Geo-Information Science and Remote Sensing

Thesis Report GIRS-2024-36

The Human Element: Integrating socio-economic variables in a physics-based energy demand model

Krijn Horstmanshoff

08-04-2024

WAGENINGEN
UNIVERSITY & RESEARCH

The Human Element: Integrating socio-economic variables in a physics-based energy demand model

Author: Krijn Horstmanshoff

Registration number: 1170376

Supervisor: Dr. Jessica Wreyford

A thesis submitted in partial fulfilment of the degree of Master of Science

at Wageningen University and Research,

The Netherlands.

08-04-2024

Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2024-36
Wageningen University and Research
Laboratory of Geo-Information Science and Remote Sensing

Preface

Dear reader,

Before you lies my Master thesis for the master Geo-Information Science at Wageningen University & Research. Over the years, I developed a passion for combining Social Science and Geo-Information Science, providing a human-interaction perspective on GIS. In this thesis on energy demand, I tried to combine these two interests into one end-product. I gained a lot of knowledge in the field of thermodynamics of buildings and how to apply these concepts in a long-term data science project. This also helped me in developing my Python competences.

I would like to thank my supervisor Dr. Jessica Wreyford for the insightful guidance and the flexibility she offered during the research process. She offered me the freedom to find out for myself what I find interesting and what I would like to research. I would also like to thank my friends in the Thesis Room for their feedback and the joyful coffee breaks in between the thesis writing.

Abstract

Housing is one of the largest contributors to CO_2 emissions. Currently, the average household uses 21% of their CO_2 emissions for their heating needs. For policymaking, it is therefore crucial to develop strategies to reduce these emissions. In order to develop such strategies, access to accurate information on where energy is used is vital. Energy demand models can help provide this information. However, in existing research a strong dichotomy exists between physics-based models, that primarily focus on the thermodynamics of buildings, and socio-economic models, that primarily focus on the people that live in the buildings. This division results in a degree of unexplained variation in model predictions. This thesis aims to incorporate socio-economic principles into a physics-based energy demand model. Results do not show a significant improvement in model performance. This is possibly due to the fact that socio-economic data is only available at a relatively high spatial level, as it would pose ethical concerns otherwise.

Contents

1 Introduction

1.1 Research background

From the end of the 19th century onwards, there have been periods of rapid population growth in the Netherlands (Komlos, 1990). This resulted in a large demand for housing, which caused the Dutch housing stock to increase by 269% since the 1950's (Centraal Bureau voor de Statistiek (CBS), 2023c; Van de Woestijne, 1933). The large quantity of houses built during this time period vary in their quality of construction and insulation (Cammen & Klerk, 2012). The current housing stock, therefore, consists of a wide range of building types from different time periods that vary in insulation techniques and heating efficiency (Nieboer & Filippidou, 2017).

Housing is one of the largest contributors to CO_2 emissions (Dahlström et al., 2022). Currently, the average household uses 21% of their CO_2 emissions for their heating and housing needs (Dubois et al., 2019). This high emittance is posing a notable obstacle to achieving the goal of the Dutch government to reduce CO_2 emissions by 95% in 2050. To find a solution to this emittance problem, policymakers are not only focusing on using more renewable energy, but also focus on reducing the overall amount of energy that households consume. Therefore, policymakers from various layers of government are developing energy saving programs and retrofitting programs to make houses and their inhabitants more energy efficient (Gemeente Amsterdam, n.d.; Ministerie van Economische Zaken en Klimaat, 2023). A national aim is set to improve the insulation of the 1.5 million houses with the lowest energy efficiency (Mininsterie van Binnenlandse Zaken en Koninkrijksrelaties, n.d.). To effectively support these programs, it is crucial to have accurate information on the energy demand of buildings.

Assessing the energy demand of houses remains a challenge, as there is little public energy demand data available on a building-scale. Energy demand models have the potential to provide more insights into how and where energy is consumed. This knowledge can help policymakers to access where energy-saving potential is the greatest (Yang et al., 2020). It can also help policymakers into developing effective retrofitting policies for the housing stock (Afaifia et al., 2021; Ali et al., 2020). Additionally, predictions into future energy demand can enable policymakers in developing smart strategies that lead to a high energy efficiency and resource optimization of energy throughout an urban area (Condotta & Borga, 2018). In summary, addressing the need to reduce energy consumption in buildings requires accurate methods for estimating a building's energy usage. Energy demand models can contribute to enhancing these estimation methods.

1.2 Energy demand modelling

In general, there are four types of energy models in literature, physics-based, social, economic and socio-economic. Physics-based models aim to explain energy consumption on the basis of the characteristics of houses (Ali et al., 2020; Duminil et al., 2018; León-Sánchez et al., 2021; Silva et al., 2017; Torabi Moghadam et al., 2018). Social models focus more on the behavioral and psychological aspects of energy consumption (M. Costanzo et al., 1986; Dubois et al., 2019; Frederiks et al., 2015). Economic models focus on the effect of appliance efficiency and energy prices to explain energy consumption (Boonekamp, 2007; Haas & Schipper, 1998; Hunt & Ryan, 2015). Socio-economic models aim to link economic incentives to behavior of households (Brounen et al., 2012; Schulte & Heindl, 2017; Yang et al., 2020). Since socio-economic encompass the interaction between social and economic factors, this research will focus on expanding upon the physics-based (Section 1.2.1) and socio-economic (Section 1.2.2) models.

1.2.1 Physics-based models

Physics-based models aim to predict energy consumption through physical factors of each individual building. This way of modelling often consists of a bottom-up approach. A bottom-up model aims to predict energy demand based on the characteristics of individual buildings and then aggregating the data for all buildings to derive regional statistics (Afaifia et al., 2021; Ghedamsi et al., 2016; Yang et al., 2020). Current research most often includes the following factors as relevant to energy consumption predictions: year of construction, net floor area, window-to-wall ratio, and weather data (J. A. Fonseca & Schlueter, 2015). For deriving the values for these factors, physics-based modesl use 2D and/or 3D geographic data to determine the energy demand of buildings (Ghedamsi et al., 2016).

Two well-known applications that use 3D data are Simstadt (Duminil et al., 2018) and CitySim Pro (Kämpf & Oguey, 2023). These applications use characteristics of the building to calculate the energy demand. These models also take external factors like solar irradiance and outside temperature into account (León-Sánchez et al., 2021). However, apart from the number of addresses, no other socio-economic parameters are included in these models. This is a problem, because energy usage is not only dependent on how a house looks but is also defined by how people use their energy.

1.2.2 Socio-economic models

Socio-economic models have a different starting point to modelling energy demand. These models often start with historical energy consumption data and apply a regression model to find out relationships, for example between types of housing and their energy usage (Mata et al., 2021; Schulte & Heindl, 2017).

Hunt & Ryan (2015) developed a model where the starting point was not the aggregated energy demand of a household, but the individual services that need energy. Services encompass aspects such as heating, lighting, and access to hot water. In this model, they also include price elasticity of these services. This means that the model takes into consideration how much money people are willing to pay for a service, which in turn has implications for the total energy demand of the house.

Schulte and Heindl (2017) performed a socio-economic analysis on price- and income elasticity for residential energy demand in Germany. The model is based on German data from the “Mikrozensus”, in which about one percent of German citizens were surveyed on a plethora of subjects, including income and energy consumption. Their results show that the reaction to changes in energy prices is highly dependent on total household income. Their model shows that households with a lower income tend to decrease their energy consumption less than wealthier households.

As shown above, socio-economic models provide insights into how households behave in their energy demand. However, they fail to integrate information on building type and external factors like the weather. This increases the residual error of the model.

1.3 Problem statement

Both physics-based and socio-economic models provide valuable insights into predicting energy consumption in buildings. However, this division of models is limiting the accuracy of both. In

reality, a building's energy consumption is influenced by an interplay of factors from both physics-based and socio-economic models (Ali et al., 2020).

Attempts have been made to compare or combine socio-economic and physics-based models (Afaifia et al., 2021; Condotta & Borga, 2018; Gassar et al., 2019; Van Den Brom et al., 2019). Van den Brom et al. (2019) looked at how much of the variance in an energy consumption model could be explained by occupants characteristics. Afaifia et al. (2021) adopted a triangular approach, in which they employed 2D GIS methods to build a database containing information about building stock and energy usage. They also utilized multiple linear regression to identify critical variables impacting energy consumption and applied hierarchical clustering to better understand regional differences. Gasser et al. (2019) developed several machine learning models that used both socio-economic factors and physics-based factors. However, both models were developed on a regional scale and not on a building scale.

Fonseca & Schlueter (2015) developed a physics-based model, in which they integrated energy services. For these services, standardized consumption values were calculated based on housing type. However, other socio-economic factors that relate to the residents themselves, like income and price elasticity, are left out of this model.

In conclusion, the current integrative energy demand models contain limitations, particularly in capturing the relationship between physics-based models and socio-economic models on a building scale. Especially in bottom-up models, the socio-economic factors are frequently overlooked or incorporated at a minimal level. This resonates with Ali et al.'s (2020) discussion of their bottom-up physics-based model, in which they concluded that there is a need for social science-based research to investigate variations in energy consumption among technologically similar buildings.

1.4 Research objective and questions

The objective of this research is to investigate how socio-economic information about households, neighborhood characteristics, and price elasticity contribute to more accurate energy demand estimations on a building-scale. This leads to the following main research question:

"What is the effect of integrating socio-economic factors into physics-based energy demand models for the purpose of predicting energy consumption in residential buildings?

This research question is further subdivided into four Sub-research questions (SRQ):

SRQ1. What are the primary components of existing physics-based models and how do these models compare?
SRQ2. What major factors do socio-economic energy demand models identify as key influencers of energy consumption?
SRQ3. Which socio-economic factors can be integrated into a select physics-based energy demand model?
SRQ4. What is the accuracy of the hybrid model when comparing it to the select physics-based energy demand model?

2 Data and Methods

The following sections describe the steps that are needed to answer the research questions presented in Section 1.4. SRQ1 will be addressed through a literature review on physics-based demand modelling (Section 2.1). SRQ2 will be answered through a literature review on socio-economic models (Section 2.2). SRQ3 will be addressed by constructing a physics-based model, in which factors from socio-economic models are incorporated (Section 2.3). SRQ4 will be answered by comparing the relative accuracy of the hybrid model with the relative accuracy of the original physics-based model (Section 2.4).

2.1 Literature review of physics-based models

In this thesis a systematic review will be used to analyze physics-based energy demand models. Often, these types of models are largely of the same design. Therefore, it is feasible to conduct a systematic review (Snyder, 2019). Based on the research question, inclusion- and exclusion criteria are defined (University Libraries, University of Maryland, 2023). These inclusion- and exclusion criteria are shown in Table 2.1.

Table 2.1 The inclusion and exclusion criteria that are used for searching physics-based models.

	Inclusion criteria	**Exclusion criteria**
Approach	Bottom-up	Top down
Scale	Residential	Regional
Type	Building-stock model	Retrofitting modelling
Scope	Heating demand	Environmental policy

The criteria in Table 2.1 identify those models that aim to predict energy usage of specific residential buildings through a bottom-up energy demand model. Other types of models that follow a top-down method or models that are performed on a regional scale should be avoided, as they fall outside the scope of this research. Papers that focus more on the usability of the models for policymaking and retrofitting policies are also less relevant for this thesis and should therefore also be left out. After determining the inclusion- and exclusion terms, the systematic analysis can commence.

Firstly, a dataset consisting of papers about physics-based energy demand models is assembled. Two search engines are used to find these papers: Web of Science and Scopus. By using the query “Energy demand model AND bottom-up”, already most papers that do not base their energy demand on single buildings are left out of the research. The remainder of papers is then stored in a .RIS file, which can be used as input for a search algorithm.

For analyzing the wide range of papers, the open-source algorithm “ASReview LAB” is used (De Bruin et al., 2023). This algorithm uses Term Frequency – Inversed Document Frequency (TF-IDF) and Naïve Bayes Machine Learning Algorithms to identify which papers are more relevant than others. When reviewing abstracts and determining the relevance of papers based on inclusion and exclusion criteria, ASReview Lab rearranges the papers according to their relevance (Utrecht University, 2022). Because irrelevant papers are moved down the stack, time is saved when reading the papers. At a certain point, only irrelevant papers will show up through the algorithm. This means that most relevant papers from the search results have been identified.

From the relevant papers, a table is made that shows which factors are used in the physics-based models to predict energy demand. Extra information about the performance of the model will also be shown in this overview. Using this table, it is possible to select suitable models as input for the physics-based model that will be used in the continuation of this thesis.

2.2 Literature of socio-economic models

For answering the second sub-research question on socio-economic models a semi-systematic literature review is conducted. The semi-systematic approach is meant for analyzing topics that have been conceptualized differently and that are studied by various research groups from a wide range of disciplines (Snyder, 2019). The term “socio-economic” already implies that these are models that consists of both social and economic factors that are interrelated. Therefore, there are various ways in which the topic of energy demand is approached.

Because of this variety, a semi-systematic review seeks to identify and understand all potentially relevant concepts. However, it is still systematic in the sense that the literature review is initialized with some core concepts. However, the process is not as linear as a systematic review. When new insights are found to be valuable for answering the research question, those concepts can change. Initial concepts that will be used as searched terms are: socio-economic, space heating, efficiency, elasticity, household, and income.

The tool ASReview Lab (see Section 2.1) can still be applied to this semi-systematic review, as its only role is to order the papers in projected relevance. To scale down the number of papers that needs to be read, it is important to really focus the review on papers with an emphasis on space heating of buildings and not on other services.

2.3 Constructing the physics-based model

To get an overview of how the model is constructed, first the input data is described and its main goal within the model is explained. This is followed by a more detailed description of how the model is designed and what assumptions have been made.

2.3.1 Input data for physics-based model

The selection of input data is based upon a multitude of papers that are deemed most relevant to this research case in the literature reviews. Sokol et al. (2017) use 3D data to get the morphological attributes of the building. This paper, Todeschi et al. (2021), and Yang et al. (2020) all use a building typology to estimate additional information on each building. All these papers also use weather data to further calibrate the model. For all these different data sources, Dutch alternatives are found and downloaded. In Table 2.2, all input datasets and their sources are briefly described.

Table 2.2 Main input data sources for the physics-based model

Data name	Data category	Description	Data source
BAG	2D cadastral data	Base registration of buildings in the Netherlands. Also contains information on building function and number of addresses.	Kadaster, 2023
3DBAG	3D buildings	3D version of the BAG, containing information on height, width and volume of every building.	Peters et al., 2022
Public data energy performance contracts (Openbare data energielabels)	Energy performance contracts (EPC)	Not all houses in the Netherlands have such a contract. This is because it is only mandatory to get an EPC when a building is sold or rented. (Rijksoverheid, n.d.)	Rijksdienst voor Ondernemend Nederland (RVO), 2022
TABULA/episcope	Building typology	EU initiative, where for each member state a building typology is created. Buildings are segmented into four building types and six construction periods. Based on the type, U-values are attributed for all building elements	Nieboer & Filippidou, 2017
Sunshine duration and radiation (Zonneschijnduur en straling)	Solar radiation data	Weather dataset that contains sunshine duration and radiation data for every 10 minutes of the day.	Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2024b
Humidity and temperature (Vochtigheid en temperatuur)	Temperature data	Weather dataset that contains humidity and temperature data fir every 10 minutes of the day.	KNMI, 2024
Key figures per postal code (Kerncijfers per postcode)	Historical gas usage	Dataset that provides information on the mean gas usage per postal code.	Centraal Bureau voor de Statistiek (CBS), 2023a
S2 Geometries of windows and doors	Window-to-Wall Ratios (WWR)	Table based on field survey, where building element proportions are linked to the TABULA building types.	Yang et al., 2020

The datasets are downloaded for the City Region of Arnhem and Nijmegen. The extent of which parts of the area are included is shown in Figure 2.1 Bounding Box of the study area that is used for implementing and assessing the model.Figure 2.1. This area was chosen, because it contains a large variation in buildings of different sizes and time periods. In addition, my personal familiarity with this region is high. This can help in assessing model behavior and makes it easier to uncover striking or unexpected findings.

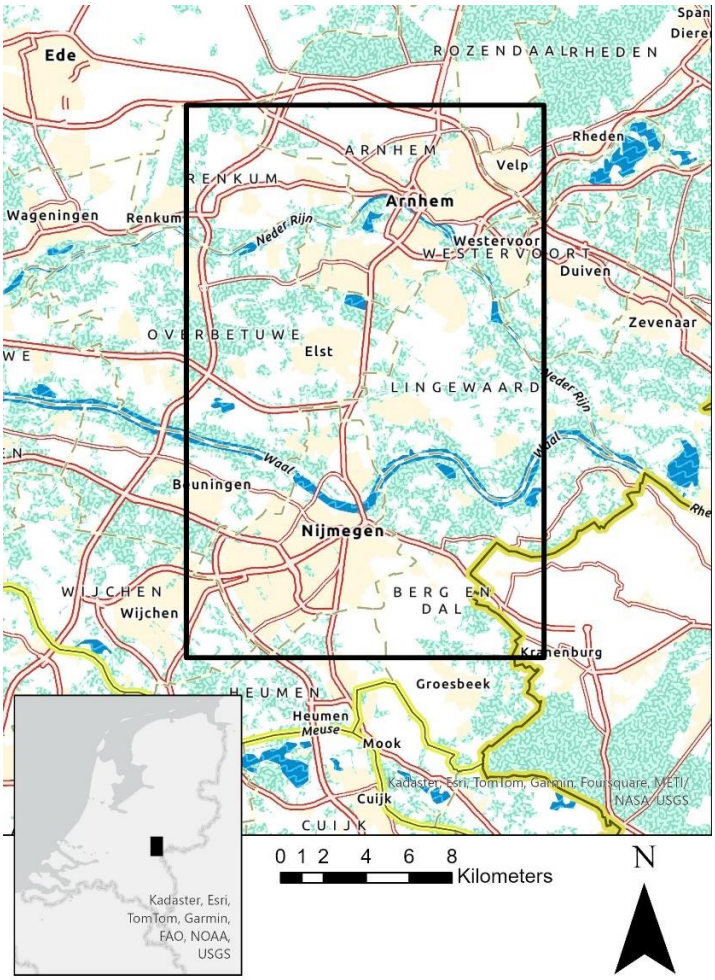


Figure 2.1 Bounding Box of the study area that is used for implementing and assessing the model.

2.3.2 Preprocessing data

To get the data ready as input for the physics-based model, some preprocessing steps are necessary. Figure 2.2 visualizes a flowchart that shows the preprocessing steps of the physics-based model.

