0.281	1 553	0.51-0.75
A+	(0.5, 0.7]	0.635	4 148	0.76-1.05
A	(0.7, 1.05]	0.938	68 326	1.06-1.60
B	(1.05 - 1.3]	1.200	166 138	1.61-1.90
C	(1.3, 1.6]	1.451	202 969	1.91-2.50
D	(1.6, 2.0]	1.785	133 696	2.51-2.90
E	(2.0, 2.4]	2.184	64 102	2.91-3.35
F	(2.4, 2.9]	2.612	31 789	3.36-3.80
G	(2.9, ∞)	3.237	13 029	>3.80

To determine the energy label from the EPI, a conversion table is given, with EPI ranges pertaining to a particular label and vice versa. These ranges are not the same as these were found in the data and their corresponding energy labels. The newer ranges are implemented from the 1st of January 2021 *Staatscourant* from (Koninkrijk der Nederlanden, 2014).

3.6 Statistical instruments

3.6.1 Multiple linear regression

To find correlations between data points, we can use linear regression. When the independent values (or: explanatory values) are multidimensional (vectors) and the dependent values (or: response values) are one-dimensional (scalars), this is called *multiple* linear regression. We are mainly interested in using multiple linear regression to *predict* rather then *test* for correlations. Below, we summarize the method and introduce the required notation, based on the course by Pardoe et al., 2021. Vectors and matrices are denoted in bold, the transpose of a matrix $\boldsymbol{A}$ is denoted as $\boldsymbol{A}^T$ and the inverse as $\boldsymbol{A}^{-1}$.

We start with n observations of the form $(\boldsymbol{x_i}, y_i)$ where $\boldsymbol{x_i}$ is a vector of length $p-1$: $(x_{i,1}, x_{i,2}, \ldots x_{i,p-1})$. We can model the observations as a linear relation:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \text{ for } i = 1, 2, \ldots, n.$$

Here, ε_i is the error term for observation i. There are p regression coefficients β_j, and β_0 is called the *intercept*. In shortened form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y}$ is the vector $(y_1, y_2, \ldots, y_n)$, $\boldsymbol{\varepsilon}$ is defined similarly, and $\boldsymbol{X}$ is the vector of observations $\boldsymbol{x_j}$ with an extra column of ones:

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots \\ 1 & x_{2,1} & x_{2,2} & \ldots \\ 1 & x_{3,1} & x_{3,2} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Given these observations, we want to find the coefficients $\boldsymbol{\beta}$ that minimize the error terms with regard to some metric. Usually, this means minimizing the residual sum of squares s^2 (or: mean squared error of the residuals):

$$s^2 = \frac{\sum_{i=1}^{n} (\varepsilon_i)^2}{n - p}.$$

This method is called the 'Ordinary Least Squares' method. In that case, the regression coefficients $\boldsymbol{\beta}$ are given by:[1]

$$\boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y},$$

and the covariance matrix S is given by:

$$S = s^2 \cdot (\boldsymbol{X}^T \boldsymbol{X})^{-1}.$$

With that established, we can generate a prediction $\widehat{y_h}$ for an input $\boldsymbol{x_h}$ that is not in the observations:

$$\widehat{y_h} = \boldsymbol{\beta}^T \boldsymbol{x_h}.$$

This prediction is not precise, since in practise there is an error term. Assuming that the errors are normally distributed with a constant variance, the following range is a prediction interval[2] for $\widehat{y_h}$ for confidence level $1 - \alpha$:

$$\widehat{y_h} \pm t_{(1-\alpha/2, n-p)} \cdot \sqrt{s^2 + \boldsymbol{x_h}^T S \, \boldsymbol{x_h}}.$$

Here, $t_{(1-\alpha/2, n-p)}$ is the $(1 - \alpha/2)$-quantile of the Students t-distribution with $n - p$ degrees of freedom. In other words, the 'true' y_h will fall within this range in $1 - \alpha$ of the predictions. We often take α as 0.05, so that the interval is the 95%-interval. For new prediction intervals, only the coefficients $\boldsymbol{\beta}$, the residual sum of squares s^2 and the covariance matrix S need to be known.

3.7 Human behaviour

Human behaviour is an important factor in predicting a dwellings energy consumption (van den Brom, 2020). It is, however, also very complex to model and predict. Additionally human behaviour mostly seems to affect energy use, which is not necessarily equivalent to building characteristics, which are the object of interest in this study. Given this complexity and questionable relevance to this research we have decided to leave human behaviour as a factor out of the scope of this research proper.

Considering, however, also the importance of understanding human behaviour when trying to understand energy use in the built environment, we have performed some preliminary tests on data available for the Netherlands to hopefully shed some more light on this important phenomenon. A short explanation of the methods used during this tangential research can be found in section 4.11, and the full results can be found in Appendix B.

[1] In practice, there are more efficient methods to determine β than inverting the $n \times n$ matrix $\boldsymbol{X}^T \boldsymbol{X}$.

[2] Not to be confused with the *confidence* interval for a linear regression: a confidence interval gives the uncertainty for the prediction the *mean* response at $\boldsymbol{x_h}$, while the prediction interval gives the uncertainty for *ne specific prediction* at $\boldsymbol{x_h}$. The confidence interval is much smaller than the prediction interval since it only has to account for the prediction error and not the true error in the observations.

4 | Methodology

The following chapter will outline the methods that were used to generate the dataset which is our final result. It will describe the data gathering process, how data is stored and treated, and how the modules work that calculate the final dataset. In Figure 4.1 the research process can be seen. The structure is based on the sub-questions formulated before and elaboration on each step is given below.

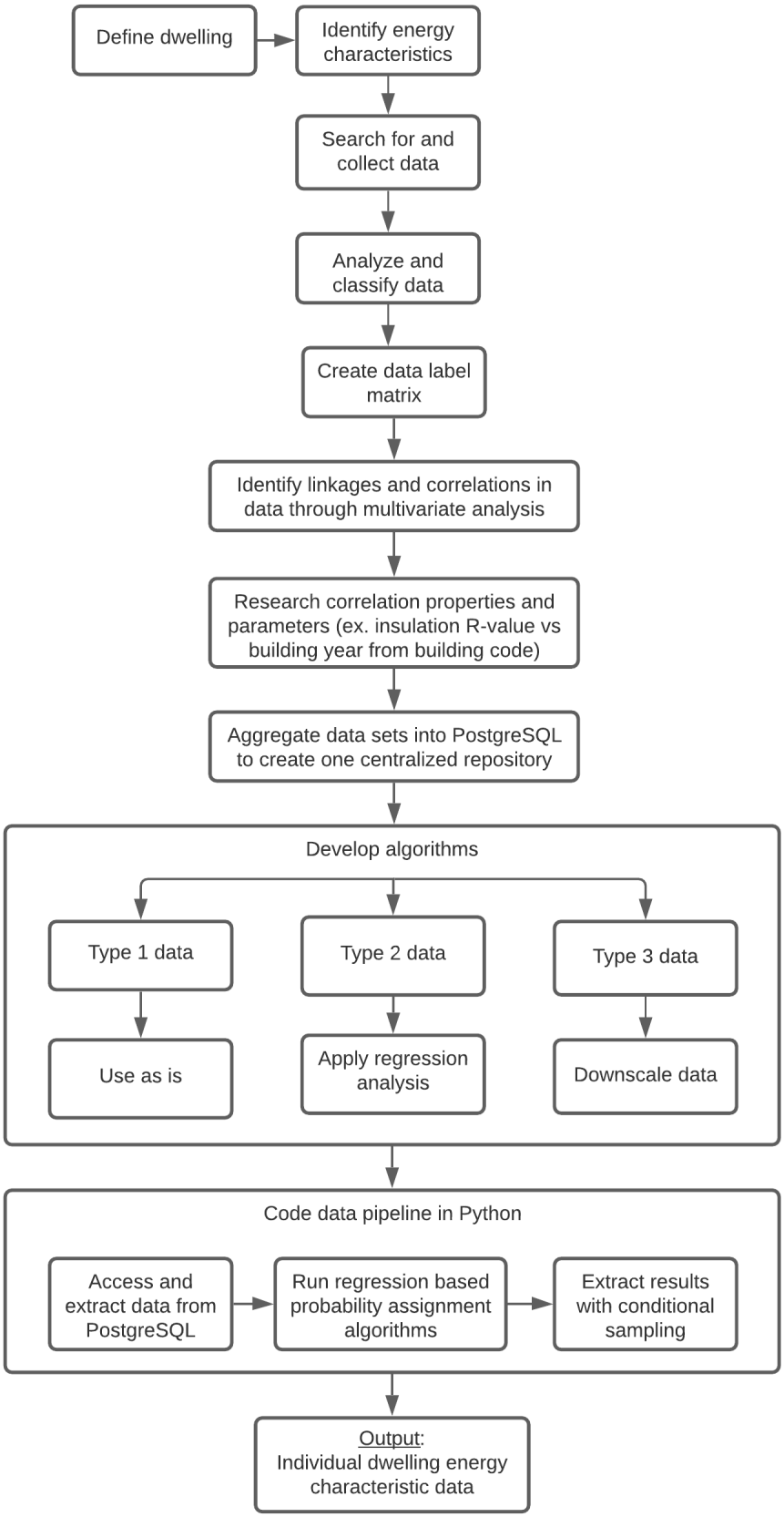


Figure 4.1: Block diagram of the research process.

4.1 Data gathering, cleaning and analysis

The BAG and WOON data sets were provided by PBL and TNO while the remaining data sets were sourced through internet searches. Going through the preliminary list of data sets we found some to be unusable, like the Nationale EnergieAtlas, which while it contained much information, was not readable for a computer. Some were inaccessible to us like the SHAERE database which after contacting Aedes we were informed that it is only available for research that directly benefits them. In other cases an iterative process occurred where doing further researching on one database led to finding more data sets.

The data sets were analysed by systematically working through the contents and while doing so compiling a matrix detailing exactly which data set contained what information. Please refer to Tables 3.4, 3.5, and 3.6 for this data label matrix. When this was done we started looking for links between data sets that would allow us to create our so-called modules. These should be seen as the building blocks from which our algorithms are built. They are the units that from one or more data sets, or output of other modules, compute a new piece of information.

A sanity check was performed on the CBS sourced gas and electricity usages. This per dwelling average gas and electricity usages were converted to m^3/m^2 and kWh/per person which were then compared to the CBS sourced benchmark electricity and gas usages. The converted average gas and electricity usages ranged up unto 120 m^3/m^2 and 6 980 kWh/person, respectively, while the benchmark consumptions were limited to 123.7 m^3/m^2 and 9380 kWh/person. Therefore, no outliers had to be excluded from the analysis.

PBL also referred us to an online map viewer which similarly to our scope provides the energy characteristics on a per dwelling level for rental (Hoogenboom, 2020)[1] and social housing (Hoogenboom, 2018)[1]. A short description on the website refers to the data being sourced from the BAG and RVO, but no further references or descriptions on how the outputs are compiled are provided. This data source can therefore only serve as a partial validation, as its accuracy cannot be confirmed. However, considering the novelty of our work and therefore the lack of comparative data, this source provides for at least some validation. Information on the rental dwellings were last updated on July 20, 2020 while the social housing data was last updated on Feb 23, 2021. Unfortunately, only information on the rental dwellings could be used for validation as this was the only data set for which the addresses could be extracted for download. 123 629 of the 753 306 records in this data set contained information on the space and water heating characteristics. Social housing downloads could not be correlated as there were no way to link the extracted data to its corresponding dwelling. Manually correlating the extracted data with the dwelling addresses online is also not practical. Outcomes of the validation is described in paragraph 5.1.

To manage all this data we decided to work in PostgreSQL, a free, open-source relational database management system, and wrote programming in Python to easily load in data sets and retrieve the stored information.

4.2 Data pipeline

The way the output data is generated is through a data pipeline. This is a python program that takes dwelling data from the BAG and, by passing the dwellings through multiple modules, generates data about the energy characteristics of those dwellings. The modules are the building blocks of the pipeline. In Figure 4.2 a graphical representation of the pipeline and its constituent modules can be found. The modules are grouped by their purpose, which can be to assign probabilities for certain heating installations being present in a dwelling, to assign probabilities for the insulation values of the dwelling, or to do supporting tasks like gathering data or sampling the probabilities. In each module, information is appended to the dwelling object for further use, as either input for other modules or as resulting output.

The modules are divided into two types. There are "normal" modules, through which dwellings pass, and Regional modules, through which regions pass. In these regional modules, data that is available on a PC6 or neighbourhood spatial scale in retrieved. This data is then assigned to the region in question,

[1] These two references are clickable links to the online viewer.

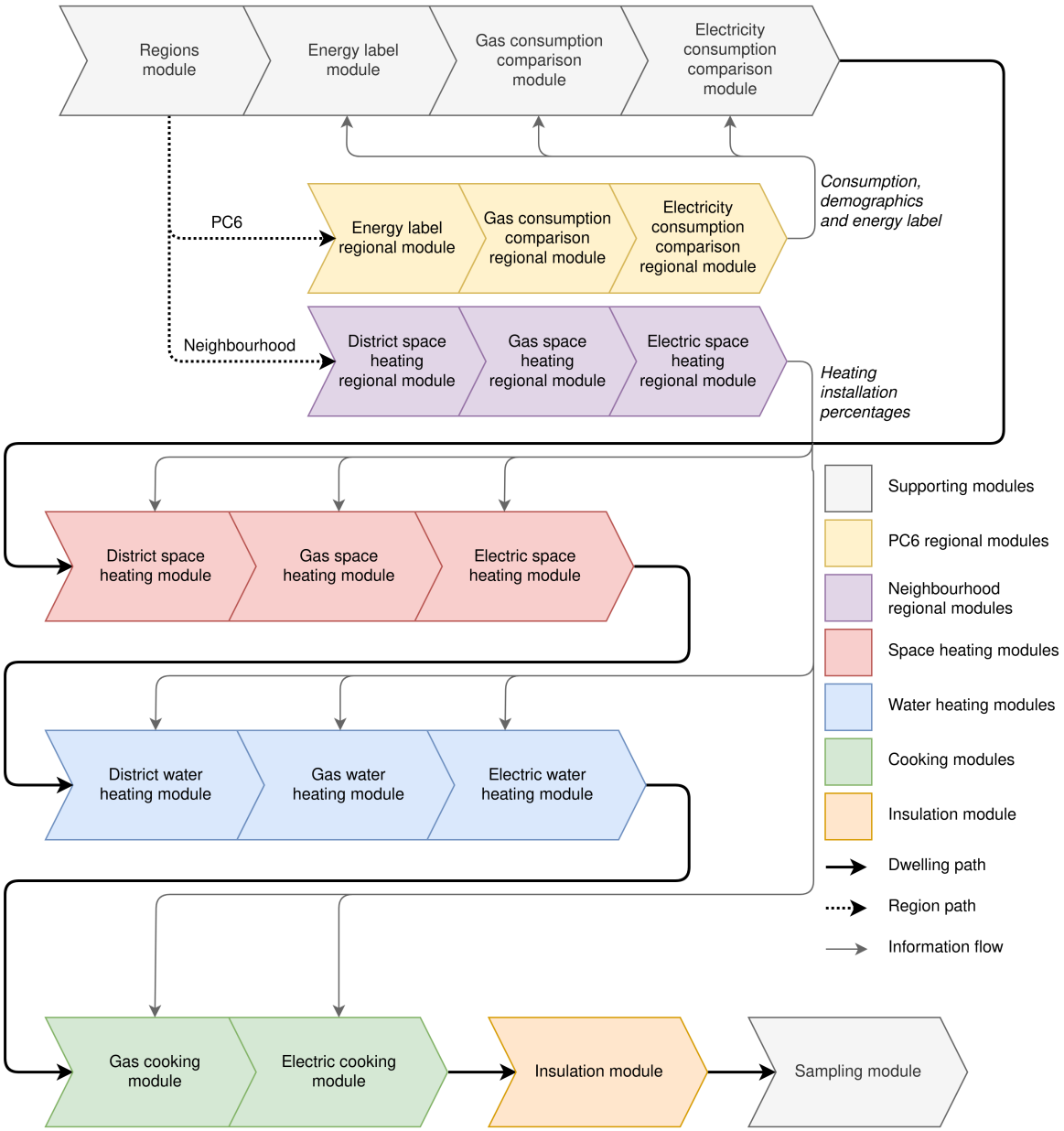


Figure 4.2: A diagram of the data pipeline showing the path a dwelling takes.

which makes it so that for every dwelling in that region the data is available, while only having to retrieve it once. More information on the region objects can be found in Section 4.3.

In the following sections, all modules will be described in more detail.

4.3 Regions module

The first module a dwelling passes through is the Regions module. In this module, the dwelling gets associated with region objects tied to its postal code and neighbourhood: a PC6 object, which represent the postal code, and a BUURT object, which represents the neighbourhood, if these do not already exist. If there already are objects for the relevant postal code and/or neighbourhood, the dwelling replaces its placeholder object in the region object.

Whenever a region object is created, it adds a placeholder object for all dwellings located in the region to itself. These placeholders are the same class as dwellings, except for the fact that they contain less information. Their use is to be able to calculate certain characteristics of the region before all dwellings

have passed through the pipeline.

After the addition of the placeholder objects, the newly created regions passes through all regional modules that are associated with the kind of region, which can be found in Figure 4.2. For a PC6 object, those are the energy label and comparison modules. In the regional energy label module, the average imputed energy label is calculated. The regional comparison modules look up and calculate information that is needed for when the dwelling in the postal code go through the comparison modules.

A BUURT object passed through the regional space heating modules. In these modules, the data from the CBS84983; Woningen; hoofdverwarmings; buurt 2019 dataset, encountered in Section 3.3.3, is retrieved. This information is the basis of the probability assignment for space heating, water heating and cooking. Elaboration on the way the data is used will follow in the relevant sections to come.

Thus, the use of the regions is to be able to access regional data for all dwellings in that region, without having to look it up or calculate it for each dwelling individually.

4.4 Energy label module

4.4.1 Predicting energy labels from building age, dwelling type and PC6 average

Roughly half of the dwellings in the Netherlands have an energy label. This makes it an interesting target to predict: how accurately can we predict energy label classes for a dwelling which does not already have an energy label? Furthermore, energy labels are often geographically correlated: dwellings in the same neighbourhood or in the same building complex have similar energy labels. However, as discussed in the theoretical background, many of these energy labels have not been measured through the same method. Although this may not have an immediate effect on the energy label received, it does mean that for some dwellings we have an energy label but not an Energy Performance Indicator (EPI). This is relevant because we want to use the EPI to be able to determine the average EPI and consequently know the average energy label for a postal code area.