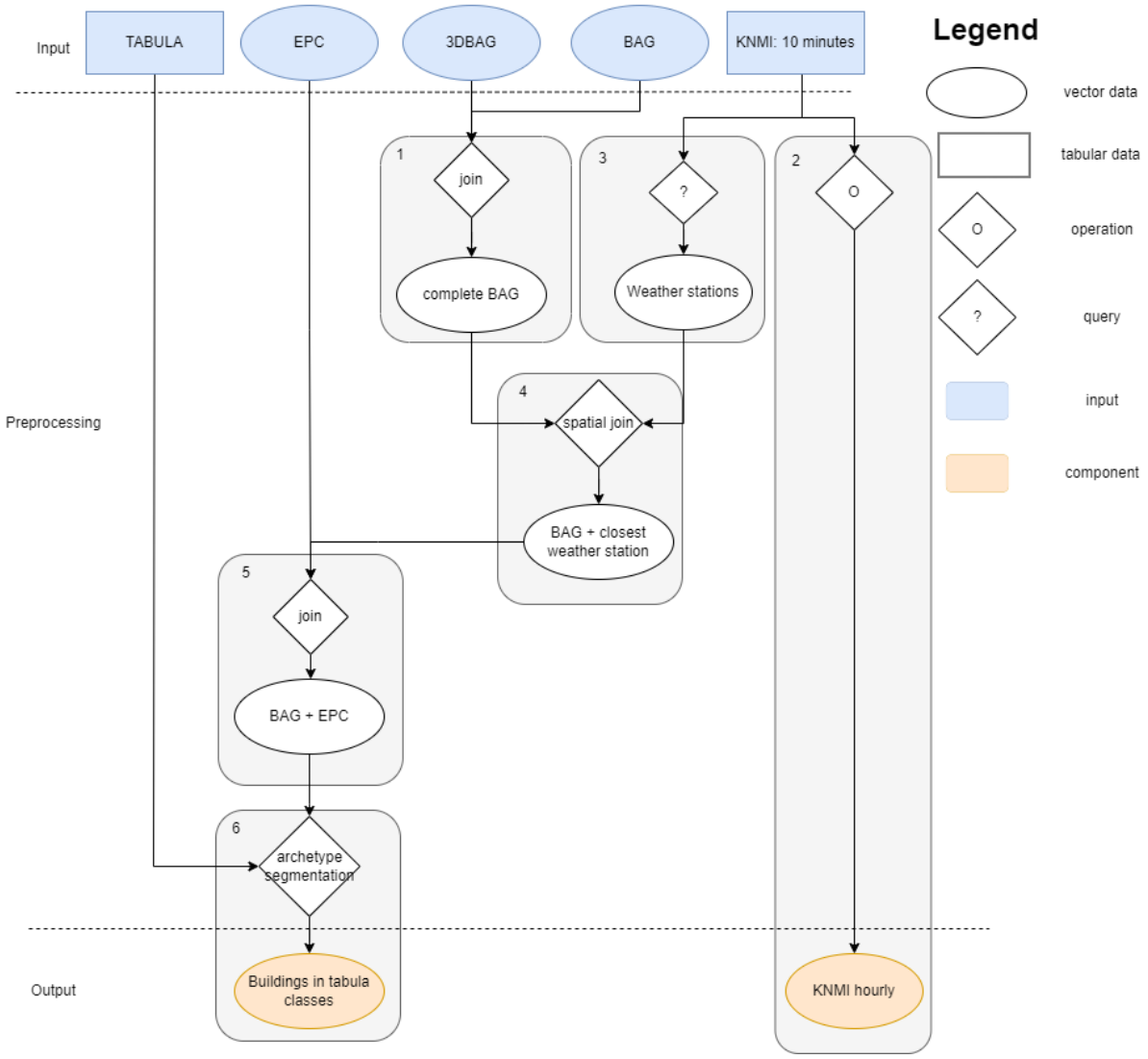


Figure 2.2 Schematic overview of the preprocessing steps of the input data.

I. Joining the 3DBAG to the regular BAG

The 3DBAG data is obtained via a WFS link. Geometric attributes from the 3DBAG are merged with those from the regular BAG dataset to acquire both spatial information and additional attributes such as building function and the count of residential units.

The building dataset is downloaded for the city region of Arnhem and Nijmegen. This choice was made because this region has a diverse building stock of various construction periods. In addition, it is a region where my personal familiarity is high, which can make it easier to come across interesting insights on the model's behavior. Though, in this thesis the model is only run for this region, it can easily be modified by changing the bounding box for the BAG and 3DBAG.

II. Getting hourly statistics for KNMI temperature and solar radiation datasets

The KNMI dataset is offered in a temporal resolution of 10 minutes. In order for the model to run in a manageable timeframe, the data is aggregated to an hourly resolution, by taking the mean of the variables for each hour. The result is exported as component data, that the model uses as input.

III. Extracting weather station locations from KNMI datasets

To assign the weather data from the closest KNMI weather station to each house in the building dataset, the locations of these weather stations are extracted from the KNMI dataset.

IV. Spatial join weather stations locations to building dataset

For each building in the building dataset, it is calculated what the closest KNMI weather station is, based on the Euclidean distance. The ID of the closest weather station is then joined to the building as a new column.

V. Joining EPC data to the houses

The available EPC data is joined to the building dataset through the BAG ID. For the buildings that do not have an EPC label, a NA value is assigned. Buildings that contain an EPC of level D or higher are considered to have had some type of renovation, while buildings with a lower or no EPC are considered to have undergone no retrofitting.

VI. Building stock segmentation based on TABULA archetypes.

The building stock dataset is segmented into four distinct archetypes: single-family homes (SFH), terraced houses(TH), terraced end houses (“hoekwoning”) (TH_End), apartment blocks (AB) and multi-family homes (MFH) This segmentation is based on a set of conditional statements that are based on the approach proposed by Yang et al. (2020). Details of these conditions can be found in Table 2.3.

Table 2.3 Conditional statements that are used for segmenting the building stock into the archetype.

Building class	No. of buildings together	Registered Addresses	Building footprint (m^2)
SFH	1	<= 2	-
TH	<=3	<= 3	-
TH_End	<=2	<=2	-
AB	<=1	>3	>1000
MFH	else		

Single family homes are mainly identified by the fact that they do not border any other house. To distinguish them from large apartment blocks the maximum amount of residences is set to two. Terraced houses are separated into end houses and middle row houses. Unlike middle row houses, end houses have one exposed wall, which in the data is counted as an outer wall but typically features fewer windows compared to the outer walls facing the street or the backyard. This configuration results in a lower overall window-to-wall ratio for end houses.

VII. Estimating number of floors, calculating conditional floor space

The 3DBAG dataset does not contain information on the number of floors a building contains. In this thesis, the floor height value is set at 3.3 meters. As maximum height, the median of the roof height is used. When visually checking the output with images of buildings, this value of 3.3 meters, showed the most correct estimation of the number of floors.

2.3.3 The physics-based model

The physics-based model is largely based on the models that were made by Todeschi et al. (2021) and Yang et al. (2020). A step-by-step guide of all the calculations is available in Appendix B.

The model works on a per building basis, which means that from the building dataset each building is analyzed separately. For each building it is calculated what the closest weather station is. Consequently, the hourly weather data of this station is linked to the building.

The model is centered around the thermal heat balance method. Equation 1 shows how the demand is calculated based on the heat balance.

$$Q_{nd} = (Q_{tr} + Q_{ve}) - \eta_{gn}(Q_{int} + Q_{sol}) \qquad \text{(Eq. 1)}$$

In this equation the energy losses are the transmission loss (Q_{tr}) and the ventilation loss Q_{ve}. The heat gains of the building are subtracted from these losses. The first component of the heat gains consist of internal sources (Q_{int}), such as residual heat from appliances and anthropogenic activities. The other component of heat gains are the solar heat gains (Q_{sol}). This is the incoming radiation that falls on a building's surface, resulting in an increase in temperature. It's relevant to note that not all gains directly translate into usable energy for the building (International Organization for Standardizaition (ISO), 2017). Thus, a gain utilization factor is derived to address this variability and calculate the hourly space heating demand of the building.

This space heating demand is then filled by the Heating, Ventilation and Air conditioning (HVAC) system of the building. Equation 2 calculates the actual demand of the HVAC system in kilowatt-hours (kWh). In the formula, the space heating demand is divided over the conditional floor space of the building. The heating losses of the distribution system per square meter are then added to this demand. This is then multiplied by the expenditure factor of the HVAC system and the total conditional floor space of the building to get the final space heating demand in kWh. To compare this result to the validation dataset, this result is converted to gas cubic meters.

$$Q_h = \left[\frac{Q_{nd}}{A_{con}} + q_{d,h}\right] e_{g,h} * A_{con} \text{ (Eq. 2)}$$

2.4 Integration socio-economic factors

For the socio-economic part of this research, postal code data by CBS is used. The dataset contains many different variables that possibly correlate with each other. To uncover which socio-economic variables play a vital role in determining the space heating demand of buildings, a random forest model is created. Through regression analysis, it is then tested if a linear relationship exists between the socio-economic variables and space heating demand.

2.4.1 Random Forest - Identifying important socio-economic variables

Preprocessing

The CBS dataset comprises three distinct spatial levels: pc4 (e.g., 1111), pc5 (1111A), and pc6 (1111AA). The pc5 dataset does not contain data on mean gas use per dwelling. This is therefore derived from the pc6 dataset by aggregating the smaller postal codes and calculating the mean gas use. Which socio-economic variables are available in the dataset and how they are represented is visualized in Table 2.4

Table 2.4 Overview of Socio-Economic Variables and their Representations in the Dataset.

Variables	**Units**	**Representation**
No. of inhabitants	No. of people	count
Gender	Gender [Male, Female]	count
Age groups	Age range [0-15, 15-25, 25-45, 45-65, 65+]	count
Background	Nationality [Dutch, Western, Non-Western]	count
No. of households	No. of people	count
Household composition	Type [single-person, multi-person without children, single-parent, two-parent]	count
Mean household size	No. of people	average
Mean property value (WOZ)	Euros (x1000)	average (x €1000)
Mean income	Classes [00-20, 00-40, 20-40, 20-60, 40-60, 40-80, 60-80, 60-100, 80-100]	Percentage groups
Social welfare payments	No. of people	percentage
Homeownership	% of total households	percentage
Renters	% of total households	percentage
Social housing	% of total households	percentage
Adress density	addresses / km^2	average
mean gas usage	m^3	average

Following this, it is checked for each separate spatial layer if there is multicollinearity. Some variables are mutually exclusive. For instance, the percentage of home-owned dwellings versus rented dwellings are opposing factors.

Variables such as age groups, gender, and household composition types are initially represented in absolute population numbers. To facilitate comparison among postal codes, these variables are transformed into relative numbers relative to the total population.

Model calibration

The final selection of variables is used as input for the Random Forest (RF) models on all three spatial scales. The data is split into a training dataset and a test dataset, where the training dataset takes up 70% of the data and the test set takes up 30%.

Using this split, two hyperparameters are tuned to reduce the degree of error in the model. The parameters that are tested are the max depth parameter and the number of trees parameter.

The model with the lowest error is used to perform on the test set. The performance metrics that are used to evaluate the model are the R-squared, the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE).

Lastly, the feature importances of the RF models are evaluated to check which socio-economic variables play a decisive role in estimating space heating demand. The feature importances are compared among the three different spatial scales. These metrics are then compared to a baseline model, to see if the implementation of the RF models has a beneficial effect compared to a random model.

2.4.2 Regression

Based on the RF models, the most important socio-economic variables are used as input for two regression models: Ordinary Linear Regression (OLS) and Ridge Regression.

Depending on the performance of the RF models, a spatial level is chosen to perform the regression on. Preferably it would be the spatial layer with the highest spatial resolution, however if the RF model performs poorly, it does not make sense to do a regression on this spatial scale. It is crucial to find a balance between statistical robustness and spatial resolution.

A first step in regression is to check if the assumptions are met. Through plotting, the assumptions of linearity and normality are checked. If variables exhibit a log-linear relationship, they are transformed to achieve linearity.

Once the assumptions have been verified, the OLS is performed with the variables that meet the assumptions. To prevent the model from overfitting, another regression model is developed, namely a ridge regression model. The ridge regression model aims to reduce the effect of multicollinearity by introducing a penalty term. This reduces the effect of the coefficients and produces a less biased result (Lin & Liu, 2017). The ridge regression is combined with a bootstrapping procedure, where the dataset is resampled 20000 times, to calculate the P-value for each of the coefficients.

After both the OLS and Ridge Regression are performed, the resulting coefficients are compared to see if large discrepancies occur. Depending on the result, the coefficients for the variables age, tenancy type and migration background are used to alter the prediction of the physics-demand model.

2.4.3 Integration of socio-economic variables

Age, Tenancy Type & Background

Because of the fact that there is only proportional socio-economic data available on a postal code level, a weighted coefficient is calculated for each variable per neighborhood. As shown in the aforementioned Table 2.4, the variables age, tenancy type and background all have subgroups with total population numbers belonging to that subgroup. For each of the subgroups a different coefficient is assigned. By multiplying the percentage of people that is part of each subgroup with the respective coefficient, you get a weighted coefficient for the entire neighborhood. This weighted coefficient is then applied to the space heating demand of all buildings within the neighborhood.

Age integrated occupancy schedule

In addition to the incorporation of regression coefficients, two alternative methods of integrating socio-economic factors are developed. What sets these approaches apart is that they are directly integrated into the physics-based model, instead of applying them in retrospect.

The first variable that is implemented is the sophistication of the occupancy schedule. The occupancy schedule in the physics-based model assumes that everyone is home between 18 and 8 the next morning. However, actual occupancy patterns vary substantially based on life stage. Therefore, the occupancy schedule is recalibrated per neighborhood, based on the age demographics proportions. This calibration is based on a paper by Mitra et al. (2020). In this paper, they used an extensive American survey to distinguish different occupancy patterns for different age groups.

Since the actual age of the inhabitants of individual houses is not known, an alternative method is used to estimate occupancy. Each of the different age groups, mentioned previously in Table 2.4, is used as a weight, to estimate the probability of individuals being present at each hour of the day. This creates a mean occupancy schedule that is adapted to the age proportions within a neighborhood. This new occupancy schedule is then used to run the physics-based demand model again.

Price elasticity of households

The final socio-economic variable to be integrated is the price elasticity of space heating demand. Schulte & Heindl (2017) analyzed how long term price elasticity of space heating demand varies between different income classes and different household compositions. The derived price elasticities from their study are then incorporated into the physics-based demand model. These price elasticities are also available in Appendix C. The price elasticities that they calculated in their model are applied in the physics-based demand model. For the actual gas price, the price index dataset by the CBS (2024) is applied. To effectively integrate price elasticity, the temporal scope of the physics-based model is extended to encompass the period from 2016 to 2019.

Three different implementations of price elasticity are developed:

- The first implementation considers the price exactly one year prior for each month of the model's run. It then adjusts the space heating demand based on the observed price increase or decrease within that time frame.
- The second implementation compares each month's price to a base value of January 2015. The space heating demand is recalculated based on the price fluctuation since 2015.
- The third implementation repeats the procedure that is used in the second implementation but uses the base year of 2010 to redefine the space heating demand.

2.5 Validation and assessment

Because there is no publicly available validation data at the building-scale level, validation must be conducted on a higher spatial level. To validate the space heating demand, the mean gas use per house column of the "Kerncijfers per postcode" dataset is used (CBS, 2023a). To ensure accurate comparison between the model predictions and the validation dataset, several steps are taken to align them to the same spatial level.

Firstly, predictions for all houses in the neighborhood are aggregated per postal code area. Similarly, the validation dataset's total demand is calculated by multiplying the mean gas demand per dwelling by the number of dwellings per postcode.

Secondly, since the model does not encompass all buildings with residential functions, the predictions are adjusted by multiplying them by the fraction of missing houses. This adjustment ensures the closest possible comparison between the validation data and the prediction dataset.

3 Results

The result section is structured in alignment with the aforementioned research questions. Section 3.1 presents findings from the literature review on papers employing a physics-based model. Section 3.2 outlines results derived from the literature review concerning socio-economic models. Following this, Section 3.3 provides the outcome of the constructed physics-based model. Section 3.4 will provide the results on the integration of socio-economic variables in the physics-based model.

3.1 Literature review of physics-based models

3.1.1 Search results

The query for papers on Web of Science and Scopus returned a total of 1583 papers. Out of these, the first 256 abstracts were reviewed using the ASReview Lab tool. However, beyond this point, the algorithm primarily presented papers that were deemed less aligned with the topic of this thesis. Of these 256 abstracts, 36 papers were selected for in-depth reading, because these papers focused on the space heating domain of energy demand modelling.

3.1.2 Input variables

The physics-based models vary in different dimensions. Some models are detailed thermodynamic models of a select group of buildings, while other models focus on developing a model that can generalize to a larger area. Some models focus on having a high temporal resolution, while other models focus more on predicting accurate monthly or yearly energy demand. Consequently, a large variation of input factors for physics-based models exists.

Figure 3.1 shows all the input factors that are used as input for the physics-based models. The most frequently mentioned factors were building typology, ground floor area, building year, and 3D geometry. The majority of the physics-based models use pre-defined building typology to segment the building stock into generalizable types. These typologies then determine the values of other factors, like the window-to wall ratio or transmittance values.

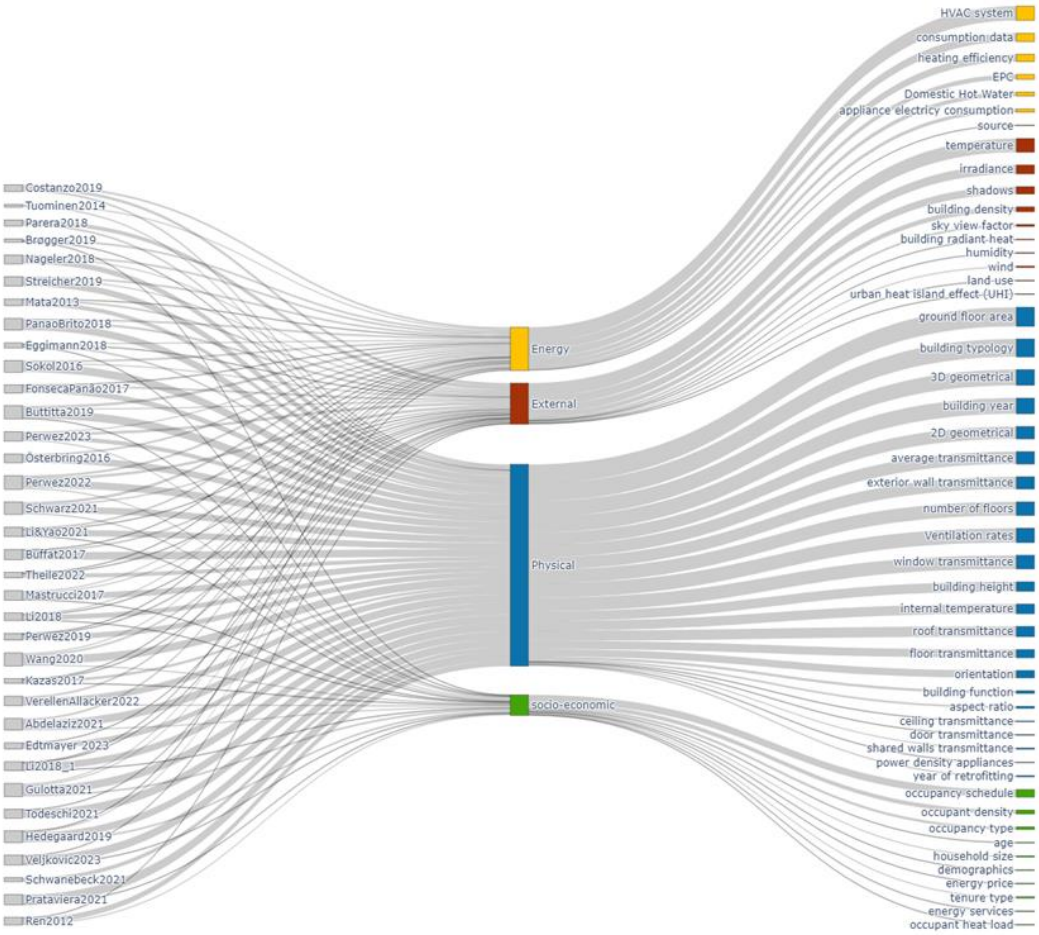


Figure 3.1 Sankey diagram, showing the categories of the input factors and the specific input factors that are used by the different physics-based demand models.

The models vary to what detail transmittance processes are included in the model. For example, the model by Ledesma (2021) uses separate U- and R_c– values for every buildings element (door, windows, ceiling, internal walls, external walls, floor and roof). Most models use a selection of these elements, of which external wall- and window transmittance are most often mentioned.