To do this, we will use multiple linear regression (see Section 3.6.1 for the theoretical background) to try to predict these labels. The reasons to use this type of regression is mainly practical: linear regression is relatively easy to apply and to reason about. And since we are interested in predicting, rather than testing hypotheses, we don't have to strictly adhere to the assumptions about linearity, independence and normality of the error terms that are required when testing hypotheses. We can simply validate our predictions on the large data set of existing energy labels.

We will go about this in the following way:

- The goal is to predict a 95% confidence and prediction interval for the energy label class of a specific dwelling, given its known characteristics.
- We assume that the dwellings that already have an energy label are representative for all dwellings in the Netherlands. Thus, we can use the existing labels to predict label classes for dwellings without an energy label. We didn't check this assumption, but we think it is reasonable given the large number of energy labels, and the fact that energy labels are mandatory for certain transactions (renting, selling, building) that apply to all types of dwellings.
- For linear regression, we need a single number as our outcome. For this we take the *energie-prestatieindex* (EPI). We convert between energy label class (G to A^{+++++}) and the EPI using the boundaries as determined from the given data in Table **??**. Specifically: for an energy label that has an EPI that has been calculated with the ISSO82.3, EP or EPA methods we will keep that value, else we convert the energy label class using the average EPI of that class. If required for presentation, we can later on transform an EPI back to an energy label class using the same boundaries again.
- We take the PC6 as the geographical scope, and use the average EPI in that PC6 as a measure for the energy labels in the surroundings of a dwelling. Remember from Table 3.1 that in 95% of the PC6's, there are between 2 and 51 dwellings. That means that the PC6 spatial level is high

enough that there usually are at least a couple of dwellings with an energy label, and that the level is low enough that the average is still relevant for specific dwellings. For dwellings in a PC6 without any energy labels, we will take the average EPI of the neighbourhood (*buurt*), or when necessary the national average.

- We consider only the independent variables for which we have information for all dwellings in the BAG. We focus mainly on dwelling type and construction year, but we discuss other variables and manipulations of these variables later on.
- The dwelling type is a categorical variable, without a logical numerical value. For linear regression, it is required that all variables are numbers. Therefore, we convert the categorical variable in different binary variables: this is called *dummy coding*. We have 6 different dwelling types, so this leads to 5 binary variables (adding a 6th binary variable is unnecessary and counterproductive since it breaks the linear independence required for the linear regression to work).

Remember from Section 3.6.1 that running the linear regression will give us the vector of coefficients $\boldsymbol{\beta}$. With an input $\boldsymbol{x_h}$ (in this case a vector containing coded characteristics of a dwelling), we can then make a prediction of the corresponding EPI $\widehat{y_h}$:

$$\widehat{y_h} = \boldsymbol{\beta}^T \boldsymbol{x_h}.$$

We previously said that the following is a prediction interval for confidence level $1 - \alpha$:

$$\widehat{y_h} \pm t_{(1-\alpha/2,n-p)} \cdot \sqrt{s^2 + \boldsymbol{x_h}^T S \, \boldsymbol{x_h}}.$$

Where ŷ is the sample estimate or the predicted value, $t_{(1-\alpha/2,n-p)}$ is the t-multiplier, and what is contained within the radical is the standard error of prediction.

This interval only holds assuming that the errors are normally distributed with a constant variance. Unless we do more rigorous statistical testing on the data set to see whether these assumptions hold, we cannot assume that this is the case for this data set. Luckily, this is no problem: we can simply check in how many of cases the actual energy label class falls within the prediction interval for the corresponding inputs. We can then increase or decrease this prediction interval until around 95% of all energy label classes in our data set are within their prediction intervals.

We tried different combinations of dependent and independent variables, see the results in Table 4.1. We aim for the correlation that provides the best fit, and thus the highest r^2 value (r^2 values lie between 0 and 1). We tried the following variants:

- We started with a simple variant, without modifying the variables, and using only the construction year, the EPI PC6 average and the dwelling type. This already gives a decent fit, with an r^2 of 0.656. Observe that the accuracy is lower than 95%, which indicates that the assumptions required for generating the prediction interval are not completely met.
- Adding the floor area and number of floors as independent variables didn't contribute: the r^2 value even decreased very slightly.
- Older buildings can be considered 'outliers' with regard to their construction year, and thus have a disproportionate effect on the generated regression. Since buildings before 1900 are likely not to differ very much from building built around 1900 – since there has been plenty of opportunity for renovation and refitting – we consider a variant where we 'cap' all construction years at 1900: if a building is older than that, we give the prediction for the building as if was from 1900. Expectedly, this improves the fit.
- We then manipulate the output variable: we never really expected the EPI to be linear (in correlation to a 'true' energy performance) to begin with. Therefore, we also tried the inverse and the natural logarithm of the EPI. And indeed: the combination with the highest r^2 turns out be the one where we take log(EPI) as our output, and also use the PC6 average of the log(EPI) as independent variable. Note that there are energy labels with and EPI below 0 in the database, but they were so few that we considered it appropriate to filter them out.

To calculate these results, we used the Python module `statsmodel`. The script to generate the results is available in our public repository[2].

Table 4.1: Results of the multiple linear regression to predict the energy label class, for various combinations of independent and dependent (output) variables. The accuracy indicates the share of correct predictions, calculated by comparing the measured values for dwellings in the energy label database with the calculated prediction intervals for these dwellings. We aim for a prediction interval that is right 95% of the time. The example prediction shows the predicted EPI, the prediction interval for the EPI and the corresponding energy label classes. The example is for a dwelling with the following properties: construction year is 2020, PC6 average EPI is 1.2 (label class B), dwelling type is a *tussenwoning*, floor area is 117 m^2 (average in the Netherlands), number of floors is 3 (modal in the Netherlands). Observe that the predicted EPI is lower than the neighbourhood average, which shows that a.o. the recent construction year leads to a higher energy label class in the prediction. This matches our intuition (and data, cf. Figure 3.3) that modern buildings have a better energy label and thus a lower EPI.

Output	Independent variables	r^2	Accuracy	Prediction
EPI	construction year, EPI PC6 average, dwelling type	0.656	92.86%	1.108 (0.455–1.761, A^{++}–D)
EPI	construction year, EPI PC6 average, dwelling type, floor area, number of floors	0.655	92.88%	1.109 (0.455–1.763, A^{++}–D)
EPI	construction year (cutoff at 1900), EPI PC6 average, dwelling type, floor area, number of floors	0.660	92.88%	1.076 (0.426–1.726, A^{++}–D)
1/EPI	construction year (cutoff at 1900), 1/(EPI PC6 average), dwelling type	0.582	97.52%	1.119 (0.812–1.802, A–D)
1/EPI	construction year (cutoff at 1900), PC6 average of 1/EPI, dwelling type	0.637	97.12%	1.134 (0.835–1.764, A–D)
log(EPI)	construction year (cutoff at 1900), log(EPI PC6 average), dwelling type	0.699	93.23%	1.107 (0.766–1.600, A–C)
log(EPI)	construction year (cutoff at 1900), PC6 average of log(EPI), dwelling type	0.707	93.20%	1.112 (0.773–1.601, A–D)

Going forward, we take the variant with the highest r^2 value, and the resulting regression formula becomes:

$$\begin{aligned}\widehat{\log(\text{EPI})} = {} & 3.845226 \\ & - 0.001914 \cdot \max(\text{construction year}, 1900) \\ & + 0.869052 \cdot (\text{average EPI in PC6}) \\ & - 0.023008 \text{ (if dwelling type is } \textit{meergezinspand hoog}) \\ & - 0.032551 \text{ (if dwelling type is } \textit{meergezinspand laag midden}) \\ & - 0.030663 \text{ (if dwelling type is } \textit{tussenwoning}) \\ & - 0.004443 \text{ (if dwelling type is } \textit{twee-onder-een-kap}) \\ & - 0.014757 \text{ (if dwelling type is } \textit{vrijstaand})\end{aligned}$$

After this, we want to find the prediction interval such that 95% of the true values lie in their predicted

[2]https://github.com/dutch-dwellings/dutch-dwelling-database/blob/main/analysis/energy_labels_multiple_linear_regression.py

ranges. Since the default accuracy of the formula is 93.20% (see Table 4.1), we need to widen the predicted intervals a little my multiplying with a calibration factor $c_{calibration}$. The prediction interval then becomes:

$$\widehat{y_h} \pm c_{calibration} \cdot t_{(1-\alpha/2,n-p)} \cdot \sqrt{s^2 + \boldsymbol{x_h}^T S \, \boldsymbol{x_h}}$$

From our regression analysis, we find $s^2 = 0.0345$. The 8×8 covariance matrix S is given in Appendix E. By trial and error, we find that for $c_{calibration} = 1.1436$, our prediction interval has an accuracy of 95.28% on the energy label data set. That is slightly larger than 95%, and thus acceptable.

4.5 Comparison modules

The aim of the comparison modules is to find a way to modify probabilities of dwellings having a certain kind of heating installation. For example, a dwelling with a high gas consumption could have a higher probability of having a gas boiler than one with a very low gas consumption. However, such a simple comparison will not work. If done this way, gas boiler allocation would be biased towards larger dwellings, as they use more gas to be able to heat more space. Thus, another approach must be used.

In these modules, we use a two step process which consists of two comparisons. First, either the gas or electricity usage of dwellings get compared against national consumption data, and afterwards the outcomes of those comparisons get compared against each other. The first comparison accounts for the fact that dwellings with different characteristics have different consumption standards. The second makes sure that we get a variable that allows for comparison between dwellings in a neighbourhood to ultimately be able to modify the heating installation probabilities.

The modification of the probabilities is done using four formulas, which map a base probability P_{base}, which is found in input data, to another probability, P, on the basis of the percentile ranking of a dwellings energy consumption in its neighbourhood R.

When P_{base} is greater than 0.5 and a higher consumption should increase the probability we use:

$$P = P_{base} + 2 \cdot (1 - P_{base}) \cdot (R - 0.5) \tag{4.5.1}$$

When P_{base} is lower than 0.5 and a higher consumption should increase the probability we use:

$$P = P_{base} + 2 \cdot P_{base} \cdot (R - 0.5) \tag{4.5.2}$$

When P_{base} is greater than 0.5 and a higher consumption should decrease the probability we use:

$$P = P_{base} - 2 \cdot (1 - P_{base}) \cdot (R - 0.5) \tag{4.5.3}$$

When P_{base} is lower than 0.5 and a higher consumption should decrease the probability we use:

$$P = P_{base} - 2 \cdot P_{base} \cdot (R - 0.5) \tag{4.5.4}$$

4.5.1 Gas usage comparison

In this module, the gas consumption of a dwelling is compared against that of other dwellings in its neighbourhood. However, the highest resolution of gas consumption data available to us is on a PC 6 spatial scale. The crux is that, because we have national usage data split out by several building characteristics, we can estimate whether a certain gas consumption is high or low for that type of building and thus differentiate between dwellings.

The first comparison uses the PC6 data as input data, as this is the lowest spatial scale we have gas consumption data on. As the average gas consumption per dwelling per year and the amount of dwellings in a postal code are known, we can calculate the total gas consumption of a postal code in a year. From the BAG we can sum the floor areas of all the buildings located in the postal code to give the total floor

area. We can use these quantities to calculate the gas consumption per floor space:

$$gas\ use\ per\ floor\ space\ in\ postal\ code = \frac{total\ gas\ use\ in\ postal\ code}{total\ floor\ area\ in\ postal\ code} = \left[\frac{m^3}{m^2}\right] \tag{4.5.5}$$

Here, we assume that the gas consumption per floor area is the same for all dwellings in the postal code. The unit of this quantity is the same as the one in the "CBS83878 Aardgaslevering vanuit het openbare net" dataset, which is used as a benchmark to compare the gas consumption against. This dataset contains gas consumption data for the 5th, 25th, 50th, 75th, and 95th percentiles of dwellings, split out by energy label, building type, construction year and floor area. This differentiation is how the same gas use per square meter can still give a different relative gas consumption between dwellings.

For every dwelling in the neighbourhood, the benchmark data associated with the four dwelling characteristics mentioned earlier is found and the data is interpolated between the percentiles given. This is done to allow for a more precise estimation of relative gas consumption, instead of only knowing in which range the dwellings usage lies. After the interpolation, the benchmark function is saved so that when another dwelling with the same characteristics passes through the module, the interpolation does not have to be done again.

After the benchmark function has been either constructed or retrieved, the gas consumption per square meter of the postal code (and thus of the dwelling) is fed into the function which outputs a number between zero and one. This is the dwelling gas usage percentile compared to the national data. A higher number indicated a higher consumption relative to other dwellings with the same energy label and building type and which floor area and construction year lie in the same range, as this percentile denotes the percentage of dwellings that have a lower gas consumption.

Once all dwellings in a neighbourhood have passed through this first comparison, a second comparison is made. This time, the comparison is between the numbers found in the first comparison. This comparison is done on the neighbourhood level because this is the spatial scale of which we know the distribution of heating installations, on which more information can be read in section 4.6. With the outcome of this second comparison we can modify the probabilities of dwellings having certain types of heating installations. The national percentiles of the dwellings are sorted into a list and for each dwelling a new percentile is found using:

$$neighbourhood\ gas\ use\ percentile = \frac{index\ in\ list}{length\ of\ list} \tag{4.5.6}$$

This means that the dwelling, which has the highest gas consumption relative to other dwellings with its characteristics will also have the highest neighbourhood percentile. However, the two rankings are not equal. It is possible to imagine a neighbourhood where all dwellings have a high gas consumption relative to other dwellings of their type, but within the neighbourhood there must still be an ordering, going from lowest (which nationally is still high) to highest.

4.5.2 Electricity usage comparison

The structure comparison module for the electricity consumption is roughly the same as that of the gas consumption comparison module. Here too there are two comparisons: first on a national basis and then within the neighbourhood. However, there are some differences in the implementation of the comparisons. This is due to the fact that the national electricity consumption data that we are using as a benchmark, the "CBS83882 Elektriciteitslevering vanuit het openbare net", has a different unit. Instead of the analogous kWh/m^2 the data is presented in kWh per person per dwelling. This necessitates the use of demographic data concerning the amount of people living in a particular dwelling.

As this is private information, the best public approximation we could find was to use PC6 level data from the CBS "Kerncijfers per postcode". The average number of residents in a postal code, rounded to the nearest positive integer, is taken to be the number of residents for every dwelling in that postal code. Again, as in the gas comparison module, the total electricity consumption and the total floor space in

the postal code are calculated and a value for the electricity consumption per square meter is found.

$$electricity\ use\ per\ floor\ space = \frac{total\ electricity\ use\ in\ postal\ code}{total\ floor\ space\ in\ postal\ code} = \left[\frac{kWh}{m^2}\right] \tag{4.5.7}$$

Then, for every dwelling in the neighbourhood, this electricity consumption per square meter is multiplied with the floor area of the dwelling in question to yield the electricity use of that dwelling. This electricity usage is divided by the amount of persons living in the dwelling to compute a value that can be compared against the national data.

$$electricity\ use\ per\ person = \frac{electricity\ use\ per\ floor\ space \cdot floor\ space}{persons\ living\ in\ dwelling} = \left[\frac{kWh}{person}\right] \tag{4.5.8}$$

The benchmark data is also different from its gas consumption counterpart in that the other dwelling characteristics are used to make a distinction between dwellings. Instead of energy label, construction year, floor area and building type, now building type, floor area and amount of inhabitants are used. The first two are available from the BAG, while the third is taken to be the same as the quantity previously divided by.

Using these characteristics, again a benchmark function is created or retrieved and an electricity usage percentile is assigned to the dwelling. These percentiles are compared in the neighbourhood in the same way as in the case of the gas consumption comparison.

4.6 Space heating

For the space heating, water heating and cooking modules, the CBS84983; Woningen; hoofdverwarmings; buurt 2019 dataset is the most important input. The dataset provides the percentage of dwellings in a neighbourhood within eight categories. The categories, along with the implications the categorisation has for space heating installations, can be found in Table 4.2. As mentioned in paragraph 3.3.3, this categorisation was developed by CBS and is based on energy labels (if available) and gas and electricity consumption data. In the space heating modules, the percentages of dwellings in a neighbourhood that have a certain type of installation is taken to be the base probability for a dwelling to fall into one of the eight categories. From this assumption, the probabilities are modified based on other data available. The specifics of these processes will be detailed in the sections below.

It should be noted that the modules calculate the probability that a certain type of installation is used in a dwelling for a specific purpose. A dwelling could get assigned a gas boiler for space heating as well as for water heating, but this only indicates that a boiler is used for both functions, not that there are two boilers present.

Table 4.2: Space heating application assignments based on CBS84983 data set

Cat.	Condition	Gas Boiler	Block Heating	District Heating	Electric Heating	Heat Pump
1	DH* with high gas use	backup	backup	yes	no	possible
2	DH with low gas use	no	no	yes	no	possible
3	DH without gas use	no	no	yes	no	no
4	EH with high gas use	no	no	no	no	Hybrid
5	EH* with low gas use	no	no	no	yes (or)	yes (or)
6	EH without gas use	no	no	no	yes (or)	yes (or)
7	Individual gas boiler	yes	no	no	no	possible
8	Block heating	no	yes	no	no	possible

* DH = District Heating; EH = Electric Heating

4.6.1 District space heating

For space heating through district heating, the probabilities associated with the categories "DH with high gas use", "DH with low gas use" and "DH without gas use" are summed to find the total probability of a dwelling having district heating, as per equation 4.6.1. If the gas use of the postal code the dwelling is located in is zero, then the probabilities of the high and low gas consumption types are set to zero.

$$P_{DH} = P_{DH\ high\ gas\ use} + P_{DH\ low\ gas\ use} + P_{DH\ without\ gas\ use} \tag{4.6.1}$$

Attempts were made to modify the probabilities based on gas consumption, but due to the fact that the individual probabilities are summed, the modification formulas did not balance each other out, creating probabilities which could be greater than one.