In addition to physical factors of the house, many models also use energy related factors as model input. The most frequently mentioned factor in this regard is the type of HVAC system, which is consequently linked to its efficiency. Some models also use existing consumption data of houses to further calibrate the model. There are models that used monthly metering data of individual houses (Brøgger et al., 2019; Kristensen et al., 2018; Perwez et al., 2022; Sokol et al., 2017; Theile et al., 2022). However, because this data is often difficult to access due to privacy constraints, other models that want to calibrate the model with energy consumption data use level from neighborhood or postal code areas(Buttitta et al., 2019; Eggimann et al., 2019; Mastrucci, Marvuglia, et al., 2017; Zhang et al., 2018).

Most models also integrate external factors into the energy demand model. The most frequent external factors were outside temperature and solar irradiance. These two weather factors have the largest influence on the internal temperature of the building, which in turn influences the space heating demand. More sophisticated models also take external factors like shading and

building density into account. The model of Costanzo et al. (2019) really laid emphasis on the external factors by also including wind, sky view factor, and humidity into the model.

Lastly, some models included socio-economic factors to further calibrate the model. This was especially pronounced in models that aimed to calculate hourly energy demand, because in order to know hourly energy demand, one needs to know the schedule of the occupant first (Abbasabadi et al., 2019; Hu et al., 2016; Palacios-Garcia et al., 2018; Veljkovic et al., 2023). These occupancy schedules were mostly based on predetermined standards. Other socio-economic factors that are sometimes included are household size and occupancy type. The other socio-economic factors in Figure 3.1, like poverty, tenure type, and income, are part of the model by Abbasabadi (2019), which actually has a socio-economic emphasis. This model, therefore, does not include many physical building factors, apart from building footprint and building year (Abbasabadi et al., 2019).

3.1.3 Performance

In order to check whether a model is good at predicting space heating demand, validation must be performed. Figure 3.2 shows the metrics that are used in existing literature to express the accuracy of the model. The abbreviations in Figure 3.2 are included into the nomenclature in Appendix D. The figure shows that the most commonly used accuracy metrics are the R-squared, Coefficient of Variation for the Root Mean Square Error (CVRMSE) and the Root Mean Square Error (RMSE). The R-squared and the CVRMSE have a great benefit, because they are unitless. Models that use these performance metrics are, therefore, easier to compare.

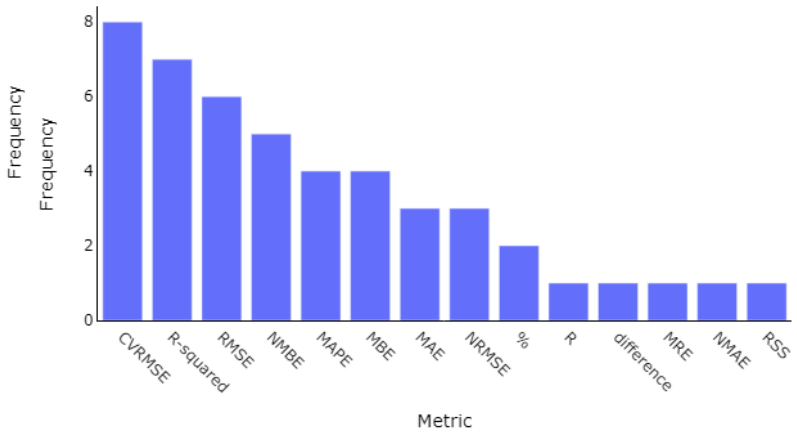


Figure 3.2 shows which models included CVRMSE and/or R-squared to express the accuracy of the model. The table shows that Todeschi et al. (2021) have the highest R-squared of all the papers that used this assessment metric. In their paper, they develop three different models: a gradient boosting Machine Learning model, a simplified GIS thermal balance model and an extensive GIS model, using the software CitySim Pro. They identified that the Machine Learning algorithm provided the highest accuracy, but also has the highest risk of overfitting. For the Machine Learning algorithm to work, you would need a labelled dataset. Therefore, they also developed a simplified physics-based GIS model. This simplified GIS model had a lower accuracy of 0.794 but is less sensitive for overfitting.

The models by Ledesma et al. (2021) and Geraldi & Ghisi (2022) show low CVRMSE values on a building scale. Both models developed a model that predicts the Energy Use Intensity (EUI) for school buildings. Because this is a relatively homogenous group of energy users, it can be expected that a lower error can be established when modelling the EUI.

Veljkovic et al. (2023) and Eggimann et al. (2019) both have good performance scores. Veljkovic et al. (2023) have a low CVRMSE value of 2.67% and Eggimann et al. (2019) have a R-squared of 0.92. However, both of these assessments have been done on the neighborhood scale, rather than the building scale. By doing this, you reduce some of the variation between buildings, which makes achieving a lower CVRMSE and a higher R-squared value more attainable.

Table 3.1 Performance measurement of physics-based demand models from literature. Empty cells imply that the paper did not disclose this information.

	R-squared	CVRMSE (%)	timescale	spatial scale
Todeschi et al., 2021	0.993		daily	building
Eggimann et al., 2019	0.92		yearly	neighborhood
Brøgger et al., 2019	0.816	105.4	yearly	building
Marvuglia, et al., 2017	0.794			building
Perwez et al., 2022	0.77	13.2	yearly	national
Schwanebeck et al., 2021	0.7		yearly	urban
Buffat et al., 2017	0.6			building
Veljkovic et al., 2023		2.67	yearly	neighborhood
Ledesma et al., 2021		4.6	yearly	building
Geraldi & Ghisi, 2022		8.17	yearly	building
Buttitta et al., 2019		10.2		building
Wang et al., 2020		11.5	yearly	neighborhood
Gulotta et al., 2021		15	yearly	national
Hedegaard et al., 2019		30		
Nageler et al., 2018		40.2	yearly	building
Sokol et al., 2017		66	yearly	building

3.2 Literature review socio-economic models

The query for socio-economic space heating demand models yielded 178 papers. Upon examining the abstracts through ASReview Lab, 24 papers were deemed relevant for this thesis, because they primarily focused on the relationship between socio-economics and space heating demand and less on other fields of energy demand. These 24 papers were reviewed, and a summary of the findings are presented in the following sub-questions.

3.2.1 Model approaches

Because the domain of socio-economics is so broad, the topic of energy consumption is analyzed from numerous perspective orientations. These orientations are service, expenditure, and socio-economic characteristics.

Service oriented

The first type of models uses services (e.g. appliances) as the starting point of the research. While these models also focus on space heating and domestic hot water, they often include a wide range of electrical devices as well. Therefore, these papers are often less relevant in the case of this research. One example of a model that aims to explain energy demand through residential services is the model by Jia et al. (2023). In this model they aim to disaggregate energy demand across appliances through a conditional demand model. By doing this, instead of only focusing on the household as a whole, more nuanced patterns of energy demand can be found. Another paper that is service oriented is the discrete choice model that was constructed by Han et al. (2022). In their research, they used microdata to distinguish the choices people make on what energy services to use and how socio-economic variables can help explain these choices.

Expenditure oriented

The second set of models takes expenditures withing the household as a starting point. These models prioritize understanding how individuals respond to price fluctuation in various energy consumption areas. They often seek to identify different elasticities based on social factors such as age and household composition. The papers by Rehdanz (2007) and Schulte & Heindl (2017) contain models that take expenditures as a starting point. In their paper, Schulte & Heindl (2017) aim to uncover how differences in household composition and household income relate to differences in energy demand elasticity. This knowledge can help in predicting energy demand alongside different kinds of households.

Socio-economic characteristics oriented

The third category of models primarily investigates the direct relationship between socio-economic characteristics and space heating demand. These models mostly do not include macro-economic trends like price fluctuations. Instead, they are focused on explaining yearly energy demand by socio-economic variables. The models are largely build on combining historical energy demand with regression methodologies or their variants, including quantile regression, ridge regression, elastic net regression (Bakaloglou & Charlier, 2021; Belaid et al., 2020; Belaid & Rault, 2021; Çebi Karaaslan & Algül, 2023; Harold et al., 2018; Lawal et al., 2021; Meier & Rehdanz, 2010; Schmitz & Madlener, 2020; Wiesmann et al., 2011). In these regression models, the aim is to find out which socio-economic characteristics of households contribute the most in explaining the variation in energy demand.

3.2.2 Performance of regression models

In Table 3.2, an overview of model performances from different papers is shown. The table shows that most OLS regression models tend to have a lower performance than more sophisticated models like Machine Learning, Quantile Regression, Random Effect Regression and Conditional Demand Models. While it looks like the model by Lawal et al. (2021) has the highest performance, it has to be noted that this model is assessed on a higher spatial level of zip codes, while the other models are assessed on the household level.

Table 3.2 Performance comparison of socio-economic models for predicting space heating demand.

Paper	Model orientation	Method	Spatial level	Temporal level	R^2
(Schulte & Heindl, 2017)	Expenditures	Quadratic expenditure system	household	annual	-
(Lawal et al., 2021)	Characteristics	Machine Learning, OLS	zip code	annual	0.95
(Belaid et al., 2020)	Characteristics	Quantile regression	household	annual	0.76
(Jia et al., 2023)	Services	conditional demand	household	annual	0.673
(Harold et al., 2018)	Characteristics	Random effect regression	household	daily	0.585
(Çebi Karaaslan & Algül, 2023)	Characteristics	OLS Regression, Quantile Regression	household	annual	0.41
(Matsumoto, 2023)	Expenditures	consumption elasticity model	household	annual	0.335
(Han et al., 2022)	Services	Discrete choice model	household	annual	0.33
(Wiesmann et al., 2011)	Characteristics	OLS Regression	household	annual	0.322
(Meier & Rehdanz, 2010)	Characteristics	OLS Regression	household	annual	0.162
(Schmitz & Madlener, 2020)	Expenditures/	Quantile regression	household	annual	0.1103

3.2.3 Significant socio-economic coefficients

When analyzing the individual effect of different factors in socio-economic models, several variables consistently show a significant influence. Figure 3.3 illustrates the percentage of research papers where specific variables have a statistically meaningful impact on energy demand. Among the most prominent socio-economic factors are age, household size and household composition. Additionally, factors related to the building itself and the climate, such as the number of rooms, heating degree days, and ground floor area, also frequently appear as significant influences.

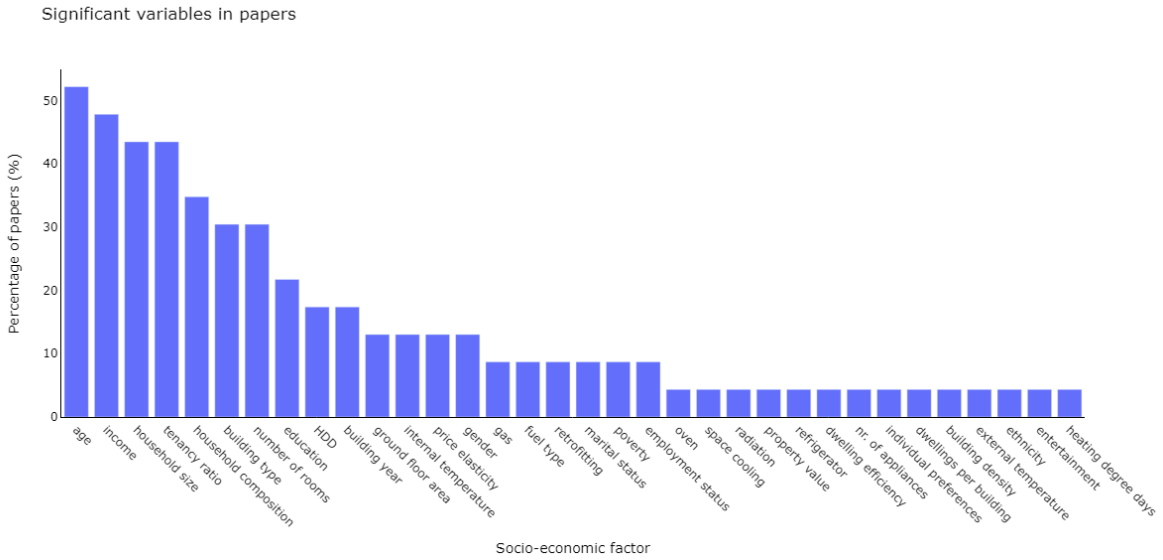


Figure 3.3 Variables from socio-economic with significant coefficients

3.3 Physics-based model

3.3.1 Archetype segmentation

Building types

The implementation of the conditional statements by Yang et al. (2020) shows a relatively accurate representation of the building stock in Nijmegen. Because no labelled data is available, the quality can only be checked visually. When looking at the dataset for Arnhem and Nijmegen it shows that the conditional statement fits relatively homogenous neighborhoods well, but performs less good in neighborhoods that are older and have a more chaotic distribution and shapes of houses. Figure 3.4 shows how this plays out in the Nijmegen neighborhoods "Neerbosch-Oost" and "Altrade".

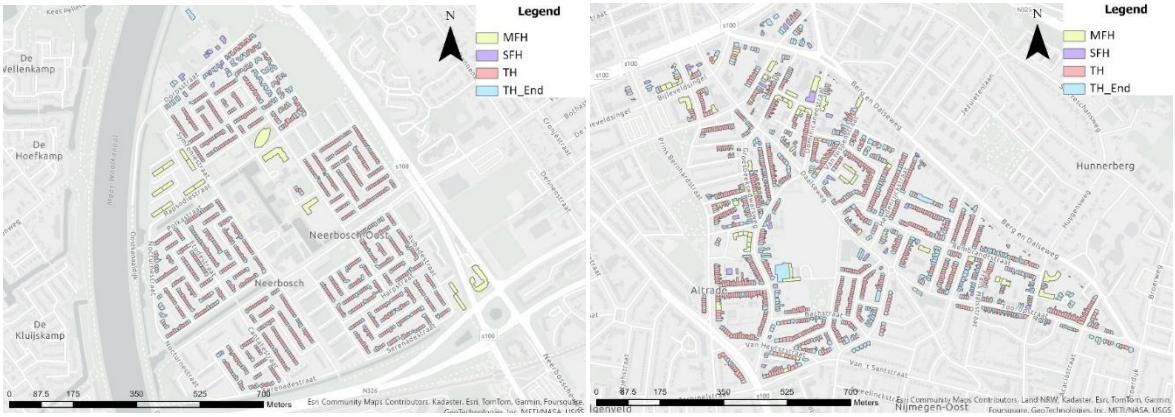


Figure 3.4 A comparison of building topologies between two neighborhoods. On the left, Neerbosch-Oost illustrates a uniform distribution of building types, while on the right, Altrade showcases a diverse range of building types (MFH = multi-family home, SFH = single family home, TH = terraced house, TH_End = terraced end house

On the Neerbosch-Oost map, which is a relatively homogenous neighborhood built during the 1970s, it shows how the model distinguishes the different building types well. On the most northern street "Dorpsstraat", the single-family homes are labelled correctly. One inconsistency in this neighborhood are the buildings on the other side of the Dorpsstraat. These buildings show

a mix between labelled terraced house, or terraced end house. When looking at Street View images by Google (2023) in Figure 3.5, it shows that these houses are attached to each other by only a garage box. This puts them somewhat in between being a middle row house and an end house, resulting in inconsistency in the segmentation.

Figure 3.5 Houses on the Dorpsstraat in Neerbosch-Oost, linked to each other by a garage box.

In Altrade, it shows that most buildings are classified correctly, however some buildings with very irregular shapes, like the large, terraced end house in the middle of the map (in blue) do not seem to be correct, due to the BAG input dataset.

The satellite image in Figure 3.6, shows a building complex in Altrade that is classified as a terraced end house and a multi-family home. It seems very unlikely that this entire building has a residential function. Consequently, this wrong typology leads to an overestimation of energy demand. These examples illustrate some of the challenges encountered in the classification of buildings, which are not limited to Altrade but also extend to other areas within the study region, predominantly resulting in an overestimation of energy demand.

Figure 3.6 Zoom-in on buildings in Altrade that are classified as terraced end house and multi-family home.

Building renovation rates

In the TABULA/episcope building typology, three distinct refurbishment scenarios are delineated, each associated with specific U-values tailored to address varying degrees of energy efficiency. These scenarios encompass the existing state, usual refurbishment and advanced refurbishment. In this thesis, buildings with an EPC label D or higher are considered to have undergone a usual refurbishment, while buildings without an EPC label or a label lower than D, are assigned to the existing state class. The refurbished state has better insulation measures than the existing state. The effect of the measures on U-values is dependent on the building age group. Figure 3.7 shows how the building stock of the Arnhem-Nijmegen region is divided into existing state and refurbishment groups. For the oldest and largest age group, it is assumed that around a third is renovated. For the two consecutive construction periods after that, it is assumed to be about half. For the newest construction periods, the renovation rates are lower, as this has not been necessary yet.

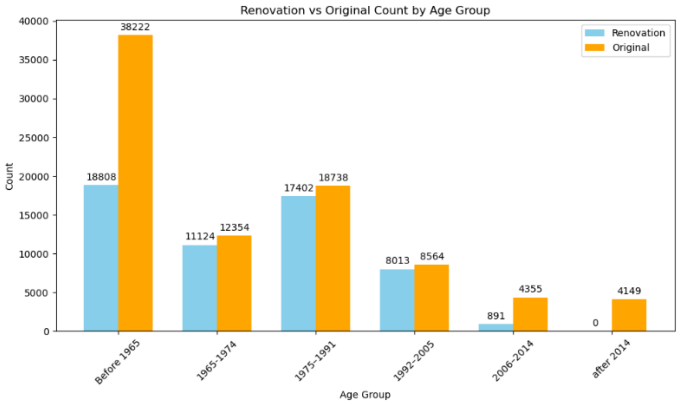


Figure 3.7 Distribution of the Arnhem-Nijmegen building stock, depicting houses categorized based on EPC data in the renovation or original state categories.

Figure 3.8 shows how some of the insulation rates in the model compare to national rates from CBS (Kloosterman et al., 2021). The graph reveals that while the model closely aligns with national standards for wall insulation, it tends to overestimate floor insulation and underestimate roof insulation. These disparities can influence the performance of the model.

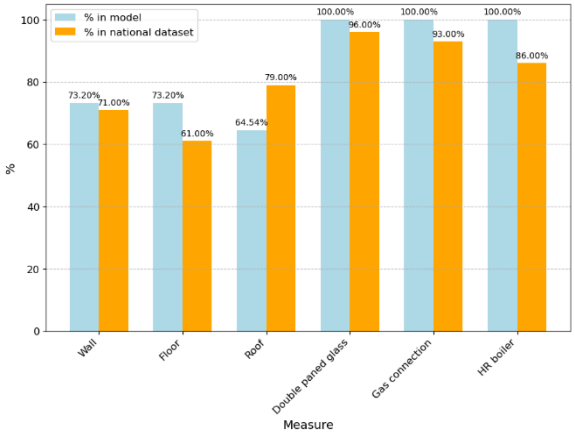


Figure 3.8 Insulation rates and HVAC system rates in the model compared to national rates.

3.3.2 Spatial findings

Yearly statistics

To understand how the model behaves spatially on building data, the average building gas use in the municipalities of Arnhem and Nijmegen is visualized in Figure 3.9. The maps reveal a distinct trend: areas near the city center and older neighborhoods tend to exhibit higher predicted gas usage, whereas regions farther from the urban core, often comprising newer buildings, show lower gas consumption.

Notably, certain postal codes display an average gas usage rate exceeding 10,000 m^3 per building, according to the model predictions. Upon inspecting the building typology dataset, it becomes evident that these postal codes only contain large apartment blocks. As the analysis considers the average gas usage per building, rather than per address, neighborhoods that predominantly contain this building type show substantially higher averages.