4.6.2 Gas space heating

In the gas space heating module the probabilities of a dwelling using either a gas boiler or block heating for space heating are computed using equations 4.6.2 and 4.6.3. The base probabilities of category 7 (Individual gas boiler) and category 8 (Block heating) are both augmented with each half of the percentage of dwelling in category 1 (District heating with high gas use). This is done because dwellings in this category can use either a gas boiler or block heating to supply additional heat in the colder months. As no data was found on the percentage of dwellings in this category that has either, the probability is split in half.

$$P_{gas\ boiler\ space} = P_{individual\ gas\ boiler} + {}^1\!/_2 \cdot P_{DH\ high\ gas\ use} \tag{4.6.2}$$

$$P_{block\ heating\ space} = P_{block\ heating} + {}^1\!/_2 \cdot P_{DH\ high\ gas\ use} \tag{4.6.3}$$

Afterwards, the probabilities are modified according to the gas use percentile in the neighbourhood, where a higher percentile increases the probability of having a gas fired installation and vice versa, following equations 4.5.1 and 4.5.2.

4.6.3 Electric space heating

In this module, the probabilities of a dwelling having either an electric heat pump, a hybrid heat pump or an electric boiler are computed.

For the electric heat pump, from the WoON survey we found that 1.55% of dwellings in the Netherlands have an electric heat pump. This would give every dwelling a 1.55% probability of having such a heat pump. However, we assume that only dwellings with an energy label of C or better will have an electric heat pump (Niessink, 2019). This reduces the amount of eligible dwellings. These dwellings have to be assigned a higher probability of having an electric heat pump, as otherwise too few heat pumps would be assigned in the sampling stage. Thus we modify the base probability using:

$$\begin{aligned} P_{electric\ heat\ pump} &= P_{electric\ heat\ pump\ \text{WOON}} * \frac{dwellings\ in\ the\ Netherlands}{eligible\ dwellings} \\ &= 0.0155 * \frac{7951730}{3359728} \\ &= 0.0367 = 3.67\% \end{aligned} \tag{4.6.4}$$

Of course, as the input variables change, the probability will change accordingly. The above probability of having an electric heat pump is the assigned to every dwelling with an energy label of C or better, while all other dwellings get assigned a probability of zero.

The probability of a dwelling having a hybrid heat pump is the same as the percentage of dwellings in the neighbourhood that have such a device. However, if the electricity use of the dwelling is high relative to dwellings of the same type, we modify the probability of the dwelling having a hybrid heat pump based on the relative gas consumption, with a higher gas consumption leading to a higher probability. This is done because of all the dwellings that have a high electricity usage, those with a hybrid heat pump will also have a relatively high gas consumption due to the built in boiler.

Equation 4.6.5 shows that the probability of a dwelling having an electric boiler is the sum of the probabilities associated with categories 5 (Electric heating with low gas use) and 6 (Electric heating without gas use). The sum of these two probabilities is then modified according to the electricity consumption within the neighbourhood. A higher electricity consumption will increase the probability of the dwelling having an electric boiler, and vice versa, so equations 4.5.1 and 4.5.2 are used.

$$P_{electric\ boiler\ space} = P_{electric\ low\ gas\ use} + P_{electric\ without\ gas\ use} \tag{4.6.5}$$

4.7 Water heating

Table 4.3 provides a summary of the CBS84983; Woningen; hoofdverwarmings; buurt 2019 data concerning water heating. These relationships are used in the water heating modules that are described below.

Table 4.3: Water heating application assignments based on CBS84983 data set

Category	Condition	Gas Boiler	Block Heating	District Heating	Electric Heating	Heat Pump
1	DH with high gas use	backup	no	possible	no	possible
2	DH with low gas use	yes	no	possible	no	possible
3	DH without gas use	no	no	possible	yes	possible
4	EH with high gas use	yes	no	no	no	no
5	EH with low gas use	yes	no	no	no	no
6	EH without gas use	no	no	no	yes (or)	yes (or)
7	Individual gas boiler	yes	no	no	no	possible
8	Block heating	possible	yes	no	no	no

4.7.1 District water heating

In the case of district heating there is no category of dwellings that definitely uses this type of heating for water heating. However, using district heating for water is only possible if there is a district heating network in place. Thus we take the probability of a dwelling using district heating for water to be the same as the probability of it using district heating for space heating, but modify this probability based on the relative gas and electricity usage. If the gas or electricity consumption is higher than average, the probability of using district heating for water heating decreases, as the likelihood that an alternative technology is used is greater. This means that equations 4.5.3 and 4.5.4 are used.

4.7.2 Gas water heating

The gas water heating module calculates the probabilities of dwellings using either gas boilers or block heating for water heating according to equations 4.7.1 and 4.7.2.

$$\begin{aligned} P_{gas\ boiler\ water} = P_{individual\ gas\ boiler} + P_{electric\ high\ gas\ use} + P_{electric\ low\ gas\ use} \\ + 1/2 \cdot (P_{block\ heating} + P_{DH\ high\ gas\ use} + P_{DH\ low\ gas\ use}) \end{aligned} \tag{4.7.1}$$

The probability of the dwelling using a gas boiler for water heating comprises the probabilities of categories 7 (Individual gas boiler), 4 (Electric heating with high gas use), and 5 (Electric heating with low gas use). And half of the the probabilities of categories 8 (Block heating), 1 (District heating with high gas use), and 2 (District heating with low gas use). The reason that for some of these categories only half the probability get added has to do with the fact that there are multiple water heating options that dwellings in these categories can use. Thus, the probability associated with that category needs to be spread out over those options. In the case of simple addition, the probabilities would not sum to one over all dwellings in the neighbourhood and some installations would be assigned more than there should be.

$$P_{block\ heating\ water} = 1/2 \cdot P_{block\ heating} \tag{4.7.2}$$

The probability of a dwelling using block heating is taken to be half of the probability of the dwelling using block heating for space heating. The probability is spread out over using block heating and using a gas boiler because dwellings in this category can use both.

4.7.3 Electric water heating

For electric water heating there are two options: the electric boiler and an electric heat pump. The probability of a dwelling using an electric boiler for water heating is equal to the sum of the probabilities of the dwelling being in category 3 (District heating without gas use) or 6 (Electric heating without gas use) as shown in equation 4.7.3. This total probability is then modified in accordance with the electricity consumption relative to other dwellings in the neighbourhood, where a high consumption increases the probability, which means equations 4.5.1 and 4.5.2 are used.

$$P_{electric\ boiler\ water} = P_{DH\ without\ gas\ use} + P_{electric\ without\ gas\ use} \tag{4.7.3}$$

The probability of a dwelling using a heat pump for water heating is set equal to the probability of the dwelling using a heat pump for space heating.

4.8 Insulation

The insulation module aims to accurately estimate the thermal resistance (R) in $m^2\ K\ W^{-1}$ of the four following insulation areas for individual dwellings:

1. Roof
2. Floor

3. Window
4. Façade

To estimate the thermal resistance, we first assign a base probability range of the R-value for the different insulation areas based on the "Besparingskentallen voor besparing in de bestaande woningbouw" (Menkveld et al., 2009). These probability ranges are also split between the dwelling types freestanding house, terraced house, corner house, and multiple household dwellings. The dwelling type “Two under 1 roof” is categorised as a corner house in this data set. The data from this source does not distinguish different types of multiple household dwellings, so all multiple household dwellings are assigned the same probability ranges of the R-value.

As older dwellings are generally poorly insulated, we also look at the probability of an insulation measure taking place (“Insulation per construction year”, 2021). A study by the RVO estimates the number of dwellings that have had an insulation measure per insulation area per year for the years 2010 to 2019 (RVO, 2021). These results can be seen in Table D.1. This allows us to determine the probability for an individual dwelling having an insulation measure on a specific insulation area in a specific year.

To increase the accuracy on a specific dwelling level we looked at the number of renovations taking place per dwelling type in the WoON survey 2018 (CBS, 2018). Here we found that freestanding houses are nearly twice as likely to have been renovated compared to the average dwelling. Based on these findings we allocated probability multipliers for all dwelling types. These multipliers can be found in Table D.2. The probability range of the R-value from the insulation measures is dependent on which area is insulated. This will be elaborated upon in the respective sections of the insulation areas.

These probability distributions however will still need to be altered based on certain requirements that dwellings legally have to comply with. The government introduced the “Bouwbesluit” or the “Building Decree” in 1992 to set building requirements for newly constructed buildings (“Bouwbesluit 2012”, 2021). One set of these requirements is on thermal resistance of a dwelling’s surface (“Bouwbesluit 2012 Artikel 5.3”, 2021). Since 1992 this Building Decree has been updated regularly to set new requirements for thermal resistance specific dwelling surfaces (“Bouwbesluit 2012 Artikel 5.3”, 2021). These requirements can be found in Table D.3. As dwellings constructed in 1992 or after will have to comply with these regulations, we recalculated the probability ranges of the R-value for dwellings constructed before 1992, and between 1992 and 2006. This is done so that all dwellings will always comply with the Building Decree.

Further data that is used to estimate the R-value of insulation areas of specific dwellings is the "Marktinformatie isolatiematerialen" or market information on insulation materials, for information on the average installed R-value of insulation materials for the years 2010 - 2019 (“Marktinformatie isolatiematerialen, isolatieglas en HR-ketels 2010-2019”, 2021). This data can be seen in Table D.4. However, for the years 2010 - 2012 this research used a different method to determine the average installed R-value of mineral and organic wool than from 2013 onward. This leads to a large variation in the R-value between these years. Therefore it was decided to estimate the average R-values for 2010 - 2012 based on the average R-values from 2013 onward. These are the corrected R-values found in Table D.4. For the years 2010 - 2012 the corrected R-values were used, but from 2013 onward the original R-values were used.

For insulation measures we make the following general assumptions:

1. A dwelling will not receive an insulation measure in the first 10 years after which it is built.
2. When an insulation measure is taken for a certain insulation area, we assume that every object in this area receives the insulation measure. So, when a window insulation measure happens, all windows of a dwelling will be replaced by HR glass.
3. Since an area is fully insulated when an insulation measure takes place, it is assumed that an area will only receive an insulation measure once.

For the specific insulation areas additional assumptions are made which will be explained in their respective sections.

4.8.1 Roof insulation

For roof insulation we can determine a base probability of the R-value based on a dwelling's build year and dwelling type. These base probability distributions can be found in Table D.9 and Table D.10. After this base probability range is determined we look at the probability of an insulation measure happening. To link an insulation measure in a specific year we look at the "marktinformatie isolatiematerialen" and assign the average R-value of insulation material sold in the year of the measure taking place ("Marktinformatie isolatiematerialen, isolatieglas en HR-ketels 2010-2019", 2021). These R-values can be found in Table D.4.

As there are two categories of insulation material used in this research, we determine the probability of a measure using one of both material categories by looking at the distribution of the materials sold in each year. These probabilities are then multiplied by the probability of an insulation measure taking place in the corresponding year and this is then linked to the corresponding R-value of this material category. These steps allow us to determine a probability range of the R-value which is added to the roof area when insulation measures take place. This probability range of the R-value is then combined with the base probability range of the R-value where no measures have taken place to obtain a full probability distribution of the thermal resistance of a roof for a specific dwelling.

4.8.2 Floor insulation

The module for floor insulation is similar to the roof insulation model with the exception of the data inputs. In this model the base probability distributions of the R-value are taken from Table D.12 and Table D.13.

Due to a lack of reliable data, there is no distinction made in insulation material used between roof insulation and floor insulation. So the methods for assigning an R-value to the probability of an insulation measure are similar to the methods in the roof insulation model.

4.8.3 Window insulation

As windows within a dwelling are not necessarily all of the same window type, there can be significant variation in the R-value of windows within a dwelling. For this reason we are not able to determine a base probability range of the R-value for the windows of a specific dwelling. In order to assign a base probability to a specific dwelling for window insulation we look at the build year of a dwelling. We assume that dwellings constructed before 1974 have single glass windows, whereas dwellings constructed between 1974 and 1992 will have double glass windows ("Insulation per construction year", 2021). For dwellings constructed after 1992 we assume that the R-value of the windows is in line with the Building Decree. The base probabilities assigned to dwellings can be found in Table D.5.

We then again look at the probability of a dwelling having received an insulation measure for the windows. From the data we know that an insulation measure for windows is defined by the installation of HR glas (hoog rendement glas or high efficiency glass) (RVO, 2021). From Table D.7 we take the R-value range of HR glass and combine this with the probability of an insulation measure taking place. This gives us a probability distribution of the R-value for windows of a specific dwelling. However, since the base probability is not defined as a range over the R-value and the R-value range of HR glass is rather small, the 95% confidence interval of the R-value for a specific dwelling will likely be very large for this insulation module.

4.8.4 Façade

For the façade insulation area there are two types of insulation measures possible. The first is insulation of the cavity walls and the second is the outside insulation of a façade. It is assumed that only dwellings constructed after 1920 are equipped with cavity walls and therefore are eligible for an insulation measure of this type ("Insulation per construction year", 2021). This is also calculated through to the probability distribution over the R-value. The dwellings constructed before 1920 are assigned to the lowest base R-value as there is no option to insulate a cavity wall. All dwellings considered in this module are eligible for the measure of outside insulation of the façade.

For the façade insulation model we take the base probability distribution over the R-value for a specific dwelling from Tables D.15, D.16 and D.17. Then we look at the probability of an insulation measure in the cavity wall and link this probability with an R-value range of 1.25 to 2.175 m^2 K W^{-1}. This range is based on the range of R-values for different insulation materials with a thickness of 0.05 m **(Source)**. As the insulation material is added to the cavity wall this R-value range is then added to the original R-value probability distribution of a specific dwelling. This will give us a new probability distribution over the R-value which will be used in the next step.

This step is to look at the probability of an insulation measure on the outside of the façade. Here we link the annual probabilities with the average installed R-value of insulation materials from "Marktinformatie isolatiematerialen" in Table D.4 ("Marktinformatie isolatiematerialen, isolatieglas en HR-ketels 2010-2019", 2021). This is similar to the roof and the floor insulation modules. The R-values corresponding to these probabilities of insulation measures are then again added to the probability distribution over the R-value which was computed in the previous step. This will then ultimately output a probability distribution over the R-value for the façade of a specific dwelling.

4.9 Cooking

Similar to the space and water heating modules, data set CBS84983 can be used to establish the conditional relationships between the presence of a heating device, electricity and gas use and the type of cooking appliance used in that dwelling. Table 4.4 lists these conditional relationships.

Table 4.4: Cooking application assignments based on CBS84983 data set.

Category	Condition	Gas	Electric
1	DH with high gas use	possible	possible
2	DH with low gas use	yes	possible
3	DH without gas use	no	yes
4	EH with high gas use	yes	no
5	EH with low gas use	yes	no
6	EH without gas use	no	yes
7	Individual gas boiler	yes	no
8	Block heating	yes	no

4.9.1 Gas cooking

The gas cooking module sums the base probabilities of all six categories that have some amount of gas consumption, as can be seen in equation 4.9.1.

This resulting probability is not modified according to the gas use, as gas use for cooking is only a small part of the gas consumption of the average dwelling, and thus total gas use is not a good indicator for assigning this type of installation.

$$\begin{aligned} P_{gas\ cooking} = P_{individual\ gas\ boiler} + P_{electric\ high\ gas\ use} + P_{electric\ low\ gas\ use} \\ + P_{block\ heating} + P_{DH\ high\ gas\ use} + P_{DH\ low\ gas\ use} \end{aligned} \tag{4.9.1}$$

4.9.2 Electric cooking

In the electric cooking module, the probabilities of the other two categories are summed to compute the electric cooking probability. These two categories are "District heating without gas use" and "Electric heating without gas use", shown in equation 4.9.2. This probability is not modified either,

for an analogous reason. The electricity required for cooking makes up such a small part of the total consumption, that looking at the total consumption will not work to predict whether a dwelling has an electric cooking installation.

$$P_{electric\ cooking} = P_{DH\ without\ gas\ use} + P_{electric\ without\ gas\ use} \tag{4.9.2}$$

4.10 Sampling

The sampling module is the final module a dwelling passes through. In this module, all the probabilities that have been computed in the previous modules are sampled to create the final output data, whether a dwelling has a certain heating installation or a certain type of insulation.

There are some conditions applied to the sampling. For example, we make sure that for each energy function (space heating, water heating and cooking) at least one installation is present in each dwelling. Without this constraint and due to the fact that we do a random sampling against the probabilities, there could be dwellings that do not have an installation for one or more of these functions. We assume that every dwelling in the Netherlands has some way to satisfy these basic needs, so we enforce the constraint.

Another constraint enforced is that district heating, electric heat pumps and block heating can only be used for water heating when they are used for space heating. The reasoning behind this choice is that the primary function of these installations is to heat space, with water heating being a possible additional use case.

The sampling modules outputs a code for each energy function which corresponds to the installations present in the dwelling.

For the insulation characteristics, the sampling module samples the probability distributions found in the insulation module. It outputs two variables per distribution: the mean R-value and the 95% confidence interval R-values.

4.11 Human behaviour

Our approach to assessing the impact of human behaviour consisted of a literature review, and some statistical tests on available data for the Netherlands. The literature review was adapted from van den Brom (2020). For the statistical tests we looked at the CBS Kerncijfers database from 2018. This database contains demographic data on the Netherlands on a PC4 level.

The reason we looked at demographic data and not directly at behavioural data is one born from pragmatism. Demographic data is easily gathered and as such widely available. Data on behaviour is significantly harder to gather and as such is very sparsely available and not for larger groups of individuals. Therefore, if we can find ways to use demographic data to predict behavioural patterns, we can more easily analyse the impact of behaviour on energy use. This also makes research for other researchers into this topic easier.

Because the format of the Kerncijfers database was slightly difficult to work with and due to time constraints, the analysis was performed by loading the dataset in a spreadsheet program (libreoffice calc to be exact) and using the built in 'correlate' function to calculate correlation coefficients. In some cases variables had to be recoded into different variables because the kerncijfers dataset usually gives the absolute number of dwellings or people in a certain category and for our analysis we usually wanted to know the share of the total that a specific category comprised. To this end specific variables were recoded into new ones by simple division. This is also explained in Appendix B.

5 | Results

Our results are two folded. PBL not only required an output dataset, but also a well-documented methodology on how the output data was produced. We have therefore meticulously documented our methods in this report, thoroughly commented our coding and allowed for open access to our PostgreSQL database. Table 5.2 comprises an overview of all the outputs of the model. The name, PostgreSQL data type and a short description are given for each output variable. An example for one dwelling can be found in Table 5.3. The variables "space_heating", "water_heating" and "cooking" have a code as their output. This code relays information about the installations present. The conversion from installation to code can be found in Table 5.1. The codes for space heating, water heating and cooking start with "sh", "wh" and "co" respectively. If a dwelling has multiple installations for the same energy functions, the codes are concatenated using an underscore, like for example "sh01_sh02" in the space heating category.

Table 5.1: Conversion form heating installations to output code.

Installation	Code
district_heating_space	sh01
gas_boiler_space	sh02
block_heating_space	sh03
electric_heat_pump_space	sh04
electric_boiler_space	sh05
hybrid_heat_pump_space	sh06
district_heating_water	wh01
gas_boiler_water	wh02
block_heating_water	wh03
electric_heat_pump_water	wh04
elec_boiler_water	wh05
gas_cooking	co01
electric_cooking	co02

Table 5.2: Output variables and details.

Output name	Data type	Description
vbo_id	character (16)	dwelling identification (verblijfsobject identificatie)
energy_label_epi_mean	double precision	Mean value of EPI distribution
energy_label_epi_95	numrange	95 percentile interval of EPI
energy_label_class_mean	energy_label_class	Energy label class
energy_label_class_95	energy_label_class_range	95 percentile interval of energy label class
district_heating_space_p	double precision	Prob. of dwelling using district heating for space heating
gas_boiler_space_p	double precision	Prob. of dwelling using a gas boiler for space heating
block_heating_p	double precision	Prob. of dwelling using block heating for space heating
hybrid_heat_pump_p	double precision	Prob. of dwelling using a hybrid heat pump for space heating
electric_heat_pump_p	double precision	Prob. of dwelling using an electric heat pump for space heating
elec_boiler_space_p	double precision	Prob. of dwelling using an electric boiler for space heating
insulation_facade_r_mean	double precision	Mean value of R-value distribution for the facade
insulation_facade_r_95	numrange	95 percentile interval for R-values for the facade
insulation_roof_r_mean	double precision	Mean value of R-value distribution for the roof
insulation_roof_r_95	numrange	95 percentile interval for R-values for the roof
insulation_floor_r_mean	double precision	Mean value of R-value distribution for the floor
insulation_floor_r_95	numrange	95 percentile interval for R-values for the floor
insulation_window_r_mean	double precision	Mean value of R-value distribution for the windows
insulation_window_r_95	numrange	95 percentile interval for R-values for the windows
district_heating_water_p	double precision	Prob. of dwelling using district heating for water heating
gas_boiler_water_p	double precision	Prob. of dwelling using a gas boiler for water heating
block_heating_water_p	double precision	Prob. of dwelling using block heating for water heating
elec_boiler_water_p	double precision	Prob. of dwelling using district heating for water heating
electric_heat_pump_water_p	double precision	Prob. of dwelling using district heating for water heating
gas_cooking_p	double precision	Prob. of dwelling using gas for cooking
electric_cooking_p	double precision	Prob. of dwelling using electricity for cooking
space_heating	character varying	Code describing space heating installation(s) present
water_heating	character varying	Code describing water heating installation(s) present
cooking	character varying	Code describing cooking installation(s) present

Table 5.3: Example of output for a single dwelling

Output name	Unit	Value
vbo_id	-	"0003010000125985"
energy_label_epi_mean	-	"1.149663070501886"
energy_label_epi_95	-	"[0.7580889452333849,1.7434961741447357]"
energy_label_class_mean	Energy label	"B"
energy_label_class_95	Energy label	"[D,A]"
district_heating_space_p	-	0
gas_boiler_space_p	-	0.927105410821643
block_heating_p	-	0.170660921843687
hybrid_heat_pump_p	-	0
electric_heat_pump_p	-	0.039221404034656
elec_boiler_space_p	-	0
insulation_facade_r_mean	K m^2/W	2.85101808042594
insulation_facade_r_95	K m^2/W	"[2.53,5.83]"
insulation_roof_r_mean	K m^2/W	3.06326857607239
insulation_roof_r_95	K m^2/W	"[2.53,5.93]"
insulation_floor_r_mean	K m^2/W	3.0216858707121
insulation_floor_r_95	K m^2/W	"[2.53,5.93]"
insulation_window_r_mean	K m^2/W	0.399795649004097
insulation_window_r_95	K m^2/W	"[0.33,0.61]"
district_heating_water_p	-	0
gas_boiler_water_p	-	0.9506749498998
block_heating_water_p	-	0.05445
elec_boiler_water_p	-	0
electric_heat_pump_water_p	-	0.039221404034656
gas_cooking_p	-	0.9405
electric_cooking_p	-	0
space_heating	-	"sh02"
water_heating	-	"wh02"
cooking	-	"co01"

5.1 Space and water heating results validation

As mentioned in paragraph 4.1, 123 629 dwelling data points found in an online viewer could be used to partially validate our space and water heating module outputs. From Table 5.4 and 5.5 it can be seen that our space heating predictions have an overall accuracy of 67% while our water heating results are overall 66% accurate. The amount of predictions is higher than the amount of dwellings because we

allow multiple installations per dwelling to be present. The dataset contained no information on space heating via electric boilers, so these cannot be included in this validation.

Table 5.4: Space heating prediction accuracy

Heating Method	Number of Predictions	Times Correct	Accuracy
District heating	16 841	11 552	69%
Gas boiler	104 790	75 674	72%
Block heating	12 802	5108	40%
Hybrid heat pump	1703	601	35%
Electric heat pump	3618	262	7%
Total	139 754	93 197	67%

Table 5.5: Water heating prediction accuracy

Heating Method	Number of Predictions	Times Correct	Accuracy
District heating	6698	3239	48%
Gas boiler	112 301	81 846	73%
Block heating	2312	252	11%
Electric boiler	8632	140	2%
Electric heat pump	155	6	4%
Total	130 098	85 483	66%

5.2 Insulation validation

To validate the insulation predictions the same data set was used as in the case of the heating installation validation. From all 123 629 dwellings, those with their glazing type known were used to validate the glass insulation predictions, the results of which are shown in Table 5.6.

Most of the information concerning the insulation of other parts of the dwelling in the data set only describes whether an extra insulation measure has taken place. However, a subset of these entries also describe the known insulation value. These R-values are always 3 K m^2/W. We compared the generated insulation values against these known values to calculate the accuracies shown in Table 5.7.

Table 5.6: Glass insulation prediction accuracy.

	Glazing type			
	Single glazing	**Double glazing**	**HR glass**	**Triple glazing**
R-value [K m^2/W]	0.175	0.333	0.625	0.833
Dwellings	30 837	65 794	53 683	6773
In 95 percentile range	23 926	62 409	50 087	2380
Accuracy	76%	95%	93%	35%
Within 0.05 of mean	13 946	59 390	50 237	2380
Accuracy	45%	90%	94%	35%

Table 5.7: Insulation prediction accuracy. The mean and range are attributes of the R-value probability distributions, which are reported in the output of the pipeline.

Component	**Dwellings**	**In 95 percentile range**	**Accuracy**	**Within 0.1 of mean**	**Accuracy**
Façade	8 990	3 996	44%	3 898	43%
Roof	4 879	2 770	56%	1 613	33%
Floor	3 943	1 960	49%	1 449	37%

5.3 Correlations in WoON survey

Since we intended to use regression models to generate a lot of our unknown data we first tried to find correlations in the data that we have. Because legal minimum building requirements for insulation have changed over the years it might be possible that insulation correlates with building year. To investigate we used the WoON survey data, specifically the derived values for insulation of groups of components. The WoON survey contains detailed measured technical information on many building components (e.g. insulation of window 1, window 2, window 3, etc.) and derived values for the value of a group on components (e.g. insulation of the windows in that dwelling).

When performing the appropriate test for correlation on all of these derived variables against build year no strong correlations can be found. Some mild and even moderate correlations can be observed, in all of these cases the Spearman test yields a higher score than the Pearson test, indicating that a non-linear model fits the data better. Notable are the correlations between outer wall insulation and build year, which are (Pearson/Spearman) (0.442/0.611), floor insulation and build year, which is (0.420/0.591), and build year and energy index, which is (-0.477/-0.719).

SPSS output for all of these tests can be found in Appendix F. In addition to that the appendix also contains scatter plots for these variables. An interesting thing that can be observed from these plots is that a suspicious amount of datapoints take the values 0 or 100 in the derived variables for insulation. Presumably this is because of the way the derived variables are calculated, however, the WoON survey's documentation does not sufficiently elaborate on the precise calculation method to allow for conclusive statements on this. This 'quirk' in the data of most points being either 0 or 100 does cast doubt on the suitability of these variables for use in analysis, and without knowing the calculation method that generated these results, it inspires hesitancy to use this data.

6 | Discussion

6.1 Main findings

Even though there is insufficient data currently available to fully describe the energy characteristics of an individual dwelling, it is possible to predict, with a level of accuracy, some energy characteristics with the data that is publicly available. We have devised a methodology and developed code that can, through multiple linear regression and the aggregation of publicly available data, predict a dwelling's space heating, water heating and cooking characteristics. Furthermore, our model can also advise on the levels of insulation present and the energy label of that dwelling. Our data output is also not as privacy-sensitive as other datasets on this spatial scale. This is because the inputs are publicly available, higher spatial scale data.

Although opportunities for validation are scarce, as there is no data set publicly available that incorporates the energy characteristics at a dwelling level for many dwellings, we can consider the limited validation we have been able to do.

For space heating installations, we have seen that the accuracy of the model is highest when it comes to gas boilers and district heating. Space heating through the use of electricity is predicted worse. This might be due to the fact that district heating is per definition highly localised, and for gas boilers we have devised the comparison methodology to be able to localise the installations. In the model, all dwellings have a base percentage of having an electric heat pump installed, which means that while the total amount of heat pumps assigned will be correct, the allocation is random and thus not localised.

The results for water heating show that for most installations the accuracy is lower than for space heating. One notable installation is the electric boiler, which only has an accuracy of 2%, while using the electricity consumption comparison method of assignment. One possible reason for the lower accuracy is that the data used, the CBS84983, categorises dwellings based on space heating installation, with each category allowing for multiple ways of water heating. This means that dwellings have lower probabilities for multiple installations, leading to less localised, and thus less accurate, results.

Concerning the prediction of the R-values of glazing, the results are accurate for double glazing and HR glass, but fall of at the high and low end of the spectrum. The inaccuracy of HR+ and HR++ predictions are caused by the lack of data on installations of triple glazing. Therefore the module currently defines all insulation measures on windows to be HR glass. We also see that we under predict the number of dwellings with single glazing. This could be caused by the assumption that an insulation measure for windows replaces all windows in a dwelling, whereas in reality this is not always the case. This leads to an over prediction of dwellings with insulation measures, and therefore less dwellings with single glazing.

The accuracies for the prediction of other insulation types are all lower than 50%. However, the sample size of the comparison is small, and the only R-values that could be checked against were 3 K m^2/W, which is not representative for the entirety of the Netherlands.

In addition to the generated results, our program allows for large data sets to be uploaded, stored and processed with ease. Data already present in our pipeline can also easily be updated with newer versions. Especially CBS data can easily be added, as those tables are downloaded and loaded into PostgreSQL automatically.

6.2 Limitations

There are multiple limitations that we were able to identify in our project. Firstly, the lack of data at our disposal restricts the links and correlations which we still expect can be made from known data points. This lack of data therefore placed bounds on our ability to find useful correlations.

Initially, we had hoped to be able to depend heavily on the WoON survey, as the database contains detailed information concerning the energy functions of dwellings. We considered that even though the

WoON survey consists of only 4 500 entries, that this would be sufficient to apply the relationships between the characteristics to larger datasets. However, this proved to be more complex than we had anticipated as the survey does not contain spatial information and the sample size is not very large. The correlations we did find were mostly weak or moderate at best. Hence, applying these links to the larger data set would lead to large margins of error, and we decided to find alternative methods.

Secondly, any assumption that is made immediately has a significant effect on the results. To be able to draw a conclusion on the type of technologies that can be found in a particular dwelling, certain assumptions have been made regarding the consumption of natural gas and or electricity. Although efforts to differentiate consumption between dwellings were made, by incorporating floor space and building type into the calculations for example, the average usage per postal code area is often the most detailed data that is available. This means that there is an unquantified degree of uncertainty in our model which can only be reduced through the deployment of statistical spatial reduction techniques instead. One suggested method of improving the above-mentioned regression is to apply dynamic downscaling.

Linked to this, an important limitation is our inability to validate the dataset we have created, as our dataset would have to be compared to known data on specific dwellings. However, often such data is not available, and if it is, it is likely that we are already using it in our calculations. This results in a limitation as it becomes difficult to assess the degree to which the data set which we have produced is valid. All in all, we considered that such limitations are unavoidable with the data that we have. We could improve on these points if we had data at a lower spatial scale. This is, however, unlikely given the nature of European privacy legislation.

There are likely more limitations to our project, but these are the ones we have been able to identify thus far.

6.3 Recommendations for future research

As the scope of this research is limited by both time constraints and the availability of data, there are multiple avenues for future research. The first of these is to use more detailed input data for the model. In this research, often averages of variables over postal codes or neighbourhoods were used, because this is the data that is publicly available. However, with increased resolution, more accurate results could be achieved. Demographic data from the Basisregistratie Persoonsgegevens, the national database of personal data maintained by the government, could be used to know how many people live in a dwelling. Meter data from energy providers could be used to know exactly how much natural gas and electricity dwellings use, so they can be better compared against the national benchmarks as well as other dwellings. Furthermore, access to the detailed energy label input data would provide some energy characteristics on individual dwelling level. This data was requested from the RVO, but we did not manage to obtain it in time. However, all this raw data could infringe privacy laws and can therefore only be used by authorised persons under controlled circumstances. It is for that reason that this data can only serve as an input or perhaps only a validation tool and cannot be used directly as an output.

Another improvement is to statistically reduce the spatial scale of dwelling gas and electricity usages from a postal code to individual dwelling level instead of using adjusted postal code averages as we did. Performing this reduction was part of the original project scope, but due to time constraint this could not be executed. The Faculty of Social and Behavioural Sciences (FSBS) at Utrecht University (UU) has a Methodology and Statistics Consultation Shop through which external assistance with statistical work can be obtained. This service is free of charge for students from FSB, but external paid services can be arranged through them as well. Mirjam Moerbeek (m.moerbeek@uu.nl) at UU can be contacted for further assistance. Another possible future resource for students requiring support with statistical work is the Living Lab Digital Humanities in the UU Library City Centre. Nora Consulting and Topscriptie are external paid resources that provide PhD and Masters students assistance with statistical analysis while CBS themselves also offer paid services that includes assistance with conducting further statistical research, collecting new data or providing access to microdata. (All the names listed above for Living Lab Digital Humanities, Nora Consulting, Topscriptie and CBS have clickable links that will redirect to their webpage.)

The accuracy of assigning the presence of a boiler could further be improved by running the assignment through a final sampling check where the sampled dwelling's gas use is compared to the expected gas use of a gas boiler in that specific circumstances. If the sampled gas use is more than what can be expected from such a dwelling, the probability of the assignment increases and visa versa. Establishing this reference gas use will require more work as it depends on a number of factors related to that that specific dwelling. Some obvious dependencies are the volume of spaces to be heated, the space and water temperature settings, the amount of occupants, the presence of a ventilation system and how often showers and baths are taken.

Validation of the current results can also be improved if the detailed data mentioned in the first paragraph are available. These real system data can be used to assess the accuracy of our model's output by comparing one against the other.

The scope of the research could be increased by including data on photovoltaic (PV) installations. The company NEO has developed an AI that recognises PV panels in satellite images and is creating a database of all PV panels in the Netherlands[1]. The presence of PV panels is a factor of importance in the energy characteristics of a building, as it drastically reduces electricity consumption and might even supply electricity to the grid.

To make the results of this research more accessible, an online viewer could be developed. [2]

Other researchers could easily find data on specific dwellings without needing to know the BAG ID of that dwelling beforehand, and laypeople could be enthused to be interested in energy in the built environment. It should be considered that the suggestions to use increased resolution data and to make an openly accessible data viewer for the results of future research are mutually exclusive as that would likely lead to a breach of privacy law.

[1] https://zonnepanelen.neo.nl/dashboard

[2] Related projects: BAG Viewer by the Kadaster, Nationale EnergieAtlas, 'Buildings' by Waag Society (visualisation of building years from BAG), EnergieLabelAtlas by Bert Spaan (original project is offline), 3D building years by Parallel.

7 | Acknowledgments

To start off with, we would like to express our sincerest gratitude to our supervisor, Dr. ir. Ioannis Lampropoulos, for his continuous and unwavering support, guidance and insightful feedback throughout this project.

Furthermore, we would like thank Folckert van der Molen from PBL and Casper Tigchelaar from TNO for sharing their expertise and guiding us through the sea of data sources available out there. Thank you for your patient support and the invaluable brainstorming sessions.

We would also like to give a word of appreciation to the course coordinators, Mrs. Elena M. Fumagalli and Mrs. Sara Herreras Martinez, for the impeccable program they provided despite the COVID-19 restrictions. We thoroughly enjoyed the guest lectures and TenneT in-house day.

Last but not least, we would like to thank our fellow students from the Porthos group for their valuable time and all the effort they put into our peer feedback sessions. Your comments and direction were superb.

Bibliography

Anderson, J. E., Wulfhorst, G., & Lang, W. (2015). Energy analysis of the built environment—a review and outlook. *Renewable and Sustainable Energy Reviews*, *44*, 149–158.

Besluit energieprestatie gebouwen (Vol. Artikel 1.1). (2006). Directie Juridische Zaken, Afdeling Wetgeving.

Bouwbesluit 2012. (2021). https://wetten.overheid.nl/BWBR0030461/2021-04-01

Bouwbesluit 2012 Artikel 5.3. (2021). https://wetten.overheid.nl/BWBR0030461/2021-04-01#Hoofdstuk5

CBS. (2018). WoonOnderzoek Nederland (WoON). https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/woononderzoek-nederland--woon--

CBS. (2020). Energiebalans; aanbod en verbruik, sector. https://opendata.cbs.nl/#/CBS/nl/dataset/83989NED/table?ts=1620118345114

CBS, PBL, RIVM, & WUR. (2020). *Isolatiemaatregelen woningen, 1982-2018*.

Chen, J., Wang, X., & Steemers, K. (2013). A statistical analysis of a residential energy consumption survey study in hangzhou, china. *Energy and Buildings*, *66*, 193–202.

Druckman, A., & Jackson, T. (2008). Household energy consumption in the uk: A highly geographically and socio-economically disaggregated model. *Energy Policy*, *36*(8), 3177–3192.

Gram-Hanssen, K. (2013). Efficient technologies or user behaviour, which is the more important when reducing households' energy consumption? *Energy Efficiency*, *6*(3), 447–457.