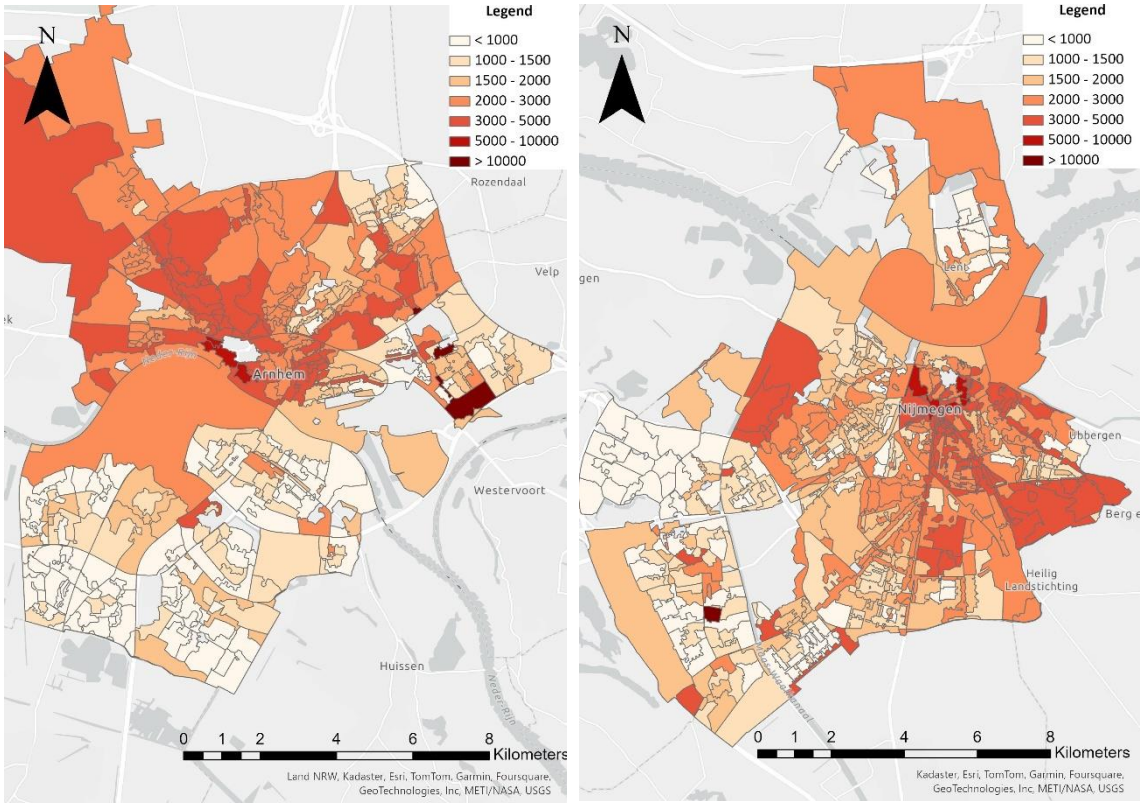


Figure 3.9 Mean gas use of all buildings within postal code area predicted by the physics-based model in Arnhem (left) and Nijmegen (right).

Temporal zoom-in

Four buildings were selected so that the model behavior can be inspected on a daily scale. These buildings are selected from the three different building classes. In Figure 3.10, the daily demand is given for each of these buildings. It can be seen that all buildings follow the same pattern. This makes sense, because all of these houses have the KNMI weather station of Deelen as the closest station. Therefore, the pattern is the same. The magnitude is what differs: it can be seen that the large apartment block consumes the most energy, followed by the old, terraced house. The old, terraced house has higher U-values, which leads to a higher demand. This is then followed by the Single-Family Home and then the new Terraced House. It makes sense that the Single-Family Home is higher, as this building has more exterior walls than the new terraced house.

Figure 3.10 Graph of daily energy demand fluctuations for four sample buildings, each belonging to a distinct building class. (MFH = Multi-Family Home, TH = Terraced House. SFH = Single Family Home)

3.3.3 Model performance

In Table 3.3, various model configurations their respective performances are delineated. The results depict varying performance across the different configurations. Notably, Run1, Run2 and Run6, have the lowest CVRMSE and NMBE values. This indicates that the actual and predicted demand rates are closely aligned. However, Run4, Run5, and Run9 boast higher R^2 values. This means that these models are better at showing the distribution of the postal codes, though their actual values are biased and are further removed from the actual predictions. These changes in performance metrics are caused by the incorporation of the 3D volume variable, replacing the approximation of the conditional floor space combined with average floor height.

Table 3.3 Performance metrics for different iterations of the model and the three different postal code layers.

		CVRMSE				R-squared				NMBE			
Run	Description	pc4	pc5	pc6	pc6_WO	pc4	pc5	pc6	pc6_WO	pc4	pc5	pc6	pc6_WO
run1	• conditional floor space HVAC • ground floor area for DHW	25.23271228	33.13826791	46.85635	44.9422	0.834352	0.678355	0.571657	0.526808	0.024103	0.136528	0.08426	0.089713
run2	• conditional floor space HVAC & DHW	28.7639672	31.48798137	48.43041	44.96509	0.848187	0.703126	0.619411	0.573908	-0.11112	0.016507	-0.03613	-0.02876
run3	• volume column HVAC • ground floor area for DHW • modified infiltration rates • applies ventilation recovery system to new buildings	45.41938288	42.8893193	66.5666	60.94276	0.836135	0.677533	0.540034	0.483476	-0.28383	-0.13436	-0.20241	-0.18995
run4	• volume column HVAC • ground floor area for DHW • TABULA infiltration constants	95.70787597	77.42682783	102.5634	87.63052	0.874623	0.745065	0.693559	0.642908	-0.78696	-0.58342	-0.65272	-0.61957
run5	• conditional floor space DHW. • volume column HVAC • applies ventilation recovery system	90.5354253	73.47591487	99.13623	85.46573	0.899932	0.757758	0.684038	0.627933	-0.74672	-0.54748	-0.62485	-0.59445
run6	Run 1, with widened occupancy schedule	27.15275083	33.34450139	50.15349	47.46725	0.830746	0.67447	0.564796	0.517398	-0.04706	0.073632	0.017346	0.024237
run7	Run 1, modified infiltration rates	28.64704198	38.26485482	50.11467	49.25299	0.818013	0.640978	0.502644	0.452102	0.124697	0.225395	0.171434	0.174861
run8	Run 3, modified infiltration rates	58.61148524	50.73991509	74.60202	66.69145	0.848197	0.697591	0.58476	0.525901	-0.42353	-0.25994	-0.32839	-0.31155
run9	Run 5, modified infiltration rates	57.938568	50.184416	74.36	66.48975	0.84943	0.699488	0.580877	0.520635	-0.41724	-0.25315	-0.32214	-0.3053

The performance of the model also differs spatially. In Figure 3.11, the standardized residuals of Model Run1 are mapped for the numeric postal code areas of Arnhem and Nijmegen. The residuals are calculated by subtracting the prediction by the true demand and then dividing it by the standard deviation. In Nijmegen a striking find is that neighborhoods that were primarily constructed after 2000 tend to be massively overestimated, while neighborhoods that were primarily constructed during the 1980s and 1990s tend to be underestimated by the model. This underestimation is also visible in Arnhem.

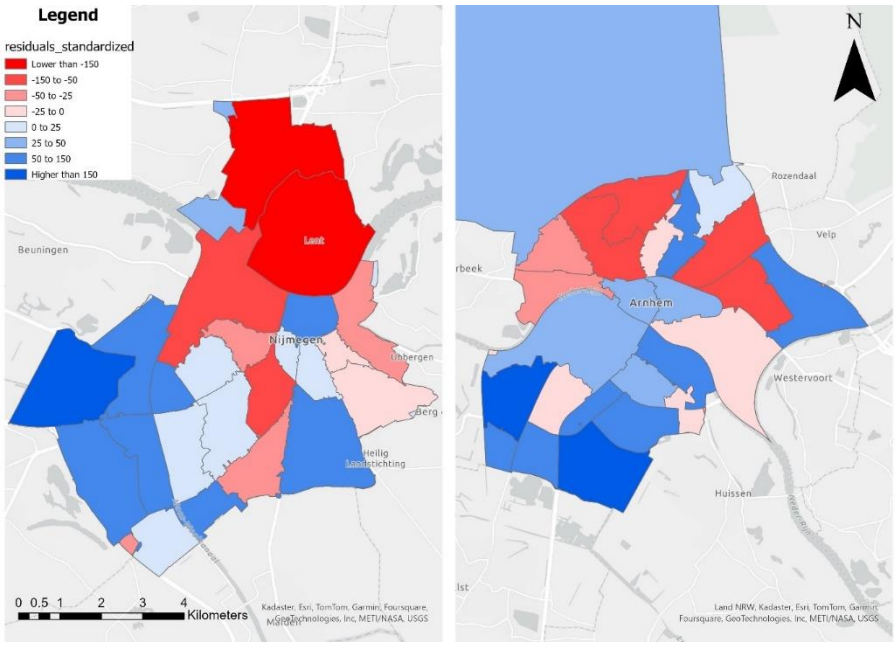


Figure 3.11 Standardized residuals for the physics-based model Run1 on the pc4 spatial level for Nijmegen (left) and Arnhem (right).

3.4 Integration of socio-economic factors

3.4.1 Important variables from Random Forest

Hyperparameter tuning

When calibrating the RF models, the hyperparameters of the number of estimators and the max depth parameter were set. A value was chosen based on the graphs visualized in Figure 3.12. In order to find a balance between reduction in RMSE and overfitting, the max-depth parameter was set at 7. This choice resulted in an approximate RMSE reduction of 100 gas m^3. While the number of estimators does not contribute to overfitting, it does influence the runtime of the model. Consequently, to optimize computational efficiency without compromising performance, the number of estimators was fixed at 40.

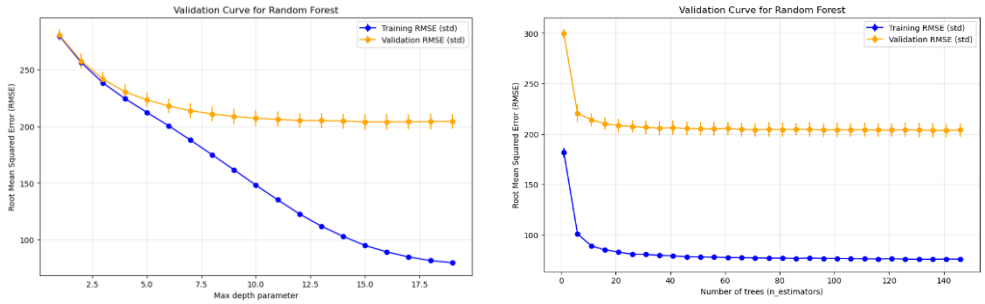


Figure 3.12 RMSE values for different settings of hyperparameters for the training dataset and the validation dataset for the Random Forest model on the pc5 level.

Performance

The Random Forest model's performance was evaluated by conducting validation across three spatial levels, with results summarized in Table 3.4. The PC4 and PC5 layers both show a reduction in RMSE and have a moderate to high R-squared value of 0.683 and 0.600, respectively. When comparing these R-squared values performances of the socio-economic models (Table 3.2), these performances are highly satisfactory.

However, the model struggles when it comes to the most local postcode layer, PC6. This is mainly due to the stringent privacy measures in place at such a fine-grained level. To protect individuals' privacy, large chunks of data are anonymized, making it challenging for the model to derive accurate insights from this highly localized data.

Table 3.4 Performance of the Random Forest model on the three different spatial postal code scales.

Postcode level	**RMSE**	**R-squared**
PC4	200.2	0.683
PC5	215.9	0.600
PC6	388.1	-0.01

Feature importance

The RF models provide a convenient means to assess which variables are most influential in explaining the variation in gas demand1. Only the feature importances for the PC4 and PC5 layers were assessed, as the PC6 model did not have any explanatory significance. The feature importances are visualized in Table 3.5. Both the pc4 and pc5 models highlight similar key variables. The most important variables include tenure type (renting or owning of dwelling), age demographics, migration background, and estimated property value (WOZ-waarde). These variables are utilized as input for the regression analysis.

Table 3.5 Feature importances for the Random Forest model on the PC4 and PC5 spatial scales.

Variable	feature	pc4	pc5
Tenure type	Home-owned dwelling	0.02	0.28
Migration background	Dutch	0.13	0.18
Age	25-45	0.19	0.17
Average property value	Average property value (WOZ)	0.06	0.14
Tenure type	Social housing	0.32	0.07
Age	65 and over	0.09	0.03
Household composition	two-parent	0.01	0.03
Total dwellings	Total dwellings	0.01	0.01
Household composition	one-person	0.01	0.01
Gender	Male	0.01	0.01
Age	15-25	0.01	0.01
Household composition	Uninhabited	0.01	0.01
Total inhabitants	Total inhabitants	0.01	0.01
Total households	Total households	0.01	0.01
Household composition	Household composition: single-parent	0.02	0.01
Age	0-15	0.02	0.01
Household composition	multi-person without children	0.03	0.01
Age	45-65	0.04	0.01
Gender	Female	0.01	0
Migration background	Western	0	0
Social benefits	Social benefits		0

3.4.2 Regression results

Because of the problem of multicollinearity, the OLS regression can only take a limited number of variables. Therefore, one feature per variable was selected as input for the model. The derived coefficients and their significance are shown in Table 3.6.

*Table 3.6 Coefficients for the OLS pc4 and pc5 levels. Significance levels p < 0.01 = ***; < 0.05 = **; <0.1 = **

Variable	PC5 coefficient	PC4 coefficient
Average property value (log)	372.80***	260.00***
Household size	-8.51	-95.02***
Home-ownership rate	3.01***	6.2150***
Age group 45-65	8.36***	20.94***

To get a better idea of how the separate features of a variable behave in a regression model, multiple iterations of a ridge regression model were executed. These results are visualized in Appendix F. The ridge models lead to a number of findings.

Firstly, across different age brackets, the coefficients reveal diverse relationships with mean gas consumption. For instance, a higher percentage of individuals aged 15-25 consistently correlates with increased gas demand, as indicated by consistently positive coefficients. Conversely, the age group of 45-65 consistently demonstrates a positive impact on gas demand, implying that higher proportions of older individuals are associated with heightened gas consumption. However, the variable representing individuals aged 65 and older shows ambiguity regarding its influence on gas demand, as none of the coefficients reach significance. Thus, conclusive interpretations regarding its effect cannot be made with certainty.

The variable representing tenure type, which is divided into homeownership and renting percentages, consistently shows a pattern. An increase in the proportion of households owning their homes correlates with higher mean gas demand, whereas a higher fraction of renting households tends to have a negative impact on mean gas demand.

Among the iterations, the variable of migration background shows that a larger share of households with a Dutch backgrounds tends to have a positive influence on the mean gas demand. In contradiction, a higher share in households with a western and non-western migration background has a negative influence on the mean gas demand.

Lastly, across the different iterations, the mean household size tends to have a negative influence on the mean gas demand.

3.4.3 Integration

Because of the large sizes of the rewritten predictions, maps showing the spatial changes in prediction values for the integration of the socio-economic variables tenancy type, cultural background, and age are included in Appendix G, H, and I. The difference is calculated by subtracting the new prediction from the old prediction. A positive percentage, therefore means that a decrease in energy demand was established, and a negative percentage means that an increase in energy demand was established.

Appendix G shows how the integration of the tenancy type variable affects the gas demand predictions. It shows that primarily neighborhoods in the cities of Nijmegen and Arnhem are expected to have a reduction in gas demand, as these neighborhoods contain a larger share of rental housing. On the other hand, neighborhoods in nearby villages like Beuningen, Bemmel, and Huissen show an increase in gas demand, as in these areas, the share of home-owned houses is larger.

Appendix H shows how the predictions change when the cultural background coefficients are integrated. It shows that urban neighborhoods with a larger share of people with a migration background, have a reduction of a maximum of around 1% in energy demand and areas with predominantly people without a migration background show a small increase in demand of around 0 to 0.3%. These areas are mostly located on the urban fringes or in the villages outside of the urban cores.

In Appendix I, the effect of integrating the age regression coefficients is mapped. The map shows that the prediction of gas demand increase in areas where primarily older people live. This is most prominently visible in the villages of Oosterbeek, Velp, and the eastern part of Nijmegen. Areas in

the urban cores of Arnhem and Nijmegen show a reduction in energy, because these areas are primarily inhabited by younger generations.

It's worth noting that the effect of the integrations on the predictions are small, typically less than 1%. This is because the coefficients applied, along with the weighted proportions for each neighborhood, soften the impact of socio-economic variables. Though the effect on the predictions themselves are small, they still have an effect on the overall performance of the model.

Performance of the model after the integration of the variables age, tenancy type, and cultural background.

The graphs in Figure 3.13, show the change in performance metrics after the integration of the three socio-economic variables. For the variable age, the effect depends on the iteration of the physics-based model. For Run1 it shows a slight reduction in CVRMSE and a small increase in R-squared for the spatial levels of pc4 and pc5, while for the Run5 predictions, the CVRSME and R-squared only get worse. For the other socio-economic integrations of cultural background and tenancy type no performance increases were found.

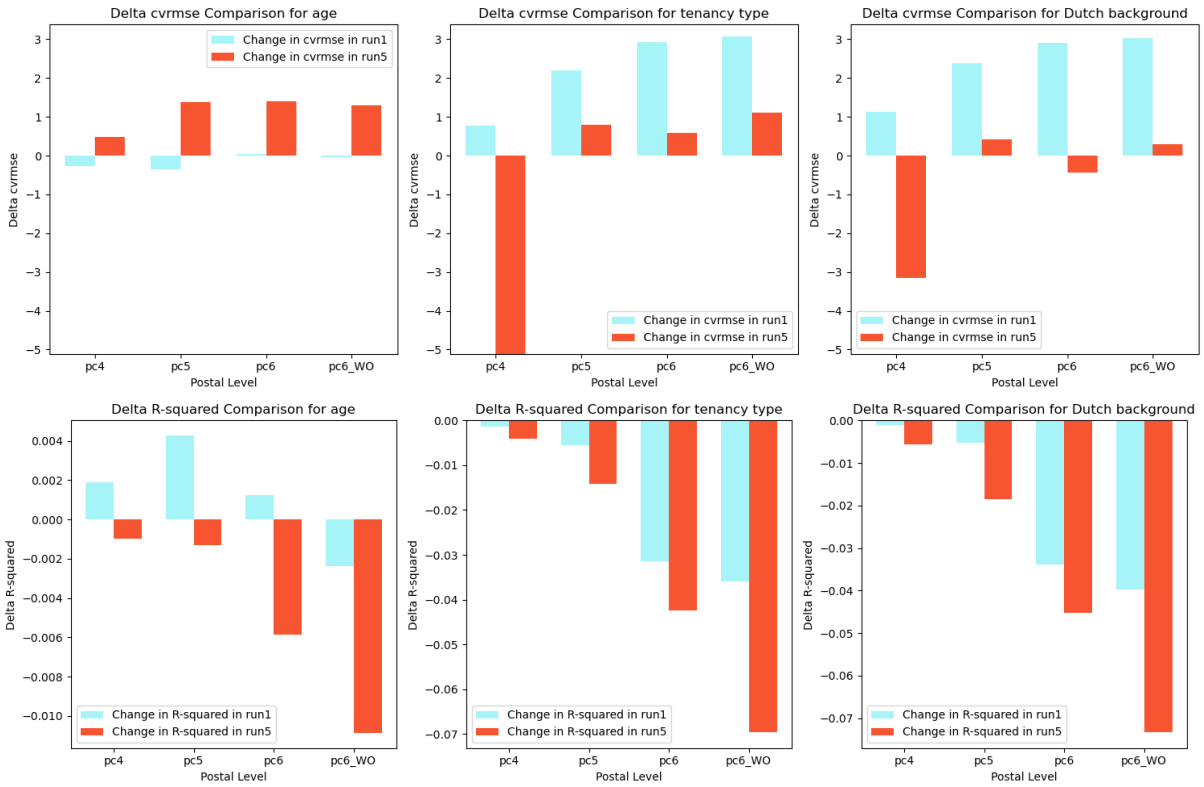


Figure 3.13 Change in performance metrics after the integration of regression coefficients of socio-economic variables age, tenancy type and Dutch background. In blue the effect on physics-based model run 1 is shown, in red on model run 5.