Guerra-Santin, O., & Itard, L. (2010). Occupants' behaviour: Determinants and effects on residential heating consumption. *Building Research & Information*, *38*(3), 318–338.

Hoegh-Guldberg, O., Jacob, D., Taylor, M., Bindi, M., Brown, S., Camilloni, I., Diedhiou, A., Djalante, R., Ebi, K., Engelbrecht, F., Guiot, J., Hijioka, Y., Mehrotra, S., Payne, A., Seneviratne, S., Thomas, A., Warren, R., Zhou, G., & Tschakert, P. (2018). Impacts of 1.5ºC global warming on natural and human systems. *Global warming of 1.5°c*. IPCC.

Hoogenboom, J. (2018). Energielabels_corporatiewoningen_publish. https://www.arcgis.com/home/item.html?id=059c8becfd1a49f58b3707915543d701

Hoogenboom, J. (2020). *Energielabels_partverhuur_publish*. https://www.arcgis.com/home/item.html?id=e57568c80f6b4d4a93a1a6e0e196c173

Insulation per construction year. (2021). https://www.isolatieprijs.nl/blog/welk-bouwjaar-woning-bepaalt-isolatiesoort

Janssen, P., & Heuberger, P. (1995). Calibration of process-oriented models [Modelling Water, Carbon and Nutrient Cycles in Forests]. *Ecological Modelling*, *83*(1), 55–66.

Jeeninga, H., Uyterlinde, M., & Uitzinger, J. (2001). Energieverbruik van energiezuinige woningen. effecten van gedrag en besparingsmaatregelen op de spreiding in en de hoogte van het reële energieverbruik.

Kaza, N. (2010). Understanding the spectrum of residential energy consumption: A quantile regression approach. *Energy policy*, *38*(11), 6574–6585.

Koninkrijk der Nederlanden. (2014). *Nr. 3661: Regeling van de minister voor wonen en rijksdienst van 31 januari 2014, nr. 2014-0000062837, tot wijziging van de regeling energieprestatie gebouwen in verband met het vaststellen van nadere voorschriften omtrent energielabels voor utiliteitsgebouwen ter implementatie van de artikelen 3 en 11, eerste tot en met vierde lid, van richtlijn 2010/31/eu van het europees parlement en de raad van 19 mei 2010 betreffende de energieprestatie van gebouwenregeling van de minister voor wonen en rijksdienst van 31 januari 2014, nr. 2014-0000062837, tot wijziging van de regeling energieprestatie gebouwen in verband met het vaststellen van nadere voorschriften omtrent energielabels voor utiliteitsgebouwen ter implementatie van de artikelen 3 en 11, eerste tot en met vierde lid, van richtlijn 2010/31/eu van het europees parlement en de raad van 19 mei 2010 betreffende de energieprestatie van gebouwen.*

Liddament, M., & Orme, M. (1998). Energy and ventilation. *Applied Thermal Engineering*, *18*(11), 1101–1109.

Majcen, D., Itard, L., & Visscher, H. (2016). Actual heating energy savings in thermally renovated dutch dwellings. *Energy Policy*, *97*, 82–92.

Marktinformatie isolatiematerialen, isolatieglas en HR-ketels 2010-2019. (2021). https://www.rvo.nl/sites/default/files/2021/02/marktinformatie-isolatiematerialen-isolatiegas-en-hr-ketels-2010-2019.pdf

Menkveld, M., Sipma, J., Leidelmeijer, K., & Cozijnsen, E. (2009). Besparingskentallen voor besparing in de bestaande woningbouw.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2018). *Regeling van de minister van binnenlandse zaken en koninkrijksrelaties, van 25 juni 2018, nr. 2018-0000380848 tot wijziging van de regeling basisregistraties adressen en gebouwen in verband met de modernisering van de basisregistratie*. Rijksoverheid.

Nationaal Expertisecentrum Warmte. (2013). *Warmte en Koude in Nederland*. https://www.rvo.nl/sites/default/files/Warmte%20en%20Koude%20NL%202NECW1202%20jan13.pdf

Niessink, R. (2019). Air source heat pump. https://energy.nl/en/search/?fwp_content_type=factsheets

O'neill, B. C., & Chen, B. S. (2002). Demographic determinants of household energy use in the united states. *Population and development review*, *28*, 53–88.

Palmborg, C. (1986). Social habits and energy consumption in single-family homes. *Energy*, *11*(7), 643–650.

Pardoe, I., Simon, L., & Young, D. (2021). *Stat 501: Regression methods*. https://online.stat.psu.edu/stat501/lesson/welcome-stat-501

Pettersen, T. D. (1994). Variation of energy consumption in dwellings due to climate, building and inhabitants. *Energy and buildings*, *21*(3), 209–218.

Rijksdienst voor Ondernemend Nederland. (2021). *Ep-online*. https://www.ep-online.nl/

Rijksoverheid. (2019). *Klimaatakkoord*. Rijksoverheid.

RVO. (2019). *Monitor energiebesparing gebouwde omgeving 2019*. https://www.rvo.nl/sites/default/files/2021/01/monitor-energiebesparing-gebouwde-omgeving-2019.pdf

RVO. (2021). Getroffen isolatiemaatreglen. https://energiecijfers.databank.nl/dashboard/dashboard/energiebesparing

Santin, O. G. (2010). *Actual energy consumption in dwellings: The effect of energy performance regulations and occupant behaviour* (Vol. 33). Ios Press.

Santin, O. G., Itard, L., & Visscher, H. (2009). The effect of occupancy and building characteristics on energy use for space and water heating in dutch residential stock. *Energy and buildings*, *41*(11), 1223–1232.

Segers, R., Niessink, R., van den Oever, R., & Menkveld, M. (2020). *Warmtemonitor 2019*. Petten: CBS TNO.

Sonderegger, R. C. (1978). Movers and stayers: The resident's contribution to variation across houses in energy consumption for space heating. *Energy and buildings*, *1*(3), 313–324.

Steemers, K., & Yun, G. Y. (2009). Household energy consumption: A study of the role of occupants. *Building Research & Information*, *37*(5-6), 625–637.

Tigchelaar, C., Daniëls, B., & Menkveld, M. (2011). *Obligations in the existing housing stock: Who pays the bill?* Petten: ECN.

Topsector Energie. (2019). *Koudevraag in nederland en europa*.

U waarde (Glas) en R waarde (Isolatie). (2021). https://www.a1-kozijn.nl/kunststof-aluminium-kozijnen/glas-en-isolatiewaarden-waarden/

van den Brom, P. (2020). Energy in dwellings: A comparison between theory and practice. *A+ BE/ Architecture and the Built Environment*, (03), 1–258.

van der Molen, F., van Polen, S., van den Wijngaart, R., Tavares, J. L., van Bemmel, B., Langeveld, J., & Hoogervorst, N. (2021). Functioneel Ontwerp Vesta MAIS 5.0. https://www.pbl.nl/publicaties/functioneel-ontwerp-vesta-mais-50

Veelgestelde vragen energielabelverplichting woningen. (N.d.). https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/wetten-en-regels/bestaande-bouw/energielabel-woningen/veelgestelde-vragen

Visscher, H., Laubscher, J., & Chan, E. (2016). Building governance and climate change: Roles for regulation and related polices.

Vringer, K., & Blok, K. (1995). The direct and indirect energy requirements of households in the netherlands. *Energy policy*, *23*(10), 893–910.

Werner, S. (2013). District heating and cooling [Update of Werner S., District Heating and Cooling, Encyclopedia of Energy (2004), pp. 841-848.Introductory Article.]. *Reference module in earth systems and environmental sciences :*

Yohanis, Y. G., Mondol, J. D., Wright, A., & Norton, B. (2008). Real-life energy use in the uk: How occupancy and dwelling characteristics affect domestic electricity use. *Energy and Buildings*, *40*(6), 1053–1059.

Yohanis, Y. G. (2012). Domestic energy use and householders' energy behaviour. *Energy Policy*, *41*, 654–665.

Yun, G. Y., & Steemers, K. (2011). Behavioural, physical and socio-economic factors in household cooling energy consumption. *Applied Energy*, *88*(6), 2191–2200.

Zimmerman, J.-P., Evans, M., Griggs, J., King, N., Harding, L., Roberts, P., & Evans, C. (2012). *Household electricity survey: A study of domestic electrical product usage.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/208097/10043_R66141HouseholdElectricitySurveyFinalReportissue4.pdf

Appendices

A | Supplimentary data

As mentioned in the report, some data sets identified during the research proposal phase were found obsolete. The reasons are as follows:

- The 2020 Kerncijfers data below is only available on a PC4 level while the 2017 Kernsijfers data is available at a PC6 level data. We therefore rather made use of the 2017 Kernsijfers data.
- The Energieverbruiker particuliere woningen data is only available on a municipality level, while the Energielevering aan woningnen en bedrijven naar postcode provides the same energy usages on a PC6 level.
- We were not granted access to the SHAERE data.
- The Kleinverbruikersdata provided the same energy consumption data as the Energielevering aan woningnen en bedrijven naar postcode so it was not needed.
- The Nationale EnergieAtlas, WarmteAtlas and Klimaatmonitor data sets are in formats that we cannot process in ProstgresQL.

Table A.1: Data sources found in the proposal phase that were not used.

Name	Description	Organisation	Spatial level	Completeness	Year
Kerncijfers per postcode	Dutch demographics	CBS	PC4	Complete	2020
Energieverbruik particuliere woningen	Gas and electricity use of houses (includes district heating)	CBS	Municipality	Complete	2019
SHAERE	Energy characteristics of social housing	Aedes	Individual dwelling	60% of social housing stock	2020
Kleinverbruiksdata	Electricity and gas provided by grid operators	Rendo, Coteq, Liander, Enexis, Stedin, Westland Infra, Enduris	PC6 level	Complete	2020
Nationale EnergieAtlas	Maps with info on a.o. district heating networks	Government organizations	Various	Complete	Various
WarmteAtlas	Maps with info on heating	RVO	Various	Complete	Various
Klimaatmonitor	Dashboards with data on climate and energy	Rijkswaterstaat / I&W	Neighbourhood and more	Complete	Various

B | Relationship between occupant demographics, human behaviour and energy consumption

The relationship between human behaviour and energy use is one that is very important and very hard to define. The influence of occupant behaviour on the energy used in and by a building is quite significant, however, data on the behaviour of the entire population is, as of yet, unavailable. A workable proxy for data on the behaviour of the occupants of a building is data on the occupants themselves, demographic data.

Demographic data is more widely available than behavioural data, and often more reliable and of higher quality since it is much easier to collect. This is why it is useful for policy makers to know how demography can relate to energy use.

In this appendix we will first examine the existing literature on this topic and then analyse Dutch demographic data to see what findings can be done.

B.1 In literature

[This section is adapted from van den Brom (2020)]

Occupant influence on residential energy consumption is a documented phenomenon (Palmborg, 1986; Sonderegger, 1978; Steemers and Yun, 2009; (Gram-Hanssen, 2013)). That's why it is important to understand occupant behaviour when trying to predict a building's energy performance (Visscher et al., 2016). Occupant behaviour, however, is something for which large amounts of data are lacking or of insufficient scale. What we can consider are occupant characteristics i.e. demographics. Demographic data is more widely available and for that reason more useful for policy makers. What follows will be an overview of the findings of previous literature in the matter of how demographics influence energy use.

In a study in England it was found that household energy use and income correlate positively (Steemers and Yun, 2009; Druckman and Jackson, 2008). It was also found that increasing income by 1% can lead to an increase in energy consumption up to 0.63% (Vringer & Blok, 1995). This correlation was also established by Druckman & Jackson, who discovered that this effect was slightly stronger for electricity than for gas (r=0.25 & r=0.23 respectively) (Druckman & Jackson, 2008) though both correlations are weak. Household size also correlates with energy consumption, as household size increases, so does household energy consumption. Energy consumption per capita, however, goes down (Druckman and Jackson, 2008; Chen et al., 2013; Santin et al., 2009; Guerra-Santin and Itard, 2010; Jeeninga et al., 2001; Kaza, 2010; O'neill and Chen, 2002; Vringer and Blok, 1995; Yohanis et al., 2008; Yun and Steemers, 2011).

In multiple countries age was found to be the strongest acting indirect effect (Santin, 2010; O'neill and Chen, 2002; Pettersen, 1994; Yun and Steemers, 2011). Presumably this is a function of demanding comfort and average income, as people between 40 and 50 use most energy on average (Yohanis et al., 2008; Yohanis, 2012). Age of children is also an influencing factor as it was found that houses with young children ventilate less than houses with older children (Guerra-Santin & Itard, 2010).

Guera Santin (2010) also found level of education to be an influencing factor, albeit a minor one. It was found that people who had enjoyed a higher level of education set their thermostat for a lower amount of hours each day than people who had a lower level of education. Though somewhat out of the scope of this report, it has also been found that household size and the presence of teenagers significantly affect appliance energy use. Additionally it was found that tenants show a slightly stronger rebound effect than homeowners. Finally, the average energy efficiency of a house was found to effect household behaviour in that, households living in inefficient homes behaved more efficiently than households living

in efficient homes.

B.2 In our own data

When we look at the data in the Netherlands we can also see this. If we combine demographic and energy use data for example we can find some interesting results. From Table B.1 for example, we can see that a correlation exists between the value of a house and the energy that that house consumes. This correlation is moderately strong, especially between Dwelling value and electricity use.

In Table B.2 we can see that there is a correlation between household size and energy consumption as well where larger households consume more energy. This correlation is not as strong with gas consumption as it is for electricity consumption though.

Please also note that the correlation between gas and energy use is not exactly the same across these tables. This is because the Dwelling value and Persons per household variables both have a few missing cases, and these are not always the same cases. Therefore the sets of entries for gas and electricity use that were used were not entirely the same across tests and this slight difference exists.

Table B.1: Correlation values of Dwelling values, electricity use, and gas use.

	Dwelling value	**Electricity use**	**Gas use**
Dwelling value	1		
Electricity use	0.565	1	
Gas use	0.438	0.684	1

Table B.2: Correlation values of the number of persons per household, electricity use, and gas use

	Persons per household	**Electricity use**	**Gas use**
Persons per household	1		
Electricity use	0.432	1	
Gas use	0.292	0.687	1

The literature also suggests that 40-50 year olds are the greatest energy consumers. To test this we can look at the share of an area's total population that falls within an age group and see if that correlates with energy use. The results of such an analysis can be found in Table B.3. What is remarkable to see in this table is that there is a moderate positive correlation between the share of 45-65 year olds and energy use and a moderate negative correlation between the share of 25-45 year olds and energy use.This means that when an area becomes inhabited more by 45-65 year old people energy consumption is likely to increase wheres it is likely to decrease as an area becomes more inhabited by 25-45 year old people. This seems to confirm the findings in literature that middle aged people tend to consume the most energy. Of course, this simple test does not correct for many factors, an influencing factor could be that as people come nearer retirement age they earn more money and have the ability to buy larger or more wanted houses. This could cause neighbourhoods with large freestanding houses to be inhabited more by middle aged people than the national average, and as these houses tend to perform worse energetically this age group comes to use more energy.

Table B.3: Influence of the share of the total population that an age group represents in an area on the electricity and gas use in that area.

	Electricity use	**Gas use**
Share $<$ 15	0.091	-0.116
Share 15-25	0.048	0.014
Share 25-45	-0.458	-0.578
Share 45-65	0.466	0.498
Share $>$ 65	-0.011	0.235

Literature also suggest household composition affects energy usage. In Table B.4 we can see an overview of correlations that can be found in our data. In this table we can see a surprising amount of moderate correlations. We can see that as the share of one person or one parent households in an area goes up energy consumption tends to decrease, this, more so for one person households than for one parent households. Multi person and multi parent households show the opposite relationship. Perhaps this is because the former types of household tend to be smaller than the latter.

What is also striking is that the correlation of household size and energy use is rather different from the correlation between people per household and energy use from Table B.2. The cause of this could be that people per household is an indicator constructed by us by dividing the number of inhabitants by the number of households, whereas the household size indicator was already given in the data. It is, however, unusual for such a difference to cause a discrepancy of this magnitude in the results.

Table B.4: Correlation of the presence of various household compositions and energy use

	Electricity use	**Gas use**
One person households	-0.616	-0.424
Multiple person households without children	0.418	0.518
One parent households	-0.378	-0.469
Two parent households	0.600	0.339
Household size	0.606	0.379

Finally we decided to look at type of ownership and if it correlates with energy use. The results of this can be seen in Table B.5. First of all we can see that the share of multi family buildings negatively correlates with gas use. This makes sense, as multi family buildings are usually apartment buildings, where heat loss per dwelling is relatively small. Why this also affects electricity use, and why more strongly, is harder to explain.

Secondly it can also be observed that rented dwellings and bought dwellings behave almost inversely. This makes sense because they actually are each others inverse. These indicators are the percentages of dwellings in an area that are rented by their occupants or that are owned by their occupants. Under normal circumstances these percentages add up to 100%.

Finally it can also be noted that dwellings that are rented out by a renting corporation correlate not much differently with energy use than dwellings that are being rented out by private individuals. There certainly is a difference but it is rather small. Similarly the share of dwellings in an area that is uninhabited basically does not correlate with energy use.

Table B.5: Correlation of various types of ownership and energy use

	Electricity use	Gas use
Multi family buildings	-0.615	-0.501
Bought dwellings	0.627	0.602
Rented dwellings	-0.648	-0.593
Renting corporations dwellings	-0.671	-0.564
Uninhabited dwellings	-0.059	0.075

It should be noted that the one parent household indicator from Table B.4 as well as the multi family buildings, renting corporations dwellings, and uninhabited dwellings indicators from Table B.5 had a rather large number of missing values (roughly between 10 and 15%) so the accuracy of results pertaining these indicators should be considered carefully.

B.3 Replication

The above tests were performed by taking CBS PC-4 demographic data, loading it in a spreadsheet programme (libreoffice calc in this case though excel should perform similarly) then using the find and replace function to replace all instances of expunged data (given a value of -99 997) with a non numerical sign. Then run the correlate function on the appropriate variables. Some variables, such as the share of the total population in an age group had to be constructed first. In this case, that was done by dividing the amount of people in an age group by the total amount of people in an area.

Deleting all instances of expunged data made the whole analysis much simpler to perform at the cost of accuracy. When CBS expunges data this is done to protect privacy. This is only necessary when the amount of people in a pc-4 area is so small that from the uncensored data personal information could be deduced. Excluding these areas from this test by nature almost surgically excludes pc-4 areas with very few inhabitants. This can skew the results slightly and should be considered when looking at this data.

C | Method suggestions

In this appendix two methods that were, for different reasons, not accessible or fruitful in this research. The first concerns a correction to dwelling electricity consumption, separating appliance use from the total use. The second method is step-by-step pseudo code which could serve as a framework for determining dwelling geometries from the 3D BAG.

C.1 Electric space heating with correction for appliance usage

An approach that did not yield realistic results is described below. The results were not satisfactory. We believe that this had to do with discrepancies in the data, and that the methodology might be useful for later research. In this approach, we tried to separate the electricity usage due to appliances from the total electricity use. This would have given us the amount of electricity used for other functions, including space heating.

From Zimmerman et al., (2012), values for the average annual electricity consumption for several household appliances and for British dwellings with and without electric heating were gathered. The dwelling electricity consumption data is split up into different building types, which brings about a resolution increase compared to using a flat average.

First, the electricity use of the appliances that are almost certainly present in every dwelling was calculated. This was then subtracted from the electricity consumption of the household. This electricity consumption is an average of the electricity consumption of dwellings in a postal code, known at a PC6 spatial scale. This operation gives us the electricity consumption per dwelling that is available for, among other things, space and water heating. However, because we are only considering part of the electricity consumption, we cannot compare it against the benchmark anymore, as that is based on total electricity usage.

The second step was to also correct the benchmark. In Zimmerman et al., (2012), there are values given for the average electricity use for dwellings of specific types not using electricity for heating in kWh/m^2 (see Table C.1). In the module, a dwellings floor space was multiplied by the relevant electricity usage value and divided by the number of people in a dwelling. The number of people in a dwelling was taken to be equal to the average number of people per dwelling in a postcode, on a PC6 spatial scale.This operation produces the gas use per inhabitant of a dwelling, which is the same unit as the CBS benchmark. This is also average amount of energy a dwelling without any electric heating uses. This means that it could be used as a correction to the CBS benchmark data.

Table C.1: Data used for the correction of the CBS benchmark. The data concerns the annual electricity consumption of dwellings that do not use any electricity for heating (Zimmerman et al., 2012).

	Annual consumption [kWh]		
	per household	per m^2	per person
All households	3638	65	2012
Terraced house - Mid-terrace	2779	62	
Terraced house - End-terrace	3442	65	
Terraced house - Small up to $70m^2$	2894	64	
Terraced house - Medium/Large above $70m^2$	4399	52	
Semi-detached house	3847	73	
Detached house	4153	62	
Bungalow	3866	61	
Flat	2829	53	

After the correction, the new benchmark data was multiplied by the number of resident, to get the data in kWh, the same unit as the dwelling electricity use, corrected for appliance use. The two were then compared in the same way as in the case of the gas boiler. However, we found that the correction to the benchmark was far too great. Most of the time the new benchmark values came out negative, implying that most if not all of the dwellings were supplying gas for heating instead of consuming it. After some testing we found that there was a large discrepancy between the CBS usage values and the British usage value, with the British ones being 1.5 to 2 times greater.

C.2 Dwelling geometries

What follows below is a set of steps that a computer program could take to calculate, among others, the outside surface area of a building and that could possibly be extended to include roof area/shape and building volume.

We have neither the knowledge nor the skill to actually implement this, but we suggest this is a possible route a more skilled programmer could take.

Required databases:

- 2D BAG
- 3D BAG

Step 1:

Build a module that can 'see' a dwelling's shape (building in the case of apartments) and recognize which points are corner points (ends of a straight line? (would only be troublesome on buildings that have a curved part in their construction, but those are quite rare)).

Step 2:

Build a module that, when given a coordinate can look in the 3D BAG and retrieve the associated height of that point.

Step 3:

Pick a dwelling from BAG. If the dwelling has a dwellings per building < 1 refer to the building ID. If not done already perform the next steps on the building, not on the dwelling.

Step 4:

(taking heights)

Of the object selected in step 3, take the height of any corner (endpoint of a line that is not adjacent to another building) and the height of three equidistant points between these two (along said line).

Additionally, take the height of one or more points randomly withing the area of the object (probaby needed to determine roof shape)

Step 5:

(using heights to determine wall area)

If the height along a line is uniform (or within a certain margin) the wall segment can be considered to be a rectangle and its surface area is simply length x height.

If the height along a line is not uniform but linearly ascending or linearly ascending and then descending the surface area of that wall segment is (Length x Lowest Height) + (Length x ((Highest Height – Lowest Height)/2)).

[more complex roof shapes require a more complex algorithm]

When using this to determine properties of the thermal shell of a building one will still need to assume the share of outside wall that is taken up by windows.

Step 6:

(using heights to determine building volume)

If the object has a flat roof the volume of the building can be calculated as Footprint x Height.

For most slanted roofs the calculation would be (Lower Height x Footprint) + (((Higher Height – Lower Height)/2) x Footprint).

[Here too do more complex roof shapes require a more complex algorithm]

Step 7:

(determining orientation)

When given coordinates the 3D BAG can also return a vector that is at a 90-degree angle of whatever surface is at those coordinates. (3D BAG viewer does this so I assume this information can also be retrieved automatically).

If, on a slanted roof, the projection of this vector on the plane formed by the east-west axis and the north-south axis points less than 45-degrees away from south then that section of roof is south facing and could thus probably hold some solar panels.

If a roof is flat it is always suitable for solar panels.

[one can always place solar panels on a flat roof, though not always at an angle since angled panels require sturdy mounts to deal with wind forces and not all roofs are structurally sound enough to provide a mounting place. This same concern for structural soundness applies to slanted roofs as well though to a much lesser extent, not all roofs are strong enough to mount solar panels.]

D | Insulation tables

This appendix contains all data tables that were used for predicting the insulation levels of a specific dwelling. Tables D.1 and D.2 contain information on the probability of an insulation measure for a specific dwelling. Tables D.3, D.4, D.5, D.6, and D.7 contain information on R values of insulation areas and materials. Tables D.8, D.9, D.10, D.11, D.12, D.13, D.14, D.15, D.16 and D.17 contain information on the base probability distributions of the R value of their respective insulation areas. All R-values are reported in m^2 K W^{-1}.

Table D.1: Insulation measures taken per year per insulation area. Source: RVO, 2021

Year	Outside Façade	Roof Insulation	HR Glass	Cavity Wall	Floor Insulation
2010	35 858	115 398	209 353	76 784	65 807
2011	73 097	157 347	295 386	114 914	122 313
2012	7648	19 542	357 411	117 197	130 465
2013	60 548	124 367	215 387	9615	106 162
2014	84 501	155 099	266 949	131 324	138 444
2015	74 448	148 477	265 195	132 769	150 615
2016	75 802	1481	261 522	159 507	146 108
2017	85 442	164 024	259 146	15 908	164 511
2018	100 978	199 784	312 337	214 035	20 446
2019	125 197	246 325	376 869	281 276	238 257

Table D.2: Probability multipliers based on WoON 2018. Source: CBS, 2018

Dwelling type	Total weighted number of dwellings	Weighted number of renovations	Weighted percentage of dwellings with renovation	Probability multiplier
Meergezins	2 268 191	133 820	5.90%	0.88
2 Onder 1 kap	829 858	45 982	5.54%	0.83
Rijwoning tussen	1 860 359	82 831	4.45%	0.66
Vrijstaande woning	981 487	130 537	13.30%	1.98
Rijwoning hoek	1 012 225	72 908	7.20%	1.07
Total	6 952 120	466 078	6.70%	

Table D.3: Minimum R-values (thermal resistance) based on insulation area as required by the various building codes from the year 1992 to 2021. R-values are reported in m^2 K W^{-1} and U-values are reported in W m^{-2} K^{-1}. Source: "Bouwbesluit 2012 Artikel 5.3", 2021

Building Decree	Façade (R)	Roof (R)	Wall (R)	Floor (R)	Window (U)	Window (R)
BD 1992	2.5	2.5	2.5	2.5	4.2	0.24
BD 2003	2.5	2.5	2.5	2.5	4.2	0.24
BD 2012	3.5	3.5	3.5	3.5	2.2	0.45
BD 2013	3.5	3.5	3.5	3.5	1.65	0.61
BD 2014	3.5	3.5	3.5	3.5	1.65	0.61
BD 2015	4.5	6	3.5	3.5	1.65	0.61
BD 2021	4.7	6.3	3.7	3.7	1.65	0.61

Table D.4: Statistics on insulation material sold from 2010 to 2019. The R-values are the average R-values for that specific material per year. Source: "Marktinformatie isolatiematerialen, isolatieglas en HR-ketels 2010-2019", 2021

Year	Synthetics mln m^2	Synthetics R-value	Synthetics %	M&W mln m^2	M&W R-value	M&W R-value Corrected	M&W %	Average R-value
2010	5.4	2.4	34.8%	10.1	3.1	2.35	65.2%	2.4
2011	7.1	2.5	38.4%	11.4	3.2	2.46	61.6%	2.5
2012	5.1	2.7	32.7%	10.5	3.5	2.57	67.3%	2.6
2013	7.2	2.8	44.2%	9.1	2.9	2.69	55.8%	2.7
2014	9.2	2.7	47.7%	10.1	2.8	2.82	52.3%	2.8
2015	11.1	3.3	48.1%	12	2.8	2.95	51.9%	3.1
2016	15.3	3.3	52.4%	13.9	3	3.08	47.6%	3.2
2017	15.4	2.8	52.0%	14.2	3	3.23	48.0%	3
2018	11.4	2.9	44.0%	14.5	3.4	3.37	56.0%	3.2
2019	10.2	3.3	46.4%	11.8	3.8	3.53	53.6%	3.4

Table D.5: R-values for glass based on building year. Source: "Insulation per construction year", 2021, "U waarde (Glas) en R waarde (Isolatie)", 2021

Construction Year	Glass type	U-value	R-value
Before 1974	Single glass	5.7	0.175
1974-1992	Double glass	3	0.333

Table D.6: R-values for walls based on building year. Source: "Insulation per construction year", 2021

Construction year	Wall type	R-value
Before 1920	No cavity wall	0.19
1920-1974	Cavity wall with no insulation	0.35
1974-1992	Cavity wall with standard insulation	1.35

Table D.7: U and R values for various glass types. Source: "U waarde (Glas) en R waarde (Isolatie)", 2021

Type	U value (W m^{-2} K^{-1})	R value (m^2 K W^{-1}
Single Glass	5.7	0.175
Double Glass	3	0.333
HR Glass	1.6 - 2	0.5 - 0.625
HR + Glass	1.2 - 1.6	0.625 - 0.833
HR ++ Glass	<1.2	>0.833

Table D.8: Distribution of R-values for "Roof" for dwellings built before 2006 separated per dwelling type. Source: Menkveld et al., 2009

R Value	Free standing	2 under 1 / corner	Terraced house	Apartment	total
0.22	0.0%	0.0%	0.2%	0.3%	0.1%
0.39	2.0%	1.8%	2.3%	1.7%	2.0%
0.86	1.4%	5.1%	5.8%	10.3%	5.3%
0.97	24.2%	23.3%	22.2%	17.5%	22.4%
1.22	5.6%	4.1%	3.2%	2.6%	3.9%
1.3	14.8%	22.1%	27.2%	22.2%	22.6%
1.97	19.0%	16.8%	14.4%	10.3%	15.6%
2	8.6%	9.4%	7.0%	9.3%	8.4%
2.53	23.8%	16.4%	16.2%	24.8%	18.5%
2.72	0.8%	1.0%	1.5%	1.0%	1.1%

Table D.9: Roof *P*-distributions for dwellings constructed before 1992. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1	Terraced house	Apartment	Total
0.22	0.00%	0.00%	0.20%	0.40%	0.10%
0.39	2.50%	2.10%	2.80%	2.00%	2.40%
0.86	1.70%	6.00%	7.00%	12.50%	6.30%
0.97	30.20%	27.20%	26.60%	21.40%	27.00%
1.22	7.00%	4.80%	3.80%	3.20%	4.70%
1.3	18.50%	25.70%	32.60%	27.00%	27.20%
1.97	23.70%	19.60%	17.20%	12.50%	18.80%
2	10.70%	11.00%	8.40%	11.30%	10.10%
2.53	5.40%	3.40%	1.20%	9.40%	3.10%
2.72	0.20%	0.20%	0.10%	0.40%	0.20%

Table D.10: Roof *P*-distributions for dwellings constructed between 1992 and 2006. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1	Terraced house	Apartment	Total
2.53	96.70%	94.50%	91.50%	96.20%	94.30%
2.72	3.30%	5.50%	8.50%	3.80%	5.70%

Table D.11: Distribution of R values for "Floor" for dwellings built before 2006 separated per dwelling type. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1 / corner	Terraced house	Apartment	Total
0.15	0.5%	0.3%	0.5%	0.5%	0.4%
0.17	9.3%	17.9%	21.1%	14.5%	17.2%
0.32	8.6%	8.1%	7.7%	5.3%	7.8%
0.52	9.3%	11.4%	14.5%	15.0%	12.4%
0.65	13.3%	10.6%	8.7%	3.9%	9.8%
1.3	4.7%	11.3%	13.4%	16.9%	11.4%
1.4	4.2%	1.9%	1.2%	1.4%	2.0%
2	9.3%	10.7%	8.2%	6.8%	9.2%
2.15	12.8%	7.6%	5.2%	3.9%	7.3%
2.53	27.7%	19.1%	19.1%	31.4%	21.7%
2.65	0.5%	1.0%	0.5%	0.5%	0.7%

Table D.12: Floor *P*-distributions for dwellings constructed before 1992. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1	Terraced house	Apartment	Total
0.15	0.60%	0.40%	0.60%	0.60%	0.50%
0.17	11.60%	20.90%	25.30%	17.60%	20.70%
0.32	10.80%	9.50%	9.20%	6.50%	9.40%
0.52	11.60%	13.40%	17.40%	18.20%	15.00%
0.65	16.60%	12.30%	10.40%	4.70%	11.80%
1.3	5.80%	13.20%	16.10%	20.60%	13.70%
1.4	5.20%	2.20%	1.40%	1.80%	2.40%
2	11.60%	12.50%	9.90%	8.20%	11.10%
2.15	16.00%	8.90%	6.20%	4.70%	8.80%
2.53	9.90%	6.40%	3.50%	16.90%	6.40%
2.65	0.20%	0.30%	0.10%	0.30%	0.20%

Table D.13: Floor *P*-distributions for dwellings constructed between 1992 and 2006. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1	Terraced house	Apartment	Total
2.53	98.30%	95.10%	97.60%	98.50%	97.00%
2.65	1.70%	4.90%	2.40%	1.50%	3.00%

Table D.14: Distribution of R-values for "Façade" for dwellings built before 2006 separated per dwelling type. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1 /corner	Terraced house	Apartment	Total
0.36	0.2%	0.0%	0.0%	0.0%	0.0%
0.43	6.0%	9.3%	13.3%	10.9%	10.4%
1.3	2.5%	1.7%	3.2%	1.9%	2.3%
1.36	0.0%	0.0%	0.1%	0.0%	0.0%
2.11	40.8%	54.8%	52.1%	39.3%	48.0%
2.53	26.8%	18.4%	17.9%	39.2%	24.8%
2.86	23.6%	15.8%	13.4%	8.7%	14.4%

Table D.15: Façade P-distributions for dwellings constructed prior to 1920. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1	Terraced house	Apartment
0.36	1.80%	0.00%	0.00%	0.00%
0.43	46.90%	100.00%	100.00%	97.10%
1.3	19.80%	0.00%	0.00%	2.90%
1.36	0.00%	0.00%	0.00%	0.00%
2.11	31.50%	0.00%	0.00%	0.00%
2.53	0.00%	0.00%	0.00%	0.00%
2.86	0.00%	0.00%	0.00%	0.00%

Table D.16: Façade P-distributions for dwellings constructed between 1920 and 1992. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1	Terraced house	Apartment
0.36	0.00%	0.00%	0.00%	0.00%
0.43	0.00%	3.96%	9.22%	0.00%
1.3	0.00%	2.19%	4.18%	2.15%
1.36	0.00%	0.00%	0.15%	0.00%
2.11	54.81%	68.96%	67.36%	55.40%
2.53	24.03%	13.40%	10.92%	34.71%
2.86	21.16%	11.50%	8.17%	7.74%

Table D.17: Façade P-distributions for dwellings constructed between 1992 and 2006. Source: Menkveld et al., 2009

R Value	Free standing	2 Under 1	Terraced house	Apartment	Total
2.53	53.20%	53.80%	57.20%	81.80%	63.20%
2.86	46.80%	46.20%	42.80%	18.20%	36.80%

E | The covariance matrix for the multiple linear regression of energy label classes

Based on the multiple regression analysis as explained in Section 4.4.1, we derived the following covariance matrix S:

$$\begin{bmatrix} 6.838 \times 10^{-5} & -4.002 \times 10^{-8} & -4.967 \times 10^{-8} & -8.978 \times 10^{-8} & -1.047 \times 10^{-7} & -7.917 \times 10^{-8} & -3.421 \times 10^{-8} & -2.323 \times 10^{-6} \\ -4.002 \times 10^{-8} & 9.985 \times 10^{-8} & 6.236 \times 10^{-8} & 6.235 \times 10^{-8} & 6.269 \times 10^{-8} & 6.280 \times 10^{-8} & -1.053 \times 10^{-11} & -4.960 \times 10^{-9} \\ -4.967 \times 10^{-8} & 6.236 \times 10^{-8} & 9.869 \times 10^{-8} & 6.241 \times 10^{-8} & 6.231 \times 10^{-8} & 6.229 \times 10^{-8} & -6.572 \times 10^{-12} & 7.875 \times 10^{-10} \\ -8.978 \times 10^{-8} & 6.235 \times 10^{-8} & 6.241 \times 10^{-8} & 8.916 \times 10^{-8} & 6.234 \times 10^{-8} & 6.230 \times 10^{-8} & 1.352 \times 10^{-11} & 2.080 \times 10^{-9} \\ -1.047 \times 10^{-7} & 6.269 \times 10^{-8} & 6.231 \times 10^{-8} & 6.234 \times 10^{-8} & 1.780 \times 10^{-7} & 6.308 \times 10^{-8} & 2.230 \times 10^{-11} & -5.495 \times 10^{-9} \\ -7.917 \times 10^{-8} & 6.280 \times 10^{-8} & 6.229 \times 10^{-8} & 6.230 \times 10^{-8} & 6.308 \times 10^{-8} & 1.644 \times 10^{-7} & 9.880 \times 10^{-12} & -8.660 \times 10^{-9} \\ -3.421 \times 10^{-8} & -1.053 \times 10^{-11} & -6.572 \times 10^{-12} & 1.352 \times 10^{-11} & 2.230 \times 10^{-11} & 9.880 \times 10^{-12} & 1.713 \times 10^{-11} & 1.148 \times 10^{-9} \\ -2.323 \times 10^{-6} & -4.960 \times 10^{-9} & 7.875 \times 10^{-10} & 2.080 \times 10^{-9} & -5.495 \times 10^{-9} & -8.660 \times 10^{-9} & 1.148 \times 10^{-9} & 1.757 \times 10^{-7} \end{bmatrix}$$

F | More detailed analysis of the WoON survey data

This appendix functions as a repository for a collection of plots, figures, and tables that resulted from more in depth analysis of the WoON survey. A short summary of the most interesting results can be found in section 5.3

When we observe that two variables correlate, we see that they behave similarly, that is: when we see a change in variable A we also observe a change in variable B. This does not mean that these variables are actually linked! (correlation =/= causation), or that a change in one variable may explain a change in another variable. Just that they behave in a similar way. Conversely, we can state that when two variables do not correlate then one variable does not have any predictive value for the other.