Performance of the model after integration of age-based occupancy schedule and elasticity

The integration of price elasticity and age-based occupancy schedules show mixed results. These results are visualized in Figure 3.14. Overall, the R-squared decreases across all model iterations of elasticity and occupancy, which means that the model does not become better at showing the true distribution of gas demand through the urban area of Arnhem-Nijmegen.

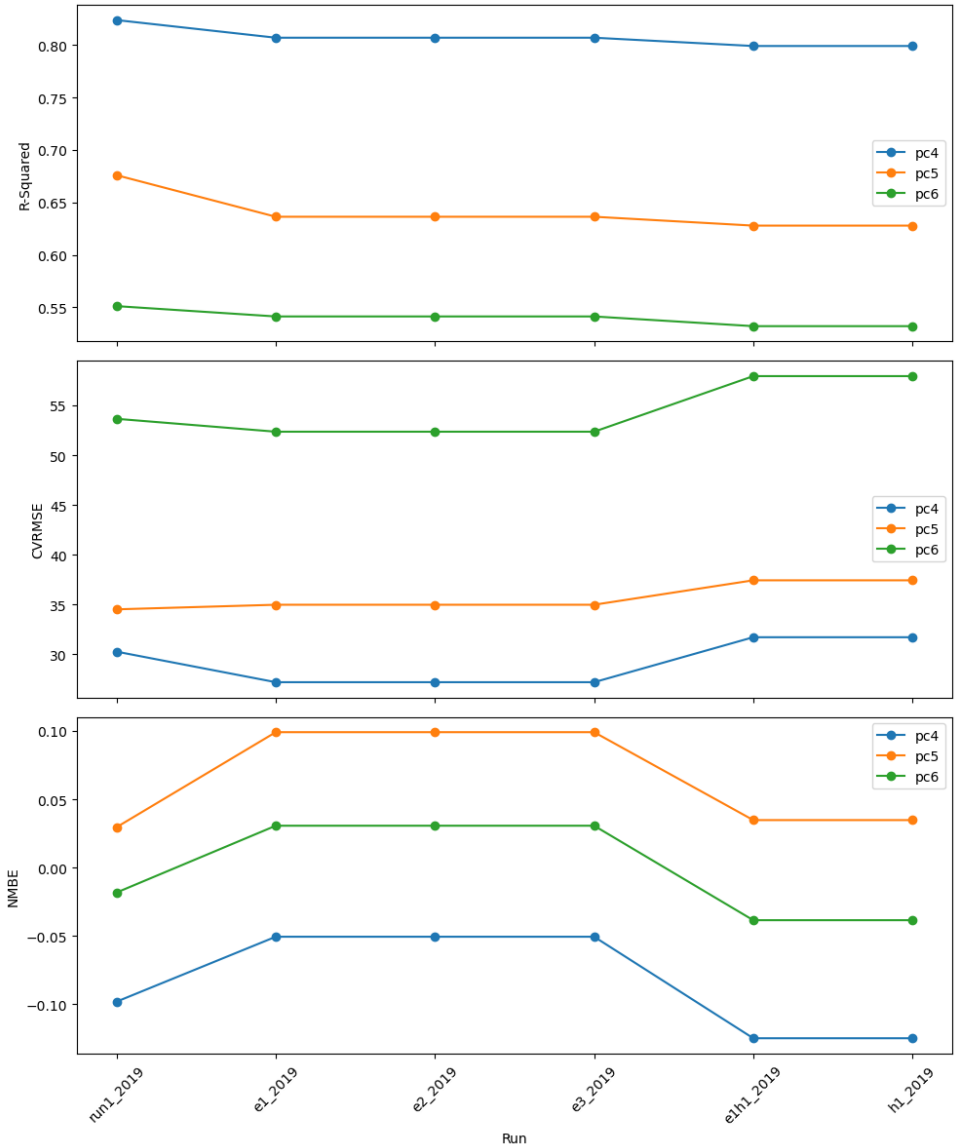


Figure 3.14 Performance metrics of the elasticity and occupancy integrated models. run1_2019 = physics-based model, e1 = elasticity base year n-1, e2 = elasticity base year 2015, e3= elasticity base year 2010, e1h1_ = elastic base year n-1, occupancy, h1 = occupancy

However, the incorporation of the elasticity in iteration E1, E2 and E3 does reduce the NMBE by 0.05 on the pc4 level. Also, the CVRMSE shrank by 5% on the pc4 level and 2% on the pc6 level. For the pc5 level, the CVRMSE stayed about the same.

The incorporation of the age-based occupancy schedule, which is model iteration H1, does not show any signs of improving the model. Consequently, the combined model of elasticity and age-based occupancy schedule (E1H1) also does not show any performance improvements.

4 Discussion

This discussion is outlined in four main sections. In Section 4.1, the data sources and the methodology are critically reflected upon. This is followed by Section 4.2, which shows the interpretation of the aforementioned results. Section 4.3 outlines the limitations of this research, which is followed by the recommendations for further research in Section 4.4.

4.1 Limitations of methodology

In this thesis, a wide range of data sources is used to predict the space heating demand within residential buildings. Each of these datasets and their respective processing come with their own simplification of reality.

3DBAG and BAG: Susceptible to Oversimplification Errors

The 3D and 2D versions of the BAG represent extensive datasets encompassing various variables. However, the richness of these datasets also renders them susceptible to erroneous data, particularly concerning the classification of building functions.

A simplification made in this research is the exclusion of buildings that serve functions beyond residential purposes. Only buildings designated solely for residential use are included in the input dataset. This exclusion criterion also includes buildings primarily intended for residential purposes, even if they contain minor areas allocated to other functions. However, this cannot be checked because the BAG only mentions which functions are present within a building and it does not mention the fraction of conditional floor area that is used by the respective functions.

In addition, within the building stock, there are also buildings that are saved as only having a residential function. However, upon further inspection, it becomes very unlikely that this is the only function present. It is difficult to assess this phenomenon for every building in the dataset, but it can cause the model to overestimate the energy demand.

Another error that is found in some buildings in the dataset is missing or incorrect data on the ground floor area. This was already mentioned by León-Sánchez et al. (2021). In their paper, they fixed this problem by geometrically modifying the 3D objects. Because in this thesis, a 2D version with 3D attributes was used, this was not possible. Therefore, errors in ground floor area, can lead to an underestimation of energy demand in certain buildings, because the calculation of ventilation loss and HVAC energy usage depend on this variable in some model configurations.

Archetype segmentation: Limited categories lead to incorrect predictions

The first set of assumptions that possibly effect the results of this research, regard the process of segmenting the building stock according to the TABULA/episcope building typology. Within this classification framework, buildings are categorized into one of three distinct building classes and five construction periods. Various parameters, such as window size, insulation quality, and heating system efficiency are assigned based on this classification. Yang et al (2020) mentioned in their paper, that assigning all these assumptions to groups of buildings, removes part of the variation in the building stock, resulting in incorrect predictions of space heating demand.

A specific problem, also mentioned by Yang et al (2020) is the assumption of TABULA that all buildings use a HR gas boiler as a heating supply system. While, for most houses this is still the case, it becomes more and more present that buildings have alternative heat sources like geothermal or heat pumps. According to CBS (2023c), at the end of 2022, 1.5 million houses had an electrical heat pump installed. Especially in neighborhoods with a larger share of newly built

houses, this fact can lead to an overestimation of gas demand by the current physics-based model.

To predict which buildings have undergone refurbishment, the Dutch EPC dataset is utilized. However, the EPC dataset of the Netherlands is characterized by incompleteness and subjectivity. As highlighted by Hettinga et al. (2023), the process of obtaining an EPC label in the Netherlands incentivize households undergoing refurbishment to get an EPC survey, while those not prioritizing it may postpone this until they sell their property. This selective participation leads to the fact that a lot of houses do not have an EPC yet.

In the methodology, without an EPC are presumed not to have undergone any refurbishment. However, this assumption may not always hold true. Refurbishments conducted without obtaining an EPC survey, often due to the associated high costs, can occur, leading to an underestimation of refurbishment rates (Interpolis, 2020). This can lead to an underestimation of refurbishment rates, which was also visible in Figure 3.8.

Weather data: Oversimplification and urban setting effects

The weather data utilized in this thesis originates from the official KNMI weather stations and is subsequently linked to each building in the dataset based on the smallest distance. However, this approach represents a simplification of reality, because many KNMI weather stations are situated outside urban areas, where weather patterns differ from those in urban settings.

Costanzo et al. (2019) emphasize the importance of incorporating comprehensive external factors into space-heating demand models, including urban morphology, density, and surface materials. Urban morphology significantly impacts wind flow within urban areas and the resulting cooling effect. Additionally, urban density and surface material influence the absorption of incoming solar radiation and the extent to which buildings are shaded by neighboring buildings. In dense urban cores these changes in environmental factors, can lead to the formation of Urban Heat Islands (UHIs), which tend to trap heat within the city.

While UHIs are often associated with their harmful effects during summer, they also influence winter conditions. According to Macintyre et al. (2021), the UHI effect reduces cold temperatures by approximately 1.5 degrees. This increase in winter temperature can lead to a decrease in space heating demand. The physics-based model developed in this thesis does not incorporate these effects, because it is using weather data from official stations outside of the urban area. This oversight could lead to an overestimation of energy consumption in urban cores.

Socio-economic data: Unethical categorizations of socio-economic phenomena

Socio-economic data is not available at the household level, as this would pose privacy concerns. Therefore, this thesis uses postal code data from three different spatial scales to integrate socio-economic data into a physics-based model. However, this approach introduces its own set of ethical considerations.

As highlighted by Hosseini et al. (2022), the aggregation of data into large datasets can foster potentially problematic relationships. Using such extensive datasets encourage users to establish statistically significant relationships between single socio-economic variables and dependent variables, such as space heating demand in this study. However, within more critically oriented social sciences, it is emphasized that phenomena seldom stem from isolated variables, but rather stem from a complex interplay of factors, regarding individual behavior, economic incentives, social norms and other influences(Frederiks et al., 2015). A great example of a social science approach that contradicts the approach taken in this thesis, is the Actor-Network Theory

paradigm. Within this paradigm, every living organism is seen as actors that take actions, irrespective of large-scale socio-economic trends (Inglis & Thorpe, 2019). Focusing on these actions first, instead of labelling people with large-scale socio-economic variables, could provide less biased findings.

Especially in the case of demographic data, these simplistic statistical relationships can possibly lead to discriminatory consequences (Hosseini et al., 2022). For instance, when a relationship is found between cultural background and space heating demand, it risks oversimplifying, suggesting that cultural background alone dictates heating needs. In reality, there is a large variety of factors that interact with each other: cultural background could possibly relate to differences in housing conditions, household sizes, working patterns, economic situations etc. It's therefore dangerous to assume these single relationships.

Another ethical question arises from the method of assigning aggregated socio-economic data to individual households within postcodes, employing a weighted mean for each variable. This procedure takes away the local variation that households. For instance, neighboring households might differ significantly in demographics, leading to diverse occupancy schedules. However, within this model, they are treated as identical entities, because of their location within the same postcode. This oversimplification could render policies derived from such models ineffective, as they fail to account for the nuanced realities of individual actors.

Validation: Dealing with differences of scale

The validation technique employed in this thesis comes with an important consideration. In this validation, all the predictions are summed per postal code. To address the issue of missing houses in the model predictions, the summed prediction is multiplied by the fraction of missing houses. However, the dwellings that are not included in the model are primarily dwellings that tend to be located in multi-purpose buildings.

Dwellings situated within buildings with mixed functions typically have smaller sizes or less exterior surface area. Consequently, the demand for energy in these types of housing units is lower than the demand in homes that solely have a residential function. By not accounting for this difference, the multiplying-by-fraction approach can lead the aligned prediction to be overestimating the energy demand of postal codes where larger varieties of building functions occur.

4.2 Interpretation of results

4.2.1 Lightweight physics-based model

Spatial findings

While the accuracy of the building segmentation along the archetype rules is not available, it can be concluded that the majority of buildings were classified correctly. Nevertheless, instances occurred where similar buildings were classified as different building types. Notably, distinguishing houses as either a Terraced End-house or a Terraced Middle Row House posed challenges, as outlined in Figure 3.5. Such inaccuracies can result in misassigned Window-to-Wall Ratios. Consequently, buildings that are classified as middle-row when they are actually end houses may inaccurately show a larger proportion of window area than is actually present. This discrepancy arises because the external wall on the free-standing side typically features fewer windows than the front and back of the building, leading to an overestimation of gas demand.

The model's predictions indicate that areas characterized by larger, older houses tend to exhibit higher gas demands compared to neighborhoods dominated by terraced and newer houses. These findings align with contextual expectations: terraced houses offer less surface area for heated air to escape through compared to single-family homes and newer houses typically have better insulation, thereby requiring less gas to main a comfortable indoor temperature.

Performance

Though the results seem to make sense contextually, it is important to check if the actual predictions make sense themselves. The results of different model iterations show varying amounts of performance. The iterations that contain specific interesting insights are Run1, Run2, Run4 and Run5.

Run 1 and 2 have the lowest CVRMSE and NMBE scores, suggesting the predictions are closest to the true values. Run1 and 2 differ in one configuration: Run 1 employs ground floor area as the reference for domestic hot water usage, whereas Run 2 utilizes conditional floor space, resulting in a doubling or even tripling of the domestic hot water prediction. This change significantly impacts the NMBE values, transitioning from a 2-10% underestimation to a 3-11% overestimation of gas demand. This is also where the model is lacking information. The model assumes an average demand per square meter for domestic hot water, which is based on the paper by Flourentzou and Pereira (2021). However, while space heating is highly dependent on a house itself, domestic hot water usage is highly dependent on the number of people that live in a house. This information was not incorporated into this model, leading to an increase in bias.

Run 4 and 5 stand out with the highest R^2 metrics. This can be attributed to their more accurate integration of building shape compared to the first two iterations. Unlike Run 1 and 2, which approximate volume by multiplying ground floor area by floor count and a standard floor height, Run 4 and 5 utilize the Level of Detail (LOD) 2.2 Volume column, accounting for the building's actual shape with angled roofs and walls. Consequently, Run 4 and Run 5 seem to perform better at showing the relative differences between different postcodes across spatial levels.

However, despite the improved portrayal of these relative differences between postcodes, Run 4 and 5 demonstrate larger bias. While the relationship among neighborhoods is better shown by Run 4 and 5, the models have a larger bias. These iterations consistently overestimate space heating demand across all neighborhoods by 50-60%. The discrepancy arises from two possible reasons. Again, the overestimation could partly be caused by the fact that the validation technique is missing some of the variation in housing size. Additionally, the inadequacy of the ventilation loss formula in Appendix B.3, which employs a constant air change rate rather than considering variables such as building quality, age, and inhabitant behavior (Sokol et al., 2017). Therefore, simply replacing the reference room height with actual volume in the formula fails to accurately calculate ventilation loss values.(Frederiks et al., 2015; Todeschi et al., 2021).

Main takeaway

Overall, it can be said that the development of a lightweight physics-based model has proven to be a breakthrough. While its predictions may not rival those of more sophisticated models in accuracy, the model efficiently calculates space heating demand across entire urban landscapes in a short amount of time. This capability allows for the identification of larger spatial patterns in energy consumption, providing invaluable insights that complement the nuances captured by more complex models. Such a model is particularly useful in exploratory research projects where the aim is to assess its feasibility and value. This applied also to this thesis, where

the lightweight model facilitated assessing how the model responds to the integration of socio-economic variables over a larger geographical area.

4.2.2 Socio-economic integration: benefits and challenges

Spatial findings

The spatial effects of the integration in this thesis show only a very low change in space heating demand predictions. What they do show, however, is that spatially the changes tend to be visible quite clearly, when variables are integrated in a separate manner.

A benefit of this integration is the fact that socio-economic variables can help correct biases inherent in the physics-based model. Specifically, it enables the identification of space heating consumption patterns that might otherwise go unnoticed. For policymaking this holds crucial importance, particularly in fields such as energy network expansion and transition, especially for the field of energy network extensification and energy transition. Spatial findings indicate that incorporating socio-economic variables could prevent the oversight of certain areas when formulating policies aimed at upgrading the energy network.

However, it is important to acknowledge the downsides as well. Integrating socio-economic variables comes with its own sets of risks, including the potential introduction of biases and discriminatory consequences. This downside is also visible in the spatial findings of the integration, as with the integration of single variables, certain neighborhoods show opposing trends, while in reality these opposing trends are not as straightforward as they appear now. This downside is also visible when looking at the performance metrics of the integration.

Performance

The findings from the integration analysis emphasize the persistent challenge of effectively incorporating socio-economic variables. Despite efforts to integrate them, the impact on space heating demand tends to be minimal. This can be attributed to the methodological approach wherein a weighted mean is calculated for each postal code area, leading to a diffusion of the variables' effects and complicating their interpretation.

Consequently, the integration does not significantly enhance the model's predictive performance. While certain model integrations may yield slight improvements, the extent of these enhancements is often marginal. Moreover, there is a legitimate question as to whether these improvements are a direct effect of the integration itself or whether they happen because of the under- or overestimating biases of the physics-based models Run 1 and 5.

Main takeaway

This thesis has contributed to understanding the influence of certain socio-economic variables on energy demand by employing a Random Forest model. The results demonstrate that the Random Forest model outperforms many existing socio-economic models identified in the literature review.

However, the thesis also sheds light on the challenges posed by limited data availability and the limitations of traditional Regression methods. Due to these constraints, integrating socio-economic variables into physics-based energy demand models did not yield positive results. Therefore, it is evident that additional measures must be undertaken to improve the efficacy of integrating socio-economic variables.

4.3 Future research

Data enrichment

This research has shown that the approach of using census data and a weighted-mean approach for integrating socio-economic variables is not effective in integrating socio-economic variables. Therefore, this study suggests several ways on overcoming this data related issue.

Conducting surveys, similar to the German “Mikrozensus” among Dutch households could provide invaluable insights into household patterns of space heating demand (Schulte & Heindl, 2017). By capturing more refined data on socio-economic factors, this approach can help getting a more nuanced understanding of the relationship between socio-economic factors and space heating demand.

Exploring smart metering data from a subset of households could offer an opportunity to uncover occupancy schedules and energy consumption patterns across diverse demographic groups. These refined schedules and patterns could then potentially be extrapolated to a larger area to assess their effectiveness in improving predictions in energy demand modeling.

Research efforts focusing on refining Dutch building typology and EPC data, hold promise in assigning more accurate building parameters and HVAC system parameters to houses. For instance, enriching building typology databases could encompass the transition from gas-based HVAC systems to other sources of energy, such as electrical heat pumps or district heating, including options like geothermal energy. This could help align the predictions of an energy demand model with real-world energy consumption trends.

Methodology recommendations

This thesis applied Random Forest, Ordinary Linear Regression (OLS) and Ridge Regression analyses to uncover patterns between socio-economic variables and space heating demand. While Random Forest is good at uncovering non-linear patterns, it cannot be used to modify physics-based predictions, as this model only offers information on feature importance, rather than direction. Regression methods were used to extract coefficients from the most important variables. However, this leads to oversimplification of relationships.

It could, therefore, be helpful to apply more elaborate statistical and survey-based models from social science like quantile regression, conditional demand models, and discrete-choice models. These more sophisticated models can possibly perform better in uncovering complex relationships between socio-economic variables and space heating demand.