It is also possible for variables to correlate a bit, when A changes B changes as well but not exactly as much as A. This is why correlation is expressed as a number between -1 and 1 where the hight of the number describes how strongly the variables correlate and the sign of the number (pos/neg) indicates the direction of the correlation.

How we measure correlation depends on the level of measurement that a variable has.

		Variable X		
		Nominal	**Ordinal**	**Continuous**
Variable Y	**Nominal**	ϕ or λ	Rank biserial	Point biserial
	Ordinal	Rank biserial	τ_b or Spearman	τ_b or Spearman
	Continuous	Point biserial	τ_b or Spearman	Pearson or Spearman

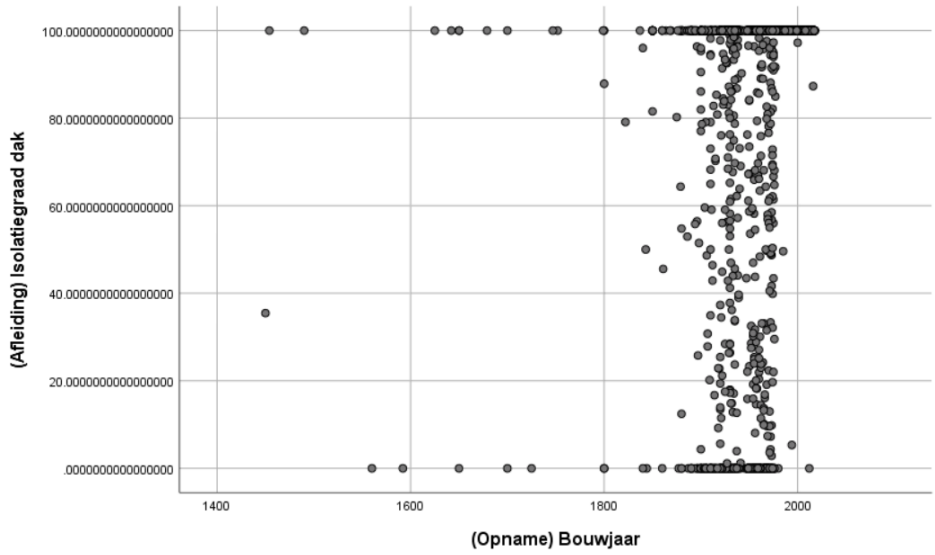
After: https://journals.sagepub.com/doi/pdf/10.1177/8756479308317006

There is no absolute definition for what counts as a strong correlation, there are only conventions and rules of thumb. Here we will follow the following scheme:

	Coefficient r	
	Positive	**Negative**
Strong	1 to 0.8	-1 to -0.8
Moderate	0.8 to 0.5	-0.8 to -0.5
Weak	0.5 to 0.3	-0.5 to -0.3
No correlation	0.3 to 0	-0.3 to 0

Datapoints 2345, 3387, and 3404 had implausible build years listed in the WoON (1005, 1005, and 19 respectively), these have been changed to more plausible values (1905 and 1980).

When trying to correlate build year and measure of insulation we can see the following:



Correlations

		(Opname) Bouwjaar	(Afleiding) Isolatiegraad dak
(Opname) Bouwjaar	Pearson Correlation	1	.322**
	Sig. (2-tailed)		.000
	N	4506	3610
(Afleiding) Isolatiegraad dak	Pearson Correlation	.322**	1
	Sig. (2-tailed)	.000	
	N	3610	3610

**. Correlation is significant at the 0.01 level (2-tailed).

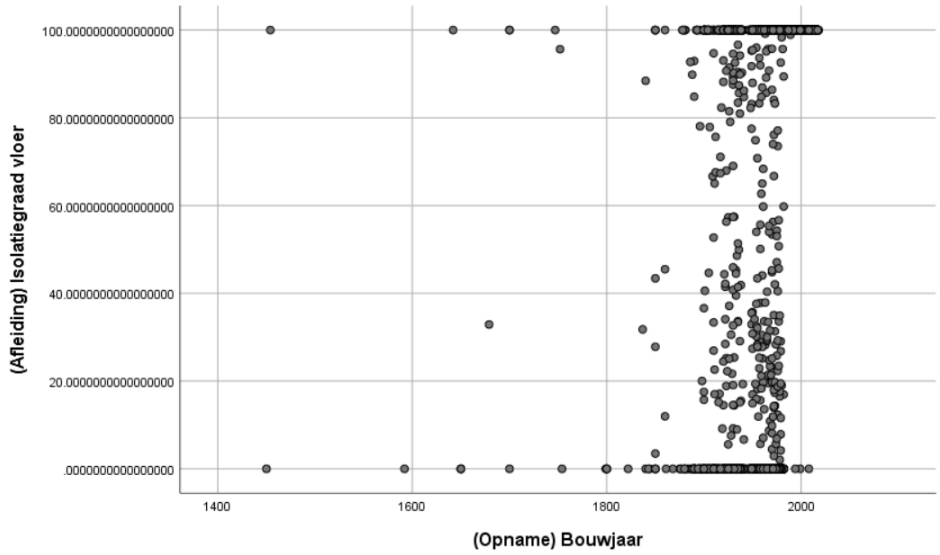
Correlations

			(Opname) Bouwjaar	(Afleiding) Isolatiegraad dak
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	.451**
		Sig. (2-tailed)	.	.000
		N	4506	3610
	(Afleiding) Isolatiegraad dak	Correlation Coefficient	.451**	1.000
		Sig. (2-tailed)	.000	.
		N	3610	3610

**. Correlation is significant at the 0.01 level (2-tailed).

As we can see there is hardly any correlation between the build year of a dwelling and the insulation of its roof. Linearly the correlation barely scratched the bottom of 'weak' and monotonically the correlation is weak at best.

When trying to correlate build year and floor insulation we can see the following:



Correlations

		(Opname) Bouwjaar	(Afleiding) Isolatiegraad vloer
(Opname) Bouwjaar	Pearson Correlation	1	.420**
	Sig. (2-tailed)		.000
	N	4506	3510
(Afleiding) Isolatiegraad vloer	Pearson Correlation	.420**	1
	Sig. (2-tailed)	.000	
	N	3510	3510

**. Correlation is significant at the 0.01 level (2-tailed).

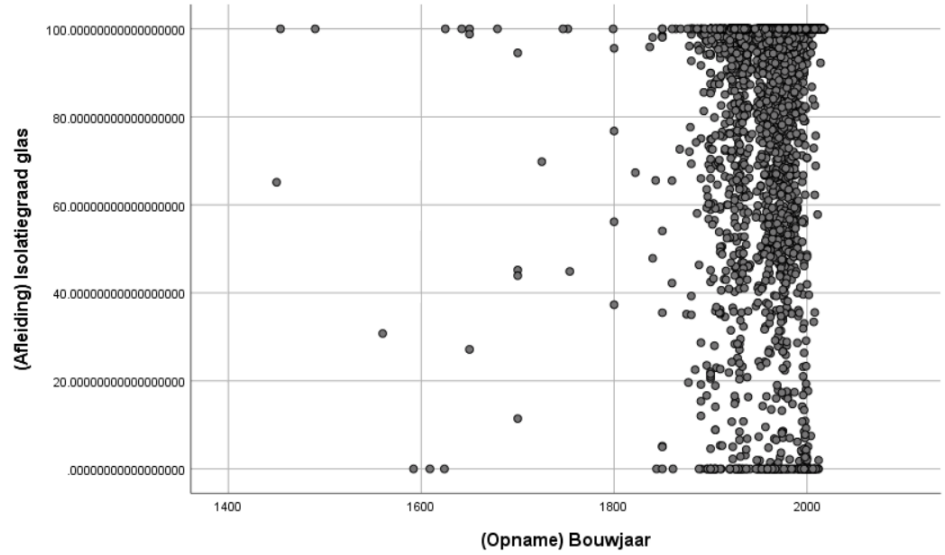
Correlations

			(Opname) Bouwjaar	(Afleiding) Isolatiegraad vloer
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	.591**
		Sig. (2-tailed)	.	.000
		N	4506	3510
	(Afleiding) Isolatiegraad vloer	Correlation Coefficient	.591**	1.000
		Sig. (2-tailed)	.000	.
		N	3510	3510

**. Correlation is significant at the 0.01 level (2-tailed).

Linearly we can see a weak correlation, and monotonically we can see a moderate correlation.

When trying to correlate build year and glass insulation these results can be observed:



Correlations

		(Opname) Bouwjaar	(Afleiding) Isolatiegraad glas
(Opname) Bouwjaar	Pearson Correlation	1	.192**
	Sig. (2-tailed)		.000
	N	4506	4503
(Afleiding) Isolatiegraad glas	Pearson Correlation	.192**	1
	Sig. (2-tailed)	.000	
	N	4503	4503

**. Correlation is significant at the 0.01 level (2-tailed).

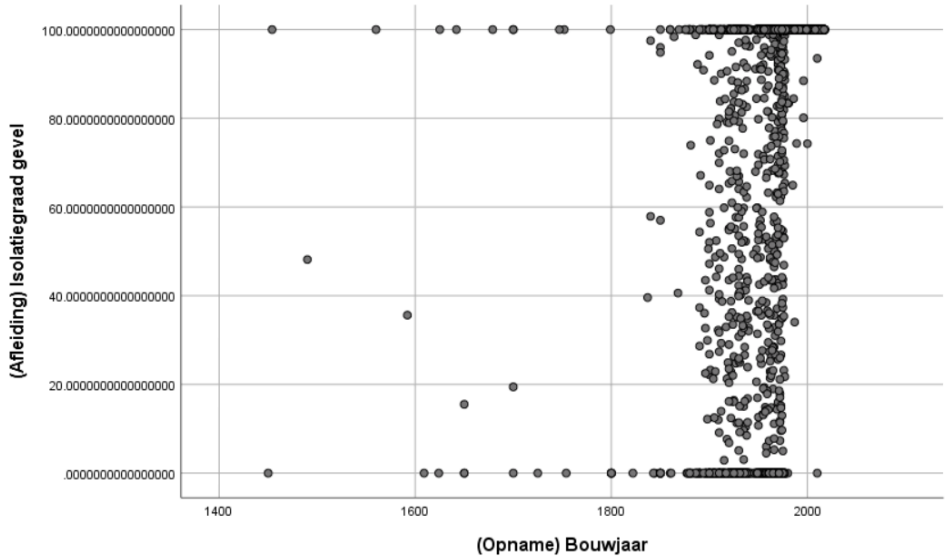
Correlations

			(Opname) Bouwjaar	(Afleiding) Isolatiegraad glas
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	.319**
		Sig. (2-tailed)	.	.000
		N	4506	4503
	(Afleiding) Isolatiegraad glas	Correlation Coefficient	.319**	1.000
		Sig. (2-tailed)	.000	.
		N	4503	4503

**. Correlation is significant at the 0.01 level (2-tailed).

Monotonically we can barely see a weak correlation.

Trying to correlate build year and outer wall insulation results in the following results:



Correlations

		(Opname) Bouwjaar	(Afleiding) Isolatiegraad gevel
(Opname) Bouwjaar	Pearson Correlation	1	.442**
	Sig. (2-tailed)		.000
	N	4506	4502
(Afleiding) Isolatiegraad gevel	Pearson Correlation	.442**	1
	Sig. (2-tailed)	.000	
	N	4502	4502

**. Correlation is significant at the 0.01 level (2-tailed).

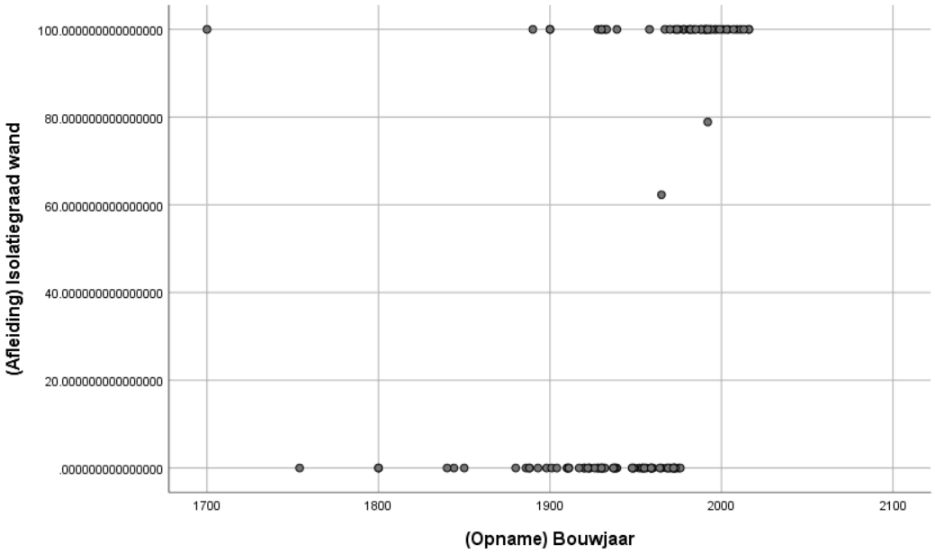
Correlations

			(Opname) Bouwjaar	(Afleiding) Isolatiegraad gevel
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	.611**
		Sig. (2-tailed)	.	.000
		N	4506	4502
	(Afleiding) Isolatiegraad gevel	Correlation Coefficient	.611**	1.000
		Sig. (2-tailed)	.000	.
		N	4502	4502

**. Correlation is significant at the 0.01 level (2-tailed).

Linearly we can observe a weak correlation but monotonically we can observe a moderate correlation.

When we try to correlate build year and inter dwelling insulation the results are as follows:



Correlations

		(Opname) Bouwjaar	(Afleiding) Isolatiegraad wand
(Opname) Bouwjaar	Pearson Correlation	1	.419**
	Sig. (2-tailed)		.000
	N	4506	107
(Afleiding) Isolatiegraad wand	Pearson Correlation	.419**	1
	Sig. (2-tailed)	.000	
	N	107	107

**. Correlation is significant at the 0.01 level (2-tailed).

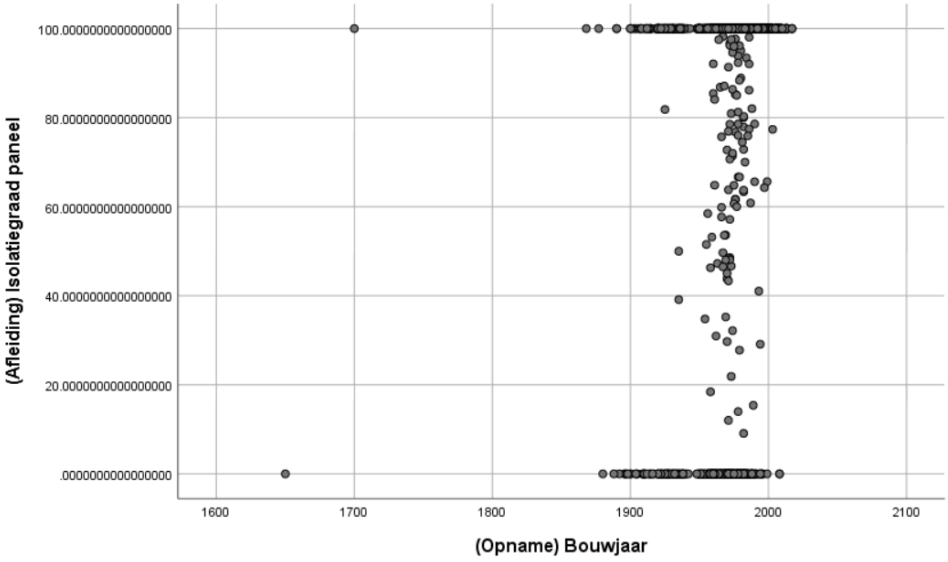
Correlations

			(Opname) Bouwjaar	(Afleiding) Isolatiegraad wand
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	.581**
		Sig. (2-tailed)	.	.000
		N	4506	107
	(Afleiding) Isolatiegraad wand	Correlation Coefficient	.581**	1.000
		Sig. (2-tailed)	.000	.
		N	107	107

**. Correlation is significant at the 0.01 level (2-tailed).

N=107 is too low to say anything in our opinion

Trying to correlate build year with inter room insulation:



Correlations

		(Opname) Bouwjaar	(Afleiding) Isolatiegraad paneel
(Opname) Bouwjaar	Pearson Correlation	1	.254**
	Sig. (2-tailed)		.000
	N	4506	1631
(Afleiding) Isolatiegraad paneel	Pearson Correlation	.254**	1
	Sig. (2-tailed)	.000	
	N	1631	1631

**. Correlation is significant at the 0.01 level (2-tailed).

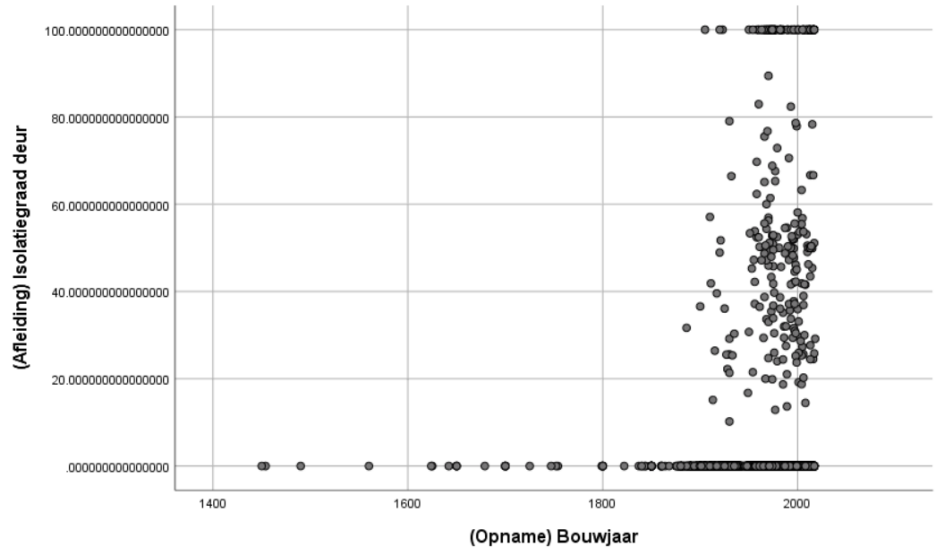
Correlations

			(Opname) Bouwjaar	(Afleiding) Isolatiegraad paneel
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	.272**
		Sig. (2-tailed)	.	.000
		N	4506	1631
	(Afleiding) Isolatiegraad paneel	Correlation Coefficient	.272**	1.000
		Sig. (2-tailed)	.000	.
		N	1631	1631

**. Correlation is significant at the 0.01 level (2-tailed).

We find no correlation whatsoever.

When looking for the correlation of build year and the insulation of doors:



		(Opname) Bouwjaar	(Afleiding) Isolatiegraad deur
(Opname) Bouwjaar	Pearson Correlation	1	.089**
	Sig. (2-tailed)		.000
	N	4506	4406
(Afleiding) Isolatiegraad deur	Pearson Correlation	.089**	1
	Sig. (2-tailed)	.000	
	N	4406	4406

**. Correlation is significant at the 0.01 level (2-tailed).

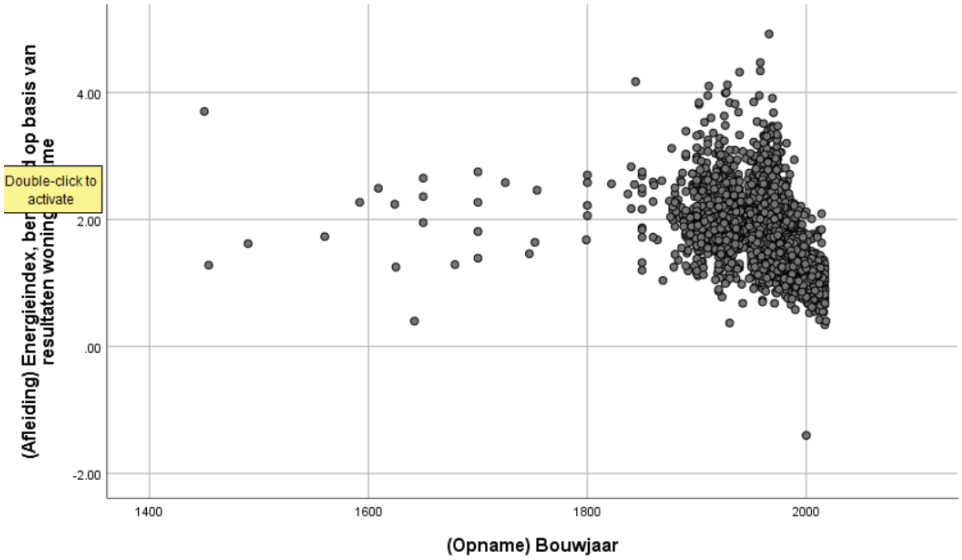
Correlations

			(Opname) Bouwjaar	(Afleiding) Isolatiegraad deur
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	.102**
		Sig. (2-tailed)	.	.000
		N	4506	4406
	(Afleiding) Isolatiegraad deur	Correlation Coefficient	.102**	1.000
		Sig. (2-tailed)	.000	.
		N	4406	4406

**. Correlation is significant at the 0.01 level (2-tailed).

We find no correlation whatsoever.

When trying to correlate build year to the energy index we find the following:



Correlations

		(Opname) Bouwjaar	(Afleiding) Energieindex, berekend op basis van resultaten woningopname
(Opname) Bouwjaar	Pearson Correlation	1	-.477**
	Sig. (2-tailed)		.000
	N	4506	4506
(Afleiding) Energieindex, berekend op basis van resultaten woningopname	Pearson Correlation	-.477**	1
	Sig. (2-tailed)	.000	
	N	4506	4506

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			(Opname) Bouwjaar	(Afleiding) Energieindex, berekend op basis van resultaten woningopname
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	-.719**
		Sig. (2-tailed)	.	.000
		N	4506	4506
	(Afleiding) Energieindex, berekend op basis van resultaten woningopname	Correlation Coefficient	-.719**	1.000
		Sig. (2-tailed)	.000	.
		N	4506	4506

**. Correlation is significant at the 0.01 level (2-tailed).

Linearly we can see a weak correlation whereas monotonically we can see a moderate correlation.

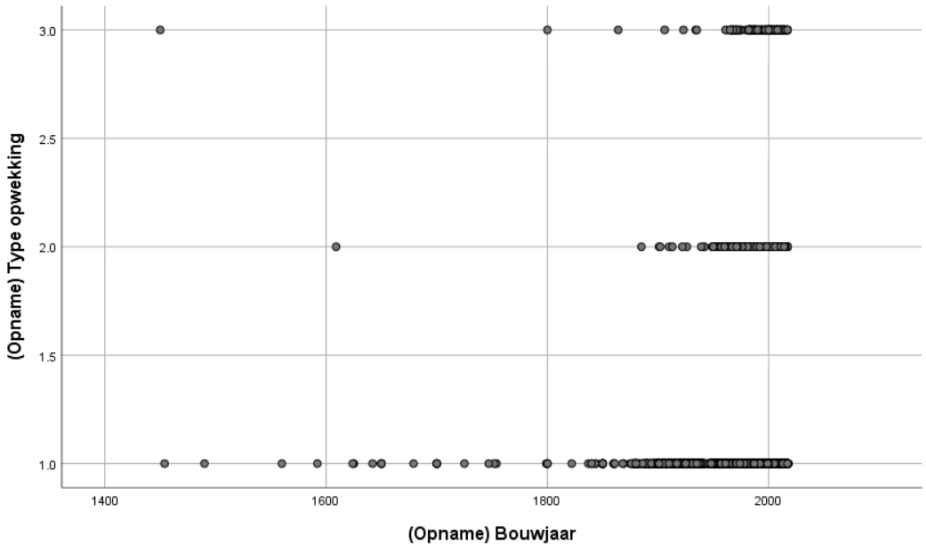
Below is a plot of build year and Source of heat:

Legend:

1 = individual generation

2 = collective generation

3 = heat from a third party

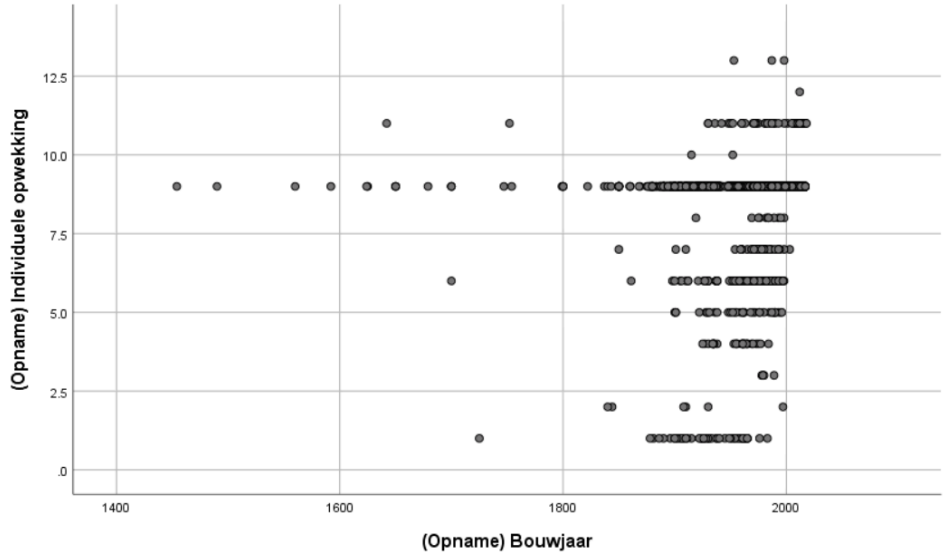


We can observe a trend where collective generation and heat from a third party occur less commonly in older dwellings than individual generation, but all three appear to occur in newer buildings.

A plot of build year and type of individual generation unit can be seen below:

Legend:

1 = "lokale gas of oliekachel"
2 = "lokale houtkachel"
3 = "lokaal elektrisch"
4 = "CR ketel of moederhaard (waakvlam)"
5 = "VR ketel (waakvlam)"
6 = "VR ketel (elektronische ontsteking)"
7 = "HR100 ketel"
8 = "HR104 ketel"
9 = "HR107 ketel"
10 = "micro-wkk"
11 = "warmtepomp - elektrisch"
12 = "warmtepomp - gas"
13 = "overig/onbekend"



We can observe that lower efficiency boilers are virtually unused in newer dwellings whereas HR-107 boilers and electric heat pumps are used in the newest dwellings.

Below is a correlation table of building year and the presence of solar panels:

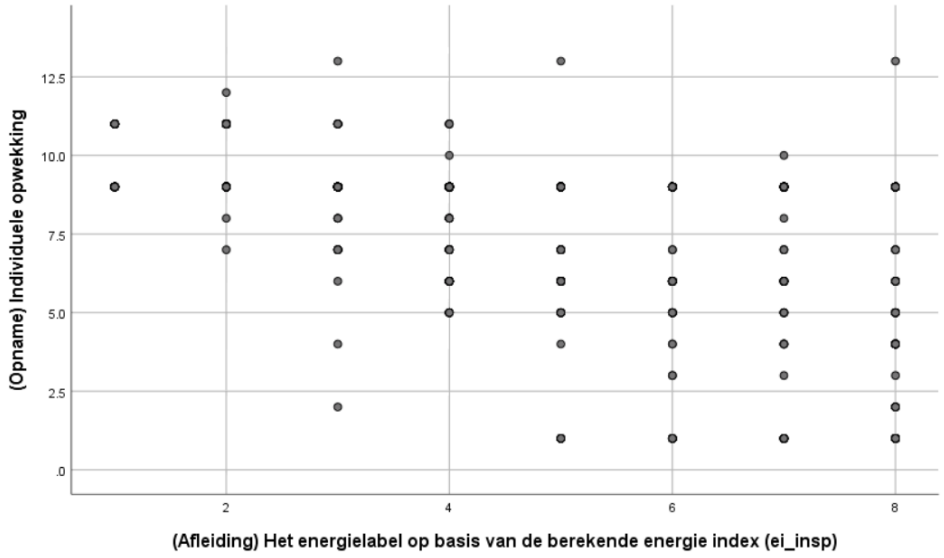
Correlations

			(Opname) Bouwjaar	(Opname) Zonnepanelen (PV-systemen) aanwezig
Kendall's tau_b	(Opname) Bouwjaar	Correlation Coefficient	1.000	.068**
		Sig. (2-tailed)	.	.000
		N	4506	4506
	(Opname) Zonnepanelen (PV-systemen) aanwezig	Correlation Coefficient	.068**	1.000
		Sig. (2-tailed)	.000	.
		N	4506	4506
Spearman's rho	(Opname) Bouwjaar	Correlation Coefficient	1.000	.084**
		Sig. (2-tailed)	.	.000
		N	4506	4506
	(Opname) Zonnepanelen (PV-systemen) aanwezig	Correlation Coefficient	.084**	1.000
		Sig. (2-tailed)	.000	.
		N	4506	4506

**. Correlation is significant at the 0.01 level (2-tailed).

We can observe that these do not correlate at all, which is somewhat to be expected as the factors that determine a rooftop's suitability for installing solar panels have hardly changed over the years.

When we try to correlate energylabel and boiler type we find:



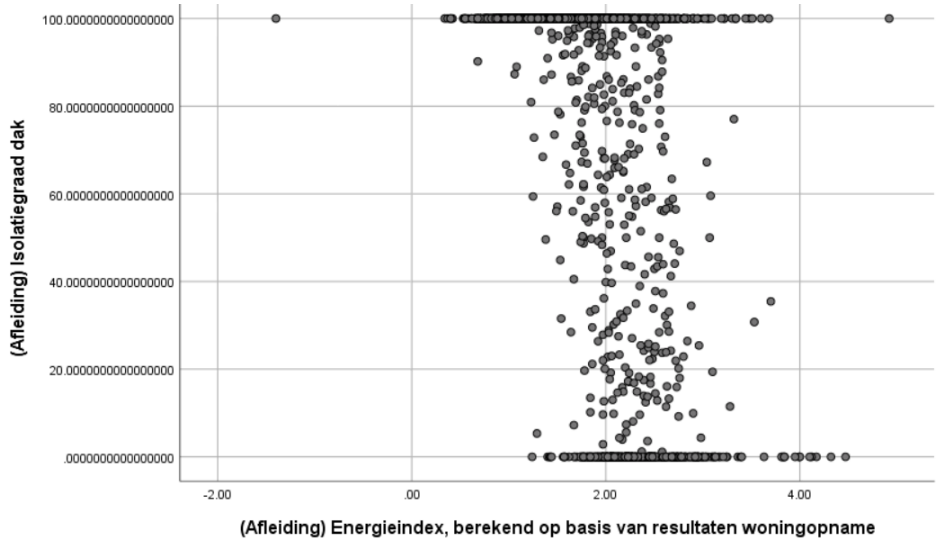
Correlations

			(Opname) Individuele opwekking	(Afleiding) Het energielabel op basis van de berekende energie index (ei_insp)
Kendall's tau_b	(Opname) Individuele opwekking	Correlation Coefficient	1.000	-.284**
		Sig. (2-tailed)	.	.000
		N	4035	4035
	(Afleiding) Het energielabel op basis van de berekende energie index (ei_insp)	Correlation Coefficient	-.284**	1.000
		Sig. (2-tailed)	.000	.
		N	4035	4506
Spearman's rho	(Opname) Individuele opwekking	Correlation Coefficient	1.000	-.321**
		Sig. (2-tailed)	.	.000
		N	4035	4035
	(Afleiding) Het energielabel op basis van de berekende energie index (ei_insp)	Correlation Coefficient	-.321**	1.000
		Sig. (2-tailed)	.000	.
		N	4035	4506

**. Correlation is significant at the 0.01 level (2-tailed).

No correlation to speak of.

Below is a correlation between energy index and insulation of the roof:



Correlations

		(Afleiding) Energieindex, berekend op basis van resultaten woningopname	(Afleiding) Isolatiegraad dak
(Afleiding) Energieindex, berekend op basis van resultaten woningopname	Pearson Correlation	1	-.541**
	Sig. (2-tailed)		.000
	N	4506	3610
(Afleiding) Isolatiegraad dak	Pearson Correlation	-.541**	1
	Sig. (2-tailed)	.000	
	N	3610	3610

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			(Afleiding) Energieindex, berekend op basis van resultaten woningopname	(Afleiding) Isolatiegraad dak
Spearman's rho	(Afleiding) Energieindex, berekend op basis van resultaten woningopname	Correlation Coefficient	1.000	-.521**
		Sig. (2-tailed)	.	.000
		N	4506	3610
	(Afleiding) Isolatiegraad dak	Correlation Coefficient	-.521**	1.000
		Sig. (2-tailed)	.000	.
		N	3610	3610

**. Correlation is significant at the 0.01 level (2-tailed).

Correlation of –5.41 is moderate, however it is also to be expected that these correlate, given that insulation is a large component in determining the energy index.

The Dutch ‘bouwbesluit’ (building decree) of 1992 was the first nation wide decree demanding certain technical standards be observed during construction of a building. New versions have gone into effect in 2003 and 2012.

Splitting the variable bouwjaar into four categories that correspond to itterations of the bouwbesluit may yield interesting results regarding insulation given that the bouwbesluit also contains demands on how well buildings should be insulated.

Insulation values in the WoON survey are given as either a continuous variable or as an ordinal variable with four categories. Since bouwjaar now also is ordinal we will calculate Kendal's tau and spearman's rho for all cases.

Kendal's tau and Spearman's rho for the derived insulation variables (continuous).

Correlations

			Bouwjaar in categorien	(Afleiding) Isolatiegraad dak	(Afleiding) Isolatiegraad vloer	(Afleiding) Isolatiegraad glas	(Afleiding) Isolatiegraad gevel	(Afleiding) Isolatiegraad wand	(Afleiding) Isolatiegraad paneel	(Afleiding) Isolatiegraad deur	(Afleiding) Energieindex, berekend op basis van resultaten woningopname
Kendall's tau_b	Bouwjaar in categorien	Correlation Coefficient	1.000	.233**	.385**	.230**	.342**	.461**	.235**	.112**	-.513**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000	.000	.000
		N	4506	3610	3510	4503	4502	107	1631	4406	4506
Spearman's rho	Bouwjaar in categorien	Correlation Coefficient	1.000	.249**	.410**	.271**	.371**	.473**	.246**	.117**	-.627**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000	.000	.000
		N	4506	3610	3510	4503	4502	107	1631	4406	4506

As we can see all of the insulation variables show some weak correlation with the build year categories, and energy index correlates moderately with the build year categories.

Kendal's tau and Spearman's rho for the derived insulation variables (in categories).

Correlations

			Bouwjaar in categorien	(Afleiding) Isolatiegraad dak in vier klassen	(Afleiding) Isolatiegraad vloer in vier klassen	(Afleiding) Isolatiegraad glas in vier klassen	(Afleiding) Isolatiegraad gevel in vier klassen	(Afleiding) Isolatiegraad wand in vier klassen	(Afleiding) Isolatiegraad paneel in vier klassen	(Afleiding) Isolatiegraad deur in vier klassen	(Afleiding) Het energielabel op basis van de berekende energie index (el_insp)
Kendall's tau_b	Bouwjaar in categorien	Correlation Coefficient	1.000	.224**	.382**	.155**	.335**	.460**	.230**	.112**	-.542**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000	.000	.000
		N	4506	3610	3510	4503	4502	107	1631	4406	4506
Spearman's rho	Bouwjaar in categorien	Correlation Coefficient	1.000	.237**	.404**	.168**	.358**	.473**	.239**	.117**	-.617**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000	.000	.000
		N	4506	3610	3510	4503	4502	107	1631	4406	4506

Differences between the correlation with the continuous variables are negligible.