By leveraging these advanced socio-economic models, the predictive performance of energy demand models can be enhanced. Moreover, employing a survey-based approach can mitigate ethical concerns raised in this thesis, as it allows for a more comprehensive understanding of energy consumption, while still respecting the principles of consent and transparency.

5 Conclusion

5.1 Research Summary

Space heating in buildings is one of the main components of household CO_2 emissions. Government initiatives aim to reduce emissions through energy-saving programs. To develop accurate policy strategies, detailed information on energy demand is vital. However, accurately assessing energy demand remains difficult. Energy demand models can help provide this information. However, in existing research a strong dichotomy exists between physics-based models, that primarily focus on the thermodynamics of buildings, and socio-economic models, that primarily focus on the people that live in the buildings. This division leads in both models to a degree of unexplained variation in model predictions. This thesis aimed at incorporating socio-economic variables into a physics-based energy demand model. The central research question that guided this investigation was:

> "What is the effect of integrating socio-economic factors into physics-based energy demand models for the purpose of predicting energy consumption in residential buildings?"

This research question was answered through a selection of methods. Initially, two comprehensive literature reviews were undertaken, focusing on physics-based and socio-economic energy demand models, respectively. Based on this literature review, a lightweight physics-based model was constructed, that could approximate energy demand for a larger area in a short amount of time. In this physics-based model, socio-economic variables were iteratively integrated through a combination of Random Forest, Regression and Elasticity analysis.

The physics-based model turned out to predict the relative differences between neighborhoods well with R-squared metrics between 0.6 and 0.9. The CVRMSE on the actual values performed a bit lower around 30%. The integration of socio-economic variables did not seem to improve the model significantly.

The fact that the integration did not improve the model could be caused by a multitude of reasons. For instance, the approach where a weighted mean per neighborhood was calculated for the coefficients might have contributed to reducing the impact of the integration. Additionally, the separate integration of socio-economic variables fails to reflect the detailed interconnections that are present in real-life scenarios. In reality, socio-economic variables are deeply intertwined, collectively shaping energy demand. Neglecting this comprehensive understanding limits the effectiveness of the model's predictions.

5.2 Answer to research question

Looking back at the aforementioned research question, the findings suggest that integrating socio-economic variables into physics-based energy demand models poses notable challenges. Firstly, the ethical dilemma surrounding privacy issues emerged as a considerable obstacle. Incorporating socio-economic factors into energy demand models requires accessing personal data, raising legitimate concerns about privacy infringement. That is why in this thesis, it was chosen to resort to postal code data on socio-economics. However, this resulted in limited available data.

This limited data availability problem posed a significant constraint in this research. Socio-economic analysis often requires fine-grained information on households. This is because

relationships between socio-economic variables and space heating demand are very comprehensive and interactive. By accessing these variables only on a higher spatial level, the relationships are diffuse and vague. This limits the effectiveness and accuracy of integrating socio-economic factors in a physics-based model.

To address this issue, several improvements can be implemented. Utilizing metering data or conducting household surveys with consent for the usage of their demand patterns can provide a more comprehensive overview, while still keeping the ethical dilemmas regarding this sensitive data in mind. This approach enables the identification of more complex interrelationships between socio-economic variables and energy demand. When combining these richer datasets with advanced socio-economic models like quantile regression, conditional demand models, and discrete-choice models, there is potential for an enhancement of the predictive accuracy of energy demand models.

In conclusion, while the integration of socio-economic variables into physics-based energy demand models holds promise for enhancing predictive accuracy, this study reveals that significant hurdles must be overcome. Simple Regression models are not sufficient in capturing the complex relationships between socio-economic variables and energy demand. Furthermore, the limited availability of fine-grained public socio-economic data complicates efforts to establish clear relationships between these variables and energy consumption.

Bibliography

Abbasabadi, N., Ashayeri, M., Azari, R., Stephens, B., & Heidarinejad, M. (2019). An integrated data-driven framework for urban energy use modeling (UEUM). *Applied Energy*, *253*, 113550. https://doi.org/10.1016/j.apenergy.2019.113550

Abdelaziz, F., Raslan, R., & Symonds, P. (2021, September 1). *Developing an archetype building stock model for new cities in Egypt*. 2021 Building Simulation Conference. https://doi.org/10.26868/25222708.2021.31009

Afaifia, M., Djiar, K. A., Bich-Ngoc, N., & Teller, J. (2021). An energy consumption model for the Algerian residential building's stock, based on a triangular approach: Geographic Information System (GIS), regression analysis and hierarchical cluster analysis. *Sustainable Cities and Society*, *74*, 103191. https://doi.org/10.1016/j.scs.2021.103191

Ali, U., Shamsi, M. H., Bohacek, M., Hoare, C., Purcell, K., Mangina, E., & O'Donnell, J. (2020). A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings. *Applied Energy*, *267*, 114861. https://doi.org/10.1016/j.apenergy.2020.114861

Bakaloglou, S., & Charlier, D. (2021). The role of individual preferences in explaining the energy performance gap. *Energy Economics*, *104*, 105611. https://doi.org/10.1016/j.eneco.2021.105611

Belaid, F., Ben Youssef, A., & Omrani, N. (2020). *Investigating the factors shaping residential energy consumption patterns in France: Evidence form quantile regression*. https://doi.org/10.25428/1824-2979/202001-127-151

Belaid, F., & Rault, C. (2021). Energy Expenditure in Egypt: Empirical Evidence Based on a Quantile Regression Approach. *Environmental Modeling & Assessment*, *26*(4), 511–528. https://doi.org/10.1007/s10666-021-09764-8

Boonekamp, P. G. M. (2007). Price elasticities, policy measures and actual developments in household energy consumption – A bottom up analysis for the Netherlands. *Energy Economics*, *29*(2), 133–157. https://doi.org/10.1016/j.eneco.2005.09.010

Brøgger, M., Bacher, P., & Wittchen, K. B. (2019). A hybrid modelling method for improving estimates of the average energy-saving potential of a building stock. *Energy and Buildings*, *199*, 287–296. https://doi.org/10.1016/j.enbuild.2019.06.054

Brounen, D., Kok, N., & Quigley, J. M. (2012). Residential energy use and conservation: Economics and demographics. *European Economic Review*, *56*(5), 931–945. https://doi.org/10.1016/j.euroecorev.2012.02.007

Buffat, R., Froemelt, A., Heeren, N., Raubal, M., & Hellweg, S. (2017). Big data GIS analysis for novel approaches in building stock modelling. *Applied Energy*, *208*, 277–290. https://doi.org/10.1016/j.apenergy.2017.10.041

Buttitta, G., & Finn, D. P. (2020). A high-temporal resolution residential building occupancy model to generate high-temporal resolution heating load profiles of occupancy-integrated archetypes. *Energy and Buildings*, *206*, 109577. https://doi.org/10.1016/j.enbuild.2019.109577

Buttitta, G., Turner, W. J. N., Neu, O., & Finn, D. P. (2019). Development of occupancy-integrated archetypes: Use of data mining clustering techniques to embed occupant behaviour profiles in archetypes. *Energy and Buildings*, *198*, 84–99. https://doi.org/10.1016/j.enbuild.2019.05.056

Cammen, H. van der, & Klerk, L. A. de. (2012). *The selfmade land: Culture and evolution of urban and regional planning in the Netherlands* (First edition). Spectrum/Uitgeverij Unieboek.

Çebi Karaaslan, K., & Algül, Y. (2023). Determinants of energy expenditures for Turkish households using quantile regression and data from an original survey in Turkey. *Environmental Science and Pollution Research*, *30*(13), 38939–38954. https://doi.org/10.1007/s11356-022-24323-8

Centraal Bureau voor de Statistiek (CBS). (2015). *Heating degree days*. https://www.cbs.nl/en-gb/news/2015/36/greenhouse-gas-emissions-down-in-a-warm-2014/heating-degree-days

Centraal Bureau voor de Statistiek (CBS). (2023a). *Kerncijfers per postcode 2022 (PC6)* [dataset]. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode

Centraal Bureau voor de Statistiek (CBS). (2023b). *Voorraad woningen; standen en mutaties vanaf 1921* [dataset]. https://www.cbs.nl/nl-nl/cijfers/detail/82235NED#

Centraal Bureau voor de Statistiek (CBS). (2023c, December 14). *Warmtepompen; aantallen, thermisch vermogen en energiestromen*. https://www.cbs.nl/nl-nl/cijfers/detail/85523NED

Centraal Bureau voor de Statistiek (CBS). (2024). *Consumentenprijzen; prijsindex 2015 = 100* [dataset]. https://opendata.cbs.nl/statline/?dl=5D403#/CBS/nl/dataset/83131NED/table

Condotta, M., & Borga, G. (2018). Urban energy performance monitoring for Smart City decision support environments. *TECHNE - Journal of Technology for Architecture and Environment*, 73-80 Pages. https://doi.org/10.13128/TECHNE-22688

Costanzo, M., Archer, D., Aronson, E., & Pettigrew, T. (1986). Energy conservation behavior: The difficult path from information to action. *American Psychologist*, *41*(5), 521–528. https://doi.org/10.1037/0003-066X.41.5.521

Costanzo, V., Yao, R., Li, X., Liu, M., & Li, B. (2019). A multi-layer approach for estimating the energy use intensity on an urban scale. *Cities*, *95*, 102467. https://doi.org/10.1016/j.cities.2019.102467

Dahlström, L., Broström, T., & Widén, J. (2022). Advancing urban building energy modelling through new model components and applications: A review. *Energy and Buildings*, *266*, 112099. https://doi.org/10.1016/j.enbuild.2022.112099

De Bruin, J., Van de Schoot, R., Ma, Y., Oberski, D., Tummers, L., Bagheri, A., Kramer, B., De Boer, J., Huijts, M., & Mödinger, C. (2023). *ASReview* (1.2.1) [Computer software]. Utrecht University. https://asreview.nl/

Dubois, G., Sovacool, B., Aall, C., Nilsson, M., Barbier, C., Herrmann, A., Bruyère, S., Andersson, C., Skold, B., Nadaud, F., Dorner, F., Moberg, K. R., Ceron, J. P., Fischer, H., Amelung, D., Baltruszewicz, M., Fischer, J., Benevise, F., Louis, V. R., & Sauerborn, R. (2019). It starts at home? Climate policies targeting household consumption and behavioral decisions are key to low-carbon futures. *Energy Research & Social Science*, *52*, 144–158. https://doi.org/10.1016/j.erss.2019.02.001

Duminil, E., Brassel, K., Nouvel, R., Benoit, A., Bruse, M., Betz, M., Alam, N., Dastageeri, H., Wate, P., Debue, P., Köhler, S., Weiler, V., Monsalvete, P., Zirak, M., Bao, K., Schneider, S., & Coors, V. (2018). *SimStadt* (0.9) [Computer software]. Hochschule für Technik Stuttgart. https://simstadt.hft-stuttgart.de/

Edtmayer, H., Fochler, L.-M., Mach, T., Fauster, J., Schwab, E., & Hochenauer, C. (2023). High-resolution, spatial thermal energy demand analysis and workflow for a city district. *International Journal of Sustainable Energy Planning and Management*, *38*, 47–64. https://doi.org/10.54337/ijsepm.7570

Eggimann, S., Hall, J. W., & Eyre, N. (2019). A high-resolution spatio-temporal energy demand simulation to explore the potential of heating demand side management with large-scale heat pump diffusion. *Applied Energy*, *236*, 997–1010. https://doi.org/10.1016/j.apenergy.2018.12.052

Flourentzou, F., & Pereira, J. (2021). Domestic hot water optimizing potential in existing or renovated multifamily residential buildings. *Journal of Physics: Conference Series*, *2042*(1), 012144. https://doi.org/10.1088/1742-6596/2042/1/012144

Fonseca, J. A., & Schlueter, A. (2015). Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. *Applied Energy*, *142*, 247–265. https://doi.org/10.1016/j.apenergy.2014.12.068

Fonseca, J. N. B., & Oliveira Panão, M. J. N. (2017). Monte Carlo housing stock model to predict the energy performance indicators. *Energy and Buildings*, *152*, 503–515. https://doi.org/10.1016/j.enbuild.2017.07.059

Frederiks, E., Stenner, K., & Hobman, E. (2015). The Socio-Demographic and Psychological Predictors of Residential Energy Consumption: A Comprehensive Review. *Energies*, *8*(1), 573–609. https://doi.org/10.3390/en8010573

Gassar, A., Yun, G. Y., & Kim, S. (2019). Data-driven approach to prediction of residential energy consumption at urban scales in London. *Energy*, *187*, 115973. https://doi.org/10.1016/j.energy.2019.115973

Gemeente Amsterdam. (n.d.). *Volg het beleid: Duurzaamheid*. Gemeente Amsterdam. Retrieved October 3, 2023, from https://www.amsterdam.nl/bestuur-organisatie/volg-beleid/duurzaamheid/

Geraldi, M. S., & Ghisi, E. (2022). Data-driven framework towards realistic bottom-up energy benchmarking using an Artificial Neural Network. *Applied Energy*, *306*, 117960. https://doi.org/10.1016/j.apenergy.2021.117960

Ghedamsi, R., Settou, N., Gouareh, A., Khamouli, A., Saifi, N., Recioui, B., & Dokkar, B. (2016). Modeling and forecasting energy consumption for residential buildings in Algeria using bottom-up approach. *Energy and Buildings*, *121*, 309–317. https://doi.org/10.1016/j.enbuild.2015.12.030

Google. (2023). *Google Street View* [Map]. https://www.google.com/maps/@51.8353882,5.8117595,3a,75y,143.54h,90t/data=!3m7!1e1!3m5!1sNlWw_dlwA65my8HaXXBGag!2e0!6shttps:%2F%2Fstreetviewpixels-pa.googleapis.com%2Fv1%2Fthumbnail%3Fpanoid%3DNlWw_dlwA65my8HaXXBGag%26cb_client%3Dsearch.gws-prod.gps%26w%3D86%26h%3D86%26yaw%3D143.53954%26pitch%3D0%26thumbfov%3D100!7i16384!8i8192?entry=ttu

Gulotta, T. M., Cellura, M., Guarino, F., & Longo, S. (2021). A bottom-up harmonized energy-environmental models for europe (BOHEEME): A case study on the thermal insulation of the EU-28 building stock. *Energy and Buildings*, *231*, 110584. https://doi.org/10.1016/j.enbuild.2020.110584

Haas, R., & Schipper, L. (1998). Residential energy demand in OECD-countries and the role of irreversible efficiency improvements. *Energy Economics*, *20*(4), 421–442. https://doi.org/10.1016/S0140-9883(98)00003-6

Han, X., Poblete-Cazenave, M., Pelz, S., & Pachauri, S. (2022). Household energy service and home appliance choices in urban China. *Energy for Sustainable Development*, *71*, 263–278. https://doi.org/10.1016/j.esd.2022.09.021

Harold, J., Lyons, S., & Cullinan, J. (2018). Heterogeneity and persistence in the effect of demand side management stimuli on residential gas consumption. *Energy Economics*, *73*, 135–145. https://doi.org/10.1016/j.eneco.2018.04.034

Hedegaard, R. E., Kristensen, M. H., Pedersen, T. H., Brun, A., & Petersen, S. (2019). Bottom-up modelling methodology for urban-scale analysis of residential space heating demand response. *Applied Energy*, *242*, 181–204. https://doi.org/10.1016/j.apenergy.2019.03.063

Hettinga, S., Van 'T Veer, R., & Boter, J. (2023). Large scale energy labelling with models: The EU TABULA model versus machine learning with open data. *Energy*, *264*, 126175. https://doi.org/10.1016/j.energy.2022.126175

Hosseini, M., Wieczorek, M., & Gordijn, B. (2022). Ethical Issues in Social Science Research Employing Big Data. *Science and Engineering Ethics*, *28*(3), 29. https://doi.org/10.1007/s11948-022-00380-7

Hu, S., Yan, D., Cui, Y., & Guo, S. (2016). Urban residential heating in hot summer and cold winter zones of China—Status, modeling, and scenarios to 2030. *Energy Policy*, *92*, 158–170. https://doi.org/10.1016/j.enpol.2016.01.032

Hunt, L. C., & Ryan, D. L. (2015). Economic modelling of energy services: Rectifying misspecified energy demand functions. *Energy Economics*, *50*, 273–285. https://doi.org/10.1016/j.eneco.2015.05.006

Inglis, D., & Thorpe, C. (2019). *An invitation to social theory* (Second edition). Polity Press.

International Organization for Standardizaition (ISO). (2017). *Energy performance of buildings—Energy needs for heating and cooling, internal temperatures and sensible and latent heat loads—Part 1: Calculation procedures* (ISO 52016-1:2017). https://www.iso.org/obp/ui/#iso:std:iso:52016:-1:ed-1:v1:en:ref:64

Interpolis. (2020, December 16). *Energielabels voor woningen: Vanaf 2021 nauwkeuriger en duurder*. https://www.interpolis.nl/magazine/wonen/energielabels-voor-woningen-vanaf-2021-nauwkeuriger-en-duurder

Jia, J.-J., Ni, J., & Wei, C. (2023). Residential responses to service-specific electricity demand: Case of China. *China Economic Review*, *78*, 101917. https://doi.org/10.1016/j.chieco.2023.101917

Kadaster. (2023). *Basisregistratie Adressen en Gebouwen* [dataset]. PDOK. https://service.pdok.nl/lv/bag/wfs/v2_0?request=getCapabilities&service=WFS

Kämpf, J., & Oguey, J. (2023). *CitySim Pro* [Computer software]. kaemco. http://www.kaemco.ch/download.php

Kazas, G., Fabrizio, E., & Perino, M. (2017). Energy demand profile generation with detailed time resolution at an urban district scale: A reference building approach and case study. *Applied Energy*, *193*, 243–262. https://doi.org/10.1016/j.apenergy.2017.01.095

Kloosterman, R., Akkermans, M., Reep, C., Wingen, M., Molnár, H., & Beuningen van, J. (2021, June 4). *Klimaatverandering en energietransitie: Opvattingen en gedrag van Nederlanders in 2020*. https://www.cbs.nl/nl-nl/longread/rapportages/2021/klimaatverandering-en-energietransitie-opvattingen-en-gedrag-van-nederlanders-in-2020/4-duurzaam-wonen

Komlos, J. (1990). Nutrition, Population Growth, and the Industrial Revolution in England. *Social Science History*, *14*(1), 69. https://doi.org/10.2307/1171364

Koninklijk Nederlands Meteorologisch Instituut (KNMI). (2024a). *Vochtigheid_en_temperatuur* [dataset]. https://dataplatform.knmi.nl/dataset/zonneschijnduur-en-straling-1-0

Koninklijk Nederlands Meteorologisch Instituut (KNMI). (2024b). *Zonneschijnduur_en_straling* [dataset]. https://dataplatform.knmi.nl/dataset/zonneschijnduur-en-straling-1-0

Lawal, A. S., Servadio, J. L., Davis, T., Ramaswami, A., Botchwey, N., & Russell, A. G. (2021). Orthogonalization and machine learning methods for residential energy estimation with social and economic indicators. *Applied Energy*, *283*, 116114. https://doi.org/10.1016/j.apenergy.2020.116114

Ledesma, G., Pons-Valladares, O., & Nikolic, J. (2021). Real-reference buildings for urban energy modelling: A multistage validation and diversification approach. *Building and Environment*, *203*, 108058. https://doi.org/10.1016/j.buildenv.2021.108058

León-Sánchez, C., Giannelli, D., Agugiaro, G., & Stoter, J. (2021). TESTING THE NEW 3D BAG DATASET FOR ENERGY DEMAND ESTIMATION OF RESIDENTIAL BUILDINGS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLVI-4/W1-2021*, 69–76. https://doi.org/10.5194/isprs-archives-XLVI-4-W1-2021-69-2021

Li, X., & Yao, R. (2021). Modelling heating and cooling energy demand for building stock using a hybrid approach. *Energy and Buildings*, *235*, 110740. https://doi.org/10.1016/j.enbuild.2021.110740

Li, X., Yao, R., Liu, M., Costanzo, V., Yu, W., Wang, W., Short, A., & Li, B. (2018). Developing urban residential reference buildings using clustering analysis of satellite images. *Energy and Buildings*, *169*, 417–429. https://doi.org/10.1016/j.enbuild.2018.03.064

Lin, B., & Liu, K. (2017). Energy Substitution Effect on China's Heavy Industry: Perspectives of a Translog Production Function and Ridge Regression. *Sustainability*, *9*(11), 1892. https://doi.org/10.3390/su9111892

Macintyre, Helen. L., Heaviside, C., Cai, X., & Phalkey, R. (2021). The winter urban heat island: Impacts on cold-related mortality in a highly urbanized European region for present and future climate. *Environment International*, *154*, 106530. https://doi.org/10.1016/j.envint.2021.106530

Mastrucci, A., Marvuglia, A., Leopold, U., & Benetto, E. (2017). Life Cycle Assessment of building stocks from urban to transnational scales: A review. *Renewable and Sustainable Energy Reviews*, *74*, 316–332. https://doi.org/10.1016/j.rser.2017.02.060

Mastrucci, A., Pérez-López, P., Benetto, E., Leopold, U., & Blanc, I. (2017). Global sensitivity analysis as a support for the generation of simplified building stock energy models. *Energy and Buildings*, *149*, 368–383. https://doi.org/10.1016/j.enbuild.2017.05.022

Mata, É., Kalagasidis, A. S., & Johnsson, F. (2013). A modelling strategy for energy, carbon, and cost assessments of building stocks. *Energy and Buildings*, *56*, 100–108. https://doi.org/10.1016/j.enbuild.2012.09.037

Mata, É., Wanemark, J., Cheng, S. H., Ó Broin, E., Hennlock, M., & Sandvall, A. (2021). Systematic map of determinants of buildings' energy demand and CO_2 emissions shows need for decoupling. *Environmental Research Letters*, *16*(5), 055011. https://doi.org/10.1088/1748-9326/abe5d7

Meier, H., & Rehdanz, K. (2010). Determinants of residential space heating expenditures in Great Britain. *Energy Economics*, *32*(5), 949–959. https://doi.org/10.1016/j.eneco.2009.11.008

Mininsterie van Binnenlandse Zaken en Koninkrijksrelaties. (n.d.). *Nationaal Isolatieprogramma*. Volkhuisvesting Nederland. Retrieved October 9, 2023, from https://www.volkshuisvestingnederland.nl/onderwerpen/nationaal-isolatieprogramma

Ministerie van Economische Zaken en Klimaat. (2023). *Rijksoverheid zet in openergiebesparing*. Rijksoverheid. https://www.rijksoverheid.nl/onderwerpen/duurzame-energie/rijksoverheid-stimuleert-energiebesparing

Mitra, D., Steinmetz, N., Chu, Y., & Cetin, K. (2020). Activity Profiles of Occupants in Residential Buildings Using the American Time Use Survey Data. *Construction Research Congress 2020*, 1067–1076. https://doi.org/10.1061/9780784482865.113

Nageler, P., Koch, A., Mauthner, F., Leusbrock, I., Mach, T., Hochenauer, C., & Heimrath, R. (2018). Comparison of dynamic urban building energy models (UBEM): Sigmoid energy signature and physical modelling approach. *Energy and Buildings*, *179*, 333–343. https://doi.org/10.1016/j.enbuild.2018.09.034

Nieboer, N., & Filippidou. (2017). *TABULA National Building Typology NL* [dataset]. https://webtool.building-typology.eu/#bm

Oliveira Panão, M. J. N., & Brito, M. C. (2018). Modelling aggregate hourly electricity consumption based on bottom-up building stock. *Energy and Buildings*, *170*, 170–182. https://doi.org/10.1016/j.enbuild.2018.04.010

Österbring, M., Mata, É., Thuvander, L., Mangold, M., Johnsson, F., & Wallbaum, H. (2016). A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model. *Energy and Buildings*, *120*, 78–84. https://doi.org/10.1016/j.enbuild.2016.03.060

Palacios-Garcia, E. J., Moreno-Munoz, A., Santiago, I., Flores-Arias, J. M., Bellido-Outeirino, F. J., & Moreno-Garcia, I. M. (2018). A stochastic modelling and simulation approach to heating and cooling electricity consumption in the residential sector. *Energy*, *144*, 1080–1091. https://doi.org/10.1016/j.energy.2017.12.082

Perera, A. T. D., Coccolo, S., Scartezzini, J.-L., & Mauree, D. (2018). Quantifying the impact of urban climate by extending the boundaries of urban energy system modeling. *Applied Energy*, *222*, 847–860. https://doi.org/10.1016/j.apenergy.2018.04.004

Perwez, U., Yamaguchi, Y., Ma, T., Dai, Y., & Shimoda, Y. (2022). Multi-scale GIS-synthetic hybrid approach for the development of commercial building stock energy model. *Applied Energy*, *323*, 119536. https://doi.org/10.1016/j.apenergy.2022.119536

Peters, R., Dukai, B., Vitalis, S., Van Liempt, J., & Stoter, J. (2022). Automated 3D Reconstruction of LoD2 and LoD1 Models for All 10 Million Buildings of the Netherlands. *Photogrammetric Engineering & Remote Sensing*, *88*(3), 165–170. https://doi.org/10.14358/PERS.21-00032R2

Prataviera, E., Romano, P., Carnieletto, L., Pirotti, F., Vivian, J., & Zarrella, A. (2021). EUReCA: An open-source urban building energy modelling tool for the efficient evaluation of cities energy demand. *Renewable Energy*, *173*, 544–560. https://doi.org/10.1016/j.renene.2021.03.144

Rehdanz, K. (2007). Determinants of residential space heating expenditures in Germany. *Energy Economics*, *29*(2), 167–182. https://doi.org/10.1016/j.eneco.2006.04.002

Ren, Z., Paevere, P., & McNamara, C. (2012). A local-community-level, physically-based model of end-use energy consumption by Australian housing stock. *Energy Policy*, *49*, 586–596. https://doi.org/10.1016/j.enpol.2012.06.065

Richardson, J., & Burdett-Gardiner, R. (2023, August 5). *What does a heat recovery system do?* The Renewable Energy Hub. https://www.renewableenergyhub.co.uk/main/heat-recovery-systems-information/how-do-heat-recovery-and-ventilation-systems-work

Rijksdienst voor Ondernemend Nederland (RVO). (2022). *GIS bestand Energielabels juli 2022* (f6799f46-7865-45d4-881a-20b7f89b734d) [dataset]. https://nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/8dff9ab0-dc82-4143-866f-2c08450abf61?tab=relations

Schmitz, H., & Madlener, R. (2020). Heterogeneity in price responsiveness for residential space heating in Germany. *Empirical Economics*, *59*(5), 2255–2281. https://doi.org/10.1007/s00181-019-01760-y

Schulte, I., & Heindl, P. (2017). Price and income elasticities of residential energy demand in Germany. *Energy Policy*, *102*, 512–528. https://doi.org/10.1016/j.enpol.2016.12.055

Schwanebeck, M., Krüger, M., & Duttmann, R. (2021). Improving GIS-Based Heat Demand Modelling and Mapping for Residential Buildings with Census Data Sets at Regional and Sub-Regional Scales. *Energies*, *14*(4), 1029. https://doi.org/10.3390/en14041029

Schwartz, Y., Godoy-Shimizu, D., Korolija, I., Dong, J., Hong, S. M., Mavrogianni, A., & Mumovic, D. (2021). Developing a Data-driven school building stock energy and indoor environmental quality modelling method. *Energy and Buildings*, *249*, 111249. https://doi.org/10.1016/j.enbuild.2021.111249

Silva, M. C., Horta, I. M., Leal, V., & Oliveira, V. (2017). A spatially-explicit methodological framework based on neural networks to assess the effect of urban form on energy demand. *Applied Energy*, *202*, 386–398. https://doi.org/10.1016/j.apenergy.2017.05.113

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Sokol, J., Cerezo Davila, C., & Reinhart, C. F. (2017). Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. *Energy and Buildings*, *134*, 11–24. https://doi.org/10.1016/j.enbuild.2016.10.050

Streicher, K. N., Padey, P., Parra, D., Bürer, M. C., Schneider, S., & Patel, M. K. (2019). Analysis of space heating demand in the Swiss residential building stock: Element-based bottom-up model of archetype buildings. *Energy and Buildings*, *184*, 300–322. https://doi.org/10.1016/j.enbuild.2018.12.011

Theile, P., Kesnar, C., Czock, B. H., Moritz, M., Novirdoust, A. A., Coors, V., Wagner, J., & Schröter, B. (2022). There's no place like home – The impact of residential heterogeneity on bottom-up energy system modeling. *Energy and Buildings*, *254*, 111591. https://doi.org/10.1016/j.enbuild.2021.111591

Todeschi, V., Boghetti, R., Kämpf, J. H., & Mutani, G. (2021). Evaluation of Urban-Scale Building Energy-Use Models and Tools—Application for the City of Fribourg, Switzerland. *Sustainability*, *13*(4), 1595. https://doi.org/10.3390/su13041595

Torabi Moghadam, S., Toniolo, J., Mutani, G., & Lombardi, P. (2018). A GIS-statistical approach for assessing built environment energy use at urban scale. *Sustainable Cities and Society*, *37*, 70–84. https://doi.org/10.1016/j.scs.2017.10.002

Tuominen, P., Holopainen, R., Eskola, L., Jokisalo, J., & Airaksinen, M. (2014). Calculation method and tool for assessing energy consumption in the building stock. *Building and Environment*, *75*, 153–160. https://doi.org/10.1016/j.buildenv.2014.02.001

University Libraries, University of Maryland. (2023, September 15). *Systematic Review*. https://lib.guides.umd.edu/SR/steps

Utrecht University. (2022, June 3). *ASReview LAB explained*. YouTube. https://www.youtube.com/watch?v=k-a2SCq-LtA

Van Den Brom, P., Hansen, A. R., Gram-Hanssen, K., Meijer, A., & Visscher, H. (2019). Variances in residential heating consumption – Importance of building characteristics and occupants analysed by movers and stayers. *Applied Energy*, *250*, 713–728. https://doi.org/10.1016/j.apenergy.2019.05.078

Veljkovic, A., Pohoryles, D. A., & Bournas, D. A. (2023). Heating energy demand estimation of the EU building stock: Combining building physics and artificial neural networks. *Energy and Buildings*, *298*, 113474. https://doi.org/10.1016/j.enbuild.2023.113474

Verellen, E., & Allacker, K. (2022). Developing a Building Stock Model to Enable Clustered Renovation—The City of Leuven as Case Study. *Sustainability*, *14*(10), 5769. https://doi.org/10.3390/su14105769

Wang, C.-K., Tindemans, S., Miller, C., Agugiaro, G., & Stoter, J. (2020). Bayesian calibration at the urban scale: A case study on a large residential heating demand application in

Amsterdam. *Journal of Building Performance Simulation*, *13*(3), 347–361. https://doi.org/10.1080/19401493.2020.1729862

Wiesmann, D., Lima Azevedo, I., Ferrão, P., & Fernández, J. E. (2011). Residential electricity consumption in Portugal: Findings from top-down and bottom-up models. *Energy Policy*, *39*(5), 2772–2779. https://doi.org/10.1016/j.enpol.2011.02.047

Woestijne, van de, W. J. (1933). *De Gids. Jaargang 97*. Digitale bibliotheek voor de Nederlandse letteren (DBNL). https://www.dbnl.org/tekst/_gid001193301_01/_gid001193301_01_0042.php

Yang, X., Hu, M., Heeren, N., Zhang, C., Verhagen, T., Tukker, A., & Steubing, B. (2020). A combined GIS-archetype approach to model residential space heating energy: A case study for the Netherlands including validation. *Applied Energy*, *280*, 115953. https://doi.org/10.1016/j.apenergy.2020.115953

Zhang, W., Robinson, C., Guhathakurta, S., Garikapati, V. M., Dilkina, B., Brown, M. A., & Pendyala, R. M. (2018). Estimating residential energy consumption in metropolitan areas: A microsimulation approach. *Energy*, *155*, 162–173. https://doi.org/10.1016/j.energy.2018.04.161

Appendix

Appendix A. U-values (in W / m^2K) of a building without retrofitting assumed, extracted from the TABULA dataset.

Building period	**Building class**	**Window**	**Roof**	**Wall**	**Floor**
Before 1965	SFH	2.9	1.54	1.61	1.72
	TH	2.9	2.08	2.22	2.44
	MFH	2.9	1.54	1.61	1.72
	AB	2.9	1.54	1.61	1.14
1965 – 1975	SFH	2.9	0.89	1.45	2.33
	TH	2.9	0.89	1.45	2.33
	MFH	2.9	0.89	1.45	2.33
	AB	2.9	0.89	1.45	1.37
1975-1990	SFH	2.9	0.64	0.64	0.64
	TH	2.9	0.64	0.64	1.28
	MFH	2.9	0.64	0.64	1.28
	AB	2.9	0.64	0.64	0.54
1990-2005	SFH	1.8	0.36	0.36	0.36
	TH	1.8	0.36	0.36	0.36
	MFH	1.8	0.36	0.36	0.36
	AB	1.8	0.36	0.36	0.32
2005-2014	SFH	1.8	0.27	0.27	0.27
	TH	1.8	0.27	0.27	0.27
	MFH	1.8	0.27	0.27	0.27
	AB	1.8	0.27	0.27	0.25
After 2014	SFH	1.8	0.21	0.21	0.27
	TH	1.8	0.21	0.21	0.27
	MFH	1.8	0.21	0.21	0.27
	AB	1.8	0.21	0.21	0.25

Appendix B. Step-by-step guide that shows how the space heating demand is calculated in the physics-based model

1. Extract one building

The model is a function that works on a per building basis. Therefore, it uses a for loop that iterates over each building to calculate the total space heating demand.

2. Joining the weather data to the building

The preprocessed temperature- and radiation dataset is joined to the building, based on which weather station is closest. By joining the weather data, a single file is created that contains both the hourly weather data and the building attributes, which is convenient to calculate the hourly space heating demand.

3. Transforming spatial data to Numpy arrays

Spatial data takes up a lot of memory and processing power. Therefore, each of the variables in the dataset is converted to a Numpy array.

4. Setting up a default occupancy schedule

To account for the fact that houses are not occupied the entire day, a standard occupancy schedule is applied to all buildings. Similarly to the assumption by Yang et al. (2020), it is assumed in this model that residents are home during 18 and 8 o'clock the next day.

5. Calculation of hourly transmission loss (Q_{tr}) through building elements.

The space heating demand is determined by four main indicators that together form the heat balance. On the one hand these are the transmission loss and ventilation loss and on the other hand, are the internal heat gain and solar heat gain. The formulas that are used in this section of the model, are based on the model by Yang et al (2020). For consistency, the answers to all formulas are stored in kilowatt-hours (kWh). Some formulas are simplified, because not all data is present in the TABULA dataset. The following formula is used to calculate the transmission loss of the building.

$$Q_{tr} = \begin{cases} b \cdot A \cdot U \cdot (T_{int} - T_{ext}) & if\ (T_{int} > T_{ext}) \\ 0 & if\ (T_{int} \leq T_{ext}) \end{cases} \quad \text{(Eq. B.1)}$$

In this formula *b* is an adjustment factor of the transmission loss. For the floor it is 0.5, for other building materials its value is 1. A is the area of the building element. U is the U-value that belongs to the building element. The U-values are assigned according to the building segmentation from the preprocessing steps. The applied U-values are visualized in Appendix A. These parameters are then multiplied by the temperature difference. For calculating the temperature difference, a base temperature of 18 degrees Celsius is used, as this is the standard that the CBS uses to calculate heating degree days (Centraal Bureau voor de Statistiek (CBS), 2015). If the external temperature is higher than the internal temperature, there is no transmission loss.

6. Calculation of hourly ventilation loss (Q_{ve}) of building envelope

V1 : Without volume column, standard infiltration rates TABULA.

$$Q_{ve,t} = \begin{cases} \rho_a c_a \cdot A_{con} \cdot 2.5m \cdot \left(n_{ve,use} + n_{ve,infiltration}\right) \cdot (T_{int,t} - T_{ext,t}) & if\ T_{int,t} > T_{ext,t} \\ 0 & if\ T_{int,t} \leq T_{ext,t} \end{cases} \quad \text{(Eq. B.2)}$$

This formula is used to calculate the ventilation loss of the building. It multiplies the heat capacity of air (ρ), which is 1200 J / (m^3K). This is multiplied with a number of parameters. A_{con} is the conditional floor space, which is all the floor space within a building. The 2.5 meters is the reference room height that is used by the TABULA dataset. Together they estimate the volume of the building. $n_{ve,use}$ and $n_{ve,infiltration}$ are the air change rate by use and the airflow rate by infiltration respectively. In the TABULA dataset they are both set at a constant of 0.4. This set of parameters is then multiplied by the temperature difference.

In model iteration 3: With volume column, standard infiltration rates TABULA

$$Q_{ve,t} = \begin{cases} \rho_a c_a \cdot V \cdot \left(n_{ve,use} + n_{ve,infiltration}\right) \cdot (T_{int,t} - T_{ext,t}) & if\ T_{int,t} > T_{ext,t} \\ 0 & if\ T_{int,t} \leq T_{ext,t} \end{cases}$$ (Eq. B.3)

Because in this thesis 3D data is used for the attributes, it is interesting to check if replacing the A_{con} * 2.5m, by the actual volume attribute of the building improves the performance of the model.

In model iteration 7: With volume column, infiltration paper Todeschi2021

$$Q_{ve,t} = \begin{cases} \rho_a c_a \cdot V \cdot n_{ve,infiltration} \cdot (T_{int,t} - T_{ext,t}) & if\ T_{int,t} > T_{ext,t} \\ 0 & if\ T_{int,t} \leq T_{ext,t} \end{cases}$$ (Eq. B.4)

Yang et al. (2020) uses constants from TABULA to estimate the building infiltration rates. However, in reality infiltration rates depend a lot on the building quality, which relates to when a building was built. Todeschi et al. (2021), use a variation of different infiltration rates. These are visualized in Table B.1 Estimated infiltration rates per building period, according to Todeschi et al. (2021) B.1. By running the model in different ventilation configurations, the most accurate configuration can be selected as the final model.

Table B.1 Estimated infiltration rates per building period, according to Todeschi et al. (2021)

Building period	infiltration rate
Before 1965	0.65
1965 - 1975	0.55
1975 – 1990	0.4
1990 – 2005	0.35
After 2005	0.3

7. Calculation of hourly internal heat gain

$Q_{int} = q_{int} A_{con}$ (Eq. B.5)

Internal heat gain is the heat that is all the heat that is created from sources within the building. These are anthropogenic heat and heat from appliances. Yang et al. (2020) estimates that the internal heat gains per square meter of conditional floor space is 3 W / m^2. This value is also adopted in this model.

8. Calculation of hourly solar heat gain (Q_{sol})

$Q_{sol} = I_{sol}(F_{east} + F_{west}) A_{window} F_{sh}(1 - F_F) F_W$ (Eq. B.6)

Houses are assumed to gain heat through the windows of the building on the eastern and western side. This simplification is applied, because it would be computationally challenging to incorporate incoming solar radiation from all angles during the day. The following parameters are used to calculate the solar heat gain:

- I_{sol} is the incoming horizontal solar radiation. This is the preprocessed KNMI radiation data.

- $F_{east\ /\ west}$ are conversion factors to estimate the incoming solar radiation on windows that are positioned vertically. The factors are 0.69 and 0.68 respectively.
- A_{window} is the total area of the windows.
- F_{sh} is a dimensionless factor that takes shading devices into consideration. It holds a value of 0.6.
- F_F is the frame fraction of the windows, which has a value of 0.3.
- F_W is a correction factor for non-scattering glazing and its value is 0.9.

9. Calculation of the heat balance ratio and the gain utilization factor (η_{gn})

Before the energy demand of the HVAC can be calculated, a dimensionless factor called the gain utilization factor needs to be calculated. The International Organization for Standardization (ISO) (2017) defines the gain utilization factor as a factor that reduces the heat gains, in order to obtain the building energy need for heating. The calculation of the gain utilization factor is dependent on three formulas. Formula B.7 calculates the heat balance ratio. It expresses the ratio between the heat gains and the heat losses. Formula B.8 calculates a time constant (τ), which is used for calculating a reference numerical parameter from ISO 13790 (α_H).

Heat balance ratio

$$\gamma_H = \frac{(Q_{int,t}+Q_{sol,t})}{(Q_{tr,t}+Q_{ve,t})} \quad \text{(Eq. B.7)}$$

The heat balance ratio in equation B.7 is calculated by dividing the sum of the heat gains by the sum of the heat losses.

$$\tau = \frac{C_m A_{con}}{H_{tr}+H_{ve}} \text{ (Eq. B.8)}$$

The time constant is computed to determine the rate of heat transfer within a building. It uses the building's heat capacity ($C_m A_{con}$). For C_m a constant value of 45 $\mathrm{Wh/(m^2K)}$ is used, similarly to Yang et al. (2020). This is then divided by the sum of the heat transfer- and heat ventilation coefficients.

Reference numerical parameter (EN ISO 13790)

$$a_H = a_{H,0} + \frac{\tau}{\tau_{H,0}} \quad \text{(Eq. B.9)}$$

In formula B.9, the reference numeric parameter from ISO 13790 is applied. The value for $a_{H,0}$is a constant with a value of 1, and $\tau_{H,0}$ has a constant value of 15. The time constant (τ), therefore defines the outcome of this formula.

Gain utilization factor

$$\eta = \begin{cases} \frac{1-\gamma_H^{a_H}}{1-\gamma_H^{a_H+1}} & if\ \gamma_H > 0\ and\ \gamma_H \neq 1 \\ \frac{a_H}{a_H+1} & if\ \gamma_H = 1 \\ \frac{1}{\gamma_H} & if\ \gamma_H < 0 \end{cases} \quad \text{(Eq. B.10)}$$

Based on the outcome for the heat balance ratio (γ_H), the formula is chosen to calculate the gain utilization factor (η). If the heat balance ratio is positive, but not one, the first formula is used. If the ratio is exactly one, the second formula is used and if the ratio is negative the last formula is used.

10. Calculation of hourly space heating demand (Q_{nd})

$Q_{nd} = (Q_{tr} + Q_{ve}) - \eta_{gn}(Q_{int} + Q_{sol})$ (Eq. B.11)

When the gain utilization factor has been calculated, the space heating demand of the building can be calculated. The demand is defined as the sum of the heat losses, subtracted by the heat gains which is multiplied by the gain utilization factor.

11. Calculation of hourly space heating demand of space heating system

Without ventilation recovery

$Q_h = \left[\frac{Q_{nd}}{A_{con}} + q_{d,h}\right] e_{g,h} \cdot A_{con}$ (Eq. B.12)

With ventilation recovery

$Q_h = \left[\frac{Q_{nd}}{A_{con}} + q_{d,h} - \eta_{gn}\left(\eta_{ve,rec_{tr}} \cdot Q_{ve}\right)\right] e_{g,h} \cdot A_{con}$ (Eq. B.13)

To find out how much energy the space heating system uses to fulfill the demand of the building, formula B.12 or B.13 is used. Formula B.12 does not take the presence of a ventilation recovery system into account, while formula B.13 does. Both formulas divide the space heating demand over the conditional floor space and then add the expected distribution losses of the space heating system. This is then multiplied by the heating expenditure coefficient of the space heating system. The TABULA dataset assumes that all buildings use a high-efficiency boiler complying with HR107 standards. This means that a constant value of 1.05 is used for all buildings.

Formula B.13 takes into account that newer buildings contain a ventilation heat recovery system $\left(\eta_{ve,rec_{tr}} \cdot Q_{ve}\right)$. This kind of system takes out the residual heat of outgoing warm air and adds this heat to the incoming fresh air (Richardson & Burdett-Gardiner, 2023). By doing this, the space heating system will be used less.

12. Calculating domestic hot water (DHW) demand.

$Q_w = Q_{nd,w} \cdot e_{g,w} \cdot A_{gfa}$ (Eq. B.14)

In the validation dataset, only the total gas usage is provided, which includes consumption for domestic hot water (DHW) purposes, such as showering or cleaning. DHW demand is more sensitive to the occupants of the house and is therefore more challenging to predict. In this physics-based model, the DHW demand ($Q_{nd,w}$) is assumed to be constant at 20.8 kWh/m^{2}a. This value is derived from the study conducted by Flourentzou & Pereira (2021). This parameter is then multiplied by the expenditure factor of the DHW system ($e_{g,w}$) and the ground floor area of the house (A_{gfa}).

13. / 13. Calculating yearly gas space heating demand of building.

$Q_{gas} \frac{\sum_{h=1}^{i=\sim 8760} Q_{h,h} + Q_w}{9.77}$ (Eq. B.15)

The final step is to sum the hourly demand of both the space heating system and the DHW system and divide it by the caloric value of gas in the Netherlands to convert the output from kWh to gas m^3.

Appendix C. The price elasticities developed by Schulte & Heindl (2017), slightly interpolated to match the income classes from the CBS postal code dataset.

household_type	income_class	elasticity
Eenpersoons	0-25	-0.205
Eenpersoons	25-50	-0.313
Eenpersoons	50-75	-0.411
Eenpersoons	75-100	-0.616
Eenouder	0-25	-0.215
Eenouder	25-50	-0.294
Eenouder	50-75	-0.378
Eenouder	75-100	-0.584
Meerpersoons zonder kinderen	0-25	-0.281
Meerpersoons zonder kinderen	25-50	-0.413
Meerpersoons zonder kinderen	50-75	-0.542
Meerpersoons zonder kinderen	75-100	-0.845
Tweeouder	0-25	-0.32
Tweeouder	25-50	-0.463
Tweeouder	50-75	-0.587
Tweeouder	75-100	-0.861

Appendix D. Nomenclature that shows the explanation of all abbreviations used in this thesis.

Abbreviation	**Meaning**
CVRMSE	Coefficient of Variation Root Mean Square Error
RMSE	Root Mean Square Error
NRMSE	Normalized Root Mean Square Error
MAPE	Mean Absolute Percentage Error
NMBE	Normalized Mean Bias Error
MBE	Mean Bias Error
RSS	Residual Sums of Squares
NVF	Normalized Variation Factor
CV	Coefficient of Variation
NMAE	Normalized Mean Absolute Error
R	Correlation
MAD	Mean Absolute Deviation
MRE	Mean Relative Error
HVAC	Heating, Ventilation and Air-Conditioning
ISO	International Standardization Organization
EPC	Energy Performance Contract
WWR	Window-to-Wall Ratio
DHW	Domestic Hot Water
EUI	Energy Use Intensity
TF-IDF	Term Frequency – Inversed Document Frequency
SFH	Single Family Home
TH	Terraced House
MFH	Multi Family Home
AB	Apartment Block

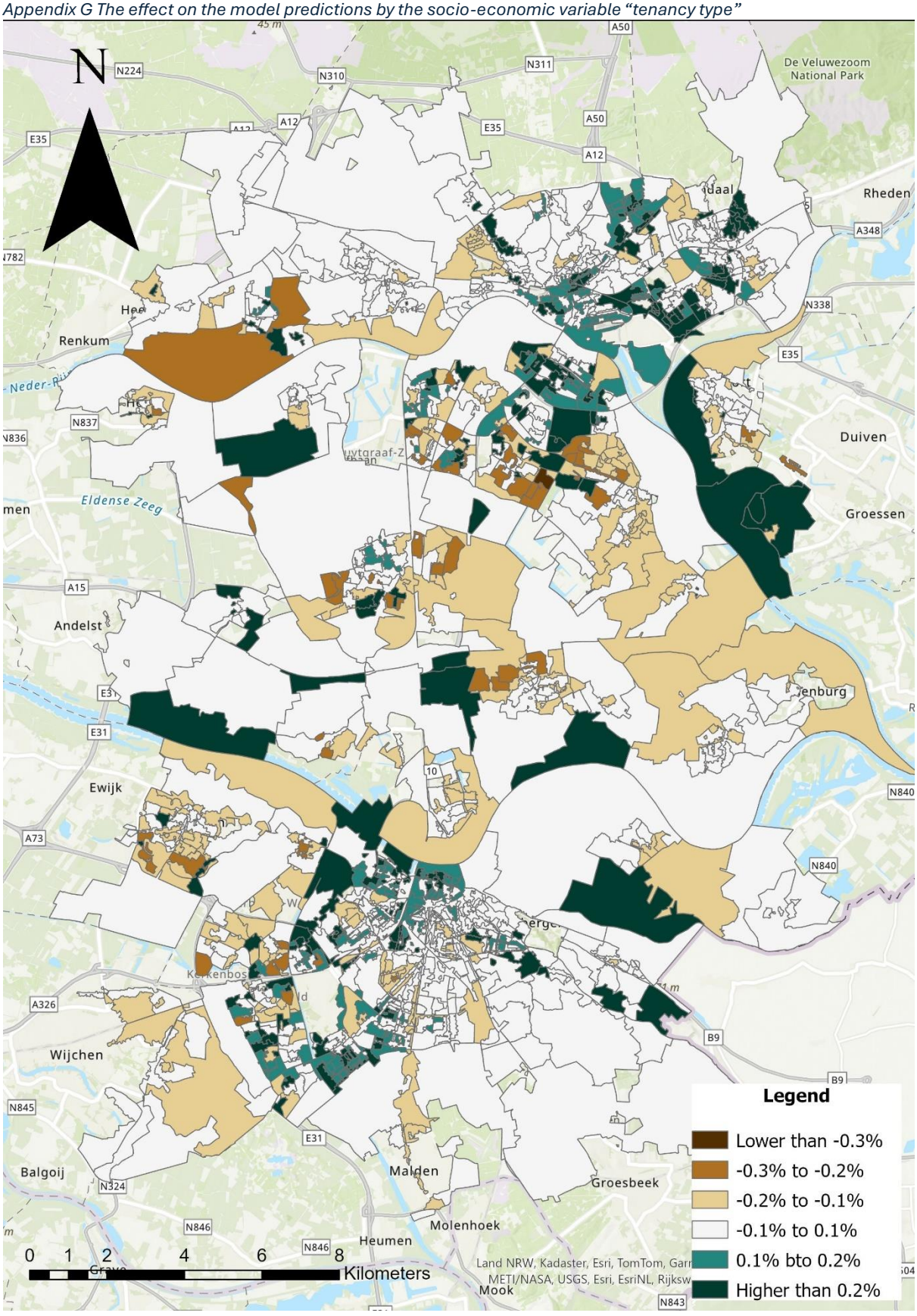
Appendix E. Physics-based models analyzed in the literature review.

Author	**DOI**
J. N. B. Fonseca & Oliveira Panão, 2017	10.1016/j.enbuild.2017.07.059
Verellen & Allacker, 2022	10.3390/su14105769
Eggimann et al., 2019	10.1016/j.apenergy.2018.12.052
Oliveira Panão & Brito, 2018	10.1016/j.enbuild.2018.04.010
Ren et al., 2012	10.1016/j.enpol.2012.06.065
Buttitta & Finn, 2020	10.1016/j.enbuild.2019.109577
Gulotta et al., 2021	10.1016/j.enbuild.2020.110584
Mata et al., 2013	10.1016/j.enbuild.2012.09.037
Schwanebeck et al., 2021	10.3390/en14041029
Sokol et al., 2017	10.1016/j.enbuild.2016.10.050
Perera et al., 2018	10.1016/j.apenergy.2018.04.004
Edtmayer et al., 2023	10.54337/ijsepm.7570
V. Costanzo et al., 2019	10.1016/j.cities.2019.102467
Hedegaard et al., 2019	10.1016/j.apenergy.2019.03.063
Kazas et al., 2017	10.1016/j.apenergy.2017.01.095
Wang et al., 2020	10.1080/19401493.2020.1729862
Prataviera et al., 2021	10.1016/j.renene.2021.03.144
Nageler et al., 2018	10.1016/j.enbuild.2018.09.034
Perwez et al., 2022	10.1016/j.apenergy.2022.119536
Theile et al., 2022	10.1016/j.enbuild.2021.111591
Schwartz et al., 2021	10.1016/j.enbuild.2021.111249
Streicher et al., 2019	10.1016/j.enbuild.2018.12.011
Li & Yao, 2021	10.1016/j.enbuild.2021.110740
Abdelaziz et al., 2021	10.26868/25222708.2021.31009
Mastrucci, Pérez-López, et al., 2017	10.1016/j.enbuild.2017.05.022
Perwez et al., 2022	10.26868/25222708.2021.30586
Buffat et al., 2017	10.1016/j.apenergy.2017.10.041
Veljkovic et al., 2023	10.1016/j.enbuild.2023.113474
Tuominen et al., 2014	10.1016/j.buildenv.2014.02.001
Österbring et al., 2016	10.1016/j.enbuild.2016.03.060
Brøgger et al., 2019	10.1016/j.enbuild.2019.06.054
Todeschi et al., 2021	10.3390/su13041595
Li et al., 2018	10.1016/j.enbuild.2018.03.064

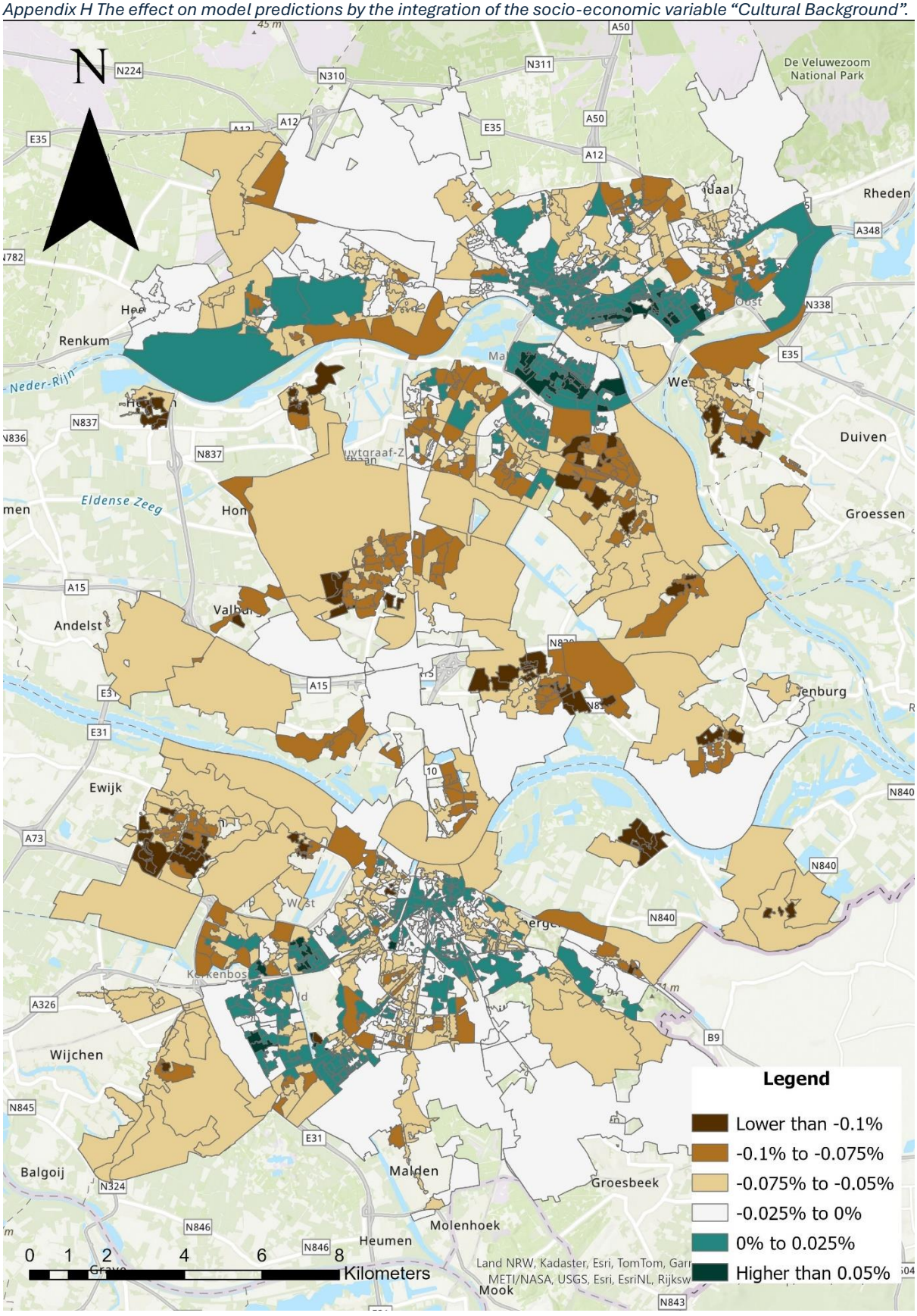
Appendix F Coefficients derived from ridge regression iterations with a variation in input variables.

	Iteration 1		*Iteration 2*		*Iteration 3*		*Iteration 4*		*Iteration 5*		*Iteration 6*	
Feature	**Coeff.**	**P**	**Coeff.**	**P**	**Coeff.**	**P**	**Coeff.**	**P**	**Coeff.**	**P**	**Coeff.**	**P**
WOZ-waardewoning_log	181.73	0.00	216.87	0.00	230.86	0.00	202.19	0.00	183.31	0.00	201.54	0.00
Huishoudgrootte	-21.74	0.65	-102.36	0.00	-117.04	0.00	33.71	0.46	-30.35	0.55	-226.94	0.00
Koopwoning	0.22	0.89	5.31	0.00	5.03	0.00	0.42	0.84	0.24	0.91	6.24	0.00
Huurwoning	-5.86	0.01	NA	NA	NA		-5.89	0.01	-5.81	0.02	NA	NA
Nederlandseachtergond	-2.26	0.02	5.17	0.00	5.40	0.00	2.99	0.00	NA	NA	NA	NA
Westersemigratieachtergrond	-5.74	0.00	NA	NA	NA		-3.34	0.00	-4.45	0.00	-4.46	0.00
Niet-westersemigratieachtergrond	-6.46	0.00	NA	NA	NA		NA	NA	-4.39	0.00	-5.43	0.00
tot 15 jaar_%	13.26	0.45	NA	NA	NA		-4.47	0.63	13.10	0.46	NA	NA
15 tot 25 jaar_%	30.45	0.06	11.07	0.00	NA		17.31	0.02	30.07	0.06	NA	NA
25 tot 45 jaar_%	7.26	0.66	-20.35	0.00	-21.22	0.00	-7.11	0.43	7.14	0.67	-24.14	0.00
45 tot 65 jaar_%	27.22	0.09	9.40	0.00	7.95	0.00	14.90	0.04	27.03	0.10	NA	NA
65 jaar en ouder_%	26.25	0.11	NA	NA	-1.29	0.49	12.18	0.08	25.98	0.11	-3.30	0.06

Appendix G The effect on the model predictions by the socio-economic variable "tenancy type"



Appendix H The effect on model predictions by the integration of the socio-economic variable "Cultural Background".



Appendix I The effect on the model predictions off the integration of the variable age.